Description of document:	Certain Materials Regarding the Total Information Awareness Program (TIA) (2001-2002)
Request date:	11-July-2002
Released date:	18-September-2014
Posted date:	27-October-2014
Note:	For list of records requested see following page
Source of document:	OSD/JS FOIA Requester Service Center Office of Freedom of Information 1155 Defense Pentagon Washington, DC 20301-1155 Fax: (571) 372-0500 FOIA Online Submission Form

Records requested in the original 11-July-2002 request:

1) A copy of any Strategic Plans for DARPA's Information Awareness Office (IAO)
2) A copy of each weekly status report on the activities of the IAO
3) A copy of each monthly status report on the activities of the IAO
4) A copy of each quarterly status report on the activities of the IAO
5) A copy of any DoD or DARPA or IAO documents (including notes, reports, memos, letters, correspondence. etc.) that discuss any matters associated with civil rights, ethical, legal or constitutional protections issues associated with the IAO programs such as Genoa, EELD (Evidence Extraction and Link Discovery), WAE (Wargaming the Asymmetric Environment), TIDES (Trans-lingual Information Detection, Extraction and Summarization), HID (Human Identification at Distance), Bio-Surveillance, and/or any associated or related programs.
6) A copy of any DoD or DARPA or IAO e-mails that discuss any matters associated with civil rights, ethical, legal or constitutional protections issues associated with the IAO programs Genoa, BELO (Evidence Extraction and Link Discovery), WAE (Wargaming the Asymmetric Environment), TIDES (Trans-lingual Information Detection, Extraction and Summarization), HID (Human Identification at Distance), Bio-Surveillance, and/or any associated or related programs.
7) A copy of the agenda and meeting minutes from IAO staff meetings (since Jan. 1, 2002)
8) A copy of Mr. Poindexter's appointment calendar from Feb. 1, 2002 to July 11, 2002.

DEPARTMENT OF DEFENSE
OFFICE OF FREEDOM OF INFORMATION
1155 DEFENSE PENTAGON
WASHINGTON, DC 20301-1155

SEP 18 2014

Ref: 02-F-1690

This responds to your July 11, 2002, Freedom of Information Act (FOIA) request to this office in which you requested eight specific items from the Information Awareness Office, Defense Advanced Research Projects Agency (DARPA). A copy of your request is enclosed for ease of reference. DARPA conducted several searches which located documents partially responsive to your request. For your request items 1-5 and 7, no responsive documents were found. We have now completed our release determination reviews, the results of which are as set forth, below and annotated within the responsive documents (PDF format), loaded on the enclosed CD-ROM disks. Please note that some documents are barely legible. We apologize for the poor quality; however, these documents are the only copies still in existence. We acknowledge that the processing of this request has been protracted and regret the delay of our response. Your patience in this matter is appreciated.

A number of documents were found to be either originated by other government agencies, or which contain primarily equities of these agencies; accordingly, we have transferred these documents, with our recommendations, to these agencies, addresses as follow, for their review and direct response to you.

Armed Forces Health Surveillance Center, 503 Robert Grant Avenue, Silver Spring, MD 20910

Walter Reed National Military Medical Center, 8901 Rockville Pike, Bethesda, MD 20889

CDC/ATSDR, Attn: FOIA Office, MS: D54, 1600 Clifton Road, N.E., Atlanta, GA 30333

Ms. Jan R. Walker, Special Assistant for External Relations, DARPA's Initial Denial Authority and I have determined that the information in some of the grayed-out segment portions within selected pages of the documents on the CD-ROM disks is exempt from release pursuant to 5 U.S.C. § 552 (b)(4) as it is proprietary information, or information obtained from sources on a confidential basis which, if disclosed, could result in competitive harm, or would impair the government's ability to obtain like information in the future. Within one document, Exemption (b)(4) was found to apply to information we deleted as it is protected by copyright. This information includes a cartoon, a quote from a newspaper, and a quote from a book, items used

solely in an internal DARPA presentation for which permission was not obtained for public release. Other information, within a number of documents, was determined to be exempt from release in its entirety pursuant to 5 U.S.C. § 552 (b)(5) as the information is deliberative in nature, part of the decision making process, and contains subjective evaluations, opinions, and recommendations that, if released, would inhibit the decision making process. A total of 56 pages have been withheld. Some additional information was found to be exempt and withheld from release pursuant to 5 U.S.C. § 552 (b)(6) as its release would constitute a clearly unwarranted invasion of the personal privacy of individuals. Lastly, some information has been withheld pursuant to 5 U.S.C. § 552 (b)(3), which applies to information specifically exempted by a statue establishing particular criteria for withholding. In this instance, the statues are 10 U.S.C. § 424 for protection of organizational and personnel information for DIA, NRO and NIMA, and 50 U.S.C. § 403(g), Section 6 of the CIA Act of 1949, pertaining to CIA functions and information. Accordingly, all of the cited information is withheld pursuant to 5 U.S.C. § 552 (b)(3), (4), (5) and (6).

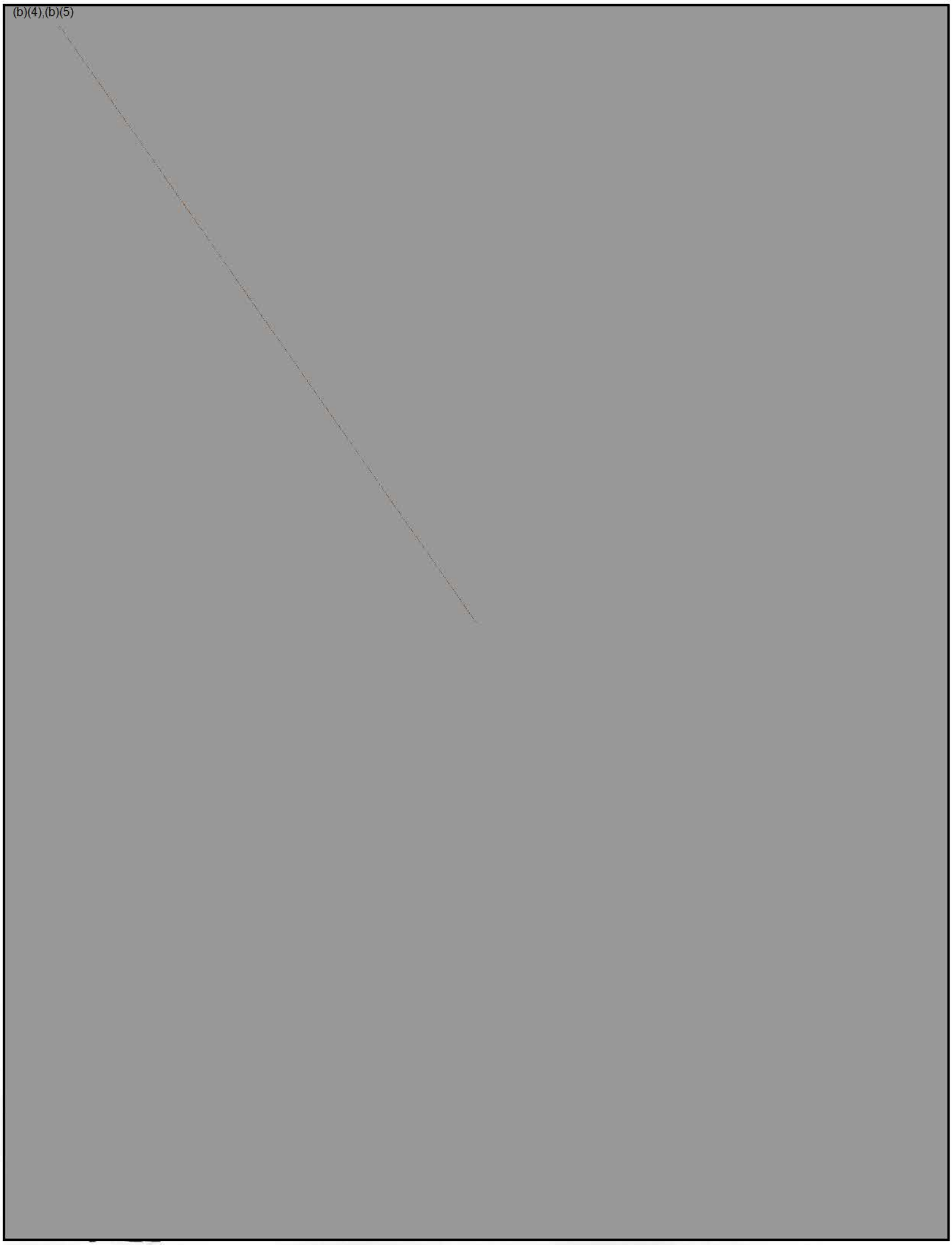
If you are not satisfied with this action, to include the partial no record response, you have the right to appeal to the appellate authority, the Director of Administration, Office of the Secretary of Defense, by writing directly to OSD/Joint Staff Freedom of Information, ATTN: Appeals Office, 1155 Defense Pentagon, Washington, D.C. 20301-1155. Your appeal must be postmarked within 60 calendar days of the date of this response. Alternatively, you may also submit your appeal electronically, within 60 calendar days of the date of this response, at the following link: http://pal.whs.mil/palMain.aspx. Your appeal should cite case number 02-F-1690 and be clearly marked "Freedom of Information Act Appeal."

Sincerely,

FOIA Analyst

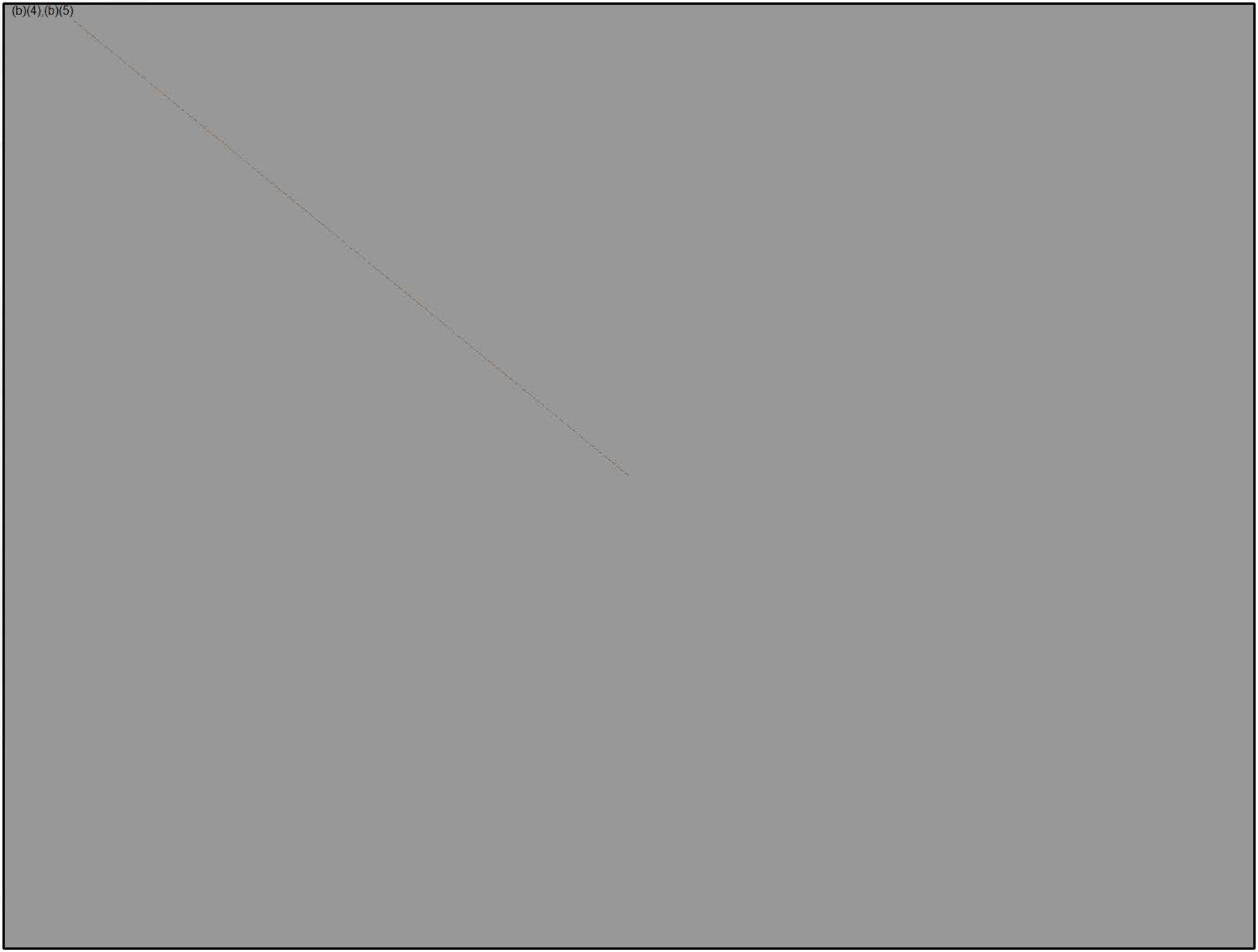
for Will Kammer
Chief

Enclosures:
As stated


(b)(4),(b)(5)


(b)(4),(b)(5)


(b)(4),(b)(5)


(b)(4),(b)(5)

Improving Security and Privacy

Strategies and technologies for our Brave New World

Lt Col Doug Dyer, PhD

DARPA/IAO

Impetus

During the 1980's, we believed that we could achieve a level of information systems security by allowing systems to remain stove-piped... But security is orthogonal to interoperability!

Isn't it possible that information technology can be used to simultaneously improve our defenses against terrorists and personal privacy?

Technology forces change

Computer Hardware

- **Processor**
  - **logic capacity: about 30% per year**
  - **clock rate: about 20% per year**
- **Memory**
  - **DRAM capacity: about 60% per year (4x every 3 years)**
  - **Memory speed: about 10% per year**
  - **Cost per bit: improves about 25% per year**
- **Disk**
  - **capacity: about 60% per year**
- **Networks**
  - **capacity: about 70% per year**

Video

- **Government Video**
  - **2002: 2M video cameras**
  - **2006: 6M video cameras**
- **Business Video**
  - **???**
- **Personal Video**
  - **???**

Software

- **Speech-to-Text**
  - **Word-error rate decreasing 5% per year**
  - **Faster**
- **Web**
  - **Doubling in size periodically**
  - **XML**
  - **Semantic Web**
- **Data mining**
- **Pattern recognition technologies**

It's not about thinking `outside the box,' it's about recognizing the new box.
-- Paraphrased from Josh Lederberg

① 

② A→ B→ C …

Advantages of the Brave New World

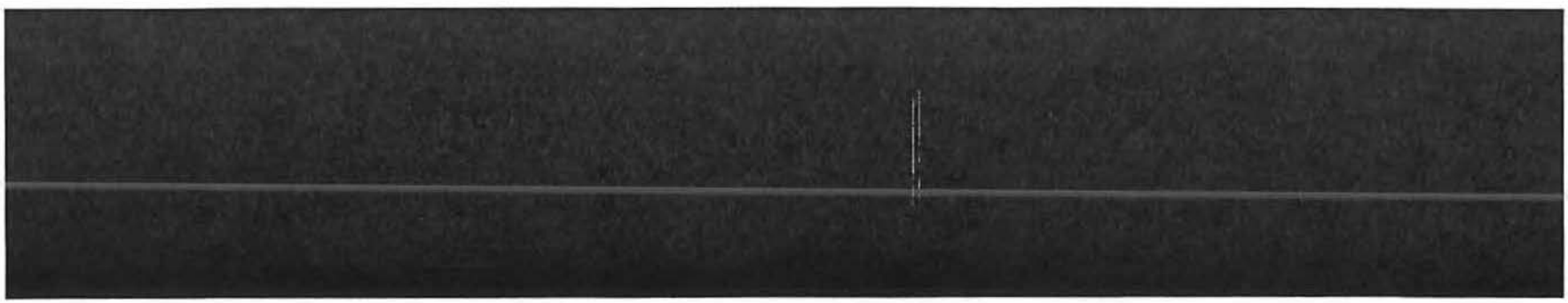
- **Artificial memory**
  - Digital files – better than paper for indexing, storing, duplicating, sharing
  - Digital histories
    - More complete accounting and credit/blame assignment
    - Process improvement
- **Auto-Coordination… of amazing economic utility**
  - People you deal with will have access to your goals, know your constraints
  - You will know all the available vendors, products, services; pros and cons
  - Price knowledge will be universal, and markets will be efficient
- **Identity and profiling**
  - Avoid invasive security checks and investigations
  - Identity theft is impossible
- **Tracking Location, and Activities**
  - Know where your kids are and what they're up to…

Technologies for protecting privacy

- **"Anonymization"**
  - Universally-Unique Identifiers to protect identity, but enable access via subpoena (or other legal authority)
- **Privacy Filters**
  - Much information is not relevant to protecting the US from foreign terrorists
    - Example or irrelevant info: Your medical record
    - Example or relevant info: Statistical knowledge of over-the-counter drug sales
  - Models dictate the information required
  - Filters can also be used in the context of law or policy (e.g., no collection on US citizens)
- **Expunging Agents**
  - Delete information that slips through filters or can be inferred to be irrelevant based on aggregate data
  - Can be used to adjust for changing laws and policy

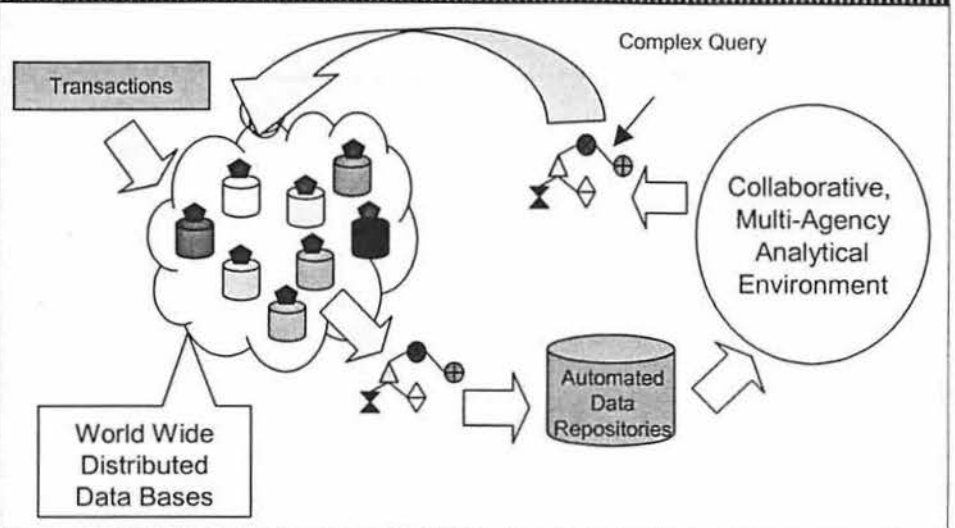
Technologies for protecting privacy

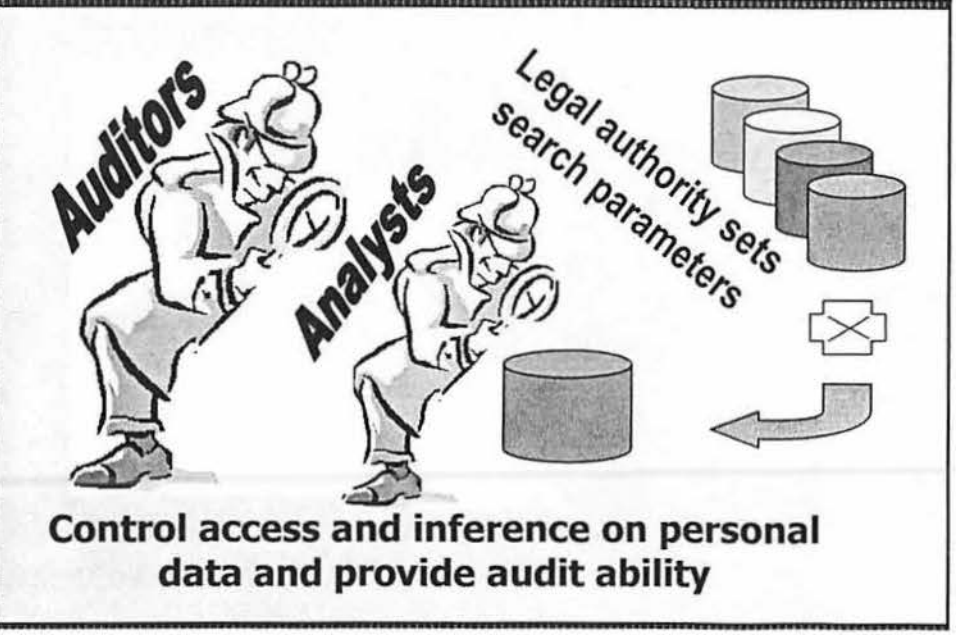
- **Information integration and partitioning** (selective revelation)
  - Access to information based on the role of the requestor
    - Example: Your doctor has access to your medical record but your insurer does not
    - Currently we release information in an ad hoc way without consistent policy
    - Current information systems are not connected, and connecting them without addressing this issue will result in inconsistent access
- **Multi-level security technologies**
  - For structured data, this is less of a problem
- **Temporal databases and immutable accounting**
  - Guarantees an audit trail for all data updates and accesses
  - Allows enforcement of the expectation that personal information won't be misused



Databases and Privacy

- **Genisys**
  - Creating ultra-large repositories populated with transactional & other data to facilitate collaborative analysis and sharing
  - Focusing on integrating and restructuring existing legacy DBs with new scalable distributed repositories and new info feeds
- **Privacy Initiatives**
  - Technology to render data anonymous & re-identify individuals as authorized
  - Security with Privacy ISAT summer study
- **Several Complimentary Seedling Projects**
  - Developing DB of synthetic transactions
    - Billions of transactions from sim model or seeded from real-world commercial data
    - Use for initial development & experiments
    - Insert red team transactions in experiments
  - Inventory of existing transactional DBs
  - Database appliance concept:
    - Query mechanisms for access control, legal authority, releasability, audit trail, etc



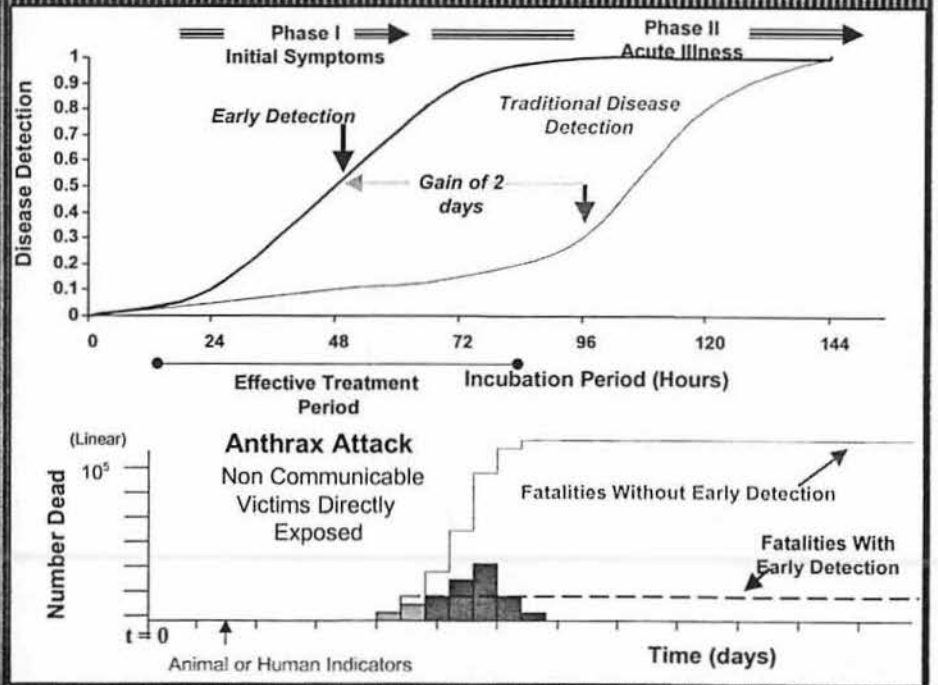


Control access and inference on personal data and provide audit ability

BioSurveillance

Goal: Develop advanced analytics & novel data for early warning

- First Sign of Biological Attack May Be Patients Falling Ill
- BioS Monitors Animal, Drugstore, & Other Novel Data for Alert
- Focuses on Clinical Data for Early Confirmation of Covert Attack
  - GOAL: Identify 1st Cases in Time to Treat, Save > Half the Potential Killed
- Operational Prototype in National Capital Area
  - Used for Biosurveillance at the Presidential Inaugural
- Alternative Prototype Being Deployed Norfolk
  - Intended to Assist Atlantic Fleet Afloat & Ashore
- University of Pittsburgh RODS System Advanced
  - Developed by AHRQ
  - Reviewed by POTUS
  - Deployed to Olympics
  - Adding Novel Data Sources
  - Developing Analytics
  - New BioWar Simulation
- Automating Data Anonymization to Ensure Privacy.



Data and Privacy

Privacy

- **Privacy is a key enabling concept of the Biosurveillance Program**
- **Privacy of Records Must Be Assured in Order to Safeguard Our Data Sources**
- **Medical Record Privacy Must Be Ensured to Comply with HIPAA**
- **De-Personalization Is NOT Equal to Anonymization!**
  - If Unique Identifiers Can Be Derived Upon Aggregation of the Data and Analysis
- **Privacy Must Be Automatically Assured, Preferably Before Data Leaves Originator**
- **However, Neutralizing Data excessively reduces Its Value Significantly**
  - May Especially Neutralize Disease Clustering Capability (for Local Outbreaks)
- **We Must Create "Smart" Anonymization that Is Automatic, Privacy-Assured, and Content-Laden!**
  - Our Research May Support Privacy technology for Wider Community as Well

Domain Data Characteristics

- Some Data Is Limited
  - Little Actual Data on Epidemiology of Biological Attacks
    - Simulations May Be Assumption Driven
  - Limited Data on Disease Progression of Pathogens of Interest
    - Unethical to Experiment (Further) in Humans
    - Little Data on Disease Prodromes in Possible Sentinel Animals
- Some Data Sources Are Inconsistent
  - Often Relies on Hand Entry
  - Lagging Data, Incomplete
  - Data Sources Are Voluntary, May Incur Expenses in Delivery
- Excess and Shifting Data Due to Disease Prevalence Background
- Rigorous Need for Data Privacy
  - Data Sharing Positively Discouraged

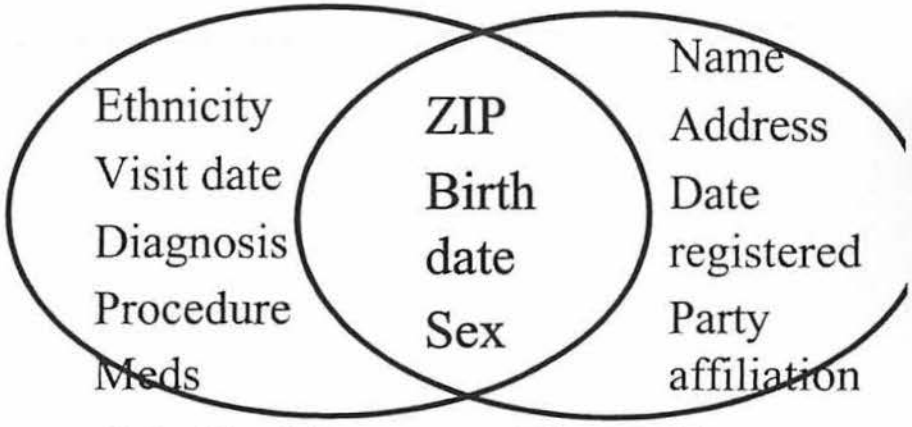
Privacy Concerns from a Legal Perspective

- Fourth Amendment of the Constitution
- Compelling state of interest
- Reasonable expectation of privacy
- Doctor-patient relationship
- Additional uses beyond direct surveillance, leading to class action lawsuits
- Legal authority to share data (grocer to Pitt) and also further secondary sharing (Pitt to others)
- Decisions made in ignorance regarding identifiabiltiy of the released data

Data Privacy

Looks anonymous, but ...

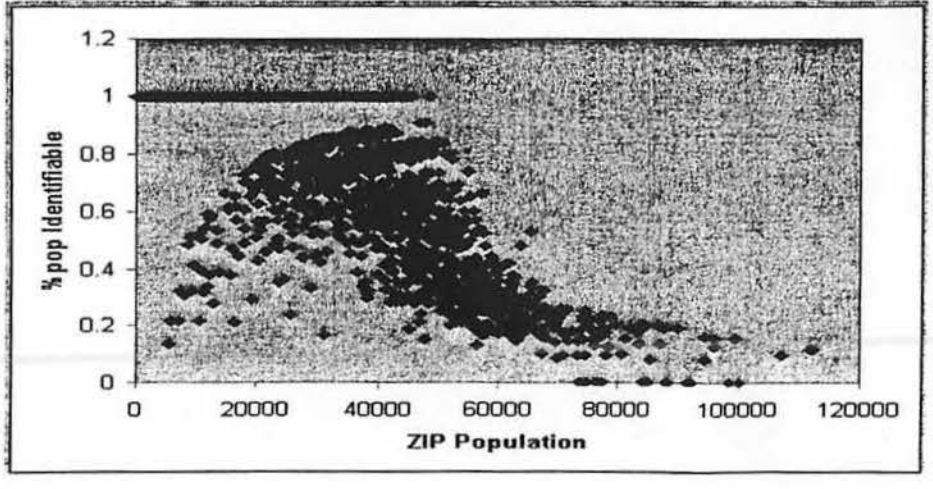
ZIP	Birth	Sex	Ethnicity
33171	7/15/71	m	Caucasian
02657	2/18/73	f	Black
20612	3/12/75	m	Asian



Privacy Concerns

- Privacy Critical to Program: HIPAA, Integrity
- Must Be Assured Automatically
- Neutralizing Data Reduces Its Value Greatly Eliminates Disease Clustering Capability
- Creating "Smart" Anonymization that Is Automatic, Privacy-Assured, and Content-Full

{*date of birth, gender, 5-digit ZIP*} uniquely identifies 87.1% of USA pop.



Data Privacy Technology

We have computational technology to address Data Privacy issues and are continually developing new technology. Here are examples:

1. Identifiability server:
   determines the identifiability of data.

2. Datafly, *k*-Similar:
   Computational systems that render data sufficiently anonymous; adheres to HIPAA.

3. Reasonable cause control:
   Identifiability of provided data automatically adjusted to match status of detection system.

Uses of BioWar

Realistic and wide range of data for detection algorithms to learn on

Assessing potential cost/benefit of different data

Planning

- Assessing potential impact of different alert, privacy, or containment policies

Intelligence

- Useful for what if analysis to plan given possible attack scenarios

Training

- Displaying what to look for from early detection standpoint, order in which indicators will occur, severity of indicators

Evaluation and Metrics

- **Historical Data Comparison for Natural Outbreaks – Ground Truth**
  - Leverage Annual Influenza Outbreak Historical Records for Indicators
  - Use NCA Scarlet Fever Outbreak for Clustered Proxy
  - Independent examples from developers. Same cities / different cities
- **Evaluate Disease Models to Differentiate Military Pathogens from Flu**
- **Expert Panel Review and Validation**
- **Develop ROC Curves Using Simulated Data**
- **Specificity vs Sensitivity Analyses**
- **Identify Existing Animal Disease Progression Models**
- **Privacy/Anonymity Review**
- **Simulation**
- **Red Team?**

4. System Evaluation

System Component	*Metrics & MOEs*	*Evaluation Method*
(a) Data Sources	-Timeliness	-Measure data time-late statistics
	- Utility/value	-Sensitivity analysis using simulation to measure effect on detection and classification performance
	- Collection effectiveness	- Evaluate complexity, automation, dependability, logistic support, security/ privacy (Analysis)
(b) Security & Privacy Protection	-Disclosure control effectiveness	-Measure degree of anonymity
	-Effect of anonymity on det/class performance	-Simulation vs degrees of anonymity
(c) Complete System	-Response time -Probability of detection and correct ID -Probability of false alarm	- Natural background/disease events with & - without simulated outbreaks
	-Utility and User acceptance	-Questionnaire/ review board

V-S Phase 1.5 Evaluation -- Aug '02

Concept and design

- Includes promising non-traditional data sources
- Supports operational readiness
- Satisfactory to local, regional, and state public health officials

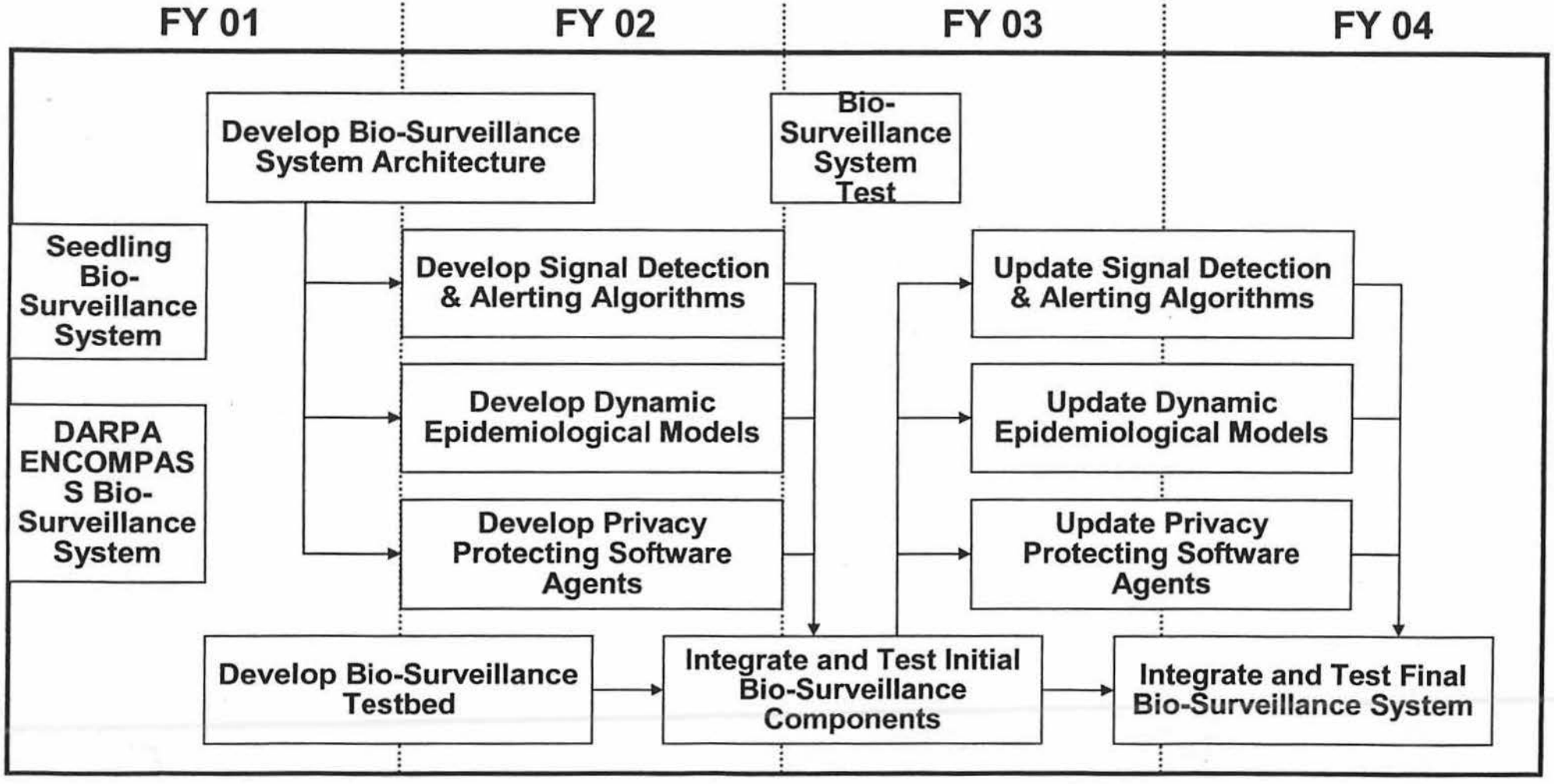
Progress on prototype and emulation

- Selected components functional
- Plausible emulation of additional components
- Disease models implemented for emulation

Real-world data acquisition

- Effectively utilizes existing data sources
- Adaptable to nascent data sources
- Maintains privacy and confidentiality

Bio-Surveillance Program Plan



Goal: What DoD Has at the End

- Two Prototype BioSurvellance Systems in DoD Cities
  (National Capital Area and Norfolk)
  - Near-term effort for National Capital Area
    - Possible Expansion to Other Cities, dependent on $ and data availability
- Knowledge of Effective Data Sentinels (i.e., what data are useful/effective indicators/predictors)
- Inference Algorithms for Detection of Potential Bio-Attacks
- Privacy Protection Algorithms
- Disease Progression Simulations

ALL TECHNOLOGY TRANSFERABLE TO OTHER GOVERNMENT AGENCIES

~~For Official Use Only/Originator Controlled~~

Issues

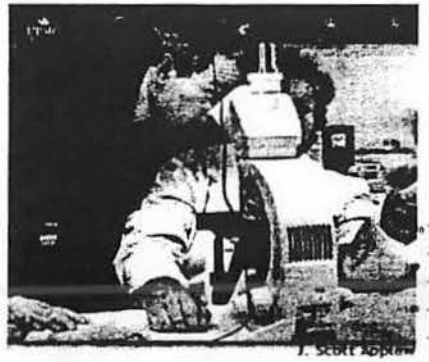
DARPA Bio-Surveillance Program

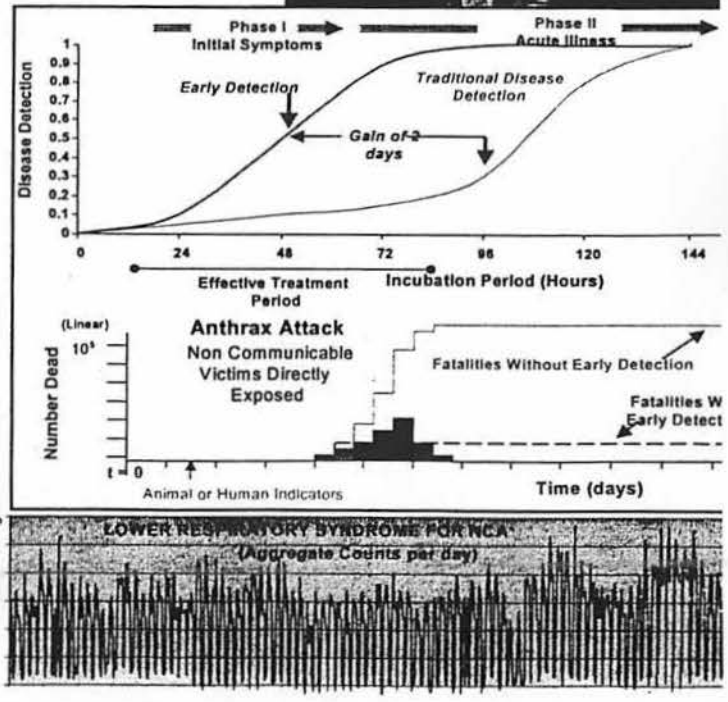
- Not Enough Funds to Continue All Four Performers
  - Probable Neck Down of Prototype Systems in 18 Months
- Data Sensitivity
  - Non-Traditional Data Sources Prefer to Keep a Low Profile, High Privacy
- Sentinel Animals Show High Potential Utility
  - However, Disease Manifestation and Relevance to Humans Not Well Characterized
- Validation
  - Challenging Since Actual Military-Type Pathogen Outbreaks Rare

Biosurveillance Program

Develops Advanced Analytics & Novel Data for Early Warning

- First Sign of Biological Attack May Be Patients Falling Ill
- BioS Monitors Animal, Drugstore, & Other Novel Data for Alert
- Focuses on Clinical Data for Early Confirmation of Covert Attack
  - GOAL: Identify 1st Cases in Time to Treat, Save > Half the Potential Killed
- Operational Prototype in National Capital Area
  - Used for Biosurveillance at the Presidential Inaugural
- Alternative Prototype Being Deployed in Norfolk
  - Intended to Assist Atlantic Fleet Afloat & Ashore
- University of Pittsburgh RODS System Advanced
  - Developed by AHRQ
  - Reviewed by POTUS
  - Deployed to Olympics
  - Adding Novel Data Sources
  - Developing Analytics
  - New BioWar Simulation
- Automating Data Anonymization to Ensure Privacy.



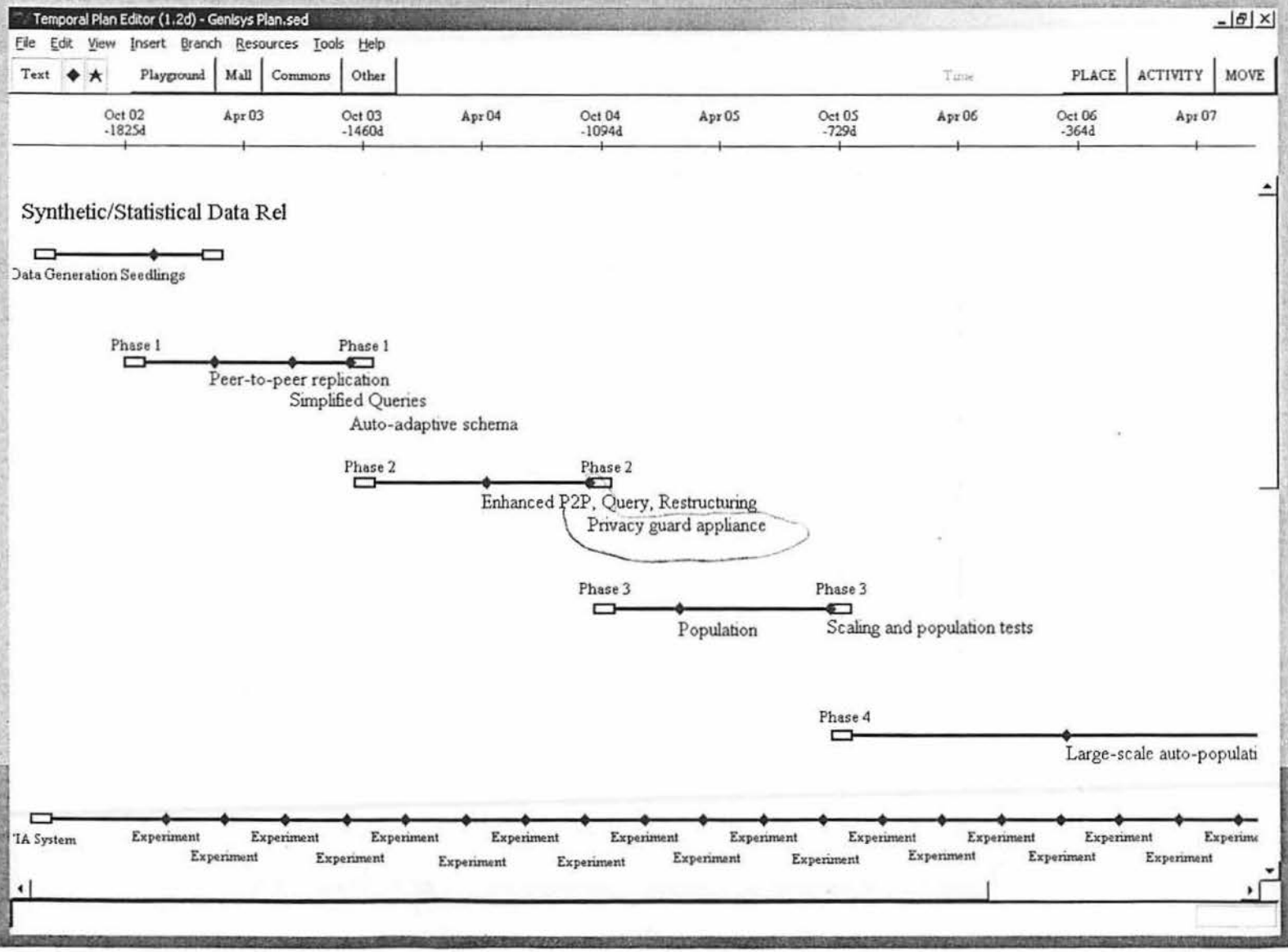


Synthetic and Anonymized Data

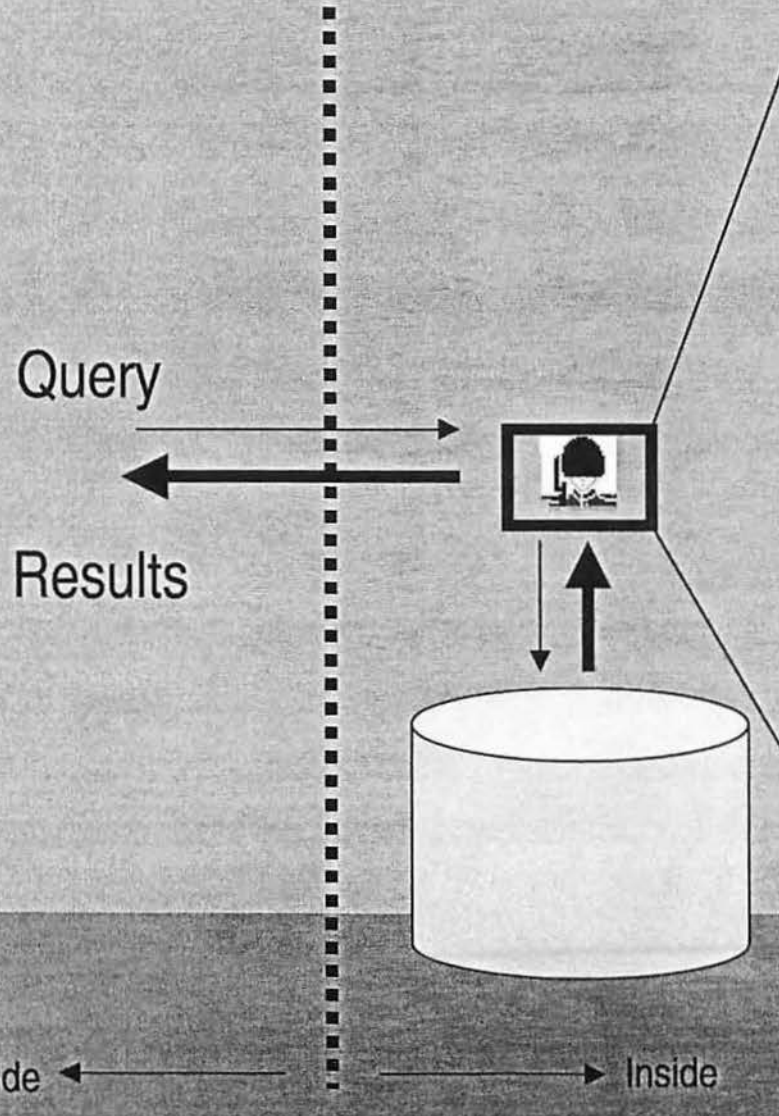
Experimental infrastructure

- MAYA Seedling: Create synthetic transaction data grounded in an accurate model of the world
  - "Vanilla world"
- IET Seedling: Create synthetic transaction data that looks anomalous, but is in fact, benign
  - "French vanilla world"
- Acxiom Seedling: Project an identity-masked set of transaction data from real commercial sources
  - "Real world minus names/address identifiers"
  - Could act as a seed for synthetic data generation
  - Can be used in a mix-and-match manner with synthetic data

Genisys/TIA Roadmap



Privacy Guard Appliance



Trusted Appliance:

Auditable guard/accounting device

- Access control
- Publishable business rules
- Complete accounting
  - Accessors
  - Queries
  - Results
  - Mix-and-match according to policy
  - Multiple methods of ensuring third-party access to accounting
    - **Fan-out publication**

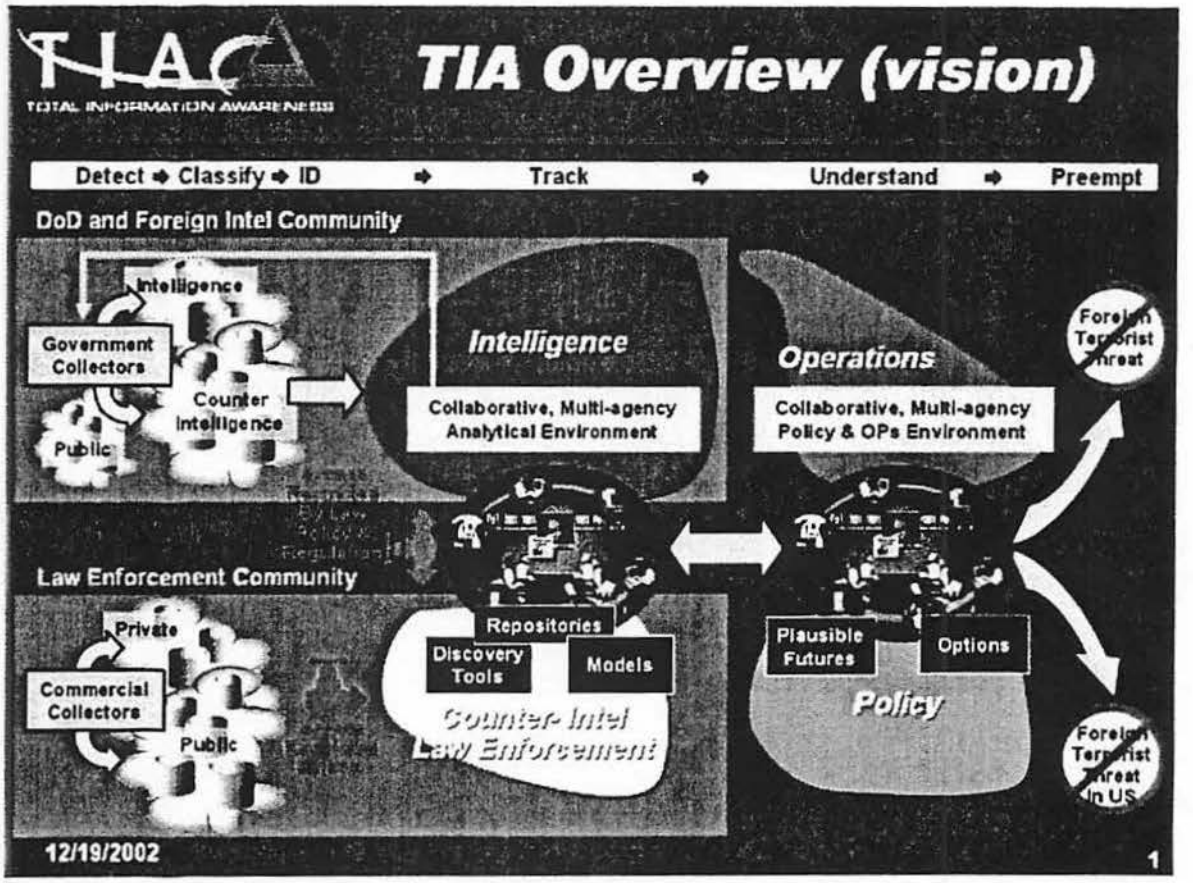
INFORMATION AWARENESS OFFICE

DARPA

Home | News | Programs | Solicitations

TOTAL IN

Total Information Awareness (TIA) System



Program Objective:

The Total Information Awareness (TIA) program is a FY03 new-start program. The goal of the Total Information Awareness (TIA) program is to revolutionize the ability of the United States to detect, classify and identify foreign terrorists – and decipher their plans – and thereby enable the U.S. to take timely action to successfully preempt and defeat terrorist acts. To that end, the TIA program objective is to create a counter-terrorism information system that: (1) increases information coverage by an order of magnitude, and affords easy future scaling; (2) provides focused warnings within an hour after a triggering event occurs or an evidence threshold is passed; (3) can automatically queue analysts based on partial pattern matches and has patterns that cover 90% of all previously known foreign terrorist attacks; and, (4) supports collaboration, analytical reasoning and information sharing so that analysts can hypothesize, test and propose theories and mitigating strategies about possible futures, so decision-makers can effectively evaluate the impact of current or future policies and prospective courses of action.

Program Strategy:

The TIA program strategy is to integrate technologies developed by DARPA (and elsewhere as

appropriate) into a series of increasingly powerful prototype systems that can be stress-tested in operationally relevant environments, using real-time feedback to refine concepts of operation and performance requirements down to the component level. The TIA program will develop and integrate information technologies into fully functional, leave-behind prototypes that are reliable, easy to install, and packaged with documentation and source code (though not necessarily complete in terms of desired features) that will enable the intelligence community to evaluate new technologies through experimentation, and rapidly transition it to operational use, as appropriate. Accordingly, the TIA program will work in close collaboration with one or more U.S. intelligence agencies that will provide operational guidance and technology evaluation, and act as TIA system transition partners.

Technically, the TIA program is focusing on the development of: 1) architectures for a large-scale counter-terrorism database, for system elements associated with database population, and for integrating algorithms and mixed-initiative analytical tools; 2) novel methods for populating the database from existing sources, create innovative new sources, and invent new algorithms for mining, combining, and refining information for subsequent inclusion into the database; and, 3) revolutionary new models, algorithms, methods, tools, and techniques for analyzing and correlating information in the database to derive actionable intelligence.

Planned Accomplishments:

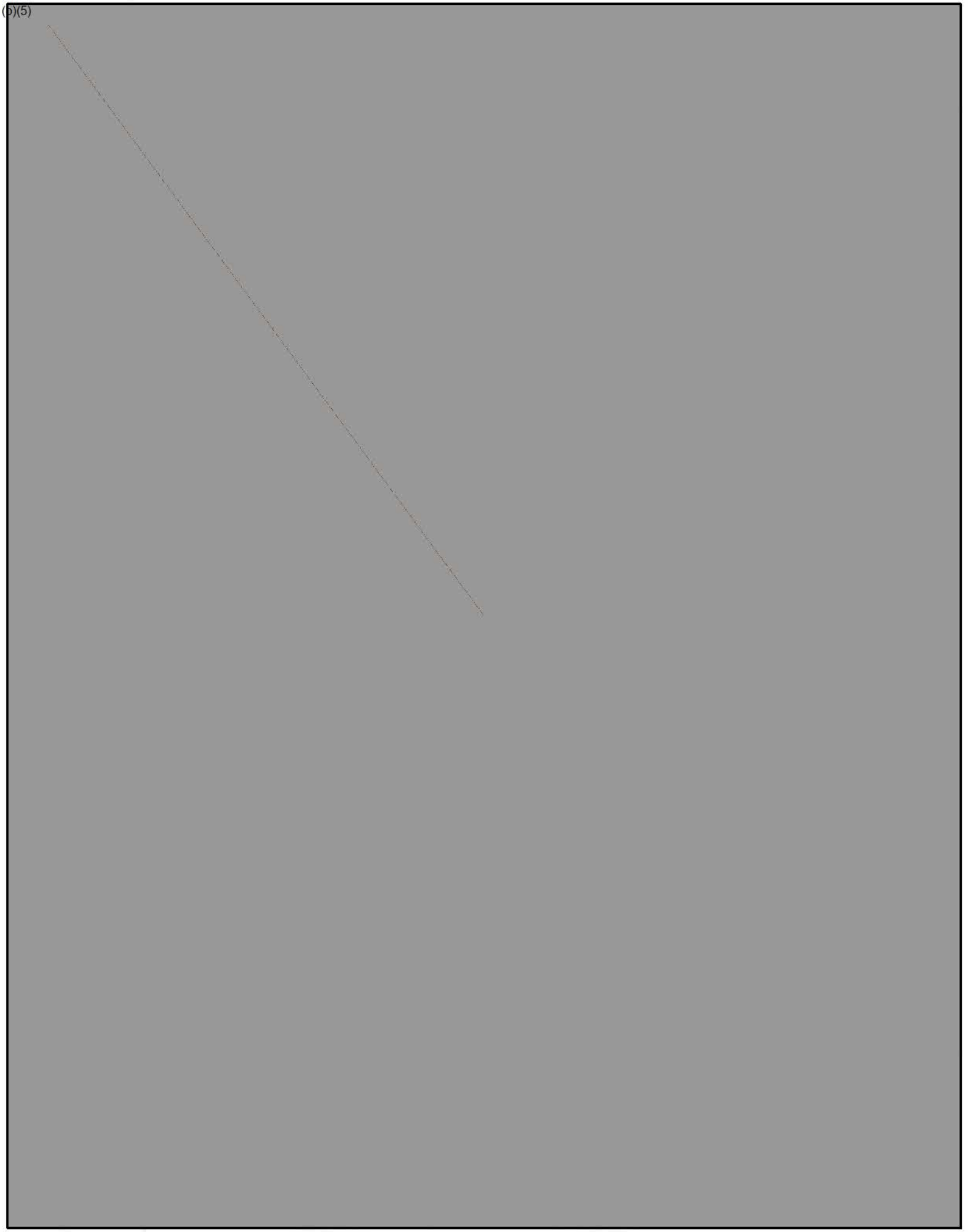
- TBA

More Information

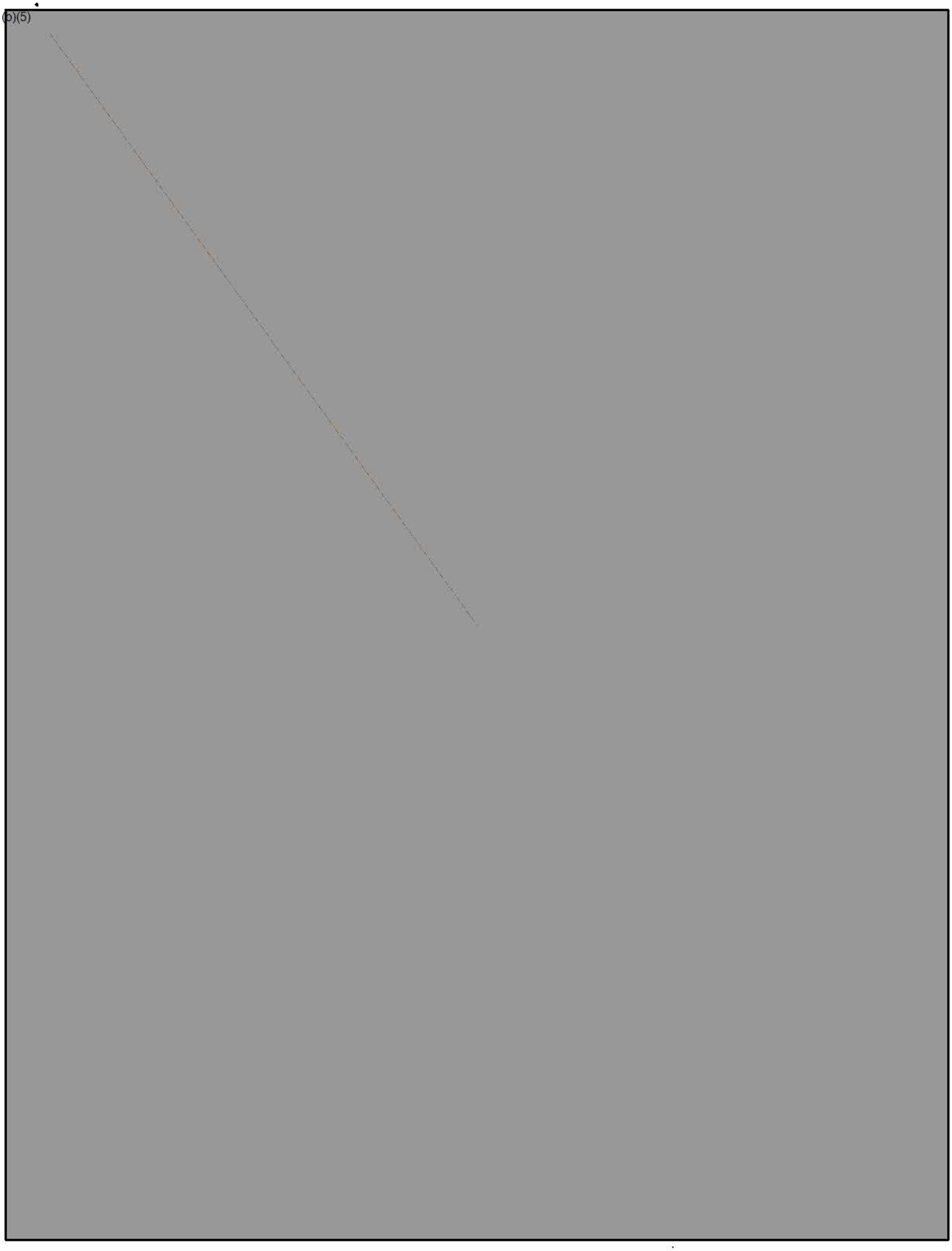
- **More Information**

Home News Programs Solicitations

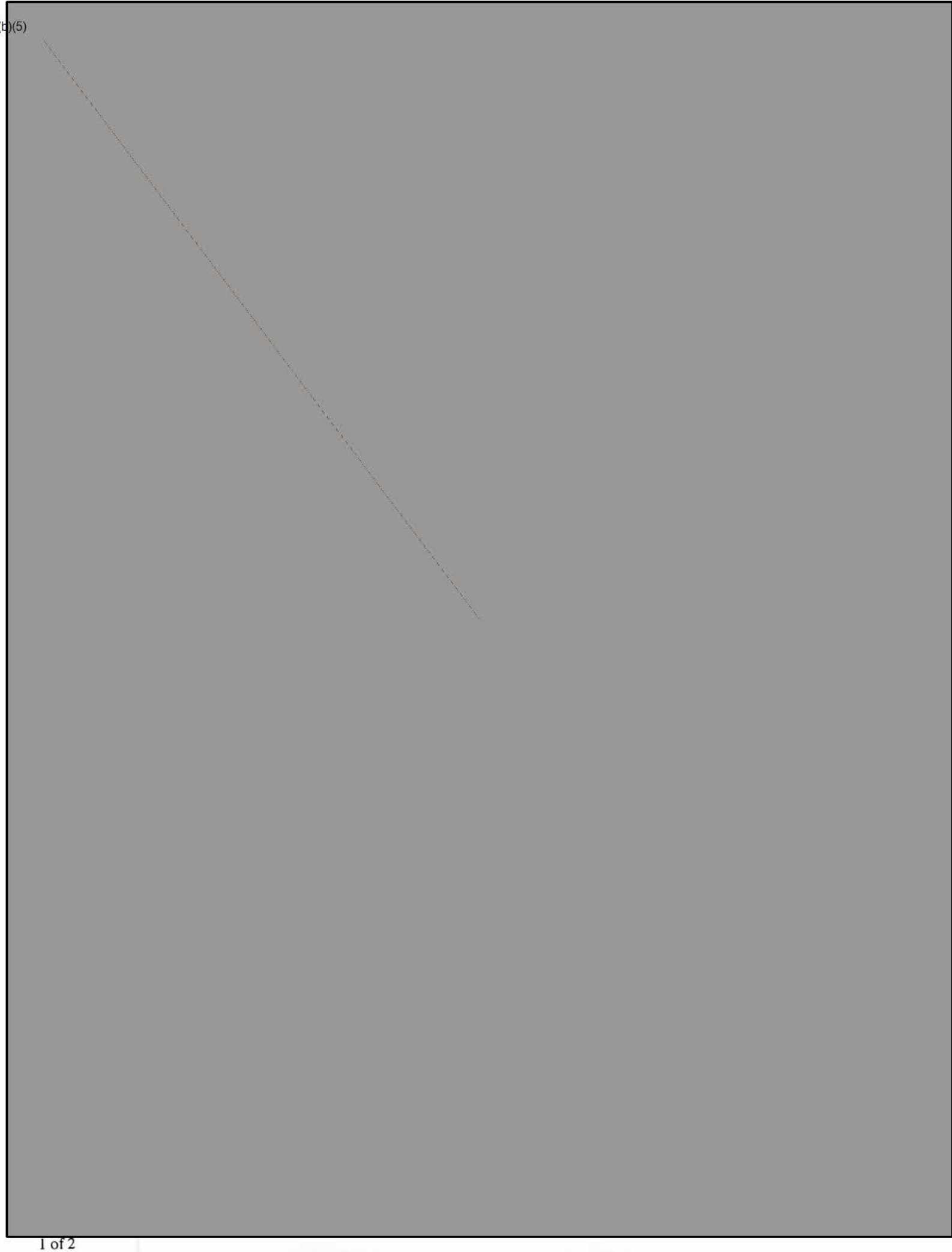
(b)(5)

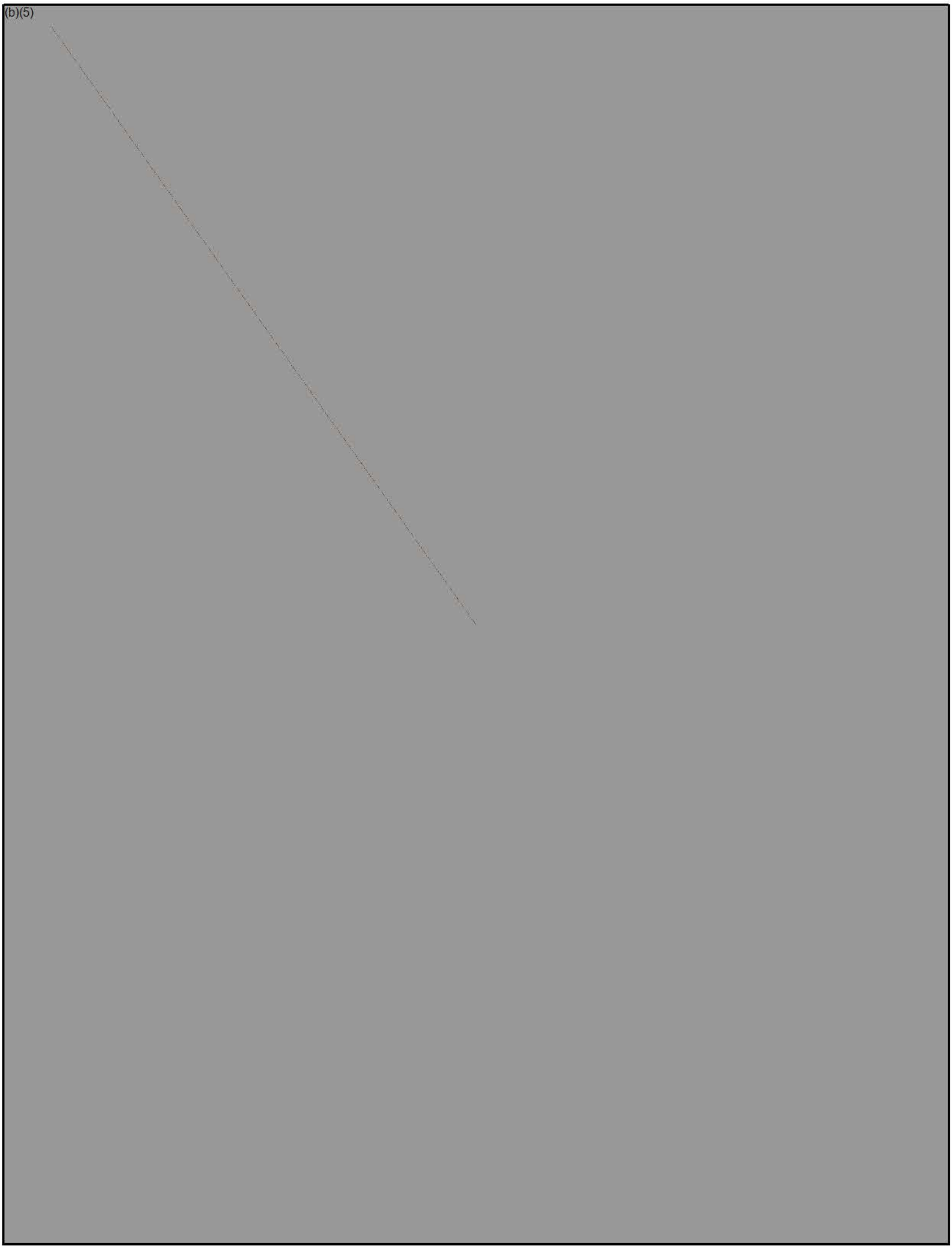


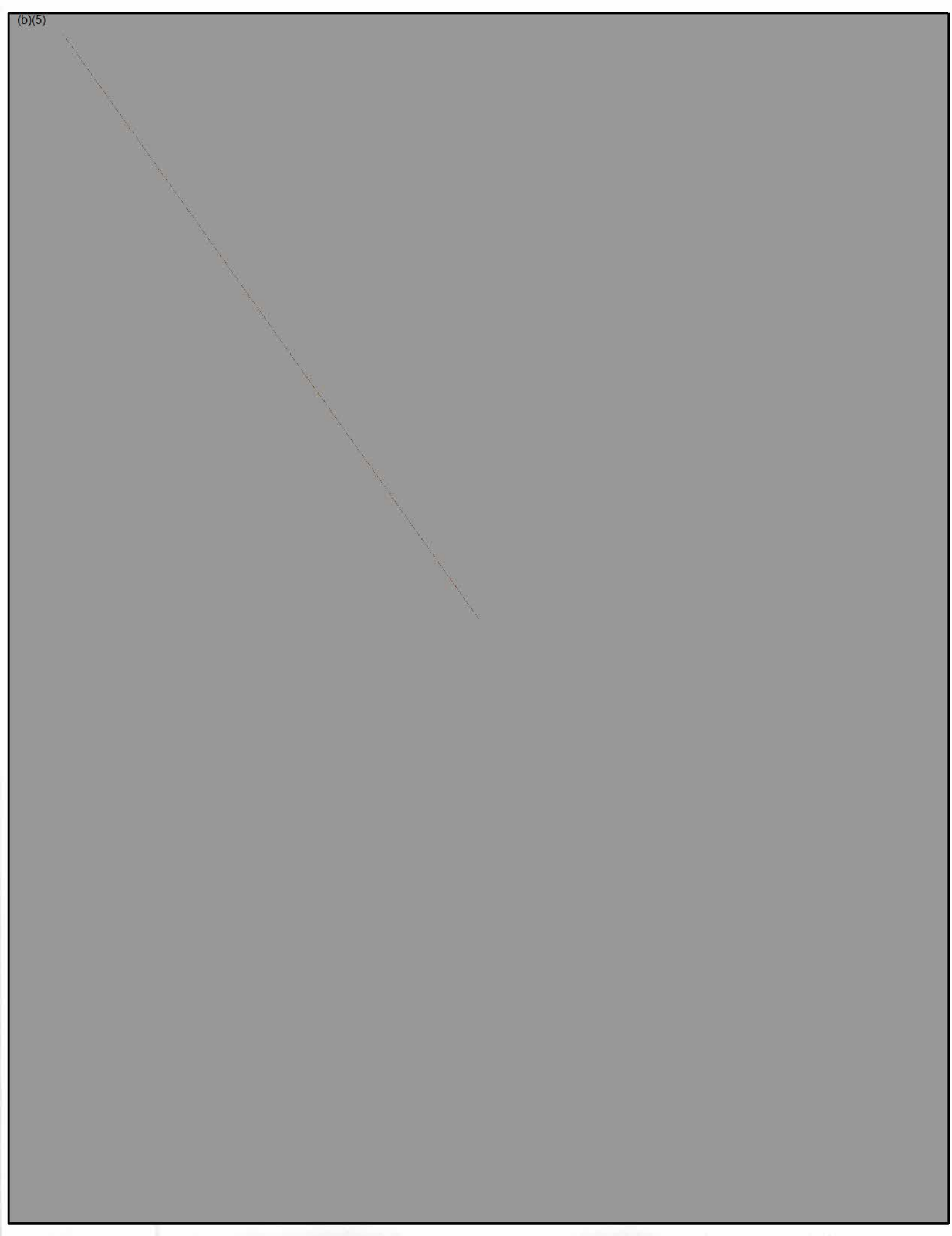
(b)(5)

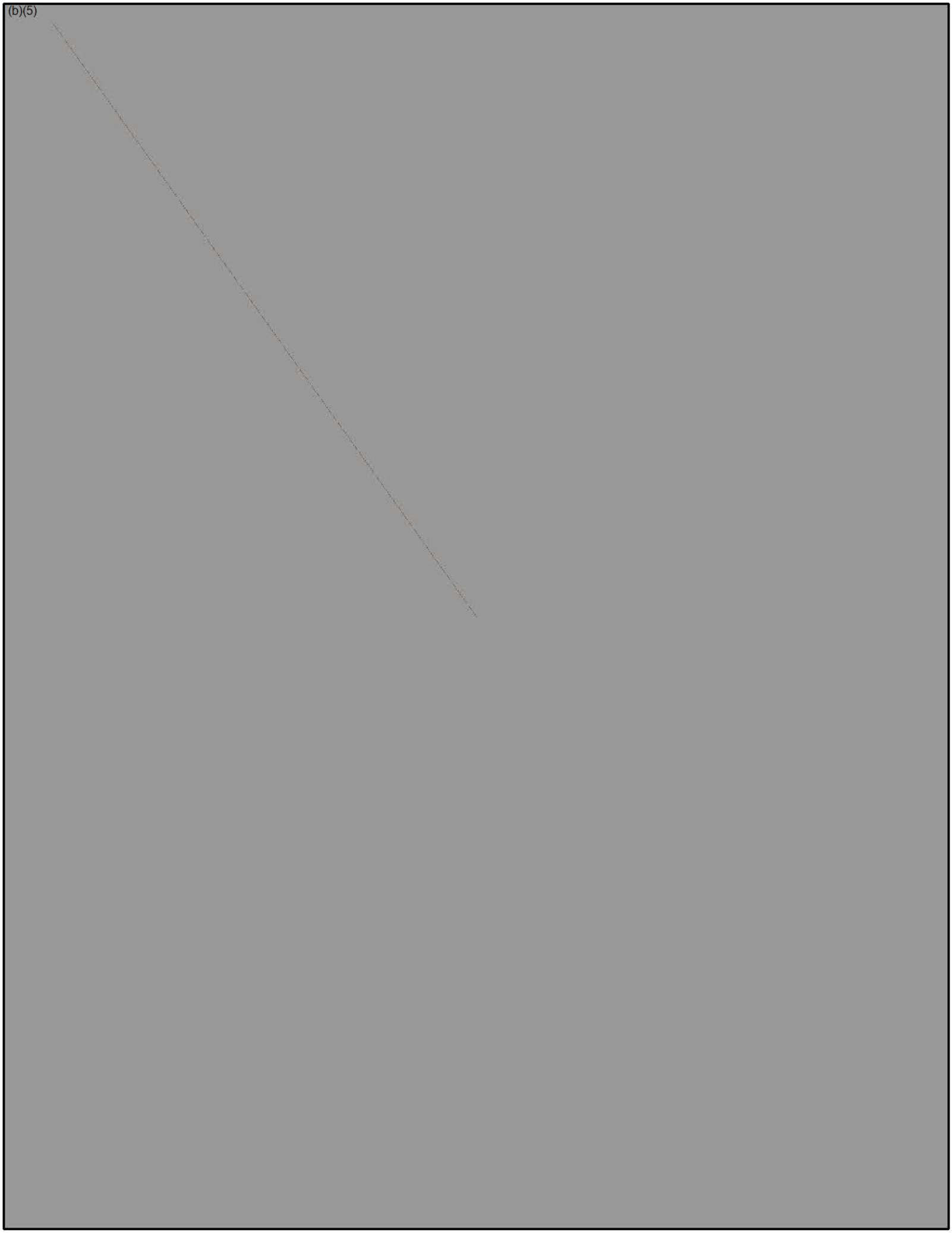


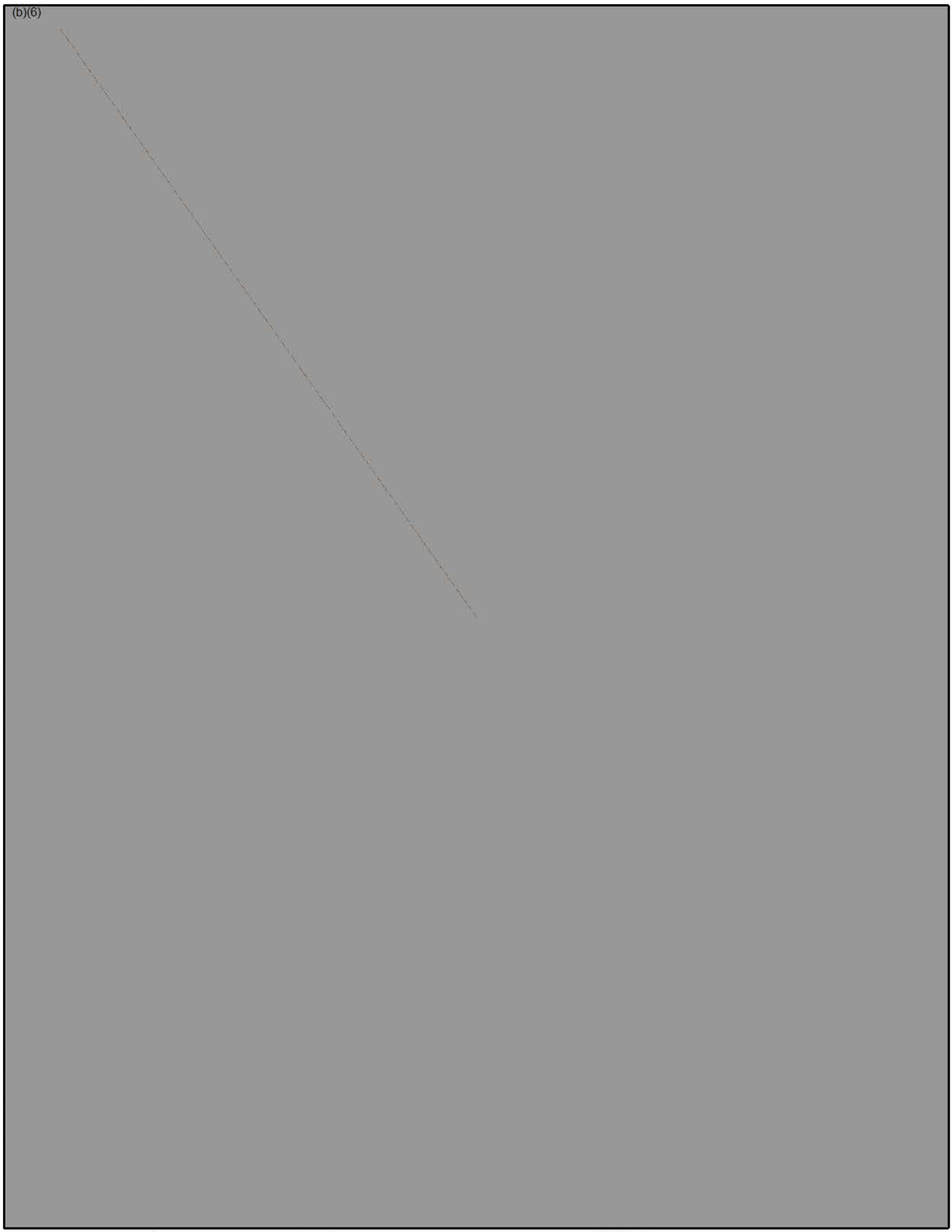
(b)(5)

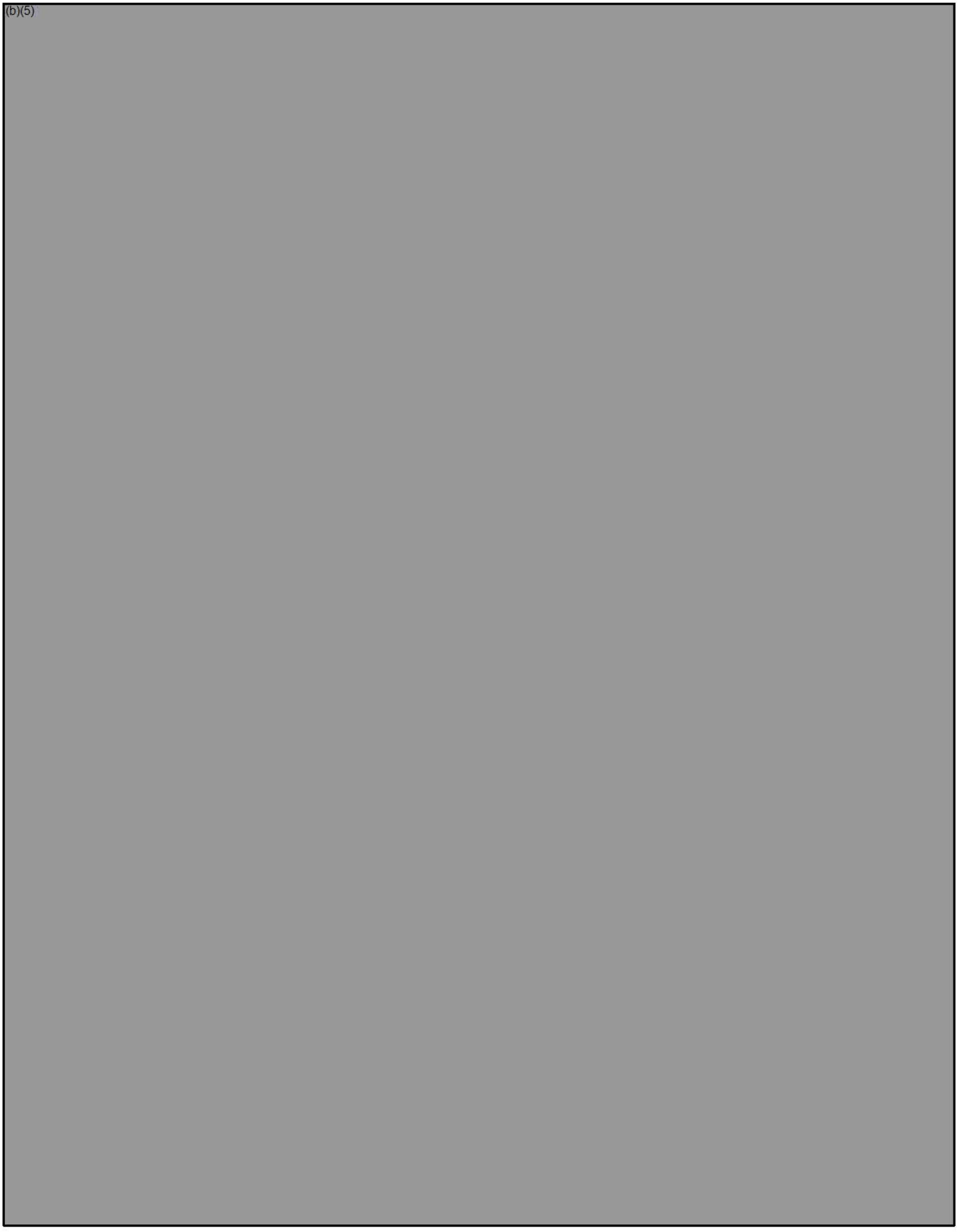


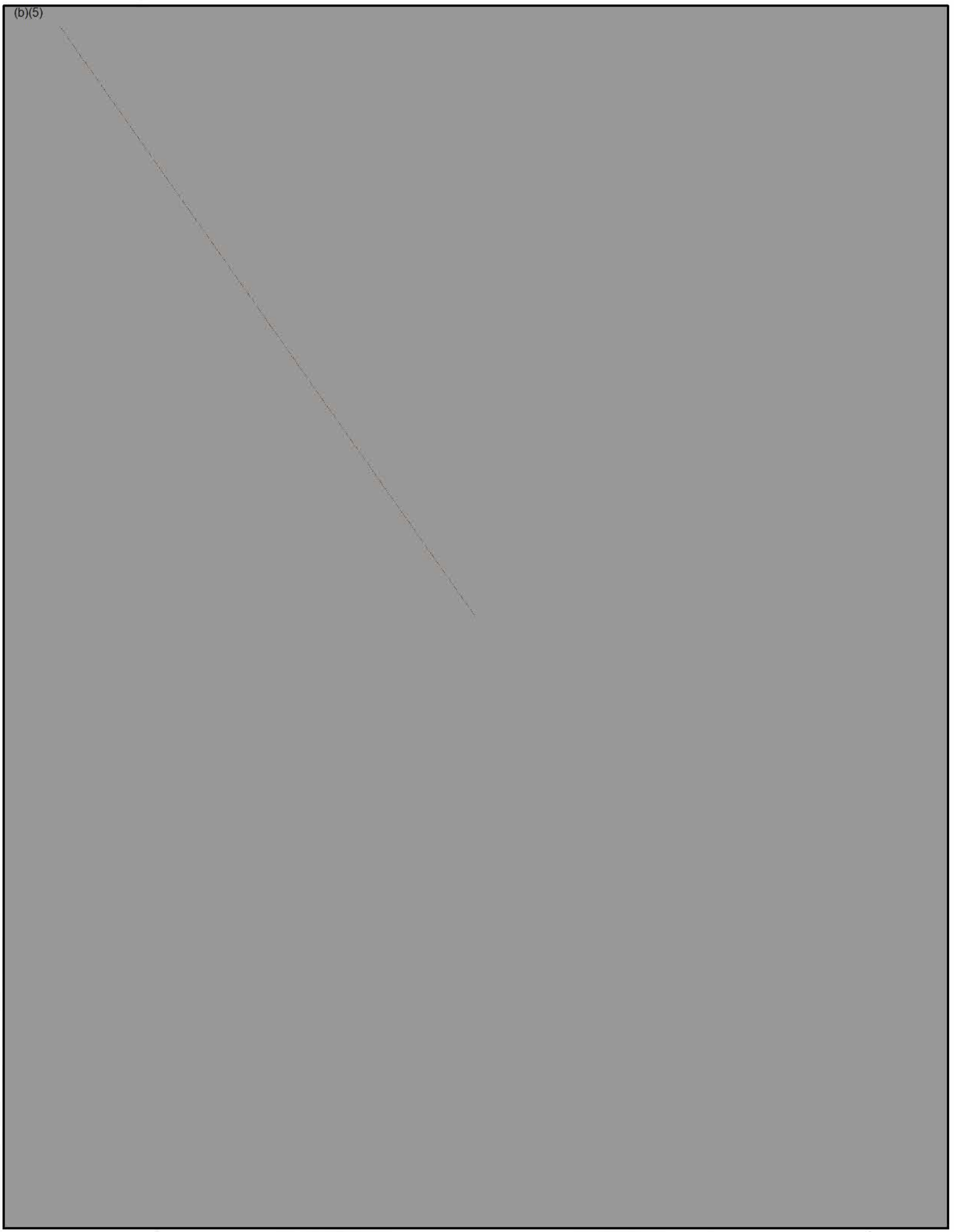

(b)(5)


(b)(5)

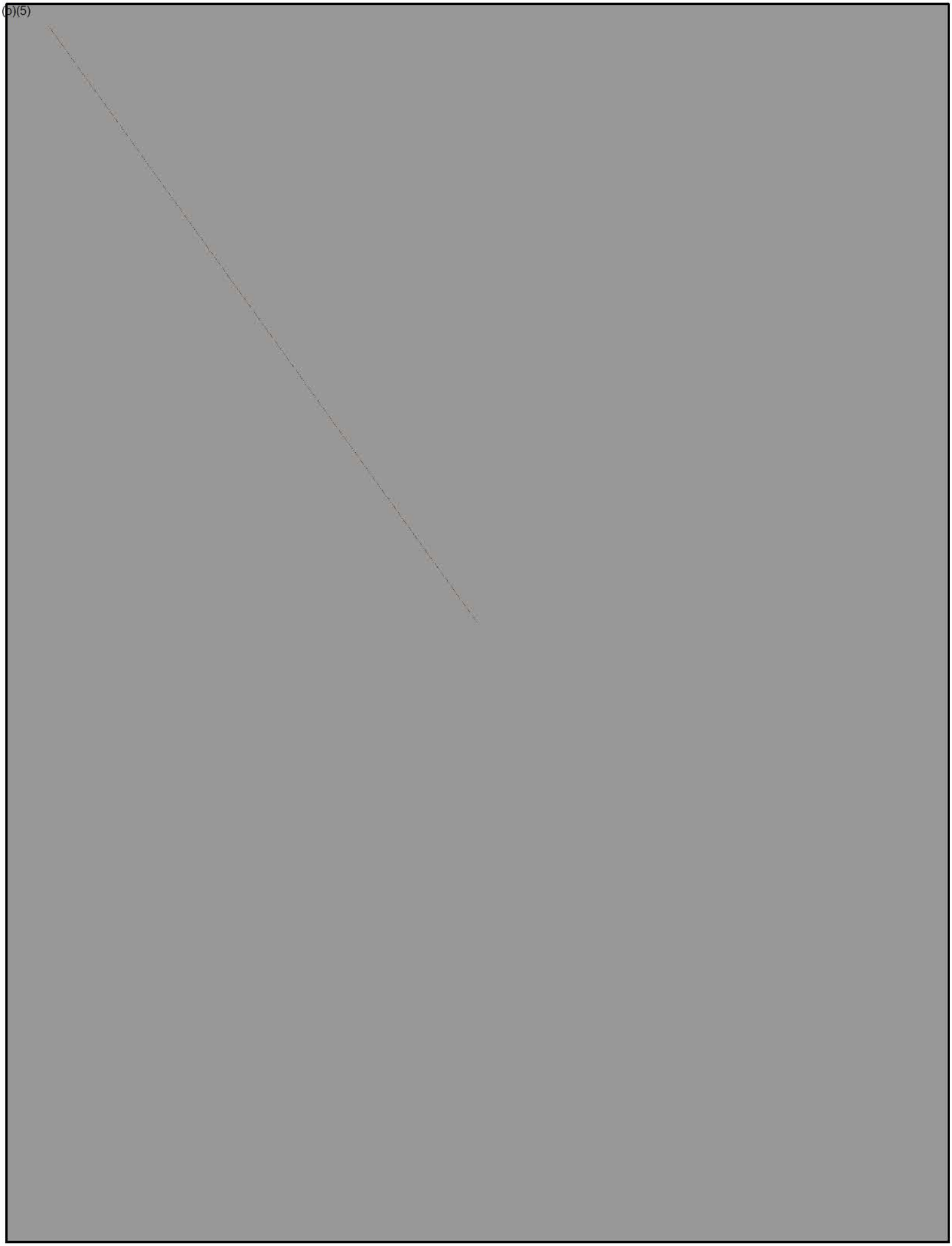

(b)(5)

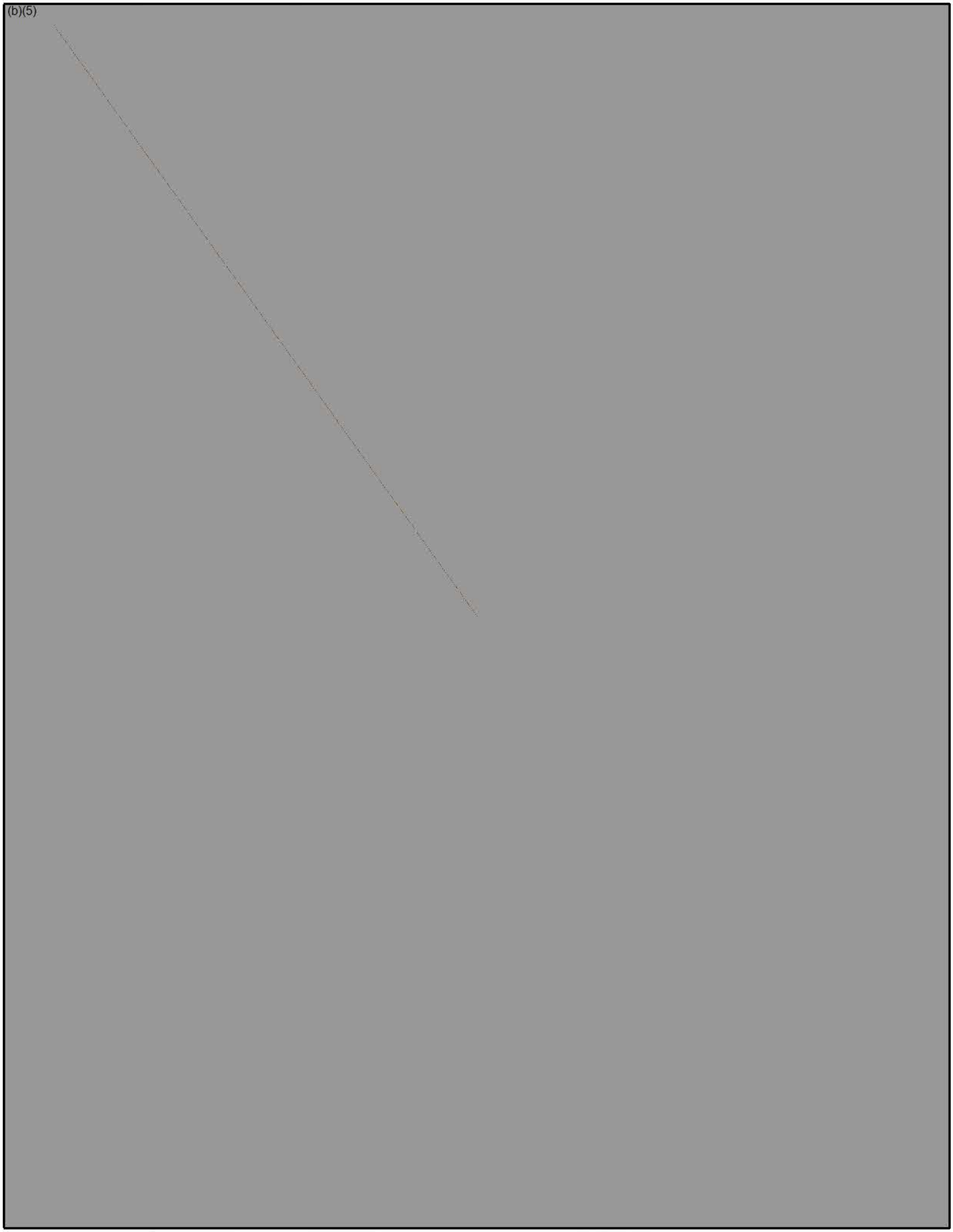

(b)(6)

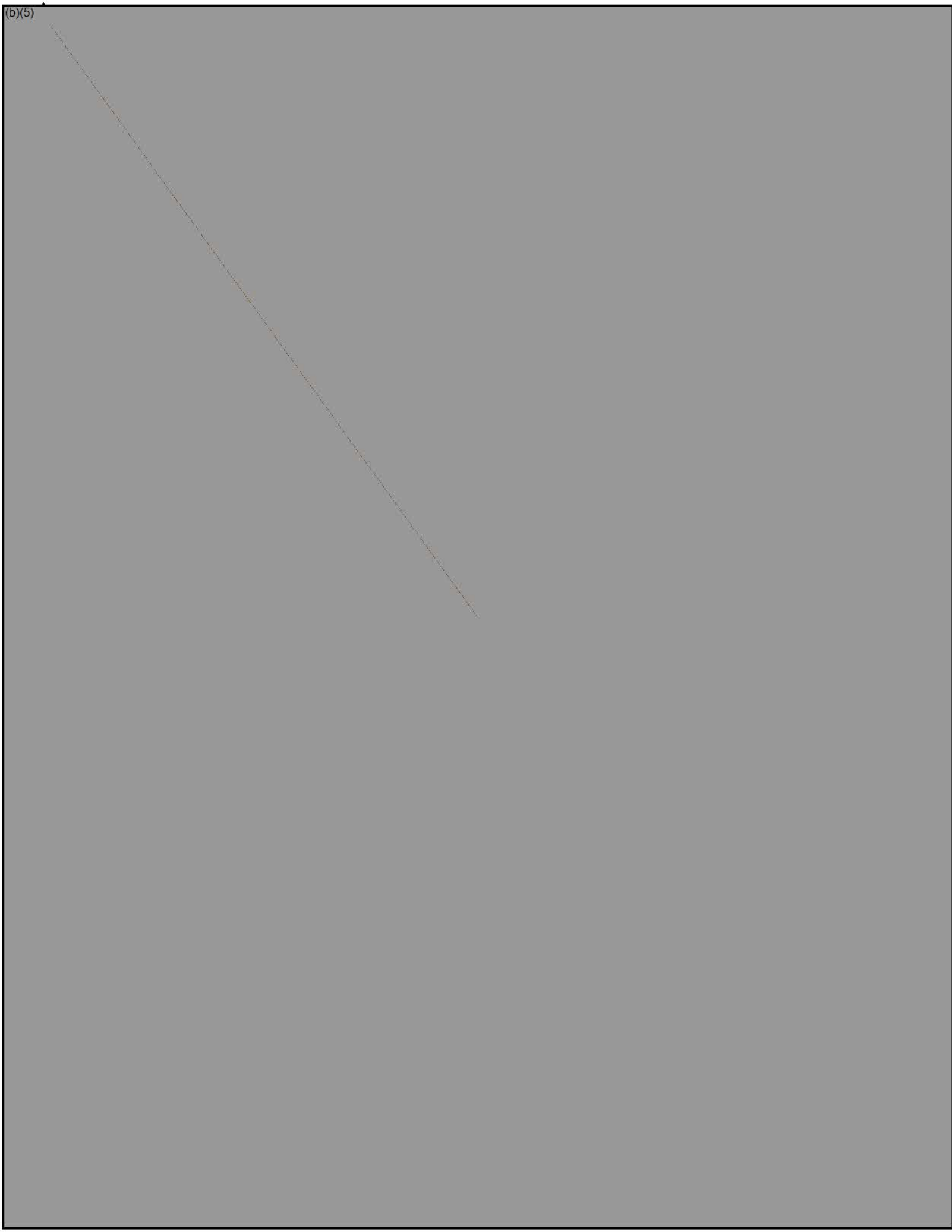

(b)(5)


(b)(5)

(b)(5)




(b)(5)


(b)(5)

Privacy

- Organizational Issues
  - Data sharing agreements
  - Signed agreements
  - Trusted brokers
- Technical Approaches
  - De-identified data sets
  - Encrypted lookback

Data Sharing Agreements

- Specifies details of data sharing
  - Reason for data sharing
  - Data to be provided
  - Steps required to guard privacy
  - Allowable uses of data
- Protects both data provider and recipient
  - Provider has legal recourse if privacy violated
  - Recipient has written permission to use data
- Consult legal department when drafting agreements

Signed Agreements

- All parties named in agreement must sign
- Each data provider will want modifications made to standard agreement
- Apprise all people with access to data of their responsibilities for maintaining privacy of data
- IRB approval must be obtained separately

Trusted Broker

- Required to satisfy privacy stipulations in data sharing agreements
- Secure repository for identified data
- Access granted to minimal number of people
- Audit logging turned on
- Machine physically secured
- Encrypted transmission of data

De-identified Data Sets

- Remove identifying information for research
  - Names, addresses, ages
  - Unique id numbers (e.g., SSN)
- Replace with non-unique or random values
  - Random id number
  - Truncated latitude and longitude
  - Grouped values (e.g., age decile)
- Delete tables that map identifying information to de-identified values
- Ensure set of values per record is unlikely to result in a unique identification

Encrypted Lookback

- May need to re-identify a person from de-identified data in the interest of public health
- Use an encrypted value to map de-identified data to identifying information
- Key for decryption stored in data provider

Genisys
DARPATech 2002 Speech

Good afternoon <morning>. I'd like to tell you about a significant information technology problem that I think you'll all understand, about its root cause and its impact on our ability to combat terrorism, and what DARPA is doing about it. It turns out, we're not getting anywhere near the full potential from our computers. Today, we use computers in the wrong way, namely to format, store, and send information... but the work is all done by people and products are consumed by people. In fact, most digital information consists of natural language, sounds, graphics, images, and other formats that can only be interpreted by humans. Is it any wonder we can't get our computers to provide more automation, be smarter in dialog, or offer any kind of intelligent assistance? To work, algorithms need structured data, that is variables with values. Yet if you analyze your own computer today, you'll probably find that of all the valuable data on your hard drive, less than one percent exists in any structured form.

Lack of automation and the structured data required for automation is absolutely not a new problem. In fact DARPA and other organizations have been researching ways to make algorithms understand natural language, and to add semantic markup to web pages, so they may be interpreted by software agents, for example. However, the quintessential source of structured information has been and is the database. This afternoon <morning>, I'd like to explain why current databases are inadequate, describe why databases are essential to combating terrorism, and tell you about a new DARPA program called Genisys aimed at totally re-inventing database technology.

Relational database technology still in use today was created in the mid 1970s when processors were slow and expensive, networks were created with modems, and disks were tiny in comparison to the computing infrastructure we have now. As a result, database implementations stressed time and space efficiencies at the expense of flexibility and ease-of-use. Although current databases are relatively straightforward, they still require careful design of data structures before they are populated, and both users and application programs still have to know a great deal about the design in order to access information. In addition, once the design of data structures is complete, changing it is quite difficult because interfaces used by applications must be re-written for the new design. Furthermore, no conventions are used when creating terms to name real-world entities and their attributes. Terms are invented ad hoc by database designers, and as a result, people who are new to the database have some difficulty interpreting the information it contains. For these reasons, it's more difficult than it needs to be to store, share, and use information in a database, and this has stymied innovation in general and the creation of large-scale automation and intelligent processing algorithms in particular. You may have personal experience creating a database---if you have Microsoft Office installed you already have a relational database called Microsoft Access---even for simple projects, most people find relational databases difficult to use.

Lack of intelligent automation and databases that enable it hurts every one of us who wants greater efficiency in our day-to-day business, but it's a particularly acute problem for combating terrorism. There are many terrorist organizations, and they each have many targets and attack methods from which to choose. Combating terrorism requires us to sift through gigabytes of information, recognize patterns, and share information between organizations. Doing this quickly enough to matter requires automation and the databases which support it.

By analyzing a number of terrorist attacks , it's clear that having timely information is key. The sooner you know about an attack, the higher the probability that you will be able to preempt or mitigate it, if necessary to respond and recover from an incident, and if possible prosecute or take appropriate military action against the terrorists. Fast, certain response is one of the best methods of deterring future attacks by other terrorists because it raises the possibility that their actions will only result in failure, humiliation, and punishment.

It turns out that before every attack, critical events occur. Supplies must be purchased, weapons may be developed and rehearsals conducted, and elements must be positioned. In almost every case, separate "systems," if we use that term loosely, capture partial evidence about these events. Examples include records of purchases and other transactions, messages and communiques, facilities and ownership, travel

itineraries, and relationships that can be inferred. These events are often reconstructed after the fact to support criminal prosecution, but we'd really like to be able to compile the information before an attack, rather than afterwards. Unfortunately, we aren't tapped into these systems. Our intelligence organizations are not optimized for terrorist threats, so we'll need to build new systems and populate new databases to get the coverage we now need. An example helps to illustrate:

In 1995, Aum Shinri Kyo attacked the Tokyo subway system using sarin gas, killing 11 and sending hundreds to the hospital. This attack is a prototypical, perhaps extreme example of predictive precursor events. Prior to the 1995 attack, Aum Shinri Kyo cultists led by Shoko Asahara had tried to buy US and Russian chemical munitions. When that failed, they engaged in a substantial weapons development program aimed at mass effects. They created elaborate development facilities for producing sarin, purchased expensive equipment, and accumulated huge chemical stockpiles. Following several malicious attempts with ineffective agents, they created a test facility in Australia and tested sarin by killing sheep whose remains were later discovered. Noxious chemical leaks were reported by neighbors near their development facility in Japan, but police still made no arrests. Asahara broadcast his intentions clearly via radio. And months before the subway attack, cultists used sarin in an overt attempt to kill three judges in the city of Matsumoto. In this example, just as in the case of 9/11, fragments of information, known by different parties in different organizations, contained a predictive pattern that could have been identified had the information been shared and properly analyzed.

To address these issues, we've created the Genisys program. Genisys has three goals. First, we'd like to be able to integrate and, if desirable, restructure legacy databases. Second, we want to dramatically increase the coverage of vital information by making it easy to create new databases and attach new information feeds automatically. This is new, multi-media, broad-spectrum information that doesn't exist in a structured database today. Third, we want to create brand new database technology based on simple, scalable, distributed information stores we call repositories. In contrast to today's databases, repositories will be able to represent a broad array of information which varies in terms of structure, certainty, and format and accessing information will be easier. Operationally, we will focus Genisys on the problem of combating terrorism. In the context of a larger counter-terrorism information system, Genisys repositories will both supply and receive information. Initially, repositories will be populated with synthetic data to support experimentation and rapid prototyping, but our intention is to iteratively develop and transition the technology, using operational feedback to focus future research. This is a model we call "assured transition." Now I'd like to describe in technical detail two of these program goals: database integration and the core repository technology.

There are three well-known methods of integrating databases today. As applications are built, new interfaces can be created as needed. As the number of applications and databases increase, this method requires a quadratically increasing number of interfaces, limiting scalablity. Alternatively, a software agent known as a query mediator may be used to translate application queries into queries that are understood by the different databases. This approach improves scalability somewhat, but often the complexity is simply moved into the agent. A third approach involves manual re-engineering of the relevant databases and subsequent manual re-implementation of the interfaces. For mature systems, this is sometimes worthwhile, but it involves the greatest amount of re-engineering and hence the greatest cost. A primary technical limitation shared by all of these approaches is the lack of design conventions for creating terms that name real-world entities and their attributes.

A new technical approach identified early in Genisys is to use domain ontologies, which are the terms and relationships that are associated with the concepts in a particular model of the world. Once there is an ontology, it's possible to define a naming convention and reduce the cost of designing the database. It's also possible to create a database crawler, similar to a web crawler, that discovers the structural design, or schema, of a legacy database. Using the existing schema and ontolgies, it is possible to create tools that help people map the old schema into the new one, reducing the cost of re-engineering and integrating databases dramatically. It's also possible to create a simpler query language than today's SQL, and perhaps in addition, automated tools for translating interfaces. Using this approach, we believe legacy databases and systems can be transformed and integrated cheaply. Now let's look at the design criteria and goals for new repository technology we are defining.

Today, processors are a commodity, high-speed networks are ubiquitous, and we're beginning to think of disk space as infinite and free. Today's infrastructure suggest new design criteria that will result in repositories that are friendlier and more capable than current databases. For example, current databases support analytical queries very well, but they have several drawbacks that make them inflexible and hard to use. First, they require a priori data modeling. That is, the database must be designed before it can be populated. Moreover, once designed and applications begin to use it, changes to the design are difficult because they require, in turn, modifications to the interfaces for applications. Repository technology will change this by making applications independent of the physical structure of the database. For example, queries will no longer refer to table names in a relational implementation. To provide even more flexibility, we want to exploit search technology so that applications do not even need to know the physical location of data. These two improvements will make creating and using structured data easier. To enhance sharing, we're interested in intelligent processing to help enforce policy while automatically granting access to people who should have it. We also want machine assistance for declassifying data, releasing the sharable content while protecting the sensitive parts. Inside the repository, we want to be able to represent uncertainty in some natural way. Finally, we want automation for restructuring the content to increase performance, add new information types, or solve new problems. Clearly, this is a long list of goals, and we may need to trade some things off to achieve them. Of paramount importance are scalability and ease-of-use, because ultimately, we want to a lot of structured content relevant to combating terrorism.

Specifically to address the privacy concerns of those not connected with terrorism, repositories need technologies for protecting personal privacy. We have three methods relevant to protecting privacy. First, we can exploit partitioning to separate identity information from transactions that people conduct, only reforming this association when we have evidence and legal authority. We can also use partitioning to project specific information authorized for particular the role of the requestor. In this way, the Center for Disease Control, for example, could be permitted access to recent statistical medical information, while others could not. Second, we will develop and employ information filters to keep information that is not relevant out of the repository. Filters could be used to implement laws and policy that regulate the kinds of information recorded and who it pertains to. Finally, we will use software agents to mine the information in the repository, form associations from content, and expunge information found to be unrelated to combating terrorism. Filters are not perfect, and this last method will help ensure privacy when filters fail or when the combination of bits of information make it clear that the information has no utility. These three method are initial steps toward achieving personal privacy and security. They are probably inadequate, and we are soliciting additional new ideas in this area.

We are just getting started in Genisys and don't yet have tested designs that could implement these characteristics, but we believe that our goals are attainable. With Genisys, we'll be able to reduce the cost of integrating databases so that different DOD and other government agencies will be inherently better at coordinating. In addition, increasing coverage from many different sources by auto-populating large information repositories means that we're much more likely to have the information we need to prevent or respond to an attack. Other programs at DARPA will use content in the repository to match patterns and support analysis, and thus we'll have a higher probability of recognizing attacks. The goal for Genisys technology is to make it cheap and simple to build these auto-population and reasoning systems while still accommodating scale, representing uncertainty, projecting data for sharing, and helping to manage the information automatically.

In March, we published a Broad Area Announcement, DARPA BAA02-08 and have made an initial round of proposal selections. We're now in the process of getting selected offerors under contract. We welcome your good ideas and support. At this point in the program, there are still opportunities to join the team for anyone with the potential for outstanding contributions.

Thank you.

Bio-Surveillance System Program Briefing

Ted Senator
DARPA/ISO
November 2001

Objective

- **Detect Pathogen Release at Least Two Days Earlier than Current Practice (stretch goal is 24 hours after attack)**
  - Disease dependent
  - Current practice relies on alert clinicians who
    - Have seen few or no cases
    - Look for common diagnoses/diseases
  - Discriminate Between Intentional Pathogen Release and Natural Outbreak of Disease (e.g., Flu)
  - Continuous and Event-Oriented Monitoring
- **Identify Type of Attack (e.g., Anthrax, Smallpox, Plague, Botox)**
- **While Protecting Privacy of Individual Source Information, including medical records**

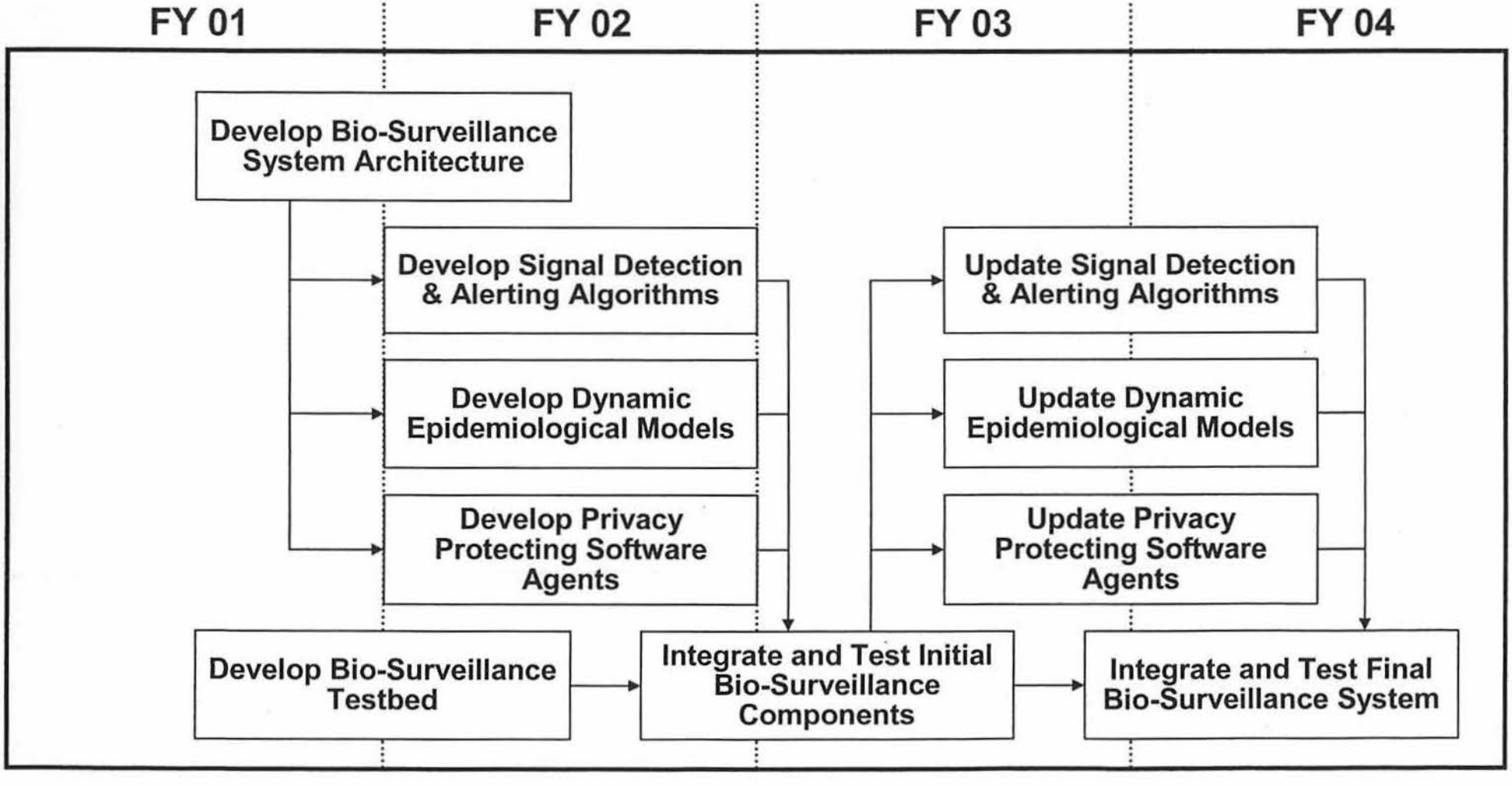
How

- **Use Unconventional Data Sources as Early Sentinels**
  - Data Mining to Identify Value of Sources
- **Normalize for "Natural" Disease Outbreaks**
- **Two Stage Detection**
  - Autonomous Alerting of Epidemiologist
  - More Extensive Decision-Aids for Alerting
- **Privacy Protection Algorithms**

Goal: What DoD Has at the End

- **Two Prototype BioSurvellance Systems in DoD Cities (National Capital Area and Norfolk)**
  - Short-term effort for National Capital Area
  - Possible Expansion to Other Cities, dependent on $ and data availability
    - San Diego
    - Omaha
    - Albuquerque
- **Knowledge of Effective Data Sentinels (i.e., what data are useful/effective indicators/predictors)**
- **Inference Algorithms for Detection of Potential Bio-Attacks**
- **Privacy Protection Algorithms**
- **BioWar Simulation**

Bio-Surveillance Program Plan



Bio-Surveillance System Program Briefing

Ted Senator
DARPA/ISO
November 2001

Objective

- **Detect Pathogen Release at Least Two Days Earlier than Current Practice (stretch goal is 24 hours after attack)**
  - Disease dependent
  - Current practice relies on alert clinicians who
    - Have seen few or no cases
    - Look for common diagnoses/diseases
  - Discriminate Between Intentional Pathogen Release and Natural Outbreak of Disease (e.g., Flu)
  - Continuous and Event-Oriented Monitoring
- **Identify Type of Attack (e.g., Anthrax, Smallpox, Plague, Botox)**
- **While Protecting Privacy of Individual Source Information, including medical records**

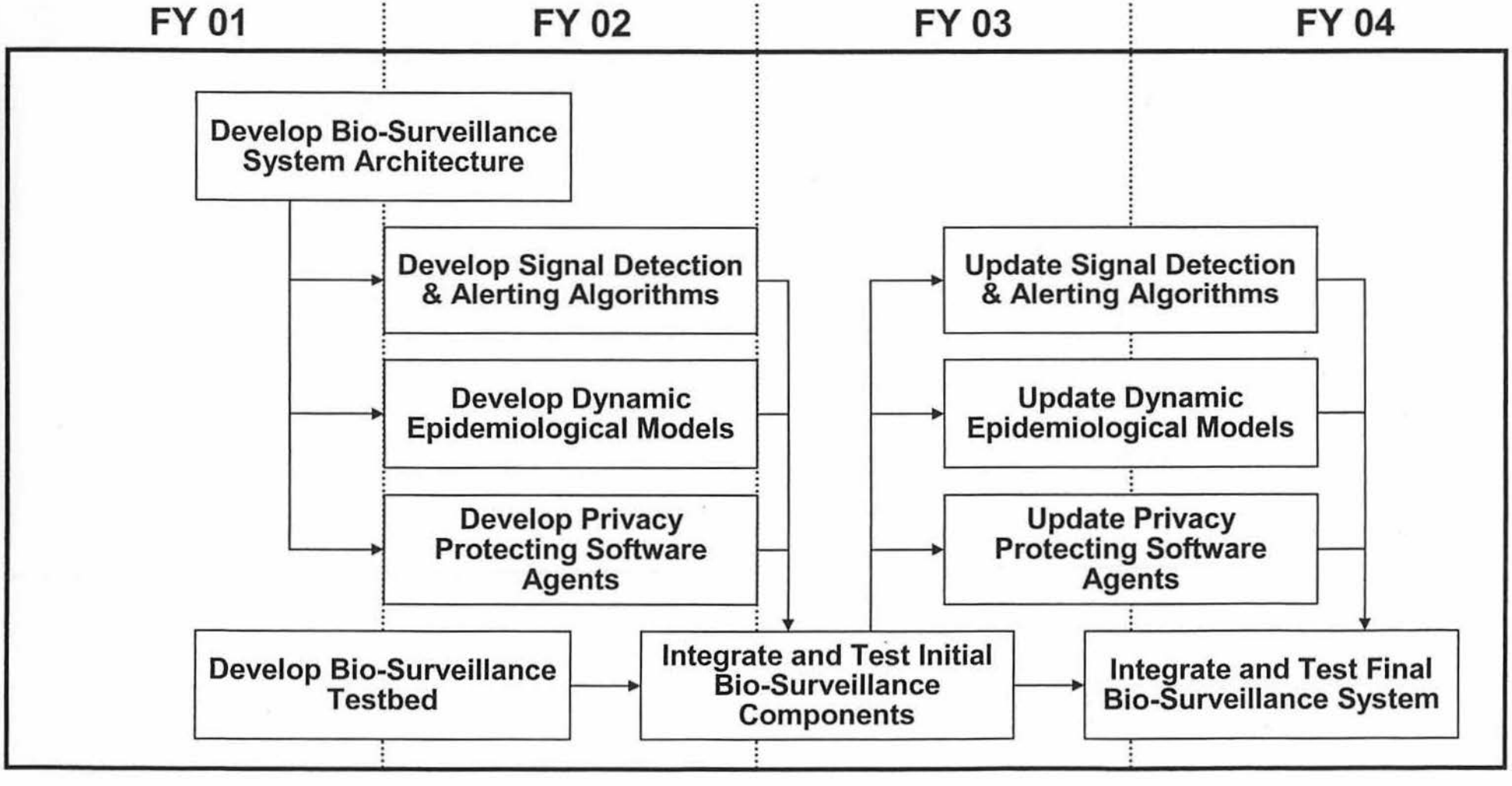
How

- **Use Unconventional Data Sources as Early Sentinels**
  - Data Mining to Identify Value of Sources
- **Normalize for "Natural" Disease Outbreaks**
- **Two Stage Detection**
  - Autonomous Alerting of Epidemiologist
  - More Extensive Decision-Aids for Alerting
- **Privacy Protection Algorithms**

Goal: What DoD Has at the End

- **Two Prototype BioSurvellance Systems in DoD Cities (National Capital Area and Norfolk)**
  - Short-term effort for National Capital Area
  - Possible Expansion to Other Cities, dependent on $ and data availability
    - San Diego
    - Omaha
    - Albuquerque
- **Knowledge of Effective Data Sentinels (i.e., what data are useful/effective indicators/predictors)**
- **Inference Algorithms for Detection of Potential Bio-Attacks**
- **Privacy Protection Algorithms**
- **BioWar Simulation**

Bio-Surveillance Program Plan



trohrkemper

From: Siegrist, David [Siegrist@PotomacInstitute.org]
Sent: Thursday, May 30, 2002 8:22 PM
To: 'Joe Lombardo (E-mail)'; 'Gene McClellan (E-mail)'; 'Chung-Sheng Li (E-mail)'; 'Mark Musen (E-mail)'; 'Mike Wagner (E-mail)'; 'Karen Cheng (E-mail)';
(b)(6) [redacted] Rich.Wojcik@jhuapl.edu'; 'GeneMc2@aol.com'; 'Liuzzi Raymond A CIV AFRL/IFTD'; 'mcam@us.ibm.com'; 'Latanya Sweeney'
Cc: rwong(contr-iao); tsenator; rlinn(contr-ixo)
Subject: Privacy Protections Work Product and Panel

Privacy Product Spec.xls

Assuring Privacy
Session Product: Table (see attached): Showing privacy technology for Developmental and Deployed systems with advantages and disadvantages. Also, an Issue List, including specification of alternative resolutions, and action items for resolution.
NOTE: See Sheet 2 of attachment as well. (Please advise me of suggested improvements on this draft format ASAP. Thanks.)

Prior Input: Identification of Active Privacy Protection development efforts that are part of PI overview brief.

Sample questions that may be addressed in the 4 team presentations:

Need for Privacy Protection in our DEVELOPMENTAL Program
List what each project is doing in its developmental project to assure privacy
Approach to any IRB's, NDA's, etc.

Need for Privacy Protection in a DEPLOYED Program
What Privacy technology development is being conducted
What assumptions are being made?
What is the Concept of Operations?
What is the tradeoff between sensitivity and privacy?
List what each project is doing in its planned deployed system to assure privacy
- List Advantages and Disadvantages of Each
- Identify Metrics and Ways to Measure Successful Privacy Protection

Are there a Set of Privacy methods that all the Projects should apply for Developmental and Deployed Projects?

List Any Issues Identified, with Alternatives for Resolution and Action Items

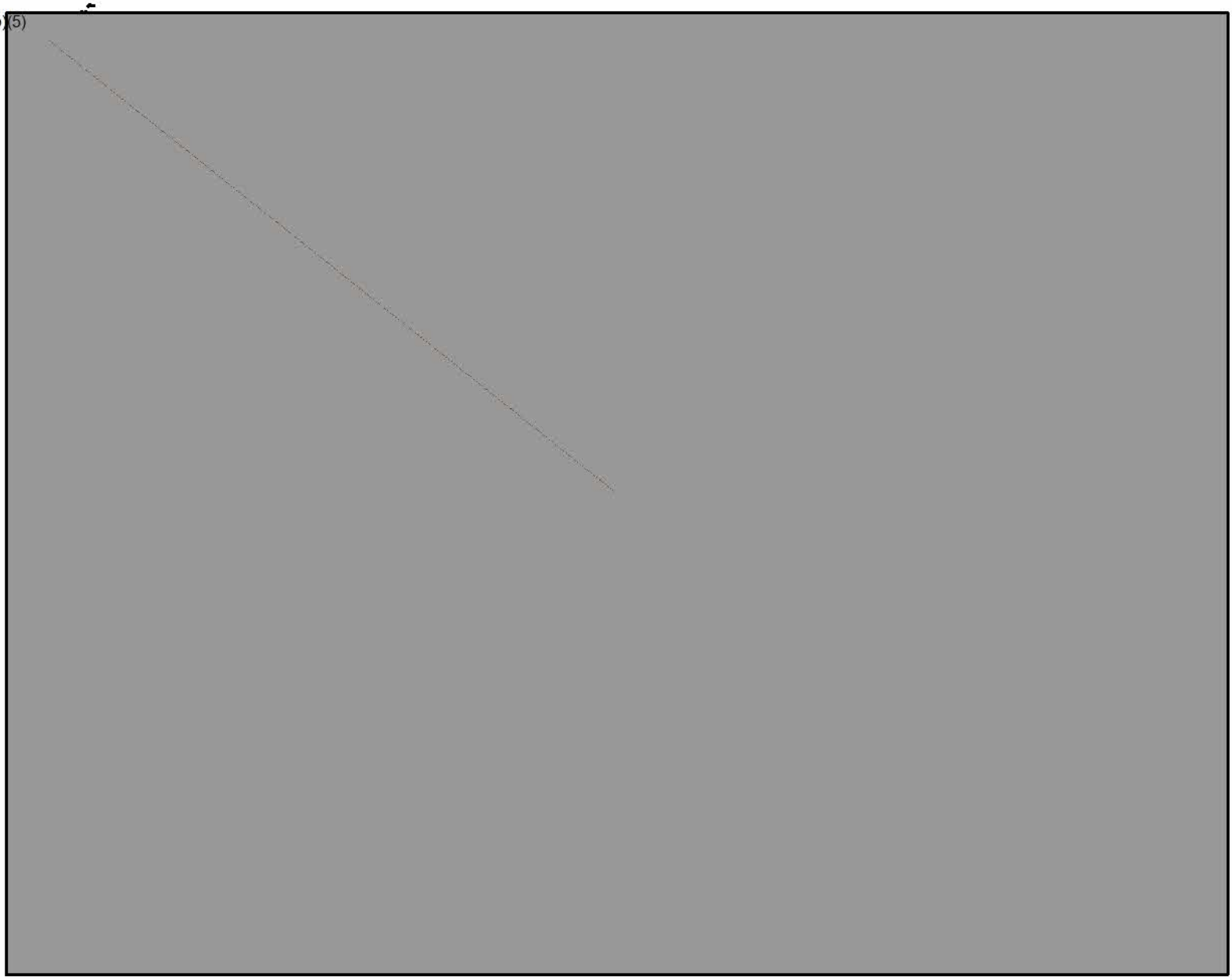
(b)(6)

(b)(5)

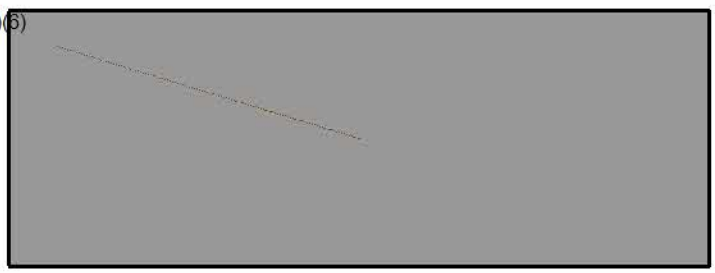
Subject:

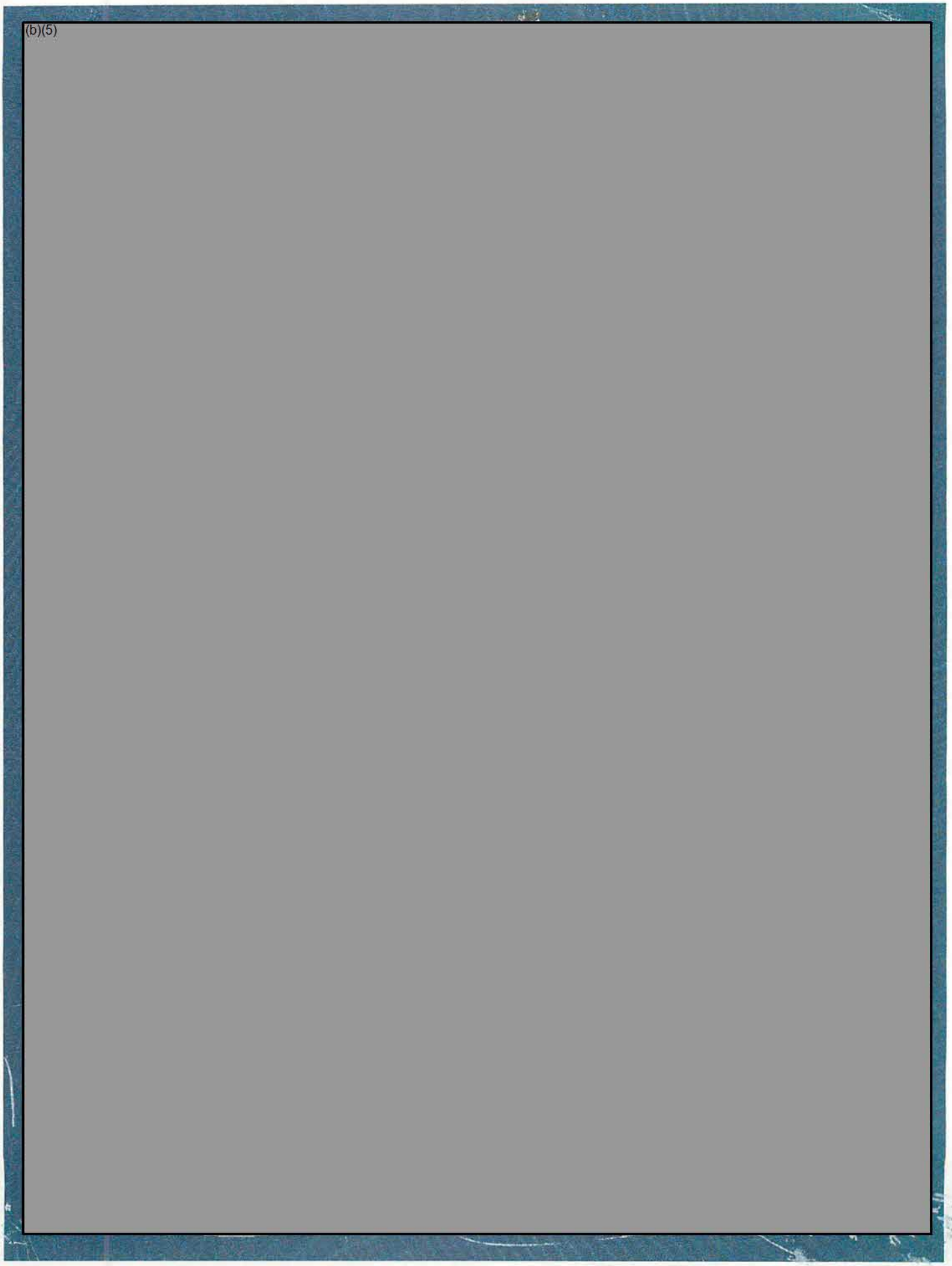
(b)(6)

(b)(5)

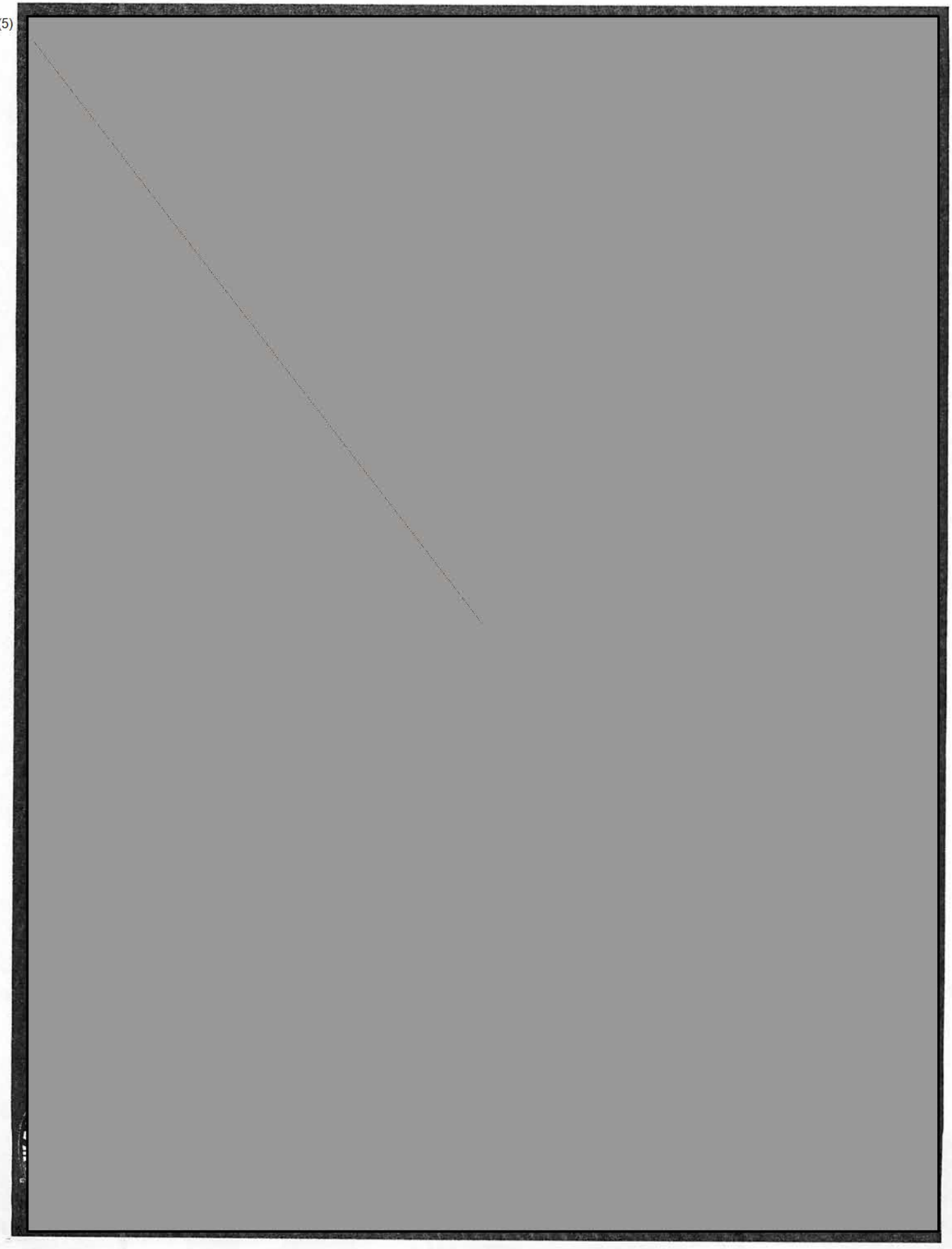


(b)(5)

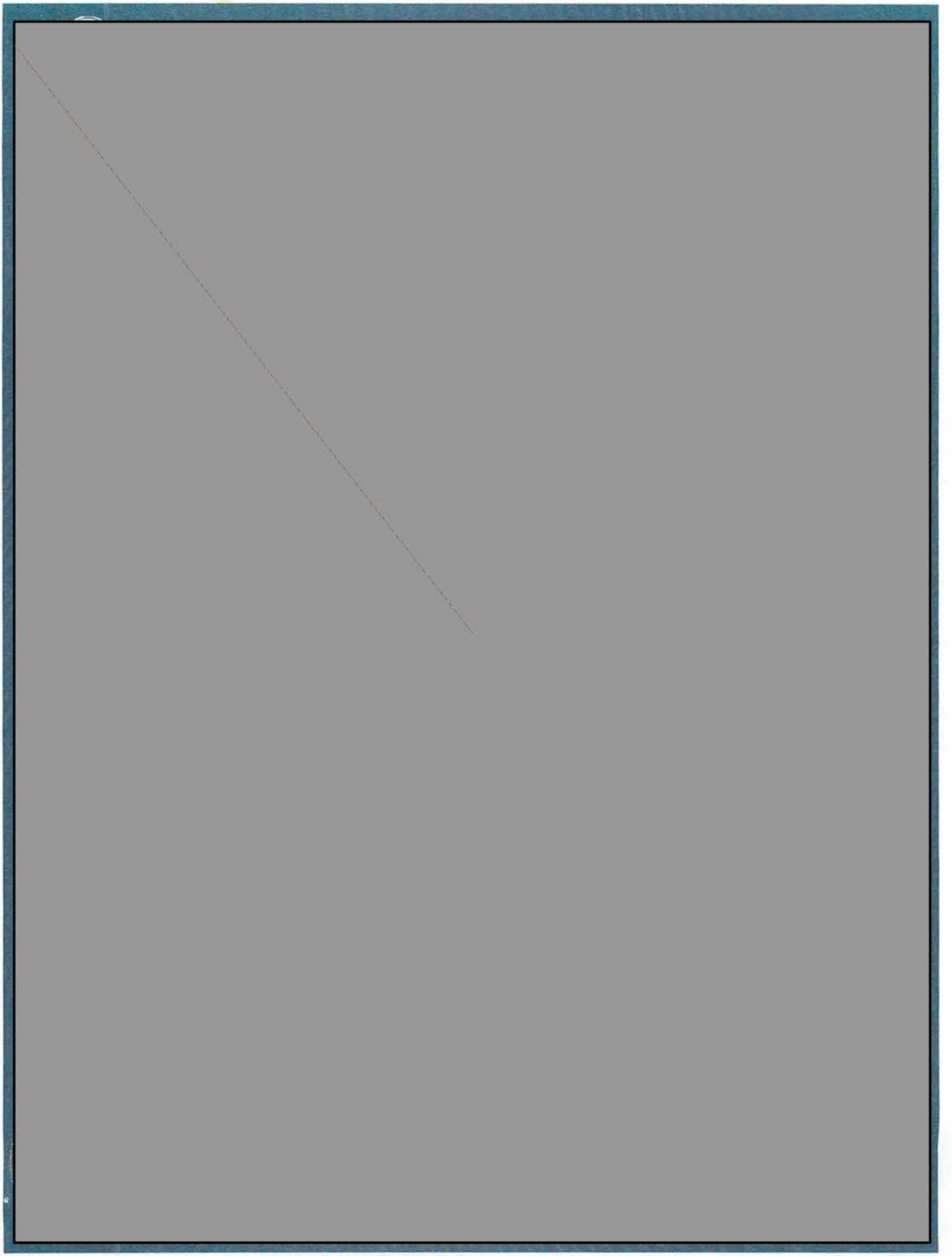



(b)(5)

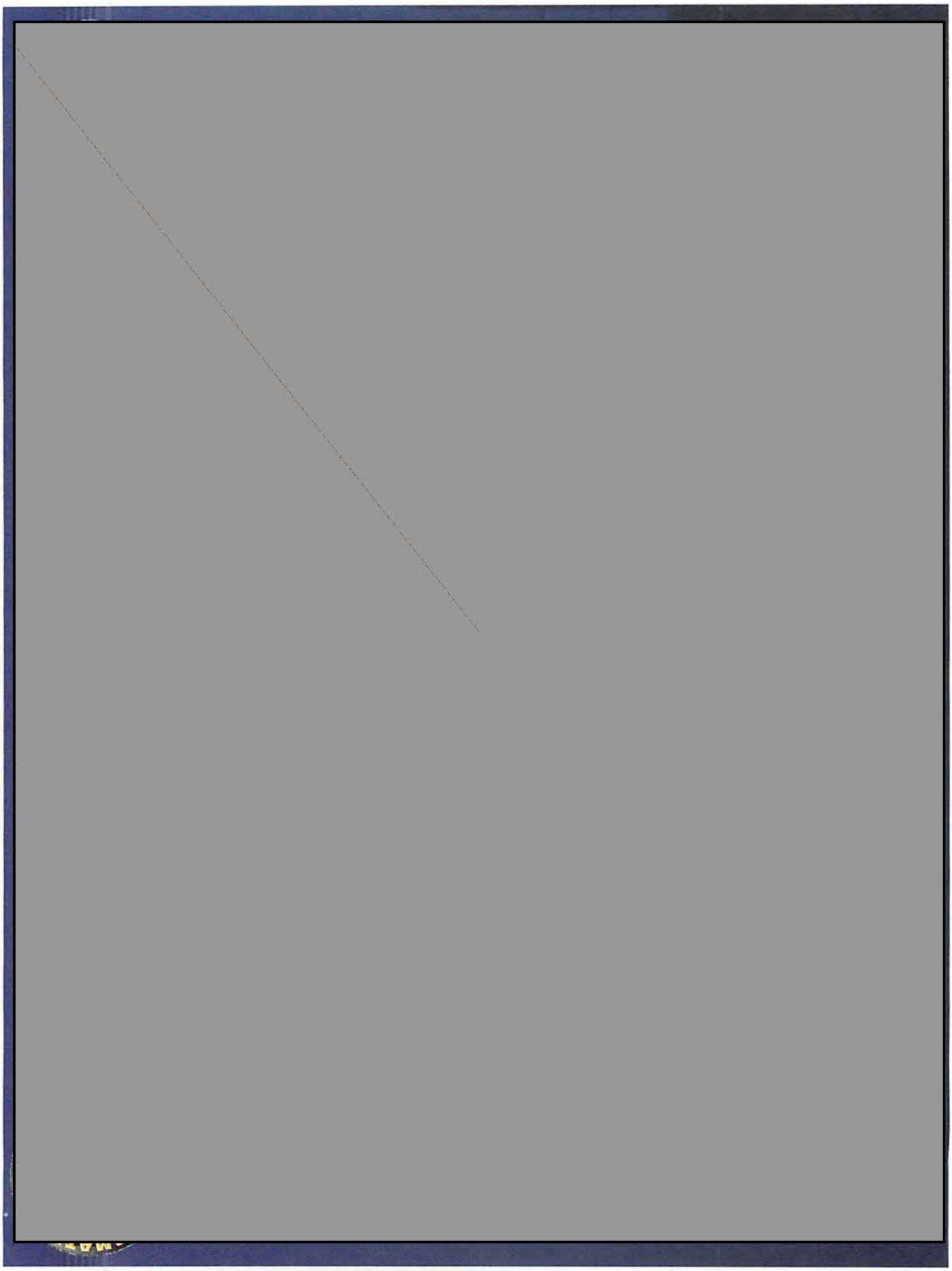
(b)(5)



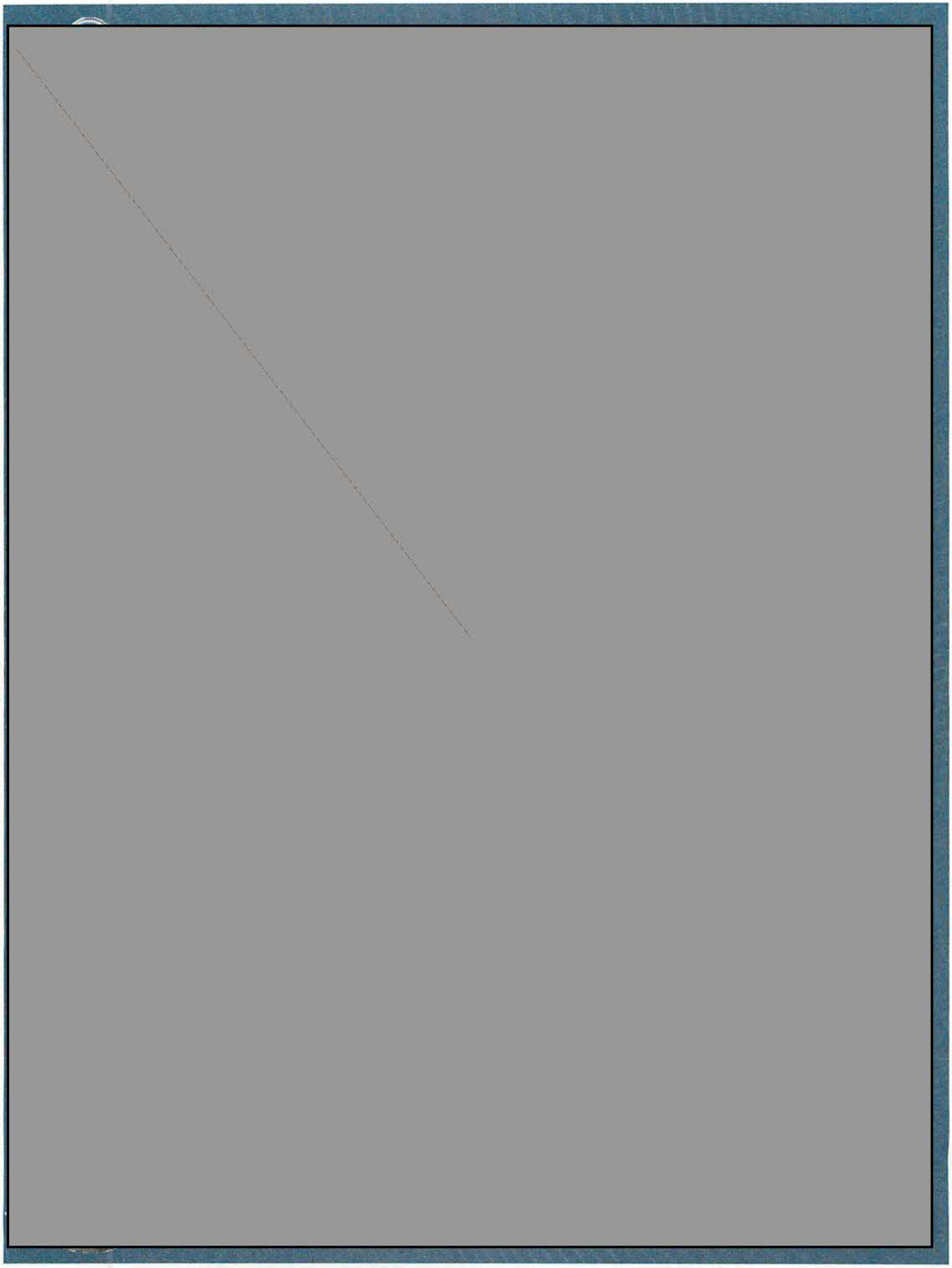
(b)(5)



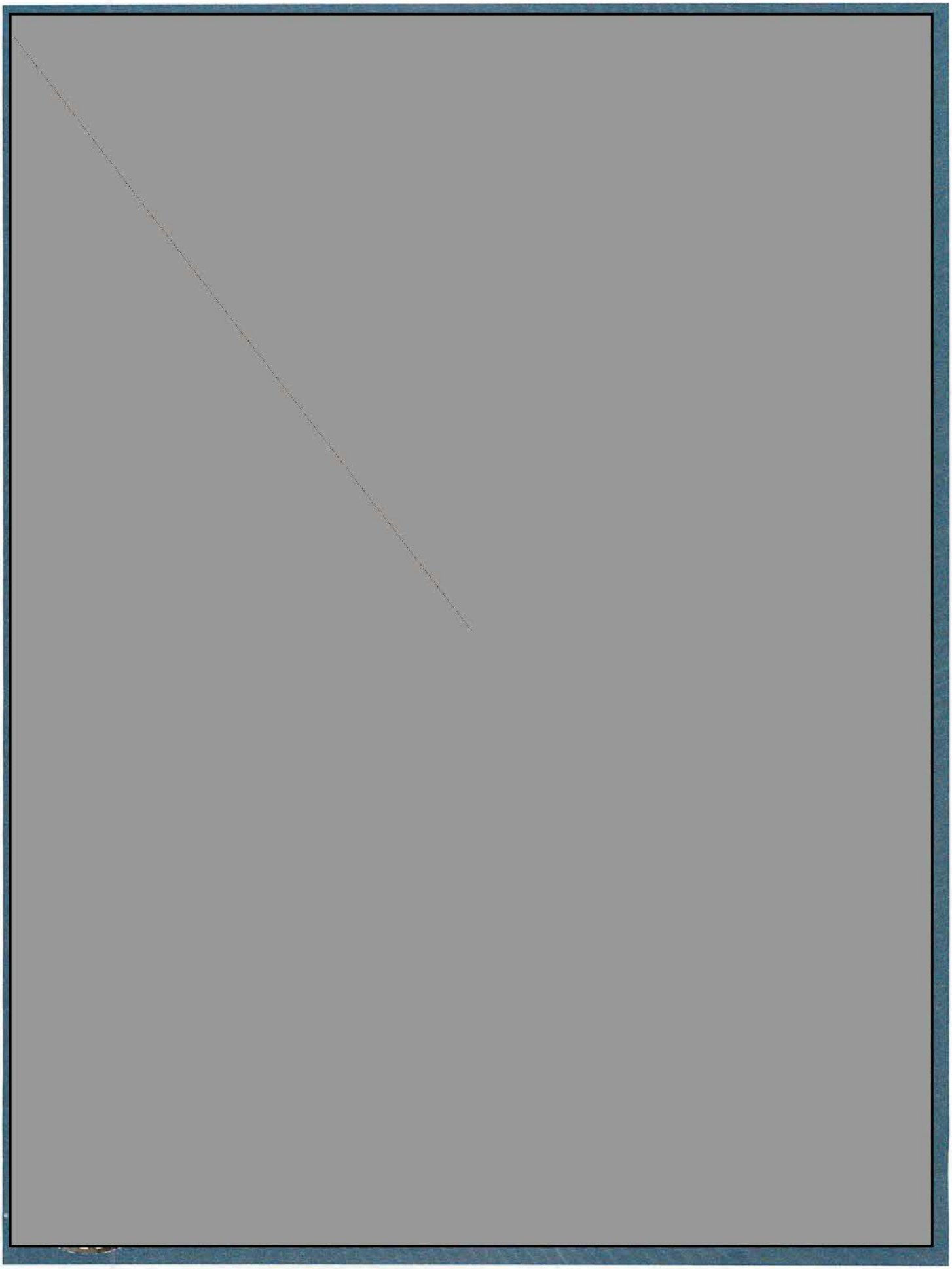
(b)(5)



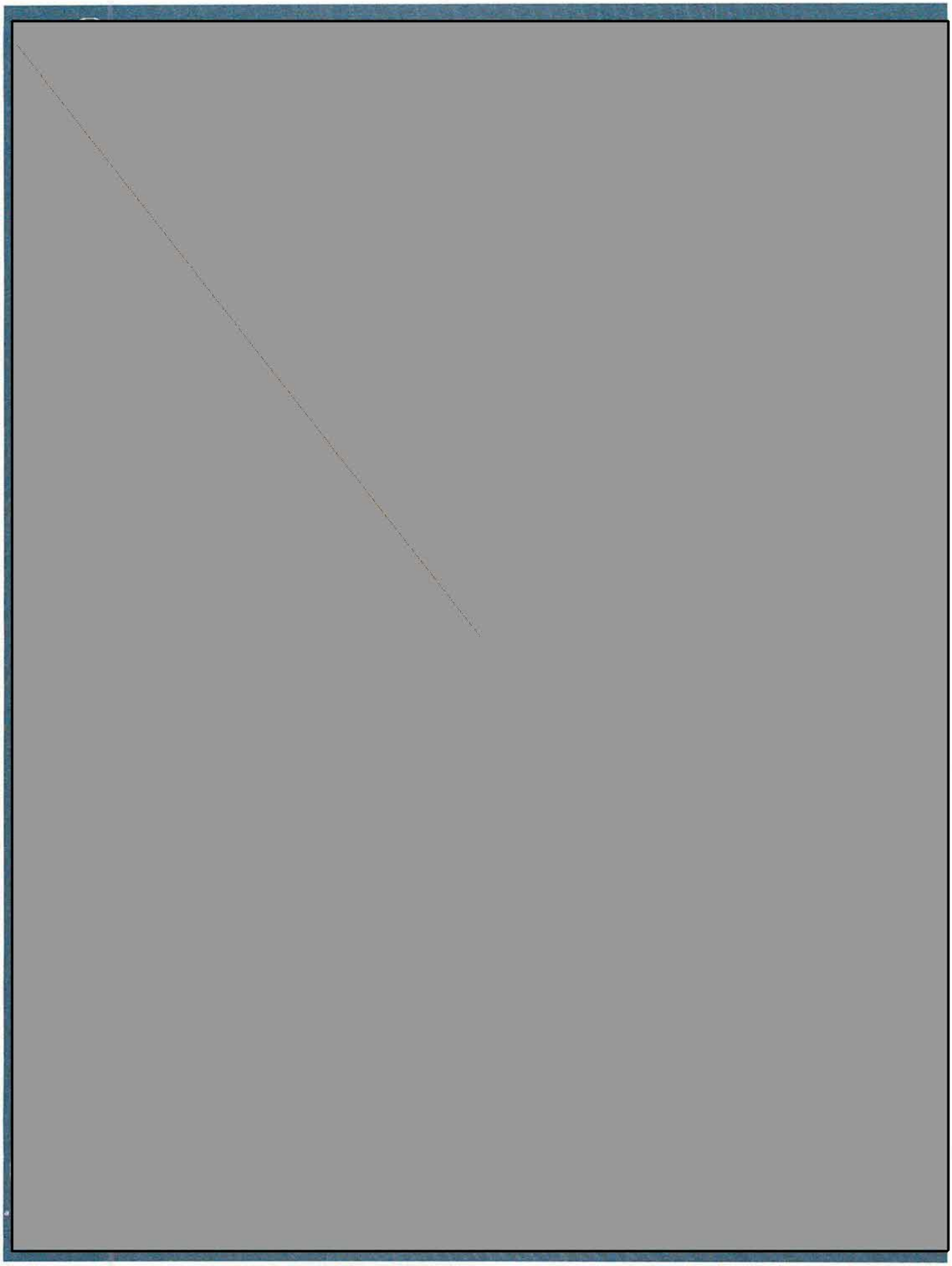
(b)(5)



(b)(5)



(b)(5)



February 01, 2002

Friday

February 2002

S	M	T	W	T	F	S
					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28		

March 2002

S	M	T	W	T	F	S
					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30
31						

(b)(6)

[redacted]

8 am — 7:00am-9:00am TeleWork (Home)

9 00 — Travel (Home to Work)

10 00 — Updated: IAO Office Issues/Discussions (1014)

11 00

DDRE (S&T) mtg (1014)

12 pm

Reserved (DARPA 1014)

1 00

2 00 — 2:15pm-2:30pm Staff Car to DDR&E

DDR&E Colloquium - car dep G1 at 2:15pm (ZAI -1815 North Ft Meyer Drive, Suite 509, Rosslyn metro stop)

3 00

4 00

5 00

Notes

jpoindexter

February 04, 2002

Monday

February 2002

S	M	T	W	T	F	S
					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28		

March 2002

S	M	T	W	T	F	S
					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30
31						

(b)(6) [redacted]

8 am — 7:00am-9:00am TeleWork (Home)

9 00 — Travel (Home to Work)

10 00 — NSA Visit (Ft. Meade)

1 00 — Discussion of SETA RFQ (1014)

Notes

jpoindexter

February 05, 2002

Tuesday

February 2002

S	M	T	W	T	F	S
					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28		

March 2002

S	M	T	W	T	F	S
					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30
31						

8 am	7:00am-9:00am TeleWork (Home)
	Travel (Home to Work)
9 00	Paul Polski/FAA (DARPA 977)
10 00	
	Alan Frederick (DARPA 1014)
11 00	ADM Bill Studeman Mtg (DARPA 1014)
12 pm	Reserved (DARPA 1014)
1 00	
2 00	Fred Schatzman (DARPA 1014)
3 00	Steven Simon (DARPA 1014)
4 00	HORUS (ISX)
5 00	

Notes

jpoindexter

February 06, 2002

Wednesday

February 2002

S	M	T	W	T	F	S
					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28		

March 2002

S	M	T	W	T	F	S
					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30
31						

8 am — 7:00am-9:00am TeleWork (Home)

9 00 — Travel (Home to Work)

10 00

Tom Garvey (DAPRA 1014)

11 00

12 pm — Reserved (DARPA 1014)

1 00

2 00 — BAA Review (DARPA 1014)

3 00

(b)(6) 4 00 — 4:00pm-10:00pm Depart Office [redacted]

5 00

Notes

jpoindexter

February 07, 2002

Thursday

February 2002

S	M	T	W	T	F	S
					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28		

March 2002

S	M	T	W	T	F	S
					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30
31						

8 am	7:00am-9:00am TeleWork (Home)
9 00	Travel (Home to Work)
10 00	Reserved (DARPA 1014)
11 00	
12 pm	Staff Car to FBI
1 00	FBI (Hoover Bldg)
2 00	Staff Car to DARPA
3 00	
	David Signori (DARPA 1014)
4 00	
5 00	
(b)(6)	7:30pm - 8:30pm [redacted]

Notes

jpoindexter

February 08, 2002

Friday

February 2002

S	M	T	W	T	F	S
					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28		

March 2002

S	M	T	W	T	F	S
					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30
31						

8 am — 7:00am-9:00am TeleWork (Home)

9 00 — Travel (Home to Work)

10 00 — Review of Recommended EARS Proposal (3701-712 Conference Room (Secret))

11 00 — Updated: Dr. Arthur Fleming-Dahl, PM candidate (Dr. Poindexter's office)

12 pm — Reserved (DARPA 1014)

1 00

2 00

3 00 — IAO System Activities (3701-977 DIRO Only Conference Room (Secret))

4 00

5 00

Notes

jpoindexter

February 11, 2002

Monday

February 2002

S	M	T	W	T	F	S
					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28		

March 2002

S	M	T	W	T	F	S
					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30
31						

8 am — 7:00am-9:00am TeleWork (Home)

9 00 — Travel (Home to Work)

10 00

Congressional Briefings (DARPA 966)

(b)(6) 11 00 [redacted] ((DARPA 1014))

12 pm — Reserved (DARPA 1014)

1 00

Updated: Congressional Briefings (IAO

2 00 — Yukie Novik/ISR (DARPA 1014)

3 00 — Groove Space Organization (DARPA 1014)

4 00

5 00

6:00pm - 7:00pm Bill Crowell Meeting(DIRO)

7:00pm - 9:00pm Dinner(TBD)

Notes

jpoindexter

February 12, 2002

Tuesday

February 2002

S	M	T	W	T	F	S
					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28		

March 2002

S	M	T	W	T	F	S
					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30
31						

Lincoln's Birthday (United States)

Lincoln's Birthday (United States)

8 am — 7:00am-9:00am TeleWork (Home)

Travel (Home to Work)

9 00 — Updated: Charles Wayne re EARS Proposal Seletion (3701-970 DIRO Conference Room (Secret))

10 00

11 00 — Reserved (DARPA 1014)

12 pm — Staff Car to NRO

1 00 — ADIT Demo (NRO Meadows 1 Facility)

2 00

3 00 — Staff Car Return

4 00 — Jack Thorpe/ISAT (TeleCon)

5 00

Notes

jpoindexter

February 13, 2002

Wednesday

February 2002

S	M	T	W	T	F	S
					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28		

March 2002

S	M	T	W	T	F	S
					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30
31						

Ash Wednesday

8 am — 7:00am-9:00am TeleWork (Home)

9 00 — Travel (Home to Work)

10 00 — Dr. John Phillips (DARPA SCIF)

11 00

12 pm — Drive (Langley)

1 00 — Meeting w/ Jim Simon (Chairman, DCI Homeland Security Intelligence Council) (CIA HQ Old 6th Floor)

2 00 — Drive (DARPA)

Updated: TIA System Concept Planning (3701-712 Conference Room (Secret))

Security Brief (DARPA 1014)

3 00

4 00

5 00

(b)(6) 6:30am - 7:00am [redacted]

Notes

jpoindexter

February 14, 2002

Thursday

February 2002

S	M	T	W	T	F	S
					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28		

March 2002

S	M	T	W	T	F	S
					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30
31						

Valentine's Day (United States)

Valentine's Day (United States)

8 am	7:00am-9:00am TeleWork (Home)
9 00	Travel (Home to Work)
10 00	Gina Genton/FBIS (DARPA 723)
11 00	Telecon w/ Brian Mora (DARPA 1014)
12 pm	Reserved (DARPA 1014)
1 00	
2 00	Victoria Stavridou/SRI (DARPA 1014)
3 00	Jaymie Durnan (DARPA 1014)
4 00	
5 00	EARS Planning (1014)
	6:30pm - 9:30pm Dinner with Ariel Merari(ANC)

Notes

jpoindexter

February 15, 2002

Friday

February 2002

S	M	T	W	T	F	S
					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28		

March 2002

S	M	T	W	T	F	S
					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30
31						

8 am
- 7:00am-9:00am TeleWork (Home)

9 00
- Travel (Home to Work)

10 00

11 00
- Robert Parker/ISI (ISI)

12 00

1 00
- Updated: Review Asymmetric Threat Brief for Congressional Staff (Director IAO)
- Reserved (DARPA 1014)

2 00

3 00

4 00

5 00

Notes

jpoindexter

February 18, 2002

Monday

February 2002

S	M	T	W	T	F	S
					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28		

March 2002

S	M	T	W	T	F	S
					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30
31						

IAO Offsite (Lansdown)

President's Day (United States)

President's Day (United States)

8 am — 7:00am-9:00am TeleWork (Home)

IAO Offsite (Lansdowne)

9 00

10 00

11 00

12 pm

1 00

2 00

3 00

4 00

5 00

6:00pm - 7:30pm Dinner(Vintage Room)

Notes

jpoindexter

February 19, 2002

Tuesday

February 2002

S	M	T	W	T	F	S
					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28		

March 2002

S	M	T	W	T	F	S
					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30
31						

IAO Offsite (Lansdown)

8 am — 7:00am-9:00am TeleWork (Home)

IAO Offsite (Lansdowne)

9 00

10 00

11 00

12 pm

1 00

2 00

3 00

4 00

5 00

Notes

jpoindexter

February 20, 2002

Wednesday

February 2002

S	M	T	W	T	F	S
					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28		

March 2002

S	M	T	W	T	F	S
					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30
31						

8 am — 7:00am-9:00am TeleWork (Home)

(b)(6) 9 00 — Updated: continuation--IAO/Charles Wayne re EARS Proposal Seletion (3701-970 DIRO Conference Room (Secret))

10 00 — Travel (Home to Work)

11 00 — Prep for HASC Day (DARPA)

12 00

1 00

2 00

3 00

4 00

5 00

Notes

jpoindexter

February 21, 2002

Thursday

February 2002

S	M	T	W	T	F	S
					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28		

March 2002

S	M	T	W	T	F	S
					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30
31						

8 am — 7:00am-9:00am TeleWork (Home)

run (3701-970 DIRO Conference Room (Secret))

9 00

10 00

11 00

12 pm — Updated: Re: Discuss SCI w/OMO (SCIF, 10th Floor)

1 00

2 00 — MetaCarta (DARPA 1014)

3 00 — HARP (DARPA 1014)

4 00

5 00

7:00am - 8:00am Travel(Home to Work)

Notes

jpoindexter

February 22, 2002

Friday

February 2002

S	M	T	W	T	F	S
					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28		

March 2002

S	M	T	W	T	F	S
					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30
31						

8 am — 7:00am-9:00am TeleWork (Home)

9 00 — Travel (Home to Work)

10 00 — ITO Discussions (TBD)

11 00 — Updated: Treasury Visit

12 pm — Reserved (DARPA 1014)

1 00

2 00 — Rob McHenry (DARPA 1014)

3 00 — Kaye Pulzone - SID (1014)

4 00

5 00

Notes

jpoindexter

February 25, 2002

Monday

February 2002

S	M	T	W	T	F	S
					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28		

March 2002

S	M	T	W	T	F	S
					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30
31						

8 am — 7:00am-9:00am TeleWork (Home)

9 00 — Travel (Home to Work)

10 00 — Prep for SASC Day

11 00 — BAA Review (DARPA 1014)

12 pm

1 00 — Fred Schatzman (DARPA 1014)

2 00

3 00

4 00 — 4:45pm-5:30pm Updated: I Briefings dry run Version II (3701-970 DIRO Conference Room (Secret))

5 00

Notes

jpoindexter

February 26, 2002

Tuesday

February 2002

S	M	T	W	T	F	S
					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28		

March 2002

S	M	T	W	T	F	S
					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30
31						

SASC Day (DARPA)

8 am — 7:00am-9:00am TeleWork (Home)

8 am — Travel (Home to Work)

9 00 — Updated: HASC/SASC- Congressional Committee briefings-Darpa-FNAL AGENDA (3701-970 DIRO Conference Room (Secret))

10 00

11 00

12 pm

1 00

2 00

3 00

4 00

5 00

Notes

jpoindexter

February 27, 2002

Wednesday

February 2002

S	M	T	W	T	F	S
					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28		

March 2002

S	M	T	W	T	F	S
					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30
31						

8 am
- 7:00am-9:00am TeleWork (Home)

9 00
- Travel (Home to Work)

10 00

11 00
- Jim Kelly (DARPA 1014)

12 pm
- IAO Seedlings review (1014)

1 00
- Seedling brief (1014)

2 00
- Groove (DARPA 1014)

3 00
- Brian Sharkey (1014)
- EARS Progress Report (1014)

4 00

5 00

Notes

jpoindexter

February 28, 2002

Thursday

February 2002

S	M	T	W	T	F	S
					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28		

March 2002

S	M	T	W	T	F	S
					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30
31						

8 am — 7:00am-9:00am TeleWork (Home)

9 00 — Travel (Home to Work)

10 00 — Jim Miller (DARPA 1014)

11 00

12 pm — reserved - 1014

1 00

2 00 — Troy Crites (DARPA 1014)

3 00 — System Activities Review (3811 Suite 850)

4 00

5 00

Notes

jpoindexter

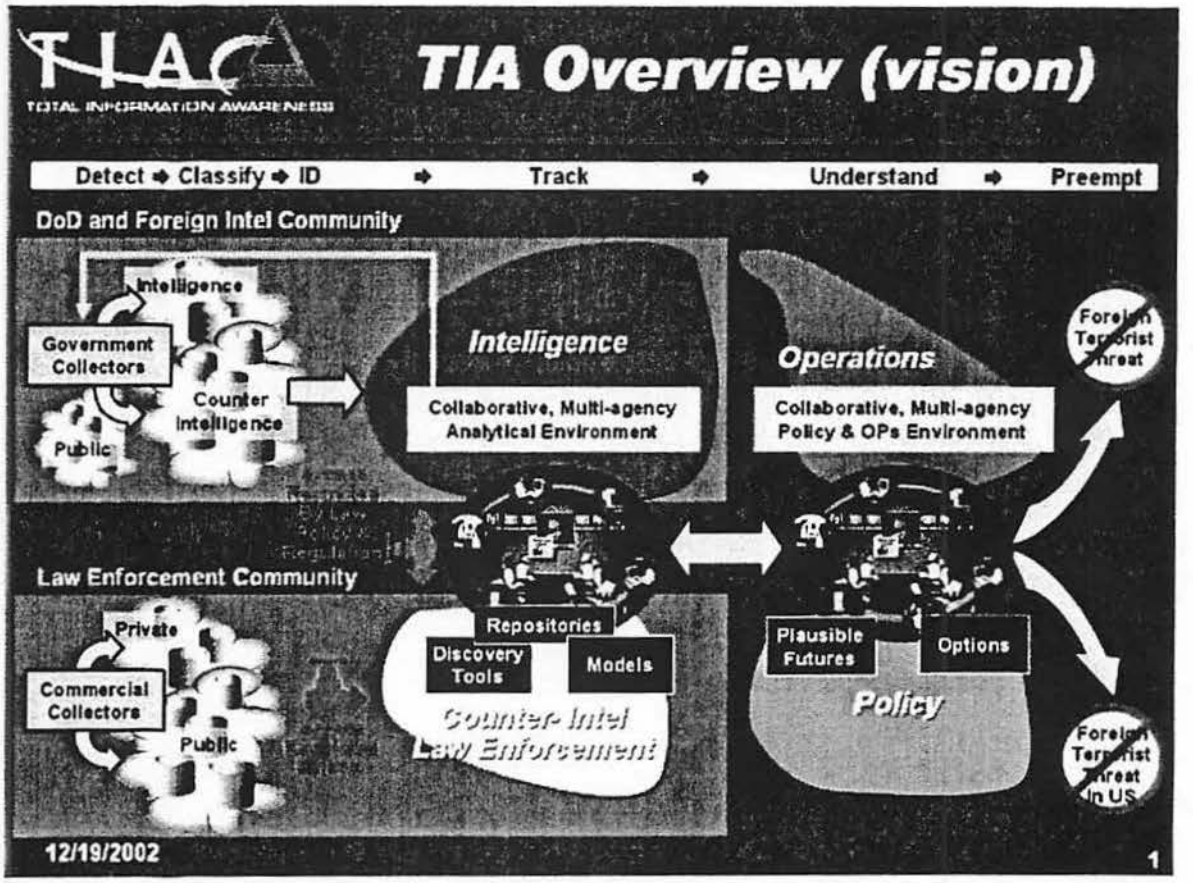
INFORMATION AWARENESS OFFICE

DARPA

Home | News | Programs | Solicitations

TOTAL IN

Total Information Awareness (TIA) System



Program Objective:

The Total Information Awareness (TIA) program is a FY03 new-start program. The goal of the Total Information Awareness (TIA) program is to revolutionize the ability of the United States to detect, classify and identify foreign terrorists – and decipher their plans – and thereby enable the U.S. to take timely action to successfully preempt and defeat terrorist acts. To that end, the TIA program objective is to create a counter-terrorism information system that: (1) increases information coverage by an order of magnitude, and affords easy future scaling; (2) provides focused warnings within an hour after a triggering event occurs or an evidence threshold is passed; (3) can automatically queue analysts based on partial pattern matches and has patterns that cover 90% of all previously known foreign terrorist attacks; and, (4) supports collaboration, analytical reasoning and information sharing so that analysts can hypothesize, test and propose theories and mitigating strategies about possible futures, so decision-makers can effectively evaluate the impact of current or future policies and prospective courses of action.

Program Strategy:

The TIA program strategy is to integrate technologies developed by DARPA (and elsewhere as

appropriate) into a series of increasingly powerful prototype systems that can be stress-tested in operationally relevant environments, using real-time feedback to refine concepts of operation and performance requirements down to the component level. The TIA program will develop and integrate information technologies into fully functional, leave-behind prototypes that are reliable, easy to install, and packaged with documentation and source code (though not necessarily complete in terms of desired features) that will enable the intelligence community to evaluate new technologies through experimentation, and rapidly transition it to operational use, as appropriate. Accordingly, the TIA program will work in close collaboration with one or more U.S. intelligence agencies that will provide operational guidance and technology evaluation, and act as TIA system transition partners.

Technically, the TIA program is focusing on the development of: 1) architectures for a large-scale counter-terrorism database, for system elements associated with database population, and for integrating algorithms and mixed-initiative analytical tools; 2) novel methods for populating the database from existing sources, create innovative new sources, and invent new algorithms for mining, combining, and refining information for subsequent inclusion into the database; and, 3) revolutionary new models, algorithms, methods, tools, and techniques for analyzing and correlating information in the database to derive actionable intelligence.

Planned Accomplishments:

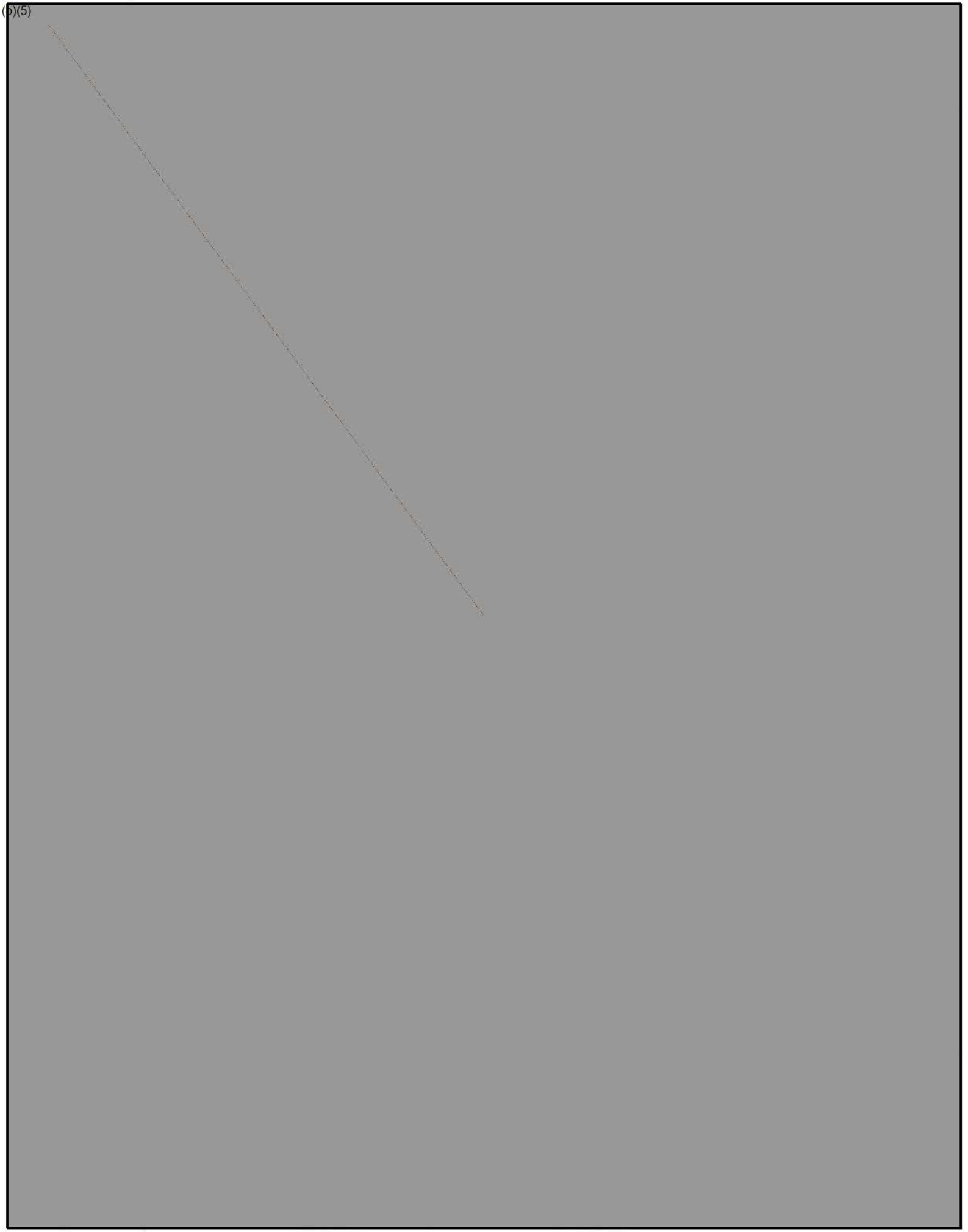
- TBA

More Information

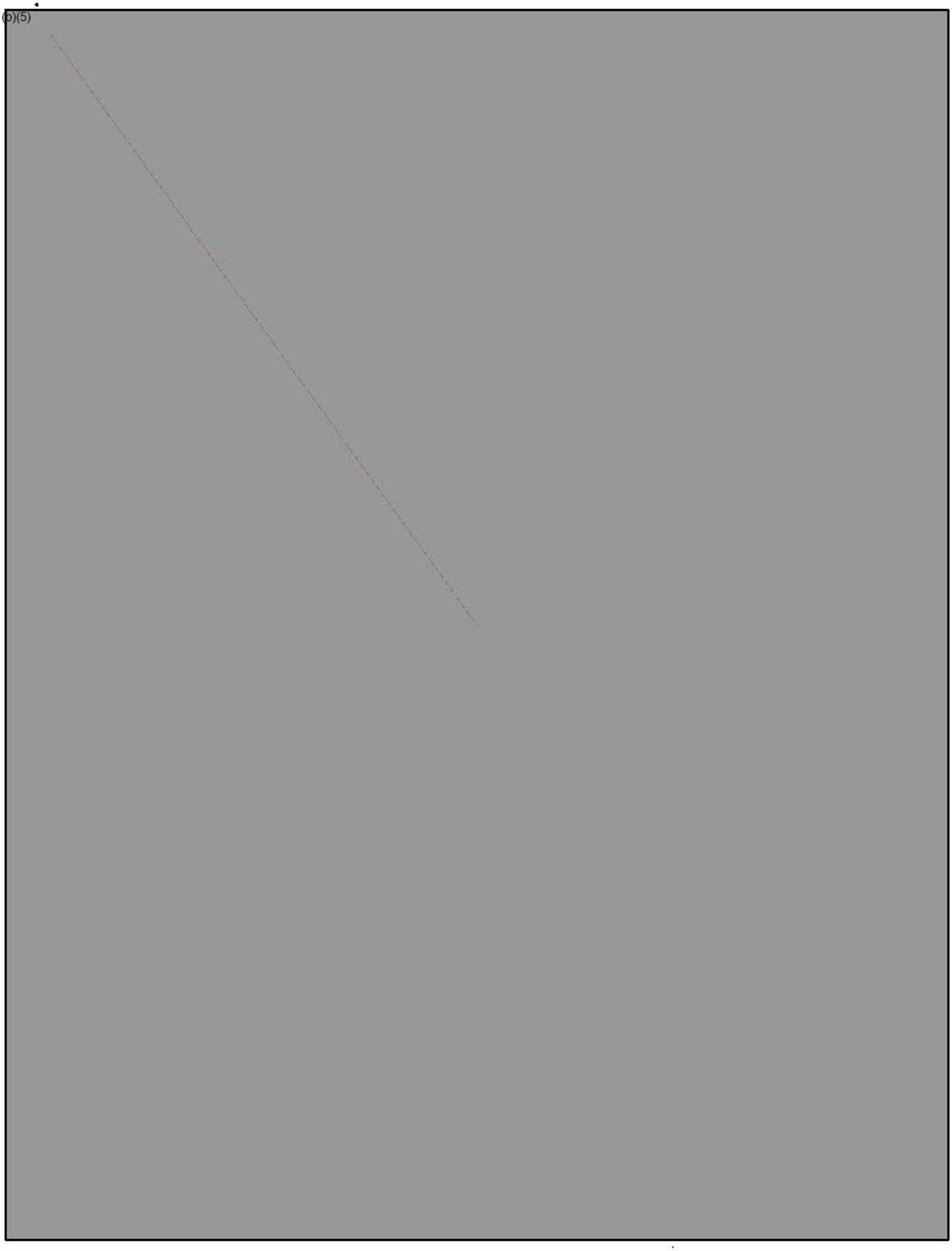
- **More Information**

Home News Programs Solicitations

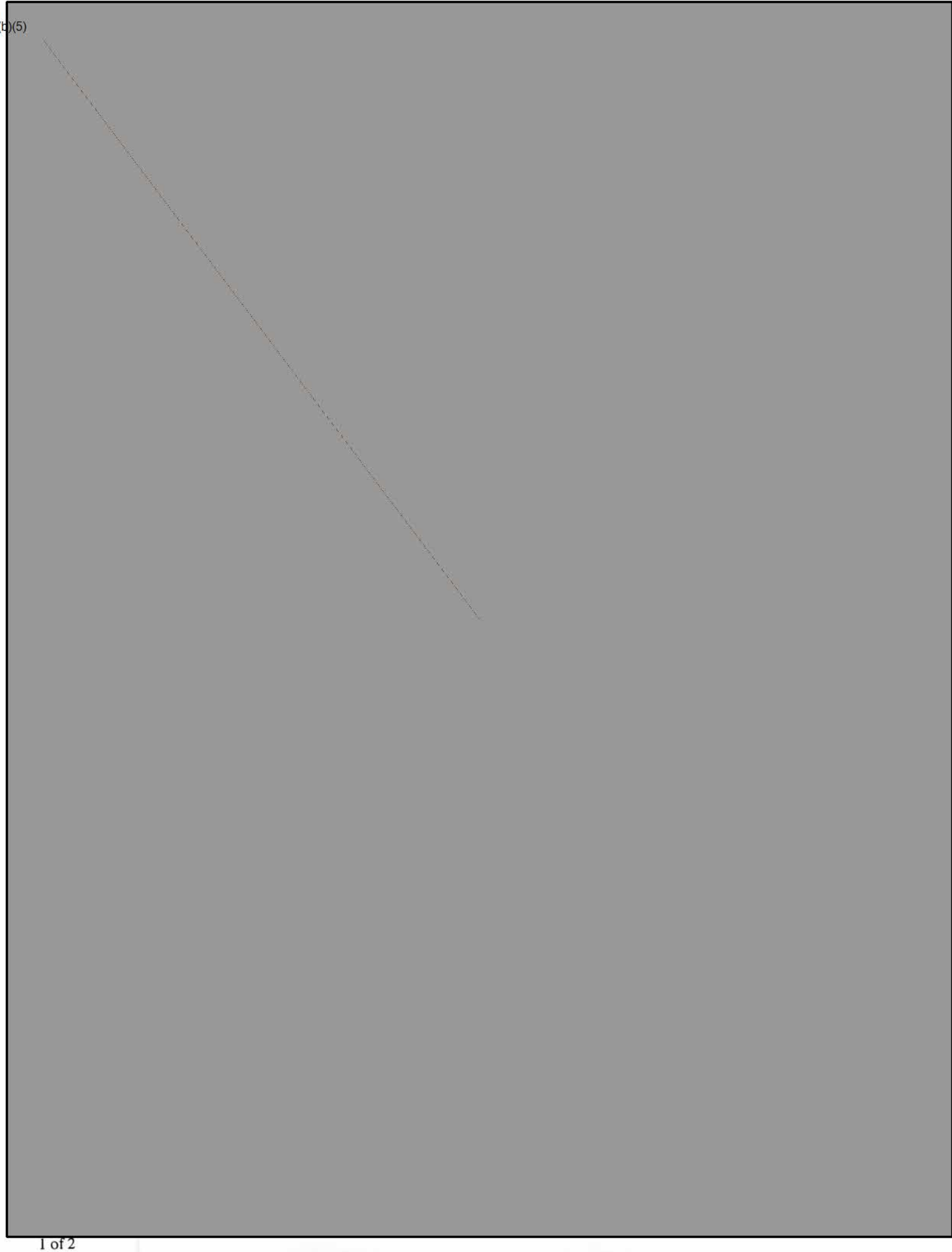
(b)(5)

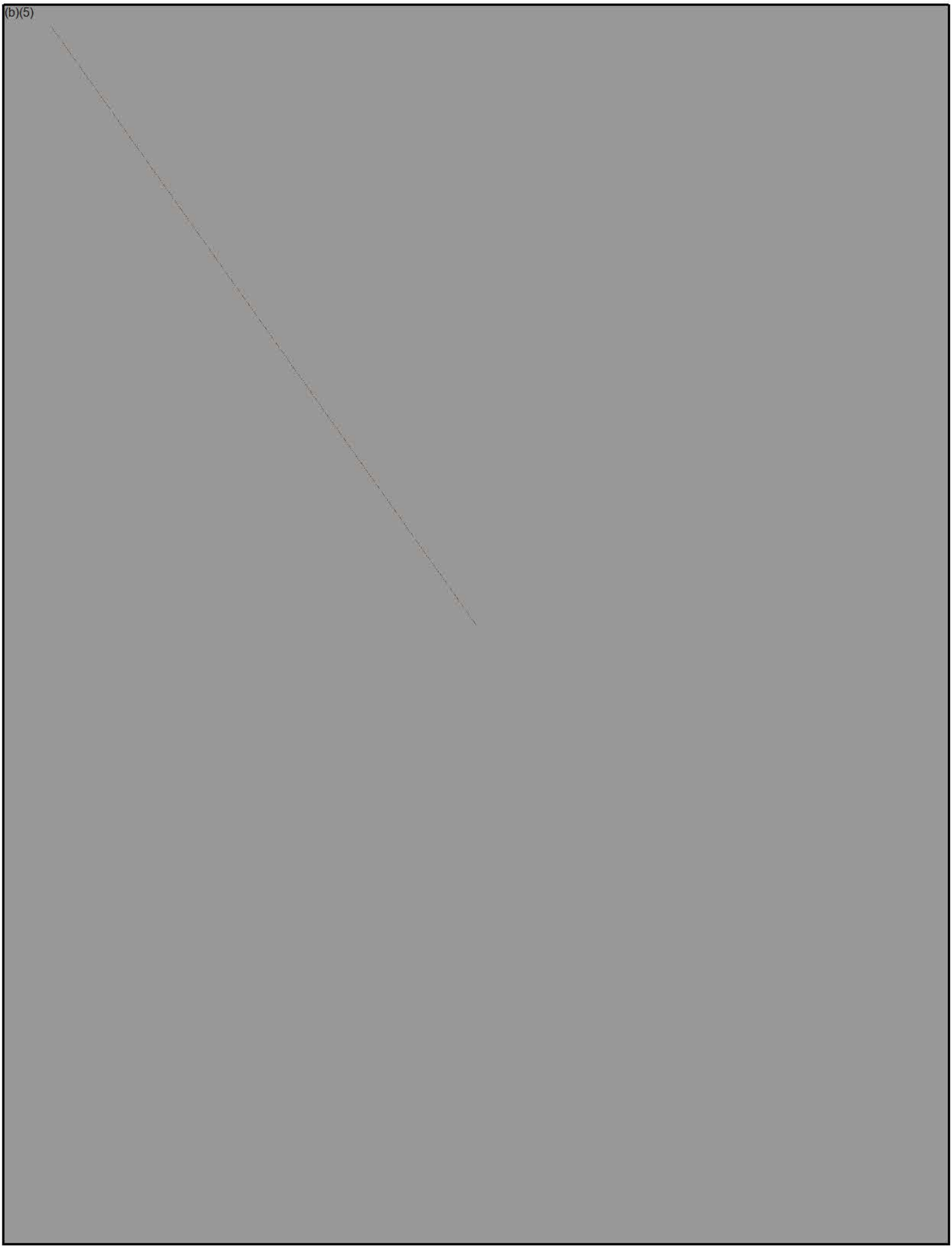


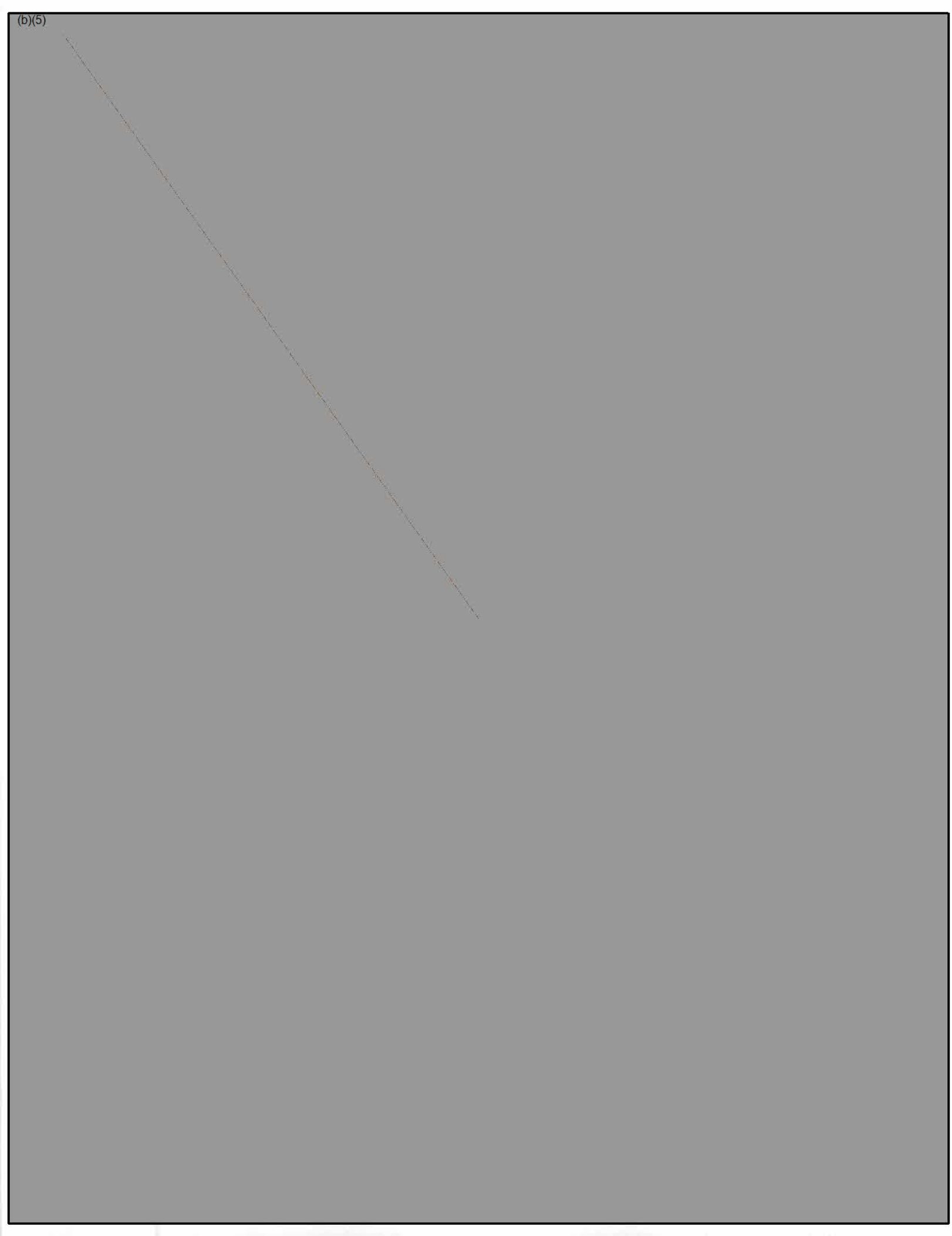
(b)(5)

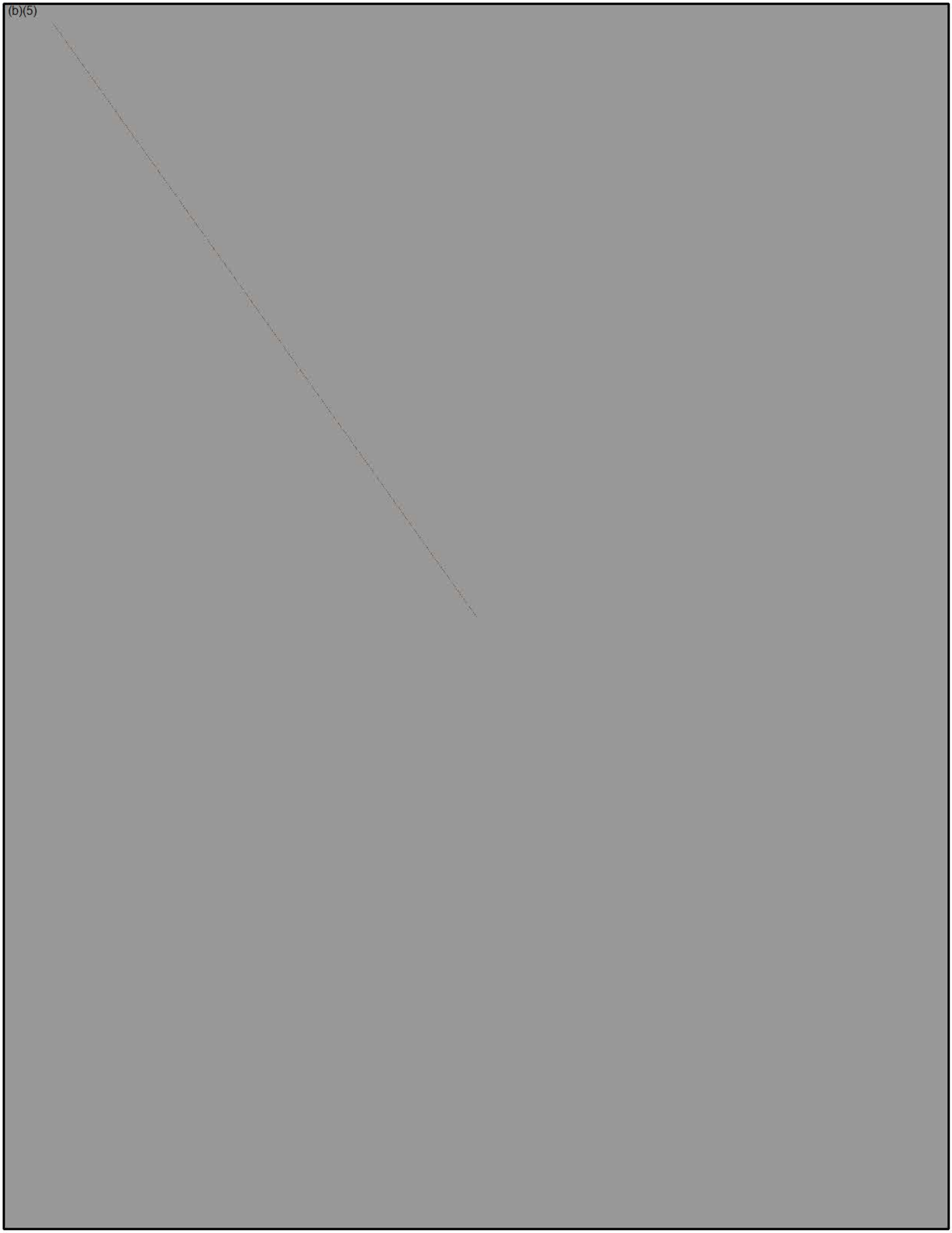


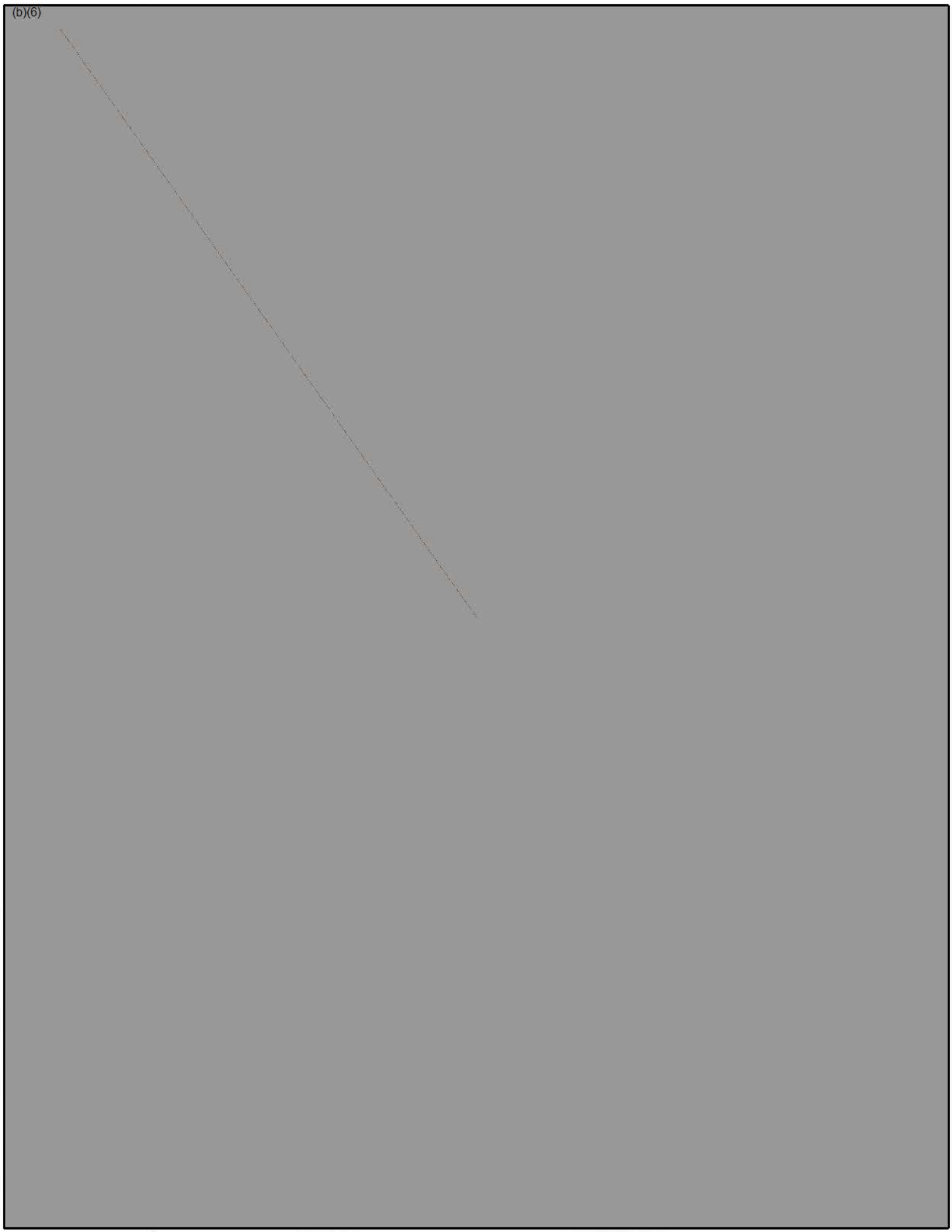
(b)(5)

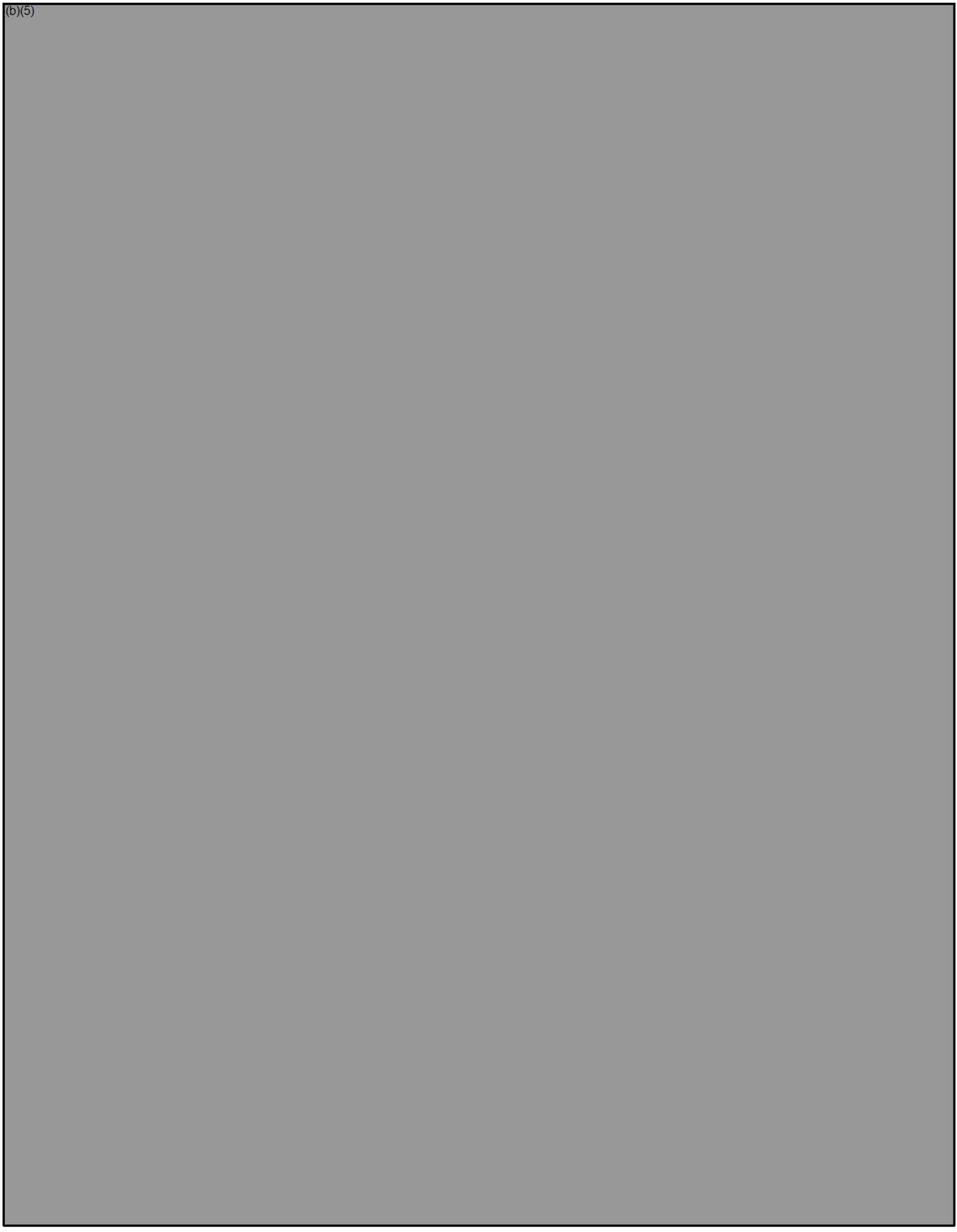


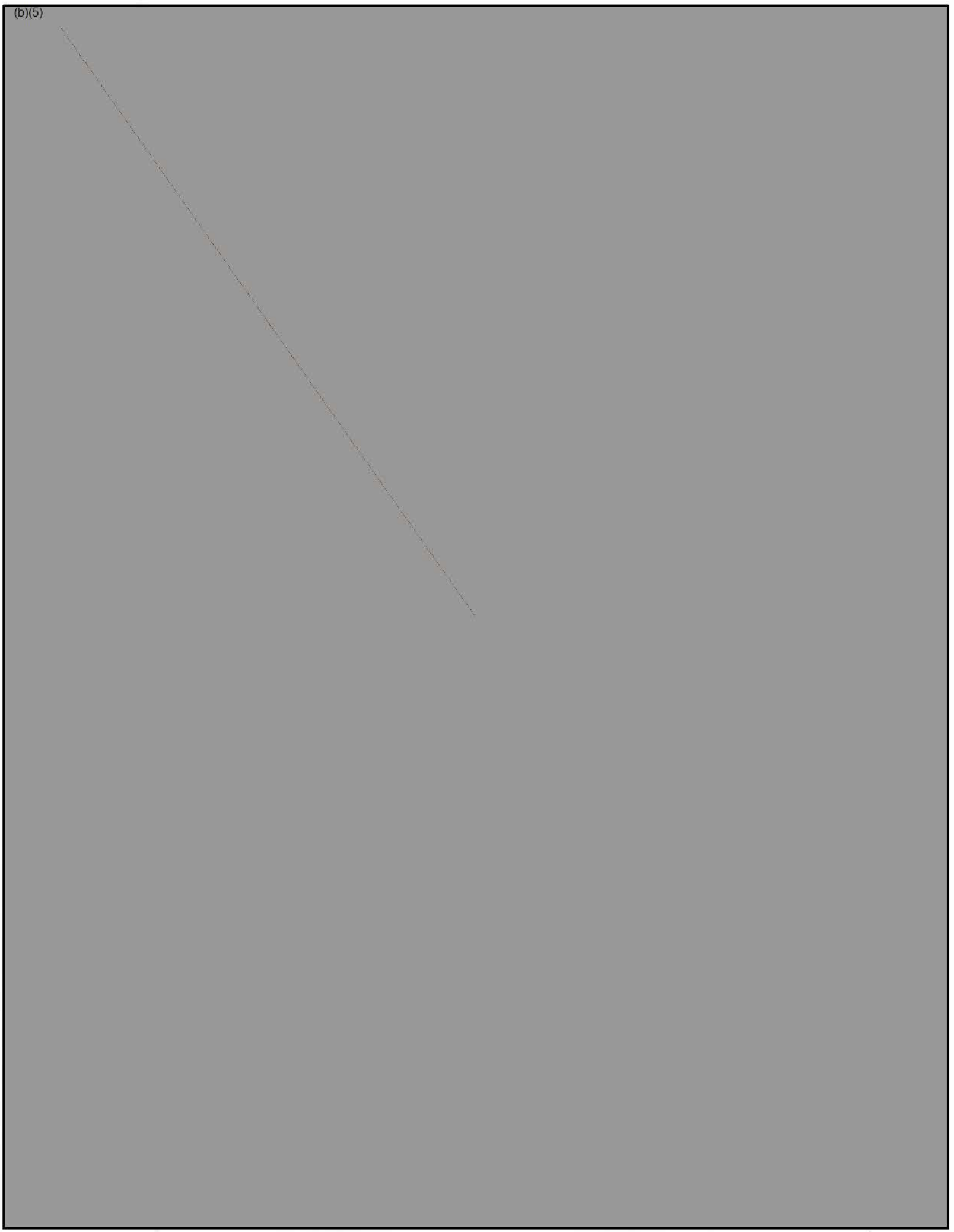

(b)(5)


(b)(5)

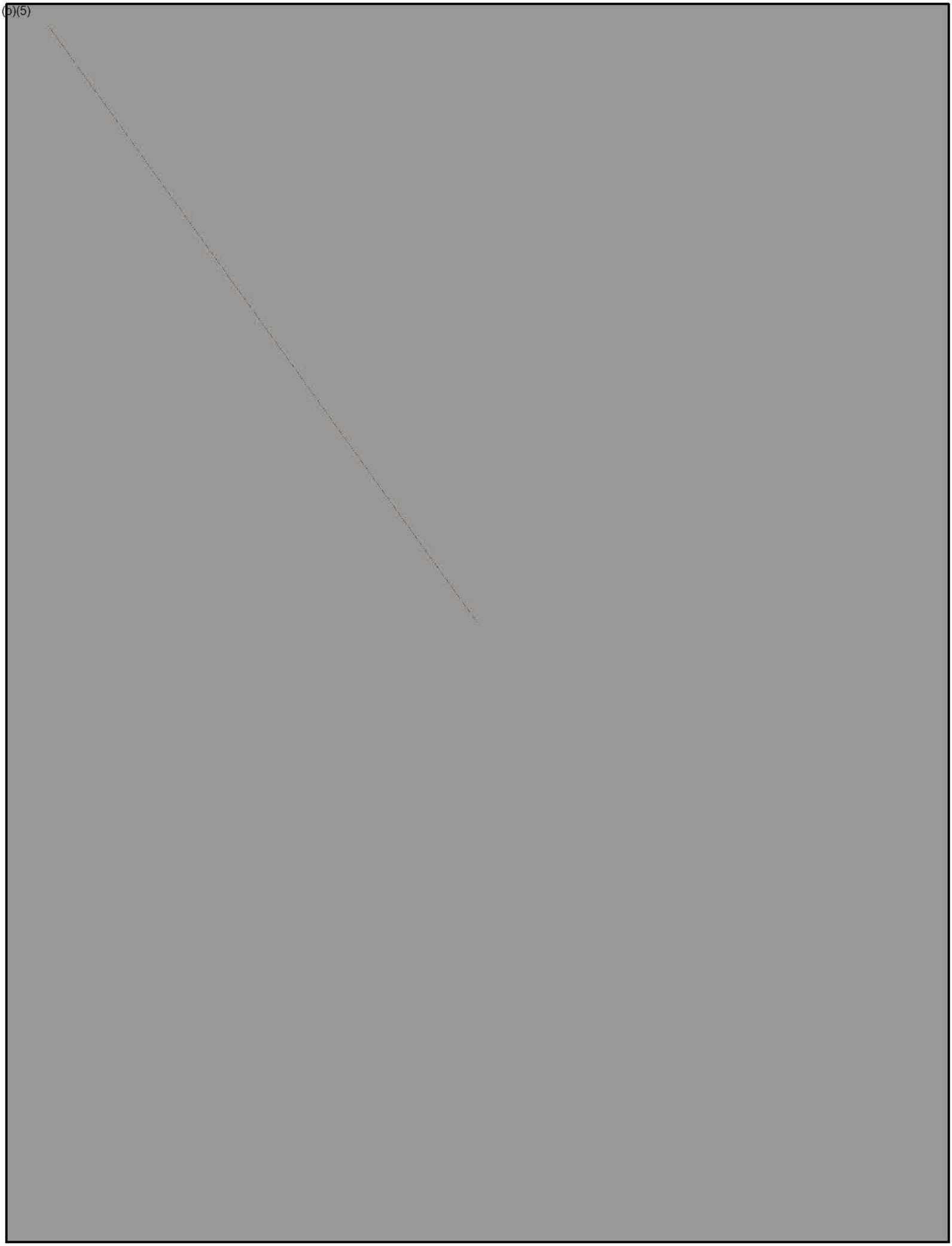

(b)(5)

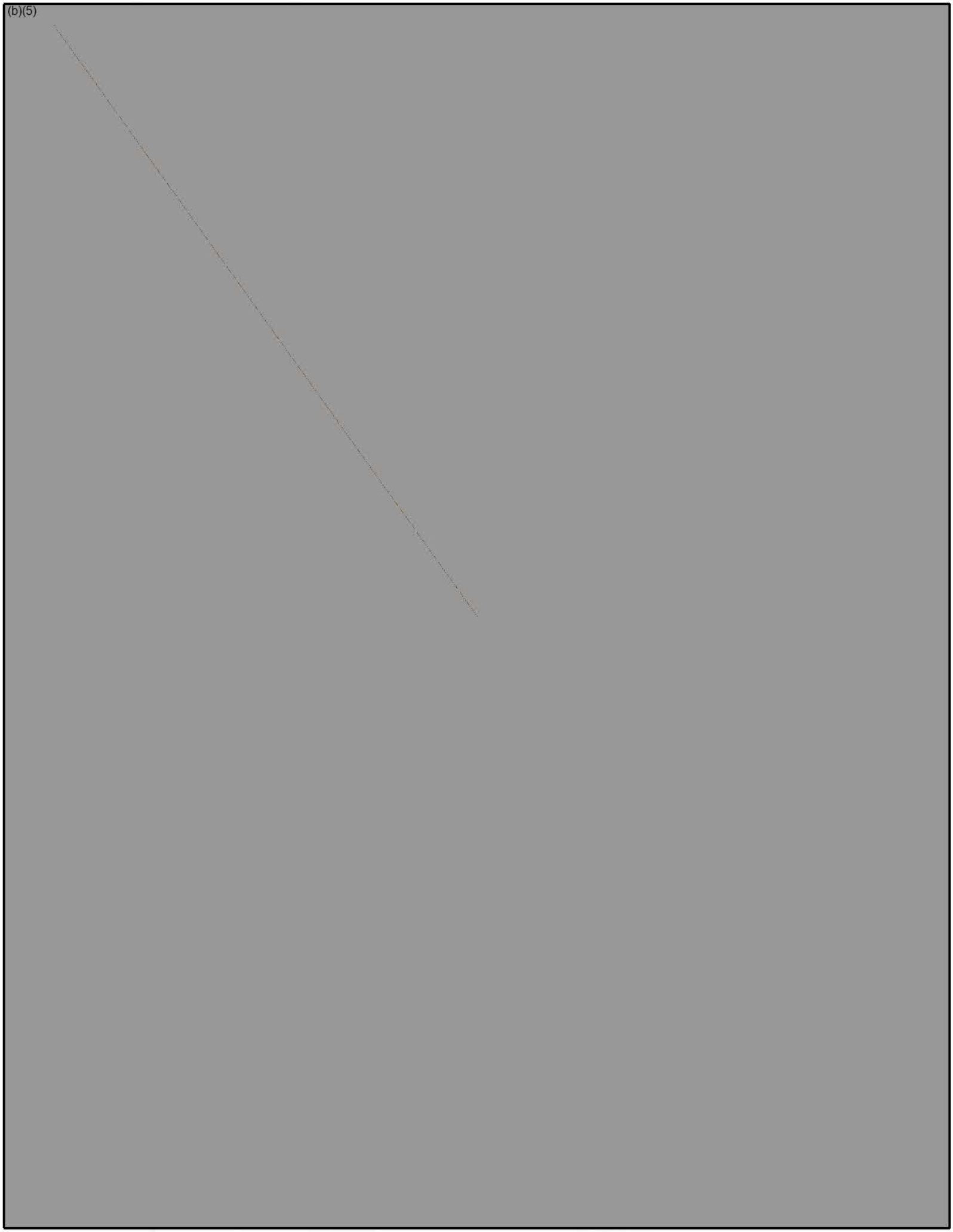

(b)(6)

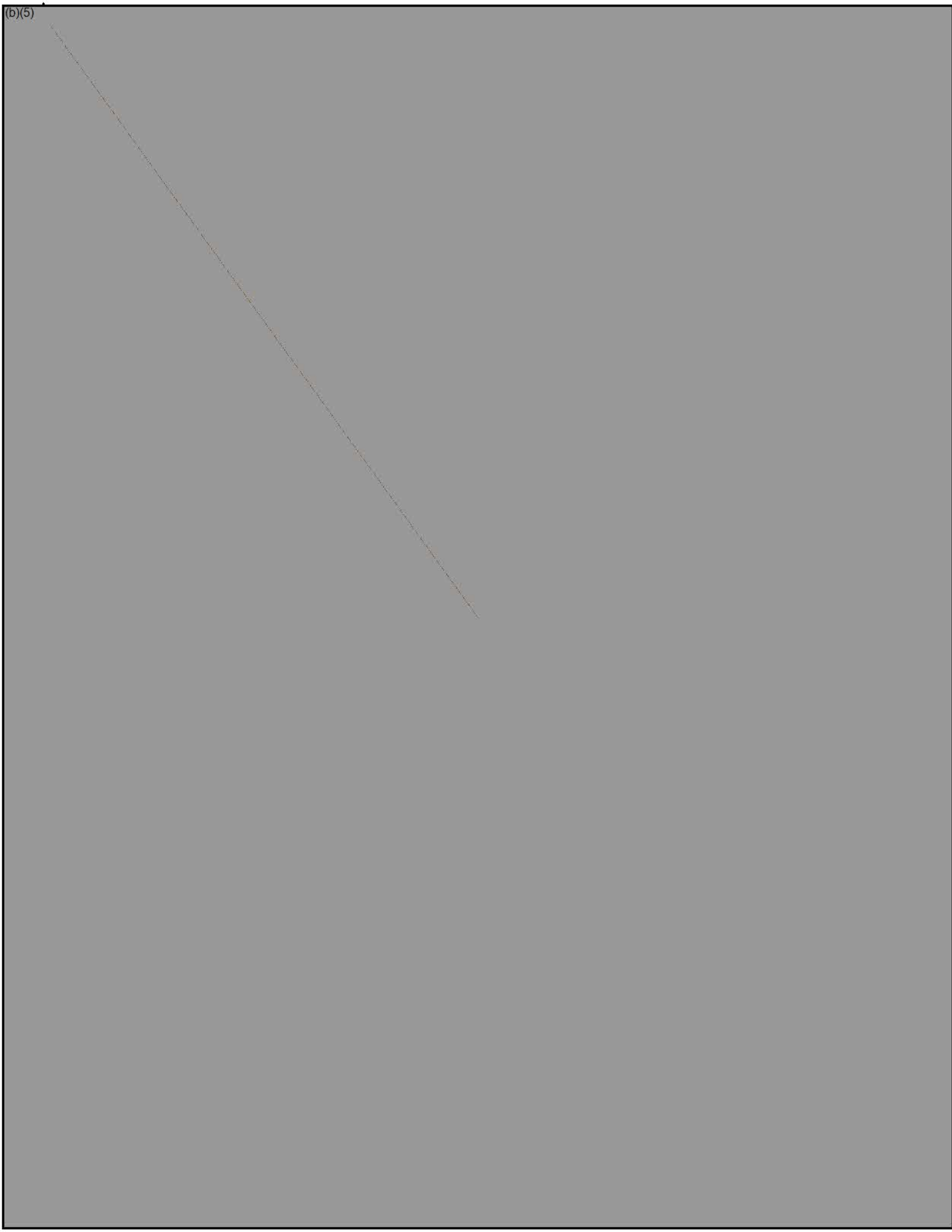

(b)(5)


(b)(5)

(b)(5)




(b)(5)


(b)(5)

March 01, 2002

Friday

March 2002

S	M	T	W	T	F	S
					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30
31						

April 2002

S	M	T	W	T	F	S
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30				

8 am — 7:00am-9:00am TeleWork (Home)

9 00 — Travel (Home to Work)

10 00 — Genomic IT Issues Meeting (1014)

11 00

11:45am-1:15pm Shaun Jones (Gaffneys)

12 00

1 00

2 00

3 00

4 00

5 00 — Meeting (DARPA 970)

Notes

jpolndexter

March 04, 2002

Monday

March 2002

S	M	T	W	T	F	S
					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30
31						

April 2002

S	M	T	W	T	F	S
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30				

8 am	7:00am-9:00am TeleWork (Home)
9 00	Travel (Home to Work)
10 00	Travel
	ADM Bowman Briefing (Wash Navy Yard)
11 00	
	Travel
12 pm	Revise BAA (Home)
1 00	
2 00	
3 00	
4 00	
5 00	

Notes

jpoindexter

March 05, 2002

Tuesday

March 2002

S	M	T	W	T	F	S
					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30
31						

April 2002

S	M	T	W	T	F	S
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30				

(b)(6)

8 am
- 7:00am-9:00am TeleWork (Home)
- 7:30am-8:30am [redacted]
- eMail (Home)

9 00
- Travel (Home to Work)

10 00

11 00
- Rosie Rosenberg (DARPA 1014)

12 pm
- Roy Reed (Gaffney's)

1 00
- Catch Up (1014)

2 00
- Dr. Ron Kostoff/ONR (DARPA 1014)
- Kathleen Kiernan, Chair of the Law Enforcement Working Group (1011)

3 00
- Robert Lautrup (3701-977 DIRO Only Conference Room (Secret))

4 00

5 00

Notes

jpoindexter

March 06, 2002

Wednesday

March 2002

S	M	T	W	T	F	S
					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30
31						

April 2002

S	M	T	W	T	F	S
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30				

8 am
- 6:00am-9:00am Travel to JSOC
- 7:00am-9:00am TeleWork (Home)

9 00
- 9:00am-6:30pm Updated: Visit to JSOC (Fayetteville, NC)

10 00

11 00

12 pm

1 00

2 00

3 00

4 00

5 00

- 6:30pm - 9:00pm Travel back from JSOC

Notes

jpoindexter

March 07, 2002

Thursday

March 2002

S	M	T	W	T	F	S
					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30
31						

April 2002

S	M	T	W	T	F	S
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30				

8 am — 7:00am-9:00am TeleWork (Home)

9 00 — Congressman Weldon (2466 Rayburn House Office Bldg)

10 00 — travel

Reserved - 1014

11 00

12 pm

1 00 — Larry Schuman (1014)

2 00 — DARPA Tech Planning Meeting (1011)

3 00 — System Activities Review (3811 Suite 850)

4 00

5 00

Notes

jpoindexter

March 08, 2002

Friday

March 2002

S	M	T	W	T	F	S
					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30
31						

April 2002

S	M	T	W	T	F	S
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30				

	Car Inspection (Home)
8 am	7:00am-9:00am TeleWork (Home)
9 00	Travel (Home to Work)
10 00	High Performance Computing Modernization Program briefing (3701-970 DIRO Conference Room (Secret))
11 00	Reserved - 1014
12 pm	
1 00	Cy Krolick (Syntek)
2 00	travel from Syntek
	Futures Markets Applied to Prediction (FutureMAP) (1014)
3 00	Dennis Gormley (1014)
4 00	
	Charles
5 00	

Notes

jpoindexter

March 11, 2002

Monday

March 2002

S	M	T	W	T	F	S
					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30
31						

April 2002

S	M	T	W	T	F	S
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30				

8 am	7:00am-9:00am TeleWork (Home)
9 00	Travel (Home to Work)
10 00	DSB (4001 Suite 500 (703)248-7726)
11 00	
12 pm	Fred Schatzman - Rich Taylor (1014 - lunch)
1 00	Mark Lazaroff (1014)
2 00	AlphaTech (1014)
3 00	reserved - 1014
4 00	Danny Hillis/Applied Minds (DARPA 1014)
5 00	

Notes

jpoindexter

March 12, 2002

Tuesday

March 2002

S	M	T	W	T	F	S
					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30
31						

April 2002

S	M	T	W	T	F	S
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30				

8 am	7:00am-9:00am TeleWork (Home)
9 00	Travel (Home to Work)
10 00	John Peterson, Arlington Institute (DARPA)
11 00	11:15am-12:15pm Jaymie Dernan (3E944)
12 pm	reserved - 1014
1 00	Review Briefs Planned for 14 Mar Aldrigde visit
2 00	Manny Aparicio (DARPA 1014)
3 00	reserved - 1014
4 00	Conference call with Mike McConnell
5 00	

Notes

jpoindexter

March 13, 2002

Wednesday

March 2002

S	M	T	W	T	F	S
					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30
31						

April 2002

S	M	T	W	T	F	S
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30				

8 am	7:00am-9:00am TeleWork (Home)
9 00	Travel (Home to Work)
10 00	IAO Meeting - NSA Cook (3701-1043 Conference Room (Unclassified))
11 00	11:30am-11:45am Nancy Kassner - Director DARPA Security (1014)
12 pm	Shaun Jones and Heather (Gaffney's)
1 00	Ravi Kanuri - Lockheed Martin (977 - Need LCD)
2 00	reserved - 1014
3 00	Tom Fetterolf (1014)
4 00	
5 00	

Notes

jpoindexter

March 14, 2002

Thursday

March 2002

S	M	T	W	T	F	S
					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30
31						

April 2002

S	M	T	W	T	F	S
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30				

8 am — 7:00am-9:00am TeleWork (Home)

9 00 — Travel (Home to Work)

10 00

10:30am-10:45am Barbara Meyrowitz - CMO's IAO rep (1014)

11 00 — reserved - 1014

12 pm — travel

1 00 — Updated: Pete Aldridge - Space Brief and More (3701-970 DIRO Conference Room (Secret))

(b)(3) 50 USC §403 (g) Section 6

travel

2 00

3 00 — System Activities Review (3811 Suite 850)

4 00

5 00

Notes

jpoindexter

March 15, 2002

Friday

March 2002

S	M	T	W	T	F	S
					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30
31						

April 2002

S	M	T	W	T	F	S
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30				

SAC Day (DARPA)

8 am
- 7:00am-9:00am TeleWork (Home)
- Travel (Home to Work)

9 00
- Updated: SAC- David Morison Congressional Committee briefings-Darpa (3701-970 DIRO Conference Room (Secret))

10 00
- Bruce Gibeson "movie guy" (1014)

11 00

12 PM

1 00

2 00

3 00
- Bryan Gabbard/ARETE (DARPA 1014)

4 00
- Tip Clifton 703-414-9630 (1014)

5 00
- Discusss Security Support to TTO and IAO (983)

Notes

jpoindexter

March 18, 2002

Monday

March 2002

S	M	T	W	T	F	S
					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30
31						

April 2002

S	M	T	W	T	F	S
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30				

8 am — 7:00am-9:00am TeleWork (Home)

9 00 — Travel (Home to Work)

10 00 — SIPRnet Install (DARPA 1014)

11 00 — Dr. Ed Thompson (1014)

12 pm — 12:15pm-1:15pm Senior Management Mtg (3701-970 DIRO Conference Room (Secret))

1 00 — TSA-FAA (3701-977 DIRO Only Conference Room (Secret))

2 00 — on travel - US Airways flight 6420 departs Washington Reagan (DCA) at 1630 arrives Boston (BOS) at 1756. (Hotel: Royal Sonesta (617) 806-4200, 5 Cambridge Parkway, Cambridge, MA 02142)

3 00 — ISAT (Conference Call 800-457-0265 969081)

4 00

5 00

Notes

jpoindexter

March 19, 2002

Tuesday

March 2002

S	M	T	W	T	F	S
					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30
31						

April 2002

S	M	T	W	T	F	S
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30				

8 am — 7:00am-9:00am TeleWork (Home)

TIDES Reviews at MITRE POC:
Lynette Hirschman, 781-271-7789 &
BBN POC John Makhoul,
617-873-3332 (Bedford & Cambridge)

9 00

10 00

11 00

12 pm

1 00

2 00

3 00

4 00

5 00 — 5:00pm-9:00pm travel - US Air flight 6535 departs BOS at 1900 arrives DCA at 2044.

Notes

jpoindexter

March 20, 2002

Wednesday

March 2002

S	M	T	W	T	F	S
					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30
31						

April 2002

S	M	T	W	T	F	S
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30				

8 am	7:00am-9:00am TeleWork (Home)
9 00	Travel (Home to Work)
10 00	Abe Wagner (DARPA 1014)
	Yaneer Bar-Yam (DARPA 1014)
11 00	11:15am-1:00pm reserved - 1014
	11:15am-11:30am Updated: Chat (101
12 pm	
1 00	Todd Carrico, Ph.D. (1014)
2 00	Karen Stephenson (DARPA 1014)
3 00	Updated: J.Foster Discussions (983)
4 00	Marjory Blumenthal - National Academy of Sciences (1014)
5 00	

Notes

jpoindexter

March 21, 2002

Thursday

March 2002

S	M	T	W	T	F	S
					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30
31						

April 2002

S	M	T	W	T	F	S
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30				

8 am — 7:00am-9:00am TeleWork (Home)

9 00 — Travel (Home to Work)

10 00 — Chuck Hoover (1014)

Dick O'Neill - The Highlands Forum (1014)

11 00

12 pm — reserved - 1014

travel - G1

1 00 — Carl Hunt (GMU - (703) 993-4333)

2 00

3 00 — System Activities Review (3811 Suite 850)

travel

4 00

(b)(6) [redacted]

5 00

Notes

jpoindexter

March 22, 2002

Friday

March 2002

S	M	T	W	T	F	S
					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30
31						

April 2002

S	M	T	W	T	F	S
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30				

8 am — 7:00am-9:00am TeleWork (Home)

9 00 — Travel (Home to Work)

10 00

Scott Fisher (DARPA 1014)

11 00

Mark Lazaroff - DSB slides

12 pm

1 00 — Tom Conroy and Howard Schue (1014)

2 00 — George Singley and Duane Andrews (1014)

(b)(6) 3 00 — travel to [redacted]

(b)(6) [redacted]

4 00

5 00

Notes

jpoindexter

March 25, 2002

Monday

March 2002

S	M	T	W	T	F	S
					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30
31						

April 2002

S	M	T	W	T	F	S
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30				

8 am — 7:00am-9:00am TeleWork (Home)

9 00 — Travel (Home to Work)

10 00 — Paul Kozemchak (DARPA - 1014)

Imagining Time - 1014

11 00

12 pm — car to Ft, Meade

1 00 — (b)(3):50 USC §402 Note [redacted]

2 00

3 00

4 00

5 00 — car return

Notes

jpoindexter

March 26, 2002

Tuesday

March 2002

S	M	T	W	T	F	S
					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30
31						

April 2002

S	M	T	W	T	F	S
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30				

8 am — 7:00am-9:00am TeleWork (Home); Travel (Home to Work)

9 00 — JCAG/OSD(C3I) includes transportation both ways (Crystal City)

10 00

11 00 — Imagining Time - 1014

12 pm — Lunch: Poindexter, Wishner (Flat Top Grill)

1 00 — Program Update - Larry Willis (1014)

2 00 — Tom Meyer (1014)

3 00 — Imagining Time - 1014

4 00 — Charles Wayne (TeleCon)

Stephen Cross - SEI CMU (1011)

5 00

Notes

jpoindexter

March 27, 2002

Wednesday

March 2002

S	M	T	W	T	F	S
					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30
31						

April 2002

S	M	T	W	T	F	S
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30				

8 am	7:00am-9:00am TeleWork (Home)
9 00	Travel (Home to Work)
10 00	DSB/Precision Compellence (4001 500)
11 00	
	travel to Rand
12 pm	Rand (Rand)
1 00	
2 00	
3 00	
4 00	
5 00	

Notes

jpoindexter

March 28, 2002

Thursday

March 2002

S	M	T	W	T	F	S
					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30
31						

April 2002

S	M	T	W	T	F	S
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30				

8 am 7:00am-9:00am TeleWork (Home)

9 00 Travel (Home to Work)

10 00 10:00am-10:50am Air Force SAB Meeting - Maj. Kimmelman (1560 Wilson Blvd. room 430 - 703-243-0016)

11 00 11:00am-11:15am travel

11:15am-12:30pm Roy Reed, Brian Sharkey and Rich Peze (1014 - lunch from downstairs)

12 pm Doug Cupo and Sandy Mangold, Integrated ISR at National Security Space Architect (3701-SSCO SMALL Conference Room (SCI/SAP))

1 00

2 00 Dr. Ted Drzewiecki (1014)

3 00 System Activities Review (3811 Suite 850)

4 00

5 00 Front Office Review

Notes

jpoindexter

March 29, 2002

Friday

March 2002

S	M	T	W	T	F	S
					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30
31						

April 2002

S	M	T	W	T	F	S
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30				

(b)(6)

8 am — 7:00am-9:00am TeleWork (Home)

9 00

10 00

11 00

12 pm

1 00

2 00

3 00

4 00

5 00

Notes

jpoindexter

April 01, 2002

Monday

April 2002

S	M	T	W	T	F	S
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30				

May 2002

S	M	T	W	T	F	S
			1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30	31	

(b)(6)

8 am 7:00am-9:00am TeleWork (Home)

9 00

10 00

11 00

12 pm

1 00

2 00

3 00

4 00

5 00

Notes

jpoindexter

April 02, 2002

Tuesday

April 2002

S	M	T	W	T	F	S
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30				

May 2002

S	M	T	W	T	F	S
			1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30	31	

Tax Prep (Home)

8 am — 7:00am-9:00am TeleWork (Home)

9 00

10 00

11 00

(b)(6)

12 pm — Steve Dorfman (Home)

1 00

2 00

3 00

4 00

5 00

Notes

jpoindexter

April 03, 2002
Wednesday

April 2002

S	M	T	W	T	F	S
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30				

May 2002

S	M	T	W	T	F	S
			1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30	31	

8 am — 7:00am-9:00am TeleWork (Home)

9 00 — Updated: JWAC-Blue (Dahlgren)

10 00

11 00

12 pm

1 00

2 00

3 00

4 00

5 00

Notes

jpoindexter

April 04, 2002

Thursday

April 2002

S	M	T	W	T	F	S
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30				

May 2002

S	M	T	W	T	F	S
			1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30	31	

Bluebird Bottom

8 am — 7:00am-9:00am TeleWork (Home)

9 00 — Travel (Home to Work)

10 00 — travel

Updated: FTTTF - Karen Waterman (Crystal City)

11 00 — Imagining Time - 1014

12 pm

1 00 — Gary Koob, ITO - Discussions (1014)

2 00 — 2:00pm-2:45pm Bob Schrag - IET

3 00 — System Activities Review (3811 Suite 850)

4 00

5 00

Notes

jpoindexter

April 05, 2002
Friday

April 2002

S	M	T	W	T	F	S
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30				

May 2002

S	M	T	W	T	F	S
			1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30	31	

8 am
- 7:00am-9:00am TeleWork (Home)

9 00
- Travel (Home to Work)

10 00
- Updated: DARPA brief to Creighton Greene brifings (SASC) (3701-SSCO LARGE Conference Room (SCI/SAP))
- Mike Siracuse (DARPA 1014)

11 00

12 pm
- Imagining Time - 1014

1 00
- BAA Source Selection Process (3701-712 Conference Room (Secret))

2 00
- Leanne (1014)
- Updated: Review Session 4/5 for - Newt Gingrich visit 4/9 (3701-970 DIRO Conference Room (Secret))

3 00

4 00
- Tip Clifton (1014)

5 00

Notes

jpoindexter

April 08, 2002

Monday

April 2002

S	M	T	W	T	F	S
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30				

May 2002

S	M	T	W	T	F	S
			1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30	31	

8 am	7:00am-9:00am TeleWork (Home)
9 00	Travel (Home to Work)
10 00	Richard Williams, OSD C3I Security and William Coleman, OSD C3I (1014)
11 00	Imagining Time - 1014
	travel
12 pm	Intelligent Information Factory (IIF) demo - DARPA van reserved (Global Vision Center at the Crystal City facility - 3rd floor)
1 00	
2 00	
	travel
3 00	Craig Partridge Mtg (1014)
4 00	Tom Garwin - Hicks & Associates (1014)
5 00	5:00pm-5:15pm Kristen Numrych (1014)

Notes

April 09, 2002

Tuesday

April 2002

S	M	T	W	T	F	S
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30				

May 2002

S	M	T	W	T	F	S
			1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30	31	

8 am	7:00am-9:00am TeleWork (Home)
9 00	Travel (Home to Work)
10 00	Raytheon - Greg Shelton, Sr. VP of Information Technology and Mel Peters 703-284
	Imagining Time - 1014
11 00	
12 pm	Richard Perle (983)
1 00	Tony Oettinger - Conference Call, Tony will initiate (1014)
2 00	Briefing for Newt Gingrich (DARPA)
3 00	
4 00	
5 00	

Notes

jpoindexter

April 10, 2002

Wednesday

April 2002

S	M	T	W	T	F	S
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30				

May 2002

S	M	T	W	T	F	S
			1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30	31	

8 am — 7:00am-9:00am TeleWork (Home)

9 00 — Travel (Home to Work)

10 00 — Proposal for Institute for Physical Sciences (3701-712 Conference Room (Secret))

11 00 — Imagining Time - 1014

12 pm

1 00

2 00 — Brian Sharkey (4001 - 450 4th floor)

3 00

4 00

5 00

Notes

jpoindexter

April 11, 2002

Thursday

April 2002

S	M	T	W	T	F	S
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30				

May 2002

S	M	T	W	T	F	S
			1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30	31	

8 am — 7:00am-9:00am TeleWork (Home)

9 00 — Travel (Home to Work)

10 00 — Mike McConnell and Art Fritzson (1014)

11 00 — Imagining Time

12 pm — travel

(b)(3) 10 USC §424 (b)(3) 10 USC §455 (b)(3)

1 00 — NRO - [redacted]

2 00 — travel

3 00 — System Activities Review (3811 Suite 850)

4 00

5 00

Notes

jpoindexter

April 12, 2002

Friday

April 2002

S	M	T	W	T	F	S
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30				

May 2002

S	M	T	W	T	F	S
			1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30	31	

Moving Day

8 am — 7:00am-9:00am TeleWork (Home)

9 00 — Travel (Home to Work)

10 00 — Seedlings (1014)

Jim Hooley (1014)

11 00 — Imagining Time

12 pm — Move Office to 6th Floor

1 00 — Tim Hoechst, VP of Technology Oracle (3701-712 Conference Room (Secret))

2 00

3 00

4 00

5 00

Notes

jpoindexter

April 15, 2002
Monday

April 2002

S	M	T	W	T	F	S
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30				

May 2002

S	M	T	W	T	F	S
			1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30	31	

Tax Day (United States)
Tax Day (United States)

8 am — 7:00am-9:00am TeleWork (Home)

9 00 — Travel (Home to Work)

10 00 — Imagining Time - 675

11 00 — Mike Siracuse (DARPA)

12 pm — 12:15pm-1:15pm Senior Management Mtg (3701-970 DIRO Conference Room (Secret))

1 00 — lunch

2 00 — Richard Clark, DoD E-Business (675)

Imagining Time - 675

3 00

4 00 — IDA SENSE Program - Dr. Richard H. White (675)

5 00

Notes

jpoindexter

April 16, 2002

Tuesday

April 2002

S	M	T	W	T	F	S
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30				

May 2002

S	M	T	W	T	F	S
			1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30	31	

8 am — 7:00am-9:00am TeleWork (Home)

9 00 — DARPA Car to Ft. Washington Meeting with Rich Peze (Parking Level G1)

9 00 — Travel (Home to Work)

10 00 — Rich Peze - DARPA car reserved (Ft. Washington)

11 00

12 pm

1 00 — travel

1 00 — Updated: Michael Saylor (MicroStrategy CEO) (3701-723 Conference Room (Secret))

2 00

3 00 — Dr. Gerald Post Brief (712)

4 00 — Craig Fields (DARPA)

5 00

Notes

jpoindexter

April 17, 2002
Wednesday

April 2002

S	M	T	W	T	F	S
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30				

May 2002

S	M	T	W	T	F	S
			1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30	31	

8 am	7:00am-9:00am TeleWork (Home)
9 00	Travel (Home to Work)
10 00	DSB Nuclear Weapons - brief from 10:30 - 12:00 (4001 - 4th floor)
11 00	
12 pm	Imagining Time - 675
1 00	
2 00	William Wright - 703-414-1054 (675)
	Prof. Zue, Director, MIT Laboratory for Computer Science
3 00	
	Dr. Joe Mitola (675)
4 00	Paul Grazinina, President - Analytical Graphics, 1-800-220-4785 (1014)
	Updated: Service Chiefs Program (IAO) (Dr. Poindexter's Office)
5 00	

Notes

jpoindexter

April 18, 2002

Thursday

April 2002

S	M	T	W	T	F	S
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30				

May 2002

S	M	T	W	T	F	S
			1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30	31	

ISAT (IDA)

8 am	7:00am-9:00am TeleWork (Home)
9 00	Travel (Home to Work)
10 00	ISAT TIA Briefing (IDA)
11 00	
12 pm	travel
	lunch
1 00	Ken Johnson (DARPA)
2 00	Paul Johns - AUSA, TVCS (675)
3 00	System Activities Review (3811 Suite 850)
4 00	
5 00	

Notes

jpoindexter

April 19, 2002

Friday

April 2002

S	M	T	W	T	F	S
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30				

May 2002

S	M	T	W	T	F	S
			1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30	31	

	Naval Academy Foundation (Annapolis)
8 am	7:00am-9:00am TeleWork (Home)
9 00	Travel (Annapolis)
10 00	NAF Exec Comte (USNA Officers Club)
11 00	
12 pm	
1 00	
2 00	
3 00	
4 00	
5 00	
	6:30pm - 10:30pm NAF Dinner(Loews Annapolis)

Notes

jpoindexter

April 22, 2002
Monday

April 2002

S	M	T	W	T	F	S
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30				

May 2002

S	M	T	W	T	F	S
			1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30	31	

8 am	7:00am-9:00am TeleWork (Home)
	Travel (Home to Work)
9 00	Updated: IAO information visualization environment meeting (3701-643 Conference Room (Secret))
10 00	IAO Staff Meeting (3701-643 Conference Room (Secret))
11 00	
12 pm	IAO Luncheon (643)
1 00	Imagining Time - 675
	1:15pm-1:30pm Updated: meet w/Dr.
2 00	travel
(b)(6) 3 00	Dragon Exp - [redacted] (INSCOM)
4 00	
5 00	travel
	6:30pm - 9:30pm Father Mike Dinner(Home)

Notes

jpoindexter

April 23, 2002
Tuesday

April 2002

S	M	T	W	T	F	S
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30				

May 2002

S	M	T	W	T	F	S
			1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30	31	

8 am	7:00am-9:00am TeleWork (Home)
9 00	Travel (Home to Work)
10 00	
11 00	Imagining Time - 675
12 pm	BAA Review Meeting (675)
1 00	Dave Signori - Rand (675)
2 00	Jeff Gaynor - Special Assistant for Homeland Security DASD (Security and Information Operations)OASD(C3I) (1014 - (703) 697-2242)
3 00	hold
	Dr. Ron LaPorte (675)
4 00	Clint Kelly (head of IR&D investments for SAIC), Gordon Oehler and Brian Sharkey (675)
5 00	

Notes

jpoindexter

April 24, 2002
Wednesday

April 2002

S	M	T	W	T	F	S
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30				

May 2002

S	M	T	W	T	F	S
			1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30	31	

8 am — 7:00am-9:00am TeleWork (Home)

9 00 — Travel (Home to Work)

10 00 — Andy Marshall and Jim Thomas (675)

11 00 — Imagining Time - 1014

travel

12 pm — Raytheon (Reston)

1 00

2 00

3 00

4 00 — travel

5 00

Notes

jpoindexter

April 25, 2002
Thursday

April 2002

S	M	T	W	T	F	S
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30				

May 2002

S	M	T	W	T	F	S
			1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30	31	

8 am — 7:00am-9:00am TeleWork (Home)

8 am — Travel (Home to Work)

9 00 — CPOF Collaboration Exercise (ISX - 4301 N. Fairfax)

10 00

11 00 — Roy Reed and Rich Peze (DARPA 675)

12 pm — eRoom Demo (3811 Hicks & Assoc. conf room)

1 00 — Eileen Preisser Meeting at the request of Dr. Tether (675)

2 00 — John L. Petersen, President THE ARLINGTON INSTITUTE (675 - he needs an LCD)

Travel (CIA)

3 00 — System Activities Review (3811 Suite 850)

3 00 — Demo (CIA)

(b)(3) 50 USC §403 g) Section 6 — [redacted] CIA)

4 00 — Demo (CIA)

5 00

Notes

jpoindexter

April 26, 2002

Friday

April 2002

S	M	T	W	T	F	S
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30				

May 2002

S	M	T	W	T	F	S
			1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30	31	

DIRO Offsite (TBD)

8 am

Offsite - Quantico Marine Corps Station (Quantico MC Base, Virginia)

7:00am-9:00am TeleWork (Home)

9 00

10 00

11 00

12 pm

1 00

2 00

3 00

4 00

5 00

6:00am - 7:30am Travel(Home to Quantico)

Notes

jpoindexter

April 29, 2002
Monday

April 2002

S	M	T	W	T	F	S
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30				

May 2002

S	M	T	W	T	F	S
			1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30	31	

8 am — 7:00am-9:00am TeleWork (Home)

9 00 — Travel (Home to Work)

10 00 — Art Fritzson

11 00 — Shaun Jones - 675

12 pm

1 00 — Travel to AIR

2 00 — WAE review of AIR (AIR)

3 00

4 00

5 00

Notes

jpoindexter

April 30, 2002

Tuesday

April 2002

S	M	T	W	T	F	S
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30				

May 2002

S	M	T	W	T	F	S
			1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30	31	

(b)(6) [redacted]

8 am — 7:00am-9:00am TeleWork (Home)

9 00 — Travel (Home to Work)

10 00 — MITRE (McLean)

11 00

12 pm

1 00

2 00

3 00 — travel

4 00 — MAYA Synthetic Data Design: Peter Lucas, Jeff Senn, Steve Roth (3701-723 Conference Room (Secret))

5 00

Notes

jpoindexter

May 01, 2002
Wednesday

May 2002

S	M	T	W	T	F	S
			1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30	31	

June 2002

S	M	T	W	T	F	S
						1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30						

(b)(6) [redacted]

8 am — 7:00am-9:00am TeleWork (Home)

9 00

10 00 — Nahum Gershon - MITRE, Pager: 888 431 8593 (675)

Eric Rickard - SRI, (703) 236-5017 (675)

11 00 — Brian Sharkey - working lunch from 11:30 - 12:30 then others from Brian's group will join (675)

12 pm

1 00 — travel

2 00 — Visit to US Coast Guard HQ - Joe Stewart (Buzzard's Point)

3 00 — travel

4 00 — Darrell dryrun (3701-629 Conference Room (Secret))

5 00 — Bob Gormley (675)

Notes

jpoindexter

May 02, 2002

Thursday

May 2002

S	M	T	W	T	F	S
			1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30	31	

June 2002

S	M	T	W	T	F	S
						1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30						

(b)(6) [redacted]

8 am — 7:00am-9:00am TeleWork (Home)

9 00

General Wesley Clark (675)

10 00 — Computer Refresh - Discussion (675)

11 00 — Imagining time - 675

12 pm — DARPA projects brief to Gen Kernan, CINC, JFCOM (3701-970 DIRO Conference Room (Secret); 3701-SSCO LARGE Conference Room (SCI/SAP))

1 00

2 00

3 00 — System Activities Review (3811 Suite 850)

4 00 — Updated: Richard Clarke, President's Critical Infrastructure Protection Board (3701-970 DIRO Conference Room (Secret))

5 00 — Updated: Discuss IAO Public Affairs (John's office)

Notes

jpoindexter

May 03, 2002

Friday

May 2002

S	M	T	W	T	F	S
			1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30	31	

June 2002

S	M	T	W	T	F	S
						1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30						

(b)(6) [redacted]

8 am	7:00am-9:00am TeleWork (Home)
9 00	Travel (Home to Work)
10 00	Brian Witten - video (3701-SSCO SMALL Conference Room (SCI/SAP))
11 00	Jim Opfer,President and CEO, LaunchPower (675)
12 pm	IAO Informal Lunch (678)
1 00	(b)(3) 10 USC §424 [redacted] DIA, Director of Collections for DoD (675)
2 00	DARPATech Plans (675)
	Dr. Douglas Maughan - PM ATO (675)
3 00	hold
4 00	
5 00	

Notes

jpoindexter

May 06, 2002

Monday

May 2002

S	M	T	W	T	F	S
			1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30	31	

June 2002

S	M	T	W	T	F	S
						1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30						

8 am — 7:00am-9:00am Telework (Home)

9 00 — Travel

10 00 — 10:00am-10:15am Coordination of schedules for the week (675)

Imagining time - 675

11 00

12 pm — 12:15pm-1:15pm Updated: Senior Management Mtg (3701-970 DIRO Conference Room (Secret))

1 00

lunch

2 00

Charles Herzfeld - Consultant to Dr. Tether (675)

3 00

hold

4 00 — Phil Barger (675)

5 00

Notes

jpoindexter

May 07, 2002

Tuesday

May 2002

S	M	T	W	T	F	S
			1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30	31	

June 2002

S	M	T	W	T	F	S
						1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30						

8 am — 7:00am-9:00am Telework (Home)

9 00 — Travel

10 00 — PC ISR Experience (Aerospace Building across from WF - NRO)

11 00 — travel

12 pm — Lunch with Zachary Lemnios (Gaffney's)

1 00 — Movie (675)

2 00 — Imagining time

3 00 — Sal Stolfo (675)

(b)(6) 4 00 — [redacted] - Mathematician, referred by Shaun Jones (675)

5 00 — TIA (675)

6:00pm - 7:00pm Parney Albright - OSTP - 202-456-6072(675)

Notes

jpoindexter

May 08, 2002

Wednesday

May 2002

S	M	T	W	T	F	S
			1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30	31	

June 2002

S	M	T	W	T	F	S
						1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30						

	HOLD - Working from home
8 am	7:00am-9:00am Telework (Home)
9 00	Travel
10 00	
11 00	
12 pm	
1 00	
2 00	
3 00	
4 00	
5 00	

Notes

jpoindexter

May 09, 2002

Thursday

May 2002

S	M	T	W	T	F	S
			1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30	31	

June 2002

S	M	T	W	T	F	S
						1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30						

	Bluebird Bottom
8 am	7:00am-9:00am Telework (Home)
9 00	Travel
10 00	Joe M. Allbaugh, Director of FEMA (500 C Street, Room 828, SW Washington, DC)
11 00	travel
	Imagining Time - 675
12 pm	Seibel Demo
1 00	Dr. Lyle Cox - 304-813-8461 (1014)
2 00	hold
3 00	System Activities Review (3811 Suite 850)
4 00	
5 00	5:00pm-7:00pm travel (DL 620/DL 1241 Depart 7:15pm IAD Arrive 10:11pm PNS (via ATL))

Notes

jpoindexter

May 10, 2002

Friday

May 2002

S	M	T	W	T	F	S
			1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30	31	

June 2002

S	M	T	W	T	F	S
						1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30						

Ken Ford - U West Florida (Pensacola, FL)

8 am — 7:00am-9:00am Telework (Home)

U of W Florida (Pensacola)

9 00

10 00

11 00

12 pm

1 00

2 00

3 00

3:30pm-9:00pm travel (DL 798/DL 244 Depart PNS 3:40pm Arrive 8:17pm IAD (via ATL))

4 00

5 00

Notes

May 13, 2002

Monday

May 2002

S	M	T	W	T	F	S
			1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30	31	

June 2002

S	M	T	W	T	F	S
						1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30						

8 am — 7:00am-9:00am Telework (Home)

Travel (Savannah - United flight 7751 leaves IAD at 925 arrives Savannah at 1105)

9 00

Travel

10 00

11 00 — 11:00am-7:00pm EELD Meeting - Sailing (Hotel: Hilton Savannah DeSoto, phone (912) 232-9000)

12 pm

1 00

2 00

3 00

4 00

5 00

7:00pm - 8:30pm Dinner with Brian Sharkey (meet in Hotel at 7:00)

Notes

May 14, 2002
Tuesday

May 2002

S	M	T	W	T	F	S
			1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30	31	

June 2002

S	M	T	W	T	F	S
						1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30						

EELD PI Meeting (Savannah)

8 am — 7:00am-9:00am Telework (Home)

9 00

10 00

11 00

12 pm

1 00

2 00

3 00

4 00

5 00

Notes

jpoindexter

May 15, 2002
Wednesday

May 2002

S	M	T	W	T	F	S
			1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30	31	

June 2002

S	M	T	W	T	F	S
						1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30						

	Return (Savannah)	Notes
8 am	7:00am-9:00am Telework (Home)	
9 00	Travel	
10 00	Travel (Savannah - United flight 7246 leaves Savannah at 1035 arrives IAD at 1210)	
11 00		
12 pm	travel to DARPA from IAD	
1 00	Updated: [redacted] ncoming and outgoing Defence Attache	
	INSCOM	
2 00		
3 00	Fred Shatzman (675)	
4 00	Jerry Woodall - will call to reschedule (DARPA 675)	
5 00		

(b)(3):10 USC §424

jpoindexter

May 16, 2002
Thursday

May 2002

S	M	T	W	T	F	S
			1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30	31	

June 2002

S	M	T	W	T	F	S
						1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30						

8 am — 7:00am-9:00am Telework (Home)

9 00 — Travel

10 00 — 10:00am-10:45am Louis Solomon - President, LPS Associates (678)

11 00 — PEs for new IAO programs (Genisys, Genoa II and TIA) (966)

Imagining Time (675)

12 00 — travel

1 00 — JCAG/FTTTF Brief (Crystal City - 1755 Jefferson Davis Drive, Suite 1200)

2 00

3 00 — System Activities Review (3811 Suite 850)

travel

4 00 — (b)(6) (INSCOM - Meet at Front Desk at 4:00)

5 00

6:00pm - 8:00pm Dinner with Shaun and George Poste(Citronelle - 3000 M Street - 202-625-2150)

Notes

jpoindexter

May 17, 2002
Friday

May 2002

S	M	T	W	T	F	S
			1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30	31	

June 2002

S	M	T	W	T	F	S
						1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30						

(b)(6)

Notes

8 am — 7:00am-9:00am Telework (Home)

9 00 — Travel

10 00 — Multi-Sector Crisis Management Consortium - briefing (Ballston Metro Center Office Tower, Suite 800, 901 North Stuart Street)

11 00

12 pm — Reception - Multi-Sector Crisis Management Consortium (Ballston Metro Center Office Tower, Suite 800, 901 North Stuart Street)

1 00 — 1:00pm-1:15pm Dr. Brian Gabbard - ARETE, cell: 818-519-3438 (675)

hold

2 00 — John Hoyt - CONUS CT information sharing and data mining, Bruce Blacar and D'Arcy Morgan will attend with John Hoyt (678)

3 00 — hold (675)

Art Bruckheim, Re: The email that Dr. Tether sent to Dr. Poindexter (675)

4 00 — Imagining Time (675)

5 00

jpoindexter

May 20, 2002
Monday

May 2002

S	M	T	W	T	F	S
			1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30	31	

June 2002

S	M	T	W	T	F	S
						1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30						

(b)(6)

8 am — 7:00am-9:00am Telework (Home)

7:30am-8:30am Travel

ISAT Security with Privacy (IDA)

9 00

10 00

11 00

12 pm

1 00

2 00

3 00

4 00

5 00

Notes

jpoindexter

May 21, 2002

Tuesday

May 2002

S	M	T	W	T	F	S
			1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30	31	

June 2002

S	M	T	W	T	F	S
						1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30						

(b)(6) [redacted]

8 am	7:00am-9:00am Telework (Home)
9 00	Travel
10 00	Victoria Stavridou (675)
11 00	Larry Jackel - IPTO Interview (675)
	Rod McDaniel (Gaffney's)
12 pm	
1 00	hold
	travel
2 00	Financial Crimes Enforcement Network (FinCEN) - Gammon, Lynda (2070 Chain Bridge Rd., Suite 200, Vienna, VA 22182)
3 00	
4 00	travel
5 00	

Notes

jpoindexter

May 22, 2002
Wednesday

May 2002

S	M	T	W	T	F	S
			1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30	31	

June 2002

S	M	T	W	T	F	S
						1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30						

8 am — 7:00am-9:00am Telework (Home)

9 00 — Travel

10 00 — Bioterrorism - David Siegrist - Potomac Institute for Policy Studies (675)

11 00 — Mike Siracuse (675)

12 pm — IAO Informal Lunch (IAO Atrium)

(b)(6) 1 00 — [redacted]675)

2 00 — lunch

BAA Source Selection meetings - Present revised roadmap and first cut of selection recommendations (678)

3 00

4 00 — Seedling Project White Paper (675)

5 00

Notes

jpoindexter

May 23, 2002

Thursday

May 2002

S	M	T	W	T	F	S
			1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30	31	

June 2002

S	M	T	W	T	F	S
						1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30						

8 am — Experiment & Results (INSCOM)

7:00am-9:00am Telework (Home)

9 00

travel

10 00

10:45am-11:45am SASA Briefing - 10:45 for 45 minutes (DIAC)

11 00

11:45am-12:15pm travel

12 pm — Teradata Demo (3811)

1 00 — Reviews

2 00 — Travel

3 00 — System Activities Review (3811 Suite 850)

4 00

5 00 — Joe Eash (672)

6:30am - 7:30am Travel

Notes

jpoindexter

May 24, 2002

Friday

May 2002

S	M	T	W	T	F	S
			1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30	31	

June 2002

S	M	T	W	T	F	S
						1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30						

8 am — 7:00am-9:00am Telework (Home)

9 00 — Updated: Greg Walters, HAC-D

Travel

10 00 — IAO Actual Time (970)

11 00

12 pm

1 00

2 00 — Mark Stefik, PARC (DARPA 675)

3 00 — Mr. Peng-Yam TAN - Head, Defense Technology Office (US), Embassy of Singapore and his Assistant (675 - Embassy will pass visit request to SID)

4 00 — Bill Tang - DARPA/MTO Smart Dust and Speech Chip (3701-643 Conference Room (Secret))

5 00

Notes

jpoindexter

May 27, 2002

Monday

May 2002

S	M	T	W	T	F	S
			1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30	31	

June 2002

S	M	T	W	T	F	S
						1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30						

Memorial Day (United States)
Memorial Day (United States)
Sailing (Annapolis)

8 am — 7:00am-9:00am Telework (Home)

9 00 — Travel

10 00

11 00

12 pm

1 00

2 00

3 00

4 00

5 00

Notes

May 28, 2002

Tuesday

May 2002

S	M	T	W	T	F	S
			1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30	31	

June 2002

S	M	T	W	T	F	S
						1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30						

8 am — 7:00am-9:00am Telework (Home)

9 00 — Updated: Israeli Perspective on Suicide Bombers (SAIC, 4001 N. Fairfax Dr, Suite 600)

Travel

10 00

11 00

12 pm — 12:15pm-1:45pm Lunch Poindexter & Wishner (Dan & Brad's in the Hilton)

1 00

2 00 — Reviews (675)

3 00 — Norm Willox, President and COO of Lexis Nexis, Paul Colangelo and Dennis Walters the Technical guys will attend as well. (675)

4 00 — Mike Steele (DARPA 675)

5 00

Notes

jpoindexter

May 29, 2002
Wednesday

May 2002

S	M	T	W	T	F	S
			1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30	31	

June 2002

S	M	T	W	T	F	S
						1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30						

8 am — 7:00am-9:00am Telework (Home)

9 00 — Travel

10 00 — travel - directly from home to Chantilly

(b)(6) 11 00 — PC ISR Experience with [redacted]

(b)(6) 12 pm — travel - Bring [redacted] back from meeting

1 00 — lunch (675)

(b)(6) Updated: [redacted] Chantilly)

2 00

3 00 — IAO Review (675)

4 00

5 00

Notes

jpoindexter

May 30, 2002
Thursday

May 2002

S	M	T	W	T	F	S
			1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30	31	

June 2002

S	M	T	W	T	F	S
						1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30						

8 am	7:00am-9:00am Telework (Home)
9 00	Travel
10 00	Tony Ottenger and Richard Hackman - Advanced Technology Panel (ATP) (675)
11 00	
12 pm	Reviews (675)
1 00	
2 00	Earl Rubright and Brian Sharkey (675)
3 00	System Activities Review (3811 Suite 850)
4 00	
5 00	
(b)(6)	6:00pm - 9:00pm [redacted]

Notes

jpoindexter

May 31, 2002

Friday

May 2002

S	M	T	W	T	F	S
			1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30	31	

June 2002

S	M	T	W	T	F	S
						1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30						

8 am — 7:00am-9:00am Telework (Home)

7:30am-8:30am Travel

Andy Marshall, Dr. Poindexter will brief from 8:30 - 11:00 (15059 Conference Center Drive, 3rd Floor, Chantilly, VA)

9 00

10 00

11 00

12 pm

1 00 — Seedling Meeting with Bob Popp (675)

BAA Reviews (675)

2 00

3 00 — Jim Kelly (675)

4 00 — William Dawson and Jeff Ingle (675)

5 00

Notes

June 03, 2002

Monday

June 2002

S	M	T	W	T	F	S
						1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30						

July 2002

S	M	T	W	T	F	S
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30	31			

Sailing (Chesapeake Bay)

8 am — 7:00am-9:00am Telework (Home)

9 00 — Travel

10 00

11 00

12 pm — 12:15pm-1:15pm Senior Management Mtg (3701-970 DIRO Conference Room (Secret))

1 00

2 00

3 00

4 00

5 00

Notes

jpoindexter

June 04, 2002

Tuesday

June 2002

S	M	T	W	T	F	S
						1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30						

July 2002

S	M	T	W	T	F	S
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30	31			

Sailing (Chesapeake Bay)

8 am
- 7:00am-9:00am Telework (Home)

9 00
- Travel

10 00

11 00

12 pm

1 00

2 00

3 00

4 00

5 00

Notes

June 05, 2002

Wednesday

June 2002

S	M	T	W	T	F	S
						1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30						

July 2002

S	M	T	W	T	F	S
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30	31			

Sailing (Chesapeake Bay)

8 am	7:00am-9:00am Telework (Home)
9 00	Travel
10 00	
11 00	
12 pm	
1 00	
2 00	
3 00	
4 00	
5 00	

Notes

jpoindexter

June 06, 2002

Thursday

June 2002

S	M	T	W	T	F	S
						1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30						

July 2002

S	M	T	W	T	F	S
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30	31			

Sailing (Chesapeake Bay)

8 am	7:00am-9:00am Telework (Home)
9 00	Travel
10 00	
11 00	
12 pm	hold for demo
1 00	
2 00	
3 00	System Activities Review (3811 Suite 850)
4 00	
5 00	

Notes

June 07, 2002

Friday

June 2002

S	M	T	W	T	F	S
						1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30						

July 2002

S	M	T	W	T	F	S
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30	31			

Sailing (Chesapeake Bay)

8 am — 7:00am-9:00am Telework (Home)

9 00 — Travel

10 00

11 00

12 pm

1 00

2 00

3 00

4 00

5 00

Notes

jpoindexter

June 10, 2002

Monday

June 2002

S	M	T	W	T	F	S
						1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30						

July 2002

S	M	T	W	T	F	S
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30	31			

8 am — 7:00am-9:00am Telework (Home)

7:30am-8:30am Travel

CSTB Workshop - Dr. Poindexter will be speaking from 9:30 - 10:45 (NAS Constitution Avenue)

9 00

10 00

11 00

12 pm

Brian Sharkey and Bob (675)

1 00 — lunch

2 00 — IAO POM Input (675)

hold for Bob Popp only

3 00 — BAA and Pad Reviews (675)

4 00

5 00 — hold

Notes

jpoindexter

June 11, 2002

Tuesday

June 2002

S	M	T	W	T	F	S
						1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30						

July 2002

S	M	T	W	T	F	S
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30	31			

8 am	7:00am-9:00am Telework (Home)
9 00	Travel
10 00	Updated: Discuss IAO POM Input (983)
	FutureMap Workshop - (866) 288-9357 (3601 Wilson Blvd, 6th Floor)
11 00	
12 pm	
1 00	lunch (675)
	1:45pm-2:00pm Jeff Gerald
2 00	Tim Grayson - TTO (675)
3 00	Bob Parker, ISI (675)
4 00	BAA Review
5 00	

Notes

jpoindexter

June 12, 2002

Wednesday

June 2002

S	M	T	W	T	F	S
						1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30						

July 2002

S	M	T	W	T	F	S
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30	31			

The Highlands Forum (CA)

8 am — 7:00am-9:00am Telework (Home)

7:30am-3:00pm travel - United flight 845 departs IAD at 7:30 am arrives SFO at 10:05 am. Transfers to United 5411 departs SFO at 11:15 am arriving MRY at 12:00 pm. Hertz Rental Car. (meet up with Shaun Jones for lunch somewhere in the hotel area around 1:30)

9 00

10 00

11 00

12 pm

1 00

2 00

3 00

4 00

5 00 — 5:00pm-7:00pm Reception (Carmel)

5:30am - 6:30am Travel

Notes

jpoindexter

June 13, 2002

Thursday

June 2002

S	M	T	W	T	F	S
						1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30						

July 2002

S	M	T	W	T	F	S
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30	31			

The Highlands Forum (CA)

8 am — 7:00am-9:00am Telework (Home)

9 00

10 00

11 00

12 pm

1 00

2 00

3 00 — System Activities Review (3811 Suite 850)

4 00

5 00

Notes

jpoindexter

June 14, 2002

Friday

June 2002

S	M	T	W	T	F	S
						1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30						

July 2002

S	M	T	W	T	F	S
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30	31			

The Highlands Forum (CA)

Flag Day (United States)

Flag Day (United States)

8 am 7:00am-9:00am Telework (Home)

9 00

10 00

11 00

12 pm

1 00

2 00 travel - United flight 5417 now departs MRY at 2:25 pm arrives SFO at 3:00 pm. Transfers to United 220 departs SFO at 4:10 pm arriving IAD at 12:13 am.

3 00

4 00

5 00

Notes

June 17, 2002

Monday

June 2002

S	M	T	W	T	F	S
						1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30						

July 2002

S	M	T	W	T	F	S
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30	31			

8 am — 7:00am-9:00am Telework (Home)

9 00 — Travel

10 00 — 10:00am-10:15am Establish a JWICS account and new user orientation for Dr. Poin

reviews

11 00

12 pm — 12:15pm-1:15pm Updated: Senior Management Mtg (3701-970 DIRO Conference Room (Secret))

1 00

lunch

2 00 — Jean Scholtz and Sharon Laskowski - NIST (675)

3 00 — hold

Fred Schatzman

4 00 — Reviews

5 00

Notes

jpoindexter

June 18, 2002

Tuesday

June 2002

S	M	T	W	T	F	S
						1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30						

July 2002

S	M	T	W	T	F	S
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30	31			

8 am	7:00am-9:00am Telework (Home)	Notes
9 00	Travel	
10 00	Updated: PC Refresh Schedule for Dr. Poindexter (laptop) (Dr. Poindexter's Office)	Reviews
11 00		Susan Bales, Director of the Navy Fleet Fo
		lunch
12 pm		
	travel	
1 00	(b)(3)10 USC00 §424 [redacted] check in at Guard Gate inside building))	
2 00	travel	
	hold (675)	
3 00	Kathy Macdonald (675)	
	Gene Tighe of the Rockwell's DC Office (675)	
4 00	Jonathon Phillips (675)	
5 00		

jpoindexter

June 19, 2002

Wednesday

June 2002

S	M	T	W	T	F	S
						1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30						

July 2002

S	M	T	W	T	F	S
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30	31			

8 am	7:00am-9:00am Telework (Home)
9 00	Travel
10 00	travel
(b)(6)	[redacted] Special Assistant to the President (1800 G Street, 10th floor)
11 00	
	travel
12 pm	Visit Gerry Yonas VP Advanced Concepts Sandia National Lab (675)
	lunch (675)
1 00	
(b)(3) 50 USC §403(g) Section 6	Updated: Brief [redacted] CIA S&T Directorate (3701-970 DIRO Conference Room (Secret))
2 00	
3 00	
4 00	Tom Garwin - Hicks and Associates (675)
5 00	

Notes

jpoindexter

June 20, 2002

Thursday

June 2002

S	M	T	W	T	F	S
						1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30						

July 2002

S	M	T	W	T	F	S
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30	31			

(b)(6)

Leanne on vacation [redacted]

8 am 7:00am-9:00am Telework (Home)

9 00 Travel

10 00 DISA - Lt. General Harry Raduege, Jr. (DISA)

11 00

12 pm Visual Analytics lunchtime demo (3811) | travel

1 00 Brian Sharkey - Report on Germany Trip (3701-689 Conference Room (SCI/SAP Only))

2 00 Capt Lilienthal - Defense Modeling and Simulation Office (DMSO) (675)

3 00 System Activities Review (3811 Suite 850)

4 00

5 00 Mike MCConnell (675)

Notes

jpoindexter

June 21, 2002

Friday

June 2002

S	M	T	W	T	F	S
						1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30						

July 2002

S	M	T	W	T	F	S
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30	31			

Leanne on vacation (b)(6)

8 am	7:00am-9:00am Telework (Home)
9 00	Travel
10 00	hold
	Dr. Julio Torres, MG USAF (Ret) (675)
11 00	Mtg w/Jonathan Rolf - MLS Status (675)
	Reviews: Call Lyle Cox. (675)
12 pm	
1 00	Gene Gallaher (3701-643 Conference Room (Secret))
2 00	
3 00	Anna Tsao and Brian Sharkey (675)
4 00	
5 00	

Notes

jpoindexter

June 24, 2002

Monday

June 2002

S	M	T	W	T	F	S
						1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30						

July 2002

S	M	T	W	T	F	S
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30	31			

8 am	7:00am-9:00am Telework (Home)
9 00	Travel
10 00	Ted Drzewiecki and Dr. DeVere Henderson (675)
11 00	Call Lyle Cox (telephone)
	travel
12 pm	OSD - lunch (3E880)
	travel
1 00	Tim Hoechst, Senior VP Technology - Oracle Service Industries (3811
2 00	
3 00	
4 00	
5 00	

Notes

June 25, 2002

Tuesday

June 2002

S	M	T	W	T	F	S
						1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30						

July 2002

S	M	T	W	T	F	S
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30	31			

8 am	7:00am-9:00am Telework (Home)
9 00	Travel
10 00	hold
11 00	Sue Gordon and David Grossman - InQTel, Mr. Gilman Louie may attend as well (675)
12 pm	Reviews (675)
1 00	
2 00	Dr. Prosnitz - DOJ (TBD)
3 00	Imagining Time (675)
4 00	BAA Review Meeting (675) / 4:15pm-5:15pm Richard Clarke follow-up (675)
5 00	

Notes

jpoindexter

June 26, 2002

Wednesday

June 2002

S	M	T	W	T	F	S
						1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30						

July 2002

S	M	T	W	T	F	S
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30	31			

MITRE Tech Symposium

8 am — 7:00am-9:00am Telework (Home)

9 00 — Travel

10 00 — Strategic Partnership - CIA and NSA (675)

11 00

12 pm — Reviews (675)

1 00 — RAND DataBase Inventory Project (675)

2 00 — hold for ISAT - Victoria Stavridou (TBD)

3 00

4 00

5 00

Notes

jpoindexter

June 27, 2002

Thursday

June 2002

S	M	T	W	T	F	S
						1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30						

July 2002

S	M	T	W	T	F	S
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30	31			

MITRE Tech Symposium

8 am	7:00am-9:00am Telework (Home)
9 00	Travel
10 00	General Wesley Clark (675)
11 00	Tom Armour (675)
12 pm	ISX Groovy TAG demo (Hicks)
1 00	lunch (675)
	Phil Barger (675)
2 00	Imagining Time (675)
3 00	System Activities Review (3811 Suite 850)
4 00	
5 00	

Notes

jpoindexter

June 28, 2002

Friday

June 2002

S	M	T	W	T	F	S
						1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30						

July 2002

S	M	T	W	T	F	S
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30	31			

8 am	7:00am-9:00am Telework (Home) / Travel
9 00	Brian Sharkey and Scott Fisher - CENTCOM (675)
10 00	66th MI (3701-689 Conference Room (IAO/IXO Secure))
11 00	
	lunch (675)
12 pm	
1 00	
2 00	Updated: Final BAA Review (675)
3 00	
4 00	Reviews
5 00	

Notes

jpoindexter

July 01, 2002
Monday

July 2002

S	M	T	W	T	F	S
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30	31			

August 2002

S	M	T	W	T	F	S
				1	2	3
4	5	6	7	8	9	10
11	12	13	14	15	16	17
18	19	20	21	22	23	24
25	26	27	28	29	30	31

Net Assessment workshop (3811 N. Fairfax)

8 am
- 7:00am-9:00am Telework (Home)
- 7:30am-8:30am Travel
- Building Strategic Knowledge Workshop (3811)

9 00

10 00

11 00

12 pm

1 00
- travel

2 00
- INSCOM

3 00

4 00
- travel

5 00

Notes

jpoindexter

July 02, 2002

Tuesday

July 2002

S	M	T	W	T	F	S
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30	31			

August 2002

S	M	T	W	T	F	S
				1	2	3
4	5	6	7	8	9	10
11	12	13	14	15	16	17
18	19	20	21	22	23	24
25	26	27	28	29	30	31

Net Assessment workshop (3811 N. Fairfax)

8 am
- 7:00am-9:00am Telework (Home)
- 7:30am-8:30am Travel
- Building Strategic Knowledge Workshop (3811)

9 00

10 00

11 00

12 pm

1 00

2 00

3 00

4 00

5 00

Notes

jpoindexter

July 03, 2002
Wednesday

July 2002

S	M	T	W	T	F	S
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30	31			

August 2002

S	M	T	W	T	F	S
				1	2	3
4	5	6	7	8	9	10
11	12	13	14	15	16	17
18	19	20	21	22	23	24
25	26	27	28	29	30	31

8 am	7:00am-9:00am Telework (Home)
9 00	Travel
10 00	Tom Strat (675)
	travel
11 00	John Mittleman - 703-696-2923 (Rm 157 ONR Tower 3- 4015 Wilson Blvd.)
12 pm	BAA Review and lunch (675)
1 00	
	travel
2 00	Stephen Cambone, Director, PA&E (4E808)
3 00	travel
	Brian (675)
4 00	
	Anna Tsao (675)
5 00	

Notes

jpoindexter

July 04, 2002

Thursday

July 2002

S	M	T	W	T	F	S
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30	31			

August 2002

S	M	T	W	T	F	S
				1	2	3
4	5	6	7	8	9	10
11	12	13	14	15	16	17
18	19	20	21	22	23	24
25	26	27	28	29	30	31

Independence Day (United States)

Independence Day (United States)

8 am — 7:00am-9:00am Telework (Home)

9 00 — Travel

10 00

11 00

12 pm — hold for demo

1 00

2 00

3 00 — System Activities Review (3811 Suite 850)

4 00

5 00

Notes

jpoindexter

July 05, 2002

Friday

July 2002

S	M	T	W	T	F	S
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30	31			

August 2002

S	M	T	W	T	F	S
				1	2	3
4	5	6	7	8	9	10
11	12	13	14	15	16	17
18	19	20	21	22	23	24
25	26	27	28	29	30	31

8 am — 7:00am-9:00am Telework (Home)

9 00 — Travel

10 00 — working from home

11 00

12 pm

1 00

2 00

3 00

4 00

5 00

Notes

jpoindexter

July 08, 2002
Monday

July 2002

S	M	T	W	T	F	S
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30	31			

August 2002

S	M	T	W	T	F	S
				1	2	3
4	5	6	7	8	9	10
11	12	13	14	15	16	17
18	19	20	21	22	23	24
25	26	27	28	29	30	31

8 am	7:00am-9:00am Telework (Home)
9 00	Travel
10 00	Dr. Gerald Gilfoyle, APS Task Force on Countering Terrorism (675)
11 00	hold for BAA Reviews (675)
12 pm	Lunch Roy Reed (675)
1 00	
2 00	Mike McConnell (675)
3 00	Oracle Briefing (675)
4 00	IAO BAA Brief for TIA, GENISYS & GENOA II (Part one) (3701-970 DIRO Conference Room (Secret))
5 00	

Notes

jpoindexter

July 09, 2002

Tuesday

July 2002

S	M	T	W	T	F	S
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30	31			

August 2002

S	M	T	W	T	F	S
				1	2	3
4	5	6	7	8	9	10
11	12	13	14	15	16	17
18	19	20	21	22	23	24
25	26	27	28	29	30	31

8 am	7:00am-9:00am Telework (Home)
9 00	Travel
10 00	Ted Gold and Don Latham, IDA to brief on DSB Study for Terrorism (675)
11 00	Travel - DARPA Car (HLS - Nebraska Ave.)
	Updated: Tour/Mtg w/ Sr. Dir of HLS - Kenneth Piernick (Nebraska Ave facility)
12 pm	Travel - DARPA Car
1 00	Lunch (675)
2 00	Updated: Dry run for FutureMAP new start briefing - 643 (3701-643 Conference Room (Secret))
3 00	Tom Mifflin - Metron (703) 787-8700 (675)
4 00	Video Content (675)
5 00	
	travel
	6:00pm - 8:00pm Farewell Function for Xan Alexander(Fort Myer Officer's Club)

Notes

jpoindexter

July 10, 2002
Wednesday

July 2002

S	M	T	W	T	F	S
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30	31			

August 2002

S	M	T	W	T	F	S
				1	2	3
4	5	6	7	8	9	10
11	12	13	14	15	16	17
18	19	20	21	22	23	24
25	26	27	28	29	30	31

8 am	7:00am-9:00am Telework (Home)
9 00	Travel
10 00	Updated: JWICS Brief (6th Floor / Room 675)
	BAA (675)
11 00	
12 pm	Poindexter IAO DT02 Rehearsal (Room 550 then to 551 for graphics approval)
1 00	lunch
2 00	
3 00	Updated: SONAT Demo (3701-970 DIRO Conference Room (Secret))
4 00	hold
5 00	

Notes

jpoindexter

July 11, 2002
Thursday

July 2002

S	M	T	W	T	F	S
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30	31			

August 2002

S	M	T	W	T	F	S
				1	2	3
4	5	6	7	8	9	10
11	12	13	14	15	16	17
18	19	20	21	22	23	24
25	26	27	28	29	30	31

	Bluebird Bottom	Notes
8 am	7:00am-9:00am Telework (Home)	
9 00	Travel	
10 00	Imagining Time (675)	
11 00	IAO Staff Meeting	
12 00	Updated: [redacted] NSA, and Joe Mitola (3701-SSCO LARGE Conference Room (SCI/SAP Only))	
1 00	Alok Nigam - JIVA KD Toolkit (675)	
2 00	hold	
	Jim Thomas, DSD Office (675)	
3 00	System Activities Review (3811 Suite 850)	
4 00		
5 00		

(b)(3):50 USC §402 Note

jpoindexter

Murray Burke

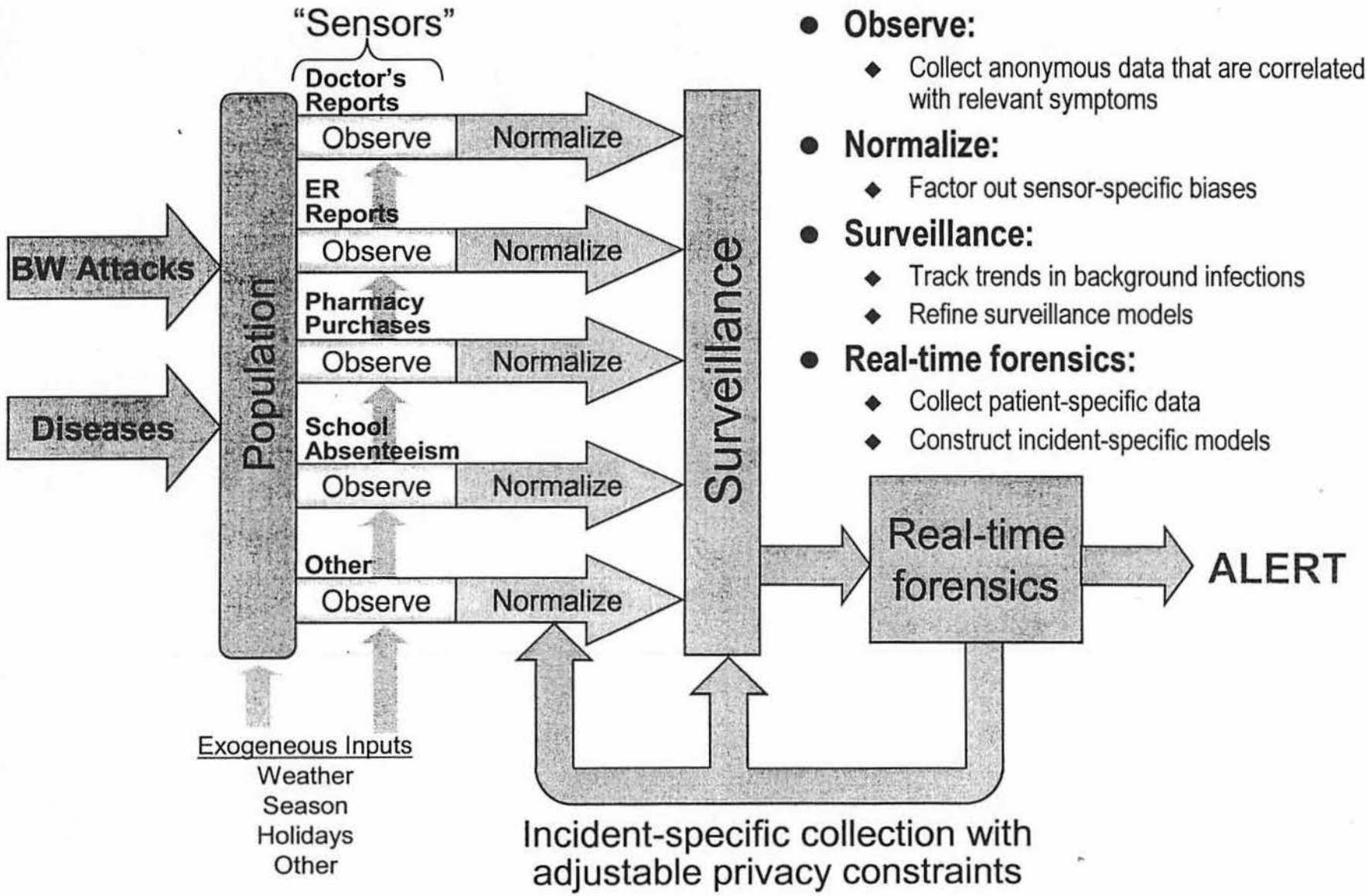
Information Technology Office

703-696-2303 • mburke@darpa.mil

Bio-Surveillance System Elements

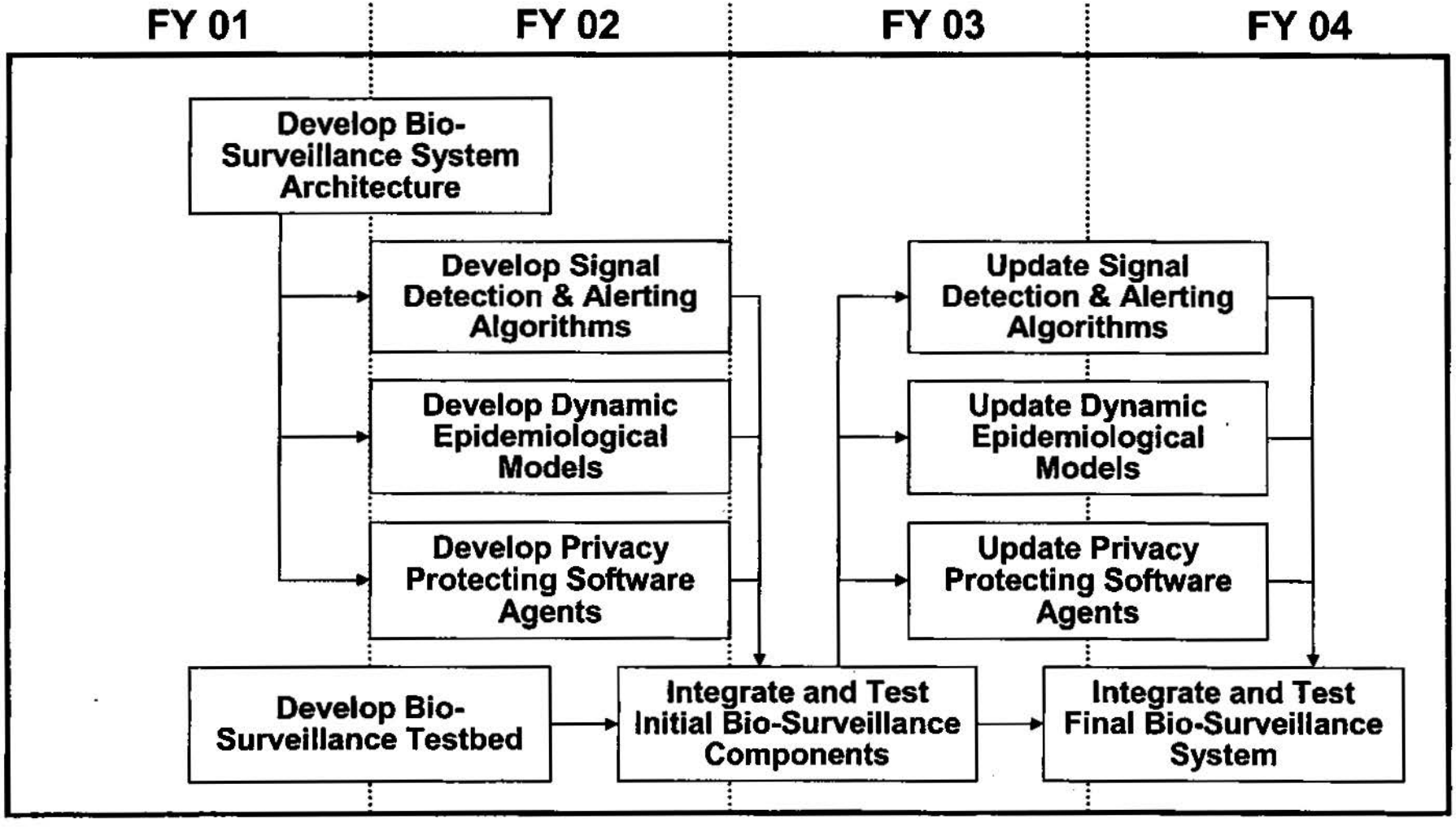
- **Autonomous Detection Algorithms**
  - existing signal detection & bayesian belief net technology
- **Disease Models for Autonomous Detection**
  - higher precision models of dynamic epidemiology
  - network models of sub-population spatial dynamics
- **Privacy Protecting Software Agents**
  - embed data monitoring agents on each data source
    - CoABS Grid and DAML protocols
  - incorporate adjustable privacy protection control in each agent
    - k-similar techniques to assure appropriate anonymity
- **Integrated Bio-Surveillance System Architecture**
  - integration of heterogeneous data sources
  - integration of bio-sensors and health monitoring data
  - dynamic management of the monitoring process

System Concept: Surveillance Triggers Specialized Processing



- **Observe:**
  - Collect anonymous data that are correlated with relevant symptoms
- **Normalize:**
  - Factor out sensor-specific biases
- **Surveillance:**
  - Track trends in background infections
  - Refine surveillance models
- **Real-time forensics:**
  - Collect patient-specific data
  - Construct incident-specific models

Bio-Surveillance Program Plan



DARPA's Initiative on Asymmetric Threat:

DARPA

Total Information Awareness

[redacted]

NY Times, Oct 7, 2001

Tom Armour
Program Manager, Information Awareness Office
tarmour@darpa.mil 703-248-1531

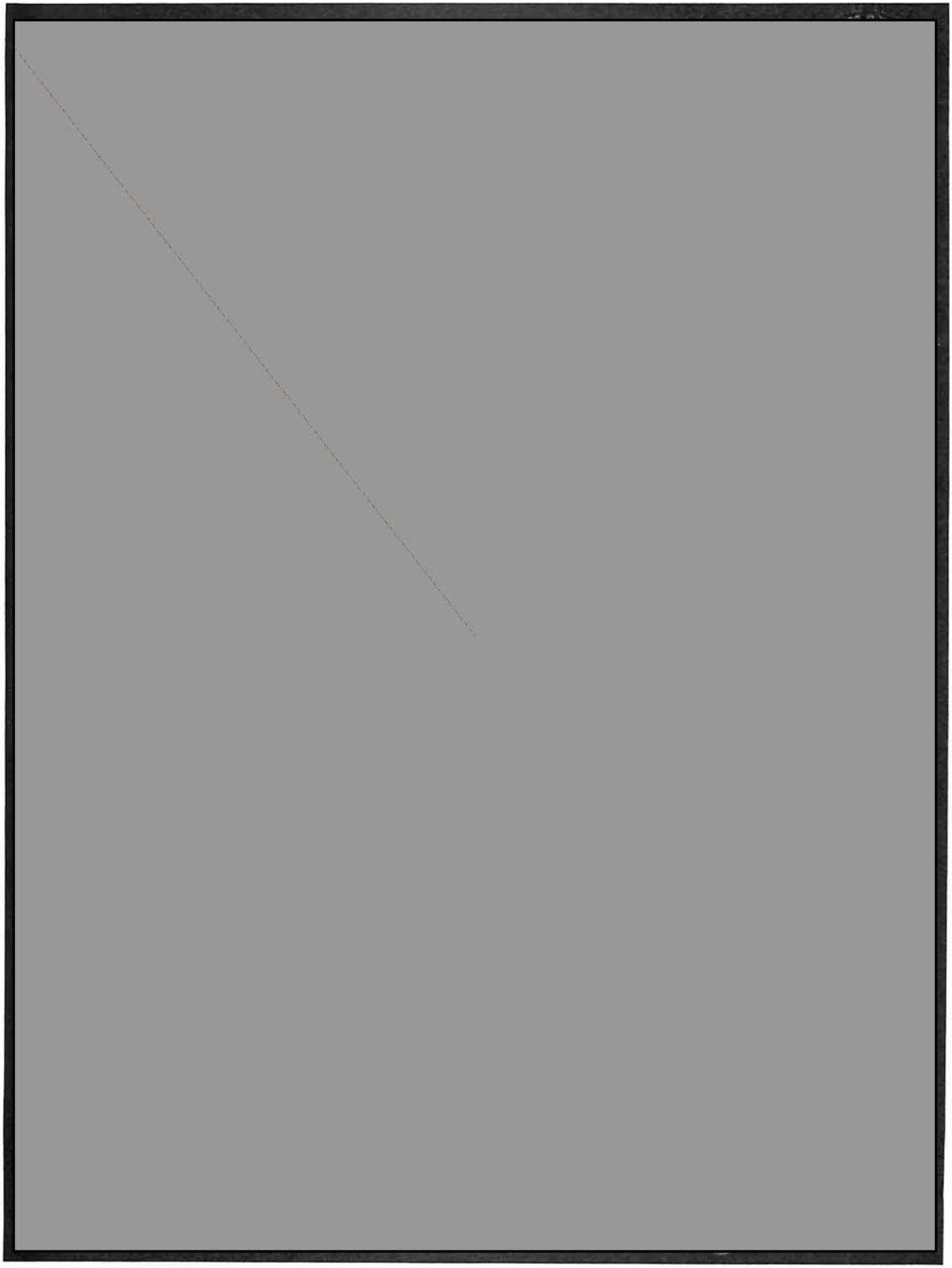
(b)(4)

DARPA

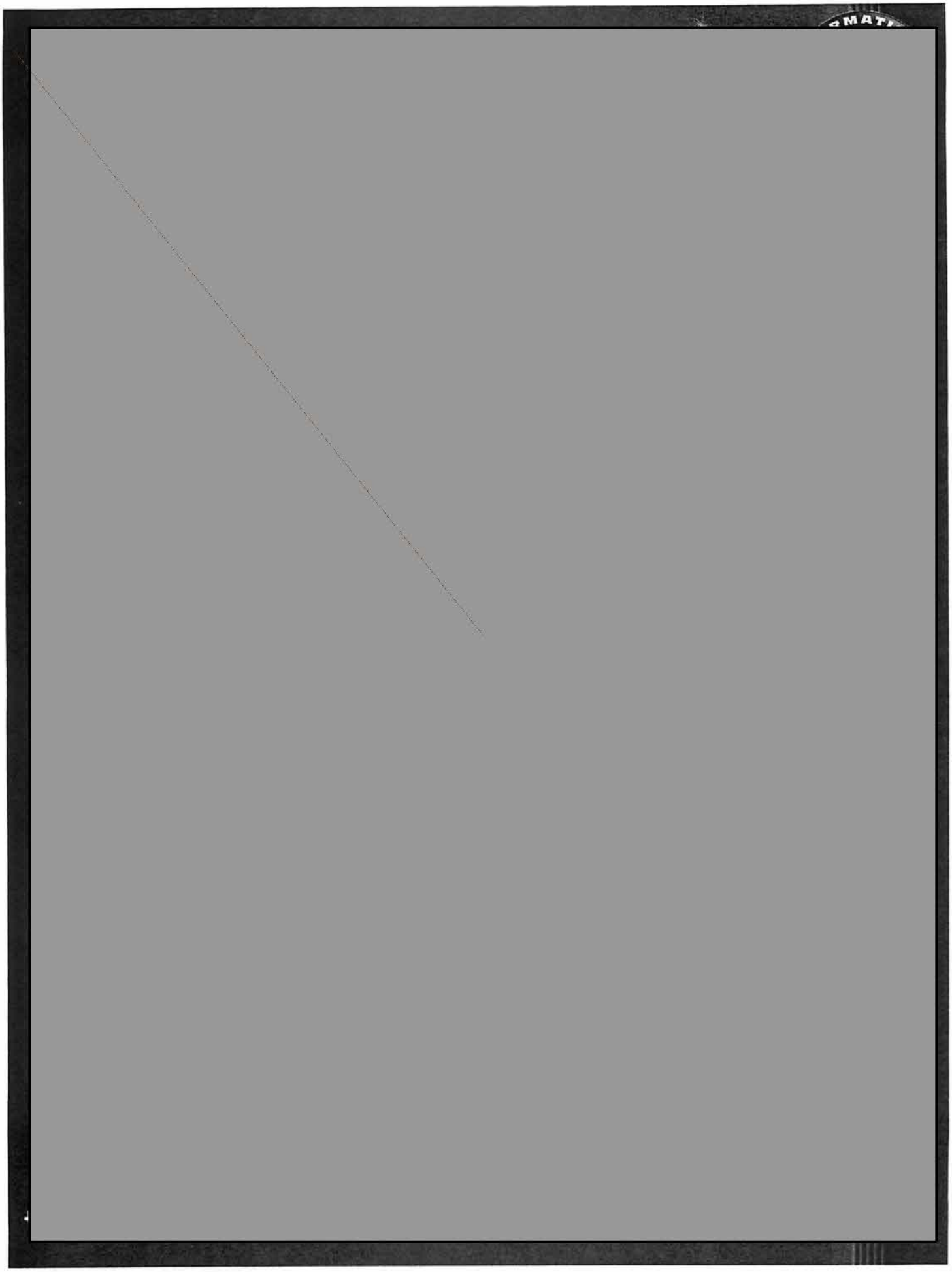
Mission Statement

The DARPA Information Awareness Office (IAO) will imagine, develop, apply, integrate, demonstrate and transition information technologies, components and proto-type, closed-loop, information systems that will counter asymmetric threats by achieving total information awareness useful for preemption; national security warning; and national security decision making.

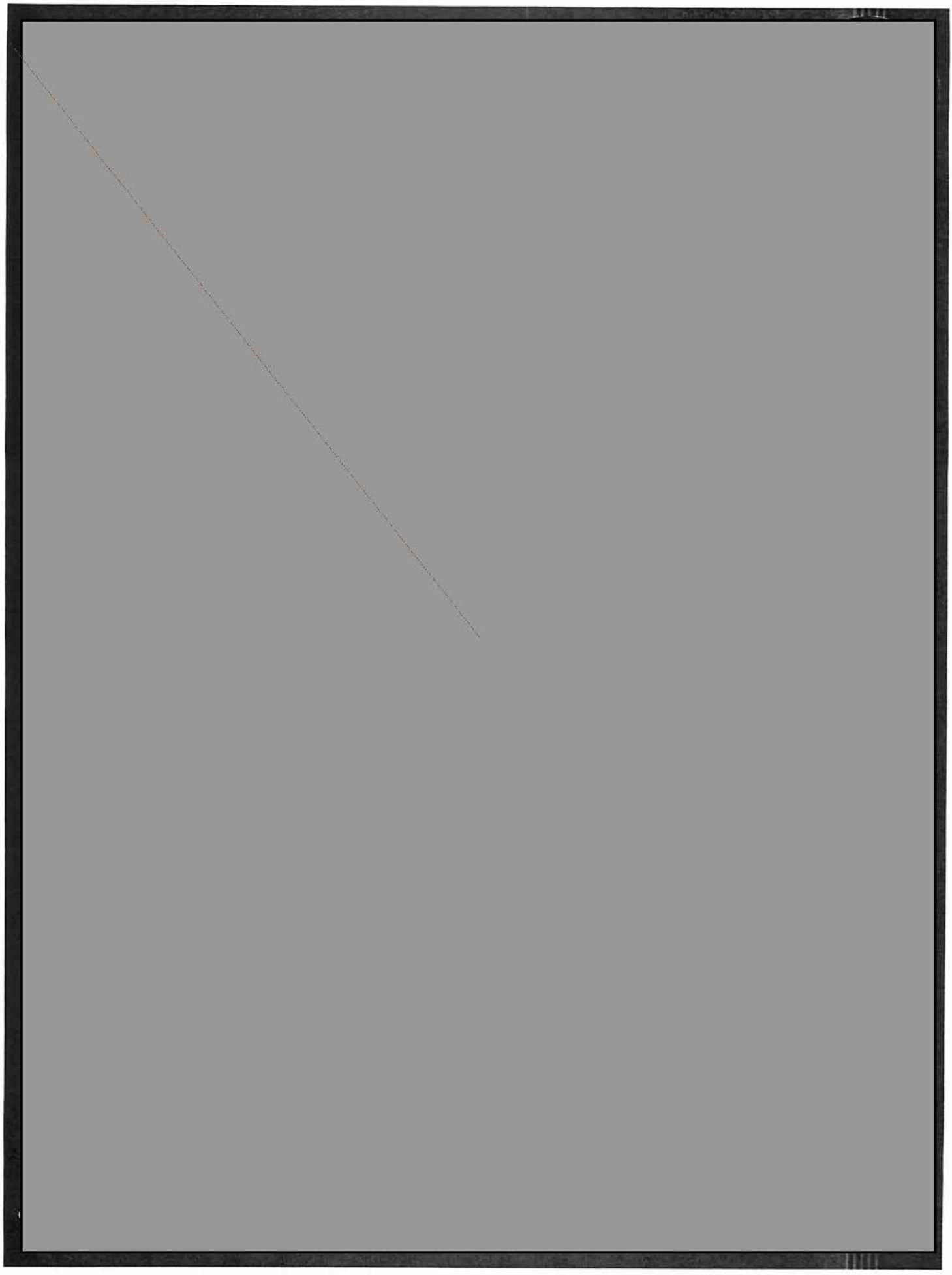
(b)(5)



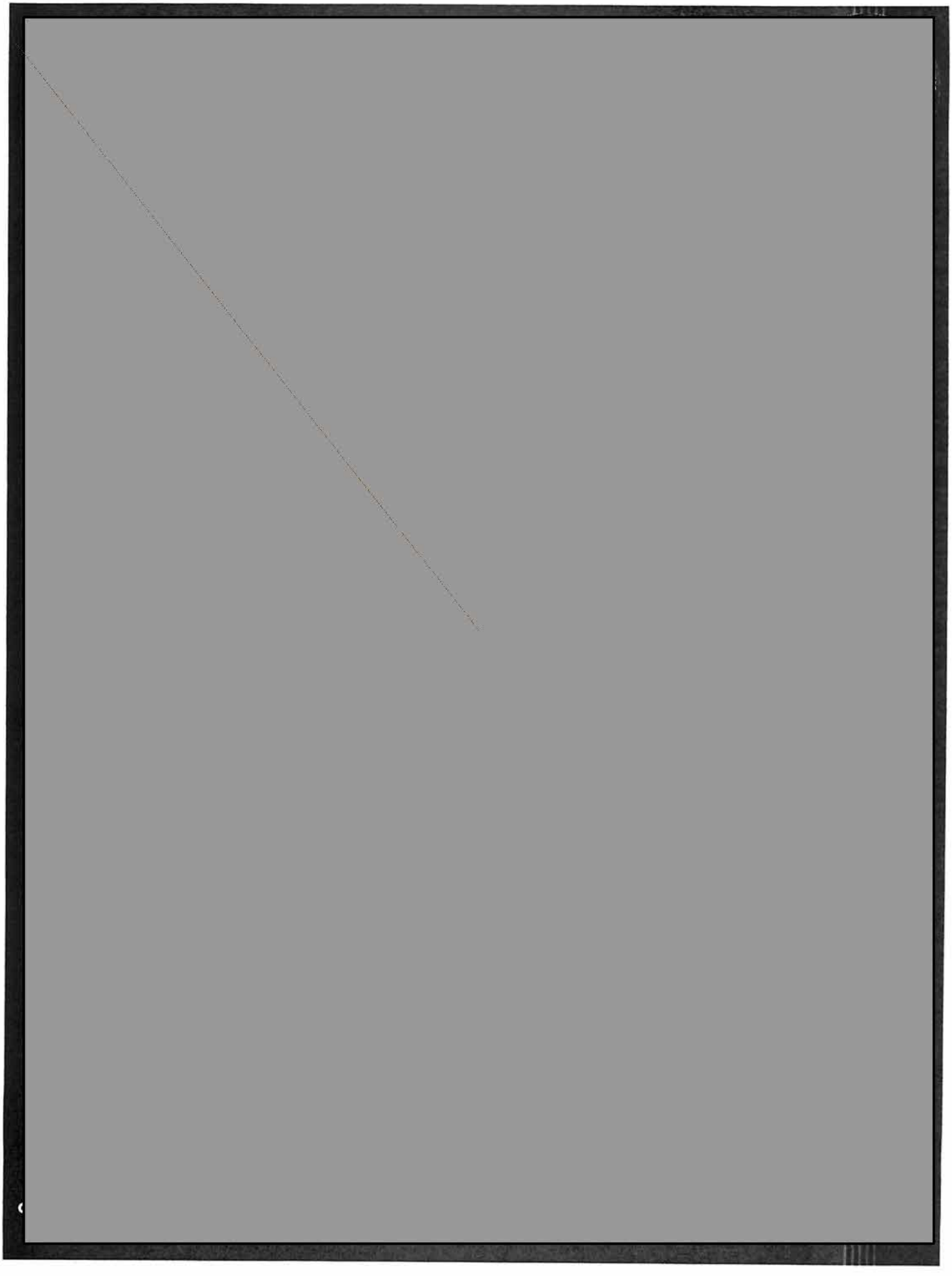
(b)(5)



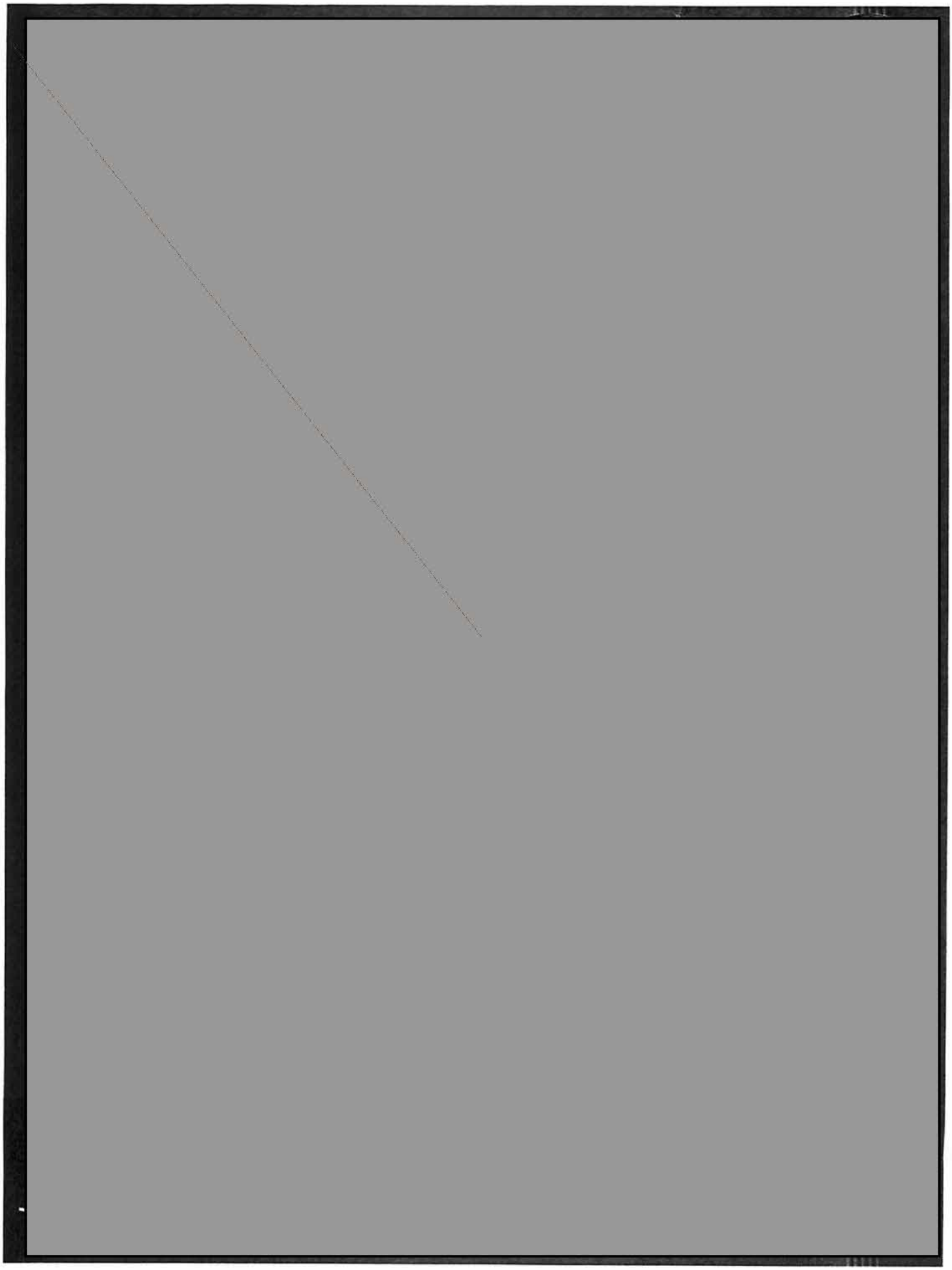
(b)(5)



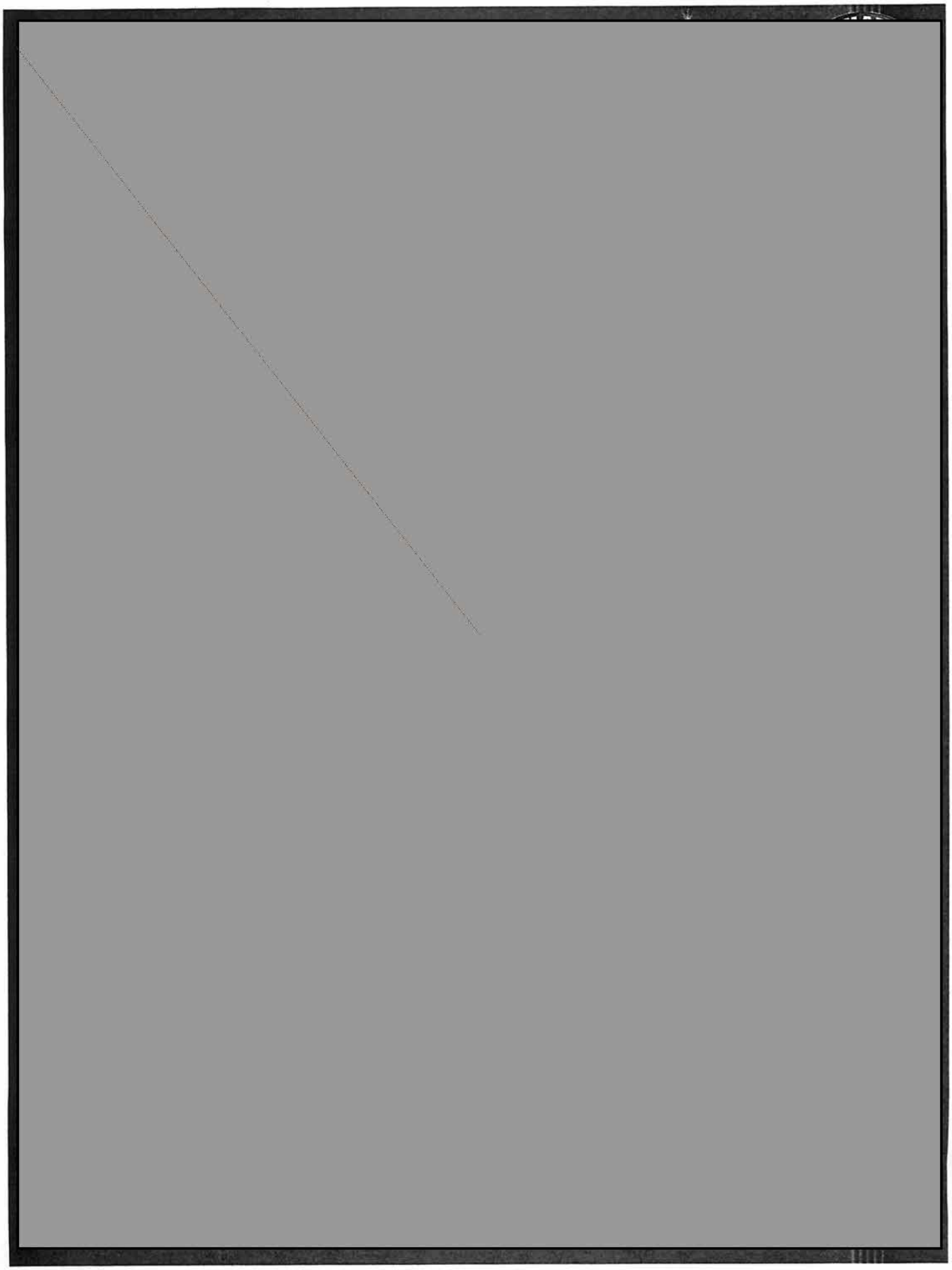
(b)(5)



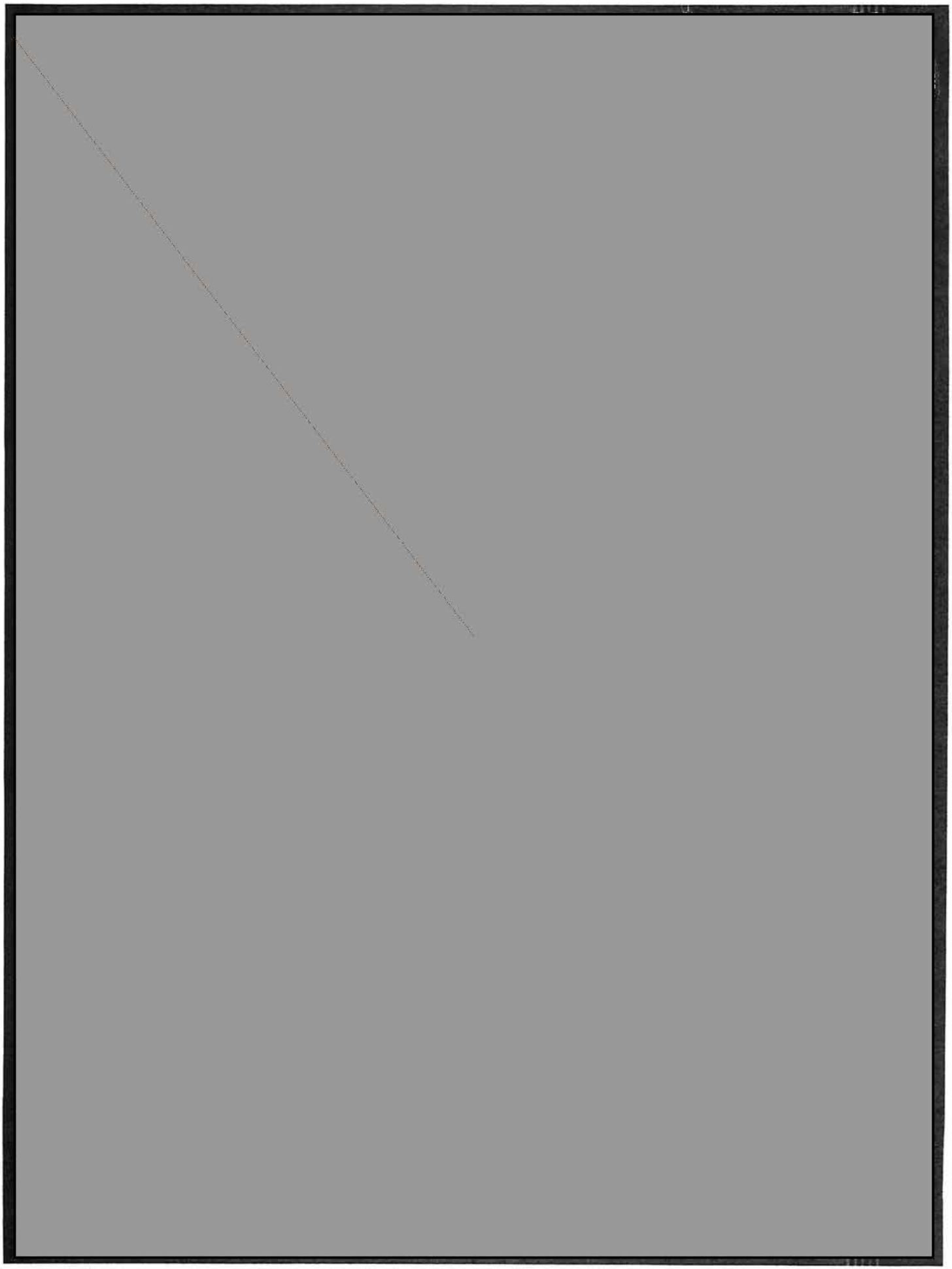
(b)(5)



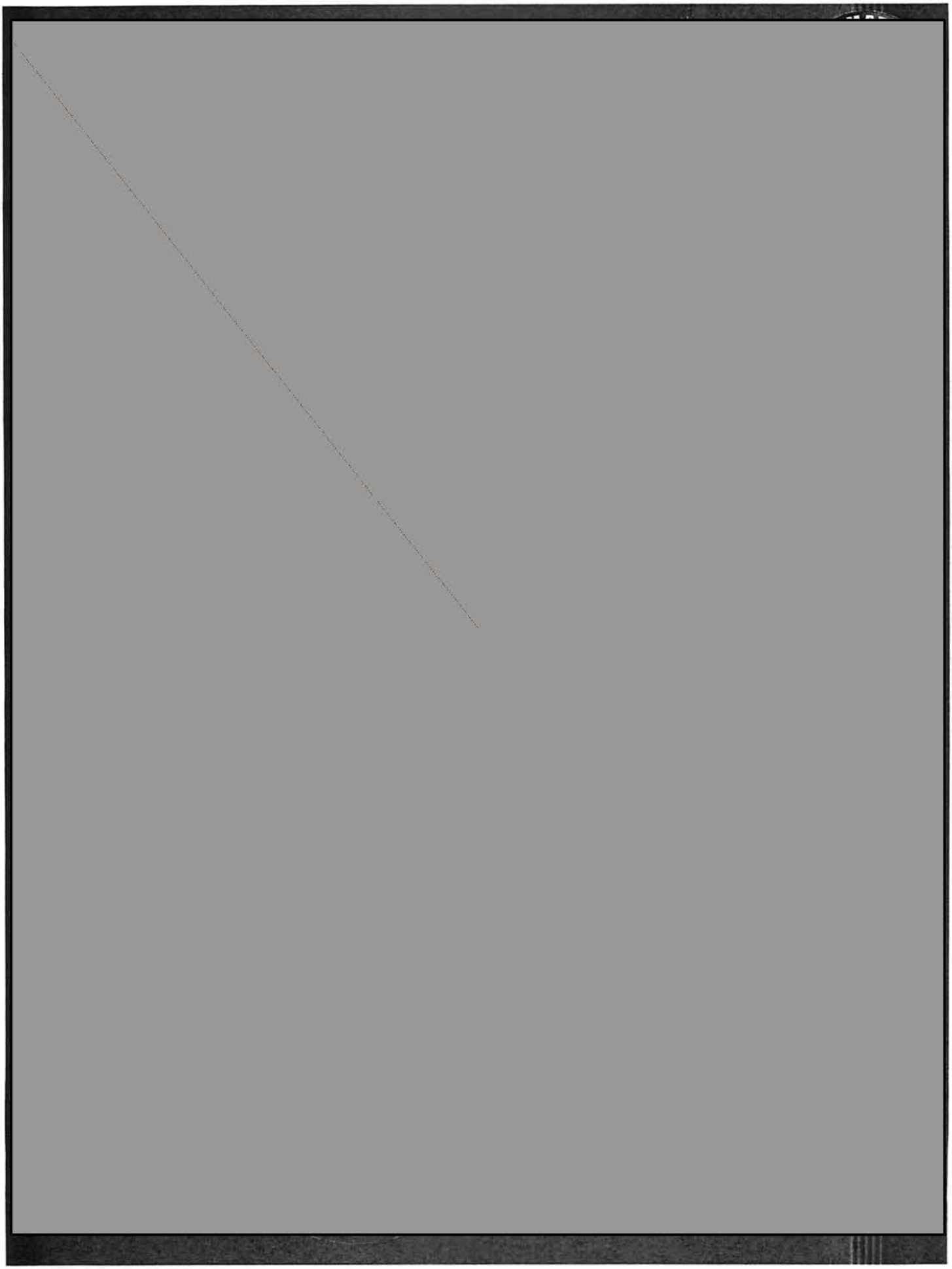
(b)(5)



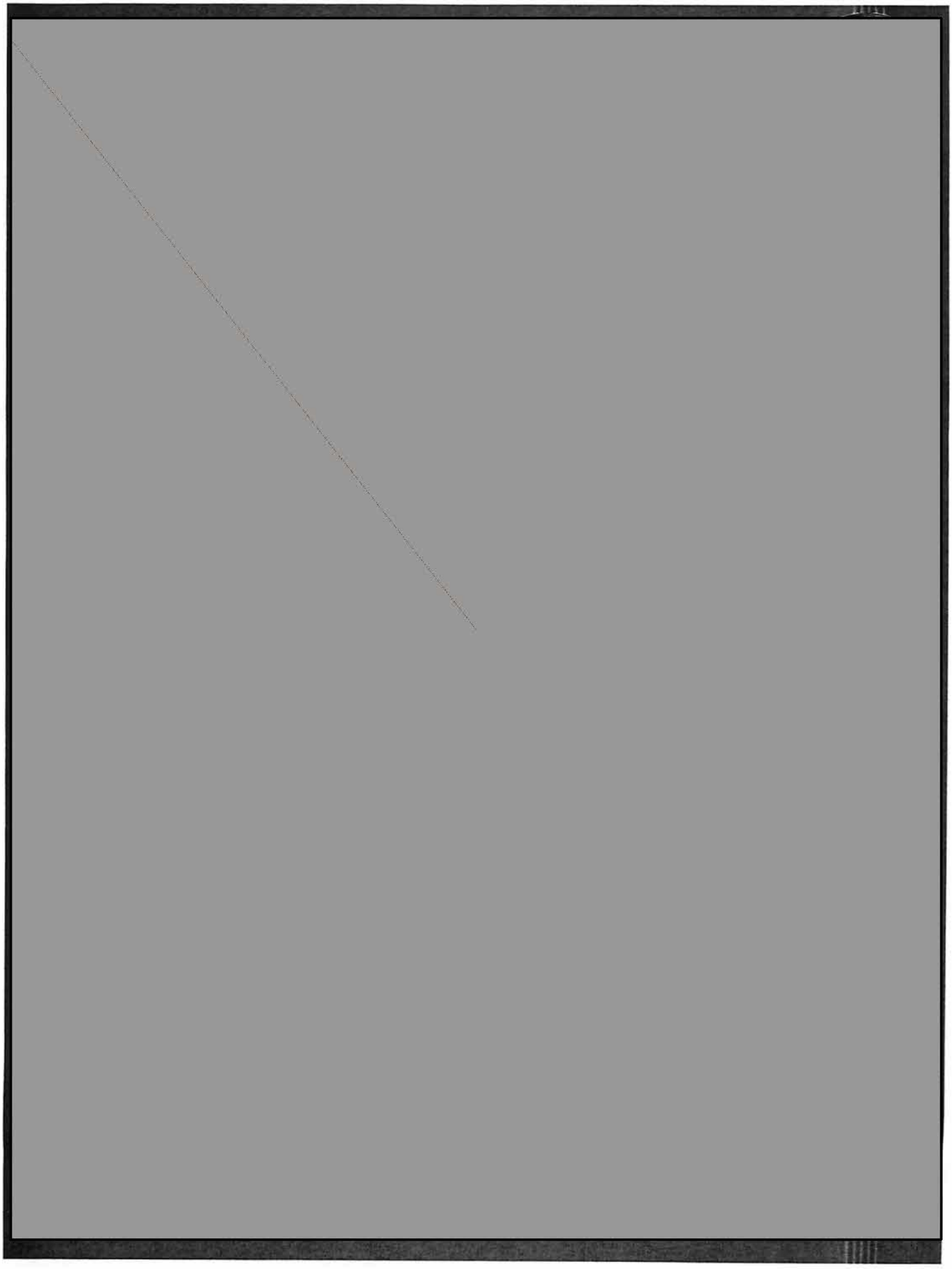
(b)(5)



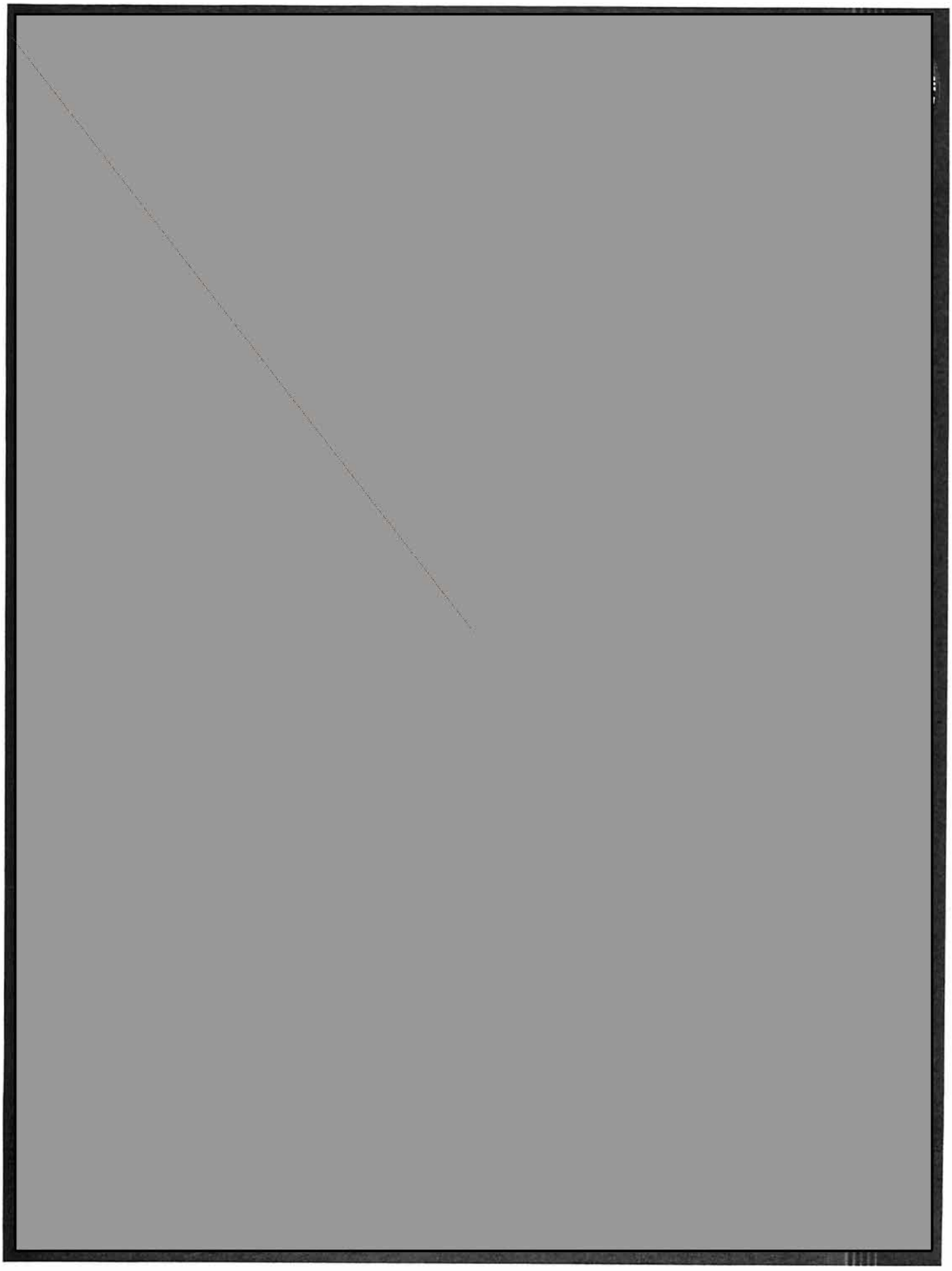
(b)(5)



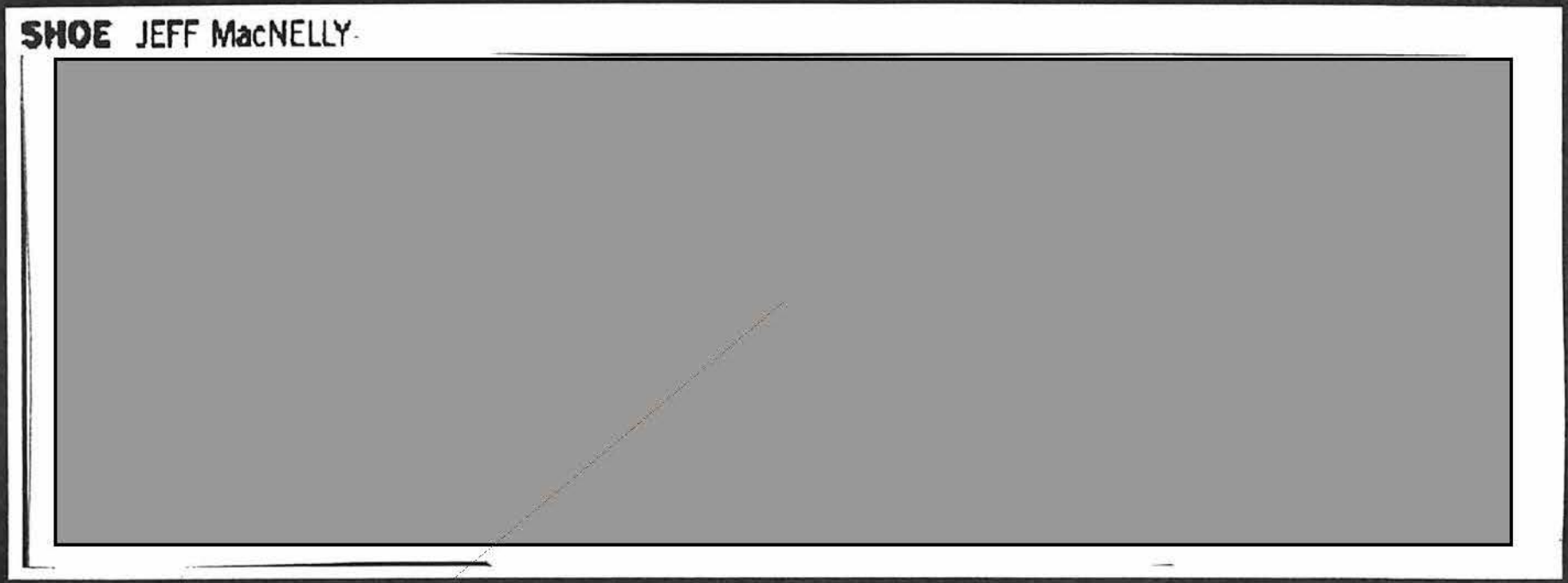
(b)(5)



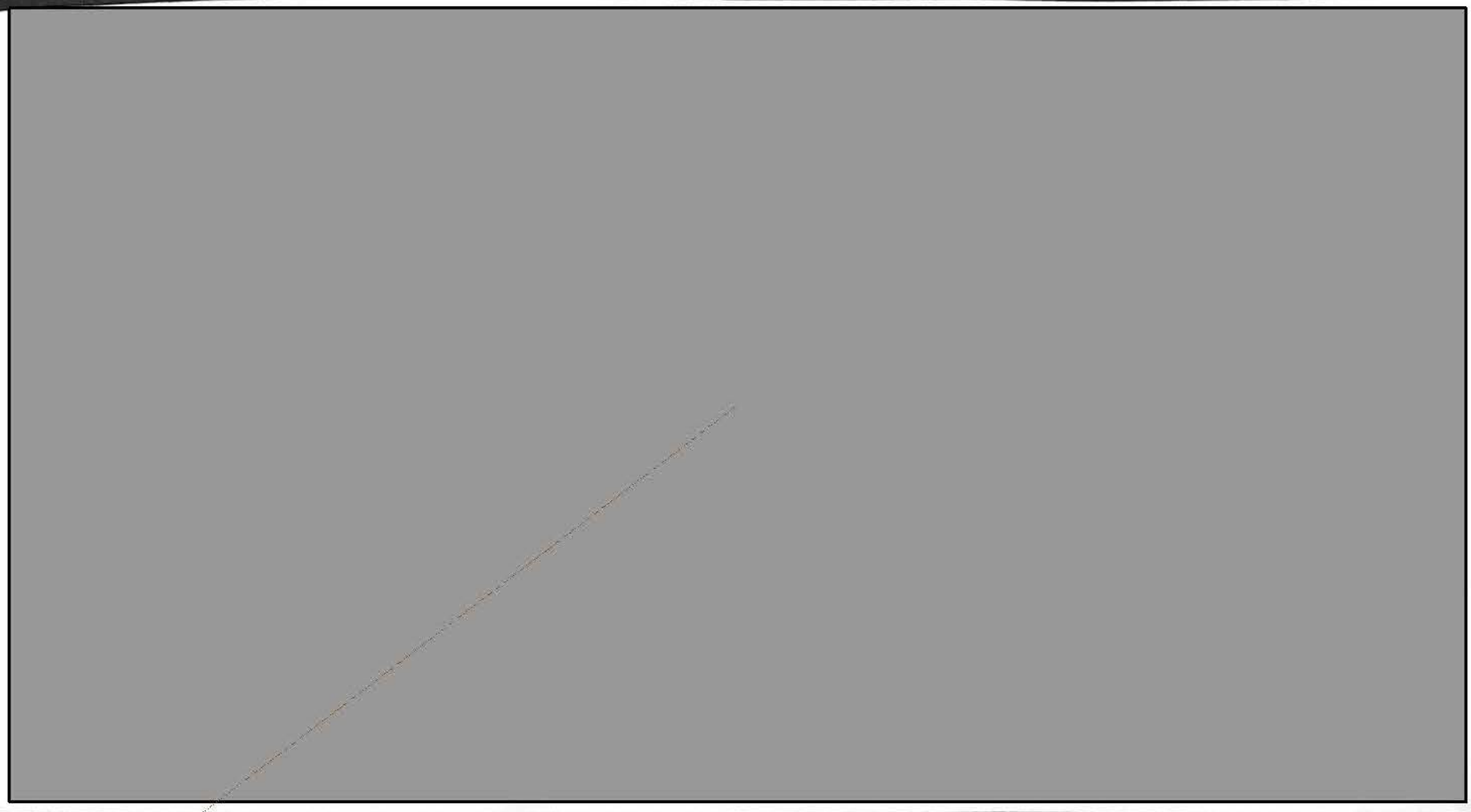
(b)(5)



On Human Reasoning

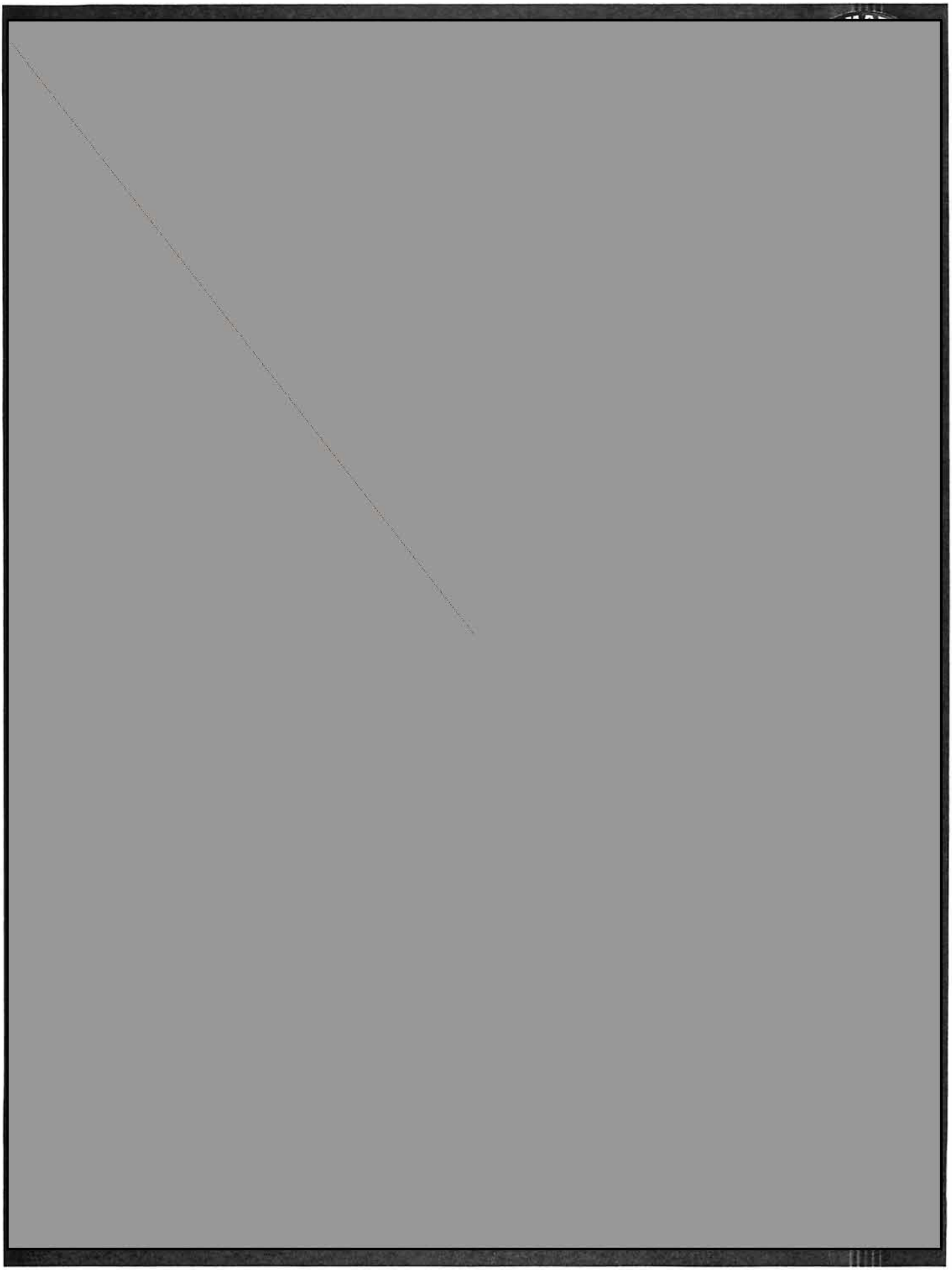


"Topsight" – Grasping the Whole

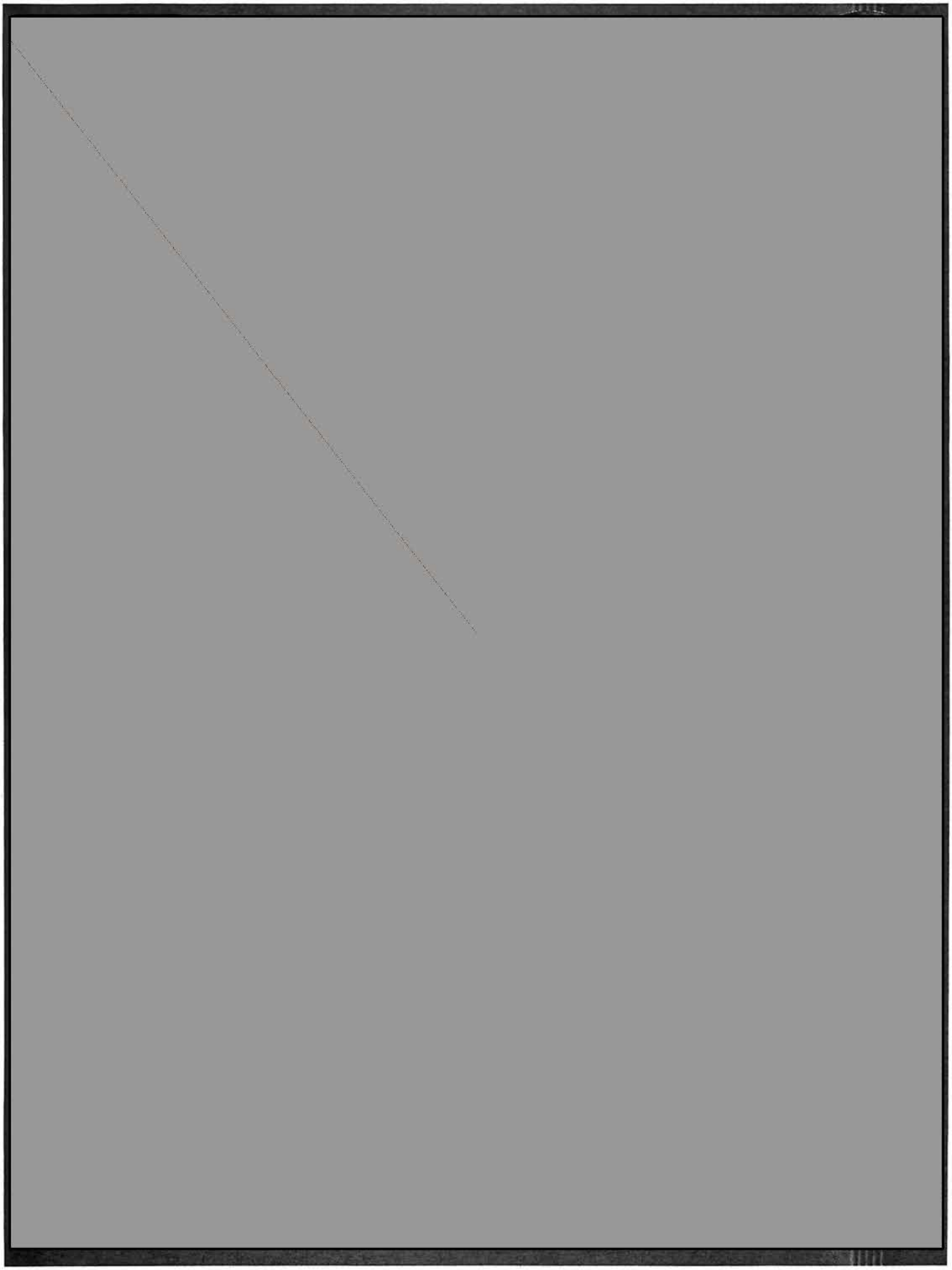


Mirror Worlds, David Gelertner, p30-35

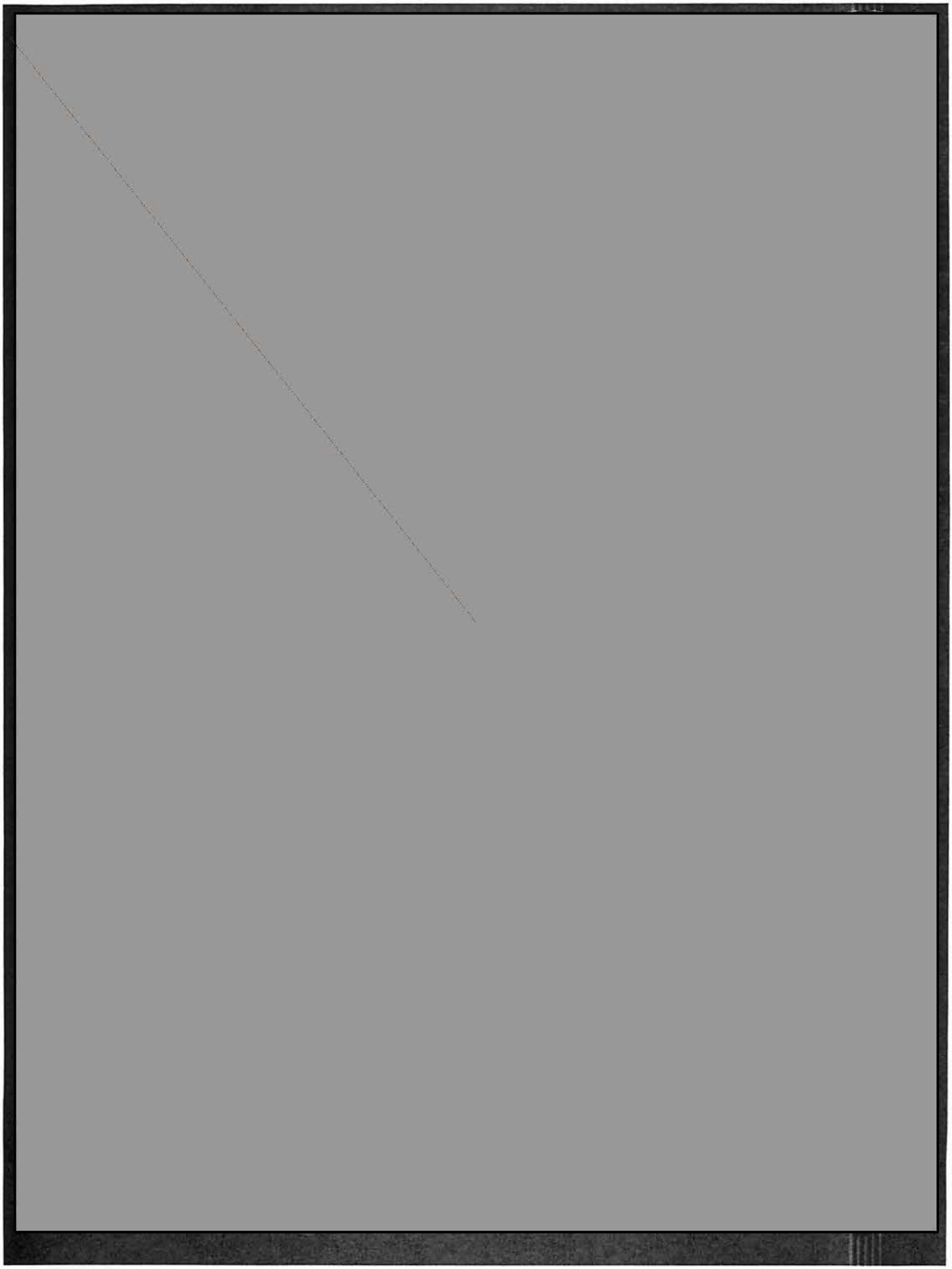
(b)(5)



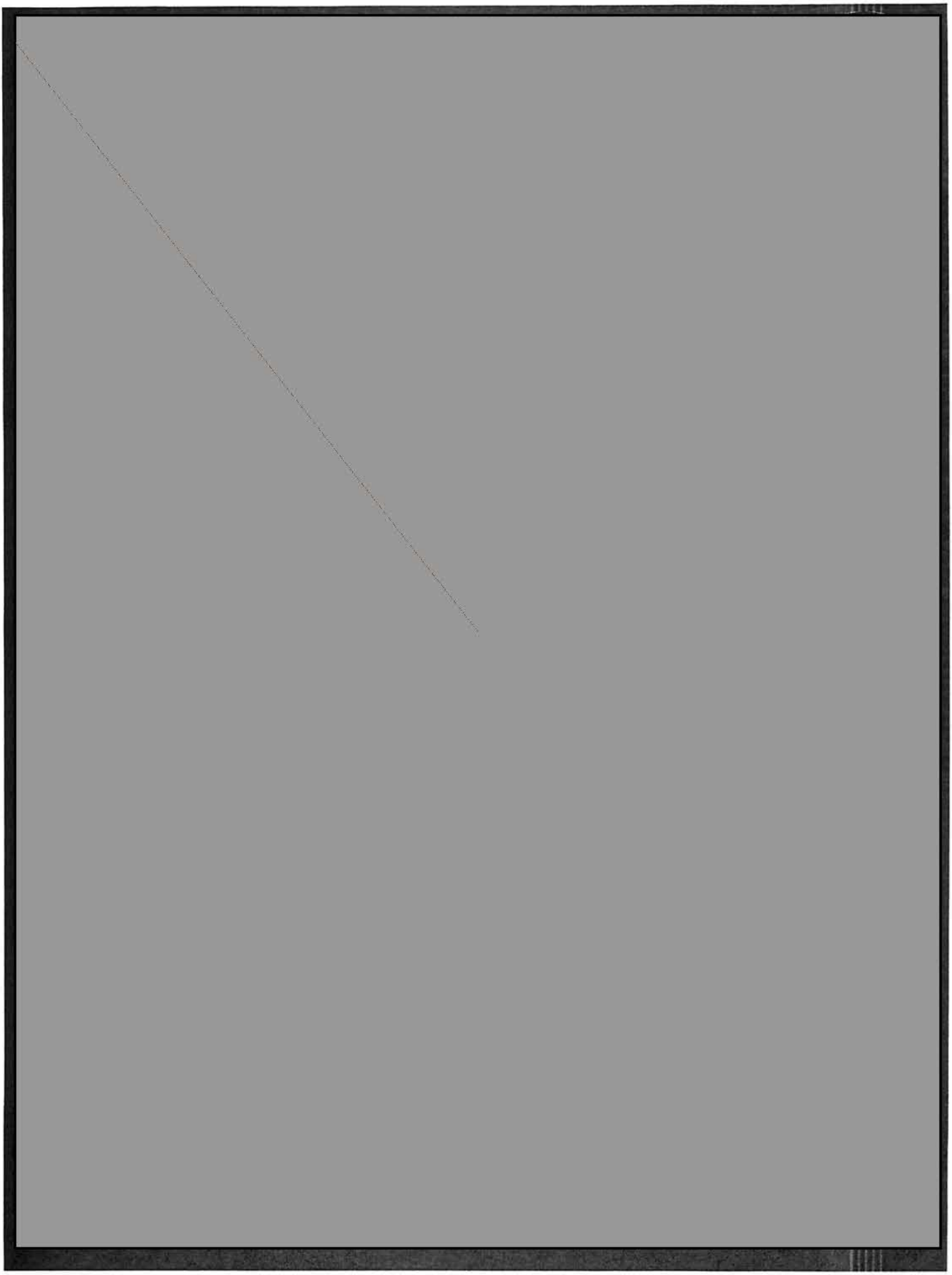
(b)(5)



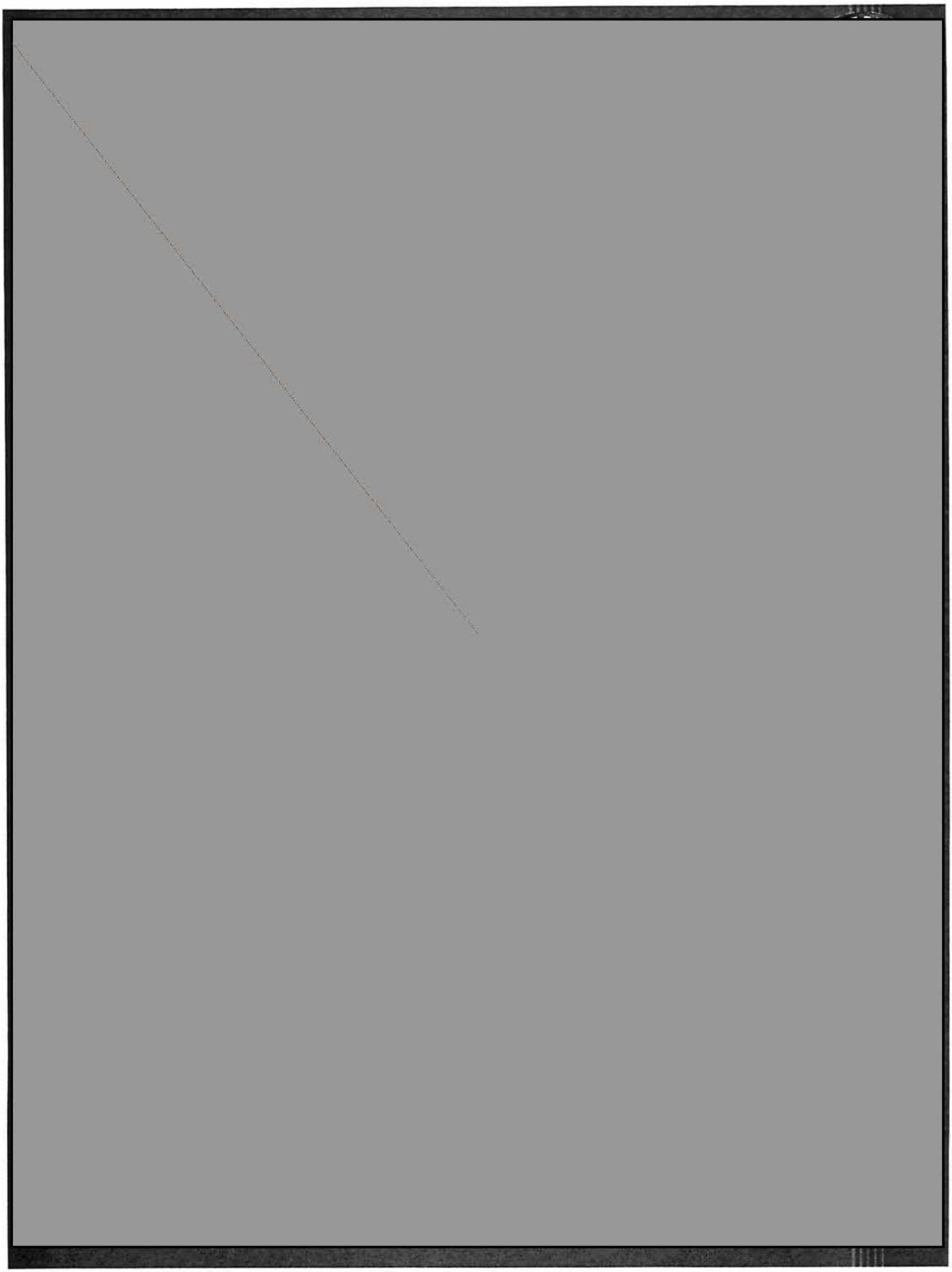
(b)(5)



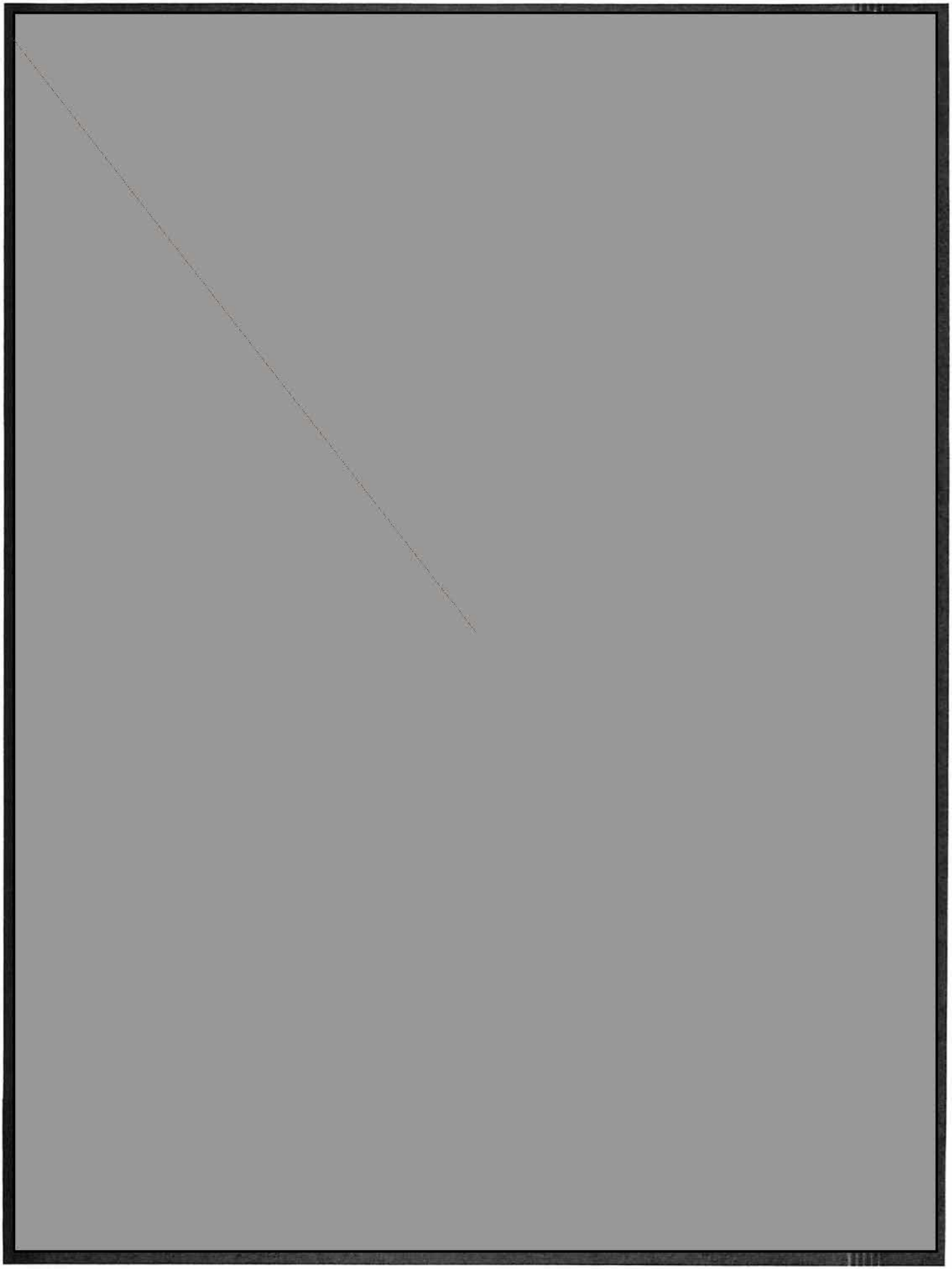
(b)(5)



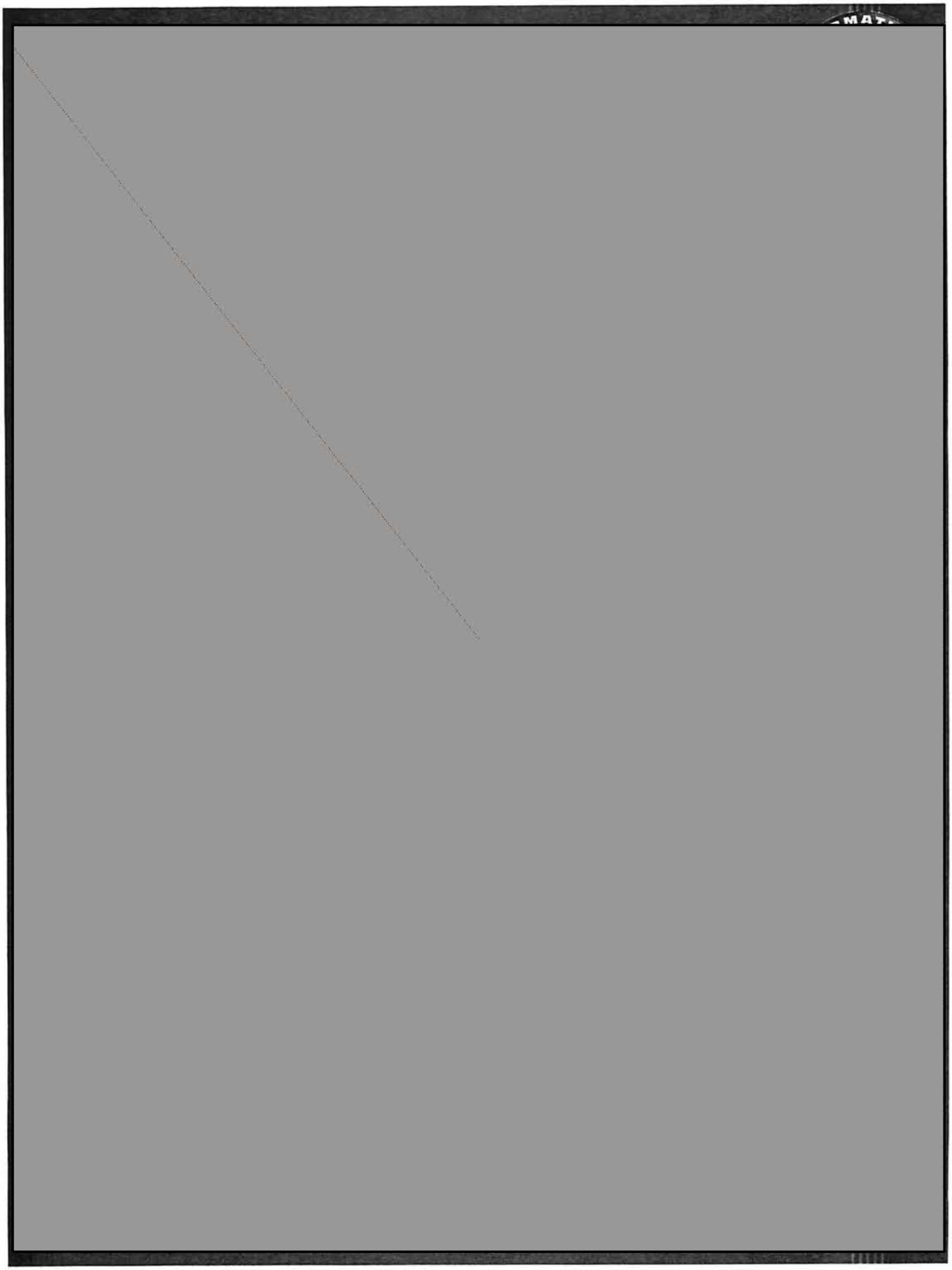
(b)(5)



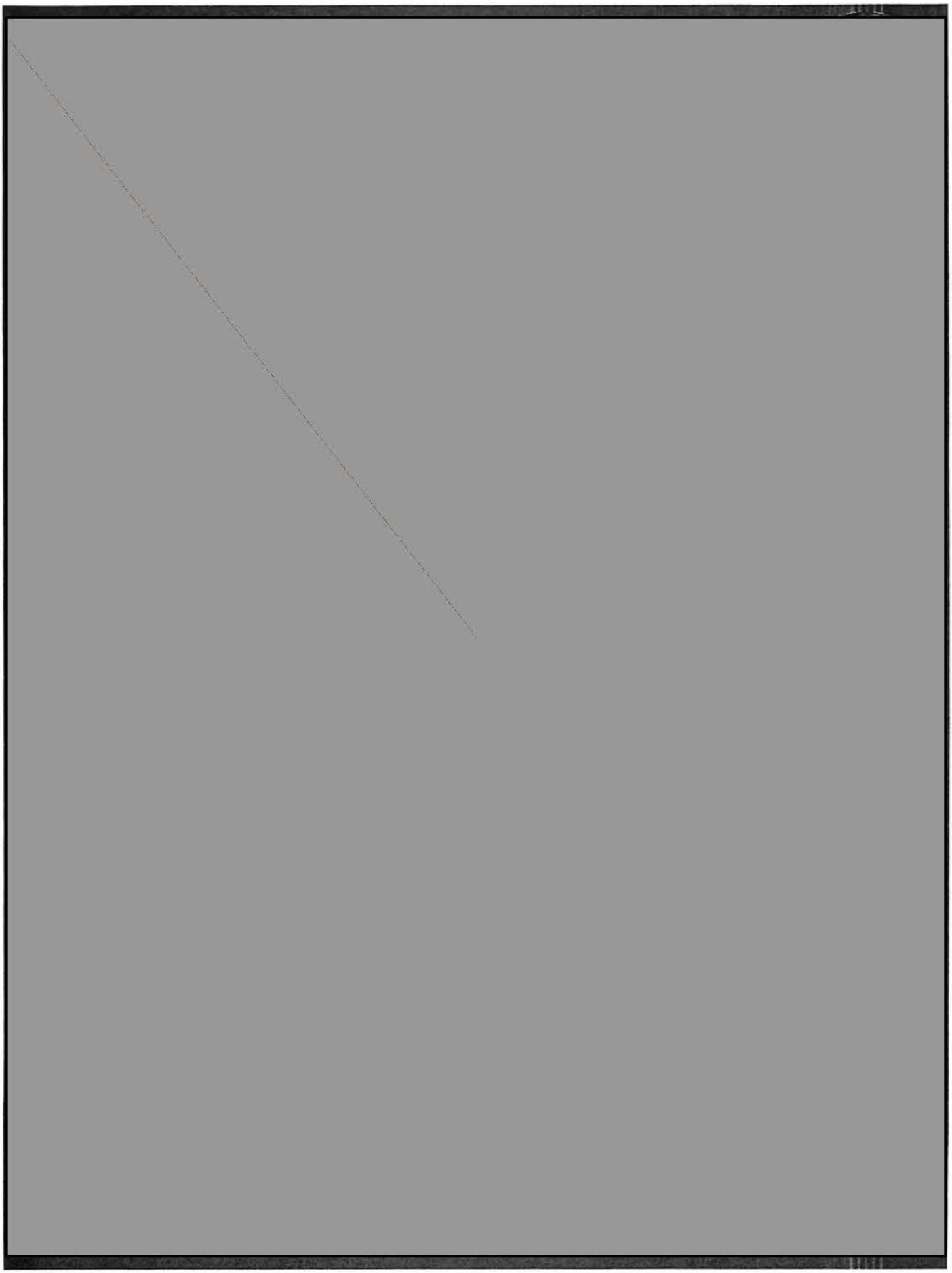
(b)(5)



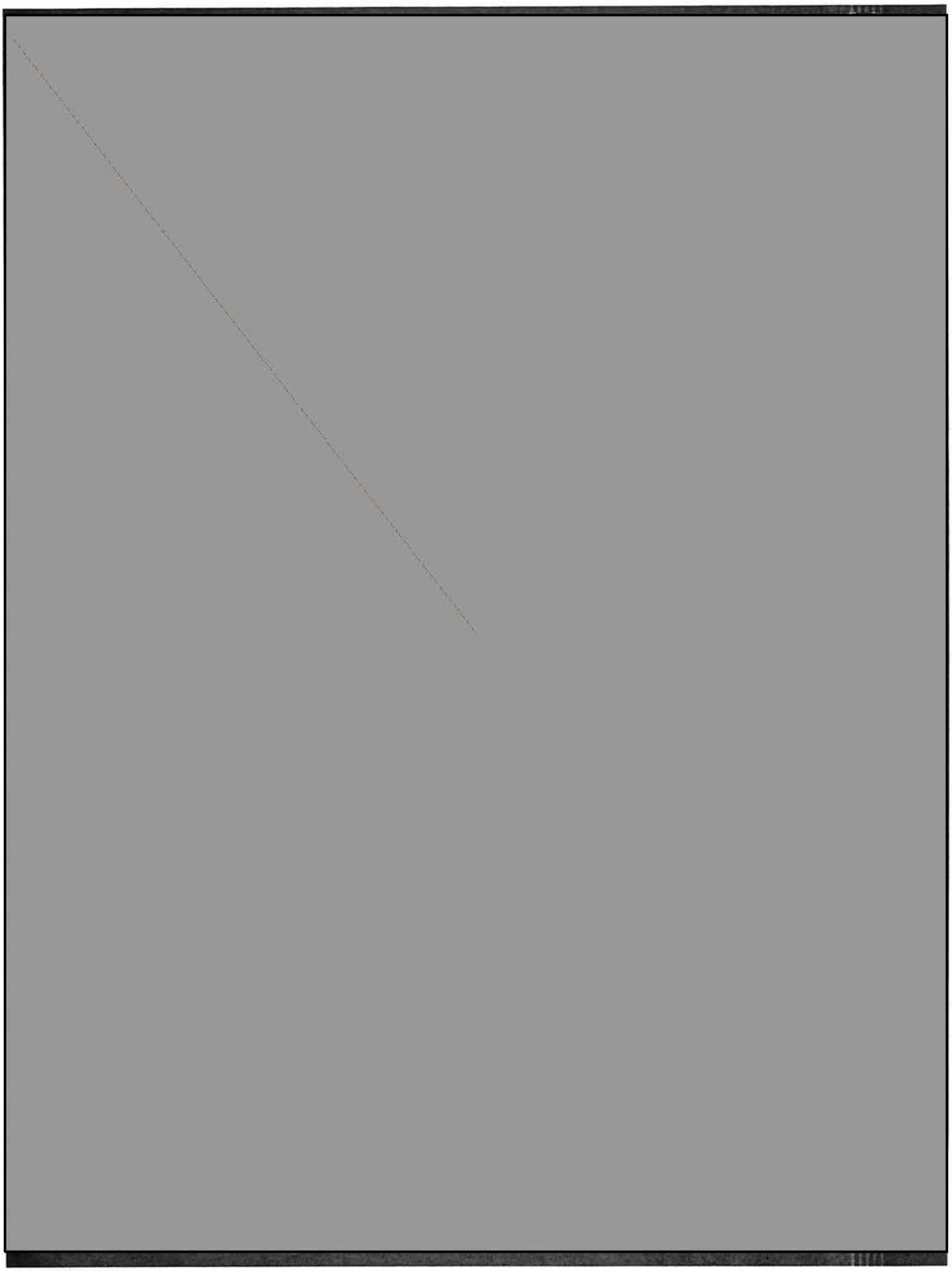
(b)(5)



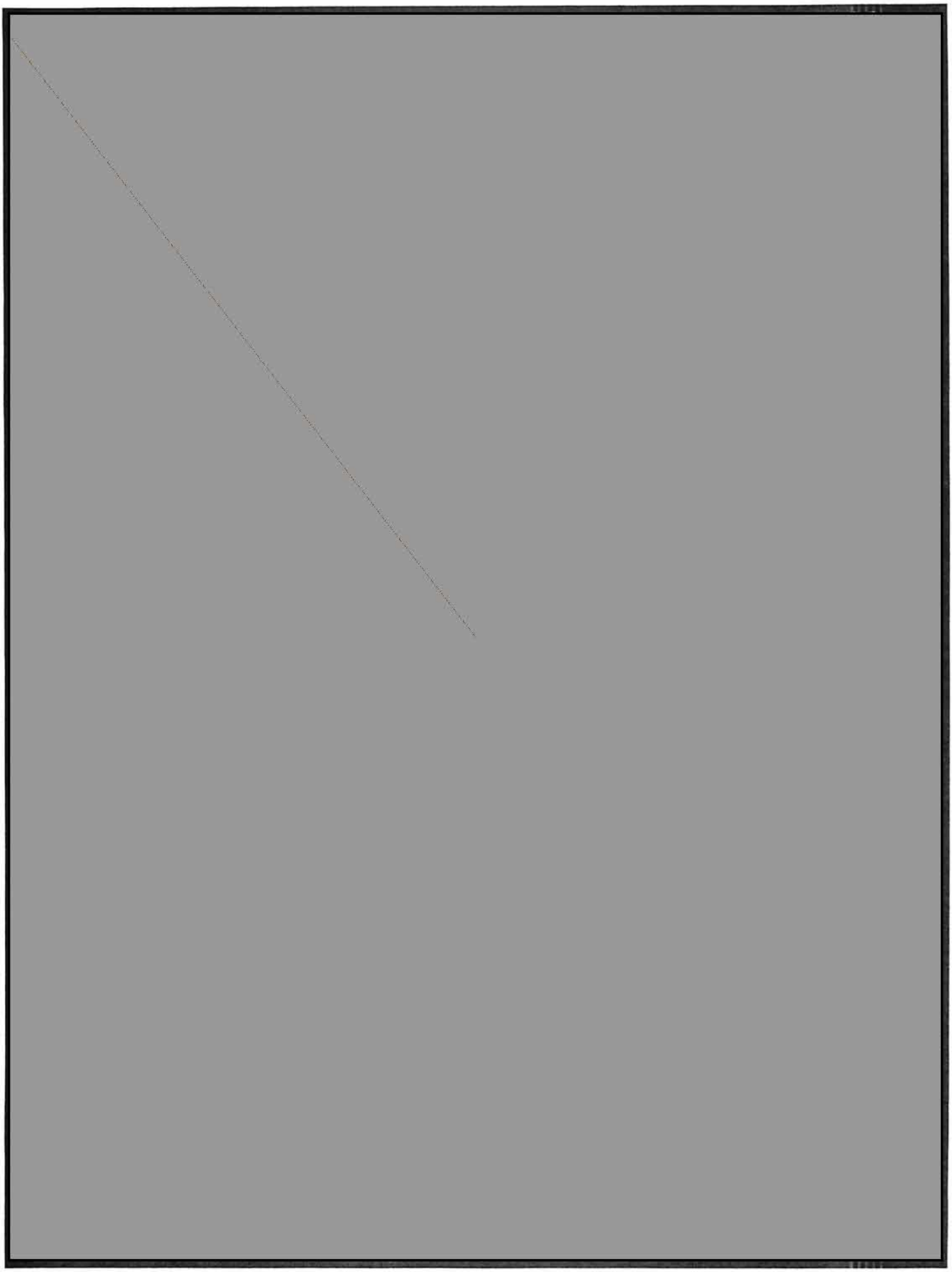
(b)(5)



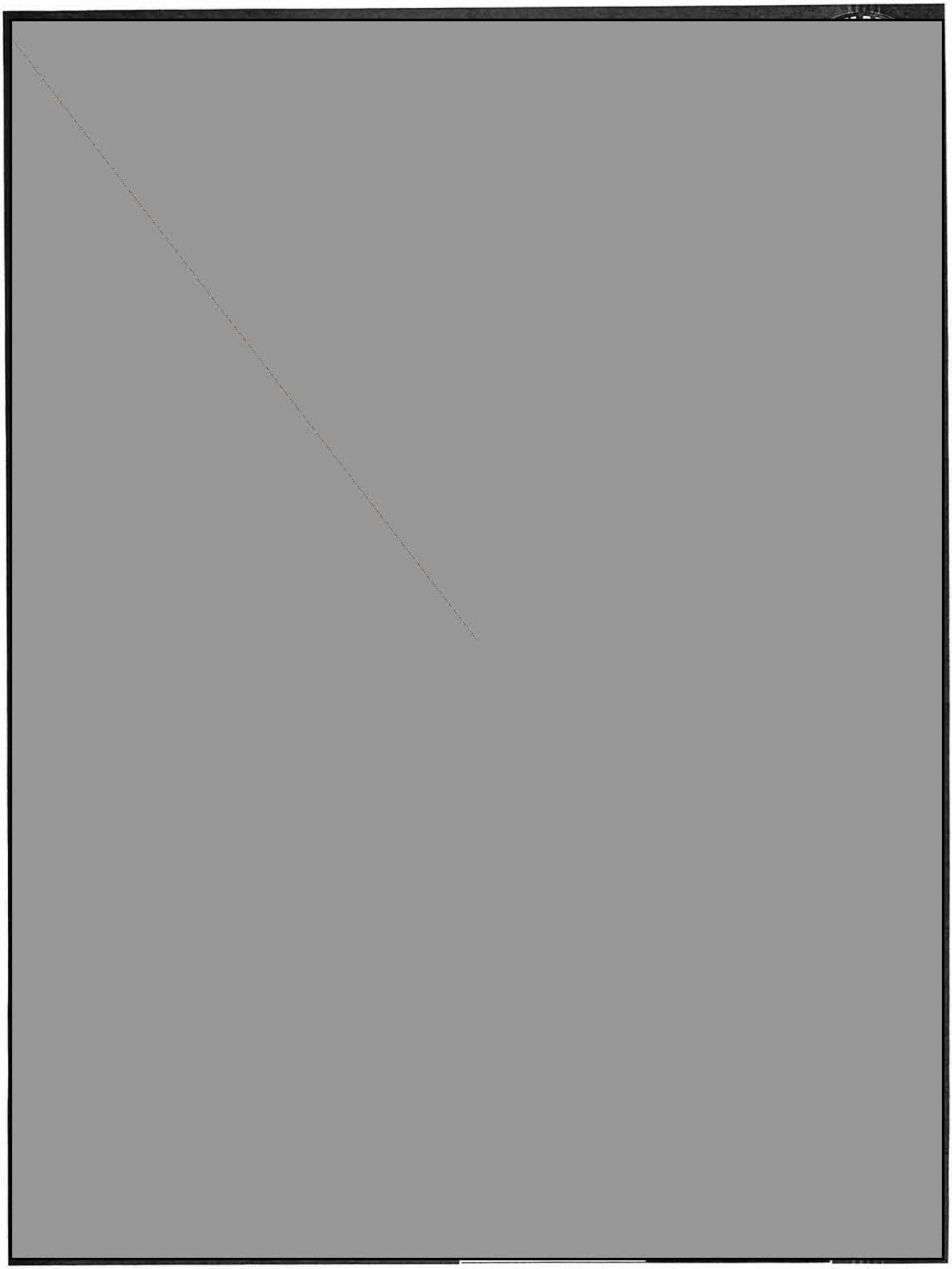
(b)(5)



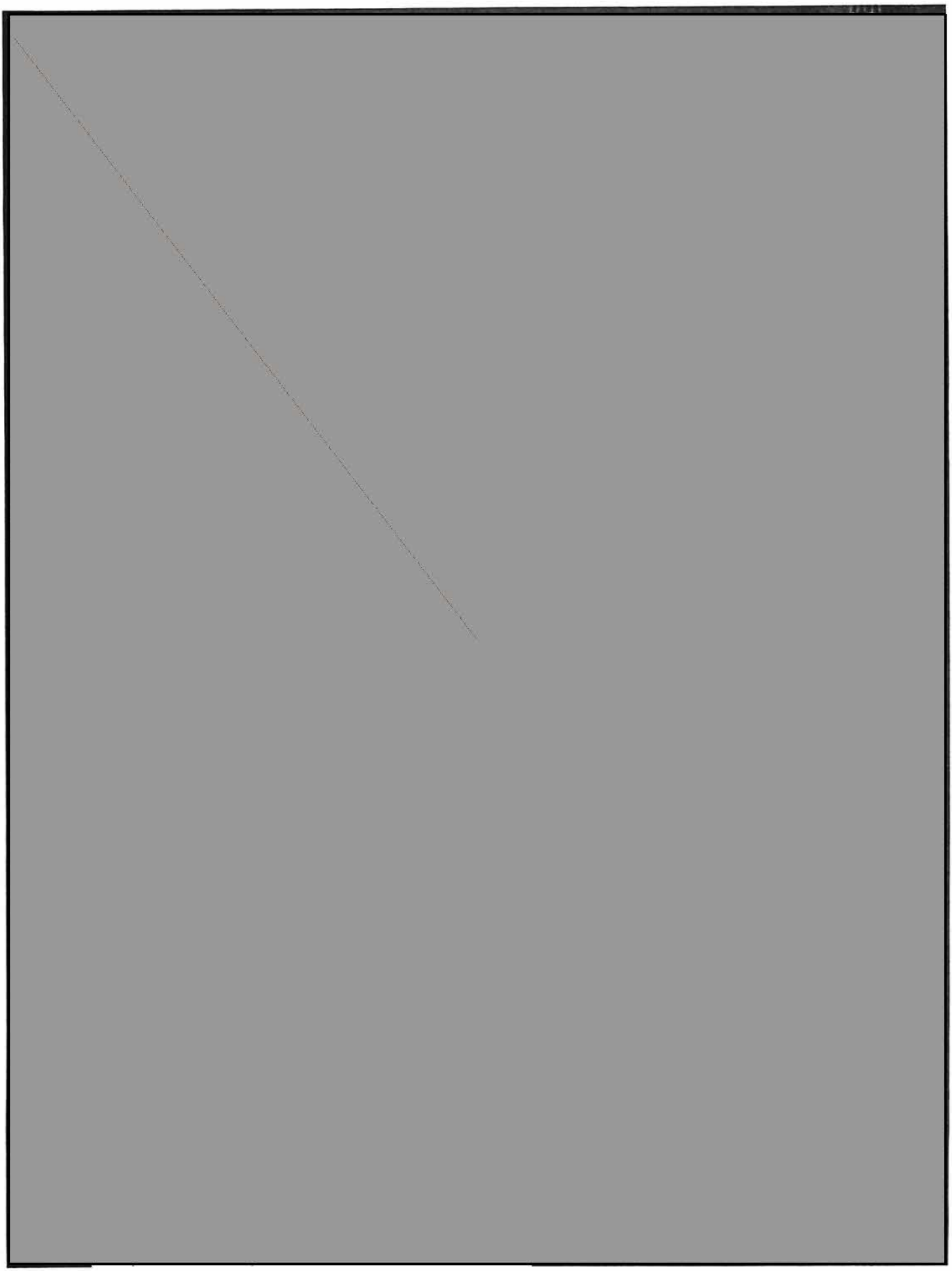
(b)(5)



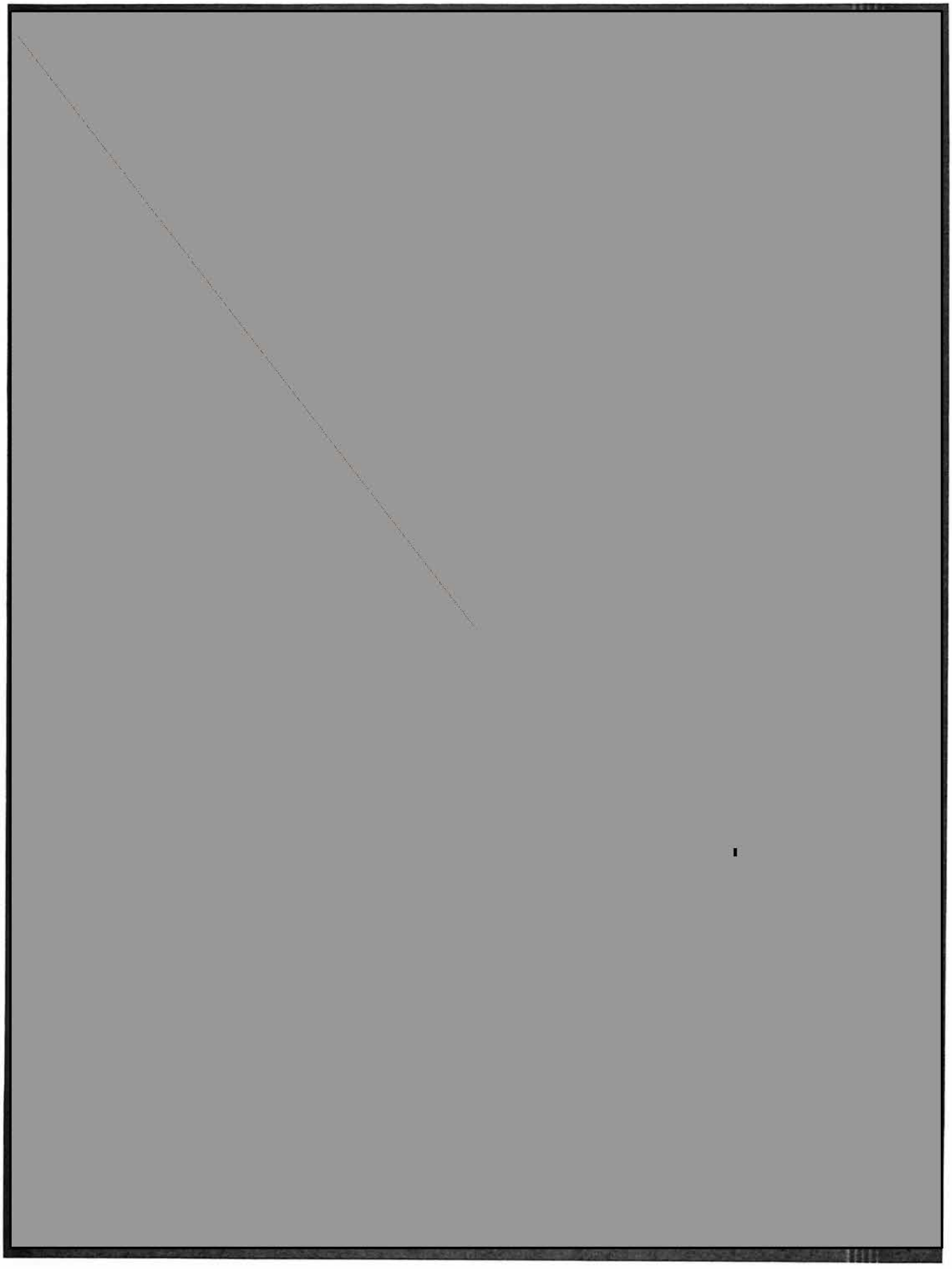
(b)(5)



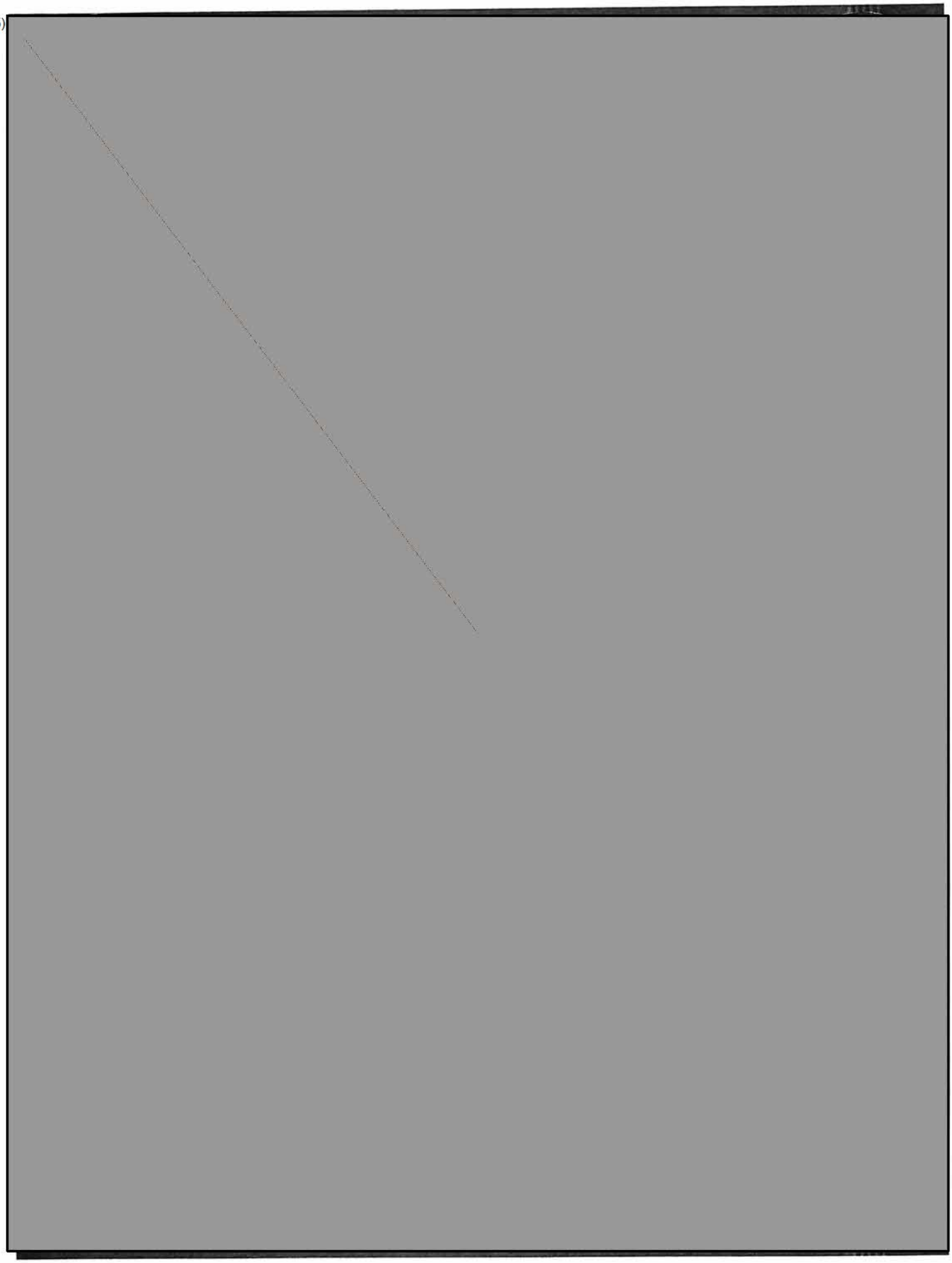
(b)(5)



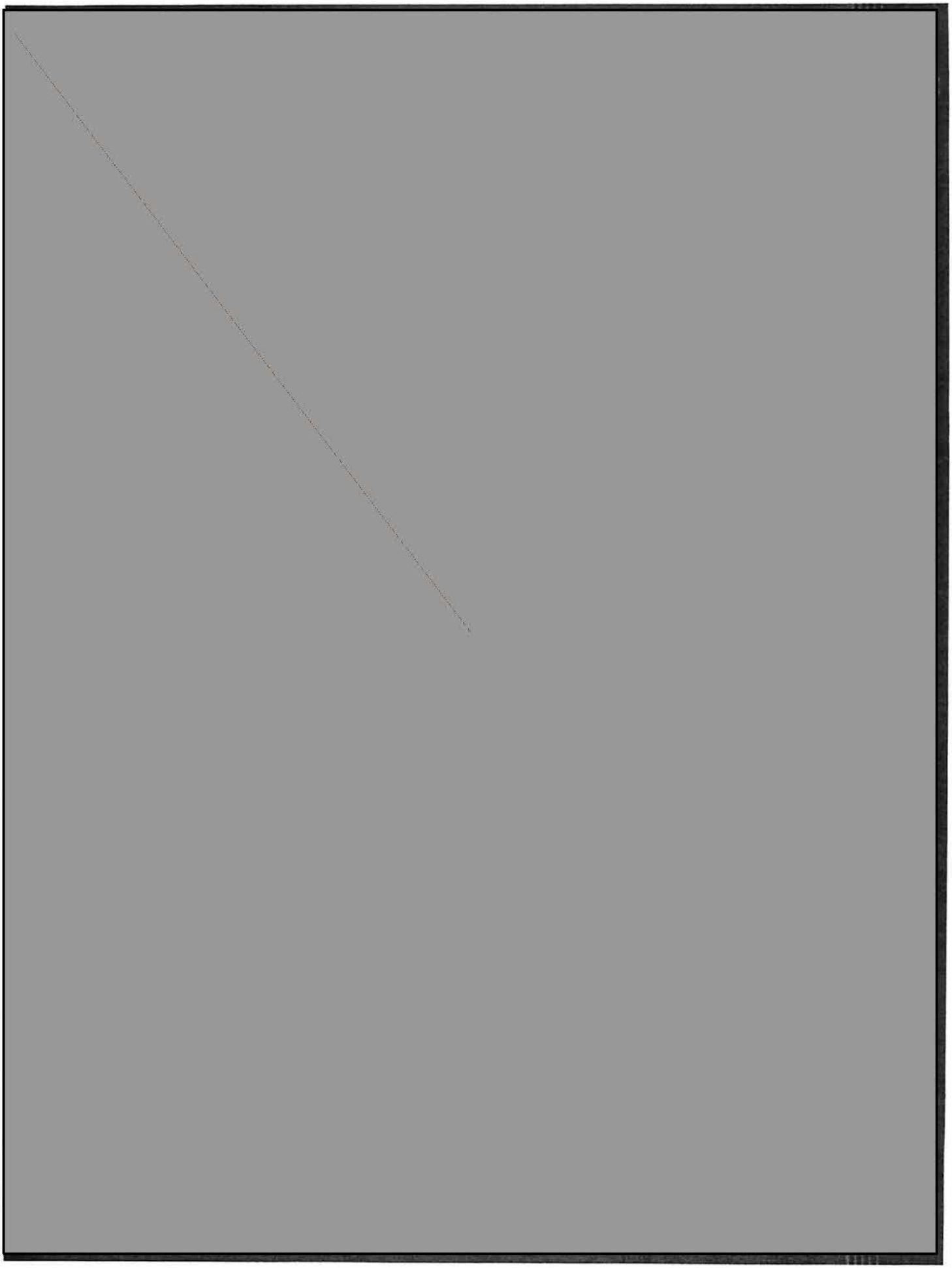
(b)(5)



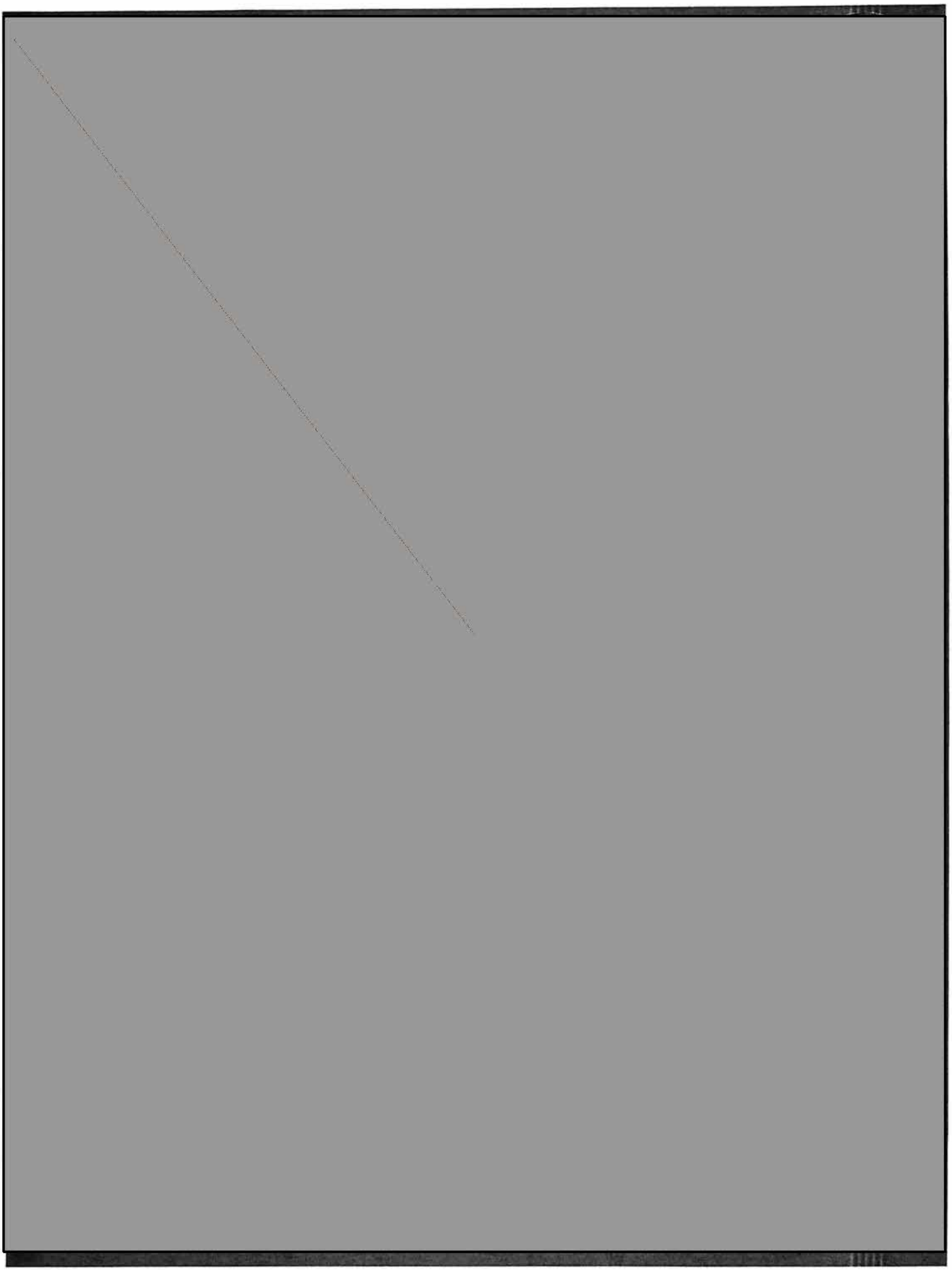
(b)(5)



(b)(5)



(b)(5)



TITLE PAGE

Congressional Inquiries concerning TIA Issues

FOLDER 1.a.1 – 1.a.7 TOC 1

CORRESPONDENCE

- 1.a.1 Senator Boxer
- 1.a.2 Senator Harkin
- 1.a.3 Senator Dorgan
- 1.a.4 Senator Schumer
- 1.a.5 Senator Durbin
- 1.a.6 Senator Levin
- 1.a.7 Senator Deal

ARBARA BOXER
CALIFORNIA

COMMITTEES:
COMMERCE, SCIENCE, AND TRANSPORTATION
ENVIRONMENT AND PUBLIC WORKS
FOREIGN RELATIONS

United States Senate
HART SENATE OFFICE BUILDING
SUITE 112
WASHINGTON, DC 20510-0505
(202) 224-3553
http://boxer.senate.gov/contact

December 16, 2002

Dr. Anthony J. Tether
Director
Defense Advanced Research Projects Agency
3701 North Fairfax Drive
Arlington, VA 22203

Dear Dr. Tether:

I am very concerned about the privacy implications of a Defense Advanced Research Projects Agency (DARPA) program called Total Information Awareness (TIA). While I understand and support DARPA's pursuit of technologies to combat terrorism, I fear components of this particular research program could have serious implications on American civil liberties.

According to news reports and information from DARPA, the goal of TIA is to pursue technologies capable of collecting, storing and analyzing vast amounts of personal data and intelligence. According to DARPA's own Information Awareness Office that manages the program, the types of records to be collected include medical, financial, educational, veterinary, communications and travel.

Given these serious concerns, I respectfully request that you address the following issues:

- Please detail all components of TIA and the goals and objectives for each of these components.
- Please detail any and all changes TIA, if implemented, could potentially require to legal, statutory and regulatory frameworks concerning privacy and governmental access to individual records, including the Privacy Act of 1974.
- What oversight is the Department of Defense or DARPA providing to ensure that taxpayer funding is not being used to develop technologies that would violate current legal, statutory and regulatory frameworks concerning privacy and governmental access to individual records?
- Has the Department of Defense complied with all Freedom of Information Act requests concerning the TIA program? If not, please explain.

1700 MONTGOMERY STREET
SUITE 240
SAN FRANCISCO, CA 94111
(415) 403-0100

312 NORTH SPRING STREET
SUITE 1748
LOS ANGELES, CA 90012
(213) 894-5000

501 'I' STREET
SUITE 7-600
SACRAMENTO, CA 95814
(916) 448-2787

1130 'O' STREET
SUITE 2450
FRESNO, CA 93721
(559) 497-5109

600 'B' STREET
SUITE 2240
SAN DIEGO, CA 92101
(619) 239-3884

201 NORTH 'E' STREET
SUITE 210
SAN BERNARDINO, CA 92401
(909) 888-8525

DEFENSE ADVANCED RESEARCH PROJECTS AGENCY
3701 NORTH FAIRFAX DRIVE
ARLINGTON, VA 22203-1714

JAN 31 2003

The Honorable Barbara Boxer
United States Senate
Washington, DC 20510

Dear Senator Boxer:

Thank you for your letter regarding the Defense Advanced Research Projects Agency (DARPA) Total Information Awareness (TIA) project. I certainly appreciate your continued support of DARPA and our pursuit of technologies to combat terrorism.

TIA is an experimental prototype system comprising innovative technologies that will give the national security community important new capabilities to uncover foreign terrorist acts against Americans. The only data used in this project will be that which is legally usable by the U.S. Government. Two oversight boards are being established to put the appropriate safeguards in place to ensure the privacy rights of American citizens are not violated.

I am pleased with the opportunity to provide you more information on the TIA project; an information paper is enclosed that specifically addresses the issues in your letter. I believe you will find the answers address your concerns about the TIA project, and I hope the information assures you that it is an important undertaking for the war on terrorism.

Sincerely,

Anthony J. Tether
Director

Enclosure:
As stated

DEFENSE ADVANCED RESEARCH PROJECTS AGENCY (DARPA) RESPONSES TO SENATOR BARBARA BOXER'S QUESTIONS AND CONCERNS REGARDING THE DARPA TOTAL INFORMATION AWARENESS (TIA) PROJECT

Issue: Please detail all components of TIA and the goals and objectives for each of these components.

Response: TIA is a Defense Advanced Research Projects Agency (DARPA) research effort to create a prototype network that integrates innovative information technologies for detecting and preempting foreign terrorist activities against Americans.

The program is a multiyear, phased, development effort. The information technologies comprising the experimental prototype network are in three areas: advanced collaborative and decision support tools; language translation; and data search, pattern recognition, and privacy protection technologies.

The collaborative reasoning and decision-support technologies will solve existing coordination problems by enabling analysts from one agency to collaborate effectively with analysts in other agencies. A major challenge to terrorist detection today is the inability to share data from databases maintained legally. The collaborative reasoning and decision-support technologies will punch holes in these "stovepipes."

The language translation technologies will enable the rapid translation of foreign language speech and text giving terrorism analysts the ability to quickly search for clues about emerging terrorist acts. Terrorism experts believe evidence of terrorist activities can be found in open-source foreign language publications and broadcasts. The rapid translation technologies will help analysts search a significant amount of material in a much shorter period than is possible today.

The research into data search and pattern recognition technologies is based on the idea that terrorist planning activities or a likely terrorist attack could be uncovered by searching for patterns indicative of terrorist activities in vast quantities of data. For this research, the TIA project will use only data that is legally obtainable and usable by the U.S. Government. The privacy protection technologies protect the TIA network from internal and external abuses and assure users that their sources and methods cannot be discerned deliberately or accidentally.

Today, the full TIA prototype exists only as a vision. A very limited prototype has been developed to demonstrate a partial set of functionalities in the form of better collaboration and data analysis tools.

When fully developed, TIA will directly support the Department of Defense (DoD) mission to counter foreign terrorist threats outside the United States using DoD's existing intelligence sources. Congress will be fully consulted if and when the full implementation occurs before it is used operationally.

DARPA programs that are developing and evaluating technology for possible inclusion as components of the TIA prototype system are:

- Collaborative Reasoning and Decision Support Technologies
  - The Collaborative Environments and Tools to Aid Human Reasoning (Genoa-II) Program is developing technologies to enable teams of intelligence analysts and operations and policy personnel to be dramatically more effective and efficient in detecting terrorist activities. The project will apply automation to team processes so critical information-sharing tasks can be accomplished much sooner than is possible today. The program will develop tools to model current states, estimate plausible futures, support formal risk analysis, and provide for automated course-of-action planning.
  - The Wargaming the Asymmetric Environment (WAE) Program is developing threat-specific tools to enable analysts and decision makers to predict terrorist attacks. The program has three research thrusts: develop predictive methodologies and technologies that work within the complex and nonlinear characteristics of today's asymmetric adversaries; develop predictive technologies that will generalize from individuals to groups, from attack behavior to more subtle enabling behaviors/ decisions that precede an attack; and develop predictive sequence technologies to allow analysts to test a projected adversary's actions and reactions to potential intervention strategies.
- Language Translation Technologies
  - The Translingual Information Detection, Extraction, and Summarization (TIDES) Program is developing technologies to enable English-speaking operators and analysts to exploit the huge amounts of foreign speech and text available electronically, but currently unexploitable. TIDES is creating powerful new capabilities for detection (finding or discovering needed information), extraction (pulling out key information), summarization (substantially shortening what a user must read), and translation (converting foreign language material to English). Current program plans are to demonstrate capability to detect and track events described in English and Chinese news sources and create an initial capability to process Arabic text and audio sources.

- The Effective Affordable Reusable Speech-to-Text (EARS) Program is developing automatic speech-to-text transcription technologies for the English, Chinese, and Arabic languages with an output substantially more accurate than currently possible. EARS will provide passive listening technology for critical languages and media for a wide range of national security applications. The program will demonstrate the ability to port applicable EARS technology to a new language within 1 month.

- Data Search, Pattern Recognition, and Privacy Protection Technologies

  - The Evidence Extraction and Link Discovery (EELD) Program is developing a suite of technologies that will automatically extract evidence from vast amounts of unstructured textual data (such as intelligence messages or news reports) leading to the discovery of additional relevant relationships and patterns of activity that correspond to unusual events, potential threats, or planned attacks. They will allow for the identification of connected items of information from multiple sources and databases whose significance is not apparent until the connections are made. The program will exploit recent advances in language understanding software to provide a capability to automatically extract facts from textual messages, web pages, and other unstructured data sources at a performance level (90 percent accuracy) comparable to today's ability to extract entities (e.g., people, places, and organizations). The research will use only data that is legally available and obtainable by the U.S. Government or is synthetically generated.

  - The Global Autonomous Language Exploitation (GALE) Program is developing technologies to discover critical information from speech and text of many languages and automatically deliver it in actionable form to military operators and intelligence analysts. Computer programs using advanced algorithms will refine, combine, and package information from broadcasts, conversations, newswire and Internet sources, and examine the data for trends and deviations. The algorithms will discern analyst interest from past behavior (actions and reports) and issue critical alerts, reports, and pointers if the data matches the analyst's past interests. The program will leverage research from the TIDES and EARS programs.

  - The Data Repository, Data Access and Privacy Protection (Genisys) Program will make it easier and cheaper to exploit distributed information sources and improve the ability to represent uncertainty in structured data. The technologies developed under this program will enable database integration with a fraction of today's design effort and retrieve accurate answers with less knowledge of the internal structure of the data. The program will also develop technologies to support automated declassification of data.

  - The Human Identification at a Distance (HumanID) Program is developing automated, multimodal, biometric technologies to detect, recognize, and identify humans, alone or grouped, in disguise or not, at a distance (150 meters), day or night, and in all weather conditions. Automated biometric recognition technologies will provide critical early warning support against terrorist, criminal, and other human-based threats. Biometric signatures will be acquired from various collection sensors

including video, infrared, and multispectral sensors. The program is also examining the feasibility of networking biometric sensors to enable identification capabilities in large areas.

- The Biological Surveillance (BioSurv) Program will dramatically increase the DoD ability to quickly detect a clandestine biological warfare attack. The research will develop disease models to identify abnormal health detectors indicative of a biological attack. The program will perform analyses on hypothesized events to determine the indicators to detect bioterrorist releases.

- DARPA is evaluating technology research recommendations contained in the Information Science and Technology (ISAT) Security with Privacy Study. Two recommendations under investigation are immutable audit and self-reporting data. Immutable audit would track all TIA data search activities and capture data regarding these activities in a strong accounting system that could not be altered or turned off. Such audit control would make it possible to identify abuse of the TIA network as well as the identity of the abuser. Self-reporting data would protect the privacy of data even after it is retrieved from a database. For example, if data is passed to an unauthorized individual, self-reporting data will enable auditors to know who accessed it. This is required to protect the sources and methods of different agencies.

Issue: Please detail any and all changes TIA, if implemented, could potentially require to legal, statutory and regulatory frameworks concerning privacy and governmental access to individual records, including the Privacy Act of 1974.

Response: There are no plans to request a change to the Privacy Act of 1974 to accomplish the research. All research and experiments associated with the TIA system will use only data that is legally obtainable under existing laws.

Today, the full TIA prototype exists only as a vision and is in its first year of a multiyear effort. If the research is successful, organizations wanting to implement the TIA system will determine the types of information they need to make the TIA system effective and whether the data is obtainable without changing existing laws. The Department of Homeland Security (HLS) will likely be the authority to decide on implementation of a TIA-like system for domestic authorities and the type of information and access required to make it an effective system.

Issue: What oversight is the Department of Defense or DARPA providing to ensure that taxpayer funding is not being used to develop technologies that would violate current legal, statutory and regulatory frameworks concerning privacy and governmental access to individual records?

Response: DoD is establishing two oversight boards to put the appropriate safeguards in place to ensure the privacy rights of American citizens are not violated.

Issue: Has the Department of Defense complied with all Freedom of Information Act requests concerning the TIA program?

Response: Yes. DARPA has received eight Freedom of Information Act (FOIA) requests. Two have been answered, and work is underway to answer the remaining six.

Issue: A *Washington Post* editorial questioned the naming of Admiral John Poindexter to oversee the Total Information Office stating that "his record of lying to Congress hardly makes him an ideal protector of the legal system, and his conduct of Iran-contra hardly makes him an advertisement for government competence." Given the intrusive nature of this project and its potential for misuse, do you believe it is appropriate for Admiral Poindexter to lead the office charge with overseeing TIA?

Response: Dr. Poindexter was hired as a Section 1101 appointee under 5 U.S.C. 3104 note and Sec. 1101 of the Strom Thurmond National Defense Authorization Act for Fiscal Year 1999 (P.L. 105-261). Dr. Poindexter has a Ph.D. in physics and, thus, has the technical skills to lead the TIA effort. He has served at the highest levels in the Federal Government and understands how multi-agency decisions are made. These skills are needed to lead the development of effective collaborative reasoning and decision-support technologies that comprise a significant part of the TIA project.

Issue: Please detail the total budget for TIA in fiscal year 2003. What is the planned future year budget of TIA? What funds have been obligated to date? (Please break down funding by component.)

Response: The total budget for TIA and related programs in the FY 2003 President's budget is as follows:

	$ (M)
TIA Project	10
Enabling Technologies (EELD, WAE, Genisys, BioSurv, HumanID)	78
Translation Programs (EARS, TIDES, GALE)	42
Command, Control and Communications (Genoa II)	7
Total	$137

The future years' budget will be outlined in the FY 2004 President's Budget, which will be available soon.

For TIA and related programs that have recorded obligations, the funds obligated to date are:

	FY03 ($M)
Enabling Technologies (EELD, WAE, Genisys, BioSurv, HumanID)	13.1
Translation Programs (EARS, TIDES, GALE)	9.3
Command, Control and Communications (Genoa II)	3.4
TIA	7.1
Total	$32.9

Issue: Please list all non-governmental organizations and experts on civil liberties and privacy that were consulted during the development of TIA.

Response: Nongovernmental organizations and experts on civil liberties and privacy that have been consulted to date are:

- Arlington Institute
- Potomac Institute
- National Academy of Sciences
- Heritage Foundation
- Information Science and Technology Study Group
- Electronic Privacy Information Council

I, A, 2

TOM HARKIN
IOWA

United States Senate

WASHINGTON, DC 20510-1502

December 5, 2002

(202) 224-3254
Fax (202) 224-9369
TTY (202) 224-4633
http://harkin.senate.gov

COMMITTEES
AGRICULTURE
CHAIR

APPROPRIATIONS

HEALTH EDUCATION,
LABOR AND PENSIONS

SMALL BUSINESS

Honorable Donald H Rumsfeld
Secretary of Defense
1000 Defense Pentagon
Washington, DC 20301-1000

Dear Secretary Rumsfeld

I am deeply concerned about the potential privacy implications of the Total Information Awareness (TIA) program of the Defense Advanced Research Projects Agency (DARPA) I understand that DARPA is currently developing a prototype of the TIA system, which if implemented by intelligence and law enforcement agencies, would access and collect a tremendous amount of personal data on all law-abiding Americans

The TIA system would try to identify terrorists before they act by using data mining techniques to find patterns in a vast integrated database of information on Americans In a November 20, 2002 press briefing Department of Defense Undersecretary Pete Aldridge suggested they would use "transactions -- such as passports, visas, work permits, driver's license, credit card, airline tickets, rental cars, gun purchases, chemical purchases -- and events -- such as arrest or suspicious activities and so forth " He also agreed when a reporter asked if "every time I would enter or a citizen would enter a credit card, any banking transaction, any medical – I go see my doctor, any prescription, all of those things become part of this database Every time they use a telephone that call enters the database " He would not speculate whether signals intelligence also would be included

Adm John Poindexter, the Director of the DARPA Information Awareness Office, has indicated plans to eventually integrate this database with new technologies they are developing for Human Identification at Distance and Bio-Surveillance Such technologies might allow identification and tracking of individuals between transactions and events

I believe we need to do all we can to protect American citizens from future terrorist attacks However, this project causes me serious concern about how it would affect the constitutional right to privacy for all law-abiding Americans This concern is only heightened with the choice of Adm Poindexter to run the project It is not clear that Adm Poindexter, who played a highly controversial role in the Iran Contra scandal, is sensitive to the need to limit government power or to the importance of strictly following laws

Specifically, I believe that government collection and use of such data on all American citizens in a vast integrated database is unprecedented and could violate current law Thus I ask your assistance in answering the following questions

150 FIRST AVENUE NE
SUITE 370
CEDAR RAPIDS IA 52401
(319) 365-4504

210 WALNUT STREET
733 FEDERAL BUILDING
DES MOINES IA 50309
(515) 284-4574

1606 BRADY STREET
SUITE 323
DAVENPORT IA 52803
(563) 322-1338

350 WEST 6TH STREET
315 FEDERAL BUILDING
DUBUQUE IA 52001
(563) 582-2130

320 6TH STREET
110 FEDERAL BUILDING
SIOUX CITY IA 51101
(712) 252-1550

- Would full implementation of the TIA system involve government access, collection, or use of personal data on Americans that is currently prohibited under federal laws or regulations? If so, what federal laws may need to be amended, and what new information would be required, to implement effectively the elements of TIA? The federal government has recently passed rules on medical privacy and on banking and financial privacy, and government collection and use of personal data are restricted by the older Privacy Act What restrictions, if any, would apply to government access, collection, or use of personal data for the TIA system from the medical privacy rules, financial privacy rules, and Privacy Act?

- What information would be required to make the TIA system effective? Would implementation of the TIA system require or allow the government to access or collect personal data on Americans that it does not currently collect?

- What checks and balances are being designed into the TIA system to guard against internal abuse of the system and external threats such as computer hacking? In a system based on automated examination of data on all Americans, what is required for a person to become identified as a potential or suspected terrorist?

- What is the current and projected budget for DARPA's Information Awareness Office and for the Total Information Awareness program? How much do you expect it will cost to implement the TIA system? When do you anticipate the TIA system will be ready for use?

- What is Adm Poindexter's salary? Adm Poindexter proposed the TIA program when he was working for a DARPA contractor, Syntek Technologies—what role, if any, does Syntek Technologies have in the TIA program? Do they have a contract with the Information Awareness Office, and if so for what amount?

- Why is the Department of Defense becoming involved in domestic intelligence issues? Is there precedent for this? What experience does DARPA have in research on law enforcement and domestic intelligence methods, and in addressing related privacy concerns?

- Have any Defense, law enforcement, intelligence or other federal agencies inquired about or requested the capabilities of the TIA system? If so, for what purposes? Do you anticipate that this system would be used by the Department of Defense? If so, for what purposes?

Thank you for your assistance Given the great attention to this issue, I would appreciate responses to these questions by December 19, 2002

Sincerely,

Tom Harkin

Tom Harkin
United States Senator

TH/lwu

THE UNDER SECRETARY OF DEFENSE
3010 DEFENSE PENTAGON
WASHINGTON, DC 20301-3010

ACQUISITION,
TECHNOLOGY
AND LOGISTICS

January 30, 2003

The Honorable Tom Harkin
United States Senate
Washington, DC 20510

Dear Senator Harkin:

This is in reply to your recent letter to Secretary Rumsfeld regarding the Defense Advanced Research Projects Agency (DARPA) Total Information Awareness (TIA) project. The DARPA TIA project researches the use of advanced information technologies to develop capabilities to detect foreign terrorist planning activities. If successful, TIA will provide the Department of Defense (DoD) with significantly enhanced capabilities to preempt foreign terrorists from carrying out attacks against deployed U.S. forces and the U.S. homeland.

The TIA project integrates other DARPA programs in three technology areas—advanced collaborative and decision support tools; language translation; and data search, pattern recognition, and privacy protection technologies—into an experimental prototype network. The TIA project is in the first year of a phased, multiyear effort.

DARPA performers will use two types of data for the research: publicly available data, such as foreign language newspapers, journals, and broadcasts; and synthetic data, i.e., artificial information generated to resemble real-world data in which the researchers embed simulated data indicative of terrorist activities

TIA research complies with all U.S. laws, including the Privacy Act. I have directed that both an internal and external oversight board be created to monitor the TIA program to ensure no potential violations of any American's civil liberties occur.

The TIA project is new, and it will be a few years before we know whether the research will be successful. If the usefulness of the technologies is proven in a research environment, we will work with Congress to determine how best to implement the new capabilities and, of course, keep you informed as to its progress.

The enclosed information paper contains answers to the questions in your letter. I trust the information will assure you that the DARPA TIA project is an important undertaking for the war against terrorism.

Sincerely,

E. C. Aldridge, Jr.

Enclosure:
As stated

RESPONSES TO SENATOR TOM HARKIN'S QUESTIONS AND CONCERNS REGARDING THE DEFENSE ADVANCED RESEARCH PROJECTS AGENCY (DARPA) TOTAL INFORMATION AWARENESS (TIA) PROJECT

Question: Would full implementation of the TIA [Total Information Awareness] system involve government access, collection, or use of personal data on Americans that is currently prohibited under federal laws or regulations? If so, what federal laws may need to be amended, and what new information would be required, to implement effectively the elements of TIA? The federal government has recently passed rules on medical privacy and on banking and financial privacy, and government collection and use of personal data are restricted by the older Privacy Act. What restrictions, if any, would apply to government access, collection, or use of personal data for the TIA system from the medical privacy rules, financial privacy rules, and Privacy Act?

Answer: The Defense Advanced Research Projects Agency's (DARPA) Information Awareness Office (IAO) recently started research and experimentation to develop a prototype TIA system; the program is multiyear, phased effort planned to last approximately 5 years. The information technologies comprising the experimental prototype network are in three areas: advanced collaborative and decision support tools; language translation; and data search, pattern recognition, and privacy protection technologies.

The collaborative reasoning and decision-support technologies will solve existing coordination problems by enabling analysts from one agency to collaborate effectively with analysts in other agencies. A major challenge to terrorist detection today is the inability to quickly search, correlate, and share data from databases maintained legally by our intelligence, counter-intelligence, and law enforcement agencies. The collaborative reasoning and decision-support technologies will punch holes in these "stovepipes."

The language translation technologies will enable the rapid translation of foreign language speech and text giving terrorism analysts the ability to quickly search for clues about emerging terrorist acts. Terrorism experts believe evidence of terrorist activities can be found in open-source foreign language publications and broadcasts. The rapid translation technologies will help analysts search a significant amount of material in a much shorter period than is possible today.

The research into data search and pattern recognition technologies is based on the idea that terrorist planning activities or a likely terrorist attack could be uncovered by searching for patterns indicative of terrorist activities in vast quantities of data. Terrorists must engage in certain transactions to coordinate and conduct attacks against Americans; these transactions leave signatures (patterns) that may be detectable. For this research, the TIA project will use only data that is legally obtainable and usable by the U.S. Government. The privacy protection technologies protect the TIA network from internal and external abuses and assure users that their sources and methods cannot be discerned deliberately or accidentally.

Today, the full TIA prototype exists only as a vision. A very limited prototype has been developed to demonstrate a partial set of functionalities in the form of better collaboration and data analysis tools. It is too early to determine what types of data will be required for the full implementation of the TIA system in the future. However, Congress will be fully consulted if and when the full implementation occurs before it is used operationally.

Question: What information would be required to make the TIA system effective? Would implementation of the TIA system require or allow the government to access or collect personal data on Americans that it does not currently collect?

Answer: The TIA system is in its first year of an anticipated multiyear research effort. It is too early to identify accurately the types of information that would make the TIA system effective. The implementation of a TIA-like system to support the DoD mission to preempt and counter foreign terrorist threats outside the United States will not require the collection of any data that is not currently collected. The Department of Homeland Security (HLS) will likely be the authority to decide on implementation of a TIA-like system for domestic authorities and the type of information and access required to make it an effective system.

Question: What checks and balances are being designed into the TIA system to guard against internal abuse of the system and external threats such as computer hacking? In a system based on automated examination of data on all Americans, what is required for a person to become identified as a potential or suspected terrorist?

Answer: Part of the TIA research effort involves developing technologies to protect the system from internal abuses and external threats, such as computer hackers.

TIA research efforts are just beginning to examine the processes for identifying foreign individuals in foreign lands as terrorists. There are no efforts performed or considered in the TIA project to identify U.S. citizens as terrorists. The HLS, working with other Federal agencies, will likely be the authority to determine the criteria to identify terrorists in the United States.

Question: What is the current and projected budget for DARPA's Information Awareness Office and for the Total Information Awareness program? How much do you expect it will cost to implement the TIA system? When do you anticipate the TIA system will be ready for use?

Answer: The fiscal year (FY) 2003 budget for IAO is $137 million. Within this amount, TIA program funding is $10 million. The financial data for fiscal years 2004 to 2007 is in the President's FY 2004 budget, which Congress will soon receive.

At this time, we do not know how much it will cost to implement the TIA system since it is too early to determine which of the research technologies will be successful. During the first 36 months, a range of ideas will be developed via limited demonstrations and preliminary prototypes. During the final 24 months, the most promising research avenues will be extended to support production of a scalable, leave-behind system prototype.

There is a chance some technologies—such as collaboration and information visualization technologies, foreign language machine translation, and speech-to-text transcription tools—will transition to some users during the research period.

Question: What is Adm. Poindexter's salary? Adm. Poindexter proposed the TIA program when he was working for a DARPA contractor, Syntek Technologies—what role, if any, does Syntek Technologies have in the TIA program? Do they have a contract with the Information Awareness Office, and if so, for what amount?

Answer: Dr. Poindexter is serving the Government as an Administratively Determined employee hired under the Experimental Hiring Authority (Section 1101 of the Strom Thurmond National Defense Authorization Act 1998). His annual salary rate for the Washington, DC, locality in 2002 is $138,200.

Dr. Poindexter's former role as an employee of Syntek Technologies involved programmatic assistance for managing DARPA's Project Genoa, which concluded in the second quarter of FY 2002. Project Genoa developed tools and a system concept for collaborative crisis understanding and management for the national security community including commanders of the Unified Commands. Today's TIA program is a natural extension and expansion of Project Genoa, and Dr. Poindexter's previous experience lends itself well to a continued program management role on behalf of DARPA.

Syntek Technologies does not have a contract with IAO. The company has a subcontract with SRS Technologies, the principal management and advisory contractor for IAO that provides business process and administrative support. Syntek is funded to provide the full-time equivalent of 1.5 persons at an annualized rate of $327, 000 for labor and other direct costs.

Question: Why is the Department of Defense becoming involved in domestic intelligence issues? Is there precedent for this? What experience does DARPA have in research on law enforcement and domestic intelligence methods, and in addressing related privacy concerns?

Answer: The research projects in IAO will not give DoD domestic intelligence capabilities. What they will do is provide DoD with significantly enhanced capabilities to preempt foreign terrorists from carrying out attacks against deployed U.S. forces and the U.S. homeland.

DARPA should lead this research. In its 44-year history, DARPA has undertaken numerous high-risk research efforts that led to significant capabilities. Many existing information technologies—including the Internet—started as advanced DARPA research projects. DARPA has had joint programs with the Federal Bureau of Investigations (FBI) and the U.S. Customs Service to develop technology that could be used for detecting explosives and drugs at airports and sea ports. The IAO is following a similar path of technical innovation with its research into advanced information capabilities that will give the United States a decisive edge in the global war on terrorism. All Americans share the frustration associated with vague warnings of terrorist threats. DARPA believes that IAO information technology programs will help the U.S. Government reduce those generic reports to advance notice of specific threatening acts.

DARPA is alert to the need to exercise extreme caution with regard to research that approaches anywhere near law enforcement and domestic intelligence areas. As a safeguard, DoD has an oversight board monitoring the research to ensure DARPA does not violate any laws.

Question: Have any Defense, law enforcement, intelligence or other federal agencies inquired about or requested the capabilities of the TIA system? If so, for what purposes? Do you anticipate that this system would be used by the Department of Defense? If so, for what purposes?

Answer: DARPA is examining the usefulness of TIA technologies in a research environment. In conjunction with these efforts, the intelligence, counterintelligence, and law enforcement communities will be invited to test these technologies. DARPA is working with the U.S. Army Intelligence and Security Command; Office of the Secretary of Defense, Counterintelligence Field Activity, Assessment and Technology Directorate; Special Operations Command; and Joint Forces Command to facilitate experimentation in an operational environment using data already in their databases. The TIA hardware and software tools under development to analyze the operational units' existing databases do not require or give access to any information to which the Intelligence Community does not already have lawful access. DARPA has had discussions with the FBI, Central Intelligence Agency, National Security Agency, and HLS. Some of these organizations want to participate in future experiments, but the plans for their involvement are not yet finalized.

If the research is successful, a TIA-like network will provide DoD with significantly enhanced capabilities to preempt foreign terrorists from carrying out attacks against deployed U.S. forces and the U.S. homeland.

BYRON L. DORGAN
NORTH DAKOTA
713 HART SENATE OFFICE BUILDING
WASHINGTON, DC 20510-3405

COMMITTEES:
APPROPRIATIONS
COMMERCE, SCIENCE & TRANSPORTATION
ENERGY & NATURAL RESOURCES
INDIAN AFFAIRS

CHAIRMAN, DEMOCRATIC POLICY COMMITTEE

OFFICE OF THE SECRETARY OF DEFENSE

2002 NOV 27 PM 12:29

United States Senate

WASHINGTON, DC 20510-3405

STATE OFFICES:

312 FEDERAL BUILDING
THIRD AND ROSSER AVENUE
P.O. BOX 2579
BISMARCK, ND 58502

112 ROBERTS STREET, ROOM 110
P.O. BOX 2250
FARGO, ND 58107

102 NORTH 4TH STREET, ROOM 104
GRAND FORKS, ND 58201

100 1ST STREET, S.W., ROOM 105
MINOT, ND 58701

November 26, 2002

The Honorable Donald Rumsfeld
Secretary of Defense
1000 Defense Pentagon
Washington, DC 20301-1000

Dear Mr. Secretary,

I know you are aware of my strong support for our nation's defense through my work on the Defense Appropriations Subcommittee. I am committed to ensuring that our United States military, and especially our men and women in uniform, have the resources they need to defend our freedom.

But I have learned recently of one program in the Department of Defense that I can not support. And I am writing to ask you to put an end to it.

I'm referring to the Total Information Awareness (TIA) Program.

As I understand it, the Defense Department is developing a prototype system to accomplish a type of "data mining." The system, when implemented, would allow federal agencies to sift through databases tracking virtually every facet of American citizens' activities, including credit card purchases; medical, travel, and academic records; and physical characteristics, such as fingerprints, iris patterns, and gait. The system would seek out strange, unusual or questionable patterns of behavior indicative of a potential terrorist threat.

Our government simply can't be doing this! In America, it is not appropriate for the federal government to gather private information about the most sensitive details of all of the American people's lives, and rummage through this information in search of "suspicious" behavior.

I agree that the War on Terrorism is a deadly serious one. And we must employ new and aggressive techniques to win this war. But our war on terrorism does not justify a massive program of government snooping on the daily lives, daily habits and daily activities of the American people.

I believe that Congress will put an end to the Total Information Awareness Program, unless you decide to do so first. Congress will not approve additional funding to take the program beyond its early stages, nor will it approve a change in the 1974 Privacy Act – which, as I understand it, would be required for implementation of the program.

U18953 /02

For that reason, I hope the money you are now spending on this project would be terminated, and put to other uses in the Pentagon that would strengthen both our national defense and our war against terrorism.

On a final note, I think you have done an excellent job as Secretary of Defense, and I have been pleased to work with you in my capacity as a member of the Defense Appropriations Subcommittee. But I wanted to be frank about my opposition to the Total Information Awareness Project. This project is a bad idea! We cannot fight terror by diminishing our civil liberties and our right to privacy.

Sincerely,

Byron L. Dorgan
United States Senator

Today, the full TIA prototype exists only as a vision. A very limited prototype has been developed to demonstrate a partial set of functionalities in the form of better collaboration and data analysis tools. It is too early to determine what types of data will be required for the full implementation of the TIA system in the future. However, Congress will be fully consulted if and when the full implementation occurs before it is used operationally.

Question: What information would be required to make the TIA system effective? Would implementation of the TIA system require or allow the government to access or collect personal data on Americans that it does not currently collect?

Answer: The TIA system is in its first year of an anticipated multiyear research effort. It is too early to identify accurately the types of information that would make the TIA system effective. The implementation of a TIA-like system to support the DoD mission to preempt and counter foreign terrorist threats outside the United States will not require the collection of any data that is not currently collected. The Department of Homeland Security (HLS) will likely be the authority to decide on implementation of a TIA-like system for domestic authorities and the type of information and access required to make it an effective system.

Question: What checks and balances are being designed into the TIA system to guard against internal abuse of the system and external threats such as computer hacking? In a system based on automated examination of data on all Americans, what is required for a person to become identified as a potential or suspected terrorist?

Answer: Part of the TIA research effort involves developing technologies to protect the system from internal abuses and external threats, such as computer hackers.

TIA research efforts are just beginning to examine the processes for identifying foreign individuals in foreign lands as terrorists. There are no efforts performed or considered in the TIA project to identify U.S. citizens as terrorists. The HLS, working with other Federal agencies, will likely be the authority to determine the criteria to identify terrorists in the United States.

Question: What is the current and projected budget for DARPA's Information Awareness Office and for the Total Information Awareness program? How much do you expect it will cost to implement the TIA system? When do you anticipate the TIA system will be ready for use?

Answer: The fiscal year (FY) 2003 budget for IAO is $137 million. Within this amount, TIA program funding is $10 million. The financial data for fiscal years 2004 to 2007 is in the President's FY 2004 budget, which Congress will soon receive.

At this time, we do not know how much it will cost to implement the TIA system since it is too early to determine which of the research technologies will be successful. During the first 36 months, a range of ideas will be developed via limited demonstrations and preliminary prototypes. During the final 24 months, the most promising research avenues will be extended to support production of a scalable, leave-behind system prototype.

There is a chance some technologies—such as collaboration and information visualization technologies, foreign language machine translation, and speech-to-text transcription tools—will transition to some users during the research period.

Question: What is Adm. Poindexter's salary? Adm. Poindexter proposed the TIA program when he was working for a DARPA contractor, Syntek Technologies—what role, if any, does Syntek Technologies have in the TIA program? Do they have a contract with the Information Awareness Office, and if so, for what amount?

Answer: Dr. Poindexter is serving the Government as an Administratively Determined employee hired under the Experimental Hiring Authority (Section 1101 of the Strom Thurmond National Defense Authorization Act 1998). His annual salary rate for the Washington, DC, locality in 2002 is $138,200.

Dr. Poindexter's former role as an employee of Syntek Technologies involved programmatic assistance for managing DARPA's Project Genoa, which concluded in the second quarter of FY 2002. Project Genoa developed tools and a system concept for collaborative crisis understanding and management for the national security community including commanders of the Unified Commands. Today's TIA program is a natural extension and expansion of Project Genoa, and Dr. Poindexter's previous experience lends itself well to a continued program management role on behalf of DARPA.

Syntek Technologies does not have a contract with IAO. The company has a subcontract with SRS Technologies, the principal management and advisory contractor for IAO that provides business process and administrative support. Syntek is funded to provide the full-time equivalent of 1.5 persons at an annualized rate of $327, 000 for labor and other direct costs.

Question: Why is the Department of Defense becoming involved in domestic intelligence issues? Is there precedent for this? What experience does DARPA have in research on law enforcement and domestic intelligence methods, and in addressing related privacy concerns?

Answer: The research projects in IAO will not give DoD domestic intelligence capabilities. What they will do is provide DoD with significantly enhanced capabilities to preempt foreign terrorists from carrying out attacks against deployed U.S. forces and the U.S. homeland.

DARPA should lead this research. In its 44-year history, DARPA has undertaken numerous high-risk research efforts that led to significant capabilities. Many existing information technologies—including the Internet—started as advanced DARPA research projects. DARPA has had joint programs with the Federal Bureau of Investigations (FBI) and the U.S. Customs Service to develop technology that could be used for detecting explosives and drugs at airports and sea ports. The IAO is following a similar path of technical innovation with its research into advanced information capabilities that will give the United States a decisive edge in the global war on terrorism. All Americans share the frustration associated with vague warnings of terrorist threats. DARPA believes that IAO information technology programs will help the U.S. Government reduce those generic reports to advance notice of specific threatening acts.

DARPA is alert to the need to exercise extreme caution with regard to research that approaches anywhere near law enforcement and domestic intelligence areas. As a safeguard, DoD has an oversight board monitoring the research to ensure DARPA does not violate any laws.

Question: Have any Defense, law enforcement, intelligence or other federal agencies inquired about or requested the capabilities of the TIA system? If so, for what purposes? Do you anticipate that this system would be used by the Department of Defense? If so, for what purposes?

Answer: DARPA is examining the usefulness of TIA technologies in a research environment. In conjunction with these efforts, the intelligence, counterintelligence, and law enforcement communities will be invited to test these technologies. DARPA is working with the U.S. Army Intelligence and Security Command; Office of the Secretary of Defense, Counterintelligence Field Activity, Assessment and Technology Directorate; Special Operations Command; and Joint Forces Command to facilitate experimentation in an operational environment using data already in their databases. The TIA hardware and software tools under development to analyze the operational units' existing databases do not require or give access to any information to which the Intelligence Community does not already have lawful access. DARPA has had discussions with the FBI, Central Intelligence Agency, National Security Agency, and HLS. Some of these organizations want to participate in future experiments, but the plans for their involvement are not yet finalized.

If the research is successful, a TIA-like network will provide DoD with significantly enhanced capabilities to preempt foreign terrorists from carrying out attacks against deployed U.S. forces and the U.S. homeland.

THE UNDER SECRETARY OF DEFENSE

3010 DEFENSE PENTAGON
WASHINGTON, DC 20301-3010

ACQUISITION,
TECHNOLOGY
AND LOGISTICS

January 30, 2003

The Honorable Byron L. Dorgan
United States Senate
Washington, DC 20510-3405

Dear Senator Dorgan:

This is in reply to your recent letter to Secretary Rumsfeld regarding the Defense Advanced Research Projects Agency (DARPA) Total Information Awareness (TIA) project.

The DARPA TIA project researches the use of advanced information technologies to develop capabilities to detect foreign terrorist planning activities. If successful, TIA will provide the Department of Defense (DoD) with significantly enhanced capabilities to preempt foreign terrorists from carrying out attacks against deployed U.S. forces and the U.S. homeland.

The TIA project integrates other DARPA programs in three technology areas—advanced collaborative and decision support tools; language translation; and data search, pattern recognition, and privacy protection technologies—into an experimental prototype network. The TIA project is in the first year of a phased, multiyear effort.

DARPA performers will use two types of data for the research: publicly available data, such as foreign language newspapers, journals, and broadcasts; and synthetic data, i.e., artificial information generated to resemble real-world data in which the researchers embed simulated data indicative of terrorist activities.

As part of the development effort, the TIA project includes phased experiments involving operational units. For these experiments, the operational units will share and analyze information in their respective databases on suspected foreign terrorist activities. Units participating in the experiments will adhere to guidance for data sharing outlined in existing laws and policies. The TIA hardware and software tools under development to analyze the operational units' existing databases do not require or give access to any information to which the Intelligence Community does not already have lawful access.

TIA research complies with all U.S. laws, including the Privacy Act. I have directed that both an internal and external oversight board be created to monitor the TIA program to ensure no potential violations of any American's civil liberties occur.

The TIA project is new, and it will be a few years before we know whether the research will be successful. If the usefulness of the technologies is proven in a research environment, we will work with Congress to determine how best to implement the new capabilities and, of course, keep you informed as to its progress.

Sincerely,

E. C. Aldridge, Jr.

CHARLES E SCHUMER
NEW YORK

COMMITTEES
BANKING
JUDICIARY
RULES

OFFICE OF THE
SECRETARY OF DEFENSE

United States Senate

WASHINGTON, DC 20510

2002 DEC -2 PM 2:42

November 25, 2002

The Honorable Donald Rumsfeld
Department of Defense
1000 Defense, The Pentagon
Washington, DC 20301-1000

Secretary of Defense
SA0019474

Dear Secretary Rumsfeld

I am writing to express my concerns about the Department of Defense's incipient "Total Information Awareness" (TIA) program proposed by the Defense Advanced Research Projects Agency (DARPA)

Like you, I am unwavering in my commitment to use all technological means at our nation's disposal to both protect our homeland and prosecute the war on terror Unfortunately, I am deeply concerned about Admiral Poindexter's ability to lead this initiative, especially since abuse of the program could lead to unparalleled intrusion into the private lives of ordinary Americans

As you know, Admiral Poindexter's controversial past raises several legitimate concerns about whether he is the right person to lead an effort like TIA If we are to deploy a system like TIA that would allow for the intensive monitoring of electronic data of ordinary Americans, it must respect the right to privacy guaranteed by the Constitution and the laws of the land Given Admiral Poindexter's troubling history, I am skeptical of his regard for these fundamental principles

As it stands now, Congressional oversight of DARPA is the only way for the legislative branch to ensure that its activities do not transgress constitutional and legal boundaries Since Admiral Poindexter has shown himself willing to hide the truth from Congress, so long as he continues as Director of the Information Awareness Office and Project Manager for TIA, there is no way Congress can be assured that it is receiving accurate information about the project

I am also concerned that in allowing TIA's development to take place so rapidly, the Administration has once again ignored the open governing tradition of our nation Time and again, this Administration prefers to go it alone, rather than engage in serious and measured conversations with Congress regarding controversial proposals Each time this happens – whether it is with the USA-PATRIOT Act proposals, military tribunals, or the Terrorism Information and Prevention System (TIPS) – the Administration overreaches and forces Congress to jump in to pull the proposals back to more reasonable and appropriate terms It is a bad strategy We should be working together to unite the American people behind good laws that protect our security without unduly endangering out liberty Now the Administration appears set to bypass Congressional input on how the TIA should be implemented I encourage you in the strongest possible terms to come talk with Congress about how we can work together to create the best possible program

U19005 /02

DARPA is one of our nation's most valuable assets Its contribution to creating the Internet is just one example of the positive impact it has had on our way of life It would be a shame for Admiral Poindexter's involvement with TIA and its secretive development to compromise the confidence that Congress and the American people have in the organization as it seeks to protect America from terrorists

I strongly urge you to dismiss Admiral Poindexter immediately and to take a more consultative approach with Congress about the development of TIA

Sincerely,

Charles Schumer

Charles E Schumer
United States Senator

THE UNDER SECRETARY OF DEFENSE

3010 DEFENSE PENTAGON
WASHINGTON, DC 20301-3010

ACQUISITION,
TECHNOLOGY
AND LOGISTICS

January 30, 2003

The Honorable Charles E. Schumer
United States Senate
Washington, DC 20510

Dear Senator Schumer:

This is in reply to your recent letter to Secretary Rumsfeld regarding the Defense Advanced Research Projects Agency (DARPA) Total Information Awareness (TIA) project.

The DARPA TIA project researches the use of advanced information technologies to develop capabilities to detect foreign terrorist planning activities. If successful, TIA will provide the Department of Defense (DoD) with significantly enhanced capabilities to preempt foreign terrorists from carrying out attacks against deployed U.S. forces and the U.S. homeland.

The TIA project integrates other DARPA programs in three technology areas—advanced collaborative and decision support tools; language translation; and data search, pattern recognition, and privacy protection technologies—into an experimental prototype network. The TIA project is in the first year of a phased, multiyear effort.

DARPA performers will use two types of data for the research: publicly available data, such as foreign language newspapers, journals, and broadcasts; and synthetic data, i.e., artificial information generated to resemble real-world data in which the researchers embed simulated data indicative of terrorist activities.

As part of the development effort, the TIA project includes phased experiments involving operational units. For these experiments, the operational units will share and analyze information in their respective databases on suspected foreign terrorist activities. Units participating in the experiments will adhere to guidance for data sharing outlined in existing laws and policies. The TIA hardware and software tools under development to analyze the operational units' existing databases do not require or give access to any information to which the Intelligence Community does not already have lawful access.

TIA research complies with all U.S. laws, including the Privacy Act. I have directed that both an internal and external oversight board be created to monitor the TIA program to ensure no potential violations of any American's civil liberties occur.

The TIA project is new, and it will be a few years before we know whether the research will be successful. If the usefulness of the technologies is proven in a research environment, we will work with Congress to determine how best to implement the new capabilities and, of course, keep you informed as to its progress.

Sincerely,

E. C. Aldridge, Jr.

United States Senate
COMMITTEE ON ARMED SERVICES
WASHINGTON, DC 20510-6050

OFFICE OF THE SECRETARY OF DEFENSE
2002 NOV 27 PM 12:21

November 27, 2002

The Honorable Donald Rumsfeld
Secretary of Defense
The Pentagon
Washington, D.C.

Dear Mr. Secretary:

For the last several months, the Defense Advanced Research Projects Agency (DARPA) has been working on a "Total Information Awareness" (TIA) project, to develop and integrate information technologies that would enable the intelligence community to sift through multiple databases and sources to detect, classify and identify potential terrorist activities.

On November 20, 2002, Under Secretary of Defense for Acquisition, Technology and Logistics Pete Aldridge told the press that one of the objectives of the TIA program is the "discovery of connections between transactions – such as passports; visas; work permits; drivers' license; credit card; airline tickets; rental cars; gun purchases; chemical purchases – and events – such as arrest or suspicious activities and so forth." DARPA briefing materials provided to the Armed Services Committee staff indicate that the types of transactions to be covered may include communications, news services, financial, education, travel, medical, veterinary, transportation and housing transactions.

I am concerned that the application of the system contemplated by the DARPA research to transactions of the type described would not only raise serious privacy concerns, it might also be illegal and possibly unconstitutional. The Fourth Amendment to the United States Constitution protects individuals against unreasonable searches, and has been interpreted by the Supreme Court to cover wiretapping and electronic surveillance. In addition, numerous statutes safeguard the privacy of particular types of records from government intrusion. For example, the Right to Financial Privacy Act of 1978, 12 U.S.C. Section 3402, states that "no Government authority may have access to or obtain copies of, or the

information contained in the financial records of any customer from a financial institution" without authorization by the individual customer or the issuance of a lawful subpoena or warrant.

I would appreciate if you would review DARPA's TIA program to ensure that the goals and objectives of the program are consistent with the protection of the rights of American citizens. In particular, I would appreciate your views on the following questions:

- Would the application of the system contemplated by the TIA program to private communications, financial and other transactions be consistent with the requirements of the U.S. Constitution and applicable provisions of law and regulation?

- Is it your expectation that the Administration would seek changes to law or regulation before testing or implementing such a system?

Thank you for your attention to this important matter.

Sincerely,

Carl Levin

Carl Levin
Chairman

THE UNDER SECRETARY OF DEFENSE

3010 DEFENSE PENTAGON
WASHINGTON, DC 20301-3010

ACQUISITION,
TECHNOLOGY
AND LOGISTICS

January 30, 2003

The Honorable Carl Levin
Committee on Armed Services
Ranking Member
United States Senate
Washington, DC 20510-6050

Dear Senator Levin:

This is in reply to your recent letter to the Secretary Rumsfeld requesting a review the Defense Advanced Research Projects Agency (DARPA) Total Information Awareness (TIA) project to ensure that the program goals and objectives are consistent with the protection of the rights of American citizens.

The DARPA TIA project researches the use of advanced information technologies to develop capabilities to detect foreign terrorist planning activities. If successful, TIA will provide the Department of Defense (DoD) with significantly enhanced capabilities to preempt foreign terrorists from carrying out attacks against deployed U.S. forces and the U.S. homeland.

The TIA project integrates other DARPA programs in three technology areas—advanced collaborative and decision support tools; language translation; and data search, pattern recognition, and privacy protection technologies—into an experimental prototype network. The TIA project is in the first year of a phased, multiyear effort.

DARPA performers will use two types of data for the research: publicly available data, such as foreign language newspapers, journals, and broadcasts; and synthetic data, i.e., artificial information generated to resemble real-world data in which the researchers embed simulated data indicative of terrorist activities.

As part of the development effort, the TIA project includes phased experiments involving operational units. For these experiments, the operational units will share and analyze information in their respective databases on suspected foreign terrorist activities. Units participating in the experiments will adhere to guidance for data sharing outlined in existing laws and policies. The TIA hardware and software tools under development to analyze the operational units' existing databases do not require or give access to any information to which the Intelligence Community does not already have lawful access.

TIA research complies with all U.S. laws, including the Privacy Act. I have directed that both an internal and external oversight board be created to monitor the TIA program to ensure no potential violations of any American's civil liberties occur.

The TIA project is new, and it will be a few years before we know whether the research will be successful. If the usefulness of the technologies is proven in a research environment, we will work with Congress to determine how best to implement the new capabilities and, of course, keep you informed as to its progress.

Sincerely,

E. C. Aldridge, Jr.

2437 Rayburn House Office Building
Washington, D C 20515-1009
phone (202) 225-5211
fax (202) 225-8272
http //www house gov/deal

SECRETARY OF DEFENSE

2002 DEC -6 PM 3:56

Nathan Deal
Congress of the United States

Committee on
Energy and Commerce

Subcommittee on Commerce,
Trade, and Consumer Protection

Subcommittee on Health

Subcommittee on
Telecommunications
and the Internet

November 26, 2002

The Honorable Donald H Rumsfeld
Secretary
United States Department of Defense
1600 Army Pentagon
Room 3E880
Washington, D C 20310-1600

Secretary of Defense

SA0019669

Dear Secretary Rumsfeld

My office has recently been inundated with constituent correspondence regarding the Office of Information Awareness and the Defense Advanced Research Projects Agency (DARPA) I would like to request some further information on these programs

Like many others, I have seen the recent media reports regarding the Homeland Security Bill, H R 5710, which was recently passed by Congress and signed into law by the President Many of these reports discuss the government's use of Section 225 of the bill to "spy" on all electronic transfers on money, purchases, cell-phone conversations, e-mail and Internet use of all Americans It has been suggested that all of this information will be stored in a "supercomputer" run by the Department of Defense The correspondence I am receiving express fear of government spying with ordinary citizens being arrested and kept in an undisclosed location due to a computer suggesting they have made "threatening" purchases I am wanting to properly express to my constituents the role DARPA and the Office of Information Awareness will be playing in homeland security

Would you please provide me with an explanation of the Department of Defense's involvement with this issue, so I can share it with my constituents in order to ease their concerns Also, any other comments you might want to share in order to allay their fears would be appreciated

I believe new technology is a necessary component in the war in terrorism and only acts to increase cooperation between varying levels of law enforcement

Thank you for taking the time to address my concerns I look forward to your response

Respectfully,

Nathan Deal

Nathan Deal

Suite 520, Wachovia Center
340 Jesse Jewell Parkway
Mailing address P O Box 1015
Gainesville, Georgia 30503
phone (770) 535-2592
fax (770) 536-2766

Suite 108
415 East Walnut Avenue
Dalton, Georgia 30721
phone (706) 226-5320
fax (706) 278-0840

Suite 102
108 West LaFayette Square
LaFayette Georgia 30728
phone (706) 638-7042
fax (706) 638-7049

U19354 /02

THE UNDER SECRETARY OF DEFENSE

3010 DEFENSE PENTAGON
WASHINGTON, DC 20301-3010

ACQUISITION,
TECHNOLOGY
AND LOGISTICS

30 DEC 2002

The Honorable Nathan Deal
U.S. House of Representatives
Washington, DC 20515-1009

Dear Congressman Deal:

Your recent letter to the Secretary of Defense requesting information on the Defense Advanced Research Projects Agency (DARPA) and its Information Awareness Office (IAO) was forwarded to me for response. Enclosed is a short summary of DARPA's IAO activities; I hope you find the information useful.

Please assure your constituents that the Department of Defense (DoD) is *not* building a "supercomputer" to store data on American citizens. DARPA is developing an experimental prototype system comprising language translation, data search and pattern recognition, and collaborative reasoning and decision support tools to give the national security community powerful, new capabilities to uncover foreign terrorist acts against Americans. The only data used in this project will be information that is legally obtainable by the U.S. Government. The DoD has safeguards in place to ensure the privacy rights of American citizens is not violated.

The project is in its first year of a planned 5-year effort. If the technologies prove successful, the DoD will work with Congress to determine how best to implement the new capabilities.

Sincerely,

E. C. Aldridge, Jr.

Enclosure:
As stated

Defense Advanced Research Projects Agency's
Information Awareness Office
and
Total Information Awareness Project

The Defense Advanced Research Projects Agency's (DARPA) mission is to research and demonstrate innovative technologies to solve national-level problems, such as the grave terrorist threat which our nation faces. DARPA created the Information Awareness Office (IAO) in response to September 11, 2001, to research, develop, and demonstrate innovative information technologies to detect terrorist groups planning attacks against American citizens, anywhere in the world.

IAO is *not* building a "supercomputer" to snoop into the private lives or track the everyday activities of American citizens, contrary to some recent media reports. Instead, IAO is developing an experimental prototype system that largely consists of three parts—language translation technologies, data search and pattern recognition technologies, and advanced collaborative and decision support tools. Together, these three parts effectively comprise the Total Information Awareness (TIA) project.

The language translation technologies will enable the rapid translation of foreign language speech and text and give analysts from intelligence, counterintelligence, and law enforcement agencies the ability to quickly search for clues about emerging terrorist acts. The intelligence community believes it can find evidence of terrorist activities in open source foreign language publications. Rapid translation technologies will help analysts search a significant amount of material in a much shorter period than is possible today.

The research into data search and pattern recognition technologies is based on the idea that terrorist planning activities or a likely terrorist attack could be uncovered by searching for patterns indicative of terrorist activities in vast quantities of data. Terrorists must engage in certain transactions to coordinate and conduct attacks against Americans, and these transactions leave signatures form patterns that may be detectable. For this research, the TIA project will use only data that is legally available and obtainable by the U.S. Government.

A major challenge to terrorist detection today is the inability to quickly search, correlate and share data from databases maintained legally by our intelligence, counterintelligence, and law enforcement agencies. The collaborative reasoning and decision-support technologies will help solve existing coordination problems by enabling analysts from one agency to effectively collaborate with analysts in other agencies.

Today, TIA exists only as a vision. The project is in its first year of an anticipated 5-year research effort. During the first 36 months, a range of ideas will be developed via limited demonstrations and preliminary prototypes. During the final 24 months, the most promising research avenues will be extended to support the transition of a scalable leave-behind system prototype. If the project is successful, the national security community and the Department of Homeland Security will consult with Congress to determine whether the TIA system should be implemented for domestic use. The DoD will consult with Congress on how to best implement the system for protection of U.S. forces overseas.

The DoD recognizes American citizens' concerns about privacy invasions. The Department has safeguards in place to ensure the TIA project will *not* violate the privacy of American citizens. In addition, IAO will research and develop privacy protection and other technologies to protect the system from internal abuses and external threats. Their goal is development of privacy technologies and processes that ensure data is protected and used only for lawful purposes.

Some individuals have questioned the role of the DoD and DARPA in this area. In its 54-year history, DARPA has undertaken numerous high-risk research efforts that led to significant capabilities. Many existing information technologies—including the Internet—started as advanced DARPA research projects. DARPA has had in the past joint programs with the FBI and the US Customs developing technology which could be used for detecting explosives and drugs at Airports and Sea Ports. IAO follows a similar path of technical innovation with its research into advanced information capabilities that will give the United States a decisive edge in the global war on terrorism. All Americans share the frustration associated with vague warnings of terrorist threats. It is believed that IAO and its TIA project will help the U.S. Government reduce those generic reports to advance notice of specific threatening acts.

1.A.

TITLE PAGE

Briefing Materials concerning TIA BIOALERT 1A

FOLDER 1.a.1.a1.-1.a.8 TOC 3

KICKOFF BRIEFS

- 1.a.1 Bio-Surveilliance System Bio Treat Overview
  - Potomac Institute for Policy Studies
- 1.a.2 Kickoff Meeting 11-12 Oct 01
- 1.a.3 Electronic Surveillance System (ESSENCE II) Status Report
  - Johns Hopkins Applied Physics Lab
- 1.a.4 ESSENCE II Progress at DoD-GEIS
  - WRAIK
- 1.a.5 Bioterrorism Surveillance in DC Metro Area
  - George Washington University School of Public Health + Health Services
- 1.a.6 Bio-Surveillance Using Sufficiently Anonymous Data
  - CMU Lab for International Data Privacy

- 1.a.7 ESSENCE II Architecture
  - Johns Hopkins Applied Physics Lab
- 1.a.8 ESSENCE II Classification Methodology
  - CYCORP

1001

Bio-Surveillance System
Bio Threat Overview

David Siegrist
Potomac Institute for Policy Studies
October 11, 2001

1

Afghan Disease Threats

Paula Bronstein / Getty Images file

Crimean Congo Hemorrhagic Fever Victim, MSNBC

- **Returning Service Members, Aid Workers and Refugees May be Ill**
  - Current Outbreak of Crimean Congo Hemorrhagic Fever
- **Virtually No Disease Surveillance in Afghanistan in the Past: Unknowns**
- **Soviets Saw Typhoid, Infectious Hepatitis, Malaria, Cholera, & Plague**
  - Pakistan Has Rift Valley Fever

2

Afghan Anthrax Threat

- **Afghanistan Reported to Have Anthrax Vaccine Facility (10k Dose/Yr)**
  - Vaccine Plants Are Generally Extremely Useful in Producing Bioagents
- **Across Uzbekistan at Resurrection Island is a Soviet Bio Dump**
  - Virtually Unguarded. Vats of Advanced Anthrax Spores the Soviets Failed to Kill.
    - Alibek: 3X Potency of Sverdlovsk Bacteria
  - May be Available to Afghans to Use as Seed Stock for Vaccine Plant
- **Iraqi Source?**
  - Iraq Intelligence Agent Reported to Have Met M. Atta in Prague in 2000.
  - Iraq Had 4 Strains from US ATCC, Including 11966 from Fort Dietrick
  - Iraqi Powder Anthrax Never Discovered. However, B. thurengiensis at Al Hakkam.

3

Florida Anthrax Case

- **First Lethal Powder Successfully Delivered to Victim. Watershed.**
  - Hundreds of Threats Over Years. Powder Thought to Require State Sponsor.
  - May Have Come in Letter. Mail Worker and Keyboard Samples Test Positive.
- **Strain Appears to Be Engineered, But Not Antibiotic Resistant**
  - Possible Origin at Iowa State University, Ames from 1950's. Haiti or Texas Source?
- **Thinking Now Is More about Bio-Crime, vs Bio-Terrorism**
  - However, State Department Has Advised All Embassies Overseas to Stock Antibiotics
- **Anthrax Bioterrorism with Refined Powder Has Now Been Done.**
- **Possibility of Mass Casualty Application?**
  - **Less So, if Bio-Crime**
  - **If Bio-Terror, Possibility of Early Test Prior to Wider Release**

4

Smallpox

- **Smallpox Has Killed Millions Over Years. Officially Eliminated.**
  - Two Legal Repositories in US and Russia. Iraq and North Korea May Have Also.
    - Russia Used to keep 20 Tons on Hand for Strategic War. Storage Site Never Inspected.
    - Iraq: Vaccinated Troops and Had Spray Drier. Hussein Believes Saddam Had It. Camel Pox.
- **Smallpox Patients Contagious When Have Rashes, Scabs.**
  - High Fever May Prostrate Victim While Contagious. Lesions on Face, Arms.
- **If Smallpox Outbreak, Ring Vaccination Strategy**
  - 7-15 Million Doses of Vaccine Will Not Support Mass Vaccination. Create "Firebreaks"
    - Four Cardboard Boxes in walk in Freezer. Half Failed Lot Inspection but May Work.
    - "The VIG Turned Pink" Current Vaccinia Immunoglobulin May Be Ineffective. New IND.
    - First Lots of 40 Million Doses of New vaccine Due in a Year or So.
- **Ring Vaccination Strategy Relies on Early Detection!**
  - Few Early Victims May have Little Apparent Relation to One Another.

5

Authoritative Views?

- **Leading Offensive Biologists in World are Ken Alibek and Bill Patrick**
- **Bill Patrick Thinks Biggest Threat is Marburg (VHF) to a Building**
  - Thinks Meteorological Vagaries Forestall "Crop Duster" Mass Attack
- **Ken Alibek Believes Marburg Could Be Used Against Cities**
  - Also Concerned About "Chimera" Virus. Genetically Load Marburg into Smallpox.
  - Super weapon Combines Hardiness and Transmissibility of SP with Lethality of Marburg.
  - No Credible Reports This Was Achieved.
- **Special Warning Signs: Anaphylactic Shock or Down Regulated Immune Response**
  - May Be Signs of New Generation Weapons that Manipulate /Crash Immune System
- **Other Advanced Bioweapons: Bioregulators**
  - "Super Toxins, Substance P, Super Endorphins, etc.
- **Bioweapons Unlikely to Behave as Natural Forms of Disease Might Be Expected to.**

6

Backups

Bio Threats

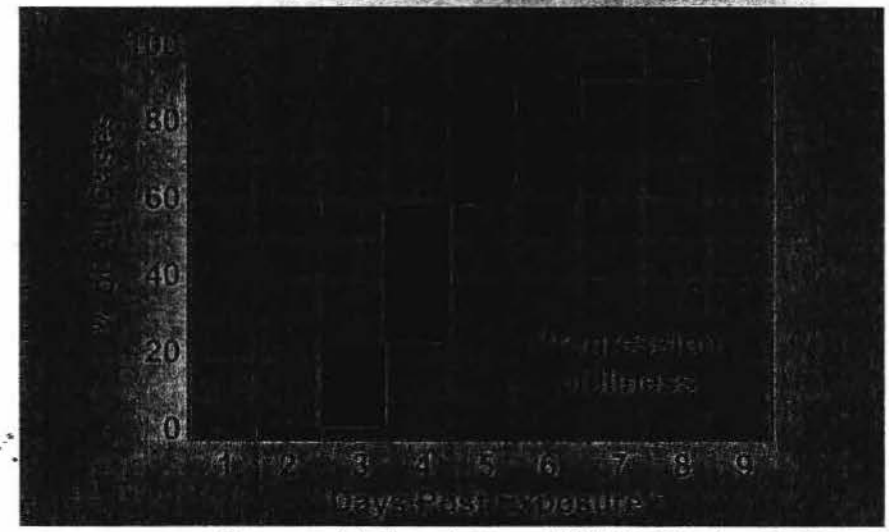
- Bio-Weapon Deployment
  - Creating an aerosol cloud
  - Releasing infected vectors, e.g. Rodents, mosquitoes or fleas
  - Contaminating food or water supplies
- Key Agents—Potential to cause large numbers of deaths and to generate panic
  - Anthrax
  - Plague
  - Botulinum toxin
  - Smallpox
  - Tularemia

Example Prodromal Syndromes

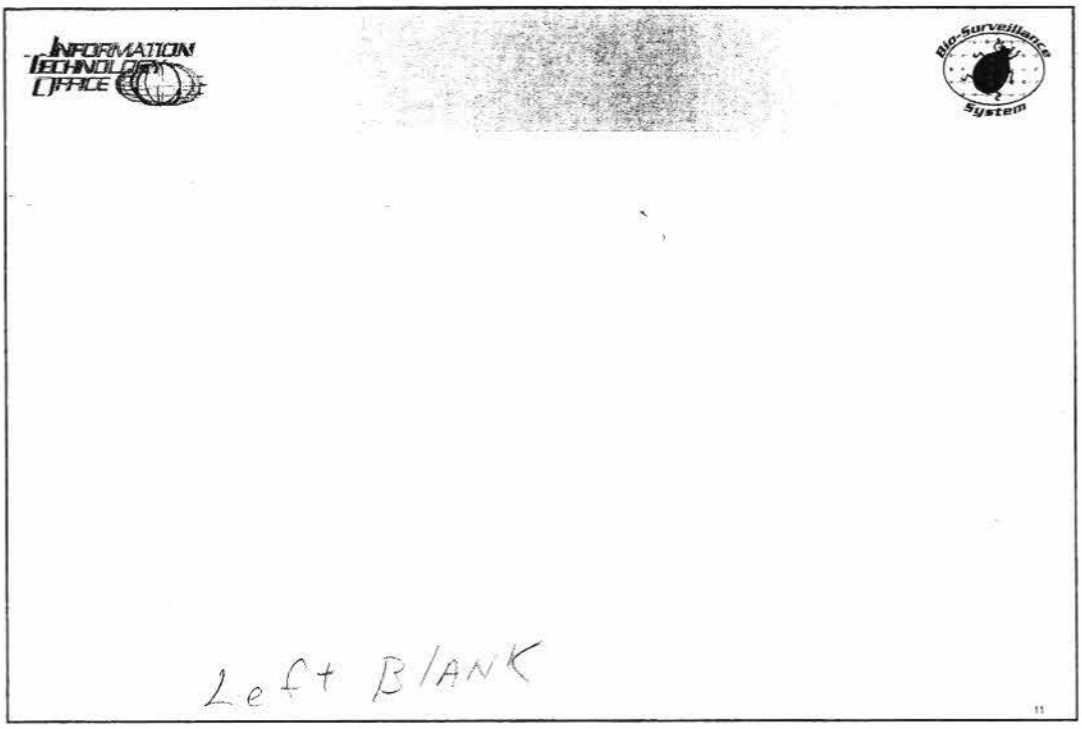
	Anthrax	Brucellosis	Glander	Plague	Tularemia	Ebola	Marburg	Venezuelan Equine Encephalitis	Q fever	Smallpox
Cardiovascular		Murmurs								
Headache		X	X	X	X	X	X	X	X	X
Other Pain		Myalgia	Myalgia	Abdomina	Myalgia	Chest	Back	Myalgia	Back	Myalgia
Drowsiness/Mal	50%	X	X	X	X	X	X	X	X	
Nausea						X	X		Few	
Fever	50%	Chills	X	X	X	X	X		X	X
Skin/Rash	Papules	Pallor		Pustules	Inflam'n					
Mucous Membra	Papules								Hyperemia	Hyperemia
Gastrointestinal					X	X	X		Some	X
Diarrhea						X	X			
Lymph Nodes	X	X	X	X	X		X			
Respiratory	Cough				Cough	Pharynx			Cough	

Percent of Cases by Day, Anthrax

Intervention by Day 4 May Reduce Casualties 50%



Source: National Guard WMD Study, 2000


INFORMATION TECHNOLOGY OFFICE
Bio-Surveillance System
Left BLANK

Bio-Surveillance Program

Kickoff Meeting
11-12 October 2001
Arlington, VA

Welcome and Introductions

Ted Senator
DARPA/ISO
(703) 696-2231
tsenator@darpa.mil

Meeting Objectives

- Ensure Common Understanding of Program Goals, Objectives, Constraints, etc., among all participants and interested parties
- Share Information Between and Among Program Participants about Technical Approach, Program Plans, Validation Strategy, Data Sources, etc.
- Identify Opportunities for Collaboration Between Program Participants
- Provide Information to be used at DARPA Internal Program Review

Agenda: Thu 11 October 2001

0730-0815	Continental Breakfast	
0815-0830	Welcome and Introductions	Burke/Senator
0830-0915	Bioterrorism Threat	Dave Siegrist
0915-0945	BioSurveillance Program Overview	Ted Senator
0945-1015	Break	
1015-1215	JHU/APL Briefing	Joe Lombardo
1215-1315	Lunch	
1315-1345	JHU/APL Discussion	
1345-1515	IBM Presentation	Chung-Sheng Li
1515-1545	Break	
1545-1615	IBM Discussion	
1615-1645	Program Validation Strategy	Dave Siegrist
1645-1700	Questions and Action Items	Ted Senator

3

Agenda-R1: Friday 12 October 2001

0730-0800	Continental Breakfast	
0800-1000	Veridian Team Approach	Gene McLellan
1000-1030	Break	
1030-1100	Veridian Approach Discussion	Ray Liuzzi
1100-1200	Mellon-Pitt Approach	Mike Wagner
1200-1230	Lunch	
1230-1330	Program Validation Discussion	Dave Siegrist
1330-1345	Mellon-Pitt Approach (cont)	Mike Wagner
1345-1430	Discussion	Jim Reilly
1430-1500	Break	
1500-1530	Data Sharing Discussion	Rey Wong
1530-1600	Wrap-Up and Action Items	Ted Senator
1600-1630	PM Feedback to PI's	Ted Senator + 4

4

Agenda-R2: Friday 12 October 2001

0730-0800	Continental Breakfast	
0815-1030	Veridian Team Approach	Gene McLellan
1030-1045	Break	
1045-1100	Veridian Approach (cont)	
1100-1115	Veridian Discussion	
1115-1200	Mellon-Pitt Approach	Mike Wagner
1200-1230	Lunch	
1230-1330	Program Validation Discussion	Dave Siegrist
1330-1415	Mellon-Pitt Approach (cont)	Mike Wagner
1415-1450	Discussion	
1445-1500	Break	
1500-1530	Data Sharing Discussion	Rey Wong
1530-1600	Wrap-Up and Action Items	Ted Senator
1600-1630	PM Feedback to PI's	Ted Senator + 4

5

Why a New Program Manager ?

- **Murray Burke Took Over DAML**
- **New BioS Program Manager is Mr. Ted Senator**
  - Heavily Involved in Bio-Surveillance Program Selection Process
  - Built Advanced Detection Systems for Money Laundering, Stock Market Violations
    - For Treasury/FinCEN, NASD Regulation
    - DARPA Agent for Navy Strategic Computing Application in mid-1980's
  - Also manage Evidence Extraction and Link Discovery Program for DARPA ISO
- **No Changes Planned to Program Goals, Objectives, etc.**

6

Program Motivation

- **Bioterrorism Becoming Much More Visible Threat**
- **Bio-Surveillance Can Make Critical Difference in Mitigating Attack**
- **This Bio-Surveillance Program is Nationally Important**
  - Already Briefed Favorably to Vice President, SecDef, & DIRO. High Profile.
  - One Prototype Already Up and Running in National Capital Area
  - Used at Inauguration
- **Contract Performers Here are an Elite Group**
  - Competition Was Fierce. Lots of Good Ideas and Capabilities. Only able to fund A+ efforts.
- **DARPA and Others Are Anticipating Great Results from the Program**

Program Considerations (1)

- **Program Emphasis: Cooperative/ Competitive Effort to Build Most Useful Prototype Operational Capability (vs Emphasis on Technology Development)**
  - Best Data Sources Should Be Identified and Shared, for Instance
  - Disease Models to Be Shared As Appropriate. No Wheel Reinvention. Others Areas for Cooperation?
- **There May Be a Programmatic Down-Select, Depending on Funds, in 2003. However, Currently Biosurveillance Appears to Be a Growth Area**
- **Cultivate Potential Transition Partners**
  - May Include LEADERS, JFCOM, DoD Disease Surveillance System, Navy Balboa, etc.
  - Transition Partners, especially those with $, Help Ensure Continuation of Efforts

Program Considerations (2)

- **This Is a DoD Program Emphasizing Force Protection. However, There May Be Civilian Benefits Also**
  - Recent GAO Report: DARPA is not in the civilian consequence management business
- **Looking for a Two-tiered Alerting System: Anomaly Detection Plus Disease Identification. Requires Baselining and Normalization.**
  - Program Stretch Goal Is Detection One Day Post Pathogen Release.
  - Very Interested in Sentinel Animals for Early Detection.
  - Privacy Protection Extremely Important
  - Low False Positives, But **No** False Negatives
- **Clearly Defined Metrics and Experiments**
  - Allow us to know if We're Making Progress
  - Support Go/NoGo Decisions, every 12-18 months

9

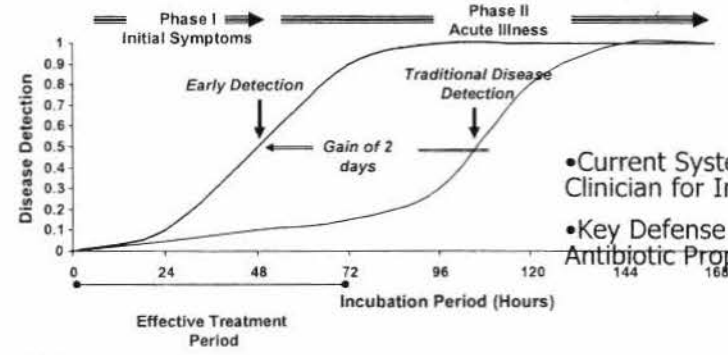
Program Briefing

Outline

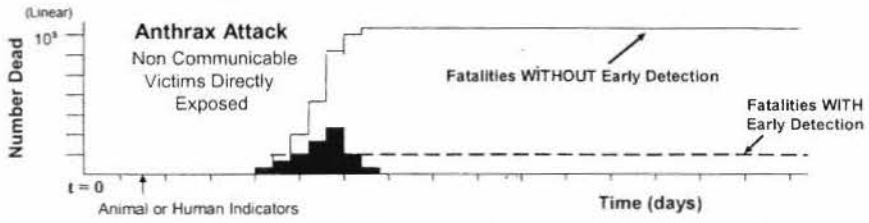
- Objective
- Military Utility
- How
- Goal
- Metrics
- Program Management
- Who
- Transition
- Funding

10

Objective: Early Detection



- Current System Relies on Alert Clinician for Initial Detection
- Key Defense Factor is Starting Antibiotic Prophylaxis Early



11

Objective

- **Detect Pathogen Release at Least Two Days Earlier than Current Practice (stretch goal is 24 hours after attack)**
  - Disease dependent
  - Current practice relies on alert clinicians who
    - Have seen few or no cases
    - Look for common diagnoses/diseases
  - Discriminate Between Intentional Pathogen Release and Natural Outbreak of Disease (e.g., Flu)
  - Continuous and Event-Oriented Monitoring
- **Identify Type of Attack (e.g., Anthrax, Smallpox, Plague, Botox)**
- **While Protecting Privacy of Individual Source Information, including medical records**

12

Military Utility

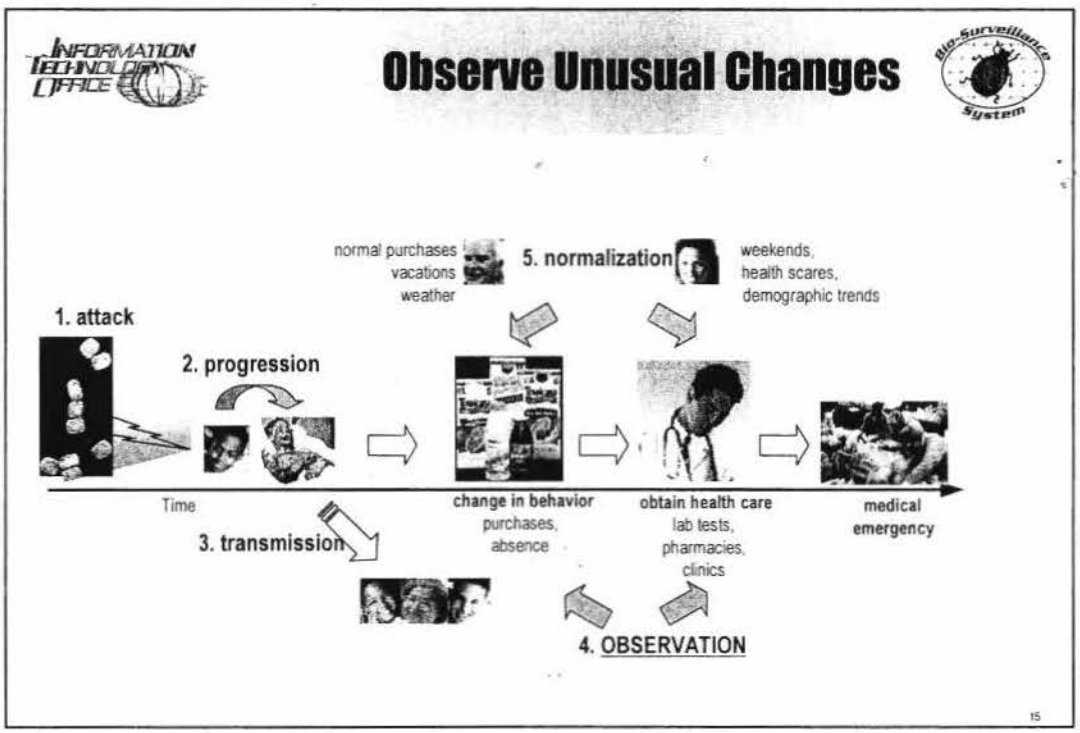
- Force Protection
- Early Detection and Effective Intervention (Treatment) Can Reduce Casualties by 50%
- Provides First-Strike Survivability
- Protects Deployment Capability
- In CONUS, military integrated with civilian population
  - Military families use civilian health care and can infect members
- Military Personnel Return from Deployment as (Unknown) Disease Carriers → Threat to US
  - Are their families, acquaintances vaccinated ?

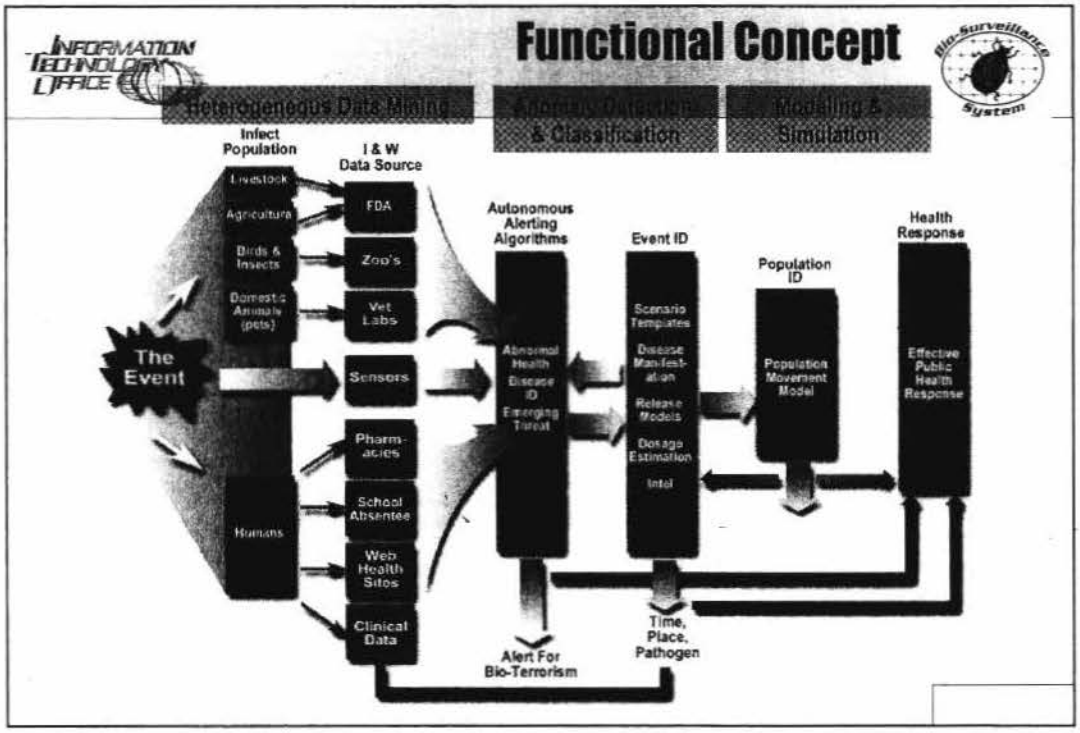
13

How

- Use Unconventional Data Sources as Early Sentinels
  - Data Mining to Identify Value of Sources
- Normalize for "Natural" Disease Outbreaks
- Two Stage Detection
  - Autonomous Alerting of Epidemiologist
  - More Extensive Decision-Aids for Alerting
- Privacy Protection Algorithms

14


Observe Unusual Changes
normal purchases
vacations
weather
5. normalization
weekends,
health scares,
demographic trends
1. attack
2. progression
Time
change in behavior
purchases,
absence
obtain health care
lab tests,
pharmacies,
clinics
medical
emergency
3. transmission
4. OBSERVATION
15


Functional Concept
Heterogeneous Data Mining
Anomaly Detection & Classification
Modeling & Simulation
Infect Population
I & W Data Source
Livestock
Agriculture
Birds & Insects
Domestic Animals (pets)
Humans
The Event
FDA
Zoo's
Vet Labs
Sensors
Pharm-acies
School Absentee
Web Health Sites
Clinical Data
Autonomous Alerting Algorithms
Abnormal Health Disease ID
Emerging Threat
Alert For Bio-Terrorism
Event ID
Scenario Templates
Disease Man-fest-ation
Release Models
Dosage Estimation
Intel
Time, Place, Pathogen
Population ID
Population Movement Model
Health Response
Effective Public Health Response

Goal: What DoD Has at the End

- Two Prototype BioSurvellance Systems in DoD Cities (National Capital Area and Norfolk)
  - Short-term effort for National Capital Area
  - Possible Expansion to Other Cities, dependent on $ and data availability
    - San Diego
    - Omaha
    - Albuquerque
- Knowledge of Effective Data Sentinels (i.e., what data are useful/effective indicators/predictors)
- Inference Algorithms for Detection of Potential Bio-Attacks
- Privacy Protection Algorithms
- BioWar Simulation

17

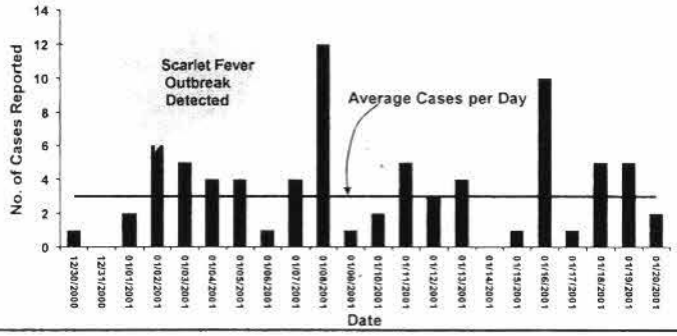
Metrics

- Historical Data Comparison for Natural Outbreaks – Ground Truth
  - Independent examples from developers
  - Same cities / different cities
- Animal Tests
- ROC Curves → Cost-Sensitive ROC Curves, Lift Curves, etc.
  - How to measure ?
- Specificity vs Sensitivity
- Expert Panel Review and Validation
- Simulations
- Geographic Clustering (e.g., Building)
- Red Team

18

Inaugural Experiment

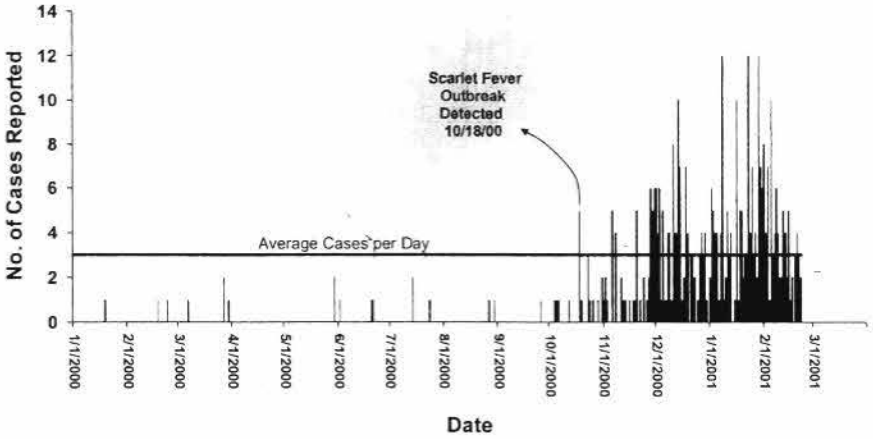
- Volunteered to Armed Forces Inaugural Committee
- Gathered historical data
  - Walter Reed – military treatment and TRICARE data
  - Over-the-counter pharmaceutical data
  - Johns Hopkins – HMO data
- No bio-terrorist attack patterns
  - Anthrax, tularemia, smallpox and botulism models
- Noted Scarlet Fever outbreak
- Contacted health services – they were not aware of outbreak as late as January 16, 2001



19

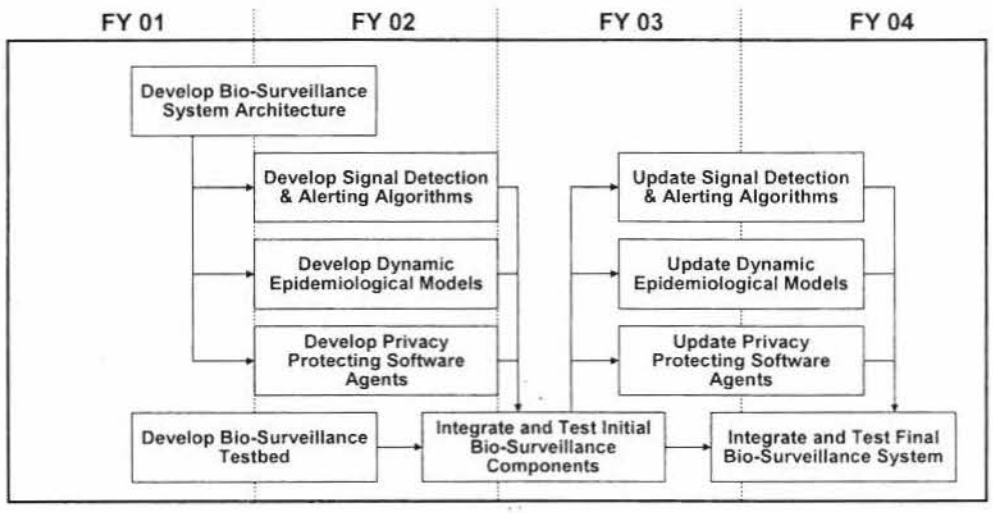
Scarlet Fever Epi Curve

Retrospectively, had we been running the system earlier, we would have detected the outbreak on October 18, 2000



20

Bio-Surveillance Program Plan



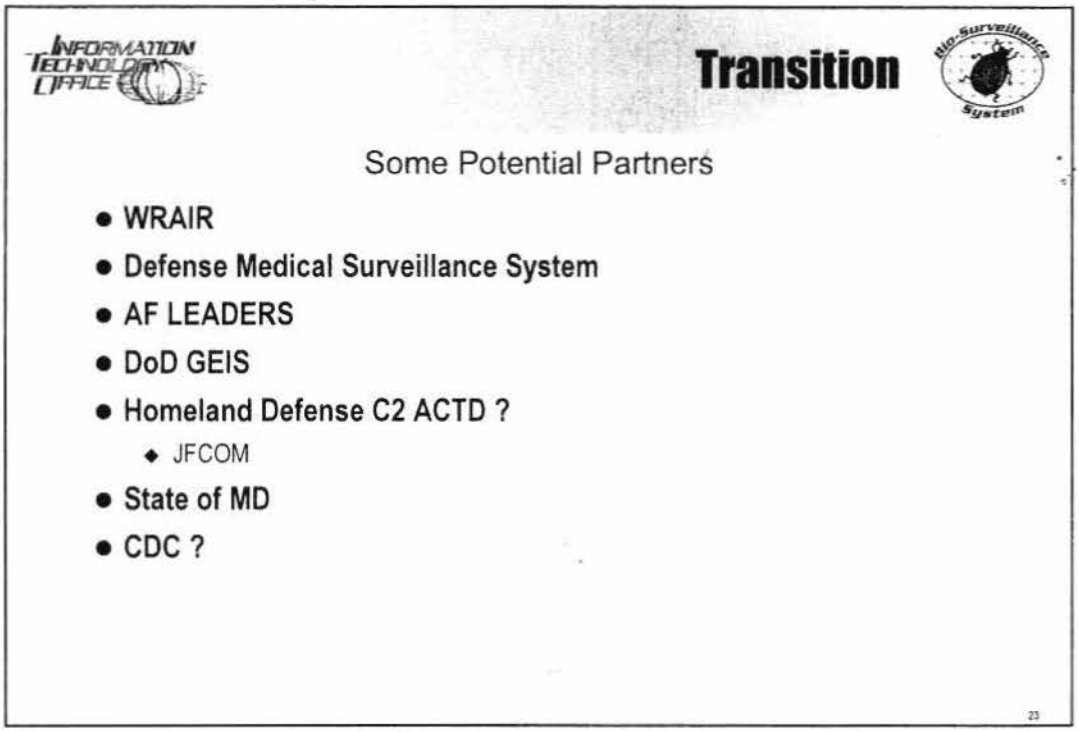
21

Who

Two Prototype Developers, Two Technology Developers, and an Independent Validator

- Johns Hopkins University / Applied Physics Laboratory
- Veridian Systems
- IBM
- Mellon-Pitt
- Potomac Institute for Policy Studies

22


Transition
Some Potential Partners
WRAIR
Defense Medical Surveillance System
AF LEADERS
DoD GEIS
Homeland Defense C2 ACTD ?
JFCOM
State of MD
CDC ?
23

CYCORP

The Johns Hopkins University

DoD GEIS Global Emerging Infections System

The George Washington University

IBM

Electronic Surveillance System for the Early Notification of Community-Based Epidemics (ESSENCE II) Status Report

Presented at the

Bio-Surveillance System Kickoff Meeting

11-12 October 2001

Joe Lombardo
The Johns Hopkins University
Applied Physics Laboratory

ESSENCE II Overview

- The ESSENCE II Project will develop an advanced outbreak detection and investigation system for early notification of abnormal health conditions caused by a covert release of a pathogen in military / civilian populations within CONUS.

- ESSENCE II will exploit non-traditional indicators of health for both the human / animal populations and apply autonomous outbreak detection algorithms to perform surveillance 24/7.

- Advanced techniques for patient anonymity, knowledge discovery, detection, classification, and system integration are being assembled by the ESSENCE II team.

Goals / Challenges

The primary goal of the ESSENCE II Project is to assemble a operational prototype of an advanced outbreak detection and investigation system for the National Capital Area at the end of FY03.

ESSENCE II will incorporate / develop:

a. **HIPAA approved anonymity algorithms to protect patient privacy**
b. **Advanced data mining techniques**
c. **Autonomous adaptive detection algorithms applied across all data sources**
d. **Rapid man-in-the-loop epidemiological investigations using rule based and probabilistic techniques**
e. **An architecture that employs a common ontology across all system components**

In FY04 the ESSENCE II system will be transferred to the Department of Defense Global Emerging Infectious Disease System for evaluation and testing.

ESSENCE II will be evaluated on its ability to:

a. **Identify naturally occurring outbreaks in the NCA area**
b. **Identify synthesized bio-events**
c. **Interface with civilian disease reporting systems**
d. **Form a collaborative environment to support a multi jurisdictional investigation**

Technical Approach

Functional Description of the ESSENCE II System

Classified sources

- Intelligence
- Other Surveillance Activities

Sensitive sources

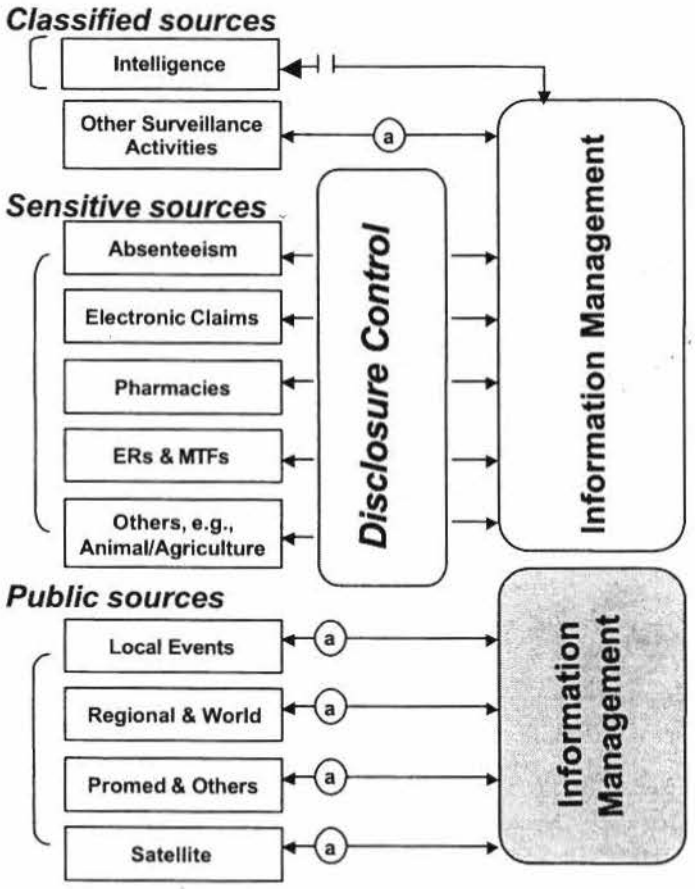
- Absenteeism
- Electronic Claims
- Pharmacies
- ERs & MTFs
- Others, e.g., Animal/Agriculture

Public sources

- Local Events
- Regional & World
- Promed & Others
- Satellite

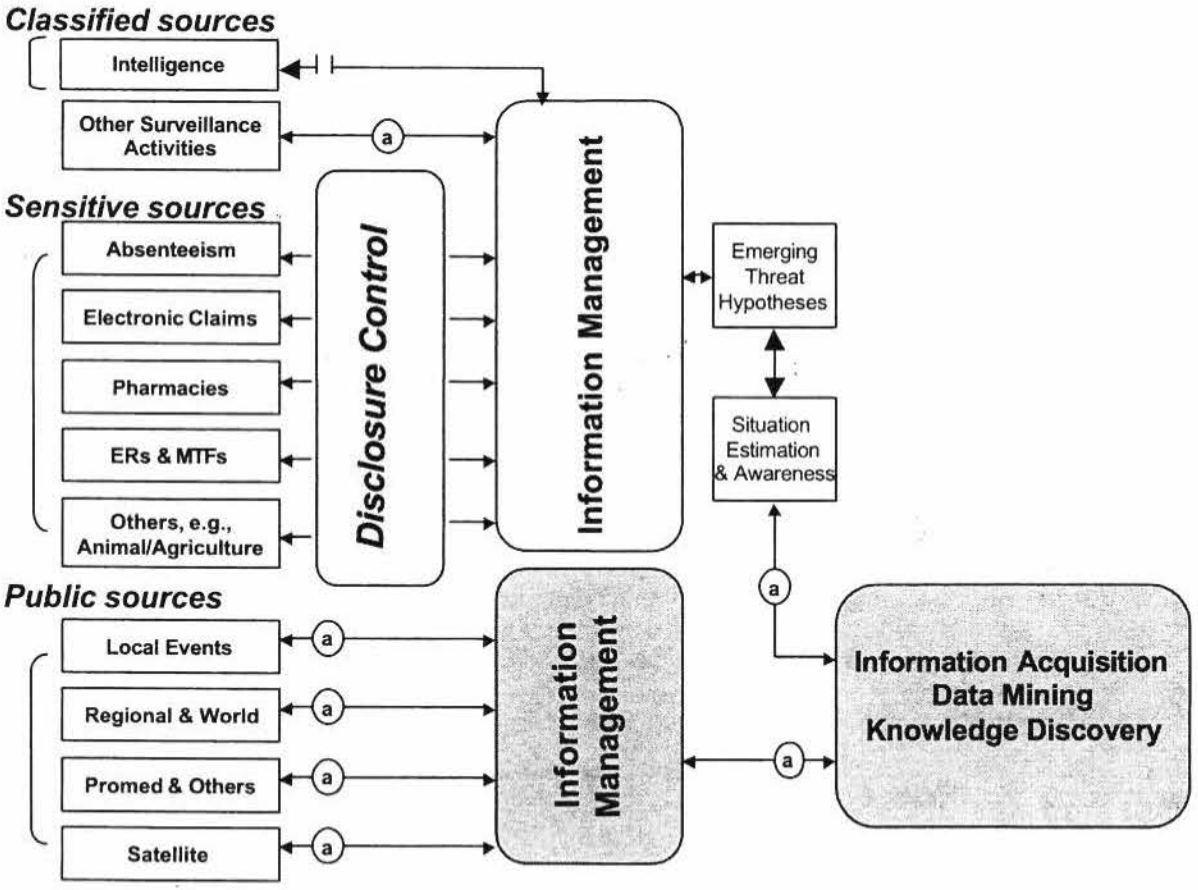
Technical Approach

Functional Description of the ESSENCE II System



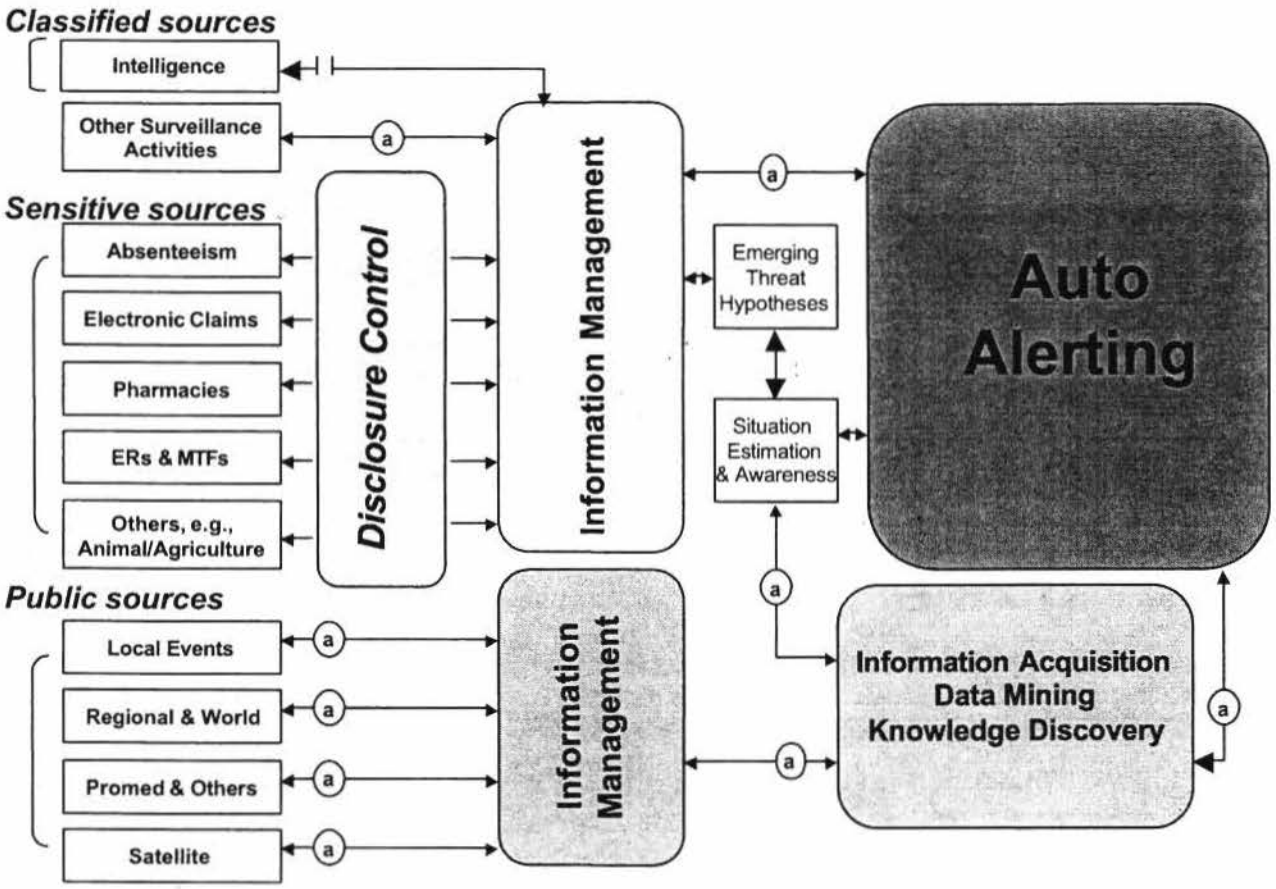
Technical Approach

Functional Description of the ESSENCE II System



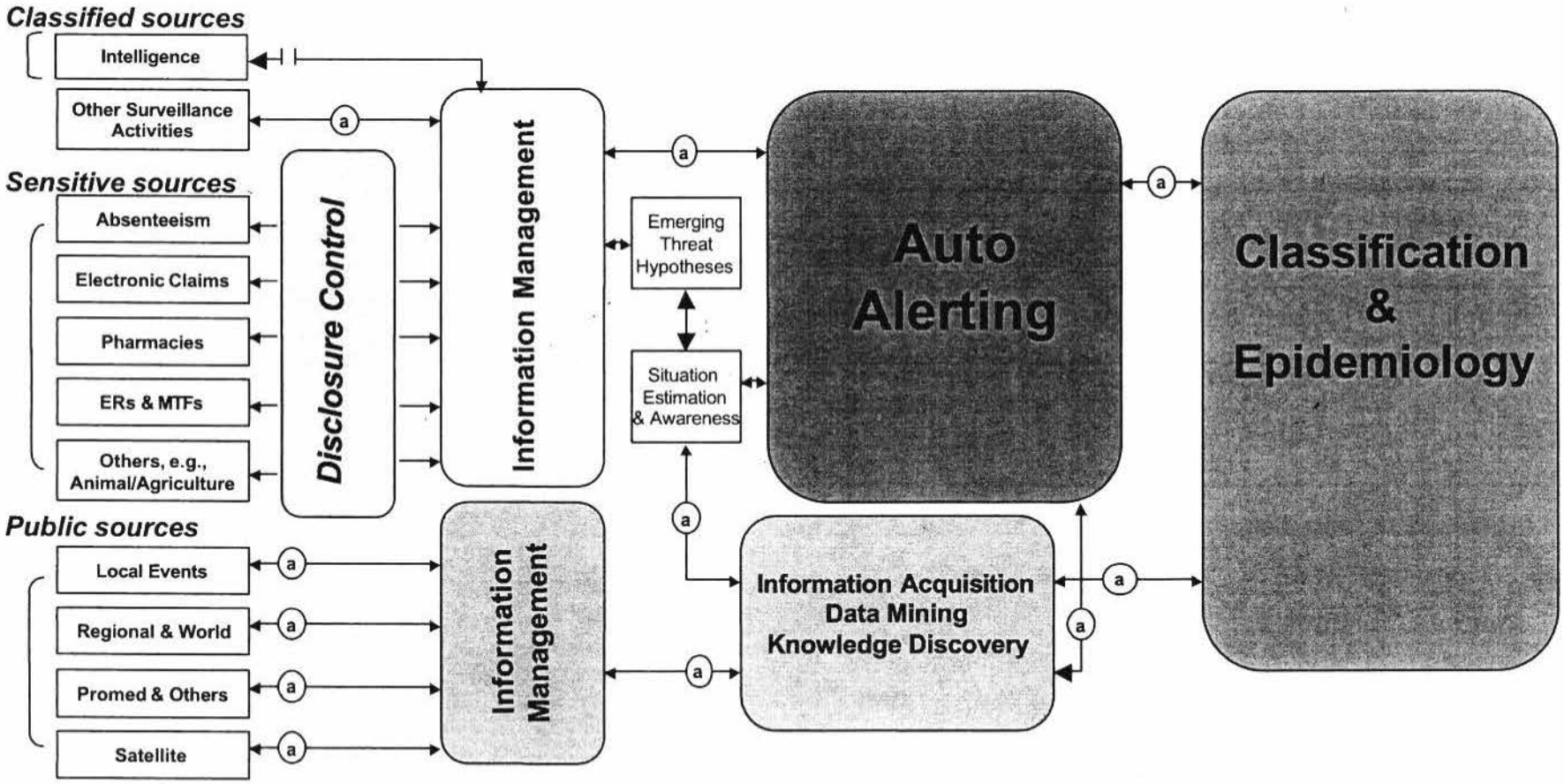
Technical Approach

Functional Description of the ESSENCE II System



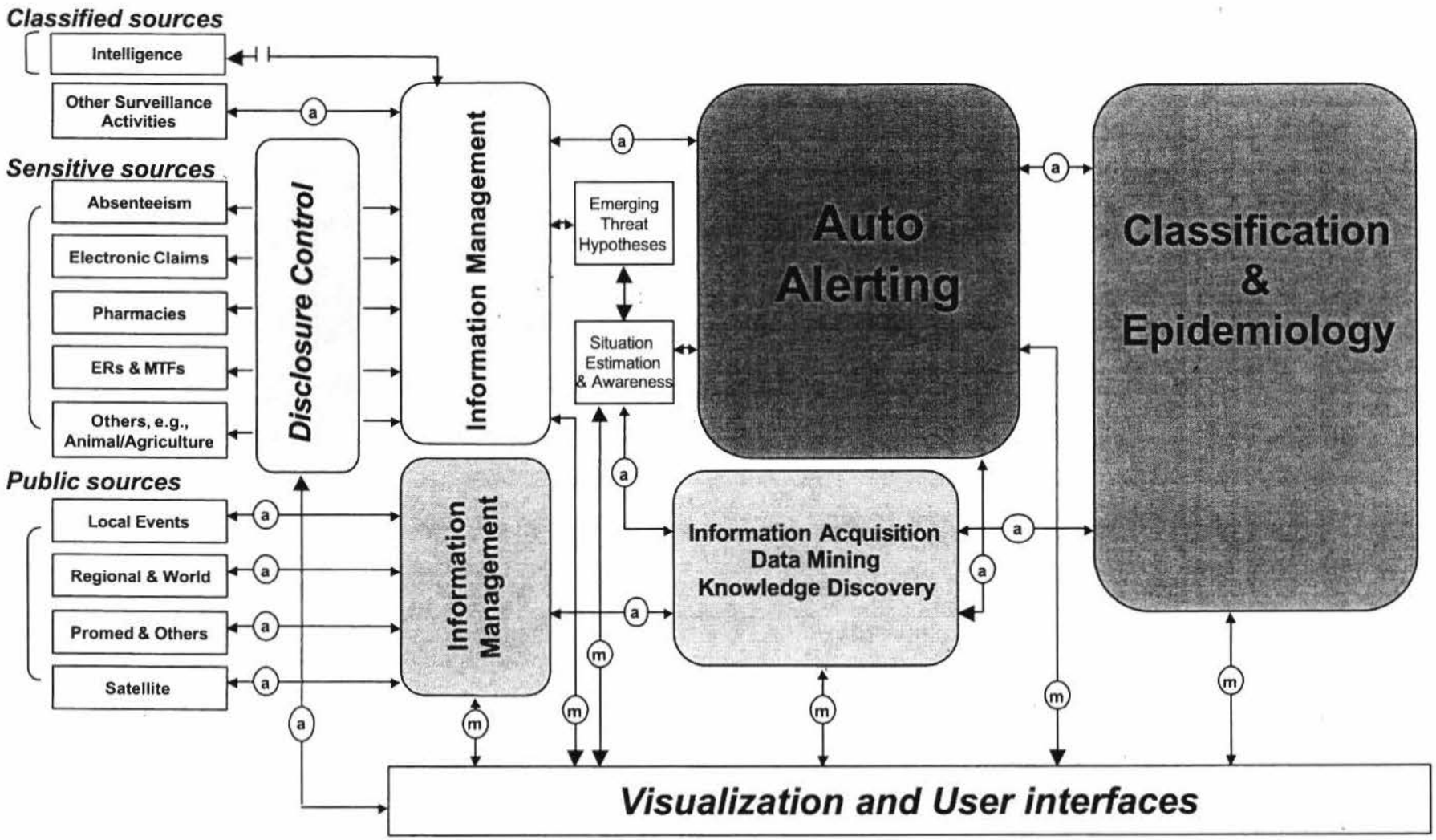
Technical Approach

Functional Description of the ESSENCE II System



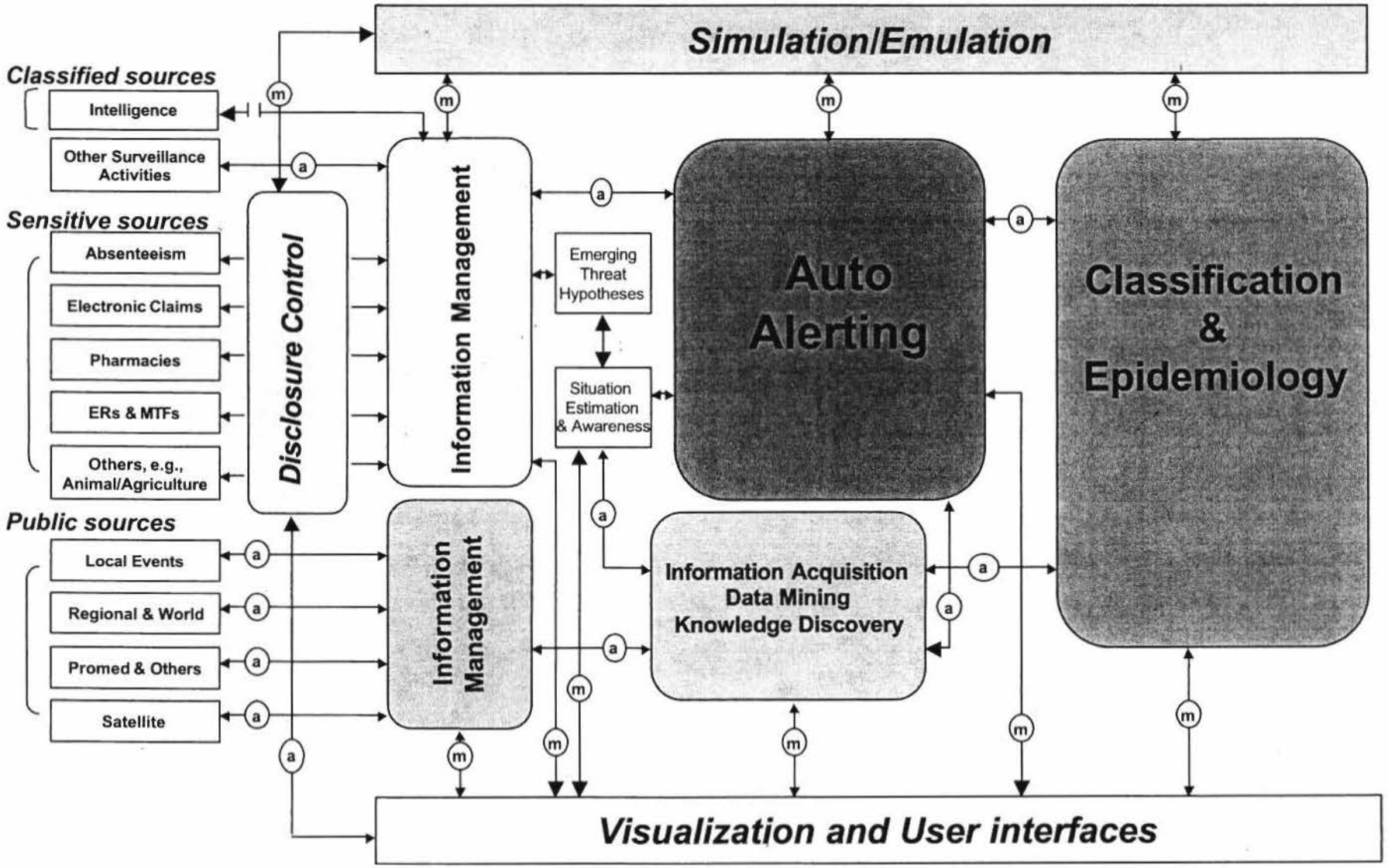
Technical Approach

Functional Description of the ESSENCE II System



Technical Approach

Functional Description of the ESSENCE II System



Preliminary Data Sources for Outbreak Detection

Sensors	Human Behaviors	HealthCare	Animal
Environmental Air	Absenteeism	Electronic Claims HMOs, BC&BS, Medicare/Medicaid	Zoos and Parks
Environmental Water	OTC Medications	Requests for lab tests	Agriculture and Animal Labs
Agriculture & Food Production	911 Calls	Sentinel Physicians Reports	Dept of Natural Resources
Gene Chips	Dial-a-Nurse calls	Emergency Medical Services	Animal Control
	Poison Center calls	Emergency Rooms Visits	National Veterinary Assistance Centers
	Health Advice Websites	Medical Examiners Reports	Animal Advice Websites

Edit View Go Communicator Help

January 2001						
Sun	Mon	Tue	Wed	Thu	Fri	Sat
	01	02	03	04	05	06
07	08	09	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30	31	February		

Prev Day | Next Day

Region Maps

- Region Status
- Region Legend

Detailed Maps and Charts

ZipCode | Region

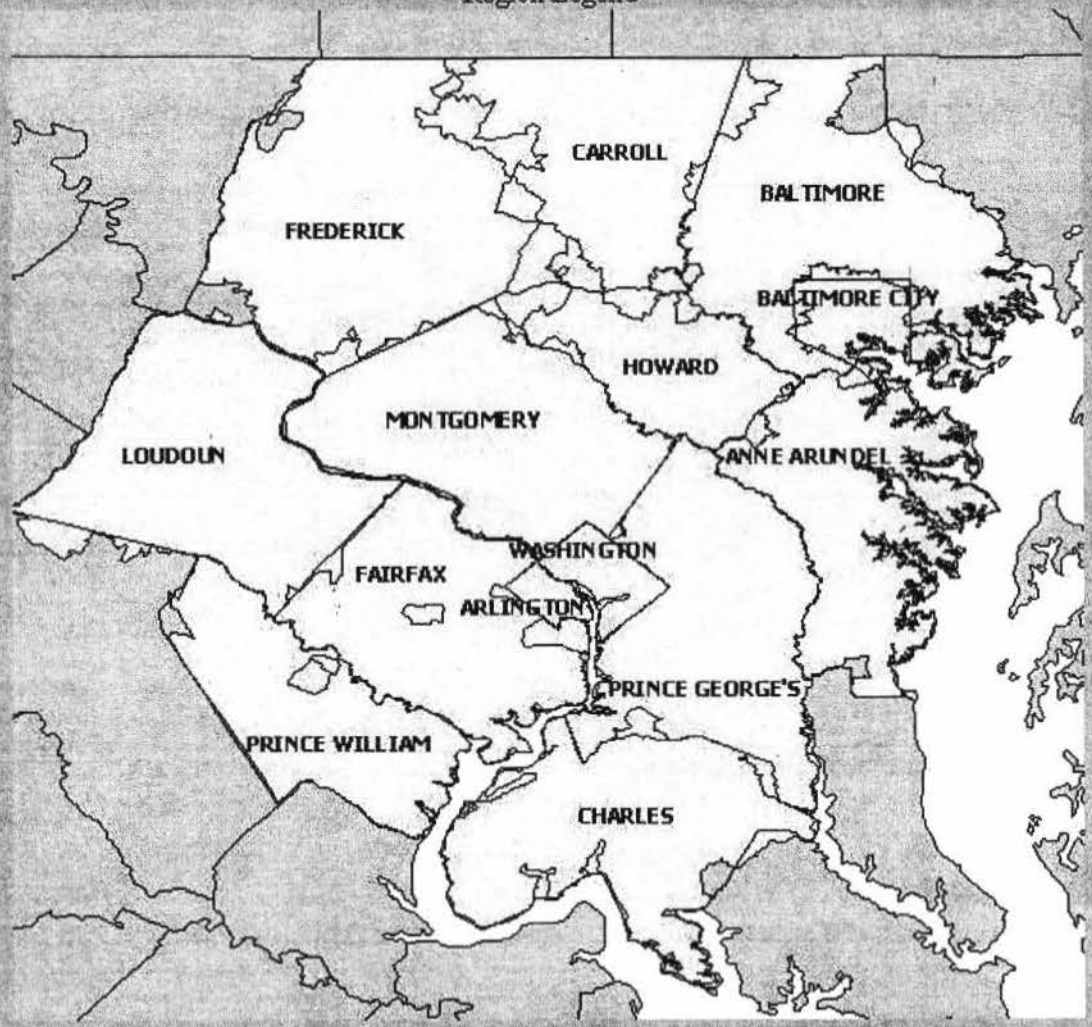
OTC-AD
OTC-FLU
ER
- URI | LRI
- UGI | LGI
- FEVER | NEURO
- CNS | DERM

SYNDROMIC
- URI | LRI
- UGI | LGI
- FEVER | NEURO
- CNS | DERM

JHU/APL - DARPA ISO
Modified Biosurveillance Prototype System

DARPA

Region Legend

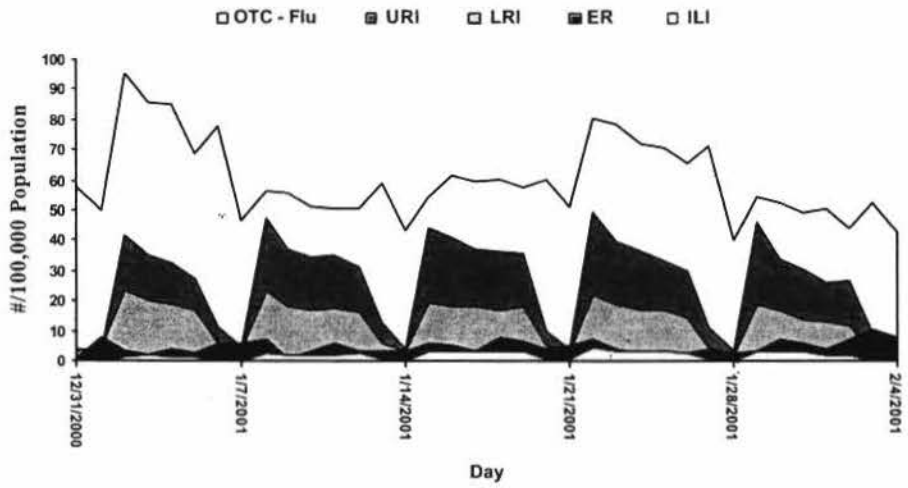


Official Use Only

Document: Done

Heterogeneous Data Issues

Daily Levels for Health Indicators Normalized by Population



- **Each data source has unique time characteristics**
  - **Time variability**
  - **Amplitude variability**
  - **Seasonal variability**
  - **Etc.**

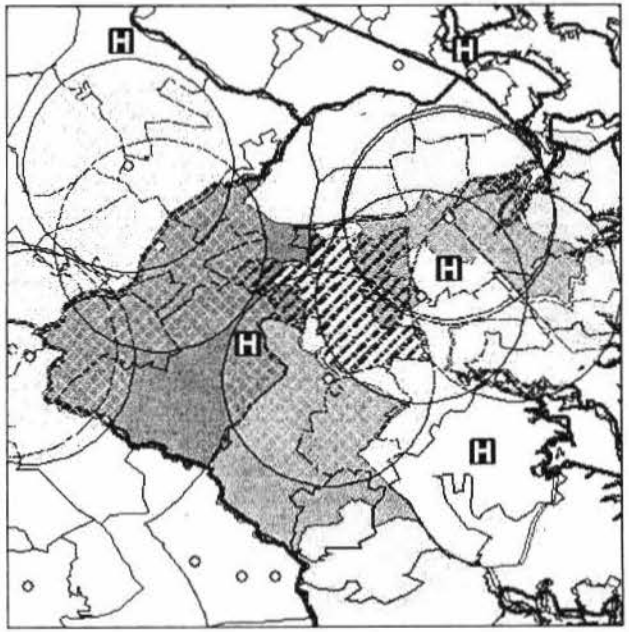
- **Each data source has unique spatial characteristics**

-Individual Zip Code

- Hospital with ER

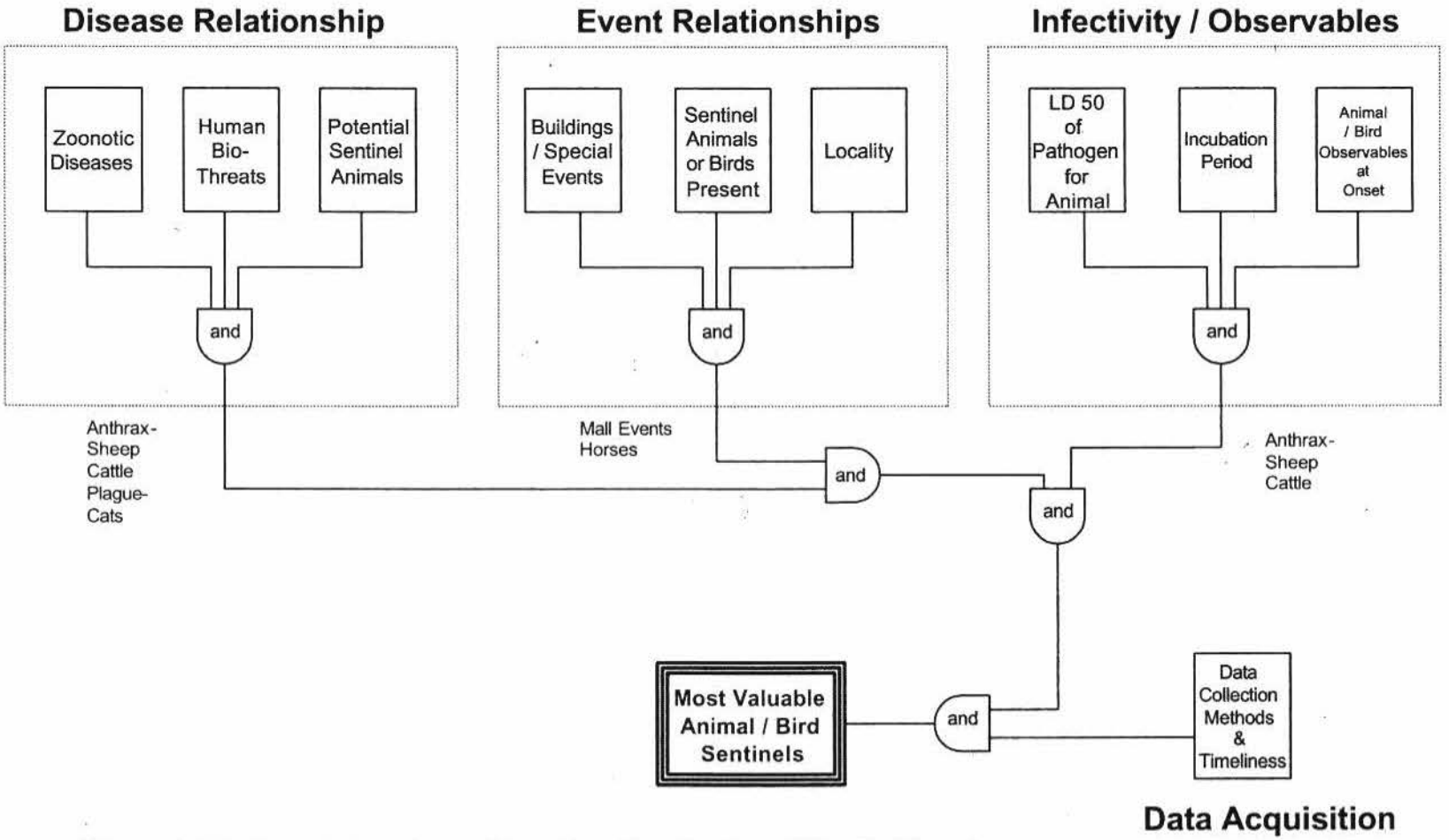
-Pharmacy Catchment area

-High School Districts



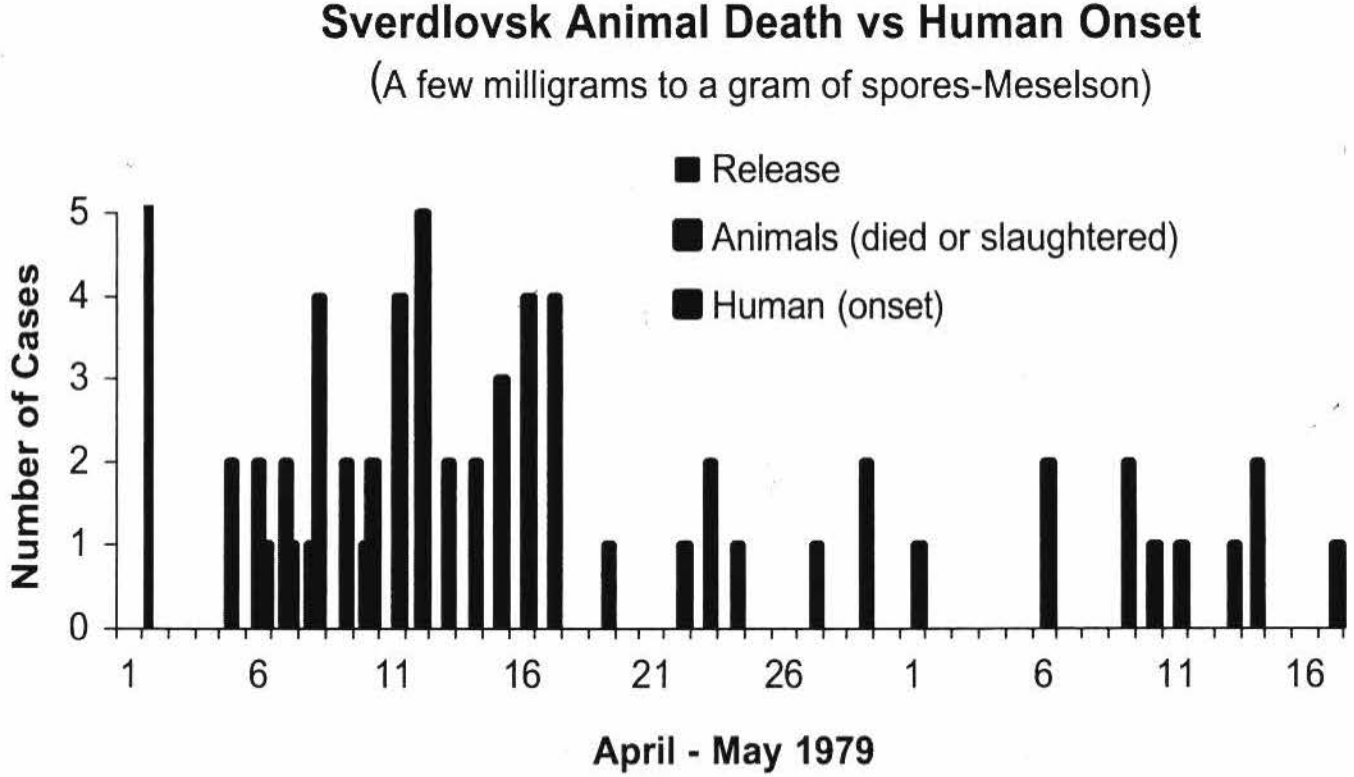
Official Use Only

Data Source Selection Criteria
(Example: Animal Indicators)



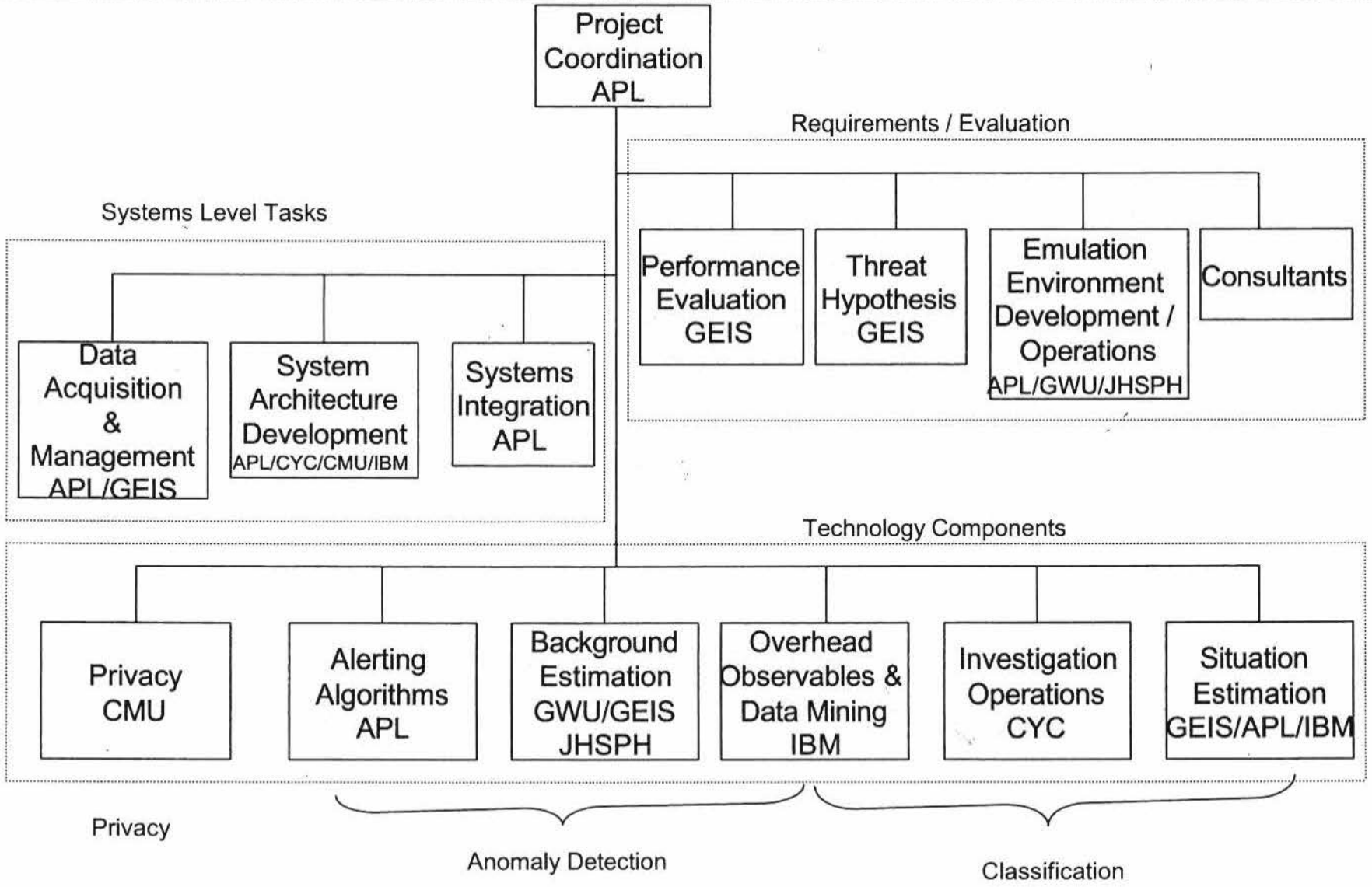
Most Valuable Sentinels Satisfy All Criteria

Early Animal Indicator at Sverdlovsk Outbreak

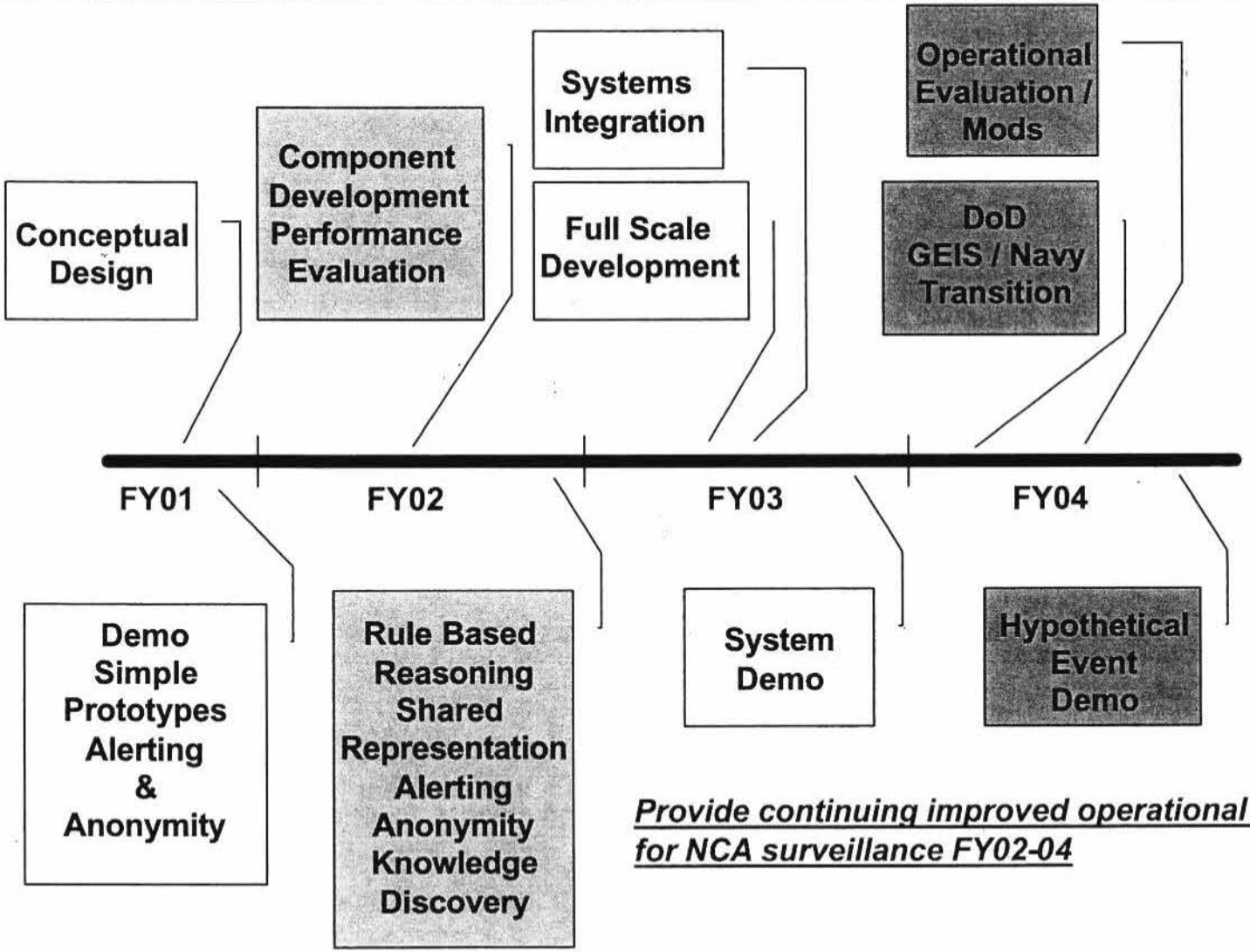


Meselson, M.. Science, Vol.266, 18 Nov.1994

ESSENCE II Team Interactions



Program Plan and Milestones



Provide continuing improved operational capability for NCA surveillance FY02-04

FY01 Accomplishments

1) Developed joint military / civilian scenario for the Baltimore / Washington corridor to evaluate outbreak detection approaches.

2) Participated in surveillance activities for the Presidential Inaugural

3) Developed NCA scenario for use in evaluating ESSENCE II components

4) Participated in a regional table top exercise to capture requirements for multi-jurisdictional outbreak detection

5) Developed transition for ESSENCE II technologies with the Maryland Department of Health and Mental Hygiene

6) Explored approaches for performing animal surveillance in the NCA as well as integrating sensor inputs.

7) Refine outbreak detection methodology based on experiences of the past year

8) Conduct daily analysis of NCA Indicators since Sept. 11th.

ESSENCE II Team Presentations

Topic	Presenter	Organization
Surveillance Requirements in the NCA	MAJ Julie Pavlin / Mike Stoto	DoD-GEIS GWU
Patient Anonymity	Latanya Sweeney	CMU
ESSENCE II Architecture	Steve Langs	JHUAPL
Outbreak Detection Methodology	Howard Burkom / Eugene Elbert	JHUAPL DoD-GEIS
Classification Methodology	Dave Gunning	Cycorp

ESSENCE II Team Presentations

Topic	Presenter	Organization
Surveillance Requirements in the NCA	[redacted] Mike Stoto	DoD-GEIS GWU
Patient Anonymity	Latanya Sweeney	CMU
ESSENCE II Architecture	Steve Langs	JHUAPL
Outbreak Detection Methodology	Howard Burkom / Eugene Elbert	JHUAPL DoD-GEIS
Classification Methodology	Dave Gunning	Cycorp

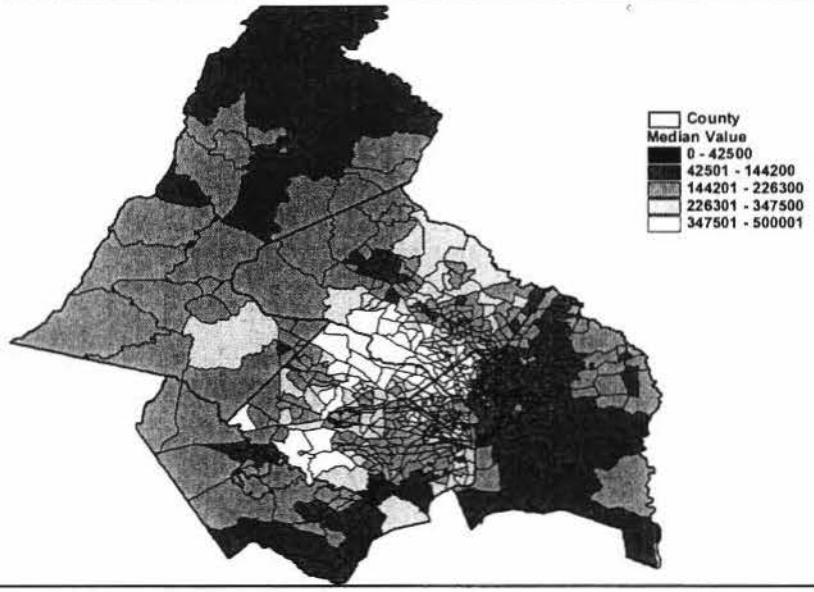
1.a.5

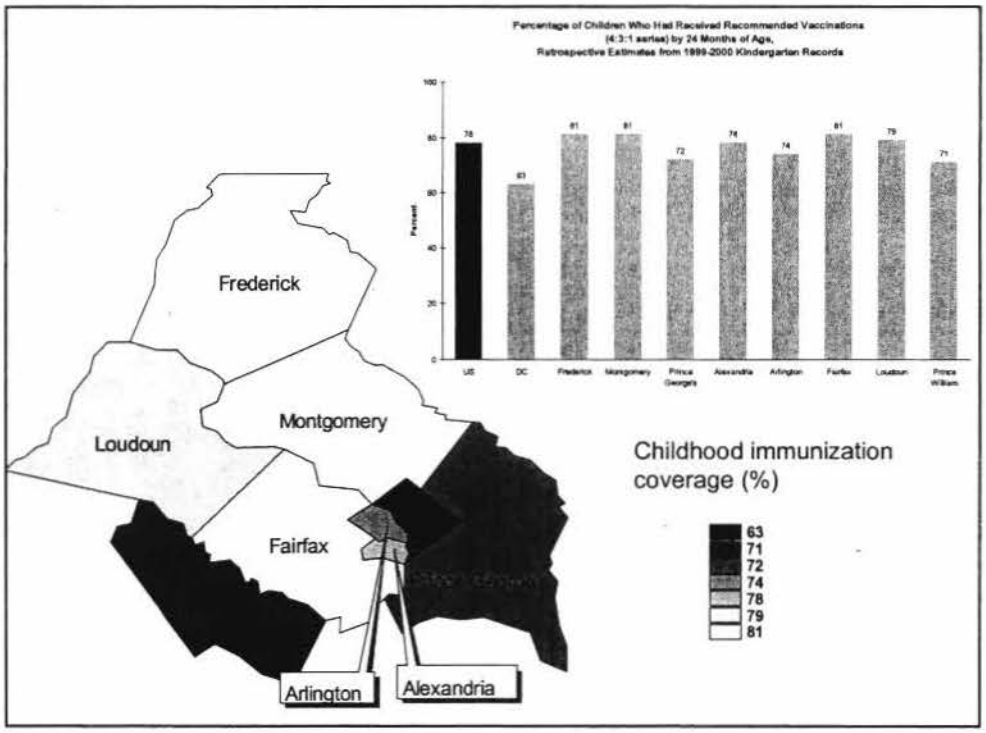
Bioterrorism surveillance in the Washington metro area

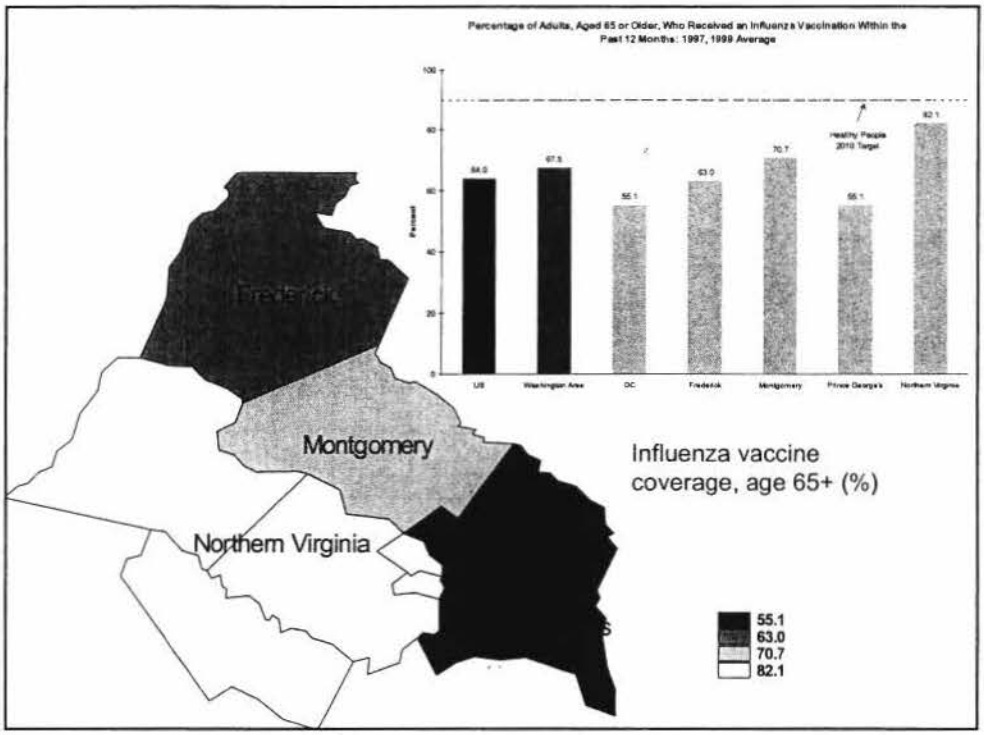
Michael A. Stoto, PhD
Virginia Foster, PhD
George Washington University
School of Public Health
and Health Services

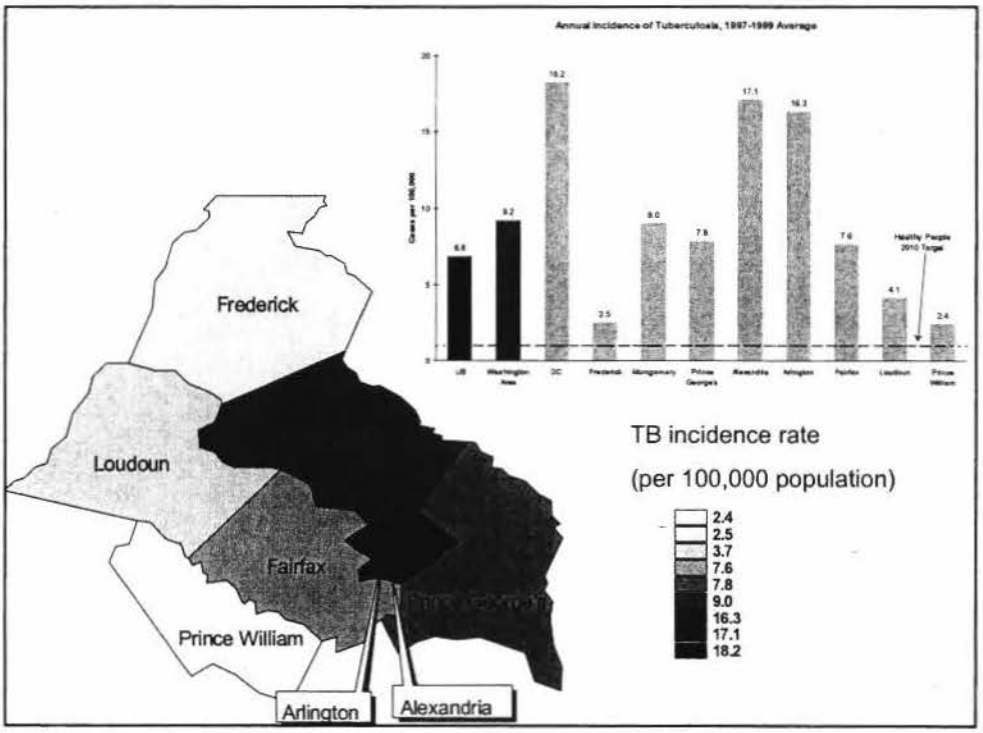
Metropolitan Washington Council of Governments

Socioeconomic diversity: Median housing value, 1990






Percentage of Adults, Aged 65 or Older, Who Received an Influenza Vaccination Within the Past 12 Months: 1997, 1999 Average
Healthy People 2010 Target
Frederick
Montgomery
Northern Virginia
Influenza vaccine coverage, age 65+ (%)
55.1
63.0
70.7
82.1


Annual Incidence of Tuberculosis, 1997-1999 Average
Healthy People 2010 Target
Frederick
Loudoun
Fairfax
Prince William
Arlington
Alexandria
TB incidence rate
(per 100,000 population)
2.4
2.5
3.7
7.6
7.8
9.0
16.3
17.1
18.2

Contributions of epidemiology and public health

- Help identify site and nature of an attack
  - Epidemiologic investigation
- Prevent secondary spread of an infectious agent to populations not exposed in the original attack
- Alleviate suffering by enabling a rapid medical response
  - Even if some patients are seen too late to treat effectively, early identification of an attack can help to find others while treatment is still effective
- Identify and remove continuing sources of biological organisms

Surveillance

- Core function of public health
- Purposes
  - identification of new disease events
  - monitor movement of a disease through a population
  - continued watchfulness over disease in the population
  - identification of infected individuals
- Types of surveillance
  - active vs. passive
  - diagnostic vs. syndromic
    - reportable/notifiable diseases
  - sentinel health events

Epidemiology of a bioterrorist attack

- Biological agents differ from other weapons in two respects
  - victims are likely to be widely dispersed in time and place when they are diagnosed
  - symptoms caused by many potential agents are difficult to diagnose
    - typically emerge as non-specific (flu-like) symptoms
    - physicians are not familiar with advanced symptoms

Possible scenario

- Release of anthrax spores at Metro Center during rush hour, hundreds exposed
- Individuals disperse to homes in Maryland, Virginia, and DC
- Days later, some develop flu-like symptoms and attempt to self-medicate
  - attack might be observable through
    - emergency room "flu" visits
    - absentee rates at downtown offices

Possible scenario

- Days later, specific symptoms occur and patients seek care
  - primary care physicians in various settings, not infectious disease specialists
  - widely dispersed throughout the metro area
  - attack might be observable through reports and analysis of "unusual cases"
- Eventually, reports from central laboratories confirm diagnosis

Surveillance requirements

- Without enhanced surveillance, a bioterrorist attack would be difficult to detect before it is too late to respond effectively
- The region needs an active surveillance system that
  - operates in real time
  - is focused on symptoms, not diagnosed disease
  - accommodates suspicions of providers
  - allows for public health involvement in epidemiologic investigation
- Research is needed to determine how to implement more effectively

Ideal surveillance system

- Draws on data from many sources
  - clinical information from private physicians' offices, emergency rooms, public clinics, etc.
  - laboratory data
  - notifiable disease reports
  - sales of prescription and nonprescription medications
  - absenteeism in schools and workplaces
  - billing, insurance claims data

Ideal surveillance system

- Real time electronic communication of large amounts of data to a central source
- Artificial intelligence capacity to sort through incoming data and identify changes that might signal an attack
- Communication system to identify and discuss small numbers of widely dispersed unusual cases

Ideal surveillance system

- Epidemiological intelligence function, based on deep knowledge of infectious diseases to
  - quickly develop case definitions
  - investigate potential outbreaks
  - separate attacks from hoaxes and natural occurrences
  - implement public health responses

Barriers to implementation

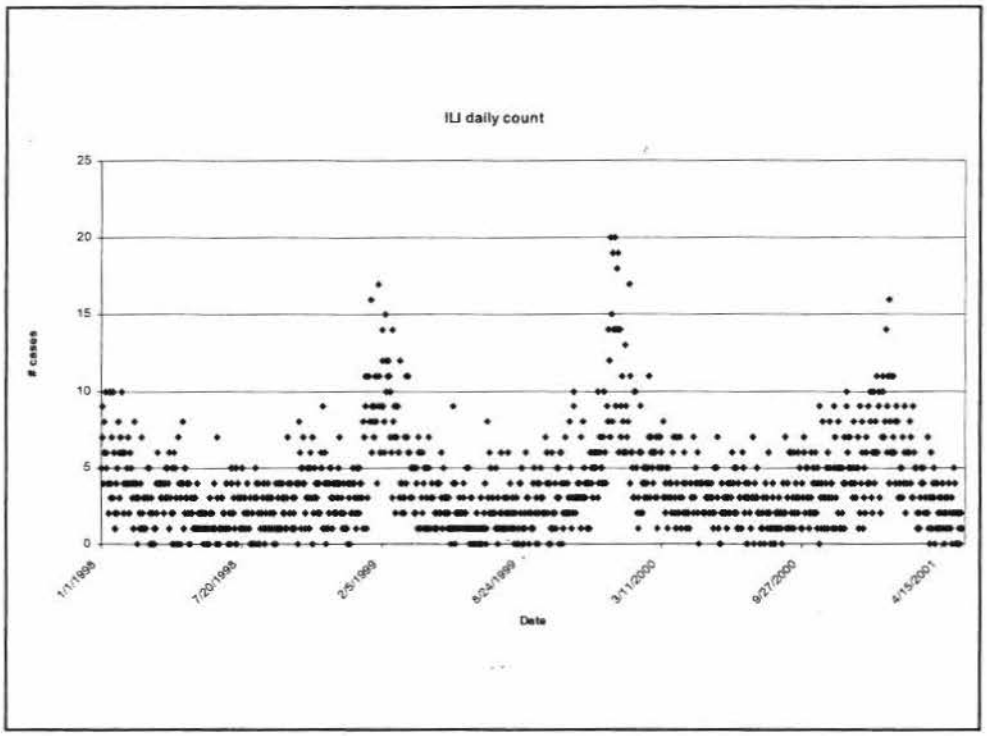
- Much of the desirable data is not computerized, even at the source
- Electronic communication systems do not exist, are incompatible, or very costly
- Data owners reluctant to share information with government
  - privacy and confidentiality of patients
  - proprietary concerns
  - costs of data sharing

Barriers to implementation

- Current infectious disease reporting system
  - state differences in what is reportable
  - reports go to Atlanta via Richmond, Baltimore, and Washington
- Difficult to determine whether an increase in number of cases is "real"
- Not just lack of resources

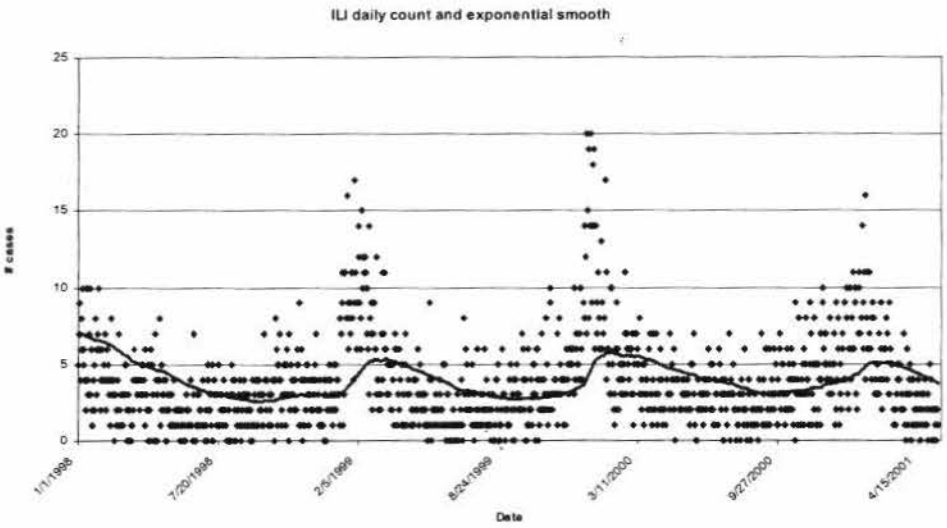
CUSUM techniques to identify "flu" outbreaks

- Initial symptoms of most agents agents are "flu-like"
- Data available
  - George Washington University Hospital Emergency Department
  - From Jan. 1, 1988 trough May 15, 2001
  - 3.5 flu seasons: Nov. 1 through May 15
- ICD-9 code groups
  - Flu (influenza)
  - ILI (influenza-like illness)
  - URI (upper respiratory infection)
  - Viral NOS (not otherwise specified)
  - Pneumonia



Exponential smoothing

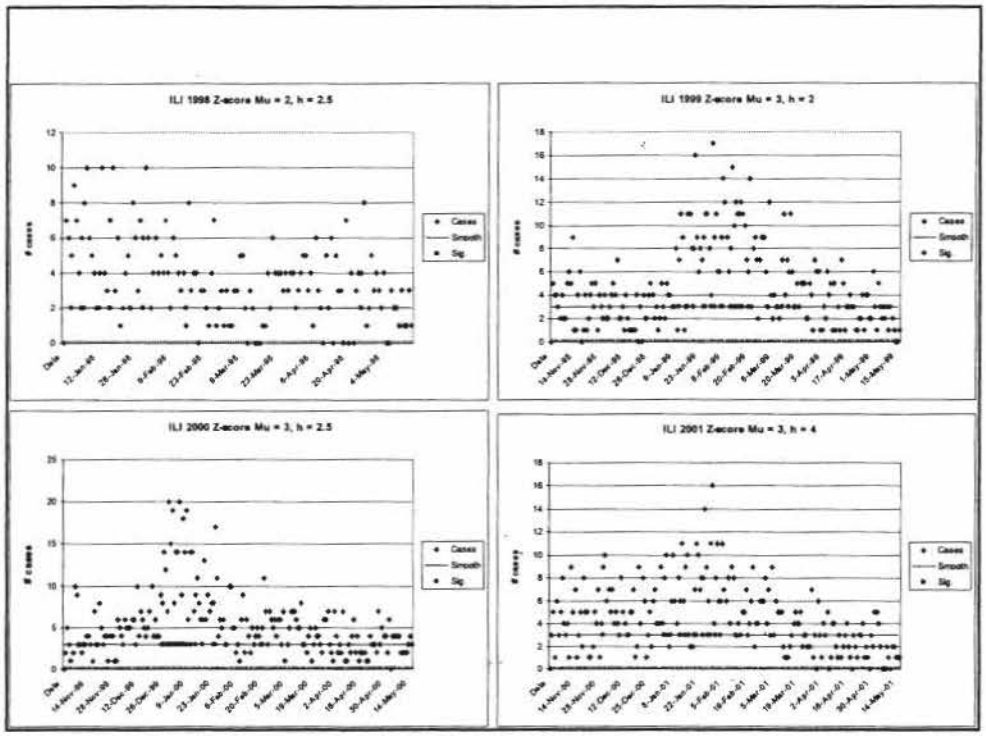
- Set w between
  - 0 (more smoothing) and
  - 1 (less smoothing)
- $y_t = wx_t + (1-w)y_{t-1}$
- non-symmetric running average



Onset of '99 and '00 seasons clear, but not '98 and '01

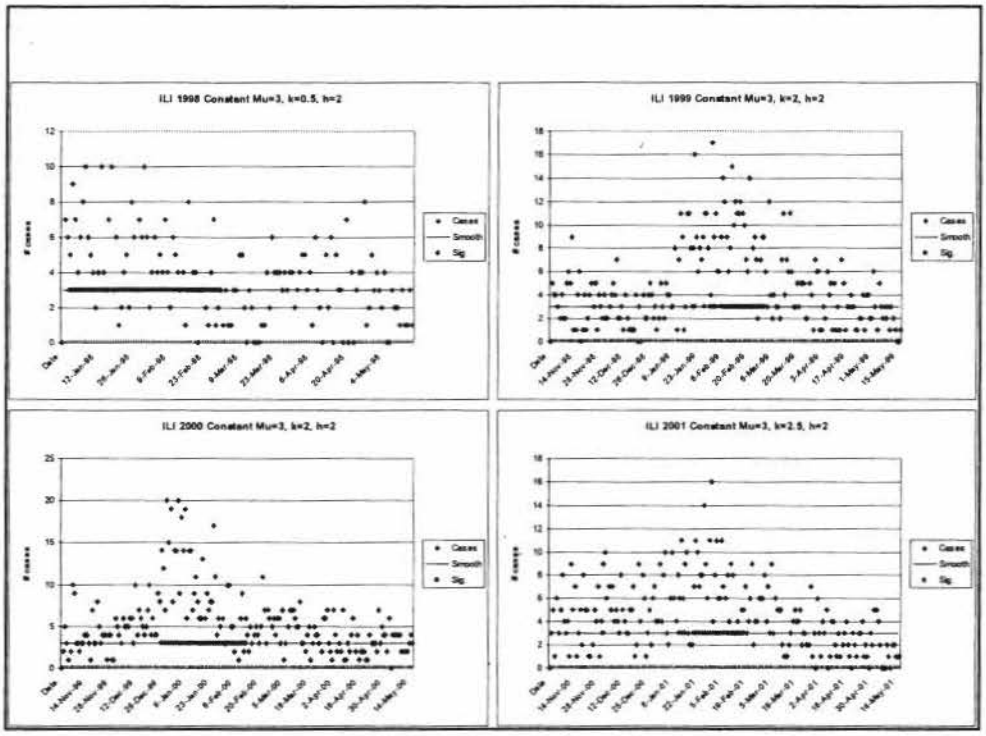
Standard outbreak detection method – Z-scores

- Estimate μ, $\sigma = \sqrt{\mu}$ (Poisson distribution)
  - $Z_t = (x_t - \mu)/\sigma$
  - Set h
  - "Outbreak" if $Z_t > h$
- Drawbacks
  - Many false positives and negatives
  - Optimal parameters vary by year
    - μ: 2 to 3
    - h: 2 to 4



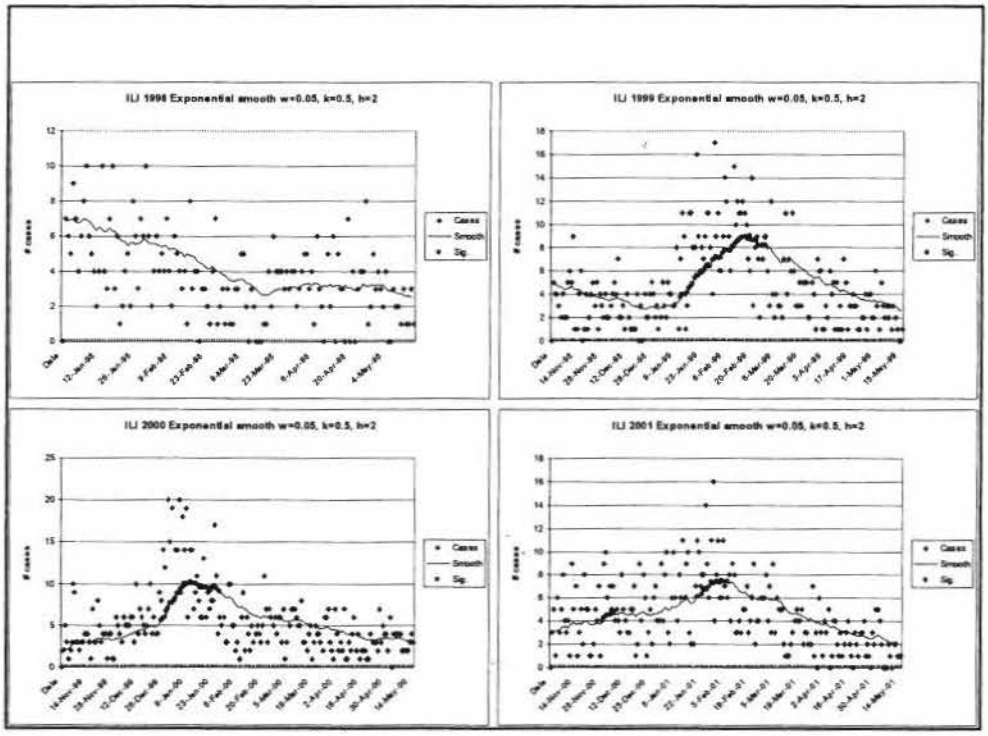
Alternative 1: CUSUM with constant background level

- Method
  - Estimate μ, $\sigma = \sqrt{\mu}$ (Poisson distribution)
  - $Z_t = (x_t - \mu)/\sigma$
  - Set k, h
  - $S_t = \text{Max}[S_{t-1} + Z_t - k, 0]$
  - "Outbreak" if $S_t > h$
- Comparison
  - Fewer false positives and negatives
  - Optimal parameters vary by year
    - k: 0.5 to 2.5



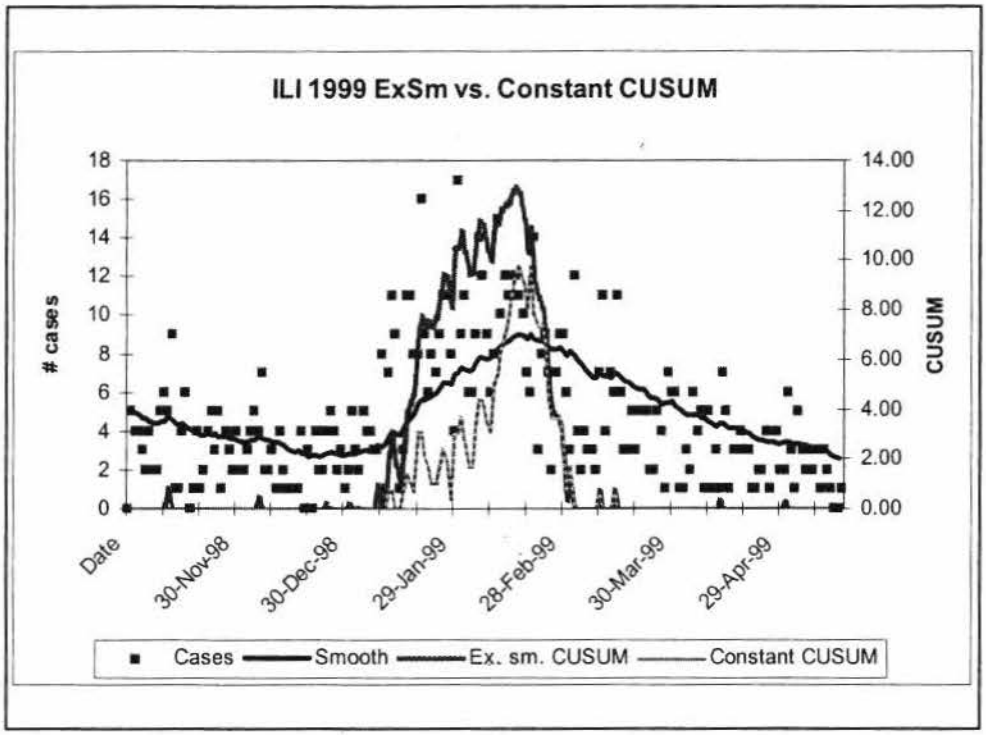
Alternative 2: CUSUM with exp. smoothing background level

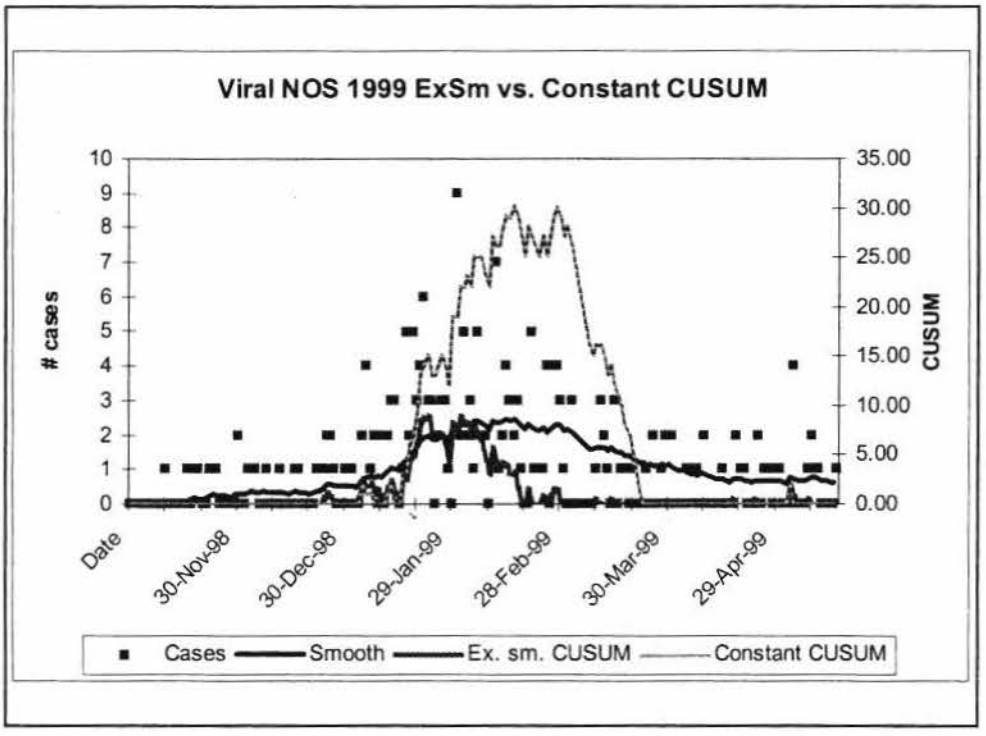
- Method
  - Set w, k, h, estimate σ from data
  - Calculate $y_t = wx_t + (1\text{-}w)y_{t-1}$
  - $Z_t = (x_t - y_t)/\sigma$
  - $S_t = \text{Max}[S_{t-1} + Z_t - k]$
  - "Outbreak" if $S_t > h$
- Comparison
  - More false positives than alternative 1
  - Unable to detect '98 outbreak
  - Optimal parameters do not vary by year

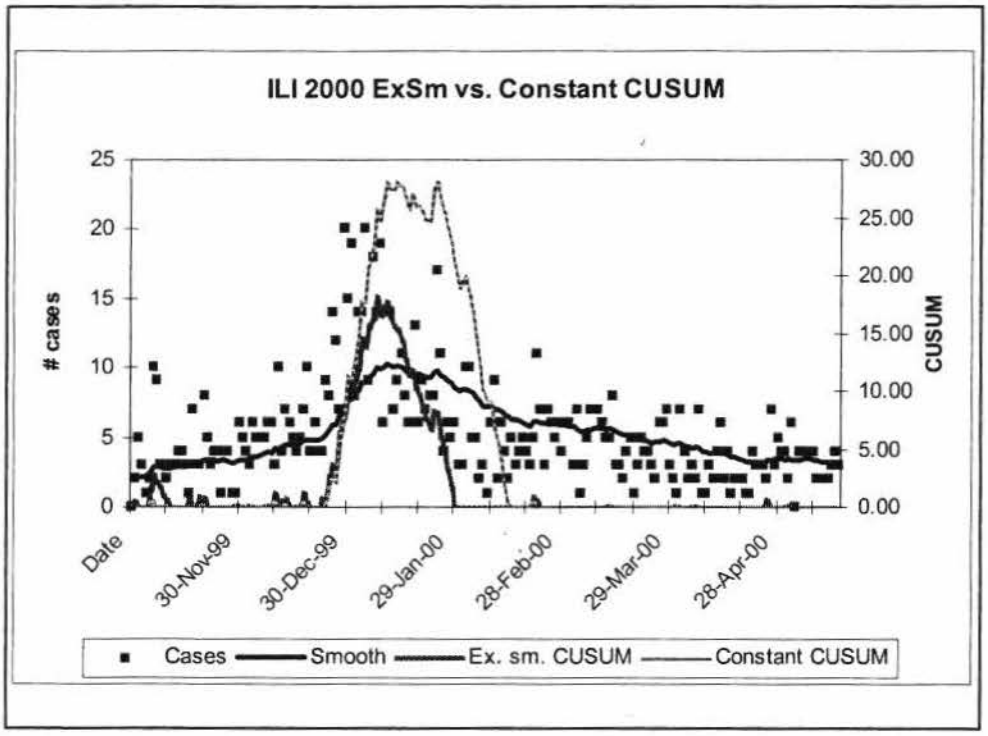


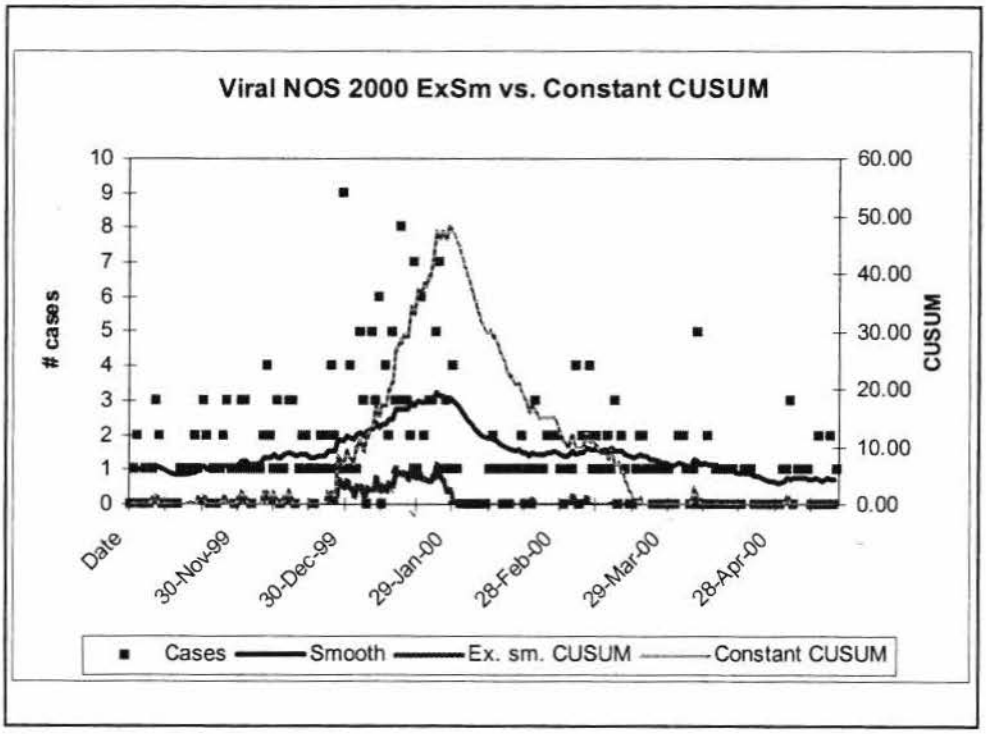
Comparison of methods

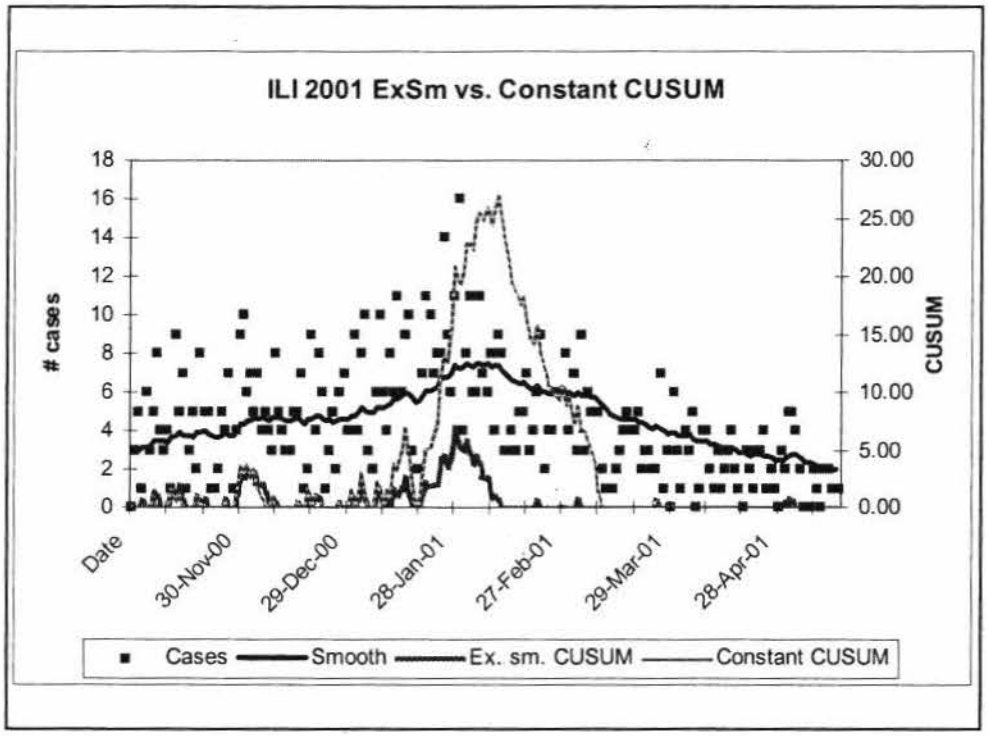
- Methods
  - CUSUM with constant base
  - CUSUM with exponential smooth base
- Data
  - Influenza-like illnesses
  - Viral not otherwise specified
  - 1999, 2000, and 2001 flu seasons

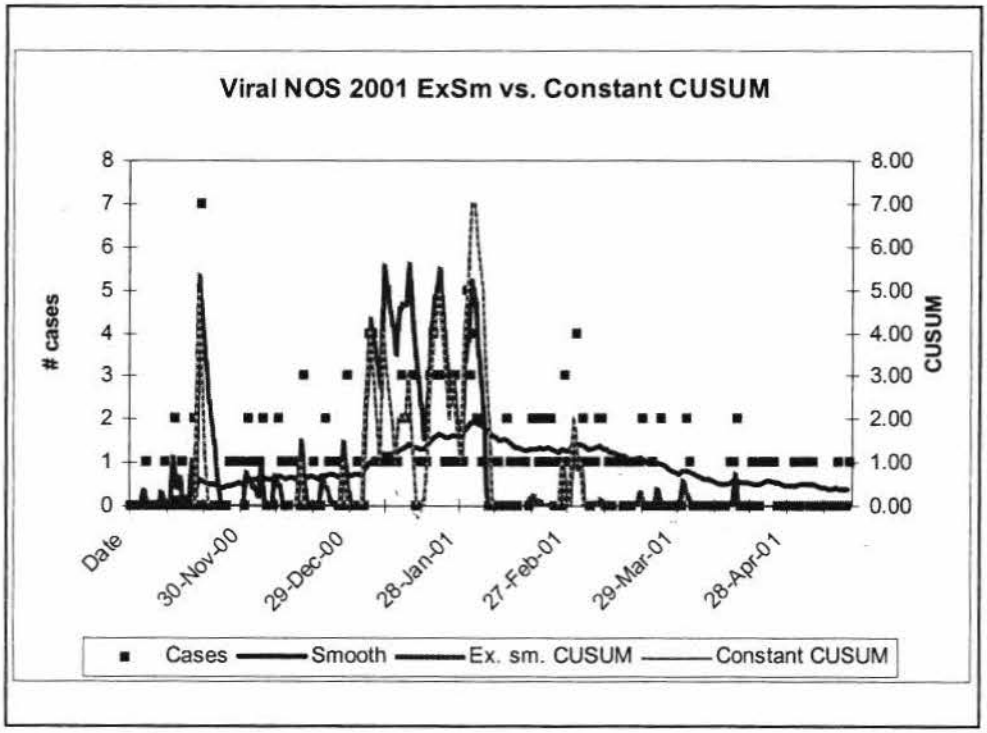

ILI 1999 ExSm vs. Constant CUSUM
cases
CUSUM
18
16
14
12
10
8
6
4
2
0
14.00
12.00
10.00
8.00
6.00
4.00
2.00
0.00
Date
30-Nov-98
30-Dec-98
29-Jan-99
28-Feb-99
30-Mar-99
29-Apr-99
Cases
Smooth
Ex. sm. CUSUM
Constant CUSUM


Viral NOS 1999 ExSm vs. Constant CUSUM
cases
CUSUM
10
9
8
7
6
5
4
3
2
1
0
35.00
30.00
25.00
20.00
15.00
10.00
5.00
0.00
Date
30-Nov-98
30-Dec-98
29-Jan-99
28-Feb-99
30-Mar-99
29-Apr-99
Cases
Smooth
Ex. sm. CUSUM
Constant CUSUM


ILI 2000 ExSm vs. Constant CUSUM
cases
CUSUM
25
20
15
10
5
0
30.00
25.00
20.00
15.00
10.00
5.00
0.00
Date
30-Nov-99
30-Dec-99
29-Jan-00
28-Feb-00
29-Mar-00
28-Apr-00
Cases
Smooth
Ex. sm. CUSUM
Constant CUSUM


Viral NOS 2000 ExSm vs. Constant CUSUM
cases
CUSUM
10
9
8
7
6
5
4
3
2
1
0
60.00
50.00
40.00
30.00
20.00
10.00
0.00
Date
30-Nov-99
30-Dec-99
29-Jan-00
28-Feb-00
29-Mar-00
28-Apr-00
Cases
Smooth
Ex. sm. CUSUM
Constant CUSUM


ILI 2001 ExSm vs. Constant CUSUM
cases
CUSUM
18
16
14
12
10
8
6
4
2
0
30.00
25.00
20.00
15.00
10.00
5.00
0.00
Date
30-Nov-00
29-Dec-00
28-Jan-01
27-Feb-01
29-Mar-01
28-Apr-01
Cases
Smooth
Ex. sm. CUSUM
Constant CUSUM


Viral NOS 2001 ExSm vs. Constant CUSUM
cases
CUSUM
8
7
6
5
4
3
2
1
0
8.00
7.00
6.00
5.00
4.00
3.00
2.00
1.00
0.00
Date
30-Nov-00
29-Dec-00
28-Jan-01
27-Feb-01
29-Mar-01
28-Apr-01
Cases
Smooth
Ex. sm. CUSUM
Constant CUSUM

Summary and conclusions

- Start of outbreaks are hard to detect!
  - especially when onset is not sharp
  - will be difficult to distinguish a bioterrorist attack from the real flu epidemic
- For ILI and Viral NOS, CUSUM methods are preferable to standard Z-score methods
  - Fewer false positives and no delay in recognizing outbreaks
- Neither exponential smoothing nor constant (Poisson) CUSUM methods are clearly preferable to the other

Additional research

- Simultaneous analysis of multiple data streams
  - different symptom and ICD code groupings
  - different hospitals and service settings
  - non-health care data
    - over-the-counter flu medications
    - absenteeism
- Geographical and socioeconomic patterns
- As background for these analyses, need to more fully understand the relationship among these indicators in "regular" flu seasons
- Run simulations in which public health officials must decide when to "call" an outbreak

Bio-Surveillance Using Sufficiently Anonymous Data

Bio-Surveillance Kick-Off Meeting

Latanya Sweeney, Ph.D.
Director, Laboratory for International Data Privacy
Assistant Professor of Computer Science & of Public Policy
Carnegie Mellon University
latanya@cs.cmu.edu
http://isos.heinz.cmu.edu/dataprivacy

CMU Team Members

- B. Joseph Barrett
- Lawrence Beasley
- Bradley Malin
- Latanya Sweeney

Cultural Computing (*new area*)

The study of computational solutions and accompanying policies to improve the adoption of computer technology by society.

Data Privacy (*new field*)

The study of computational solutions for releasing data such that the data remain practically useful while the identities of the subjects of the data are not revealed.

Laboratory for International Data Privacy
http://sos.heinz.cmu.edu/dataprivacy/

Learning information about entities...

Data Linkage ("data detectives"):

combining disparate pieces of entity-specific information to learn more about an entity

Privacy Protection ("data protectors"):

release information such that certain entity-specific properties (such as identity) are controlled; restrict what can be learned

This talk

New area and field in CS

- Immediate system goals
- Early detection
- Privacy problem
- Computational privacy solutions
- Legal model
- Live demonstration of prototype 1.0

Immediate System Goals

1. Detects outbreaks or attacks early by continuously observing the general population
2. Provides scientific guarantees of privacy protection; automated adherence to HIPAA
3. Models "probable cause" legal predicate to lower anonymity constraints as conditions warrant
4. Results in automated releases of *useful* data that are minimally invasive to privacy

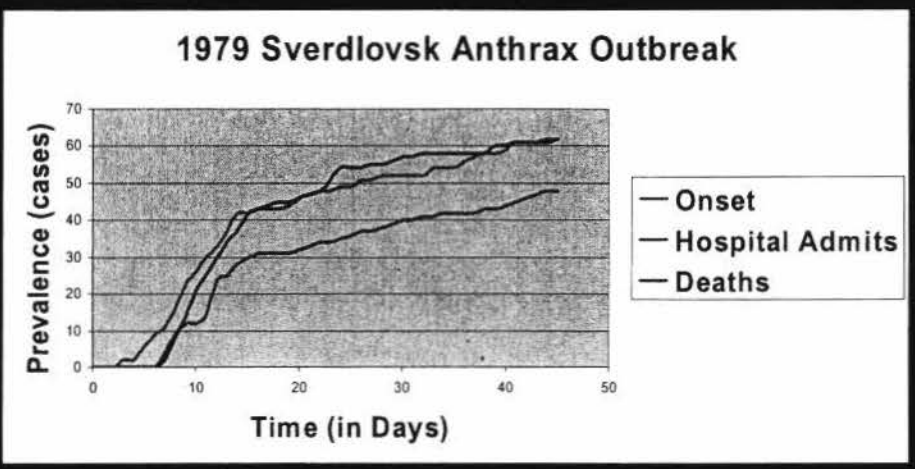
This talk

New area and field in CS

Immediate system goals

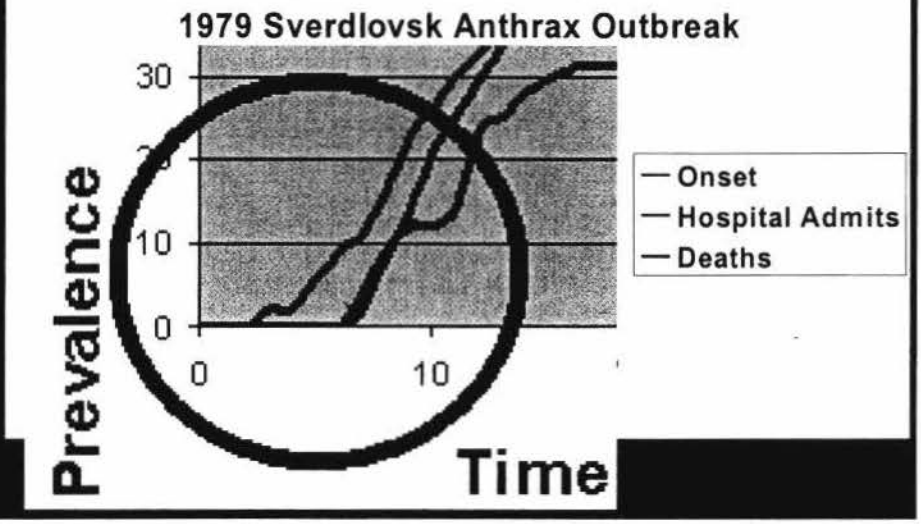
- Early detection
- Privacy problem
- Computational privacy solutions
- Legal model
- Live demonstration of prototype 1.0

Detect Early using Onset, Coordinate Deaths & Hospital Admits



Based on results reported in Guillemin, 1999.

How can we detect onset? How early on each can we predict? How does coordination help?



Continuously Observe Behaviors to Detect Onset of Symptoms

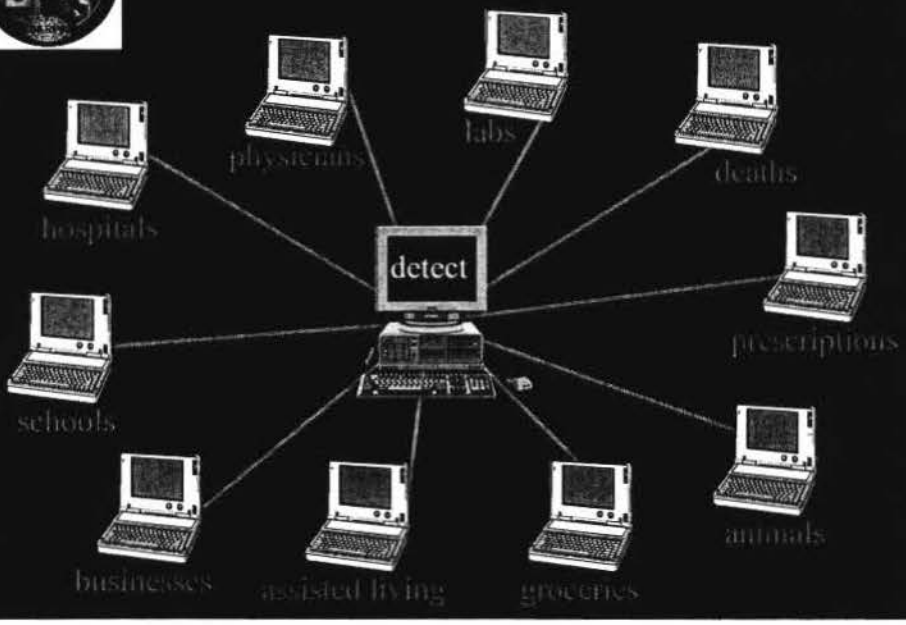
Syndromic surveillance:

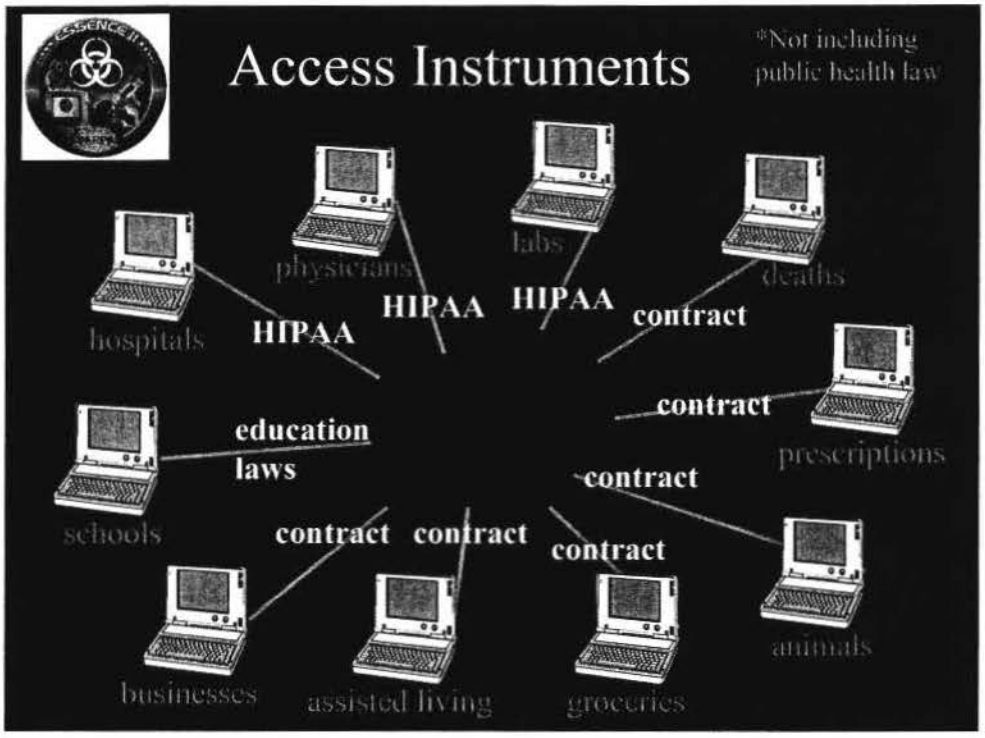
How many are acting ill?

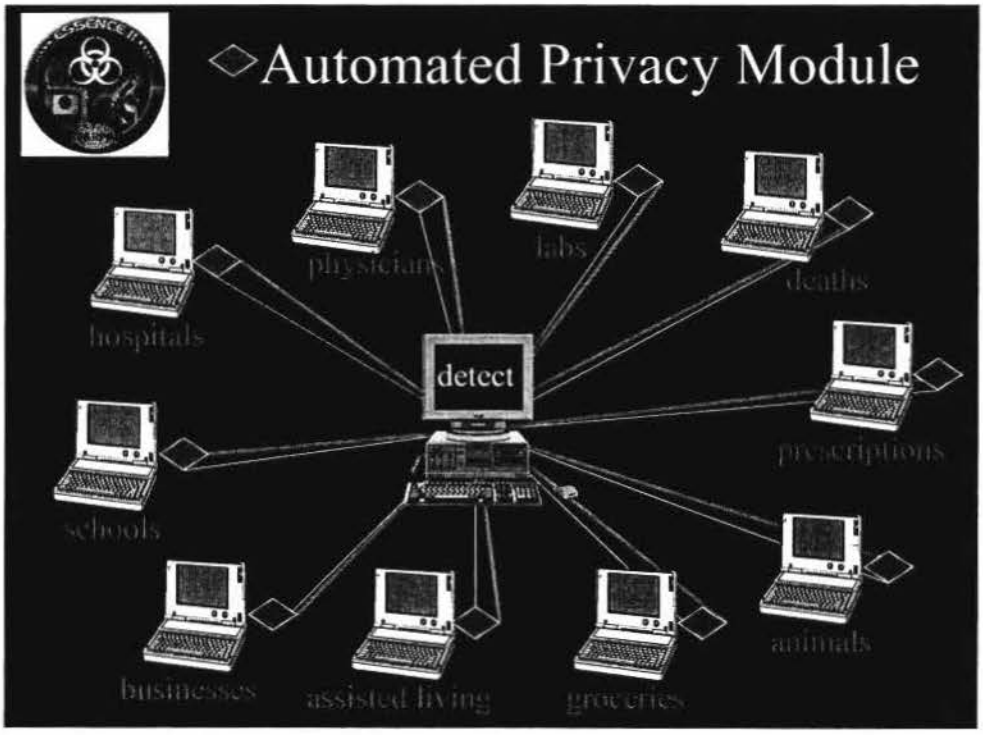
Unusual behaviors→syndromes?

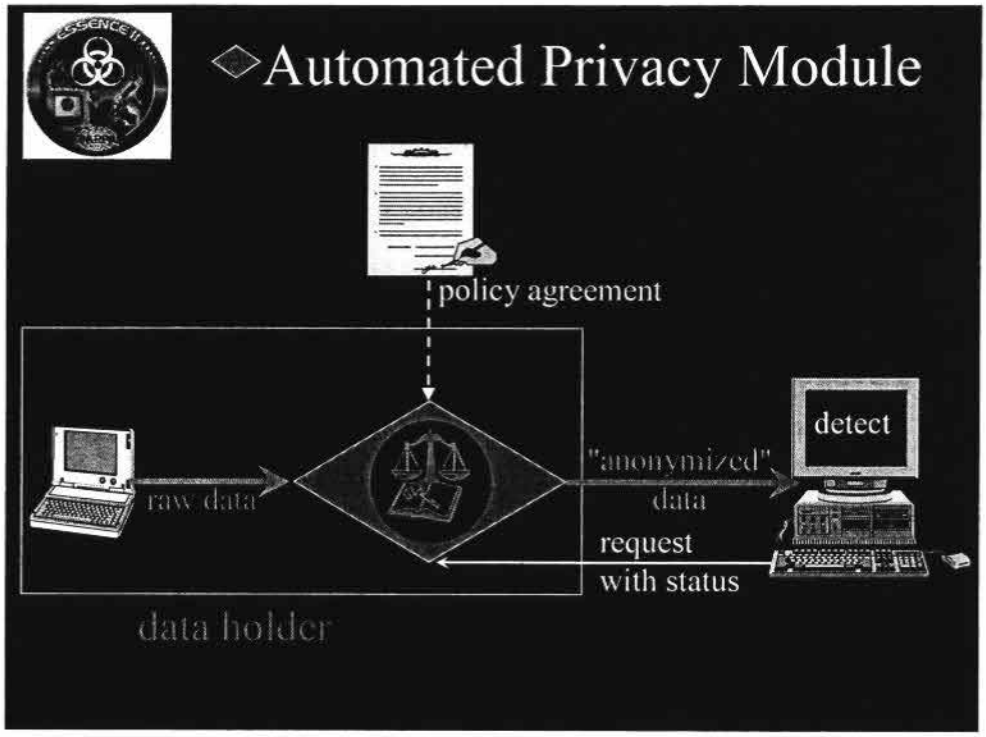
Not confirmed diagnoses!

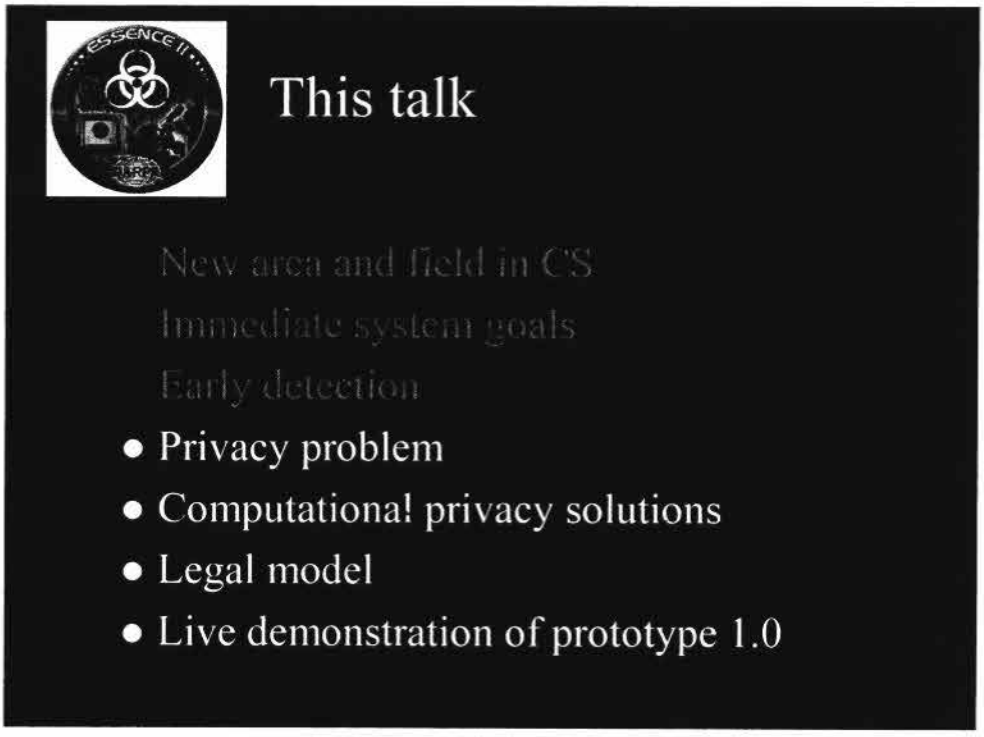
Surveillance of Secondary Data




Access Instruments
*Not including public health law
physicians
labs
deaths
hospitals
HIPAA
HIPAA
HIPAA
contract
contract
prescriptions
education laws
schools
contract
contract
contract
businesses
assisted living
groceries
animals


Automated Privacy Module
physicians
labs
deaths
hospitals
detect
prescriptions
schools
businesses
assisted living
groceries
animals


Automated Privacy Module
policy agreement
raw data
"anonymized" data
request with status
detect
data holder


This talk
New area and field in CS
Immediate system goals
Early detection
• Privacy problem
• Computational privacy solutions
• Legal model
• Live demonstration of prototype 1.0

Looks anonymous

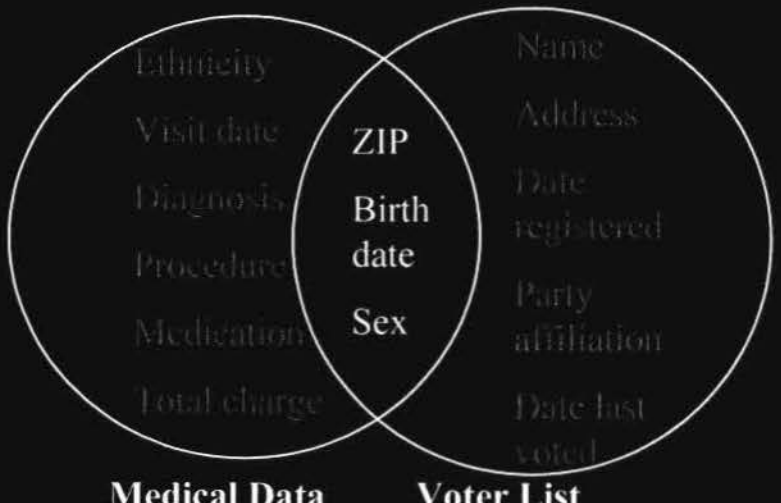
ZIP	Birth	Sex	Ethnicity
33171	7/15/71	m	Caucasian
02657	2/18/73	f	Black
20612	3/12/75	m	Asian

Health data (GIC example)

Medical Data

Population data (GIC example)

Linking to re-identify data



Uniqueness in Cambridge Voters

Birth date alone	12%
Birth date & gender	29%
Birth date & 5-digit ZIP	69%
Birth date & full postal code	97%

Birth date includes month, day and year.
Total 54,805 voters.

JLME 97

Few characteristics make a person unique

Birth includes month, day and year:

365 days x 100 years = 36,500 possibilities

Two genders and Five ZIP (5-digit) codes:

2 * 5 * 36,500 = 365,000 possibilities

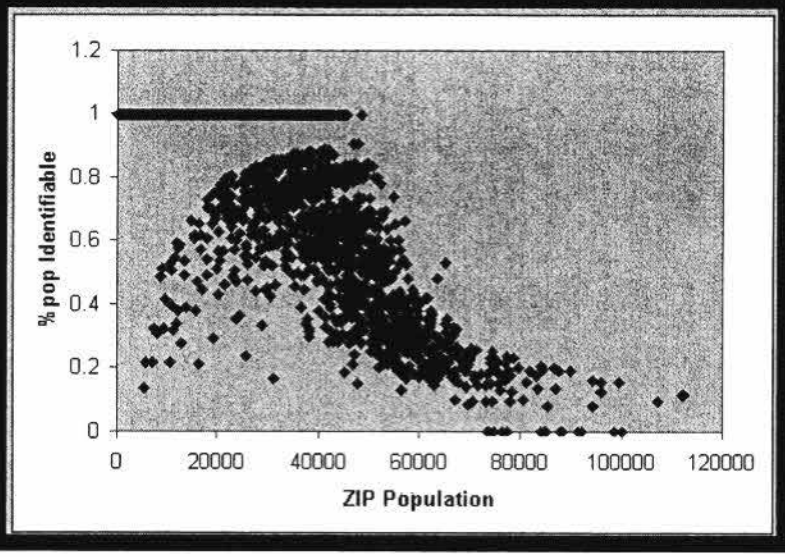
But the Cambridge Voter list had:

54,805 voters

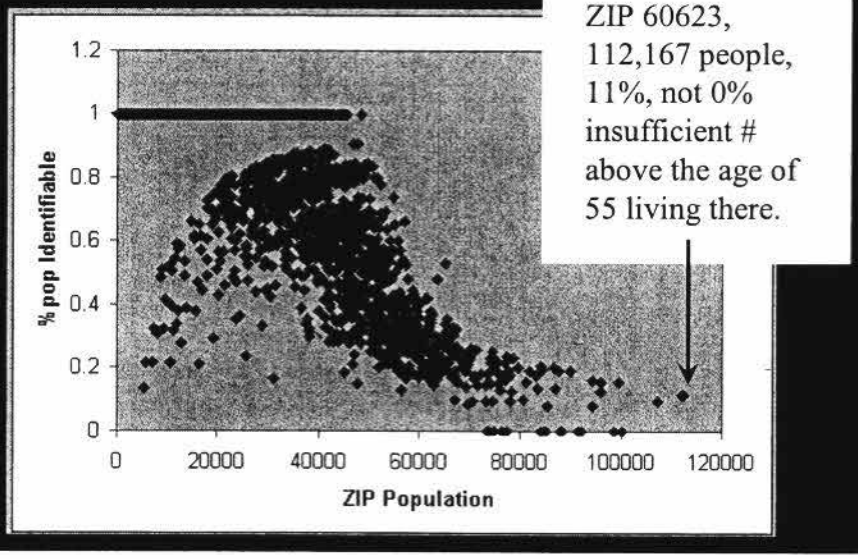
So in general, using
(birth[mon,day,yr], gender, ZIP[5-digit])
provides a unique quasi-identifier.

JLME 97

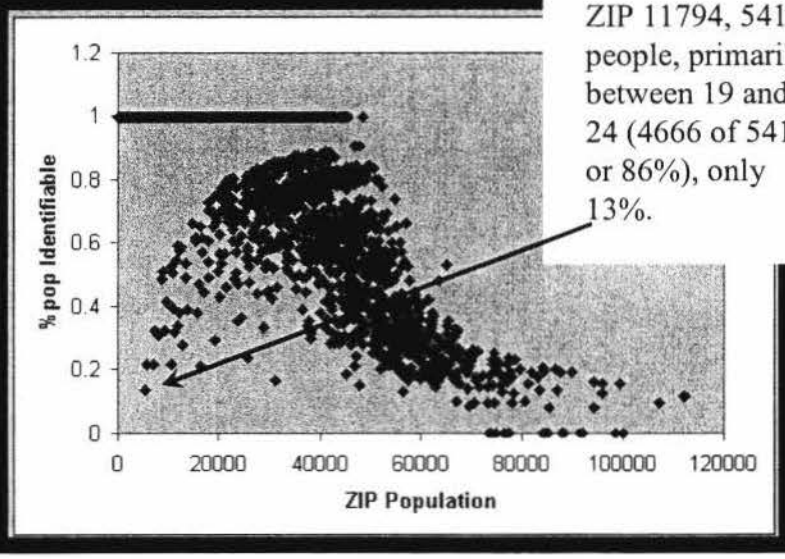
{date of birth, gender, 5-digit ZIP} uniquely identifies 87.1% of USA pop.



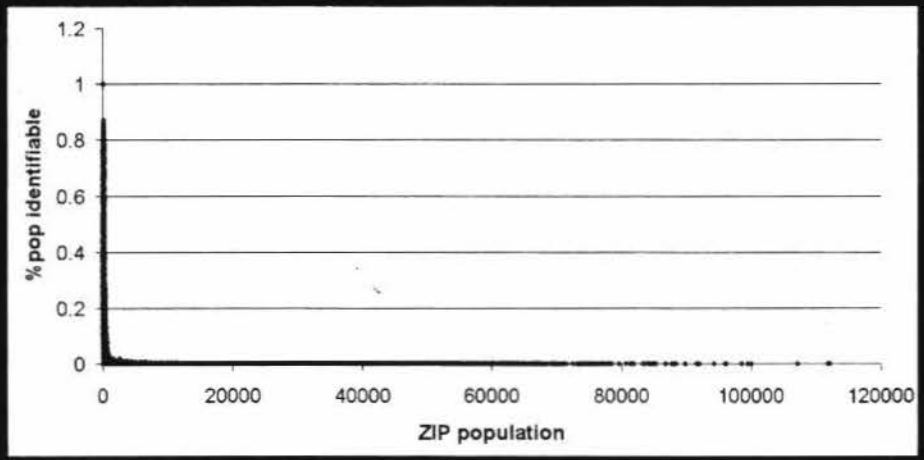
{date of birth, gender, 5-digit ZIP} uniquely identifies 87.1% of USA pop.



{*date of birth, gender, 5-digit ZIP*}
uniquely identifies 87.1% of USA pop.



Even *BirthYear* can still identify people!



This talk

New area and field in CS

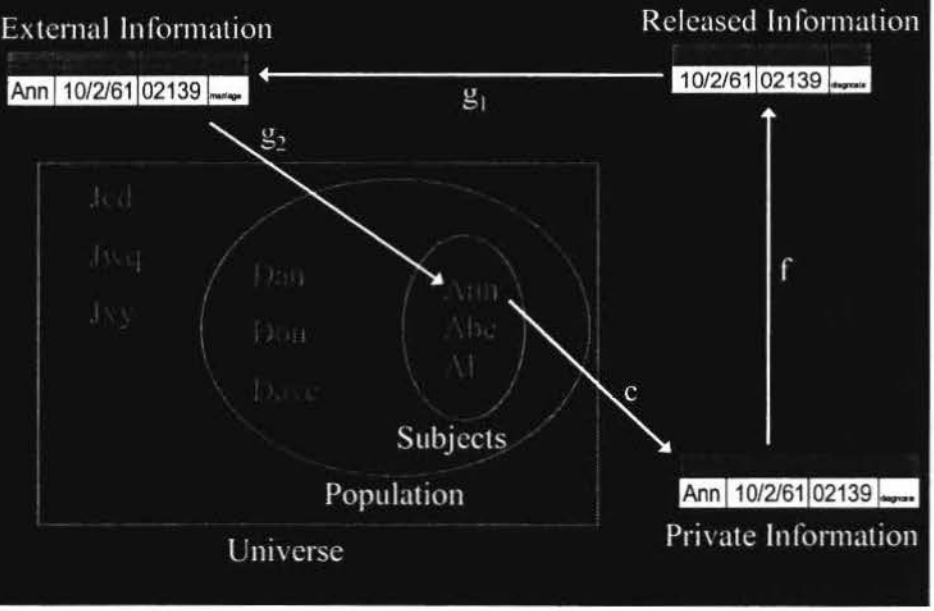
Immediate system goals

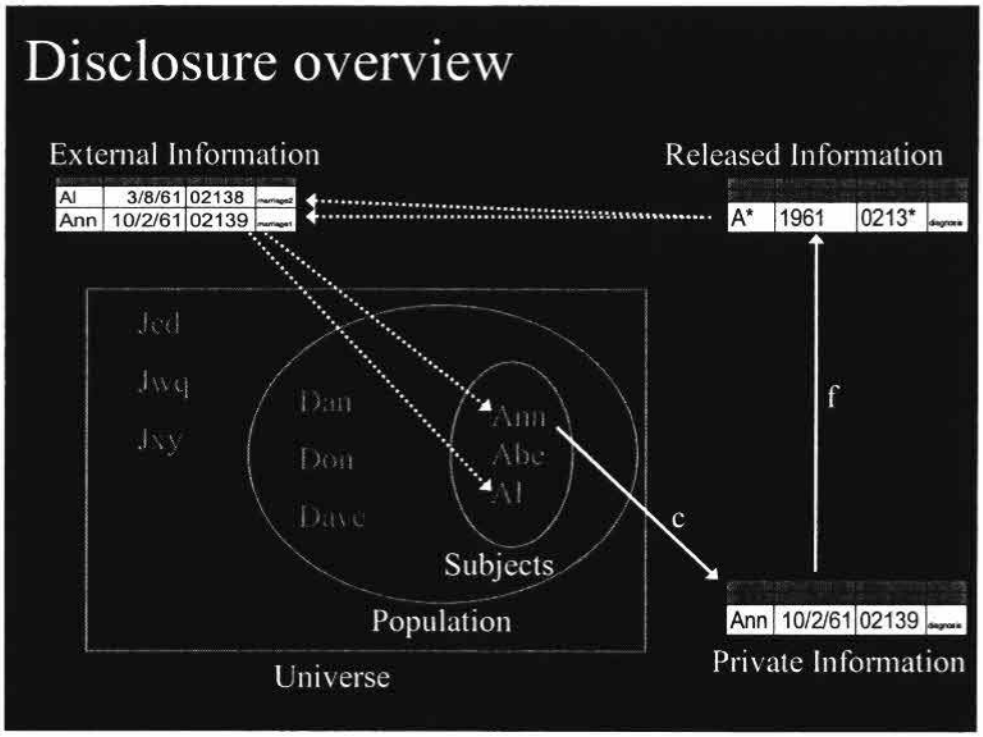
Early detection

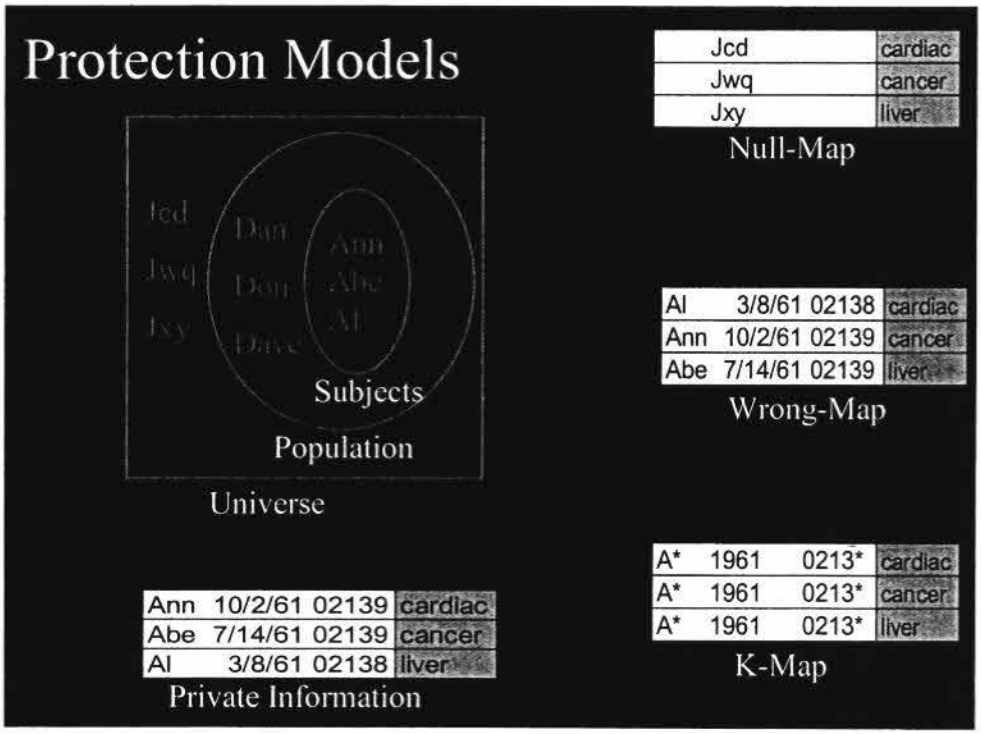
Privacy problem

- Computational privacy solutions
- Legal model
- Live demonstration of prototype 1.0

Disclosure overview




Disclosure overview
External Information
Al 3/8/61 02138
Ann 10/2/61 02139
Released Information
A* 1961 0213*
Jcd
Jwq
Jxy
Dan
Don
Dave
Ann
Abe
Al
Subjects
Population
Universe
f
c
Ann 10/2/61 02139
Private Information


Protection Models
Jcd cardiac
Jwq cancer
Jxy liver
Null-Map
Jcd
Jwq
Jxy
Dan
Don
Dave
Ann
Abe
Al
Subjects
Population
Universe
Al 3/8/61 02138 cardiac
Ann 10/2/61 02139 cancer
Abe 7/14/61 02139 liver
Wrong-Map
Ann 10/2/61 02139 cardiac
Abe 7/14/61 02139 cancer
Al 3/8/61 02138 liver
Private Information
A* 1961 0213* cardiac
A* 1961 0213* cancer
A* 1961 0213* liver
K-Map

Techniques are specific to use

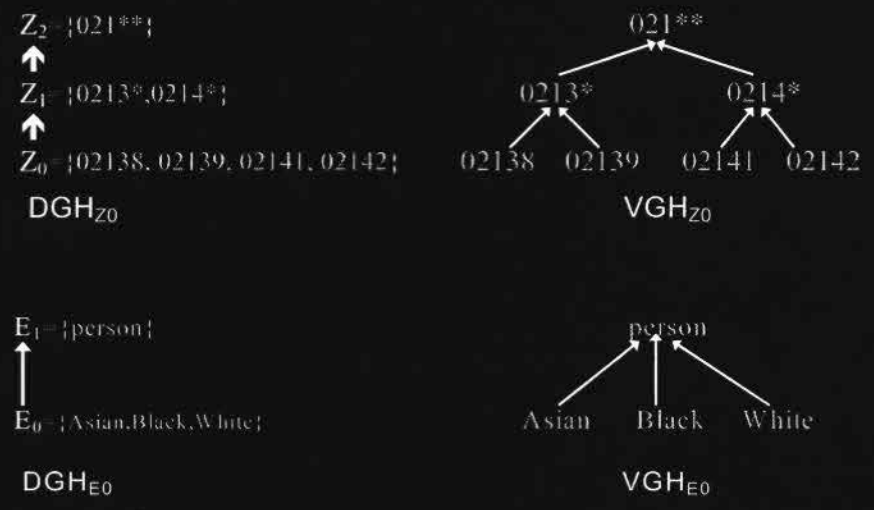
Technique	A-Data Mining	B-Statistical
De-identification	depends	depends
Encryption	depends	depends
Suppression	depends	no
Generalize values	depends	no
Swap values	no	yes
Substitution	depends	depends
Outlier to medians	no	depends
Perturbation	no	yes
Rounding	no	yes
Additive noise	no	yes
Sampling	depends	depends
Add tuples	no	yes
Scramble tuples	yes	yes

k-anonymity and *k*-map protection

- Quasi-identifier, profile $\{Birth_{0.5}, ZIP_{0.7}, Sex_{0.3}\}$
- Generalization 10/27/59 ➲ 1959
- Suppression 02139 ➲ ■■■■
- *Encryption* *3245123* ➲ *2168582*

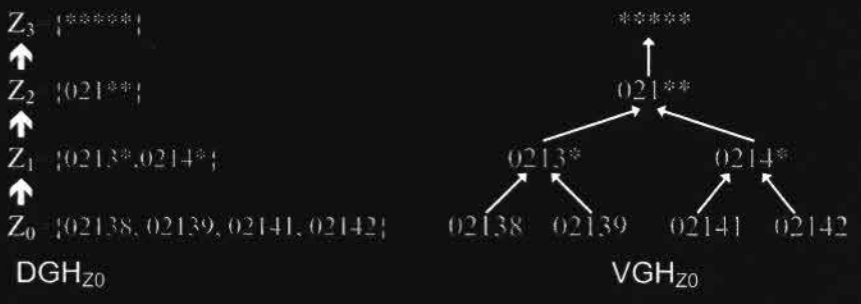
AMIA 97, IEEE II IP 97

Generalization



IEEE IFIP 97

Precision Metric



$$prec(GT) = 1 - \frac{\sum_{i=1}^{N_A} \sum_{j=1}^{N} \frac{h}{|DGH_{Ai}|}}{|N| \bullet |N_A|}$$

Precision of generalized tables

Eth_0	ZIP_1
Black	0214*
Black	0214*
Black	0213*
Black	0213*
Black	0213*
Black	0213*
White	0213*
White	0213*
White	0213*
White	0213*
White	0213*
White	0213*

Eth_1	ZIP_0
Person	02141
Person	02141
Person	02138
Person	02138
Person	02138
Person	02138
Person	02138
Person	02139
Person	02139
Person	02139
Person	02138
Person	02138

Eth_1	ZIP_1
Person	0214*
Person	0214*
Person	0213*
Person	0213*
Person	0213*
Person	0213*
Person	0213*
Person	0213*
Person	0213*
Person	0213*
Person	0213*
Person	0213*

Eth_0	ZIP_2
Black	021**
Black	021**
Black	021**
Black	021**
Black	021**
Black	021**
White	021**
White	021**
White	021**
White	021**
White	021**
White	021**

$Prec(PT)=1$, $Prec(GT_{[0,1]})=0.83$, $Prec(GT_{[1,0]})=0.75$,
$Prec(GT_{[1,1]})=0.58$, $Prec(GT_{[0,2]})=0.67$

Sample Data

SSN	Ethnicity	Birth	Sex	ZIP	Problem
819181496	Black	09/20/65	m	02141	short of breath
195925972	Black	02/14/65	m	02141	chest pain
902750852	Black	10/23/65	f	02138	hypertension
985820581	Black	08/24/65	f	02138	hypertension
209559459	Black	11/07/64	f	02138	obesity
679392975	Black	12/01/64	f	02138	chest pain
819491049	White	10/23/64	m	02138	chest pain
749201844	White	03/15/65	f	02139	hypertension
985302952	White	08/13/64	m	02139	obesity
874593560	White	05/05/64	m	02139	short of breath
703872052	White	02/13/67	m	02138	chest pain
963963603	White	03/21/67	m	02138	chest pain

HIPAA *safe harbor* results

SSN	Ethnicity	Birth	Sex	ZIP	Problem
	Black	1965	m	021**	short of breath
	Black	1965	m	021**	chest pain
	Black	1965	f	021**	hypertension
	Black	1965	f	021**	hypertension
	Black	1964	f	021**	obesity
	Black	1964	f	021**	chest pain
	White	1964	m	021**	chest pain
	White	1965	f	021**	hypertension
	White	1964	m	021**	obesity
	White	1964	m	021**	short of breath
	White	1967	m	021**	chest pain
	White	1967	m	021**	chest pain

Requires lots of diverse hospitals else 5-digit ZIP can be inferred.

Datafly results

SSN	Ethnicity	Birth	Sex	ZIP	Problem
902387250	Black	1965	m	0214*	short of breath
197150725	Black	1965	m	0214*	chest pain
486062381	Black	1965	f	0213*	hypertension
235978021	Black	1965	f	0213*	hypertension
214684616	Black	1964	f	0213*	obesity
135243442	Black	1964	f	0213*	chest pain
487620561	White	1964	m	0213*	chest pain
259003630	White	1964	m	0213*	obesity
410968224	White	1964	m	0213*	short of breath
664545413	White	1967	m	0213*	chest pain
860424429	White	1967	m	0213*	chest pain

IEEE IFIP 97, NRC 98

k-similar results

SSN	Ethnicity	Birth	Sex	ZIP	Problem
486753948	Black	1965	m	02141	short of breath
758743753	Black	1965	m	02141	chest pain
976483662		1965	f	0213*	hypertension
845796834		1965	f	0213*	hypertension
497306730	Black	1964	f	02138	obesity
730768597	Black	1964	f	02138	chest pain
348993639	Caucasian	1964	m	0213*	chest pain
459734637		1965	f	0213*	hypertension
385692728	Caucasian	1964	m	0213*	obesity
537387873	Caucasian	1964	m	0213*	short of breath
385346532	Caucasian	1967	m	02138	chest pain
349863628	Caucasian	1967	m	02138	chest pain

Idea of *k*-Similar

The idea behind the *k*-Similar algorithm is to grow clusters of tuples that are closest to each until each cluster has at least *k* tuples.

Measuring distance

$$dist(V_{x,y}) = \sum_{i=1}^{n} \frac{d_i}{|DGH_{Ai}|}$$

Z_2={021**}

↑

Z_1={0213*,0214*}

↑

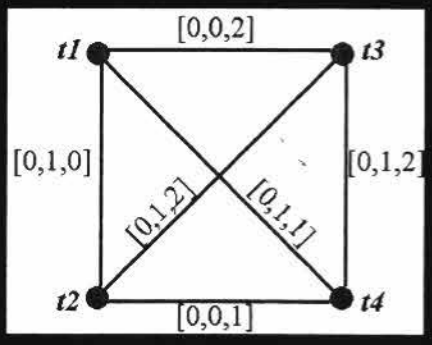
Z_0={02138, 02139, 02141, 02142}

DGH_{Z0}

VGH_{Z0}

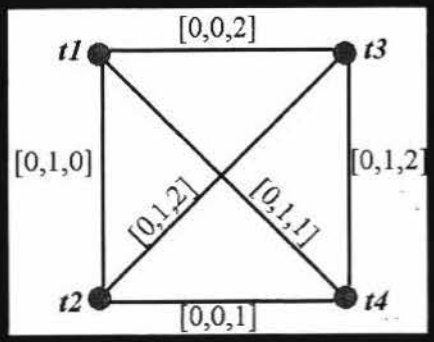
k-Similar: Clique after step 1

	A1	A2	A3
	Home ZIP	Hospital ZIP	Work ZIP
t1	02138	02138	02138
t2	02138	02139	02138
t3	02138	02138	02141
t4	02138	02139	02139



k-Similar: Minimal clusters

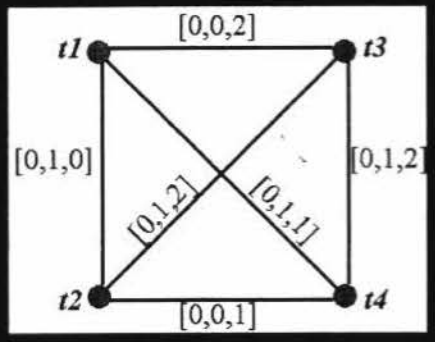
	A1	A2	A3
	Home ZIP	Hospital ZIP	Work ZIP
t1	02138	02138	02138
t2	02138	02139	02138
t3	02138	02138	02141
t4	02138	02139	02139



Mins: {*t1*, *t2*}
{*t2*, *t4*}
{*t1*, *t3*}

k-Similar: Identify isolated minimals

	A1	A2	A3
	Home ZIP	Hospital ZIP	Work ZIP
t1	02138	02138	02138
t2	02138	02139	02138
t3	02138	02138	02141
t4	02138	02139	02139

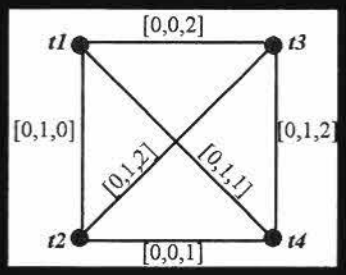


Mins: {*t1*, *t2*}
{*t2*, *t4*}
{*t1*, *t3*}

Solution:
{*t2*, *t4*}
{*t1*, *t3*}

k-Similar: Final result

	A1	A2	A3
	Home ZIP	Hospital ZIP	Work ZIP
t1	02138	02138	02138
t2	02138	02139	02138
t3	02138	02138	02141
t4	02138	02139	02139



	A1	A2	A3
	Home ZIP	Hospital ZIP	Work ZIP
t1	02138	02138	021**
t2	02138	02139	0213*
t3	02138	02138	021**
t4	02138	02139	0213*

This talk

New area and field in CS

Immediate system goals

Early detection

Privacy problem

Computational privacy solutions

- Legal model
- Live demonstration of prototype 1.0

HIPAA results

SSN	Ethnicity	Birth	Sex	ZIP	Problem
	Black	1965	m	021**	short of breath
	Black	1965	m	021**	chest pain
	Black	1965	f	021**	hypertension
	Black	1965	f	021**	hypertension
	Black	1964	f	021**	obesity
	Black	1964	f	021**	chest pain
	White	1964	m	021**	chest pain
	White	1965	f	021**	hypertension
	White	1964	m	021**	obesity
	White	1964	m	021**	short of breath
	White	1967	m	021**	chest pain
	White	1967	m	021**	chest pain

Requires lots of diverse hospitals else 5-digit ZIP can be inferred.

Datafly results

SSN	Ethnicity	Birth	Sex	ZIP	Problem
902387250	Black	1965	m	0214*	short of breath
197150725	Black	1965	m	0214*	chest pain
486062381	Black	1965	f	0213*	hypertension
235978021	Black	1965	f	0213*	hypertension
214684616	Black	1964	f	0213*	obesity
135243442	Black	1964	f	0213*	chest pain
487620561	White	1964	m	0213*	chest pain
259003630	White	1964	m	0213*	obesity
410968224	White	1964	m	0213*	short of breath
664545413	White	1967	m	0213*	chest pain
860424429	White	1967	m	0213*	chest pain

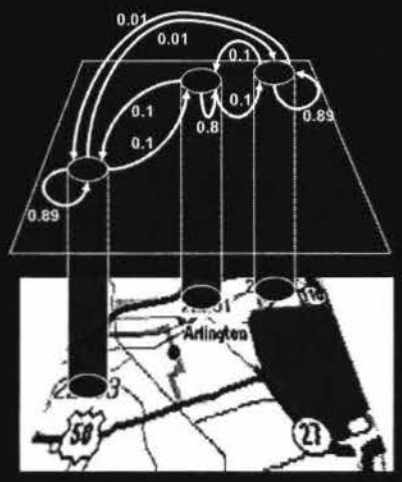
IEEE IHP 97, NKC 98

k-similar results

SSN	Ethnicity	Birth	Sex	ZIP	Problem
486753948	Black	1965	m	02141	short of breath
758743753	Black	1965	m	02141	chest pain
976483662		1965	f	0213*	hypertension
845796834		1965	f	0213*	hypertension
497306730	Black	1964	f	02138	obesity
730768597	Black	1964	f	02138	chest pain
348993639	Caucasian	1964	m	0213*	chest pain
459734637		1965	f	0213*	hypertension
385692728	Caucasian	1964	m	0213*	obesity
537387873	Caucasian	1964	m	0213*	short of breath
385346532	Caucasian	1967	m	02138	chest pain
349863628	Caucasian	1967	m	02138	chest pain

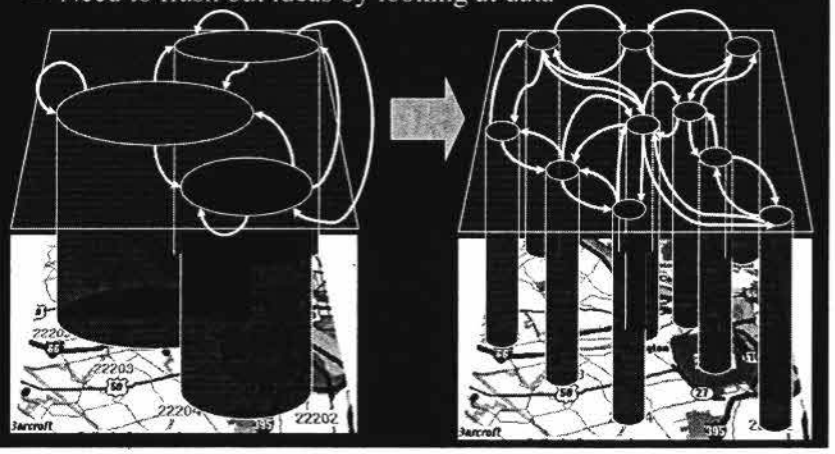
Model Interactions Between Clusters

- Contact happens both within a cluster, and between clusters
  - Strong geographical component, both work and residence
  - Model interactions between clusters
- Accurate modeling requires small clusters
  - Assume homogenous mixing within a cluster
  - Most accurate models consider individuals or households
  - Data / privacy implications

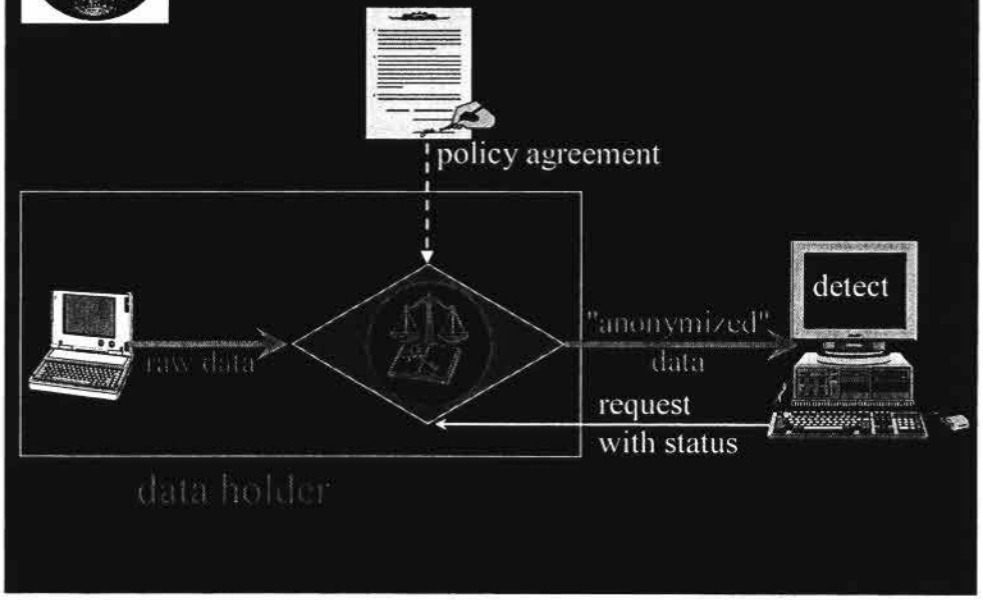


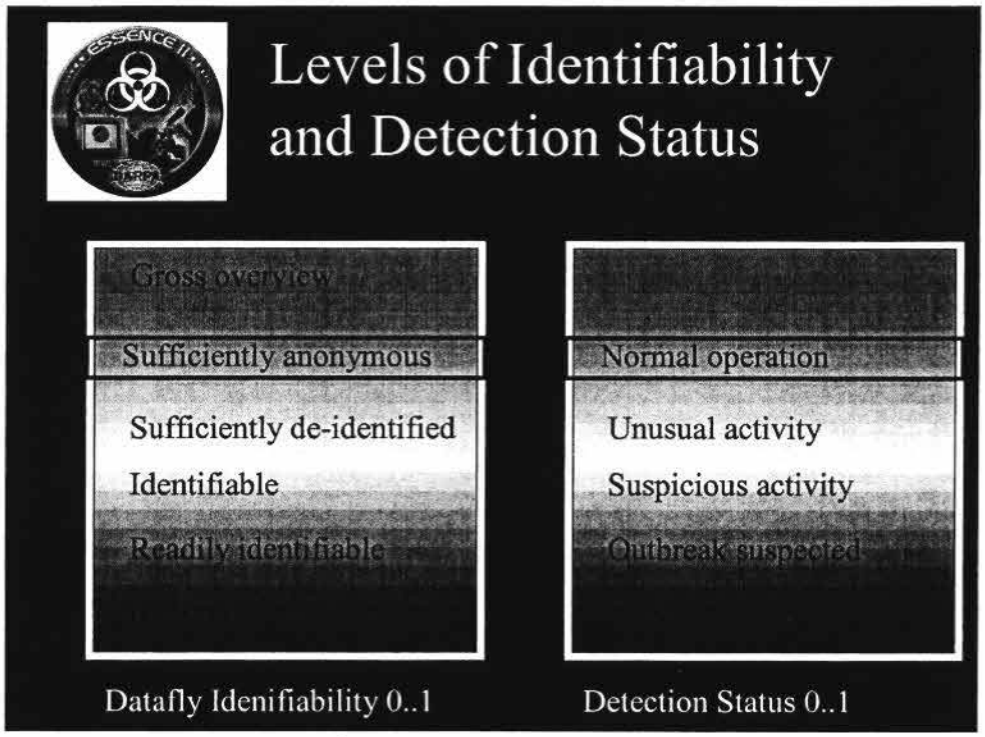
Dynamically Augment the Model When Surveillance Detects Possible Attack

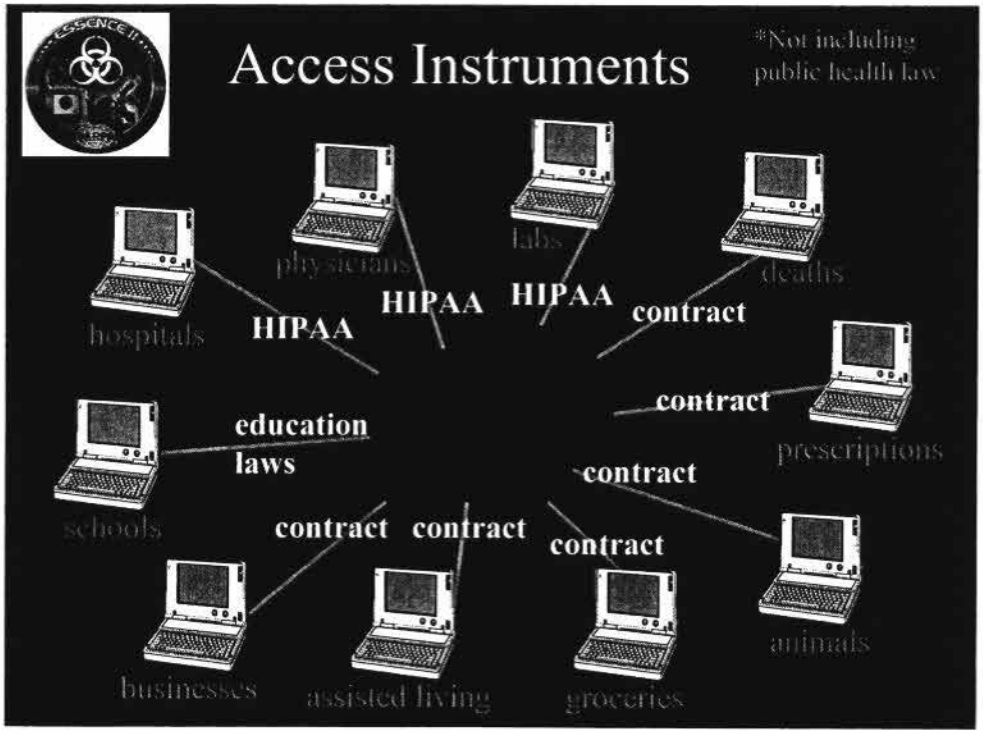
- Lower the privacy threshold when potential attack detected
  - But how often, how quickly, to what level?
  - Can we take advantage of disease-specific processing?
  - Need to flush out ideas by looking at data

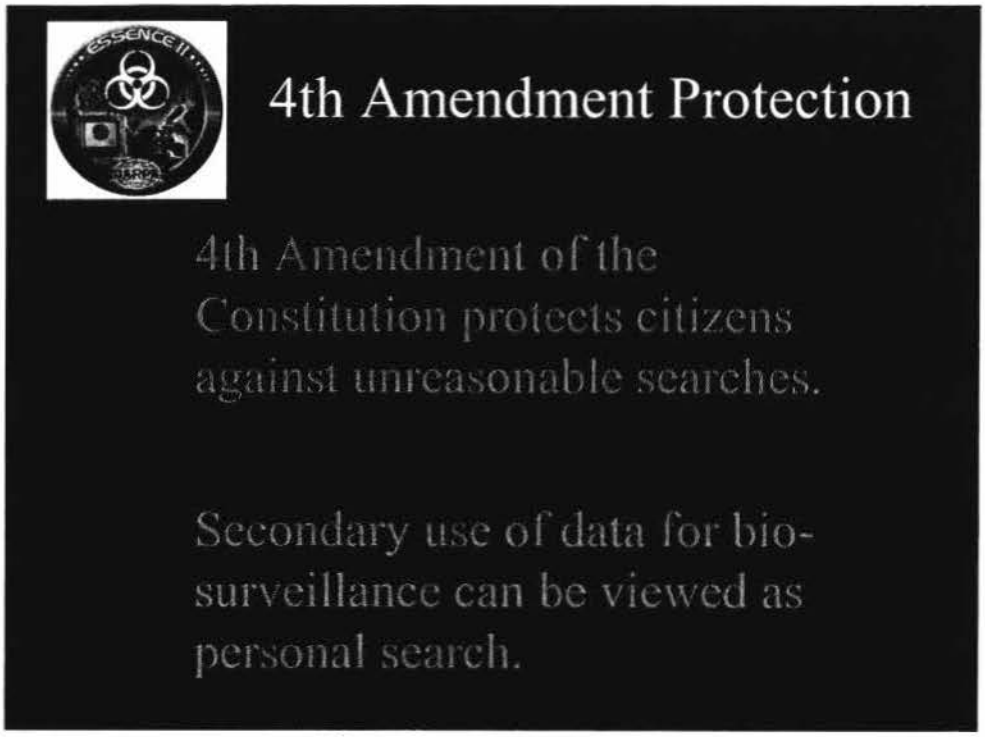


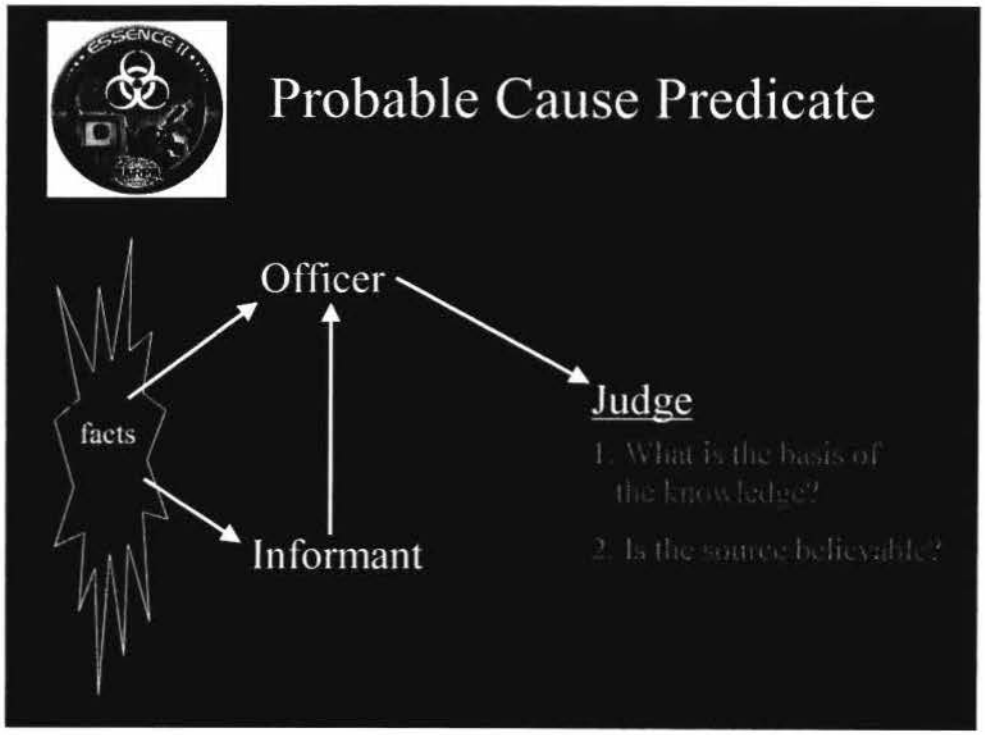
◇Automated Privacy Module

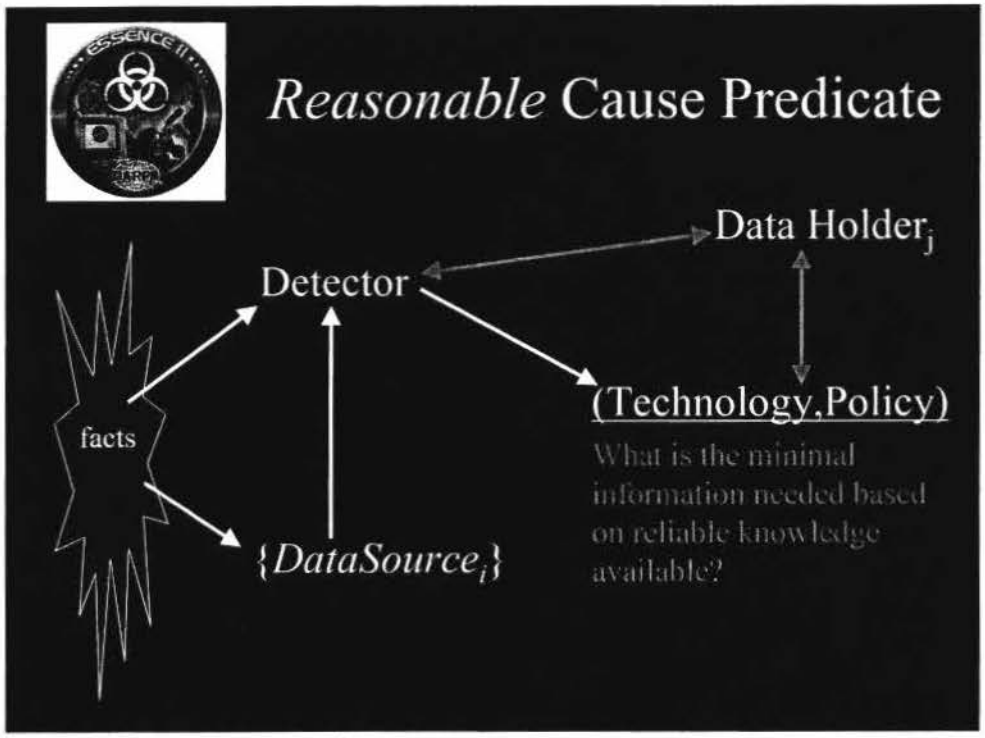



Levels of Identifiability and Detection Status
Gross overview
Sufficiently anonymous
Sufficiently de-identified
Identifiable
Readily identifiable
Datafly Idenifiability 0..1
Normal operation
Unusual activity
Suspicious activity
Outbreak suspected
Detection Status 0..1


Access Instruments
*Not including public health law
physicians
labs
deaths
hospitals
HIPAA
HIPAA
HIPAA
contract
contract
education laws
contract
schools
contract
contract
contract
prescriptions
businesses
assisted living
groceries
animals


ESSENCE II
4th Amendment Protection
4th Amendment of the Constitution protects citizens against unreasonable searches.
Secondary use of data for bio-surveillance can be viewed as personal search.


ESSENCE II
Probable Cause Predicate
Officer
facts
Informant
Judge
1. What is the basis of the knowledge?
2. Is the source believable?



Access policy as basis for metrics

$$P_R = (D, I, E, F)$$

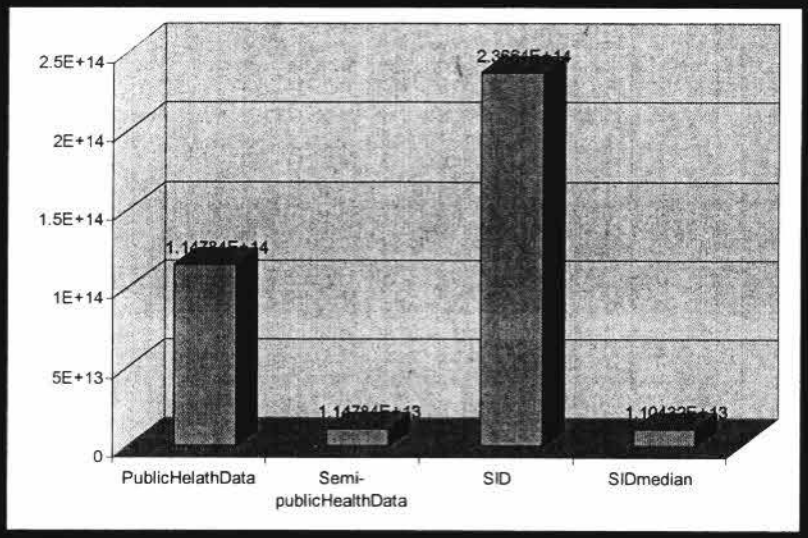
Gross DSP of a release $GI_S(P_R) = c * D * E$

Gross Maximal Risk *gross maximal risk* $= I * E$

Long-term DSP $= \Sigma GI_S^i(P) = D * (E_1 + F_1 + F_2 + \ldots + F_{n-1})$

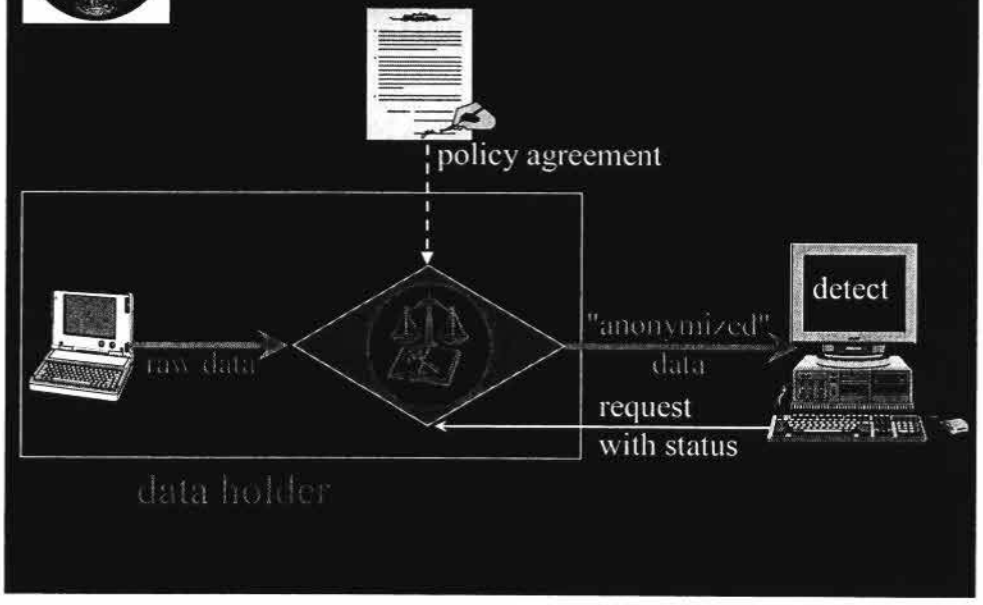
Total storage *total storage*$(P_R, t_1, t_2) = AI_S(P_R, t_1, t_2)_i * I$

Gross Maximal Risk = $I * E$



Many more people in SID considered identifiable!

◇Automated Privacy Module



Immediate System Goals

1. Detects outbreaks or attacks early by continuously observing the general population
2. Provides scientific guarantees of privacy protection; automated adherence to HIPAA
3. Models "probable cause" legal predicate to lower anonymity constraints as conditions warrant
4. Results in automated releases of *useful* data that are minimally invasive to privacy

Metrics for next steps

1. ROC curves for detection algorithms on data with varying levels of anonymity, both in isolated data sources as well as data sources in combination.
2. Policy language expressiveness and correctness measured and check-and-balance performance over time reviewed.

This talk

New area and field in CS

Immediate system goals

Early detection

Privacy problem

Computational privacy solutions

Legal model

- Live demonstration of prototype 1.0

1.a.7

Essence II Architecture

Biosurveillance Kick-off
October 11, 2001

Steve Langs
JHU/APL

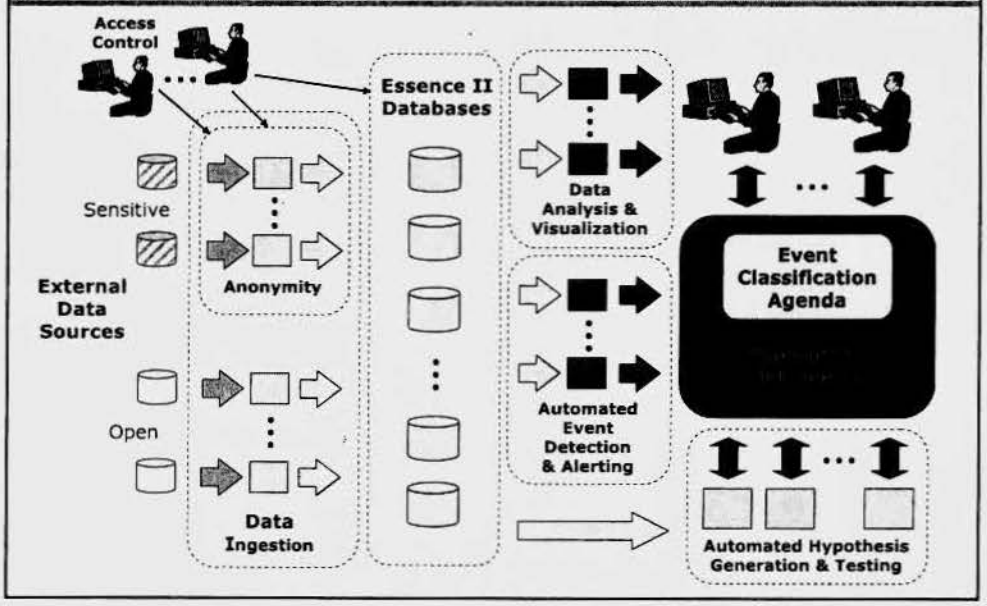
Architecture Challenges

- Support focused, agile event investigation.
- Coordination of diverse problem-solving resources.
- Continuous operation.
- Heterogeneous data sources externally controlled.
- Data anonymity and security.
- Large data volume, asynchronous data arrival.
- Collaboration among investigators and agencies.

Architecture Approach

- Mixed-initiative, blackboard-based abduction control.
- Agent-based software infrastructure.
- Autonomous detection agents, semi-autonomous reasoning agents.
- Knowledge-level component integration.
- Anonymity control at data ingestion.
- Distributed databases.
- Security - PKI, SSL, firewall, VPN.

Architecture Overview



1. a. 8 c

ESSENCE II
Classification Methodology

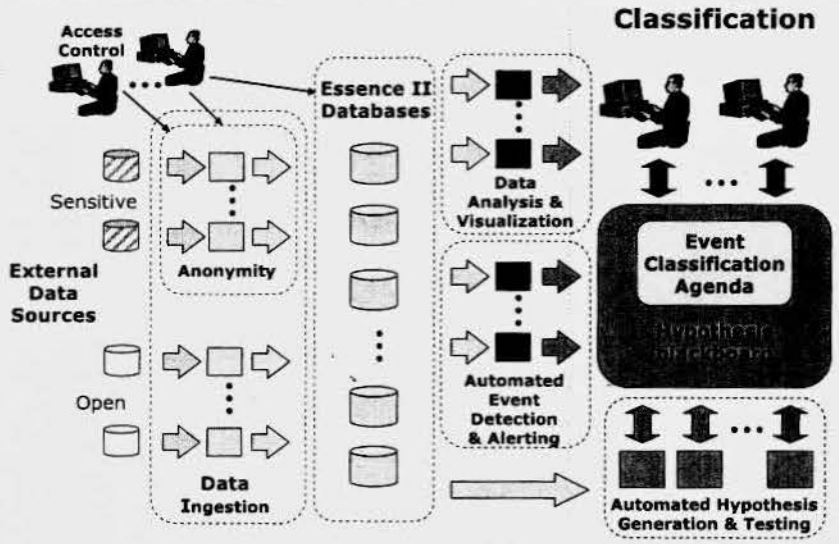
Bio-Surveillance Program Kick-Off Meeting
October 11, 2001

David Gunning

CYCORP

Architecture Overview

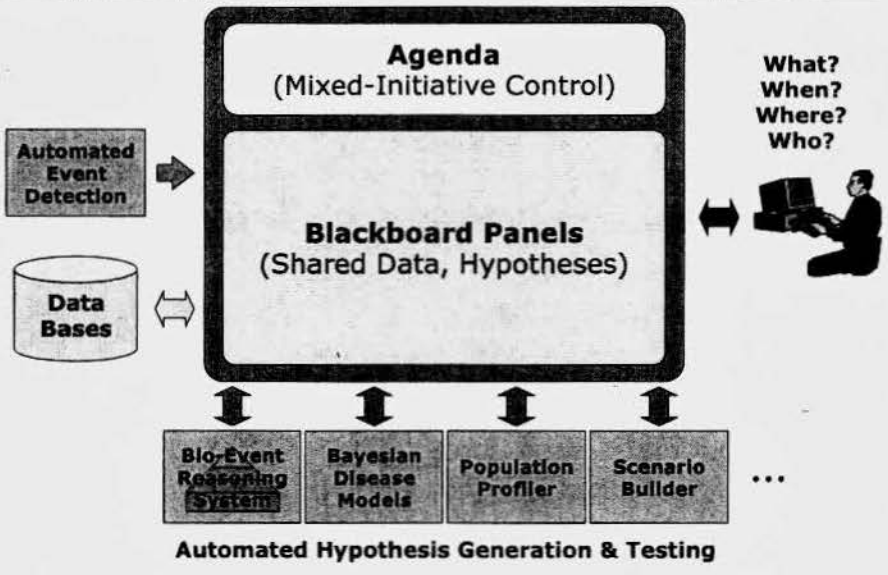
CYCORP



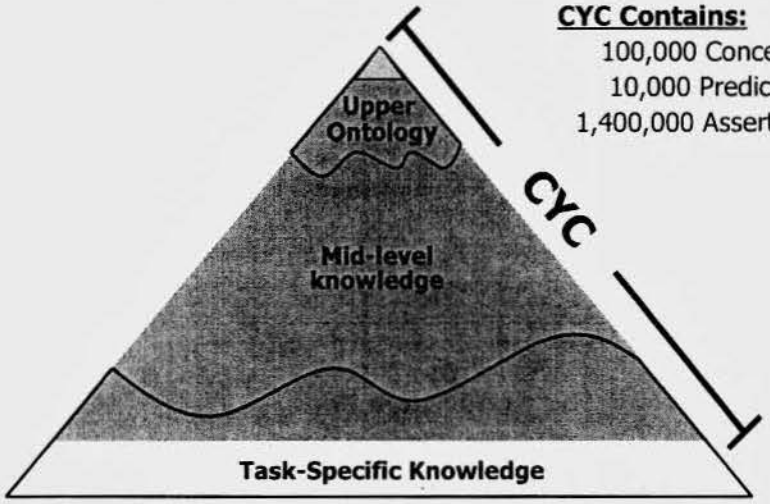
Event Classification

- What disease is active?
- What bio-agent is the cause?
- Where was the agent released?
- When was it released?
- Who has been affected?

Classification Approach



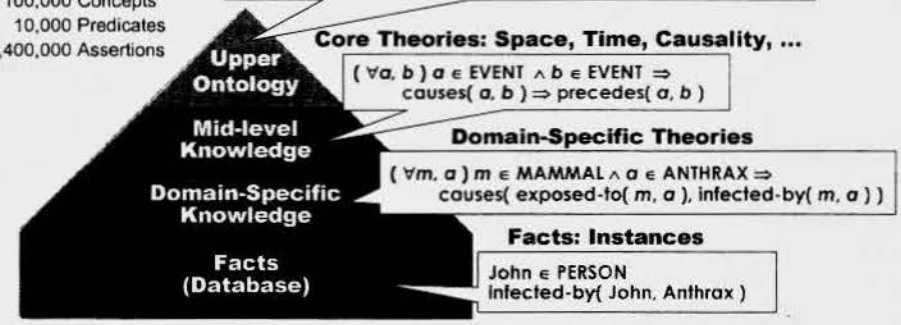
Cyc Knowledge Base



CYC Contains:
100,000 Concepts
10,000 Predicates
1,400,000 Assertions

What is in the Cyc Ontology?

CYC Contains:
100,000 Concepts
10,000 Predicates
1,400,000 Assertions



Upper Ontology: Abstract Concepts
EVENT ⊃ TEMPORAL-THING ⊃ PARTIALLY-TANGIBLE-THING

Core Theories: Space, Time, Causality, ...
(∀*a*, *b*) *a* ∈ EVENT ∧ *b* ∈ EVENT ⇒
causes(*a*, *b*) ⇒ precedes(*a*, *b*)

Domain-Specific Theories
(∀*m*, *a*) *m* ∈ MAMMAL ∧ *a* ∈ ANTHRAX ⇒
causes(exposed-to(*m*, *a*), infected-by(*m*, *a*))

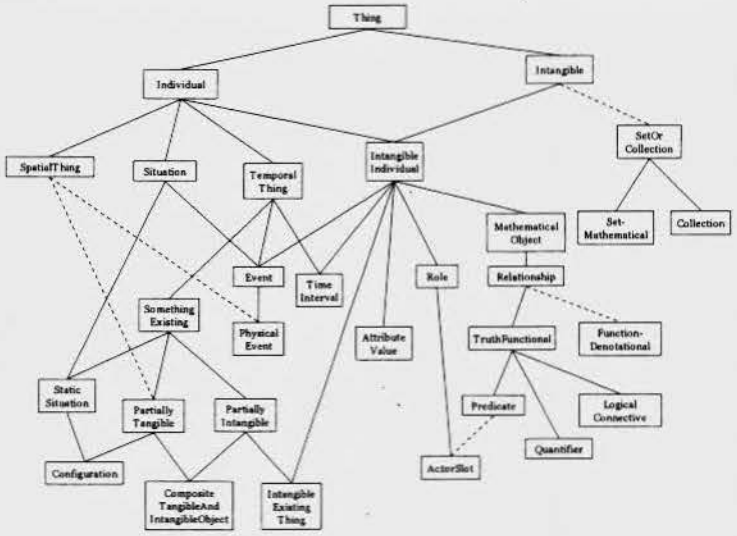
Facts: Instances
John ∈ PERSON
infected-by(John, Anthrax)

Concepts (Sets):
SUBSET ⊃ CONCEPT
Instances (Members):
Instance ∈ CONCEPT

Quantified Variables:
(∀*variable*) or **(∃ *variable*)**
Logical Operators:
¬, ∧, ∨, ⇒, ⇔

Predicates (True or False):
predicate(arg1, arg2 **)**
Functions (Return a value):
y = **function(** arg1, arg2 **)**

Cyc Upper Ontology



Cyc Mid-Level Knowledge

- Space
- Spatial Relations
- Time
- Quantities
- Physical Objects
- Parts of Objects
- Movement
- Paths, Networks
- Basic Physics
- Basic Chemistry
- Basic Biology
- Changes of State
- Transformations
- Materials

- Events
- Context
- Actions
- Actors
- Agreements
- Goals
- Skills
- Capabilities
- Intentions
- Scripts
- Actor Roles
- Scenes
- Plans
- Processes

- People
- Human Anatomy
- Human Physiology
- Mental Attributes
- Emotions
- Social Beings
- Social Events
- Human Activities
- Organizations
- Professions
- Business
- Geography
- Nations
- Cultures

- Conceptual Works
  - Books
  - Documents
  - Software
  - Databases
  - Web Pages
  - Plays
  - Movies
- Devices
- Buildings
- Construction
- Clothing
- Transportation

Domain-Specific Knowledge
(currently in Cyc)

CYCORP

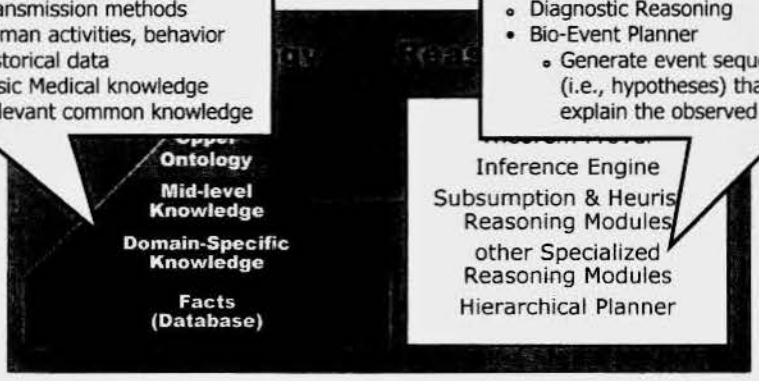
- **Biological Agents**
  - Cyc knows that biological agents can be biological toxins, or fungal, bacterial, or viral agents.
- **Toxins**
  - Cyc knows about aflatoxin, botulinum, ricin, batrachotoxin, saxitoxin, and various types of venom.
- **Bacteria**
  - Cyc knows about 100+ kinds of bacteria, including anthrax.
- **Viruses**
  - Cyc knows information about Ebola, HIV, Hepatitis (A, B, C, D, and E), and smallpox. Cyc also knows that some viruses are genetically engineered.
- **Biological Weapons**
  - Cyc knows a great deal about the collection of weapons (e.g., bombs, warheads and aerosol canisters) and how each is appropriate for dispersing different kinds of biological agents.

Bio-Event Reasoning System

CYCORP

Bio-Surveillance Ontology:
- Bio-agents, effects, symptoms
- Disease epidemiology
- Transmission methods
- Human activities, behavior
- Historical data
- Basic Medical knowledge
- Relevant common knowledge

Bio-Surveillance Reasoning:
- Question Answering
- Pro/Con Hypothesis Support
- Diagnostic Reasoning
- Bio-Event Planner
  - Generate event sequences (i.e., hypotheses) that explain the observed data



Cycorp Tasks for ESSENCE II

- **Phase I (FY01-02)**
  - Develop Initial Bio-Surveillance Ontology
  - Develop Prototype Bio-Event Reasoner
  - Implement Event Classification Blackboard
- **Phase II (FY03)**
  - Complete Bio-Surveillance Ontology and KB
  - Complete Bio-Event Reasoner
- **Phase III (FY04)**
  - Support Transition and System Implementation
  - Tailor Ontology and Bio-Event Reasoner

TITLE PAGE

Briefing Materials concerning TIA BIOALERT 1

FOLDER 2.d TOC 8

EVALUATIONS

- 2.d.1 Problem Space/Evaluation
    - Pitt-CMU
- 2.d.2 Bio-Surveillance Evaluation
    - Potomac Institute for Policy Studies

Genuflecting before the Problem Space

Condition	BIOTERRORISM LIST	REPORTABLE LIST
Anthrax	DTRA list, A-list	R
Cholera	DTRA list, A-list	R
Plague (pneumonic)	DTRA list, A-list	R
Smallpox	DTRA list, A-list	R
Tularemia (inhaled)	DTRA list, A-list	R
Hemorrhagic fever virus	DTRA list, A-list	
Ricin toxin	DTRA list, A-list	
Botulism	DTRA list, A-list	R
Brucellosis	DTRA list, A-list	R
Carbon monoxide poisoning	A-list	R
Infectious diarrhea syndrome		
Acute upper respiratory syndrome		
Typhoid	DTRA list	R
Hantaviral diseases		R
Campylobacter		R
Cryptosporidiosis		R
E. coli 0157:H7		R
Giardiasis		R
Influenza		R
Legionnaires' disease		R
Neisseria meningitis in blood or csf		R
Salmonellosis		R
Shigellosis	DTRA list	R
Yersinia		R

x

Route
Air
Food
Water
Mail
...

Contagious person-to-person aerosol: ***rubella, zoster***

Premonitory release: Boca Anthrax

Foodborne: ***Salmonella sp.***

Waterborne: ***Cryptosporidium***

Vectorborne: ***West Nile, Malaria, Lyme***

Continuous release of bioaerosol: ***Legionella, pollen***

Building contamination: ***CO***

Sexually transmitted: ***HIV, N. gonorrhea, Hep C***

Large scale bioaerosol : ***NO OUTBREAKS***

Dato, Wagner, et al. 2001

CMU-Pitt

Evaluation

Outline

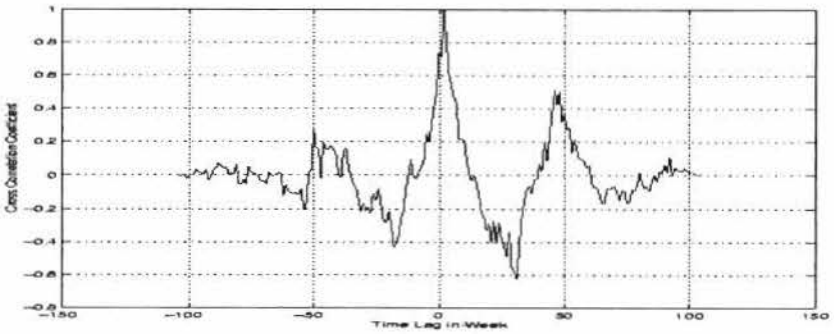
Identifying leading indicators (for each threat)

Identifying best algorithms (for each threat)

CMU-Pitt

I. Identifying Leading Indicators: Directly measuring earliness of signal

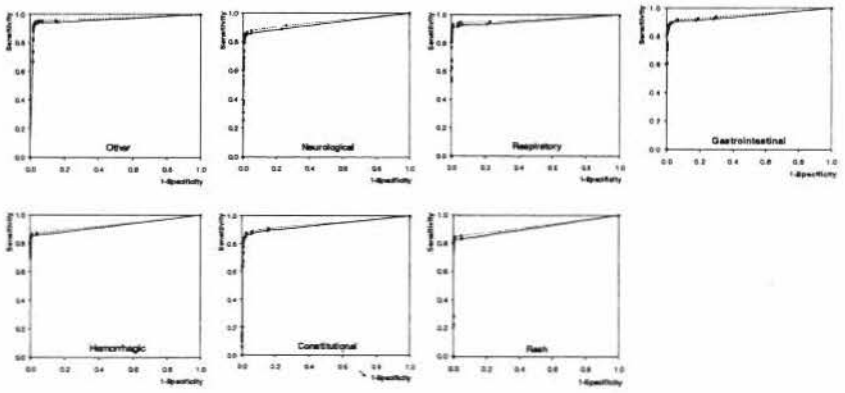
- Cross Correlation coefficient



Cross correlation between weekly counts of respiratory and weekly and Pneumonia and Influenza Deaths

CMU-Pitt

Does the Signal Mean What We Think it Means?



CMU-Pitt

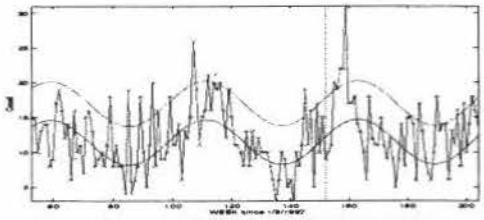
Does the Signal Mean What We Think it Means?

- 1425 randomly selected patients
- 22 of their ED reports described *acute infectious GI syndrome*

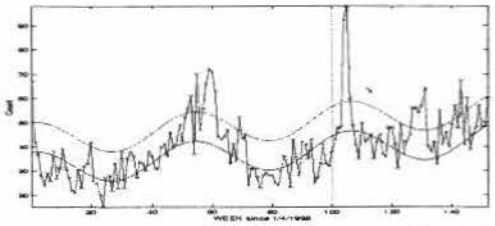
	ICD-9 Admit Diagnoses	Free-text Chief Complaints
Sensitivity	0.32	0.63
Specificity	0.99	0.94

CMU-Pitt

Identifying Leading Indicators: Measuring (Algorithm + Signal)



Gold Standard: P&I Deaths Pittsburgh 1/98 to 48/00.



Weekly counts of respiratory ICD-9 codes from 1/98 to 48/00.

CMU-Pitt

Identifying Leading Indicator: Serfling Method and Influenza

Signal	Sens.	Spec./*PPV	Timeliness (weeks)
Emergency home visits	81	75	-1.6
Sick-leave reported to national health service	29	74	-1.4
Sick-leave reported to GP	76	65	-1.2
Sick-leave reported by companies	74	67	0.0
Sentinel GP visits	67	72	-1.2
Sentinel GP visits due to ILI	69	69	-0.4
Sentinel pediatrician visits	64	65	-1.7
Hospital fatality	47	82	-1.0
Influenza-related drug consumption	57	65	-1.2
Sentinel GP overall activity	57	63	-1.2
Sentinel pediatrician overall activitiy	47	67	1.3
Emergency room respiratory chief complaints	100	50	-1.0
Emergency room viral chief complaints	100	25	-1.0

Quenel et al. 1994 *Int J Epid* 23:849-55
Tsui, Wagner et al. *JAMIA* 2002

CMU-Pitt

II. Evaluating Algorithms

- **How well does the algorithm work under ideal conditions?** (i.e. when the data are generated in a way consistent with all the model assumptions).
- **Manual-insertion proxy**
- **How well does it work under simulated, non-ideal conditions?** These tests will use simulated data, generated by a model other than the one being tested, in which ground truth is available and so we can directly measure detection performance.
- **How well does it work on real data?**

CMU-Pitt

8

2.0.2

Bio-Surveillance Evaluation

Tinman Draft

David Siegrist
Potomac Institute for Policy Studies
June 4, 2002

1

Biosurveillance Program Goals

- **Early Detection of Bio Attack**
  - Differentiate Deliberate Release Despite Non-Specific Symptoms
  - Identify Leading Indicators of Disease
  - Characterize Nontraditional Data

- **Rapid Pathogen Identification**
  - Determine Probable Deliberate Infection
  - Identify Disease
    - Clinical Data Augmentation Acceptable
- **Attack Characterization**
  - Identify Source of Outbreak
  - Determine Weapon Laydown Area and Duration
  - Identify Those Who Need Treatment
  - Project Their Overall Likely Medical Needs

2

Scope of Biosurveillance Problem
Any Venue Type Could Have Aerosol or Vector Attacks

- **Metro Areas:** JHU; Veridian; Mellon Pitt
- **Neighborhood/Mall:** JHU; Veridian; Mellon Pitt
- **Military Bases:** IBM; Veridian, JHU
- **Special Events (eg, Stadiums):** JHU; Veridian; Mellon Pitt; IBM
- **Individual Buildings:** DARPA Immune Building Program
- ***An Operationalized Program Goal***: Demo/Evaluate the Three Program Goals Against Each Box in the Matrix
- Projects Could Select Scenarios to Demonstrate Their Operational Advantages

	Possible Venues			
Challenge Modes	SPECIAL EVENT	SINGLE CAMPUS	NEIGHBORHOOD	LARGE AREA
AEROSOL	VH/VI	Hard/IMP	VH/IMP	HARD/CRIT
VECTOR FOOD/BEV	VH/IMP	Hard/IMP	VH/IMP	VH/VI
VECTOR HUMAN	VH/CRIT?	Hard/IMP	VH/VI	EH/CRIT
VECTOR ANIMAL	VH/VI	Hard/IMP	VH/IMP	VH/VI
VECTOR MECHANICL	VH/IMP	Hard/IMP	VH/IMP	EH/IMP

Degree of Difficulty: Extremely Hard (EH); Very Hard (VH); Hard (HARD).
Degree of Importance: Critical (Crit); Very Important (VI); Important (Imp).
NOTE: These Characterizations Suggestive and Subjective ONLY

3

Evaluation Progression
Primary Program Goal: Early Detection

- **Overall BioSurveillance PROGRAM (vs Project) Evaluation Goal:**
  - Early Detection, Path ID & Attack Characterization Demo'd & Eval'd in Every Box on Matrix
  - ***Goodness: Sensitivity, Specificity, Spatial/Temporal Coverage/Granularity, and Privacy***
- **Aug, 2002 Evaluation.**
  - Use Largely Synthetic Data for "Easier" Scenario
  - Early Detection of Large Area Aerosol, Min 25% of Potential Deaths Saved
- **Later: Earlier Detection, Rapid Pathogen Identification and Attack Characterization**
  - Stress Disease Models, Nontraditional Data
  - Determine Source; Who Has Been Exposed; Project Medical Needs
  - Automated, Dynamic Privacy Assurance
  - Additional Scenarios from Attack Matrix

4

Overall Evaluation Progression Overview

- **2002 Goal: 25% Deaths Saved**
  - Simulated Data; Some Non-Traditional Data Sources
  - Privacy Supported
- **2003 Goal: 50% Deaths Saved**
  - Better Simulated Data; Some Real Data; Significant Non-Traditional Data Sources
  - Include Accepted Disease Model for Military Pathogens. Differentiate Flu from Them.
  - Fully Characterize Attack. Project Med Resources Needed.
  - Privacy Assured
- **2004 Goal: 75% Deaths Saved**
  - Predominantly Real Data; Exchange Data or Analytics with Others; Real Time Runs
  - Identify Military-type Pathogen.
  - Possible Characterization: Identify Location of Smallpox Victim for Previous 3 Weeks
  - Privacy Assured Automatically

5

FY02 Evaluation

- **Data Evaluation**
  - Teams Work Individually to Best Characterize Their NT and Traditional Data
  - Document Their Work to DARPA and Potomac Institute
  - Provide Results to CMU to Improve BioWAR Simulation
  - Data/Ontology Working Group to Facilitate Process
- **Algorithm Evaluation**
  - BioWAR Provides Synthetic Data Streams for Common Algorithm Experiment
  - Looking for Sensitivity, Specificity, Granularity, Coverage and Privacy
  - Process Overseen by Evaluation Working Group

6

Synthetic Data for Common Challenge

- **Common Challenge Evaluation Mandated by PM**
- **Teams Appear Restricted from Sharing Actual Data (Proprietary)**
  - Realistic Comparison of Systems Possible Given a Sharing of Analytics
  - Necessitates Synthetic Data to Generate ROC Curves
  - Teams Using Data They've Generated/Trained on Less Credible for Evaluation
- **Near Term Approach for August: Use CMU BioWAR to Generate Data**
  - CMU Has Agreed to Provide Same
  - Interaction with Teams on Evaluation Prep Promises Improved Simulation as Well.

Detection Evaluation Progression Primary Program Goal: Early Detection

- **Aug, 2002 Algorithm Evaluation.**
  - 1. Start with Early Detection of Large Area Aerosol Anthrax Release
    - Canonical Scenario with Highest Impact, Potential for Detection
    - Stresses Anomaly Detection and Fusion Algorithms
    - Use Largely Simulated Data Superimposed on Historical Background Rates
      - Simulated Data Should NOT Be Generated by Same Algorithms that Model Uses
  - 2. Additional Evaluation: Smaller Size Attack Stressing Granularity
  - Non-Traditional Data Tested as Well if Possible
    - Needs Show Empirical Relation to Disease & Realistic Simulation Before With and WO Evaluation
  - ROC Curve Sensitivity vs Specificity to be Measured on Multiple Runs
  - Data Privacy Assurance Also Evaluated
    - Scenarios Tried With and Without Active Privacy Measures to Demonstrate Effectiveness.

Some Questions About Using BioWAR

- **Can and Will BioWAR Capture Different Demographic Parameters for Different Regions?**
- **What Mechanisms Are Used to Assign Behaviors?**
- **Does BioWAR Use a Particular Disease Model?**
- **To What Extent Has BioWAR Been Validated?**
- **Does it Accurately Represent Non-traditional Data?**
  - Can Particular Types be Instantiated before August?
- **Perception of Fairness.**
  - Is Choice of BioWAR Seen as Preferential to Any Team including Mellon Pitt?
    - [NOTE: No One Has Claimed Unfairness as of 6-13-02, so Pitt Included]

Actions to Use BioWAR for Evaluation

- **Get Data Source Types and Demographics to CMU to Instantiate**
  - Need List
- **Teams to Work with CMU to Make Simulation as Realistic as Possible**
- **Familiarization with its Data Outputs**
- **CMU to Supply Executable Simulation, Sample Data Sets ASAP**

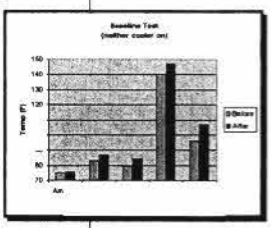
Detection Factors

- Data Source Contains Valid Early Indicator of Outbreak
  - Use Influenza or Smaller Outbreaks of Infectious Diseases As Tests
- Data Added in Simulation Is Realistic in Amount and Distribution
  - Taken from Historical Data
- Percentage Coverage of the Designated Area
  - Use Reference Sources to Determine Approximate Population Covered
- Data Fusion
  - Relating Anomalous Heterogeneous Data in a Timely Fashion
- Timeliness in Acquiring Data
  - Time the Data Is Recorded/Received After the Event
- Data-specific Response time, Minimum intensity (P_d vs SNR) and ROC (P_d vs P_{fa})
  - Use Simulated Outbreak Events Inserted in Live Database
- Probability of Recognizing Outbreak Region within Coverage Area (P_r vs SNR_{Min})
  - Simulated Output from Larger Scenario. Highest Level of Resolution Attained.
- Event Detector Response Time. Sensitivity (P_d vs SNR); ROC (P_d vs P_{fa})
  - Simulated Output from Larger Scenario
- Security & Privacy Protection, Disclosure Control Effect of Anonymity on Performance

11

August Evaluation Scenario Process Flow

- **Common Anthrax Scenario with Significant Aerosol Source**
  - Data Includes "Near Misses" as Well as Actual Outbreaks. FPs Detract Significantly.
- **Need to Establish Baseline for Early Detection. Possibilities:**
  - Zero Baseline: Goal Is to Save >25% of Total Deaths Projected from Simulated Scenario
  - Or Some Arbitrary Figure: X% of Deaths or X Days After Release
- **In Evaluation, Performers Make the "Call," Asserting Outbreak**
  - In Operation, Earlier "Alert" Will Be Used to Signal Epidemiologist
  - Outbreak Determination "Call" Is Then Assessed for Accuracy and Timeliness
    - Some Time Constant Needs Added to Determine % Deaths Saved
- **Response Assumptions Need Agreed Before. Recommendation:**
  - 8 Hours to Confirm Diagnosis; 8 Hours to Move Stockpile; 12 Hours to Break it Down and Distribute It; 8 Hours Before Effective Immunity.
  - These Are Fairly Optimistic Assumptions
  - Assume NO Lives Saved through Subsequent Medical Treatment
  - Overall: Suggest Adding 36 Hours to Performer "Call" to Determine % Saved.



12

Tinman Draft "Scoring" Suggestive of Relative Importance

- Use Non-Traditional Data Sources: Pass/Fail
- Make Early Determination of Disease Outbreak: 25 Pts.
- Minimal False Alarms: 15 Pts.
- Support Granularity (Spatial & Temporal): 20 Pts.
- Multiple Coverage of Large Target Area: 10 Pts.
- Differentiate Natural from Unnatural Outbreak: 10 Pts.
- Technology Extensible to Other Regions 5 Pts.
- Uses Advanced Technology 5 Pts.
- Active Personal Privacy Assurance: 10 Pts.

100 Pts.

- Depersonalization without Anonymity: Minus 20 Pts.
- Release Patient Information on Internet: Minus 50 Pts.
- Violate HIPAA: Possible Prison Term and Fine/ False Negative: ++ Morbidity and Mortality

Note: ROC Curve May Summarize Sensitivity/Specificity "Goodness" Using Simulation

13

BACKGROUND

14

Considerations in Using Real & Synthetic Data
From Dr. Howard Burkom, JHU

- **Use Authentic Data for Outbreaks of Flu and Uncommon Diseases**
  - No Data Modeling Is Required
  - However, Influenza Event Is Not a Point Source, Common Vehicle Outbreak
  - There Are Very Few Sufficiently Large Data Sets Containing Known Outbreaks
    - Needs Dates of Positive Laboratory Tests for Influenza. Perform Sensitivity Analyses.
    - Accurate P_d May Not Be Obtained, but Good Estimates of False Alarms Possible
  - The Use of Authentic Data Is Complicated by Both Privacy and Proprietary Issues
- **Simulating the Effects of an Outbreak on Authentic Background Data**
  - Requires Acceptance of the Underlying Model
  - Permits Statistical Analysis of Large Numbers of Synthetic Cases
  - Limit Privacy/Proprietary Issues by Randomizing Background

15

Simulated Data and Detecting an Outbreak
From Dr. Howard Burkom, JHU

- Generate Epicurves of Daily Onset of Symptoms in Victims
  - Use Accepted Sartwell Lognormal Model for the Distribution of Incubation Periods
    - Need Reasonable Estimates for the 2 Model Parameters for Military Pathogens
- Estimate the Effect of a Given Number of Clinical Cases on the Data
  - Demographically Differentiated Human Behavior. Difficult.
- Estimate the Geographical Distribution of the Clinical Cases
- Form a Testbed of Background Data. (eg., Inaugural)
- Exercise Algorithms on Backgrounds for Plausible Threshold Levels, P_{FA}
- Compute Sets of Simulated Outbreaks for Injection Into Background
- Parametric Exercise of Algorithms on the Set of Simulated Outbreaks Injected Into Real Data. Generate ROC Curves.

16

TITLE PAGE

FOLDER 2.e TOC 9

Briefing Materials concerning TIA BIOALERT 1/Folder 2.e

SIMULATED DATA

- 2.e.1 CDC Influenza Regions
    - CDC
- 2.e.2 Evaluation-Use of Authentic + Simulated Data
    - Veridian-Stanford
    - CMU

2.e.5

Evaluation – Use of Authentic and Simulated Data

Veridian-Stanford

(CMU)

June 5, 2002

D. Siegrist Questions

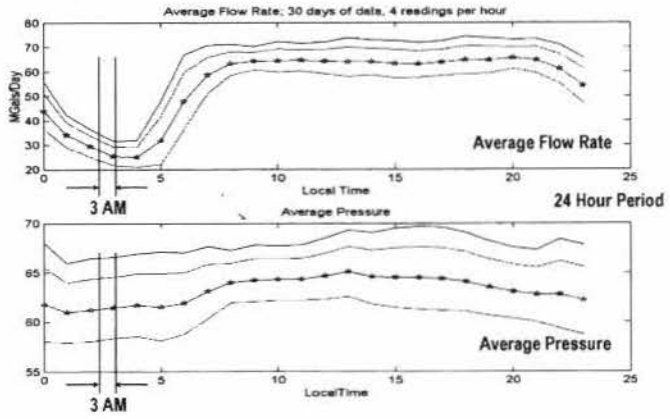
- Use Authentic Data for Outbreaks of Flu and Uncommon Diseases?
- Simulate the Effects of an Outbreak on Authentic Background Data?
- Generating Epicurves of Daily Onset of Symptoms in Victims
- Estimating the Effect of a Given Number of Clinical Cases on the Data
- Demographically and Geographically Differentiating Human Behavior

Validation Approaches

- **Techniques applicable to both authentic and simulated data**
- **Effective test of influences on individual and distributed sets of sources**
- **Develop and assess quality of normalcy models**
  - Applicable to validating authentic data
  - Compare effectiveness of characterizing bio-agent influences against normalcy models
- **Compare bio-agent based Kalman anomaly tracks against normalcy Kalman track conditions**
  - Compares performance between both filters
- **Plot specificity vs sensitivity per source or over sets of sources**
  - ROC curve
- **Two sets of examples**
  - Normalcy development
    - Water flow analysis
  - Spatial-Temporal Kalman filter (Anomaly vs. normal tracking)
    - School Absentees
    - Aspirin Sales

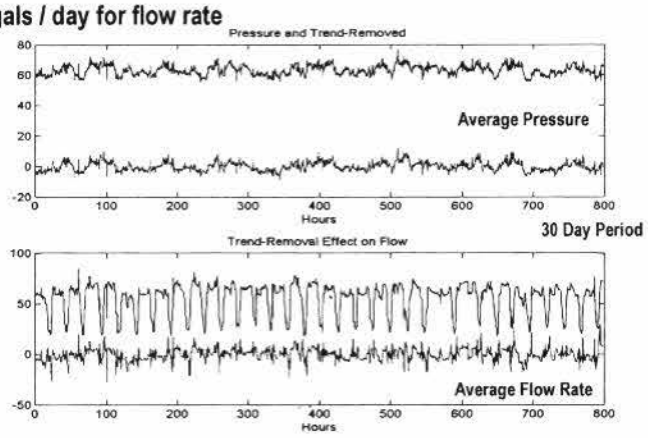
Water Flow Analysis (1) Normalcy Development

- **Determining normal characteristics**
- **Hourly flow rate and pressure over a 24-hour period, averaged over 30 days.**
- **4 readings per hour, with 1 std interval (~+/-6 MG/Day) and upper 95% prediction limit**



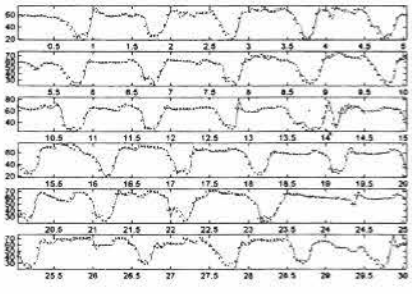
Water Flow Analysis (2) Normalcy Development

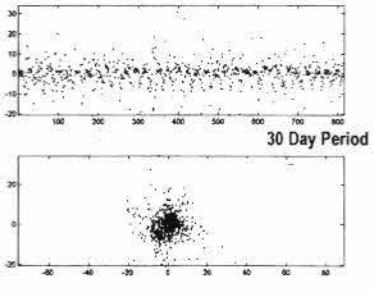
- Blue – Raw series
- Pressure and Flow rate subtracting average diurnal cycle
- Green residual is zero-mean, unimodal series of numbers;
- Std~6 Mgals / day for flow rate



Water Flow Analysis (3) Normalcy Development

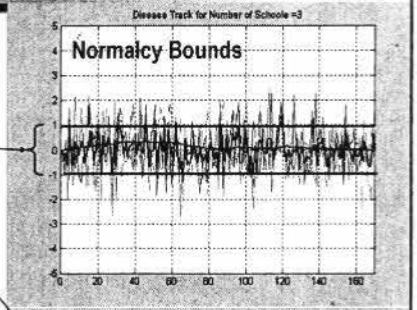
- Validating Random Fluctuations
  - Water flow rate series: Autoregressive (order 2) treatment of both local and 24 hour variability
  - Forecast residual series and Residual lag-1 correlation: σ~6 Mgals/Day



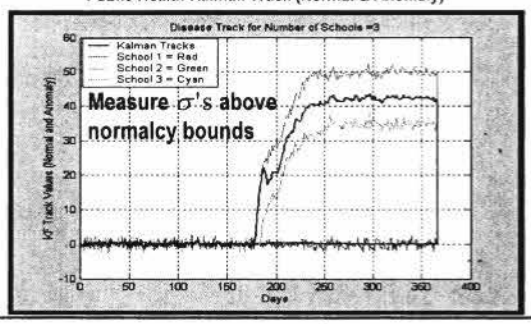


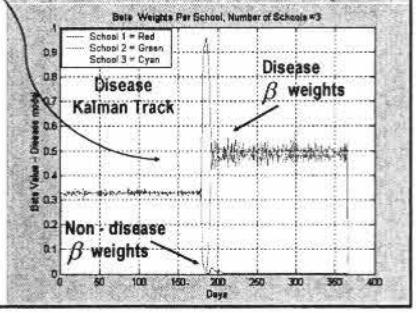
Evaluation Example School Absentee (KF Anomaly and Aggregate Scores)

- **Compare bio-agent based Kalman anomaly results against normalcy Kalman tracks**
  - Baseline normalcy bounds (+/- nσ)
  - Number of standard deviations defines intensity of bio-agent attack over time
  - Beta aggregation weights compare degree of contribution from each source



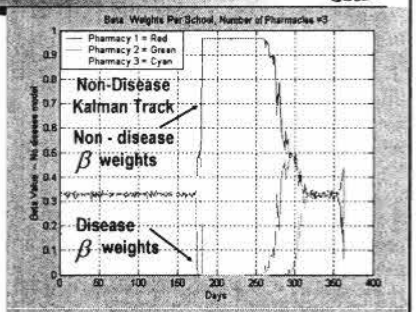
Public Health Kalman Track (Normal & Anomaly)



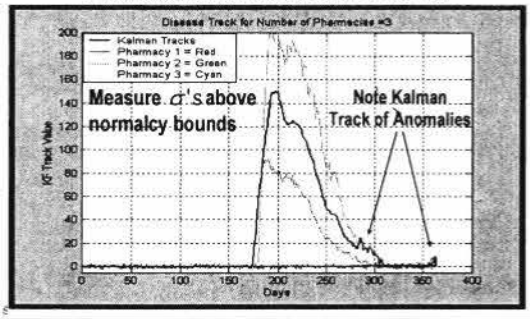


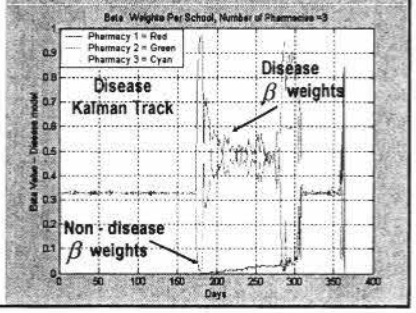
Evaluation Aspirin Example (KF Anomaly and Aggregate Scores)

- **Compare bio-agent based Kalman anomaly results against normalcy Kalman tracks**
  - Beta aggregation weights compare degree of contribution from each source
    - Non-disease and disease conditions
    - Give rise to spatial correlation and clustering



Public Health Kalman Track (Normal & Anomaly)





TITLE PAGE

Briefing Materials concerning TIA BIOALERT 1

FOLDER 2.f TOC 10

PRIVACY

- 2.f.1.a Assuring Privacy
    - IBM Research, Yorktown Heights
- 2.f.1.b Privacy Method (Developmental)
    - IBM T.J. Watson Research Center
    - Veridian-Stanford Team
    - Pitt-CMU
    - Johns Hopkins U Applied Physics Lab
- 2.f.1.c Privacy Issues/Initiatives
- 2.f.1.d Assuring Privacy
    - Veridian-Stanford Team 2nd PI Meeting

FOLDER 2.f TOC 10 – Cont'd – Page 2

- 2.f.1.e Privacy Method (Developmental)
    - IBM T.J. Watson Research Center
    - Veridian-Stanford Team
    - Pitt-CMU
    - Johns Hopkins U Applied Physics Lab

Z.F.I.A

Assuring Privacy

IBM Research
Yorktown Heights, USA

Data

- Health self-assessment (HR,Legal,IRB)
  - Age (binned), children, smoke, flu-shot, county of residence, work location, mass transit use
- Badge recording at entry (HR,Legal)
  - ID, time of entry
- Phone log (HR,Legal)
  - Originating phone, Number called, time of call
- Web log (HR,Legal)
  - Source IP address, target IP address, URL
- SDI (NDA)

IBM

Health self-assessment

- Applied k-anonymity model [Sweeney, Samarati]
  - (Uses generalizations and suppressions)
  - k=10 (specified at survey subscription)
  - Retained 4 attributes
    - Age (<46, rest), Children (y/n), flu shot (y/n), work place (haw/ykt)
    - 28 out of 387 suppressed

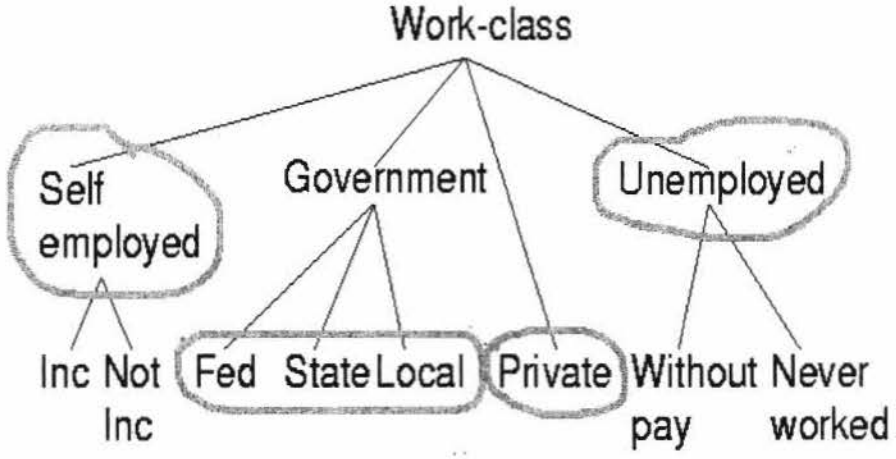
IBM

Our Extensions

- Usage based metrics
  - Tailor information loss due to generalizations and suppressions
    - E.g., anonymized table used for predictive modeling
- Allow more flexible generalizations
  - Intervals for numeric attributes, multi-level generalizations using taxonomy tree
- Aggressive exploration of solution space using genetic algorithm framework
  - Minimize loss metric while satisfying anonymity constraint

IBM

Example: Generalization



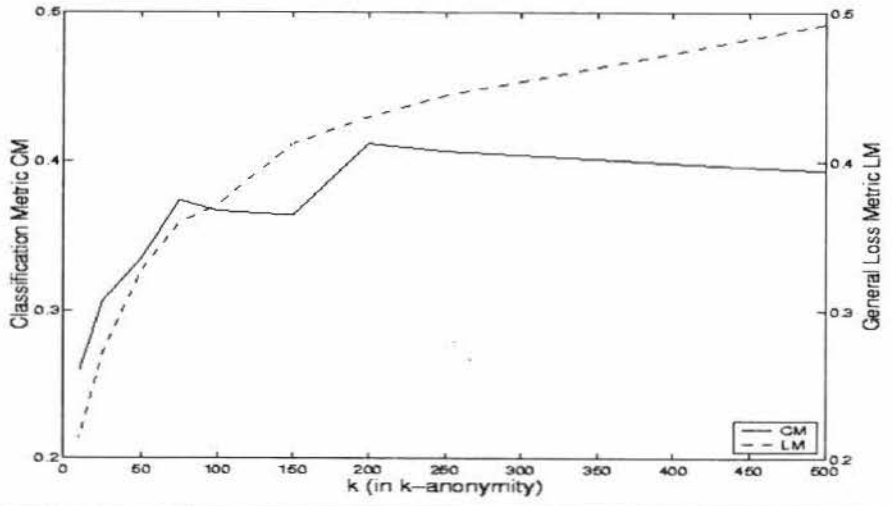
IBM

Accomplishments

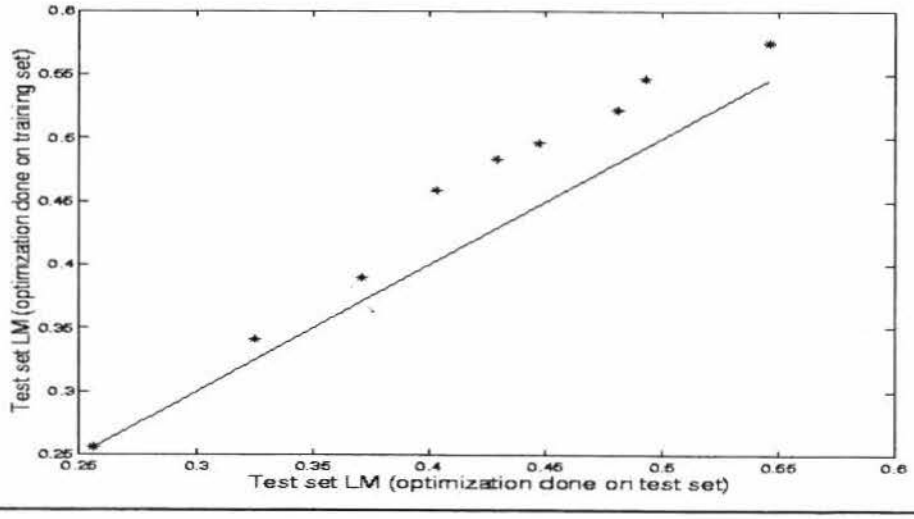
- Prototype system developed for anonymizing data
  - Applied to health survey
  - Tested using surrogate census data
    - Large search space (chromosome length = 157 bits)
    - Anonymized solutions generated (0.5M iterations of GA)
    - Quality of some solutions validated
      - Predictive modeling usage

IBM

Optimizing LM



Reuse on new data: LM



Comparison With Noise + Swapping Approach

- Generalization and suppression
  - + Truthful relationships in each record
  - - Handling many identifying attributes
  - + Handling categorical attributes

IBM

Future Work

- Inferential attribute disclosure
  - Sensitive attributes
  - Probabilistic notion of disclosure
    - Knowledge gain due to data dissemination

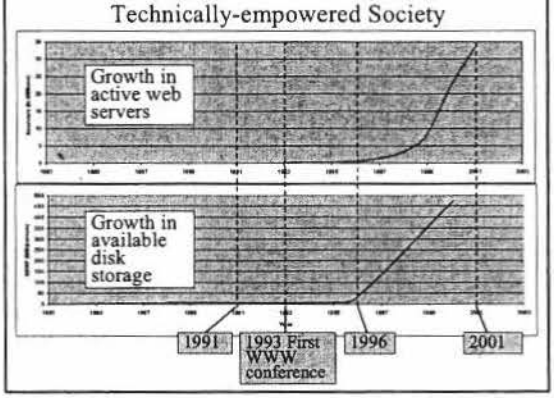
IBM

TEAM	Privacy Method (Developmental)	For Data Type	Advantages	Disadvantages
Veridian/Stanford JHU/APL IBM Mellon Pitt	Extensions to k-anonymity model (implemented in prototype privacy module)	Various data sources with numeric and categorical data (e.g., health data, survey data)	Usage based metrics tailor information loss to purpose for which data is being released. More flexible generalizations expand solution space which is aggressively searched using a genetic algorithm framework.	The k-anonymity based approach addresses the identity exposure issue but does not address the attribute exposure issue which is important to consider in the medical data context.

Enabling Data Sharing for Bio-Surveillance using Data Anonymity

Latanya Sweeney, Ph.D.
Director, Laboratory for International Data Privacy
Assistant Professor of Computer Science, Technology and Policy
Carnegie Mellon University
latanya@cs.cmu.edu
http://lab.privacy.cs.cmu.edu/people/sweeney/

Technically-empowered Society



Laboratory for International Data Privacy at CMU

Work with real-world stakeholders:
- public health agencies
- government agencies
- private corporations

Kinds of projects currently underway:
- health data
- web data
- video surveillance data
- genetic data
- census surveys
- crime data
- grocery data, and so on…

http://lab.privacy.cs.cmu.edu/

Data Anonymity (*new area*)

The study of computational solutions for releasing data such that the data remain practically useful while the identities of the subjects of the data are not revealed.

"Useful AND Secure"

Learning information about entities...

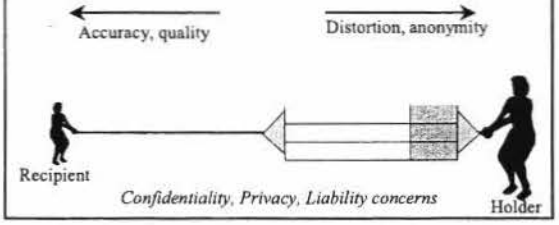
Data Linkage ("data detectives"):
combining disparate pieces of entity-specific information to learn more about an entity

Privacy Protection ("data protectors"):
release information such that certain entity-specific properties (such as identity) cannot be inferred; restrict what can be learned

Enabling Early-detection BioSurveillance: the STOP Project

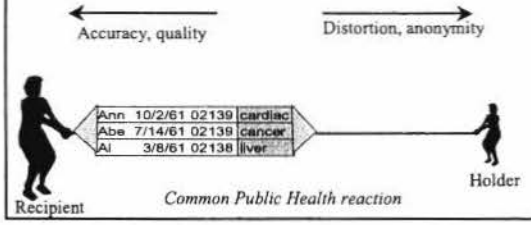
The goal of the CMU STOP project working with Essence II is to enable early detection bio-surveillance by providing tools and analyses necessary to acquiring needed data while being minimal invasive to personal privacy.

"Can't release data"



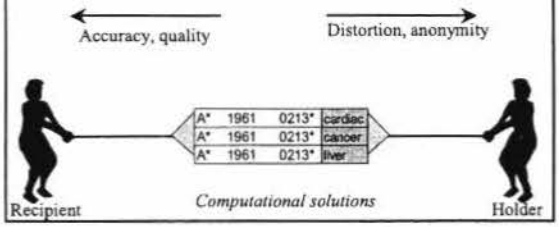
Confidentiality, Privacy, Liability concerns

"Privacy is dead, get over it"



Common Public Health reaction

"Share data while guaranteeing anonymity"



Computational solutions

Three Questions:

1. What access problems do bio-surveillance efforts face when attempting to collect non-traditional data?
2. Does public health law provide sufficient authority?
3. Do anonymity tools reduce the risk and liability? How do other instruments compare? What are the tools like?

Three Questions:

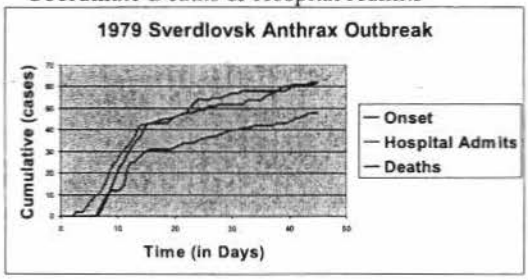
1. What access problems do bio-surveillance efforts face when attempting to collect non-traditional data?
2. Does public health law provide sufficient authority?
3. Do anonymity tools reduce the risk and liability? How do other instruments compare? What are the tools like?

Bottom line: data anonymity reduces burden and better enables collection.

This talk

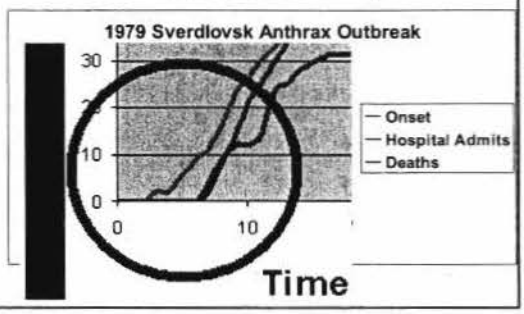
- Kinds of Access Problems
- Public Health Law survey
- Reasonable Cause Model
- Distributed Surveillance
- Examples of Very Non-Traditional Data Sources

Detect Early using Onset, Coordinate Deaths & Hospital Admits



Based on results reported in Guillemin, 1999.

How can we detect onset? How early on each can we predict? How does coordination help?



Continuously Observe Behaviors to Detect Onset of Symptoms

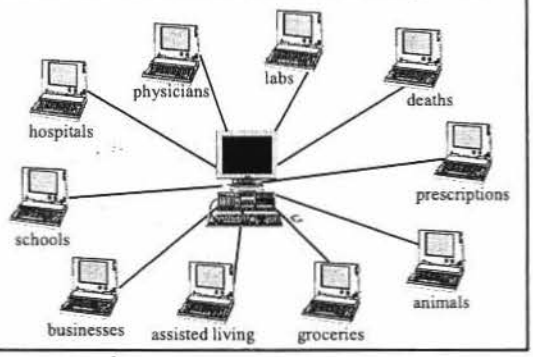
Prodromic surveillance:

How many are acting ill?

Unusual behaviors→syndromes?

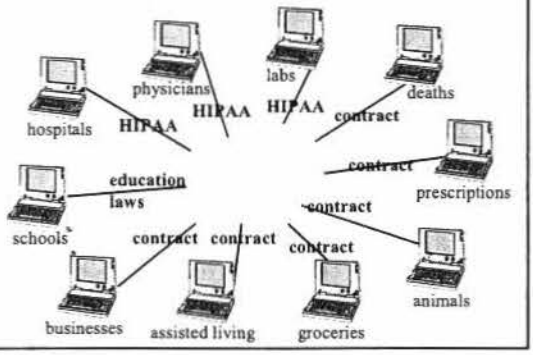
Not confirmed diagnoses!

Centralized Surveillance of Secondary Data



Access Instruments

*Not including public health law



Legal Access Instruments

Here is a list of legal instruments by which data could be collected and shared for bioterrorism surveillance.

- Public health laws
- HIPAA
- contracts
- education laws
- new legislation

Privacy Concerns from a Legal Perspective

- Fourth Amendment of the Constitution
- Compelling state interest
- Reasonable expectation of privacy
- Doctor-patient relationship
- Additional uses beyond direct surveillance, leading to class action lawsuits
- Legal authority to share data (grocer to system) and also further secondary sharing (system to others)
- Decisions made in ignorance regarding the identifiability of the released data

Privacy Concerns under Public Health Law

- reporting versus surveillance
- class action law suit risk in cases where global information is passed on for non-surveillance use
- relationship to HIPAA

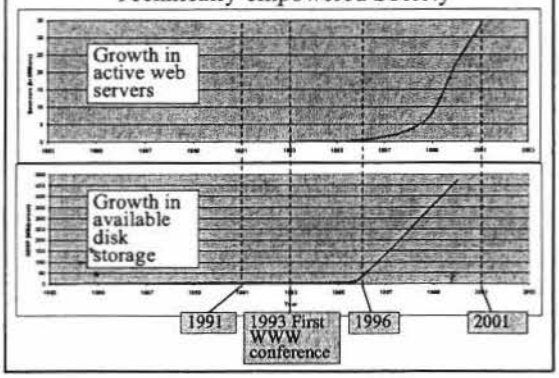
Privacy Concerns from an Ethical Perspective

- Harms and benefits
- Risks to socially-defined groups
- Ability to opt out
- Informed consent has no role
- "minimal risk"
- Observation of public behavior versus observation of public through data

Additional Privacy Concerns specific to Research

- Human subjects
- Research funding source
- Role of IRB
- Research purposes and secondary research on collected data
- Secondary sharing of data

Technically-empowered Society



Anonymity misconceptions

- incorrect belief that de-identified implies anonymous
- incorrect belief that linkage and mining are controlled by encryption
- incorrect belief that security is the same as privacy
- inability to enumerate all sources, users and uses of medical data

Anonymous data

… implies that the data cannot be manipulated or linked to identify an individual.

De-identified Data

… all explicit identifiers, such as name, address and phone number are removed, generalized or replaced with a made-up alternative.

De-identifying information provides no guarantee that the result is anonymous.

JLME 97, NRC 98

Hospital Discharge Data, *fields 1-12*

#	Field description	Size
1	HOSPITAL ID NUMBER	12
2	PATIENT DATE OF BIRTH (MMDDYYYY)	8
3	SEX	1
4	ADMIT DATE (MMDYYYY)	8
5	DISCHARGE DATE (MMDDYYYY)	8
6	ADMIT SOURCE	1
7	ADMIT TYPE	1
8	LENGTH OF STAY (DAYS)	4
9	PATIENT STATUS	2
10	PRINCIPAL DIAGNOSIS CODE	6
11	SECONDARY DIAGNOSIS CODE - 1	6
12	SECONDARY DIAGNOSIS CODE - 2	6

Hospital Discharge Data, *fields 12-25*

#	Field description	Size
13	SECONDARY DIAGNOSIS CODE - 3	6
14	SECONDARY DIAGNOSIS CODE - 4	6
15	SECONDARY DIAGNOSIS CODE - 5	6
16	SECONDARY DIAGNOSIS CODE - 6	6
17	SECONDARY DIAGNOSIS CODE - 7	6
18	SECONDARY DIAGNOSIS CODE - 8	6
19	PRINCIPAL PROCEDURE CODE	7
20	SECONDARY PROCEDURE CODE - 1	7
21	SECONDARY PROCEDURE CODE - 2	7
22	SECONDARY PROCEDURE CODE - 3	7
23	SECONDARY PROCEDURE CODE - 4	7
24	SECONDARY PROCEDURE CODE - 5	7
25	DRG CODE	3

Hospital Discharge Data, *fields 26-37*

#	Field description	Size
26	MDC CODE	2
27	TOTAL CHARGES	9
28	ROOM AND BOARD CHARGES	9
29	ANCILLARY CHARGES	9
30	ANESTHESIOLOGY CHARGES	9
31	PHARMACY CHARGES	9
32	RADIOLOGY CHARGES	9
33	CLINICAL LAB CHARGES	9
34	LABOR-DELIVERY CHARGES	9
35	OPERATING ROOM CHARGES	9
36	ONCOLOGY CHARGES	9
37	OTHER CHARGES	9

Hospital Discharge Data, *fields 38-50*

#	Field description	Size
38	NEWBORN INDICATOR	1
39	PAYER ID 1	9
40	TYPE CODE 1	1
41	PAYER ID 2	9
42	TYPE CODE 2	1
43	PAYER ID 3	9
44	TYPE CODE 3	1
45	PATIENT ZIP CODE	5
46	Patient Origin COUNTY	3
47	Patient Origin PLANNING AREA	3
48	Patient Origin HSA	2
49	PATIENT CONTROL NUMBER	
50	HOSPITAL HSA	2

Hospital Discharge by State, Part 1

	Mandate	Private (Insiders)	Semi-Private (Limited)	Semi-Public (Deniable)	Public (No Restrictions)	AHRQ SID
Alabama	N	N				
Alaska	N	N				
Arizona	Y	Y	N	Y	Y	Y
Arkansas	Y	Y	N	N	N	
California	Y	Y	N	Y	Y	Y
Colorado	N	Y	N	Y	N	Y
Connecticut	Y	Y	N	Y	Y	Y
Delaware	Y	Y	N	N*	N*	
District of Columbia	N	N				
Florida		Y	N		Y	Y
Georgia		Y	N	N	N	Y
Hawaii	N	Y	N	Y	Y	Y
Idaho	N	N				
Illinois	Y	Y	Y	Y	Y	Y
Indiana	Y	Y	N	N	N	
Iowa	Y	Y	N	Y	Y	Y
Kansas	Y	Y	N	Y	N	Y
Kentucky	Y	Y	N	Y	N	
Louisiana	N	Y	N			
Maine	Y	Y	N	Y	Y	
Maryland	Y	Y	N	Y	Y	Y
Massachusetts	Y	Y	N	Y	Y	Y
Michigan	N	Y	N	Y	N	
Minnesota	N	Y	N	Y	N	
Missouri		Y	N	Y	Y	Y
Mississippi	N	N				

Hospital Discharge by State, Part 2

	Mandate	Private (Insiders)	Semi-Private (Limited)	Semi-Public (Deniable)	Public (No Restrictions)	AHRQ SID
Montana	N	N				
Nebraska	N	Y	N	Y	Y	
Nevada	Y	Y	N	N	Y	
New Hampshire	Y	Y	N	Y	Y	
New Jersey	N	Y	Y	N	Y	Y
New Mexico	Y	Y	N	N	Y	
New York	Y	Y	N	Y	Y	Y
North Carolina		Y	Y	N	N	
North Dakota		Y	N	N	Y	
Ohio	Y	Y	N	N	N	
Oklahoma	Y	Y	N	Y	N	
Oregon		Y	N	Y	Y	Y
Pennsylvania	Y	Y	Y	Y	Y	Y
Rhode Island	Y	Y	N	Y	Y	
South Carolina	Y	Y	N	Y	Y	Y
South Dakota	N	N				
Tennessee	Y	Y	N	Y	Y	Y
Texas	Y	Y	N	N	N	
Utah	Y	Y	N	Y	Y	Y
Vermont	Y	Y	N	Y	Y	
Virginia	Y	Y	N	Y	Y	
Washington	Y	Y	N	Y	Y	Y
West Virginia	Y	Y	N	Y	Y	
Wisconsin	Y	Y	N	Y	Y	Y
Wyoming	Y	Y	N	Y	N	

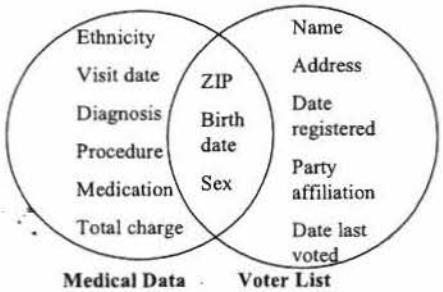
Health data (GIC example)

Medical Data

Population data (GIC example)

Voter List

Linking to re-identify data



Medical Data **Voter List**

Uniqueness in Cambridge Voters

Birth date alone	12%
Birth date & gender	29%
Birth date & 5-digit ZIP	69%
Birth date & full postal code	97%

Birth date includes month, day and year.
Total 54,805 voters.

JLME 97

Few characteristics make a person unique

Birth includes month, day and year:

365 days x 100 years = 36,500 possibilities

Two genders and Five ZIP (5-digit) codes:

2 * 5 * 36,500 = **365,000 possibilities**

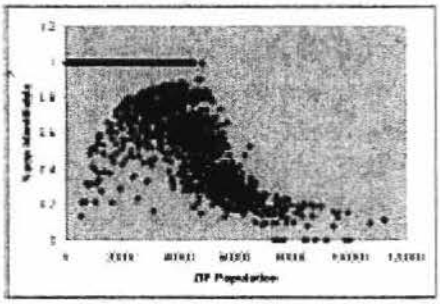
But the Cambridge Voter list had:

54,805 voters

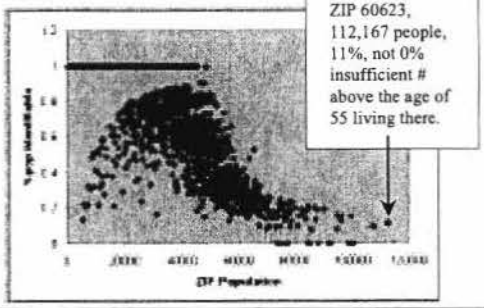
So in general, using
(birth[mon,day,yr], gender, ZIP[5-digit])
provides a **unique quasi-identifier**.

JLME 97

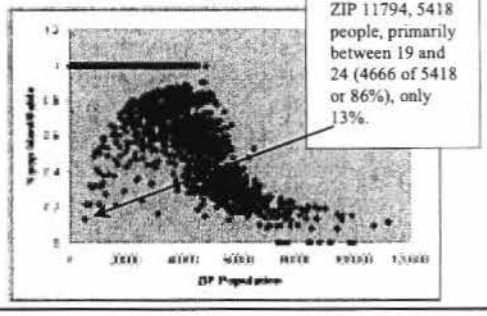
{date of birth, gender, 5-digit ZIP}
uniquely identifies 87.1% of USA pop.



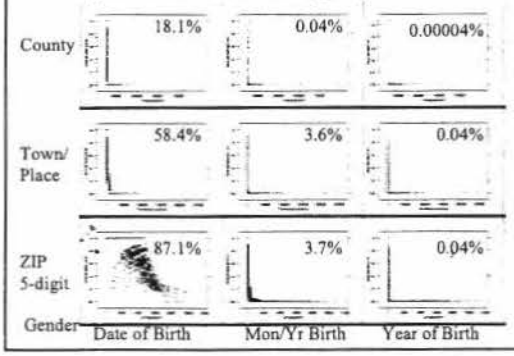
{date of birth, gender, 5-digit ZIP}
uniquely identifies 87.1% of USA pop.



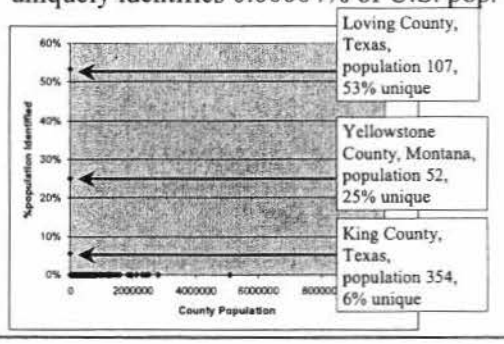
{date of birth, gender, 5-digit ZIP}
uniquely identifies 87.1% of USA pop.



Uniqueness of Demographics in U.S.



{Year of birth, gender, County},
uniquely identifies 0.00004% of U.S. pop.



Cancer registry looks anonymous

Diagnosis	DiagDate	ZIP
Kaposi's Sarcoma	1/18/91	32555
Kaposi's Sarcoma	5/12/94	37581
Kaposi's Sarcoma	3/5/92	32172
Kaposi's Sarcoma	8/8/93	30158
Neuroblastoma	4/3/91	39164

Kinds of Privacy Mechanisms

- Limited data sets
- De-identification tools
- Anonymity tools
- IRB only
- Legal authority
- Authorized agent
- Data use agreements
- None required

Access policy as basis for metrics

$$P_R = (D, I, E, F)$$

Gross DSP of a release $GI_S(P_R) = c * D * E$

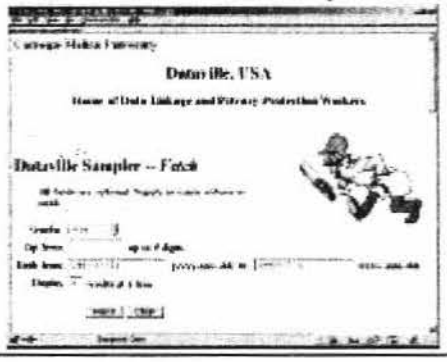
Gross Maximal Risk $gross\ maximal\ risk = I * E$

Long-term DSP $= \Sigma GI_S(P) = D * (E_1 + F_1 + F_2 + \ldots + F_{n-1})$

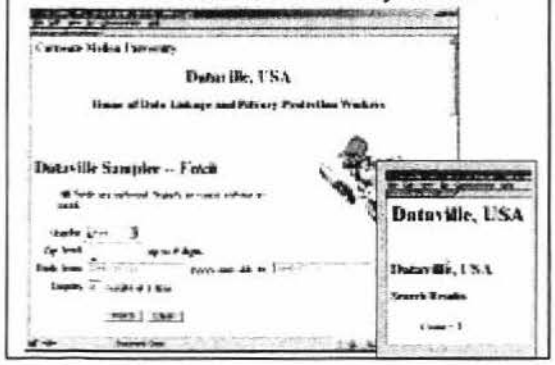
Total storage $total\ storage(P_R, t_1, t_2) = AI_S(P_R, t_1, t_2)_I * I$

Sweeney, L. 2000. *Towards all the data on all the people.*

Live Demo of Identifiability Server



Live Demo of Identifiability Server



Identifiability Server – Levels of accuracy

Identifiability of Privacy Mechanisms

Limited data sets	varies
De-identification tool	better, but varies
Anonymity tool	best
IRB only	varies
Legal authority	varies
Authorized agent	varies
Data use agreements	varies
None required	varies

Access policy as basis for metrics

$$P_R = (D, I, E, F)$$

Gross DSP of a release $GI_S(P_R) = c * D * E$

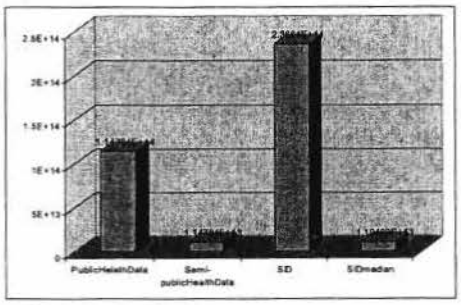
Gross Maximal Risk $gross\ maximal\ risk = I * E$

Long-term DSP $= \Sigma GI_S(P) = D * (E_1 + F_1 + F_2 + \ldots + F_{n-1})$

Total storage $total\ storage(P_R, t_1, t_2) = AI_S(P_R, t_1, t_1)_i * I$

Sweeney, L. 2000. *Towards all the data on all the people.*

Example: Gross Maximal Risk = $I * E$

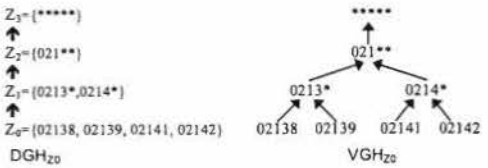


SID has greater risk.

Risk and Liability of Privacy Mechanisms

Limited data sets	varies
De-identification tool	varies
Anonymity tool	minimal
IRB only	varies
Legal authority	varies, future none
Data use agreements	varies, but often high
Authorized agent	varies, future unsure
None required	none, future unsure

Precision Metric



$$prec(RT) = 1 - \frac{\sum_{i=1}^{N_A} \sum_{j=1}^{N} \frac{h}{|DGH_{Ai}|}}{|RT| \bullet |N_A|}$$

Privacy Protection of Privacy Mechanisms

Limited data sets	varies
De-identification tool	varies
Anonymity tool	best
IRB only	varies
Legal authority	varies
Authorized agent	varies
Data use agreement	varies
None required	varies

This talk

- ✓Kinds of Access Problems
- Public Health Law survey
- Reasonable Cause Model
- Distributed Surveillance
- Examples of Some Non-Traditional Data Sources

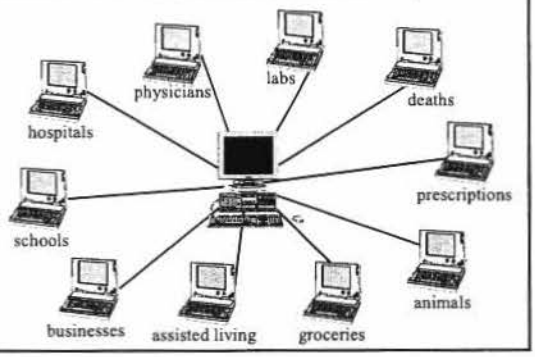
LegalSTOP: Public Health Law Survey

Laboratory for International Data Privacy

JTLP The Pittsburgh Journal of Technology Law and Policy

School of Law

Centralized Surveillance of Secondary Data



Survey 67 Questions in 6 Groups

I Central authority 1-4

II Require non-traditional data 5-16

III Coaxing 17-28

IV Further sharing 29-40

V Other Internal uses 41-52

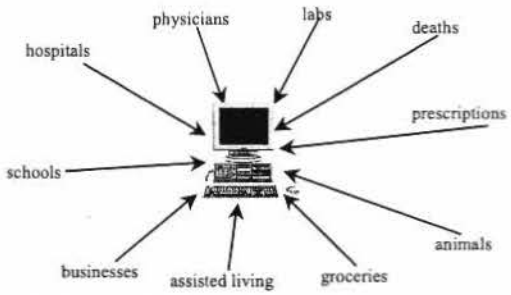
VI De-identified 53-67

I. Central Authorities

Here is a list of potential central authorities for collecting and analyzing the data for bioterrorism surveillance purposes under public health law.

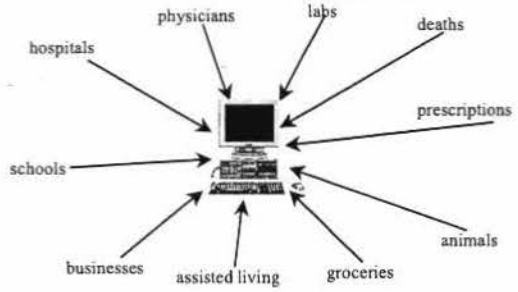
1. Public health agency
2. Trusted broker of public health agency
3. Corporation (for profit)
4. University (non-profit, non-competitor)

Case 1: Public Health Agency



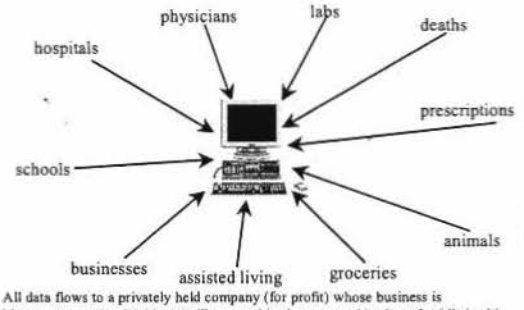
All data flows to public health agency directly under Public Health law.

Case 2: Trusted Broker (Competing) of Public Health Agency



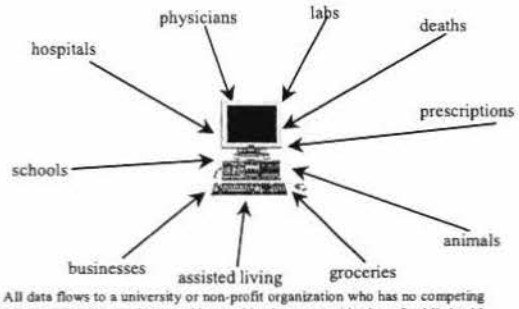
Example All data flows to one competitor by authority of public health agency.

Case 3: Private For-profit Corporation



All data flows to a privately held company (for profit) whose business is bioterrorism and/or health surveillance and is also a trusted broker of public health.

Case 4: Non-profit Corporation



All data flows to a university or non-profit organization who has no competing relationship with the data providers and is also a trusted broker of public health.

I. Central Authority Questions

1. Does the Public Health Department (Dept.) have the legal authority to **require** data on the general population **be reported to the Dept.**?
2. … to an authorized agent …
3. … to a for-profit corporation as agent …
4. … to a non-profit organization as agent…

II. Require Non-traditional Data

5. Does the Public Health Department have the legal authority to **require all *hospital visits*** be reported to the Dept. even if it does not involve a reportable disease or some other limiting factor such as public health emergency?

6- emergency room data
7- physician offices
8- medical labs
9- death data
10- prescription data
11- animal data
12- over the counter, groceries
13- video surveillance data
14- assisted living groups
15- business absenteeism
16- school absenteeism

III. Coaxing data and providing information in exchange

Example. In order to get pharmacies to provide data on medications sold, the public health agency computes global information about the industry and then provides the aggregate information back to each pharmacy in exchange for being provided the data for bio-surveillance.

Questions 17-28, data type specific

IV. Further Sharing of Collected Data

Example. Suppose a grocer provides information to public health for bioterrorism surveillance purposes. The public health agency than provides the data (in full or part) to others.

Questions 29-40, data type specific

V. Internal use for Other Purposes

Data provided to public health for bio-surveillance purposes but then used by public health internally for other reasons.

Questions 41-52, data type specific

VI. Use of De-identified and Sufficiently Anonymous Data

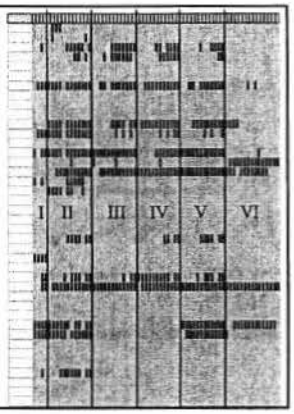
Using de-identified (or better, sufficiently anonymous data) may reduce the privacy risk and liability and enable more data sharing under existing laws.

Questions 53-67, data type specific

Overview by State

I Central authority 1-4
II Require data 5-16
III Coaxing 17-28
IV Further sharing 29-40
V Internal uses 41-52
VI De-identified 53-67

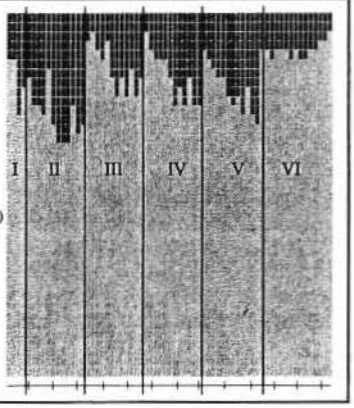
No
Uncertain
Yes



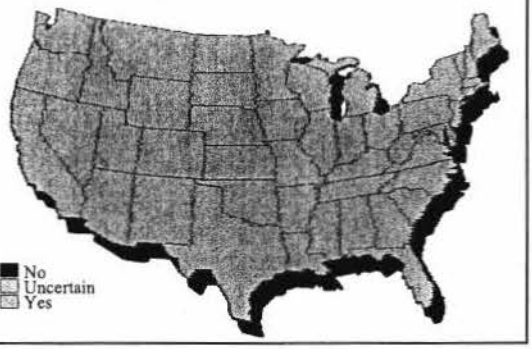
Overview by Question

I Central authority 1-4
II Require data 5-16
III Coaxing 17-28
IV Further sharing 29-40
V Internal uses 41-52
VI De-identified 53-67

No
Uncertain
Yes

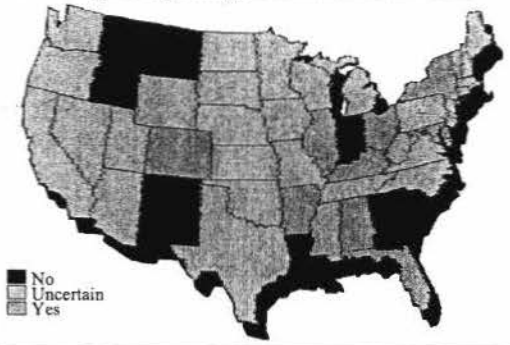


Have case reporting authority?



No
Uncertain
Yes

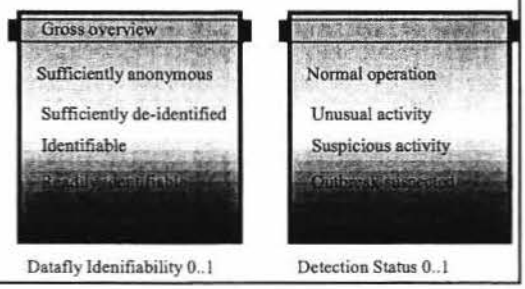
17. … require reporting on non-traditional data?



This talk

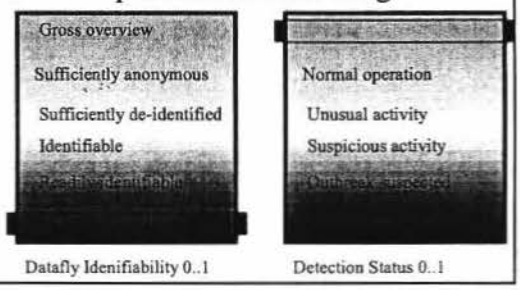
- ✓Kinds of Access Problems
- ✓Public Health Law survey
- Reasonable Cause Model
- Distributed Surveillance
- Examples of Some Non-Traditional Data Sources

Mechanical distortion decisions typically renders data useless



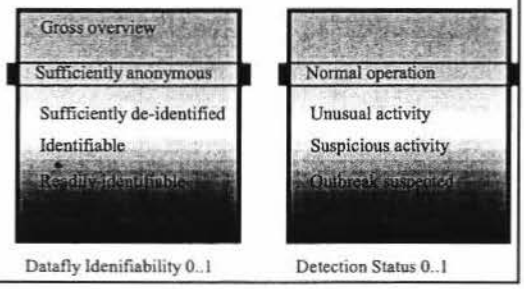
Datafly Idenifiability 0..1 Detection Status 0..1

Explicitly identified data generates privacy concerns which may ultimately prohibit data sharing



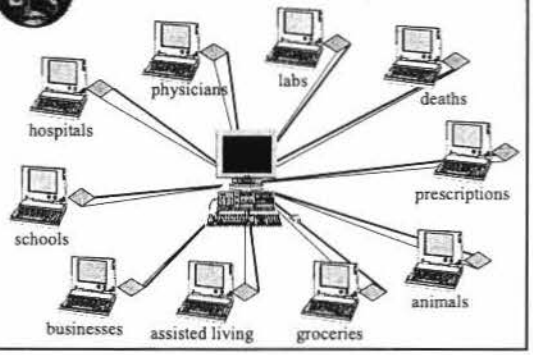
Datafly Idenifiability 0..1 Detection Status 0..1

Levels of identifiability matching detection status

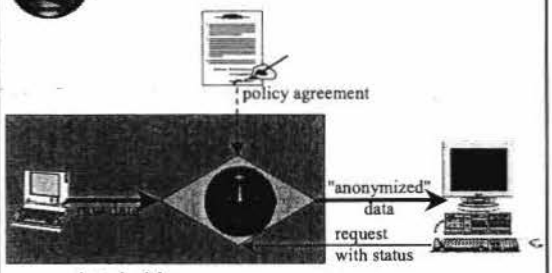


Datafly Idenifiability 0..1 Detection Status 0..1

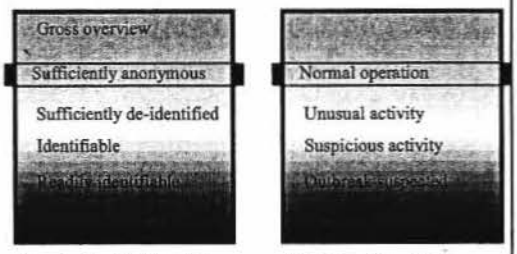
Automated Privacy Module



Automated Privacy Module

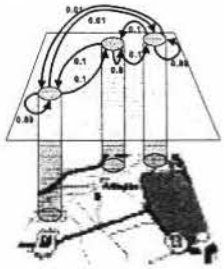


Levels of identifiability matching detection status



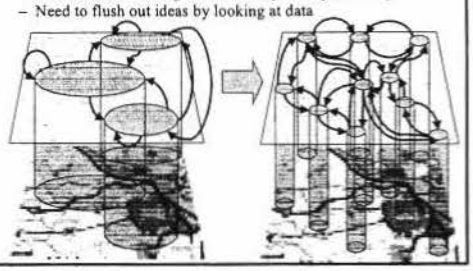
Model Interactions Between Clusters

- Contact happens both within a cluster, and between clusters
  - Strong geographical component, both work and residence
  - Model interactions between clusters
- Accurate modeling requires small clusters
  - Assume homogenous mixing within a cluster
  - Most accurate models consider individuals or households
  - Data / privacy implications

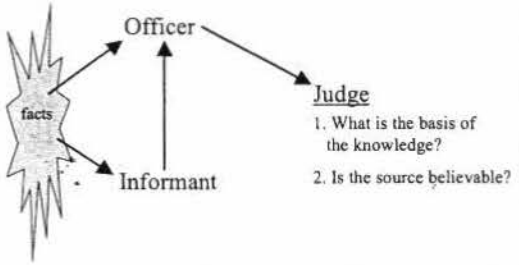


Dynamically Augment the Model When Surveillance Detects Possible Attack

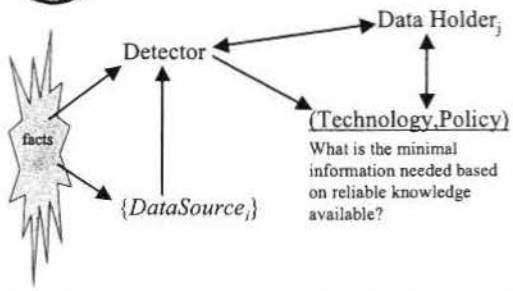
- Lower the privacy threshold when potential attack detected
  - But how often, how quickly, to what level?
  - Can we take advantage of disease-specific processing?
  - Need to flush out ideas by looking at data



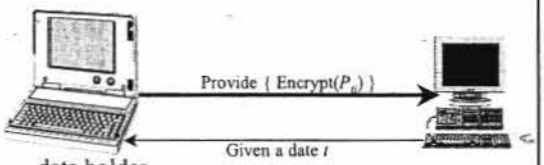
Probable Cause Predicate



Reasonable Cause Predicate



A quick possibility to combat duplication concern … provide encrypted values for identifiers



BUT! This would allow the detector to make audit trails of patients, which could lead to their identity. More thought needed…

This talk

- ✓Kinds of Access Problems
- ✓Public Health Law survey
- ✓Reasonable Cause Model
- ✓Distributed Surveillance
- Examples of Some Non-Traditional Data Sources
  - Video surveillance
  - Web use and medical claims

Analyses based on data sources

Data sources	
hospital inpatient claims emergency room visits physician office visits lab reports prescriptions	"traditional" for bioterrorism surveillance
assisted living reports school absentism business absentism animal reports	extended list of sources
grocery purchases web clickstream data death data videos of public spaces GPS auto data	very non-traditional data

Analyses based on data sources

Data sources	
• hospital inpatient claims emergency room visits • physician office visits lab reports • prescriptions	"traditional" for bioterrorism surveillance
assisted living reports school absentism • business absentism animal reports	extended list of sources
• grocery purchases • web clickstream data • death data • videos of public spaces • GPS auto data	very non-traditional data

Team Members on Video Surveillance

- Luis von Ahn*
- B. Joseph Barrett, J.D.
- Manuel Blum, Ph.D.*
- Elaine Newton
- Latanya Sweeney, Ph.D.
- Jie Yang, Ph.D.

VideoSTOP: Coughing Behavior

The goal of this project is two-fold:

1. In public spaces, count the number of people who are "coughing" by using existing video surveillance data.
2. To accomplish #1 in such a way that is minimally invasive to privacy.

For Data Type	Privacy issues
Poison center data	Data are deidentified by poison center before sending them to us.
911 data	Data providers do not have deidentification capability. Currently investigating having a third party deidentify data.
Animal data	No privacy issues
OTC data	We only get zip code information

Other privacy initiatives

NOTE: We also are in negotiations with IBM to obtain the 'K-nearest' confidentiality software

Memoranda of understanding

We execute a Memorandum of Understanding (MOU) with each data provider in order to describe the data we are expecting to receive and our responsibilities for handling the data and maintaining the privacy of the provider. Each MOU is tailored to fit the needs of each data provider and is approved by our legal council. The process of drafting the MOU and getting it signed by all parties usually requires extensive discussions between the parties and enough time to circulate the final document among those who must sign.

Researcher agreements

We require that every researcher, student, staff member that is part of our group review the data sharing agreements and memoranda of understanding with each data provider. We then require them to acknowledge that they have read such agreements and agree to abide by them.

Internal Review Board approval

The research protocols being used to analyze the various data sets have been reviewed and approved by our Internal Review Board (IRB) to ensure that we are handling the data appropriately so as to maintaining the security and privacy of the data.

(b)(4)

Biosurv database

This is a database of [redacted] Department patient admission records used by researchers developing detection algorithms. Patient privacy is being preserved by deleting all identifying information (e.g., name, ssn, address) and replacing necessary fields with random values (e.g., patient id number). Patient location information maintained in the database is generalized so as to aggregate patients (e.g., latitude and longitude values are rounded to the nearest .005 degree which aggregates patients within a square 1/3 mile on each side).

Trusted Data Broker

The goal of this project is to help develop a trusted data broker facility in [redacted] under the auspices of a governing board comprised of the County (b)(4) Department of Health and the entities providing surveillance data. The tasks include developing a hardware architecture, software architecture and the accompanying policies to the board.

Currently, the hardware architecture of the Trusted Data Broker facility consists of a Sun Enterprise 250 server running Solaris, an Oracle database and messaging interfaces to the data providers. It sits behind an Intel-based server running OpenBSD for fire walling and intrusion detection services. Network connectivity is established through IPSEC VPNs and fractional T1 leased lines.

We have developed a database schema, messaging interfaces, message parsing engines, batch processing engines, data warehousing programs, and an encrypted identifier API for this project. The database currently consists of 3 years worth of data comprised of 986401 records of people entering and leaving hospitals in [redacted] The messaging interfaces connect to the Health Level 7 (HL7) message routers of each healthcare system to provide consistency in the data. Message parsing engines parse the HL7 messages and store them in the database. Batch processing engines parse batches of hospital admission data for hospitals without HL7 and store them in the database. Data warehousing programs aggregate the data periodically for later analysis. The encrypted identifier API encrypts data using a public key infrastructure. All of the engines, programs and APIs are implemented in Java.

Clean

For Data Type	Privacy issues
Poison center data	Data are deidentified by poison center before sending them to us.
911 data	Data providers do not have deidentification capability. Currently investigating having a third party deidentify data.
Animal data	No privacy issues
OTC data	We only get zip code information

Other privacy initiatives

NOTE: We also are in negotiations with IBM to obtain the 'K-nearest' confidentiality software

Memoranda of understanding

We execute a Memorandum of Understanding (MOU) with each data provider in order to describe the data we are expecting to receive and our responsibilities for handling the data and maintaining the privacy of the provider. Each MOU is tailored to fit the needs of each data provider and is approved by our legal council. The process of drafting the MOU and getting it signed by all parties usually requires extensive discussions between the parties and enough time to circulate the final document among those who must sign.

Researcher agreements

We require that every researcher, student, staff member that is part of our group review the data sharing agreements and memoranda of understanding with each data provider. We then require them to acknowledge that they have read such agreements and agree to abide by them.

Internal Review Board approval

The research protocols being used to analyze the various data sets have been reviewed and approved by our Internal Review Board (IRB) to ensure that we are handling the data appropriately so as to maintaining the security and privacy of the data.

(b)(4)

Biosurv database

This is a database of U[redacted] Department patient admission records used by researchers developing detection algorithms. Patient privacy is being preserved by deleting all identifying information (e.g., name, ssn, address) and replacing necessary fields with random values (e.g., patient id number). Patient location information maintained in the database is generalized so as to aggregate patients (e.g., latitude and longitude values are rounded to the nearest .005 degree which aggregates patients within a square 1/3 mile on each side).

Trusted Data Broker

The goal of this project is to help develop a trusted data broker facility in [redacted] under the auspices of a governing board comprised of the County (b)(4) Department of Health and the entities providing surveillance data. The tasks include developing a hardware architecture, software architecture and the accompanying policies to the board.

Currently, the hardware architecture of the Trusted Data Broker facility consists of a Sun Enterprise 250 server running Solaris, an Oracle database and messaging interfaces to the data providers. It sits behind an Intel-based server running OpenBSD for fire walling and intrusion detection services. Network connectivity is established through IPSEC VPNs and fractional T1 leased lines.

We have developed a database schema, messaging interfaces, message parsing engines, batch processing engines, data warehousing programs, and an encrypted identifier API for this project. The database currently consists of 3 years worth of data comprised of 986401 records of people entering and leaving hospitals in

(b)(4)

[redacted] The messaging interfaces connect to the Health Level 7 (HL7) message routers of each healthcare system to provide consistency in the data. Message parsing engines parse the HL7 messages and store them in the database. Batch processing engines parse batches of hospital admission data for hospitals without HL7 and store them in the database. Data warehousing programs aggregate the data periodically for later analysis. The encrypted identifier API encrypts data using a public key infrastructure. All of the engines, programs and APIs are implemented in Java.

ASSURING PRIVACY

Veridian-Stanford Team

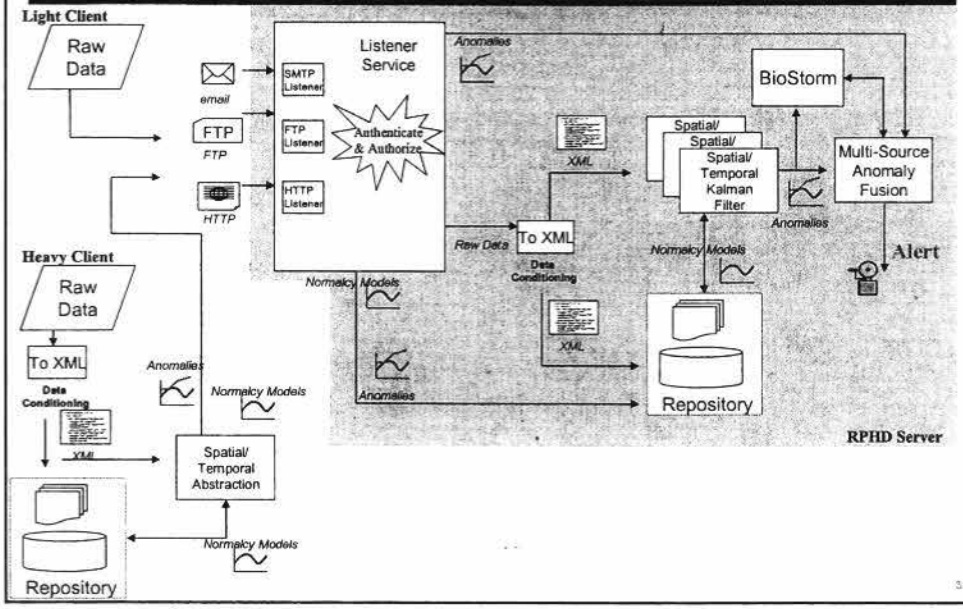
2nd PI Meeting

4-6 June 2002

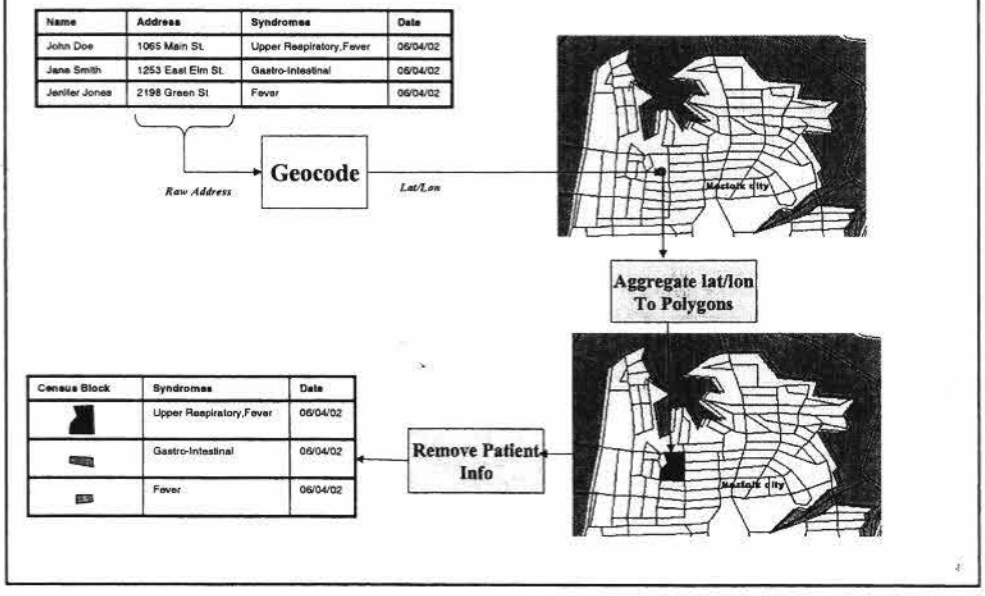
Architecture

- Allows processing of source data either locally or remotely
  - Light-weight data source client (eg: elementary school)
    - Raw data is sent to server to be processed
    - No custom software installed
  - Heavy-weight data source client (eg: hospital)
    - Raw data is processed locally
    - All data is stored locally on repository to maintain privacy
    - Only Kalman filter output: Normalcy Models and Anomalies are sent to server for fusion
- Repository:
  - All data is converted to XML/XML Schema
    - Allows easy integration and export of data
  - Repository can be either:
    - RDBMS or
    - Flat files

RPHD Data Acquisition



Ensuring Privacy in Medical Records for Remote Spatial Processing



2.F.1.5.

TEAM	Privacy Method (Developmental)	For Data Type	Advantages	Disadvantages
Veridian/Stanford	Distributed processing	CHCS	within confines of institution	Only abstracted data available elsewhere
Veridian/Stanford	Incorporation of methods developed by others	Medical records	Cost savings, benefit from experts in privacy field	none
JHU/APL				
IBM				
Mellon Pitt				

TITLE PAGE

FOLDER 2.G TOC 11

Briefing Materials concerning TIA BIOALERT 1

STATUS UPDATES

- 2.G.1a EpiSPIRE Project Update
  - IBM T.J. Watson Research Center
  - John Hopkins University
  - SDI Inc.
- 2.G.1b Scalable BioSurveillance Systems
  - Pitt-CMU
- 2.G.1c Veridan-Stanford Team Project Update

2.9.1 16

EpiSPIRE Project Update

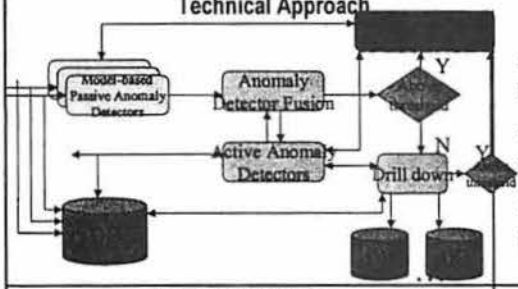
IBM T. J. Watson Research Center
Johns Hopkins University
SDI Inc.

Team

- IBM T. J. Watson Research Center
  - Sankar Basu
  - Murray Campbell
  - Yuan-Chi Chang
  - V. Guruprasad
  - Vijay Iyengar
  - Mahesh Joshi
  - Chung-Sheng Li (PI)
  - Milind Naphade
  - Ed Pednault
  - Jeanette Rosenthal
  - John R. Smith
  - Joel Wolf
  - Kung-Lung Wu
  - Philip S. Yu
- Johns Hopkins University
  - Prof. Greg Glass (co-I)
  - Elaine Florio
  - Rebecca Freeman
- SDI
  - Andrew Kress

EpiSPIRE
IBM & JHU/PI: Chung-Sheng Li

Technical Approach



Objectives and Performance Goals

- Provide real-time access of heterogeneous data sources
- Leverage multi-modal data sources to improve precision of bio-surveillance
- Improved learning speed with few training examples

Accomplishments

- Site-specific simultaneous active and passive surveillance
- Leverage environmental data for baseline
- Feature extraction and event detection from non-traditional data
- Model-based fusion of information among structured, semi-structured and non-structured data sources

Programmatics

- **05/02: Environmental modeling of flu-like disease, completion of data acquisition, initial evaluation of detection algorithm, complete privacy infrastructure**
- **08/02: Complete evaluation of detection algorithm**
- **12/02: Validation of end-to-end system, integration with JHU/APL**
- **02/03: Technology available for transition**

3

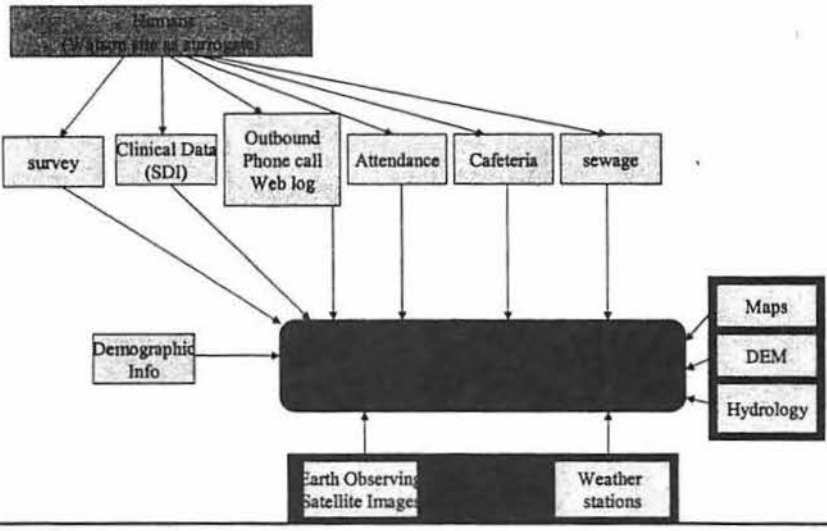
Technical Approach: New Ideas

- Methodology
  - Site-specific simultaneous active and passive surveillance
  - Use of real time environmental data in preparing the baseline of disease modeling
- Data gathering
  - Develop survey techniques for active surveillance
  - Use of outgoing phone and web logs (as opposed to incoming web log at the web site)
  - Automatic car counting through surveillance video
- Advanced Analytics
  - Differential detection technique
  - Use of multi-dimensional demographic profile of patient population as discriminator
  - Multi-modal/multi-channel techniques

4

EpiSPIRE for Site Specific Bio-surveillance

•Comprehensive catalog of all site-specific data sources

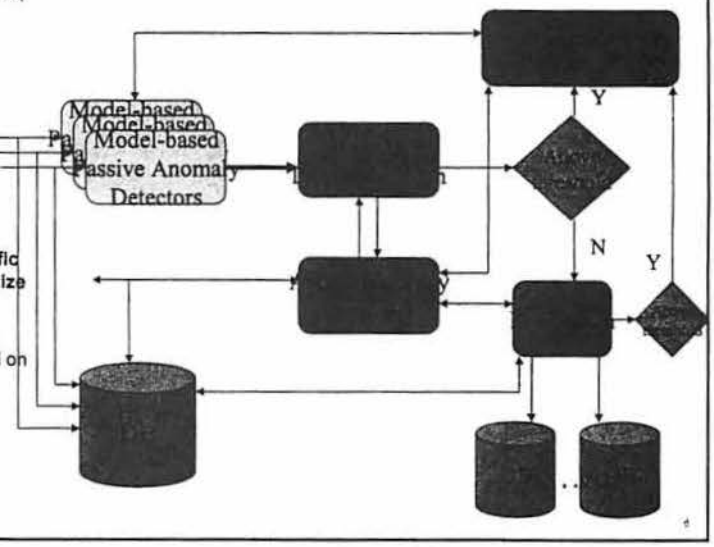


Technical Approach: Methodologies

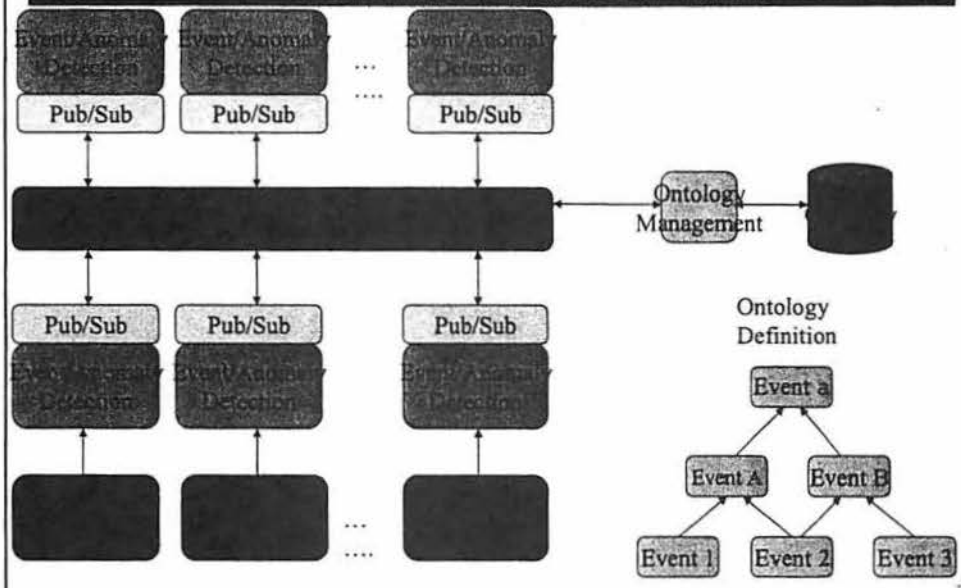
•Use of simultaneous **active** and **passive** surveillance techniques, validation through *in situ* site-specific data gathering

•**Detectable scenarios**

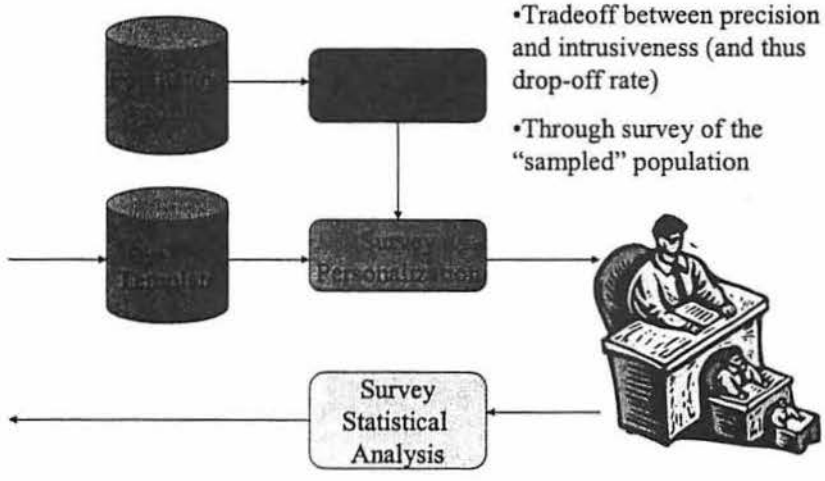
- •**Respiratory infection** (ground truth based on SDI)
  - –Survey
  - –Phone records
  - –Proxy log
- •**Anthrax for site specific attack with sufficient size**
  - –Simulated attack
  - –Emulated signature on multiple channels based on the behavior model



Technical Approach:
Event/Alert Generation – System View



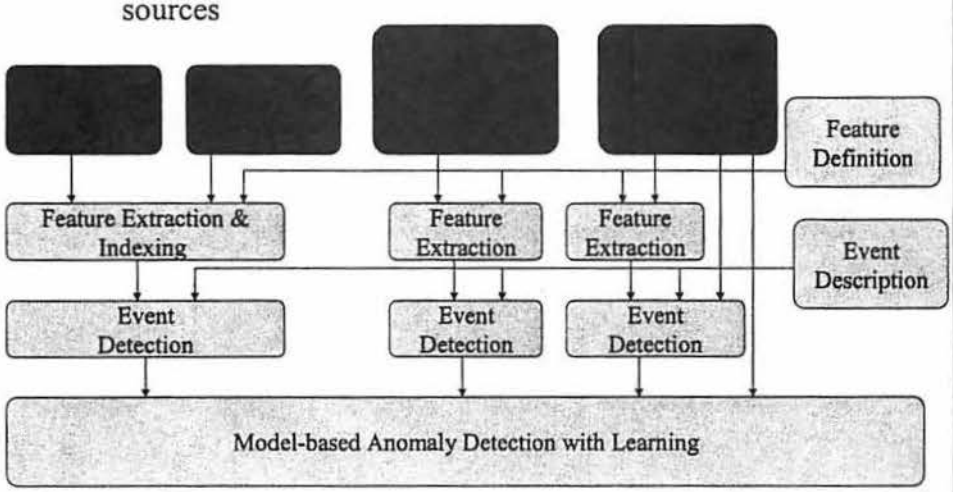
Active Anomaly Detector



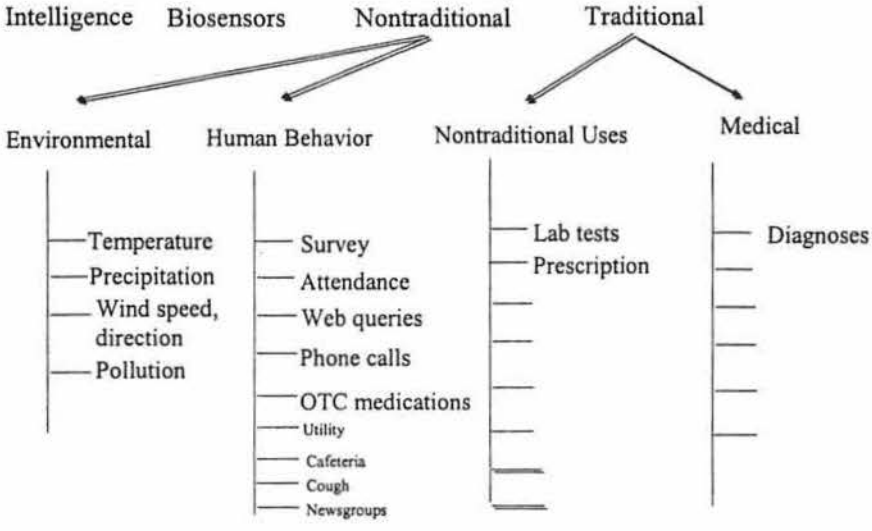
- Tradeoff between precision and intrusiveness (and thus drop-off rate)
- Through survey of the "sampled" population

Model-based Passive Anomaly Detector

•Alert generation through passive monitoring of data sources

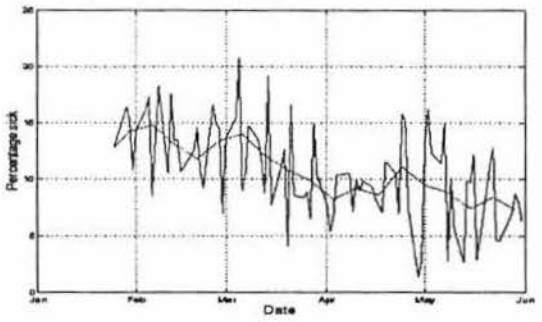


Technical Approach: Non-traditional data sources



Active Surveillance

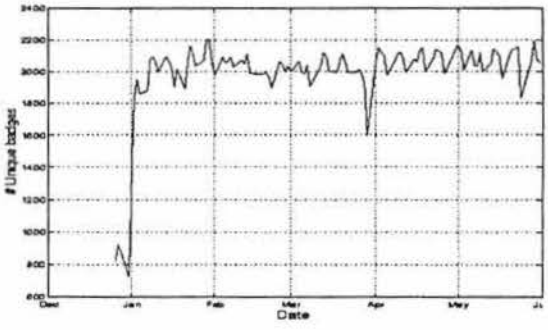
- **Hypothesis**: the increase of potential flu outbreak can be detected through active surveillance
- **Methodology**:
  - surveying employee at Watson Research Center from Westchester county for their well being
  - Each employee is expected to respond with normal, worse than normal, much worse than normal
- **Validation Strategy**:
  - Validate against SDI data



11

Badge Data

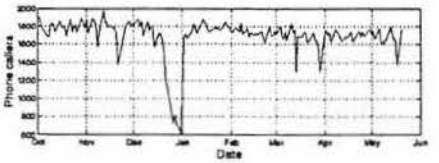
- **Hypothesis**: occurrences of potential flu outbreak can be detected through data mining on the badge access pattern.
- **Methodology**:
  - we have start archiving badge access of Watson Research Center since Dec. 13.
  - The badge access data set contains unique entry record of Watson premises.
- **Validation strategy**:
  - Validate against SDI

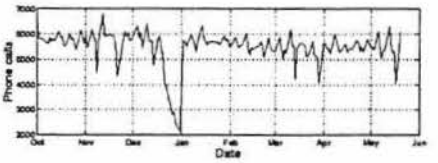


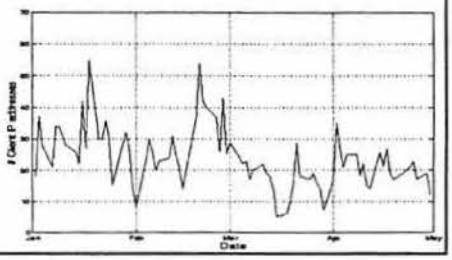
12

Firewall Server Log/Phone Records

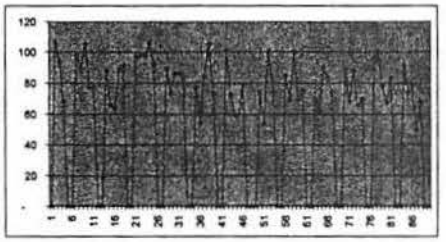
- **Hypothesis**: occurrences of potential flu (and other disease outbreaks) outbreak can be detected through the mining of the proxy server log traces and phone records
- **Methodology**:
  - We have started archiving Watson proxy log trace since end of December. Each record in the proxy log trace contains the source and destination IP addresses of web access, as well as the URL access record.
  - We identified unique access records of those URL's and phone numbers which are related to a few pre-established medical sites in westchester county
- **Validation Strategy**:
  - Validate against SDI data for Westchester county



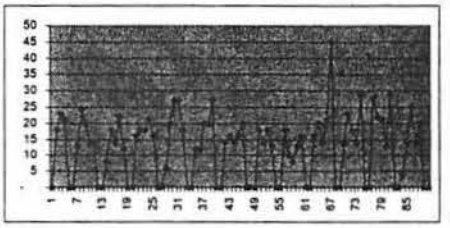




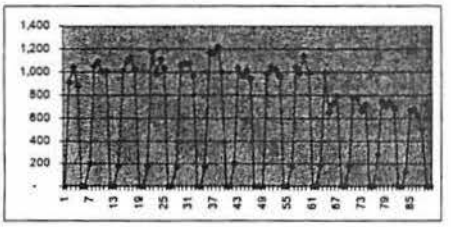
Cafeteria Data (Jan-Mar)



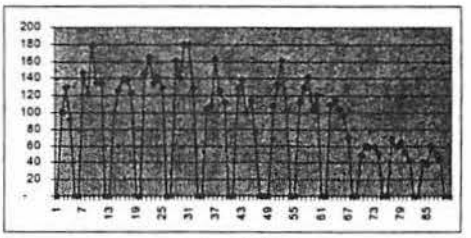
soup



Fountain juice



Total beverage

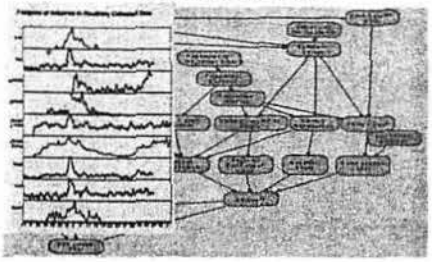


Fresh fruits

2.g.5 1.b.13

Scalable BioSurveillance Systems
Organization: Pitt-CMU PI: M.Wagner; co-PI K. Carley

Technical Approach



Objectives and Performance Goals

-Decision-theoretic surveillance system using passive and active data collection
-BioWar: Agent-based simulation Economic modeling for decision models
-Multiply number of data sources by 5
-Use historical models to reduce FPs
-Detect small outbreaks, use non spatial
-Fusing individual data with aggregate

Accomplishments

Data sets/test beds: Biosurv2, 911; poison center. Pittsburgh and Utah test beds.
Detection software: WSARE, Fast spatial scan, PANDA, and SecondStage
Simulation software: BioWar and economic model of anthrax release
Evaluations: Initial BioWar-based evaluation underway
Pitt-CMU

Programmatics

Evaluation Milestones: Ongoing tests using semisynthetic data. August program evaluations

Deliverables and Transitions: Detection and simulation software to DARPA contractors, PA and Utah

Sharing: WSARE and BioWar data=>Veridian

Who We Are

Detection
Greg Cooper MD, PhD
Andrew Moore, PhD
Mike Wagner, MD, PhD.
Rich Tsui, PhD
Weng Keen Wong
Virginia Dato MD MPH
William Hogan, MD, MS
Ed Krenzelok, PharmD
Wendy Chapman PhD
Bob Olszewski, Ph.D.
Jeremy Espino MD
Xiaoming Zeng MD
Oleg Ivanov MD, MPH
Zhongwei Liu MS
Zhen Liu, M.S.

Simulation
Kathleen Carley PhD
Doug Fridsma MD
Mark Roberts MD, MS
Scott Braithwaite, MD

Privacy
Ravi Sharma PhD
George Duncan PhD
Rema Padman PhD

Poison Center Project
Ed Krenzelok PharmD
Rita Mrvos RN
Xiaoming Zeng MD

County Health Dept.
Dave Piposzar MPH

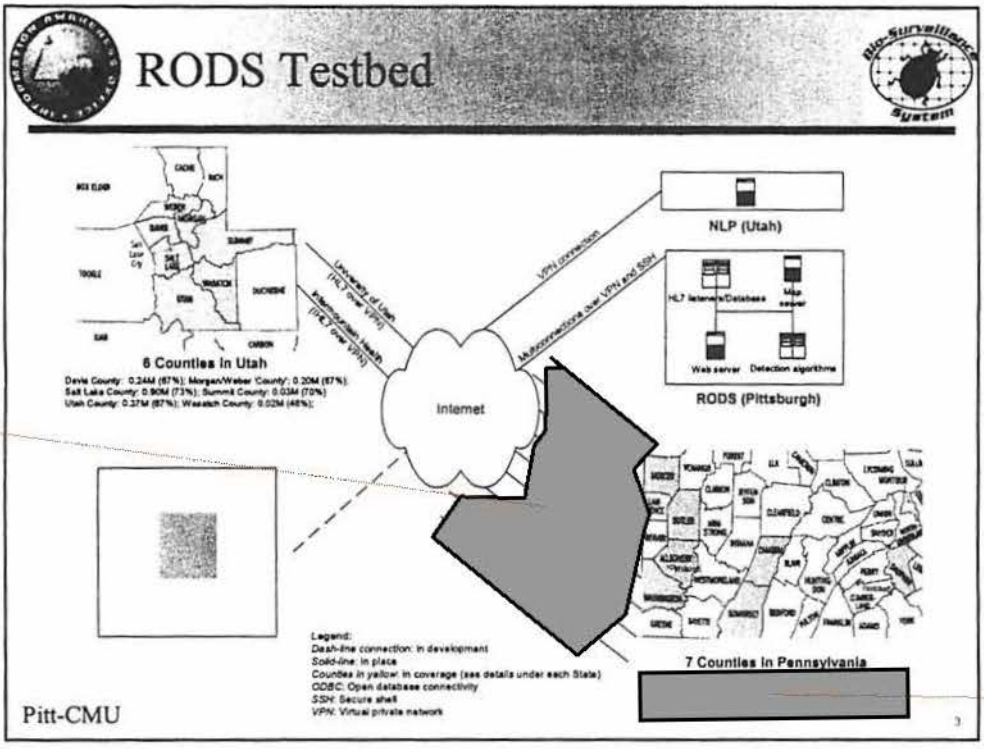
Hospital Infection Control Project
Rich Tsui, PhD
Carlene Muto, M.D., M.S.
Panos Chrysanthis PhD
Haobo Ma MD
Bill Pasculle, DSc

[redacted] (b)(4)

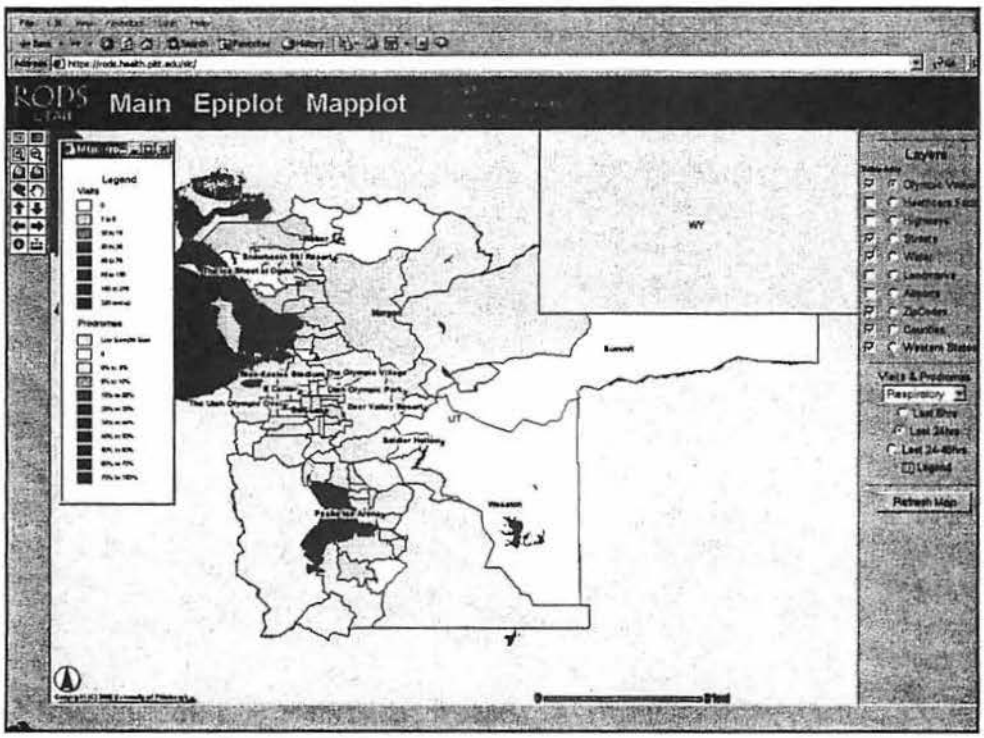
Pitt-CMU

Current Funders: DARPA, AHRQ, CDC, NLM (training grant)

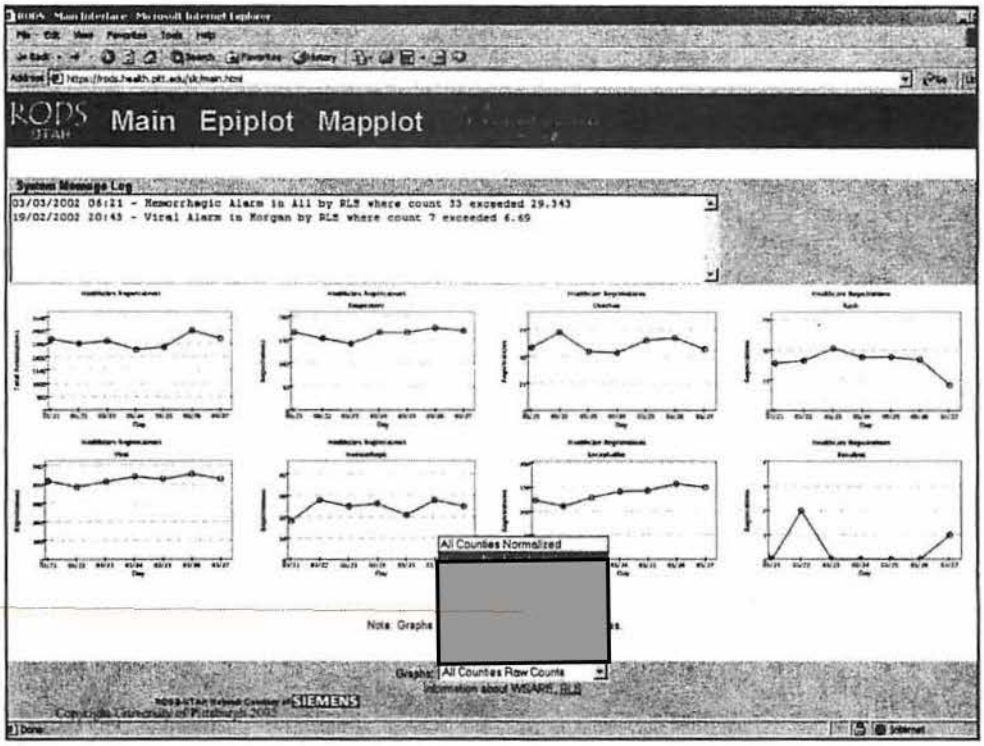
(b)(4)

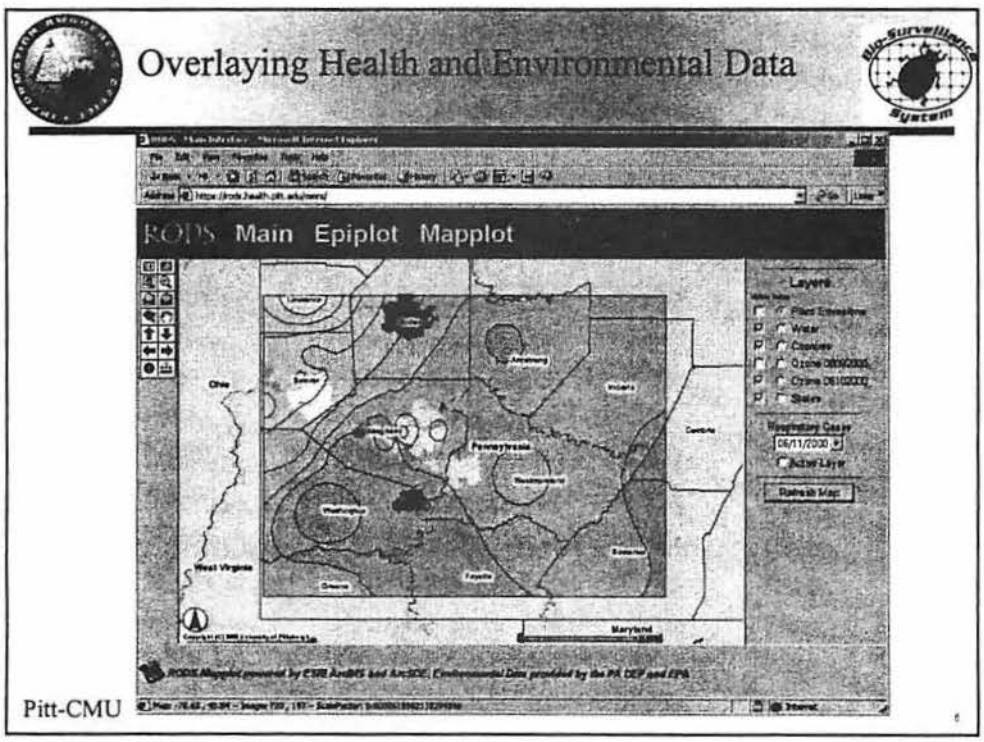

RODS Testbed
6 Counties in Utah
Internet
NLP (Utah)
RODS (Pittsburgh)
7 Counties in Pennsylvania
Legend:
Dash-line connection: in development
Solid-line: in place
Counties in yellow: in coverage (see details under each State)
ODBC: Open database connectivity
SSH: Secure shell
VPN: Virtual private network
Pitt-CMU

(b)(4)


RODS
Main
Epiplot
Mapplot
Legend
Layers
Refresh Map

(b)(4)


RODS Main Epiplot Mapplot
System Message Log
03/03/2002 06:21 - Hemorrhagic Alarm in All by RLE where count 33 exceeded 29.343
19/02/2002 20:43 - Viral Alarm in Morgan by RLE where count 7 exceeded 6.69
All Counties Normalized
Graphs: All Counties Raw Counts
SIEMENS


Overlaying Health and Environmental Data
RODS Main Epiplot Mapplot
Layers
Pennsylvania
West Virginia
Maryland
Pitt-CMU

Technical Approach

- What Are the New Ideas?
  - Algorithms
  - Bayesian/decision theoretic, economic models
  - Agent-based simulations (Biowar)
  - SecondStage

Algorithms

Method	Has Pitt/CMU tried it?	Tried but little used	Tried and used	Under development	Multivariate signal tracking?	Spatial?
Time-weighted averaging	Yes	Yes				
Serfling	Yes		Yes			
ARIMA	Yes	Yes				
SARIMA + External Factors	Yes		Yes			
Univariate HMM	Yes		Yes			
Kalman Filter	Yes	Yes				
Recursive Least Squares	Yes		Yes			
Support Vector Machine	Yes	Yes				
Neural Nets	Yes	Yes				
CuSUM	Yes	Yes				
Randomization	Yes		Yes	Yes		
Spatial Scan Statistics	Yes			(w/ Howard Burkom)	Yes	Yes
Bayesian Networks	Yes			Yes	Yes	
Contingency Tables	Yes		Yes			
Scalar Outlier (SQC)	Yes	Yes				
Multivariate Anomalies	Yes		Yes		Yes	
Change-point statistics	Yes			Yes		
FDR Tests	Yes		Yes		Yes	
WSARE (Recent patterns)	Yes		Yes	Yes	Yes	Yes
PANDA (Causal Model)	Yes			Yes	Yes	Yes
FLUMOD (space/Time HMM)				Yes	Yes	Yes

Details of these methods and bibliography available from "Summary of Biosurveillance-relevant statistical and data mining technologies" by Moore, Cooper, Tsui and Wagner. Downloadable (PDF format) from www.cs.cmu.edu/~awm/biosurv-methods.pdf

WSARE v2.0

- Inputs:
  1. Date/time-indexed biosurveillance-relevant data stream
  2. Time Window Length
  3. Which attributes to use?
- Outputs:
  1. Here are the records that most surprise me
  2. Here's why
  3. And here's how seriously you should take it

Normally, 8% of cases in the East are over-50s with respiratory problems.

But today it's been 15%

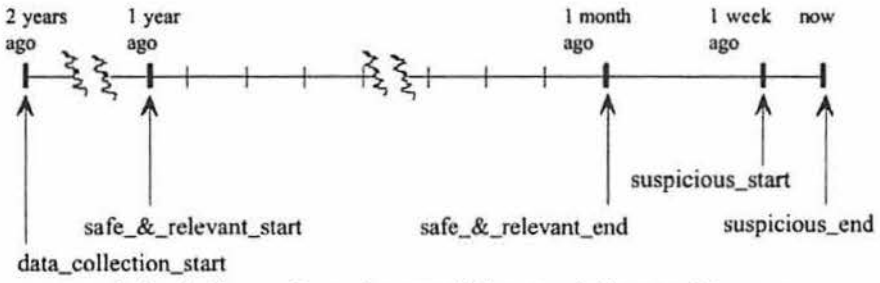
Don't be too impressed!

Taking into account all the patterns I've been searching over, there's a 20% chance I'd have found a rule this dramatic just by chance

Primary Key	...	Home			Work			Recent Weather	(Many more...)
		Large Scale	Medium Scale	Fine Scale	Large Scale	Medium Scale	Fine Scale		
h6r32		E	15						
t3q15		E	15						
t5hh5	6/2/2 14:15 Smith-field 622 Respiratory F 80s SE 15								

Pitt-CMU

PANDA (Patient-based ANomaly Detection Algorithm)



Ask whether a Bayesian predictive model learned from a 'safe and relevant period is the best predictor of current activity OR whether 2 models are better

Compute the likelihood ratio of 2 models versus one model

Ask that question for all recent time intervals exhaustively and pick the one with the biggest ratio…

Pitt-CMU

PROJECT UPDATE

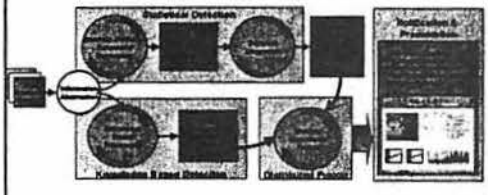
Veridian-Stanford Team

2nd PI Meeting

4-6 June 2002

Real-time Population Health Detector (RPHD)
Veridian-Stanford/McClellan

Technical Approach



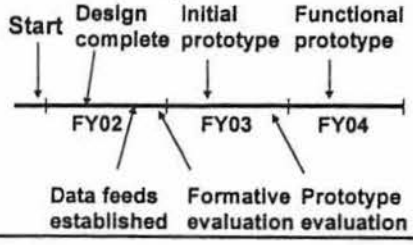
Objectives and Performance Goals

- Achieve early detection of outbreaks by use of behavioral, pre-diagnostic clinical, and animal data
- Improve detection and characterization of disease outbreaks through fusion of behavioral and medical indicators
- In initial tests, demonstrate potential of 25% lives saved; measure detection sensitivity, specificity and timeliness
- Improve ROC curves and timeliness of detection with non-traditional data sources and methods

Accomplishments

- Designed prototype for integration of statistical detection algorithms with Stanford knowledge-based methods
- Tested prototype spatio-temporal Kalman filter module in map-based GUI
- Signed MOU to acquire non-traditional data from City of Norfolk
- Performed exploratory analysis of city water data for Norfolk

Programmatics



Technical Approach -- New Ideas I

- What Are the New Ideas, How Achieved?
  - *Cough detection* as a non-traditional indicator of public health
    - Install audio system in low-noise environments such as classrooms
    - Use frame-based audio classification via feature analysis
    - Analyze coughs/person/unit time across multiple locations
    - Develop normalcy model accounting for air quality
  - *Utility usage* as a non-traditional indicator of public health
    - Have established real-time collection of water usage data at 15 min intervals for Norfolk
    - Will detect aberrations against a dynamical "normalcy" model for diurnal and weekly cycles
  - Spatio-temporal *Kalman filter* for aberration detection
    - Exploit temporal methods from econometrics and radar technology
    - Improve event detection by utilizing spatial correlations

Technical Approach -- New Ideas II

- What Are the New Ideas, How Achieved?
  - Integration of *statistical* and *knowledge-based* processing
    - Develop ontology for bio-surveillance data sources and associated Problem Solving Methods (PSMs) using Stanford's Protégé 2000
    - Utilize ontology to structure the RPHD system and invoke PSMs in real-time operation
  - *Anomaly detection* based on automated forecasting with dynamical models
    - Build adaptive background models in the Kalman filter detectors
    - Provides dynamically adjusted normalcy models for optimum aberration detection
  - *Multiple hypothesis* tracking
    - Utilize multiple Kalman filters to simultaneously evaluate alternate hypotheses regarding the state of the population health

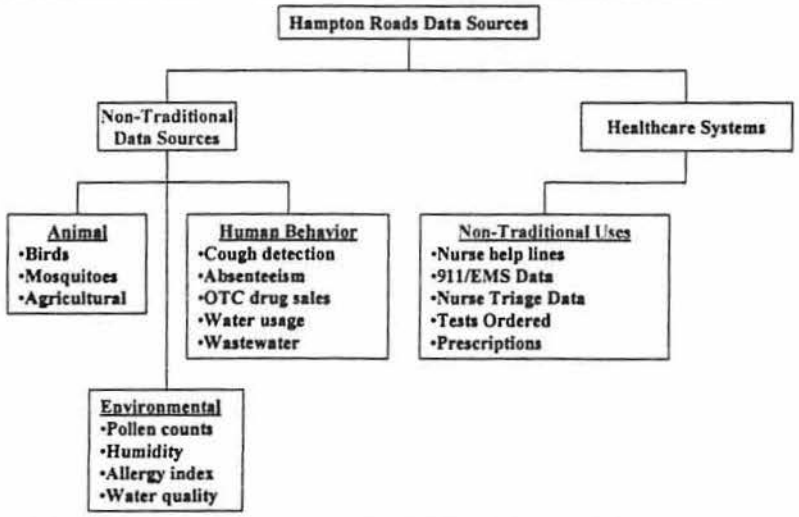
Technical Approach

- What Attack Scenarios Could be Detected?
  - Qualitatively
    - Aerosol/Large area - - can be detected early
    - Neighborhood/campus - - Small attacks can be detected in a limited geographical area if those affected do not disperse
    - Detection generally good for all vectors unless victims are widely dispersed spatially and/or temporally
  - Quantitatively
    - Detection probability depends on relative effect size as well as on the false-alarm rate that can be tolerated
      - Depends *both* the anomalous activity level following a particular attack *and* the normal variability distributed across all data sources. The former must be estimated by simulated attack data; the latter by time-series forecasting of multiple sources.
  - Specific estimates of the size of detectable attacks and the timing and reliability of detection will be estimated for Norfolk by Monte Carlo testing.

5

Technical Approach

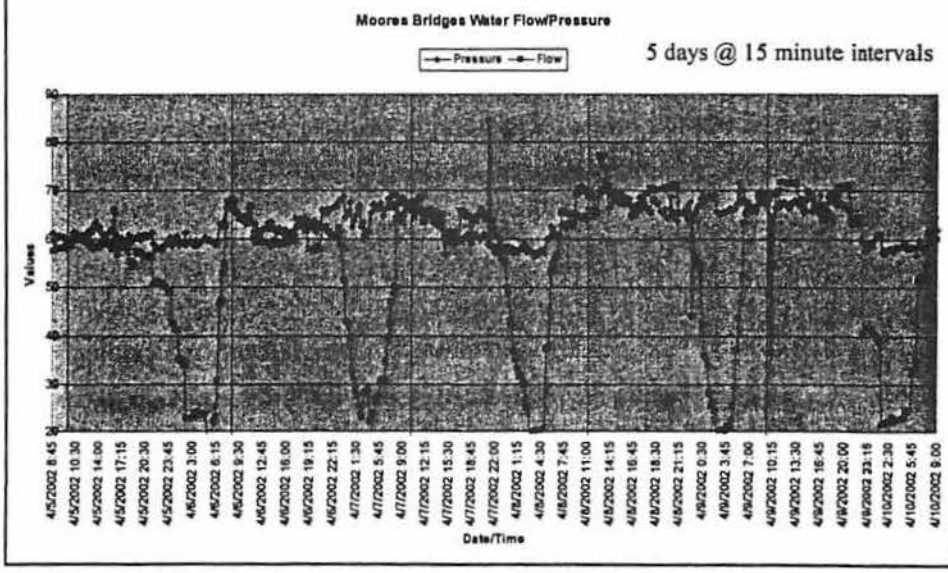
- What Are the Non-traditional Data Sources?



Technical Approach

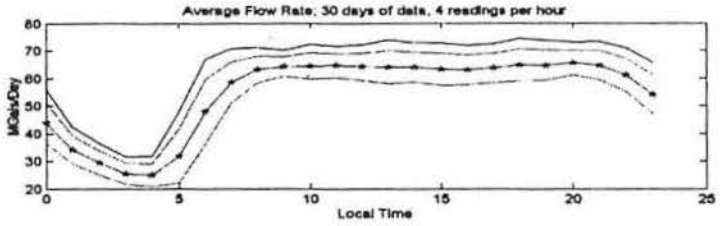
- What Non-traditional Data Have You Selected, Acquired, Characterized, and Analyzed ? How ?
  - Water usage at 15 minute intervals from Norfolk plant via e-mail data service (see next two slides)
- How Do They Serve as Leading Indicators of Outbreaks?
  - They are relatively direct consequences of the documented signs/symptoms of illness
- On What Are They Based (e.g., empirical analysis, behavioral hypothesis/research, etc.)?
  - Relevance as leading indicators is based on the progression of illness for bio-agent-induced disease
  - Detectability (SNR) is being analyzed

Sample Water Usage in Norfolk



What is Detectable in Water Usage?

30 Day Avg Flow Rate and Pressure (4 Readings Per Hour) With
1 std interval (~+/-6 MG/Day) and Upper 95% prediction limit



Population of Norfolk/Portsmouth is over 330,000.

Rough Calculation
100,000 people ill with 5 flushes (2 gal ea.) extra over 8 hours is about one standard deviation.

Technical Approach

- What Advanced Analytics are being Used/Developed for Detection?
  - Knowledge-based processing of pre-diagnostic data
  - Spatio-temporal Kalman Filter (next two slides)
  - Cough detection by clustering techniques
  - Tightly integrated architecture
- Characterization?
  - NBC CREST will be used for attack and weapon characterization

Kalman Filter Formalism for a Population Health Detector

State Vector:

$$\mathbf{x}_t = \begin{bmatrix} T_t \\ G_t \\ \vdots \\ R_t \end{bmatrix}$$

T_t = Population fever distribution

G_t = Gastrointestinal distress index

$\vdots$

R_t = Respiratory distress index

Measurement Vector:

$$\mathbf{z}_t = \begin{bmatrix} (OTC_i)_t \\ (ABS_j)_t \\ \vdots \\ (CD_k)_t \end{bmatrix}$$

OTC_i = OTC drug sales, location i

ABS_j = Absentee counts, location j

$\vdots$

CD_k = Cough rates, location k

State Vector: represents aggregate state of population health

Measurement Vector: current observations related to population health

11

Kalman Filter Formalism (continued):

Also Required:

1. *Measurement equation*: $\mathbf{z}_t = f(\mathbf{x}_t, t)$
   Relates aggregate health parameters to behavioral data (as in CMU BioWar). Can be function, result of numerical integration, lookup-table output, etc.
2. *State update equation*: $\mathbf{x}_t = g^{\mu}(\mathbf{x}_{t-1})$
   Updates state vector based on normalcy ($\mu = 0$) or known dynamics of disease μ .
3. *Estimates of errors* associated with measurements and disease models.

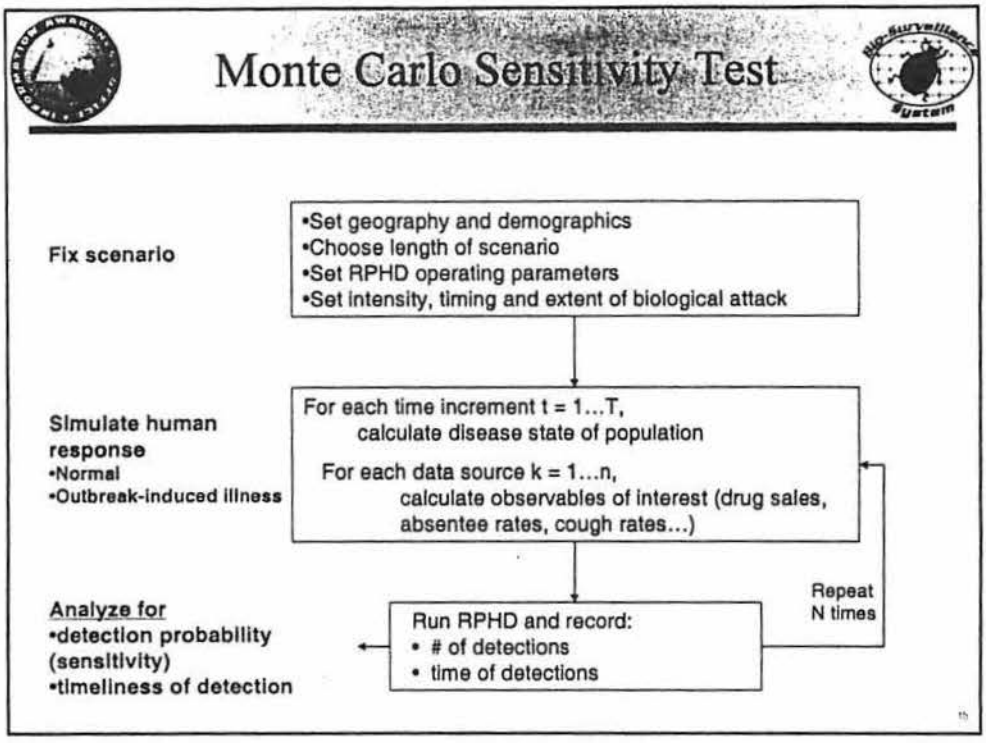
12

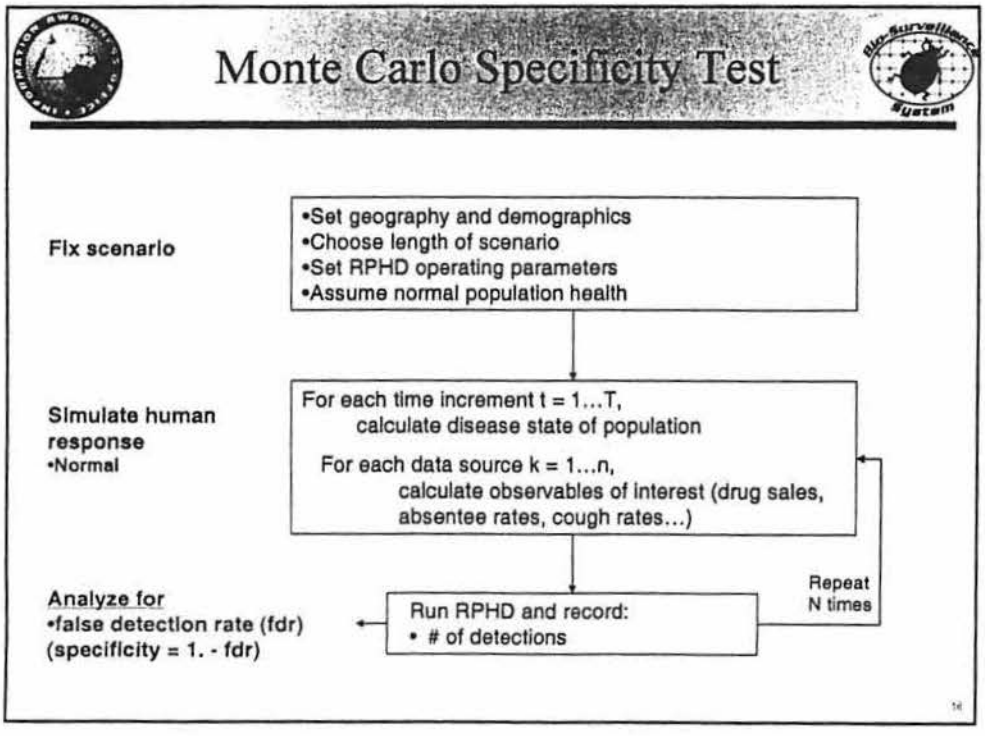
Technical Approach - - Privacy

- What New Privacy Protection Techniques are Being Developed ? How will they be Used ?
  - The Veridian-Stanford Team has not initiated any development efforts for privacy protection
  - Privacy protection will be assured by:
    - Utilization of privacy protection technology available from others
    - Developing a distributed architecture for RPHD that will work with the most sensitive patient data inside the originating institution
      - Takes advantage of a distributed network of local experts/custodians responsible for and familiar with each data source
      - Will export only abstracted/aggregated data to central public health site

Objectives and Performance Goals

- What Metrics will be used to Measure Progress?
  - *Sensitivity, specificity and timeliness* will be used to evaluate progress on RPHD
    - *Sensitivity* (detection probability for assumed attack)
    - *Specificity* (complement of false alarm probability)
    - *Timeliness of detection* relative to time of bioagent attack
    - Next two slides show Monte-Carlo method for determination of each
      - Can be done entirely with synthetic data
      - Can be done with simulated attack against real background data
      - Can be done against historical data with real outbreaks
  - Progress measured by incremental improvements in area of ROC curve
    - By adding data sources
    - By adding or improving analytical methods


Monte Carlo Sensitivity Test
Fix scenario
•Set geography and demographics
•Choose length of scenario
•Set RPHD operating parameters
•Set intensity, timing and extent of biological attack
Simulate human response
•Normal
•Outbreak-induced illness
For each time increment t = 1…T,
calculate disease state of population
For each data source k = 1…n,
calculate observables of interest (drug sales, absentee rates, cough rates…)
Repeat N times
Analyze for
•detection probability (sensitivity)
•timeliness of detection
Run RPHD and record:
• # of detections
• time of detections


Monte Carlo Specificity Test
Fix scenario
•Set geography and demographics
•Choose length of scenario
•Set RPHD operating parameters
•Assume normal population health
Simulate human response
•Normal
For each time increment t = 1…T,
calculate disease state of population
For each data source k = 1…n,
calculate observables of interest (drug sales, absentee rates, cough rates…)
Repeat N times
Analyze for
•false detection rate (fdr)
(specificity = 1. - fdr)
Run RPHD and record:
• # of detections

Objectives and Performance Goals

- What Internal Evaluations Will Be Performed in Addition to the Common August Project Evaluation?
  - Propose to use synthetic data for August evaluation
  - Additional evaluations will be performed as data sources are added
  - Will begin using real background data for evaluation in FY03
- What are the Performance Goals and Technical Milestones ?
  - Expect to demonstrate potential for 25% lives saved relative to traditional surveillance in August tests; double that figure within one year
  - Further goals are to increase number of non-traditional data sources and analyze contributions to timeliness of detection in FY03

17

Objectives and Performance Goals

- What are the Tradeoffs Between Sensitivity, Specificity, and Spatial/Temporal Coverage/Granularity?
  - Increased spatial granularity (resolution) will generally improve sensitivity through cluster analysis with modest loss in specificity
  - Qualitative characteristics of data sources listed in table below.

Data Type	Timeliness	Detectability (SNR)	Granularity	Specificity
Water Demand	poor-moderate	weak	low	low
School Absentee	moderate	moderate	moderate	low
OTC Drug	poor-moderate	weak	high	moderate
Clinic Appt Calls	poor	strong	high	moderate
Work Absentee	moderate	strong	moderate	low
Cough Rate (school)	moderate	weak-moderate	moderate	moderate
Clinic Visit (Flu Like)	poor	strong	high	high

18

Objectives and Performance Goals - I

- What Are the Key Deliverables and When Will They Be Available for Integration And/or Transition?

Technology Product	*Description*	*Potential Customers*	*Date of Availability*
RPHD	Real-time Population Health Detector	Naval Medical Center, Portsmouth Hampton Roads Regional Epidemiologist	February 2004 functional prototype December 2004 fully deployed

Objectives and Performance Goals - II

- What Are the Key Deliverables and When Will They Be Available for Integration And/or Transition?

Technology Product	*Description*	*Potential Customers*	*Date of Availability*
Cough Detector	A component of RPHD for the detection the sounds of coughing in a domestic, business, or classroom environment using frame-based sound classification.	Any Bio-Surveillance Data Collection System	January 2003
Water Usage Analysis Tool	A component of RPHD to analyze water usage data as an indicator of population health.	Any Bio-Surveillance Data Collection System	April 2003
Wastewater Analysis Tool	A component of RPHD to analyze wastewater data as an indicator of population health.	Any Bio-Surveillance Data Collection System	June 2003
Ontologies	Ontologies authored in Protégé 2000 for standard data formats such as NEDSS Logistical Data Model (NLDM).	Bio-ALIRT Program participants	Dec 2002

Accomplishments

- What Has Been Accomplished So Far ?
  - Negotiating data agreements, some data collection
  - Software development
- What Obstacles Have Been Encountered ?
  - Arrangement of MOUs and agreements
  - Scheduling meetings
- Where Has Progress Exceeded Expectations ?
  - Uniformly positive expressions of support and cooperation from both military and civilian leaders, medical, public health personnel in Hampton Roads
- What Lessons have been Learned?
  - Get to the legal department early

21

Accomplishments

- What Specific Transition Opportunities Have Been Identified And/or Exploited?
  - Who is Using Your Product Now?
    - BASIS component is part of DARPA Syndromic Surveillance System
  - Who has Committed to Using it in the Future?

(b)(4)

[redacted]

  - What Other Opportunities are there ?
    - Additional jurisdictions in Hampton Roads area
    - Additional hospital systems

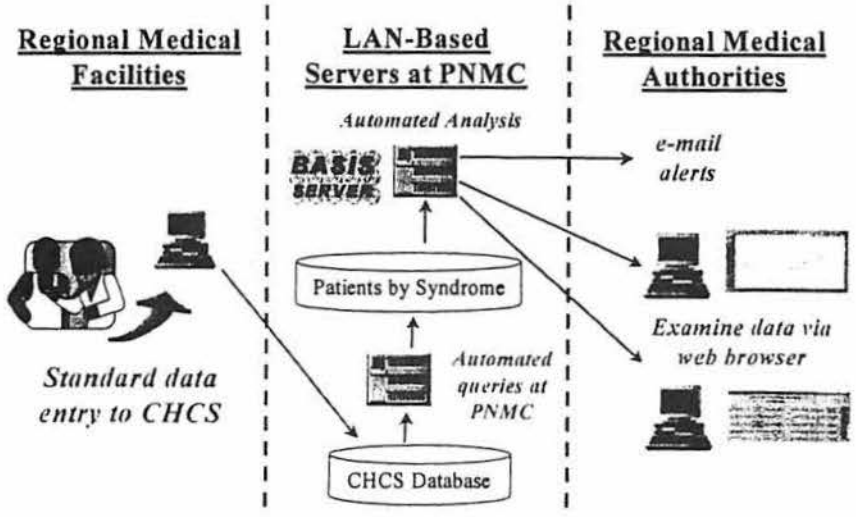
22

Accomplishments

- What Specific Transition Opportunities Have Been Identified And/or Exploited?

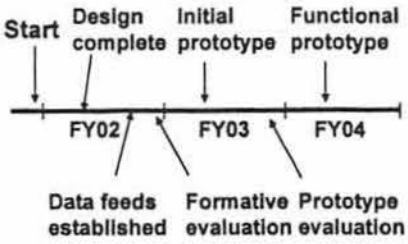
Technology Product	Description	Potential Customers	Date of Availability
BASIS	An RPHD component for automated statistical monitoring of electronic medical records using an adjustable threshold that alerts users to statistically significant deviations from a standard baseline.	Naval Medical Center, Portsmouth Other medical treatment facilities	February 2002 Status: Pending completion of MOU between AFRL Rome and NMCP

BASIS Automated Alerts for PNMC



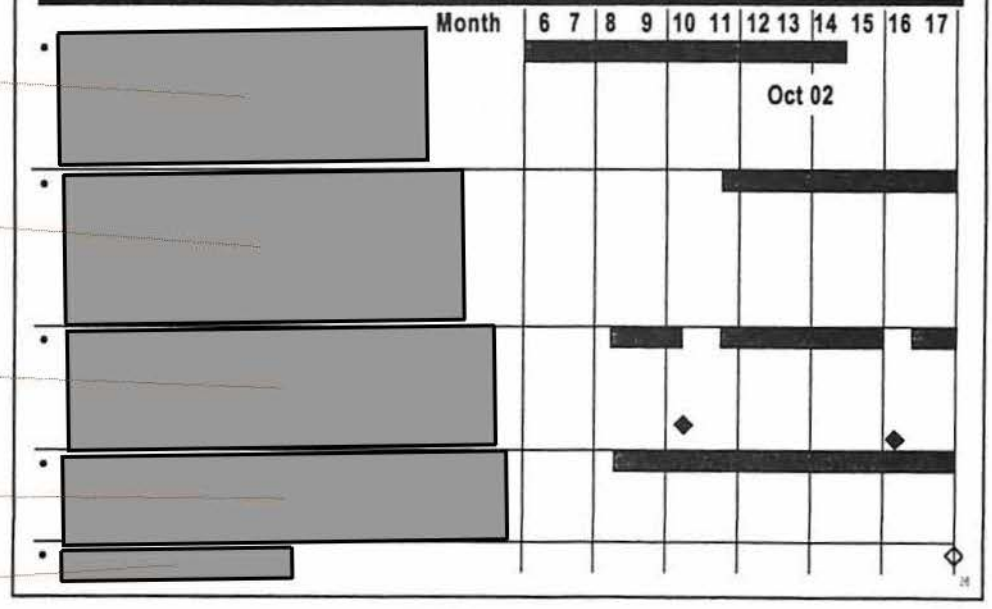
Programmatics

- Identify Current Project Plan, Including Key Milestones for Evaluations, Deliverables, and Transitions



– *Phase II details on next slide*

Research Plan -- Phase 2



Programmatics

- What Else is Needed?
  - Data Sources to be Acquired, Characterized, and Analyzed ?
    - Data involving known outbreaks for performance assessment
  - Other ?
    - Data de-identification software such as Datafly
    - Behavioral models predicting observables based on illness

Programmatics

- What, if anything, Has Changed From Your Original Plan? Why ?
  - Because of the availability of BioWar, have spent less developing simulation
  - Improved integration of Stanford and Veridian components in composite software architecture
  - Driven toward more distributed processing by privacy concerns

Programmatics - -Sharing

- How are you Sharing with Other Projects?

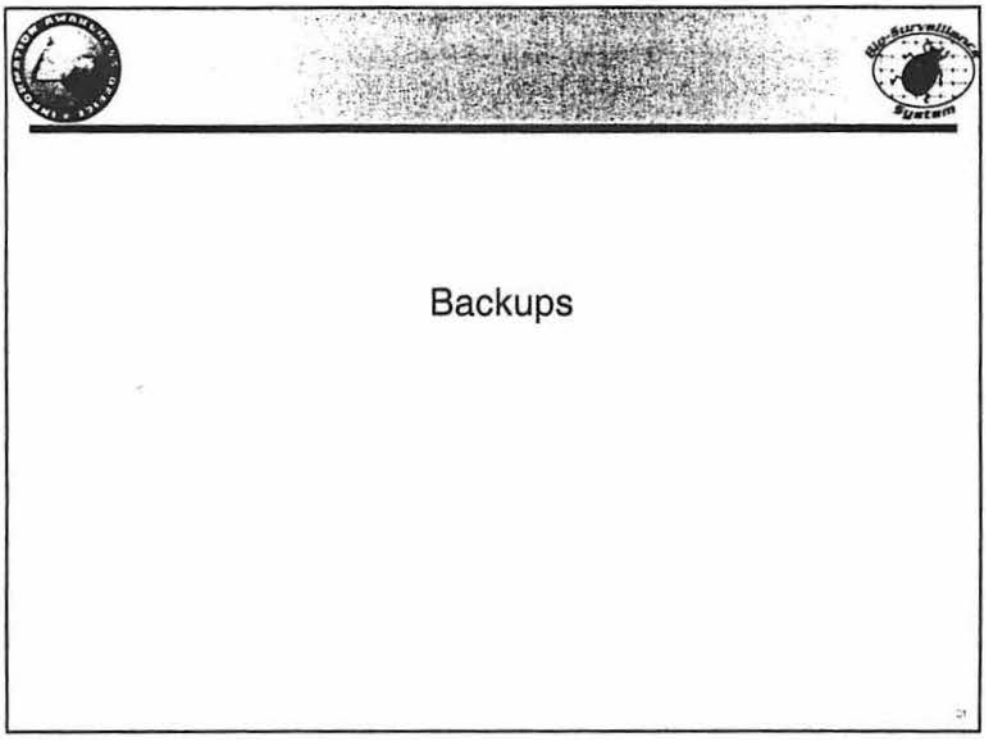
Technology Product	Description	Potential Customers	Date of Availability
AMedP-8 Models	Consequence Analytic Tools for NBC Operations: Volume 1 – Biological Agent Effects and Degraded Personnel Performance for Tularemia, Staphylococcal Enterotoxin B (SEB) and Q-Fever. Biological human response models for anthrax, plague, botulism, VEE.	CMU BioWar	Provided to CMU February 2002
NBC CREST	Medical planning tool for NBC attacks that will evaluate the impact of prophylactic treatment on morbidity and mortality as a function of time of administration - - developed for U.S. Army OTSG	DARPA Bio-Surveillance Team Validation for "Lives Saved" by early detection	June 2002 pending release by OTSG
Protégé 2000	A tool for development and management of ontologies that define conceptual data models.	Developers wishing to model data sources and knowledge required to interpret data.	Immediate

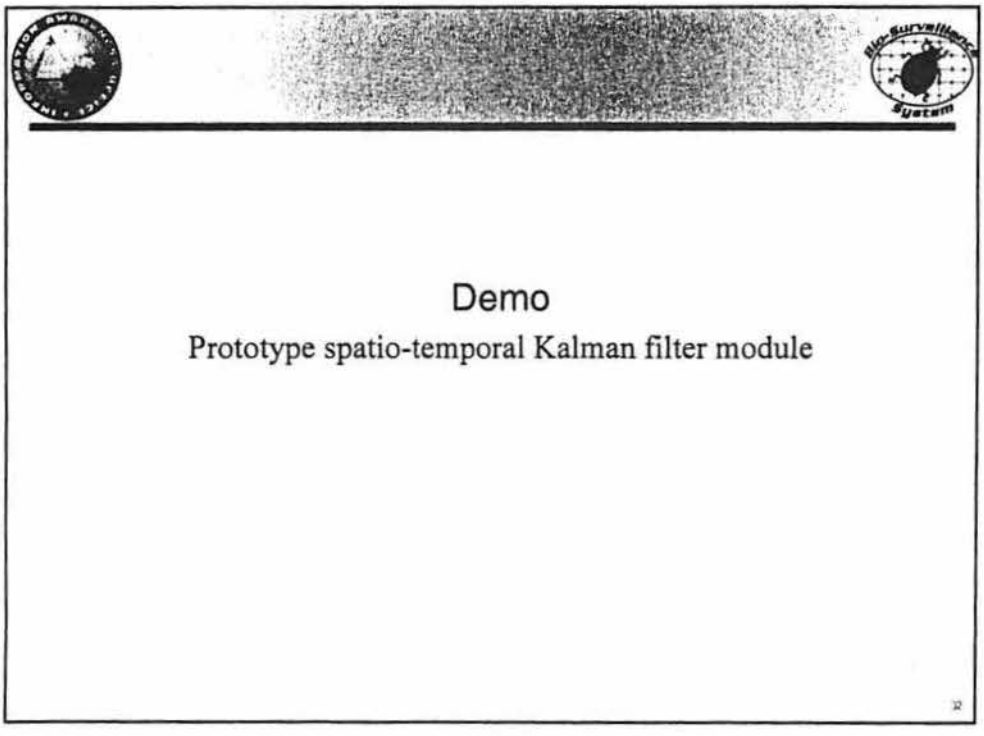
29

Programmatics

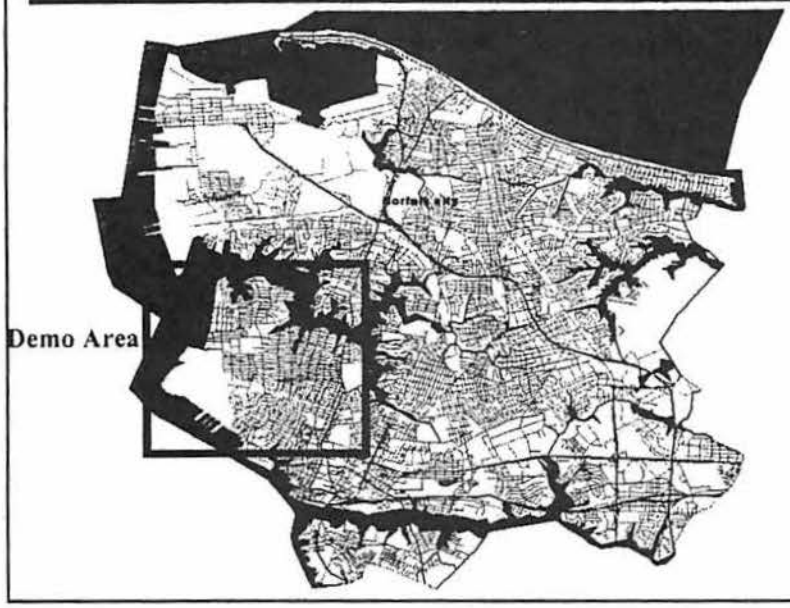
- How are you Collaborating/Sharing with Other Projects?
  - Have provided background material on models of bio-agent-induced disease to CMU
  - Have acquired BioWar from CMU
    - Tested code
    - Generated data for simulation
    - Provided constructive feedback on BioWar
  - Acquired WSARE by CMU for testing

30


Backups


Demo
Prototype spatio-temporal Kalman filter module

Norfolk, Virginia



RPHD Use of BioWar Data

- Simulation data generated using CMU BioWar software (5/7/02 delivery)
  - Current release has software bugs that prevent correct execution (e.g., run with only 3 schools)
- Experiment 1: Norfolk Simulation:
  - 3 schools
    - Monroe
    - Larchmont
    - Taylor
  - 3 drug stores each selling 5 OTC drug types
    - Cough
    - Cough / Cold
    - Cold
    - Aspirin
    - Imodium

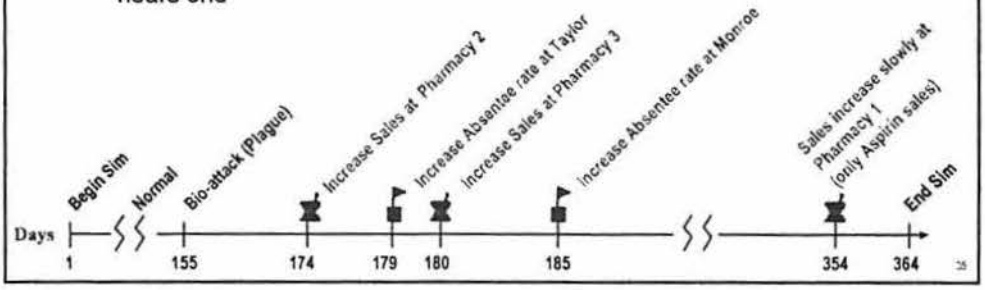
    Respiratory (all three types combined into one sales type)

    Fever

    Gastrointestinal
  - 364 days worth of "historic" data generated to provide statistics for normalcy models
  - 366 days worth of simulation data
  - 1 disease strain (plague) -- no background diseases (e.g., influenza)

Experiment 1 Simulation Scenario

- Bio-attack (Plague) occurs on day 155
- Attack occurs near Taylor school
- OTC drug sales rate increases at the closest pharmacy to Taylor (Pharmacy #2) following attack
- Absentee rate increases at Taylor school following attack (proceeds OTC drug sales increase)
- Disease begins to spread to neighboring area (Monroe school district)
- OTC drug sales begin to increase at pharmacy closest to Monroe (Pharmacy #3)
- Absentee rate increases at Monroe school
- OTC drug sales begin to increase slightly at Pharmacy #1 as simulation nears end



Experiment 1 Schools and Pharmacies

Three Schools: △
1) Monroe
2) Larchmont
3) Taylor

Three Pharmacies: ★
- Each covers 1 school zone

Bio-attack: Bubonic Plague

TITLE PAGE

Briefing Materials concerning TIA BIOALERT 1

FOLDER 2.h TOC 12

TRANSITIONS

- 2.h.1.a Transition Products
    - IBM T.J. Watson Research Center
- 2.h.1.b Transition Products List & Schedules
    - IBM T.J. Watson Research Center
- 2.h.1.c ESSENCE II Transition Products (Initial Only)
- 2.h.1.d Product Transition
    - Veridian-Stanford Team 2nd PI Meeting

Zuhl.A

12

Transition Products

IBM TJ Watson Research Center

Video Car Counting

- Count auto/truck traffic from video
- Useful for
  - Site entry/departure
  - Regional traffic levels
  - Disease spread models
- Available for download, as of June 30, via alphaworks

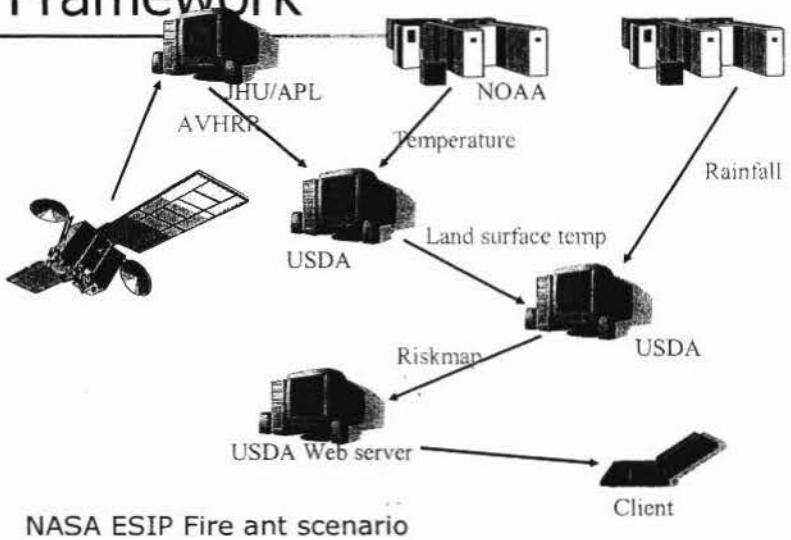
High Resolution Environmental Data

- Real-time and archived temperature and precipitation data from weather stations, with interpolation using satellite images
- Useful for
  - Modeling effects of weather on normal disease
  - Effects on attacks
- Available now for Northeast

Privacy (k-anonymity) System

- Anonymize sensitive data, using abstraction and suppression, to satisfy k-anonymity constraint
  - Usage-based metrics, more flexible generalizations, genetic algorithm search
- Useful for data with linkable fields (e.g., zipcode, DOB, etc.)
- Prototype currently running at Watson, can be made available, conditions TBD

Application Roaming Framework

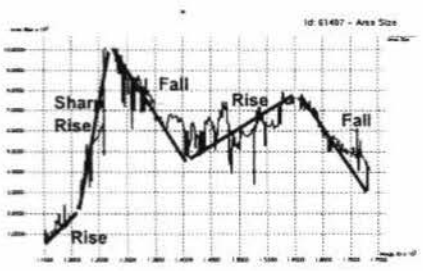


NASA ESIP Fire ant scenario

EpiMALM: Multiple Abstraction Level Mining of Time Series

- Expression of time-series behavior in terms of semantic labels at multiple abstraction levels
  - Content-based search based on time-series relatively unexplored
  - prior work on EKG, stock prices
- Extraction of time-series from imagery
- Extension of time-series language to permit expression of both intra- and inter-series and intra- and inter-object constraints (spatial and temporal)

- Example query:
  - Locate well logs sections which contain the pattern *sharply rising segment follow by flat segment follow by sharply falling segment*



2.4.1.B

Transition Products List & Schedules

Product Name	Description	Availability	Transition	Contact	Potential Customer	ACTION
IBM Video Car Counting		Now- generally available through www.alphaworks.ibm.com: 30 June 2002		Belle Tesng/ John R. Smith		
IBM Application Roaming Framework		Now- generally available through www.alphaworks.ibm.com: now		Yuan-Chi Chang		
IBM Tool for Interpolation of Daily Temperature and Moisture Data though use of AVHRR		Now		Yuan-Chi Chang		
IBM MALM	Tool for search specific temporal pattern based on multiple abstractions of features and semantics	Now		Yuan-Chi Chang		
IBM Phone Record, Proxy Log Detector, Cafeteria Data Detector, Sewage Data Detector		Estimated availability: July 2002		Murray Campbell		
IBM Active Surveillance		Estimated availability: July 2002		Murray Campbell		
IBM K-anonimity Privacy Protection Module		Estimated availability: September 2002		Vijay Iyengar		
IBM Multi-Data Source Fusion Module		Estimated availability: July 2002		John R. Smith		
MP Second Stage Detection	A client-server Java component that can be embedded into existing systems. The main initial functions are communicating requests to data collectors and investigators of epidemics, providing forms with dynamic questioning strategies for data collectors, provision of statistical output for dynamic questioning strategies for data collectors, provision of statistical output for epidemic analysis. Messaging standard is XML.	Version 1: June 2002	all contractors, NEDSS			
MP WSARE	"What's strange about recent events"	Version 1: January 2002; Version 2: April 2002; Version 3 (environment sensitive): July 2002	all contractors			
MP Spatial Scan Very Fast Algorithims for Classic Scan	Joint work with Howard Burkhom of JHU/APL	Version 1: July 2002	At minimum, JHU and Pitt/Mellon; probably all contractors			
PANDA	The PANDA (Patient-based Anomaly Detection and Assesment) system will use a Bayesian approach to event detection. As such, it will derive posterior probabilities of many possible causes of the data being observed by the system. Such probabilities will be useful in directly assessing the etiologies acting in a given regional population. Morever, these probabilities can be part of a decision theoretic system that determines when to activate an alarm and then recommends what additional data to collect. PANDA will use Bayesian networks as its core representation.	September 1, 2002	All contractors			
MP BioWAR Simulation	City level multi-agent network model for examining weaponized biological attacks. It can be used to aid detection by providing realistic data and to aid intelligence by providing understanding of how weaponized attacks might play out. Additionally, it can be used to aid response training by providing health workers with a system for learning what they might expect.	Data available to other contractors: mid July 2002. Estimated availability of version to be used by other contractors in stand-alone fashion: November 2002				
MP Bioevent Economic Models	We have developed a general purpose markov model of an anthrax bio terrorism attack that models different attack characteristics, and response strategies. The model is sensitive to time-dependent characteristics of the release and resource constraints of the response.	Version 1: July 1, 2002	All contractors			
Veridian RPHD	Real-time Population Health Detector	February 2004 functional prototype	Naval Medical Center, Portsmouth			
		December 2004 fully deployed	Hampton Roads Regional Epidemiologist			
Veridian Cough Detector	A component of RPHD for the detection the sounds of coughing in a domestic, business, or classroom environment using frame-based sound classification.	Jan-03	Any Bio-Surveillance Data Collection System			

Veridian Water Usage Analysis Tool	A component of RPHD to analyze water usage data as an indicator of population health.	Apr-03	Any Bio-Surveillance Data Collection System			
Veridian Wastewater Analysis Tool	A component of RPHD to analyze wastewater data as an indicator of population health.	Jun-03	Any Bio-Surveillance Data Collection System			
Veridian BASIS	An RPHD component for automated statistical monitoring of electronic medical records using an adjustable threshold that alerts users to statistically significant deviations from a standard baseline.	Feb-02	Naval Medical Center, Portsmouth			
			Other medical treatment facilities			
		Status: Pending completion of MOU at NMCP				
Veridian AMedP-8 Models	Consequence Analytic Tools for NBC Operations: Volume 1 – Biological Agent Effects and Degraded Personnel Performance for Tularemia, Staphylococcal Enterotoxin B (SEB) and Q-Fever.	Provided to CMU February 2002	CMU BioWar			
	Biological human response models for anthrax, plague, botulism, VEE.					
Veridian NBC CREST	Medical planning tool for NBC attacks that will evaluate the impact of prophylactic treatment on morbidity and mortality as a function of time of administration - - developed for U.S. Army OTSG	June 2002 pending release by OTSG	DARPA Bio-Surveillance Team Validation for "Lives Saved" by early detection			
ESSENCE World View	Provides real-time estimates of abnormal syndrome levels for all military medical facilities world wide. Gives preventive medical officers a heads up for potential outbreaks at their post / facility with detailed information to permit an investigation.	Currently operational	All military medical facilities	MAJ Julie Pavlin		
ESSENCE II ER Complaint Classifier	Automates the collection of chief complaint data and conversion into syndrome groupings	Currently operational for Montgonery County Hospitals	DoD-GEIS & All hospitals in the NCA	Rich Wojcik		
ESSENCE II Spatio -Temporal Detector	Forms clusters based on higher than normal levels of disease present in a geographic area using data streams from several disparat sources	Currently operational	DoD-GEIS & All the public health jurisdictions in the NCA	Howard Burkom		
ESSENCE II Watch Board	The ESSENCE II watch board provides a real-time view of the status of the data collection and outbreak detection processes for the NCA	Currently operational	will transition to DoD-GEIS in FY04 & DHMH in FY03	Rich Wojcik		
ESSENCE II Temporal Detection Suite	Performs estimates of disease levels at variable spatial resolutions in the NCA.	The initial version is currently operation, but upgrades are expected to be completed by the end of the fiscal year	Outputs currently being provided to Maryland jurisdictions will transition to DoD-GEIS in FY04 & DHMH in FY03	Howard Burkom Eugene Elbert		
ESSENCE II Datafly	Scrubs raw data to remove or modify data fields to protect the privacy of individuals	operational at the end of FY02	Move from CMU to JHUAPL for integration early in FY03	Latanya Sweeney		
ESSENCE II Messaging	ESSENCE II Messaging is needed for the individual components of the surveillance system to operate in a coordinated fashion. It provides the business rules needed for the system to function and can be used as a foundation for the development of variant surveillance systems.	Messaging is nearing completion and will be available at the end of FY02	Will be used in the Maryland & GEIS systems	Rich Wojcik		

Transition Products List & Schedules

Product Name	Description	Availability	Transition	Contact	Potential Customer	ACTION
IBM Video Car Counting	Count car/truck traffic from video. Useful for site entry/departure, regional traffic levels, disease spread models.	Available June 30, 2002 at www.alphaworks.ibm.com	All	Belle Tseng/John R. Smith	MP	
IBM Application Roaming Framework	Framework for integrating disparate data sources.	Currently available at www.alphaworks.ibm.com	All	Yuan-Chi Chang	All contractors	
IBM Tool for Interpolation of Daily Temperature and Moisture Data though use of AVHRR	High resolution environmental data, currently available for Northeast.	Now	All	Yuan-Chi Chang	All contractors	
IBM MALM	Tool for search specific temporal pattern based on multiple abstractions of features and semantics	Now	All contractors	Yuan-Chi Chang	JHU/APL	
IBM Cough Detector	Detect coughs from audio.	Currently a research prototype. Can be made available to other contractors.	All contractors	Milind Naphade	Veridian	
IBM k-anonymity Privacy Protection Module	Anonymize sensitive data using k-anonymity model. Usage-based metrics, more flexible generalizations, genetic algorithm optimization.	Currently a research prototype. Can be made available to other contractors.	All contractors	Vijay Iyengar	Veridian	
IBM Multi-channel detectors	Anomaly detectors in multi-channel data. Currently under evaluation.	Estimated availability: September 2002	All contractors	Milind Naphade	All contractors	

ESSENCE II Transition Products

Initial Transition Products Only

ESSENCE World View
ER Complaint Classifier
Spatio -Temporal Detector
Watch Board
Temporal Detection Suite
ESSENCE II Datafly
Architecture Messaging

2.A.1.c

ESSENCE II Transition Products

ESSENCE World View	Provides real-time estimates of abnormal syndrome levels for all military medical facilities world wide. Gives preventive medical officers a heads up for potential outbreaks at their post / facility with detailed information to permit an investigation.
ESSENCE II ER Complaint Classifier	Automates the collection of chief complaint data and conversion into syndrome groupings
ESSENCE II Spatio -Temporal Detector	Forms clusters based on higher than normal levels of disease present in a geographic area using data streams from several disparat sources
ESSENCE II Watch Board	The ESSENCE II watch board provides a real-time view of the status of the data collection and outbreak detection processes for the NCA
ESSENCE II Temporal Detection Suite	Performs estimates of disease levels at variable spatial resolutions in the NCA.
ESSENCE II Datafly	Scrubs raw data to remove or modify data fields to protect the privacy of individuals
ESSENCE II Messaging	ESSENCE II Messaging is needed for the individual components of the surveillance system to operate in a coordinated fashion. It provides the business rules needed for the system to function and can be used as a foundation for the development of variant su

Transition Products

PRODUCTS	IN USE BY	Dates Available
WSARE v1	Veridian	Jan 2002
v2		April 2002
v3		July 2002
FAST SPATIAL SCAN		July 2002
PANDA LR version		9/1/02
Bayesian		late fall 2002
RODS*	Utah, PA	Now
BioWar Data	Veridian	July 2002
stand-alone		November 2002
SecondStage		Now
Economic Model		July 2002

*RODS development supported by NLM, AHRQ, CDC

WSARE

Description: **Detection algorithm based on brute force search of medium dimensional spaces especially spatial and temporal**

Date available:

Version 1 available Jan 2002
Version 2 available Apr 2002

Version 3 (environment sensitive) July 2002

Transition customers:

All contractors

Contact: awm@cs.cmu.edu

Spatial Scan Very Fast Algorithms for Classic Scan (joint work with Howard Burkhom of JHU/APL)

Description:
Spatial Scan Very Fast Algorithms for Classic Scan

Date available:
Version 1 available July 2002

Transition customers:
At minimum, JHU and Pitt/Mellon: Probably all contractors

Contact: awm@cs.cmu.edu

RODS

Description:
Data collection system, interfaces and algorithms. Health system resident detection component.

Date available: now

Transition customers:
Already Utah, PA. Potentially Indiana

Contact: jhutman@cbmi.upmc.edu

BioWar

Description:

City level multi-agent network model for examining weaponized biological attacks. Can be used to aid detection by providing realistic data, to aid intelligence by providing understanding of how weaponized attacks might play out, to aid response training by providing health workers with a system for learning what they might expect.

Date available:

Input/Output list for customization by June 2002
Data available to other contractors mid July 2002
Version 1 to be used by other contractors in stand-alone fashion est. November 2002 (depends on getting list of data needs from contractors and having servers arrive on campus).

Transition customers: **Possibly other contractors, or software support firm like Aptima, or various government agencies.**

Contact: carley+@cmu.edu

PANDA

Description:

The PANDA (Patient-based ANomaly Detection and Assessment) system will use a Bayesian approach to event detection. As such, it will derive posterior probabilities of many possible causes of the data being observed by the system. Such probabilities will be useful in directly assessing the etiologies acting in a given regional population. Moreover, these probabilities can be part of a decision theoretic system that determines when to activate an alarm and then recommends what additional data to collect.

Date available:

Likelihood-ratio version of Panda will be available to all contractors on 9/1/02
The Bayesian version, will be available later in the fall of 2002

Transition customers: **all contractors**

Contact: Greg Cooper gfc@cbmi.upmc.edu

2.4.1.d.

PRODUCT TRANSITION

Veridian-Stanford Team

2nd PI Meeting

4-6 June 2002

Objectives and Performance Goals - I

- What Are the Key Deliverables and When Will They Be Available for Integration And/or Transition?

Technology Product	*Description*	*Potential Customers*	*Date of Availability*
RPHD	Real-time Population Health Detector	Naval Medical Center, Portsmouth Hampton Roads Regional Epidemiologist	February 2004 functional prototype – December 2004 fully deployed

Objectives and Performance Goals - II

- What Are the Key Deliverables and When Will They Be Available for Integration And/or Transition?

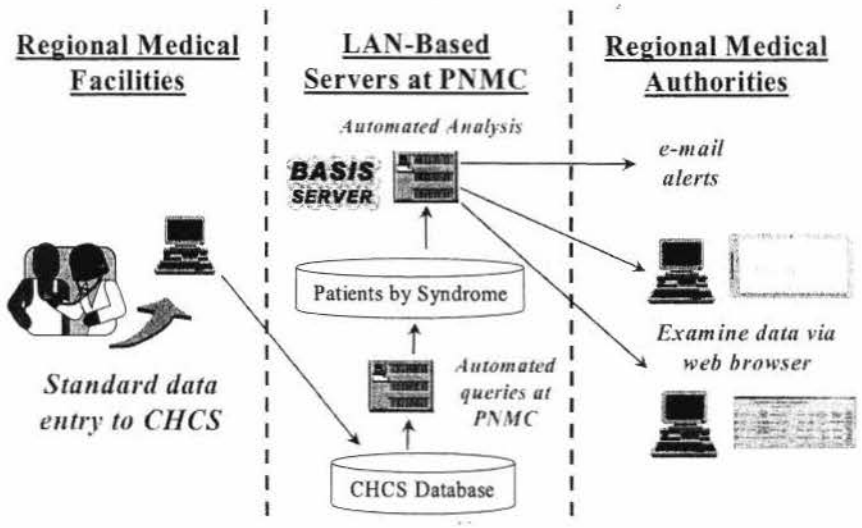
Technology Product	Description	Potential Customers	Date of Availability
Cough Detector	A component of RPHD for the detection the sounds of coughing in a domestic, business, or classroom environment using frame-based sound classification.	Any Bio-Surveillance Data Collection System	January 2003
Water Usage Analysis Tool	A component of RPHD to analyze water usage data as an indicator of population health.	Any Bio-Surveillance Data Collection System	April 2003
Wastewater Analysis Tool	A component of RPHD to analyze wastewater data as an indicator of population health.	Any Bio-Surveillance Data Collection System	June 2003
Ontologies	Ontologies authored in Protégé 2000 for standard data formats such as NEDSS Logistical Data Model (NLDM).	Bio-ALIRT Program participants	Dec 2002

Accomplishments

- What Specific Transition Opportunities Have Been Identified And/or Exploited?

Technology Product	Description	Potential Customers	Date of Availability
BASIS	An RPHD component for automated statistical monitoring of electronic medical records using an adjustable threshold that alerts users to statistically significant deviations from a standard baseline.	Naval Medical Center, Portsmouth Other medical treatment facilities	February 2002 Status: Pending completion of MOU between AFRL Rome and NMCP

BASIS Automated Alerts for PNMC



Programmatics - - Sharing

- How are you Sharing with Other Projects?

Technology Product	Description	Potential Customers	Date of Availability
AMedP-8 Models	Consequence Analytic Tools for NBC Operations: Volume 1 – Biological Agent Effects and Degraded Personnel Performance for Tularemia, Staphylococcal Enterotoxin B (SEB) and Q-Fever. Biological human response models for anthrax, plague, botulism, VEE.	CMU BioWar	Provided to CMU February 2002
NBC CREST	Medical planning tool for NBC attacks that will evaluate the impact of prophylactic treatment on morbidity and mortality as a function of time of administration - - developed for U.S. Army OTSG	DARPA Bio-Surveillance Team Validation for "Lives Saved" by early detection	June 2002 pending release by OTSG
Protégé 2000	A tool for development and management of ontologies that define conceptual data models.	Developers wishing to model data sources and knowledge required to interpret data.	Immediate

Nuclear, Biological and Chemical (NBC)
Casualty and Resource Estimation Support Tool (CREST)

NBC CREST demo

TITLE PAGE

FOLDER 1.a.1-9 TOC 13

Briefing Materials concerning TIA BIOALERT 2

MEETINGS/STATUS UPDATES

- 1.a.1 Scalable BioSurveillance Systems
    - Pitt-CMU
- 1.a.2 EpiSPIRE Project Update
    - IBM T.J.Watson Research Center
    - Johns Hopkins University
    - SDI Inc.
- 1.a.3 ESSENCE II Status Summary
    - DARPA Bio Surveillance PI Meeting-2Dec02
- 1.a.4 Bio-ALIRT Data Ontology
    - Stanford University School of Medicine
- 1.a.5 Dimentionality Red of Multiple Time Series Data
    - Veridian, Los Angeles
- 1.a.6 Are Knowledge-Based Approaches Extensible?
    - Stanford University

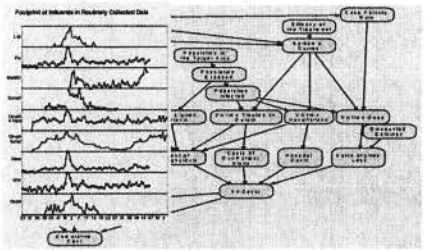
FOLDER 1.a.1-9 TOC 13 –Cont'd – Page 2

- 1.a.7 NT Data
- 1.a.8 Project Update – 3rd PI Meeting- 3-5Dec02
    - Veridian-Stanford Team
- 1.a.9 RPHD Analysis Techniques

1001

Scalable BioSurveillance Systems
Organization: Pitt-CMU PI: M.Wagner; co-PI K. Carley

Technical Approach



Objectives and Performance Goals

-Decision-theoretic surveillance system using passive and active data collection
-BioWar: Agent-based simulation
-Multiply number of data sources by 5
-Use historical models to reduce FPs
-Detect small outbreaks, using spatial and non spatial analysis
-Fusing individual data with aggregate

Accomplishments

Data sets/test beds: ED registrations, HC-4, 911, poison center, OTC, agribusiness. Indian-PA-Utah test bed.
Detection software: WSARE, Fast spatial scan, PANDA, BARD, and SecondStage
Simulation software: BioWar and economic model of anthrax release
Evaluations: with real and semi-synthetic data, and with synthetic BioWar data

Programmatics

Evaluation Milestones: Studies using real, semi synthetic, and synthetic BioWar data.

Deliverables and Transitions: RODS, WSARE and CoCo (natural language processing) released under free license. Simulation software to DARPA contractors, PA and Utah

Sharing: WSARE and BioWar data=>Veridian

Outline

- ACCOMPLISHMENTS from FY02
  - What's Working and What's Not?
  - Nontraditional Data: How Much Earlier Detection?
- PLANS for FY03
  - Quantitative Goals
  - Transition Path for the Technology
  - Ideas on Accomplishing Pathogen ID and Attack Characterization
  - Other Funding Sources and Division of Labor (if appropriate)
  - Further Evaluation/Validation Plans

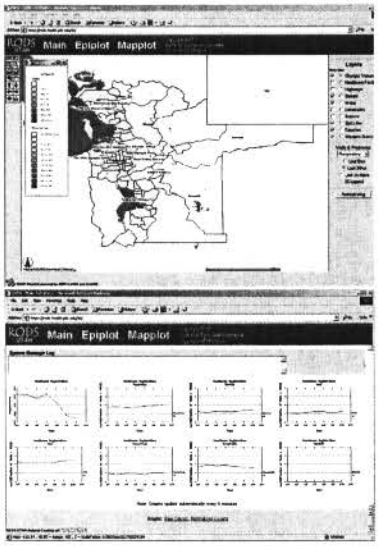
ACCOMPLISHMENTS from FY02

- RODS Winter Olympic and continued Utah Surveillance
- RODS Pennsylvania Surveillance
- Algorithms:
  - WSARE
  - BBF and PANDA
  - Fast Scan
  - BARD
- Use of BioWar in the summer evaluation
- Data Sources/N=18 Study of Electrolyte Sales

Pitt-CMU

3

RODS Winter Olympic Surveillance



Pitt-CMU

Real-time Outbreak and Disease Surveillance (RODS)

- Real-time collection of chief complaint data
- From >70% of all ED and acute care visits in Utah
- Natural language processing into prodromes
- Servers located at University of Pittsburgh
- Implemented in under 6 weeks
- >2000 uses in 2 weeks

Drs. Wagner and Moore giving a live demo of RODS to President George Bush, Secty. Thompson, Governor Ridge at RODS Lab 2/5/02

Key Contributions

- Prodromic surveillance from free text now being deployed by most states
- Use of HL7, VPN, Internet, NLP, advanced algorithms

4

RODS Pennsylvania Surveillance

- Chief complaint coverage

  Southwestern PA 50%

  Harrisburg 100%

- Retail data: 16% statewide (evenly distributed)
- Poison call center: 100% statewide

Commonwealth has funded a project to connect 100% of 170 hospitals within 3 years

BioWar

- Kathleen Carley will discuss at 2:15

What Advanced Analytics Are Being Used/developed for Detection/characterization?

Method	Has Pitt/CMU tried it?	Tried but little used	Tried and used	Under development	Multivariate signal tracking?	Spatial?
Time-weighted averaging	Yes	Yes				
Serfling	Yes		Yes			
ARIMA	Yes	Yes				
SARIMA + External Factors	Yes		Yes			
Univariate HMM	Yes		Yes			
Kalman Filter	Yes	Yes				
Recursive Least Squares	Yes		Yes			
Support Vector Machine	Yes	Yes				
Neural Nets	Yes	Yes				
CuSUM	Yes	Yes				
Randomization	Yes		Yes	Yes		
Spatial Scan Statistics	Yes			(w/ Howard Burkom)	Yes	Yes
Bayesian Networks	Yes			Yes	Yes	
Contingency Tables	Yes		Yes			
Scalar Outlier (SQC)	Yes	Yes				
Multivariate Anomalies	Yes		Yes		Yes	
Change-point statistics	Yes			Yes		
FDR Tests	Yes		Yes		Yes	
WSARE (Recent patterns)	Yes		Yes	Yes	Yes	Yes
PANDA (Causal Model)	Yes			Yes	Yes	Yes
FLUMOD (space/Time HMM)				Yes	Yes	Yes

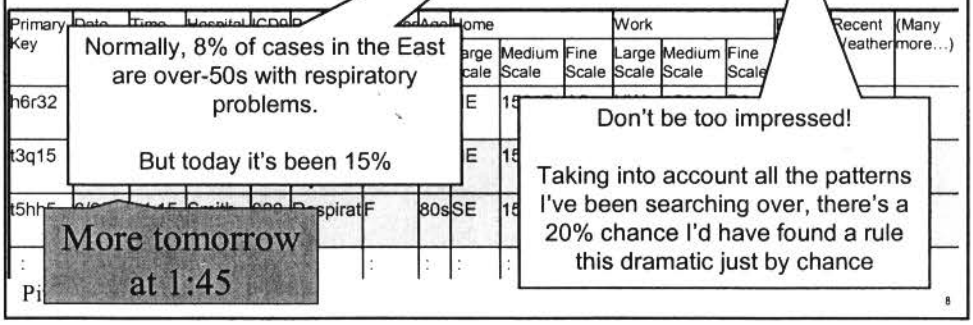
Details of these methods and bibliography available from "Summary of Biosurveillance-relevant statistical and data mining technologies" by Moore, Cooper, Tsui and Wagner. Downloadable (PDF format) from www.cs.cmu.edu/~awm/biosurv-methods.pdf

Pitt-CMU

7

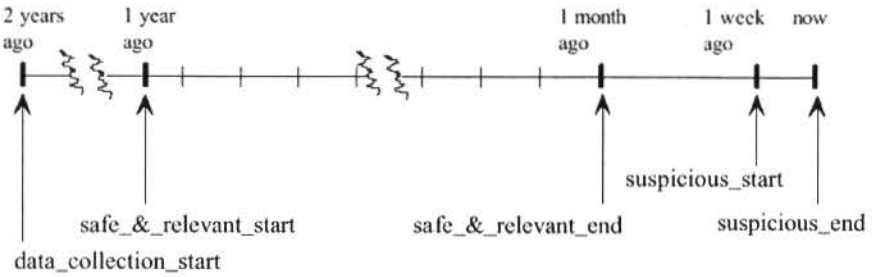
WSARE v2.0

- Inputs:
  - 1. Date/time-indexed biosurveillance-relevant data stream
  - 2. Time Window Length
  - 3. Which attributes to use?
- Outputs:
  - 1. Here are the records that most surprise me
  - 2. Here's why
  - 3. And here's how seriously you should take it



8

BBF (Biosurveillance using Bayesian and Frequentist methods)



Ask whether a Bayesian predictive model learned from a 'safe and relevant period is the best predictor of current activity OR whether 2 models are better

Compute the likelihood ratio of 2 models versus one model

More at 5:30

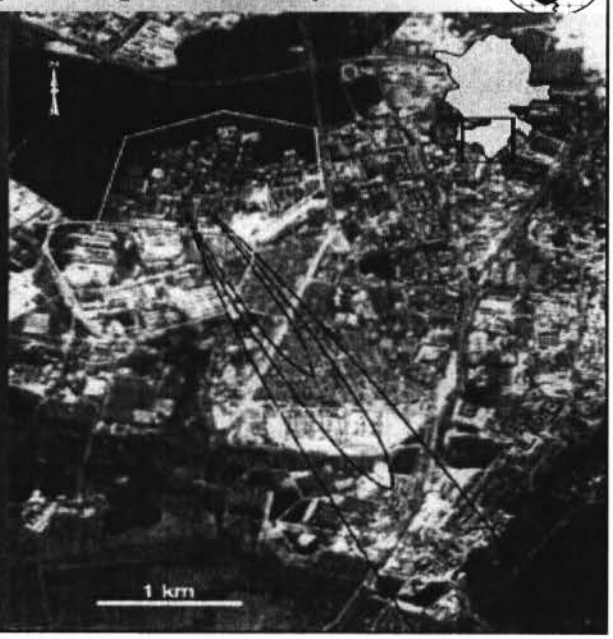
stion for all recent time intervals exhaustively
one with the biggest ratio…

Pitt-CMU

9

BARD: Temporal-Spatial Analysis

Sverdlovsk: Aerial View

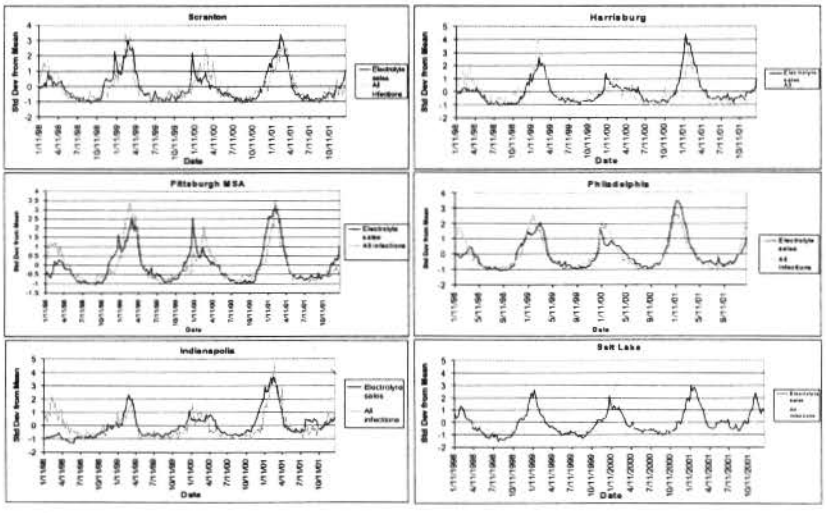


More on Thursday 10:00

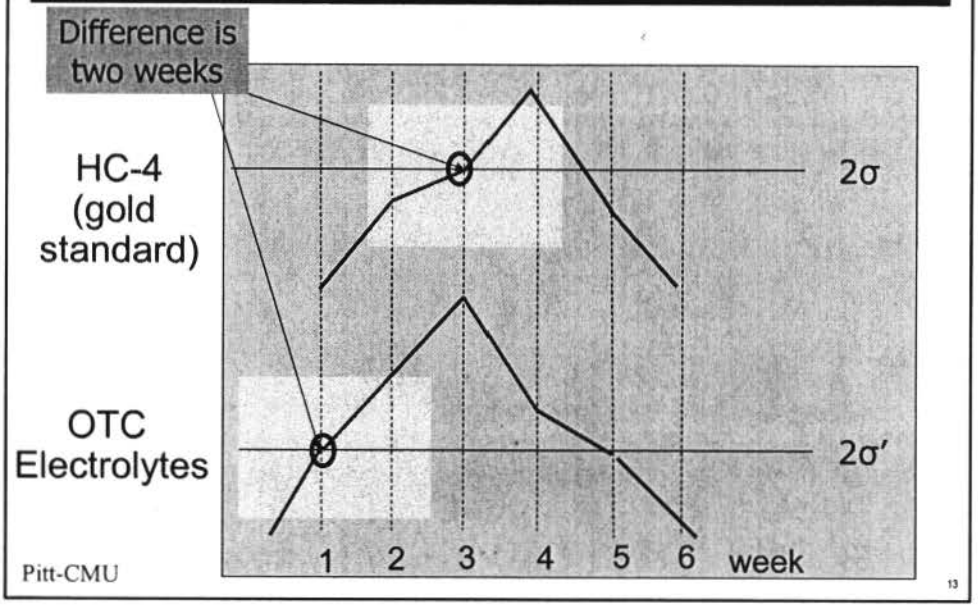
Nontraditional Data Sources

- Data Sources: How Much Faster Does Each One Serve As Leading Indicators of Outbreaks? How Do You Know?

Retail Sales of Electrolyte Products vs. "All" Pediatric Respiratory and Diarrheal Diseases.



Timeliness



Pitt-CMU

13

Mean Earliness at Different Sigma Detection Thresholds

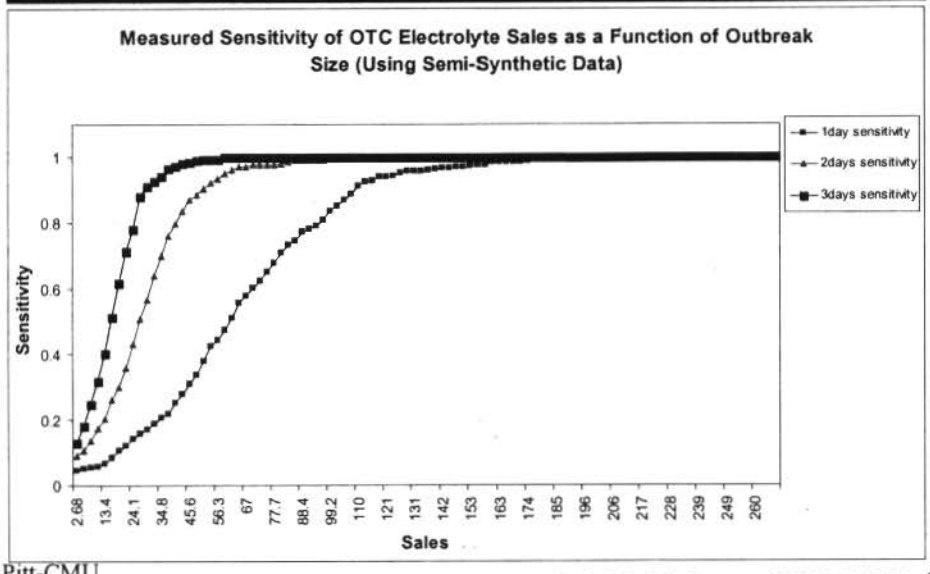
Detection Threshold	Avg. Earliness (weeks)*	Earliness 95% C.I.	Sensitivity	Specificity
2.0	2.8	(1.9-3.7)	100%	100%
2.5	2.8	(2.0-3.6)	100%	100%
3.0	2.7	(1.8-3.5)	100%	100%
3.5	2.7	(1.8-3.5)	100%	100%
4.0	2.9	(2.1-3.7)	100%	100%
4.5	3.3	(2.5-4.2)	100%	100%
5.0	3.3	(2.5-4.1)	100%	100%
5.5	3.3	(2.4-4.2)	100%	100%
6.0	3.3	(2.4-4.2)	100%	100%
6.6	3.1	(2.2-4.0)	100%	100%
7.0	3.6	(2.7-4.6)	100%	100%
7.5	3.8	(2.8-4.9)	100%	100%
8.0	3.9	(2.9-5.0)	100%	100%
8.5	3.8	(2.9-4.8)	100%	100%
9.0	3.8	(2.7-4.9)	100%	100%

* n=18 outbreaks

Pitt-CMU

14

Sensitivity of OTC Electrolyte Sales: Analysis Using Semi-synthetic Data

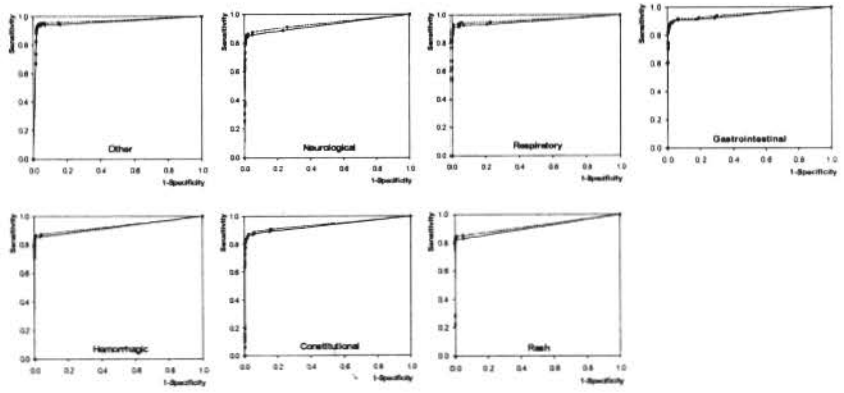


Pitt-CMU

Method of Goldenberg et al, PNAS 2002

15

Detection from Chief Complaint Data

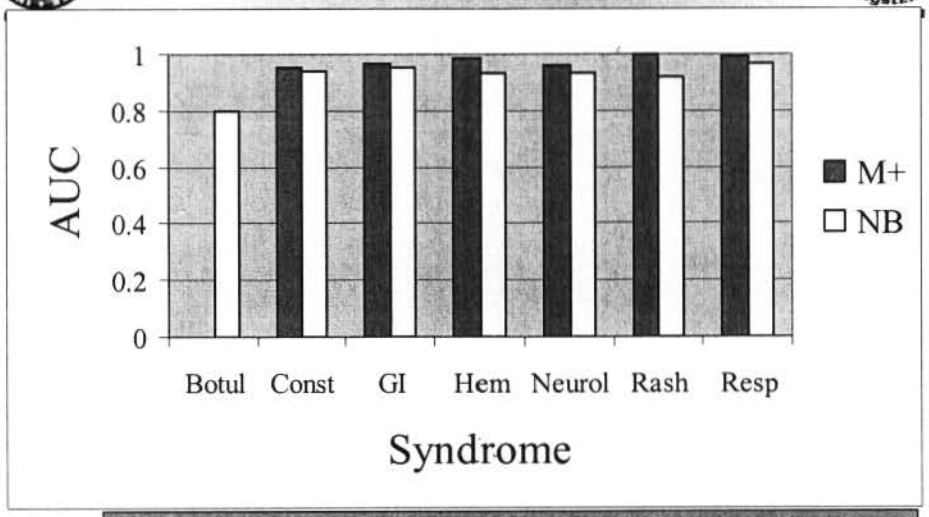


Proof 1: Text processing methods can map chief complaints to prodromes as well as experts

Pitt-CMU

16

Accuracy of Natural Language Parsing of Chief Complaints into Syndrome Groups



* Legen
Bayes, a
Pitt-CMU

Fancy, but expensive to train (M+) NLP methods work slightly better, but may not be cost effective

17

Accuracy of NLP of Free-text Chief Complaints for Cases of Acute Infectious GI Disorder

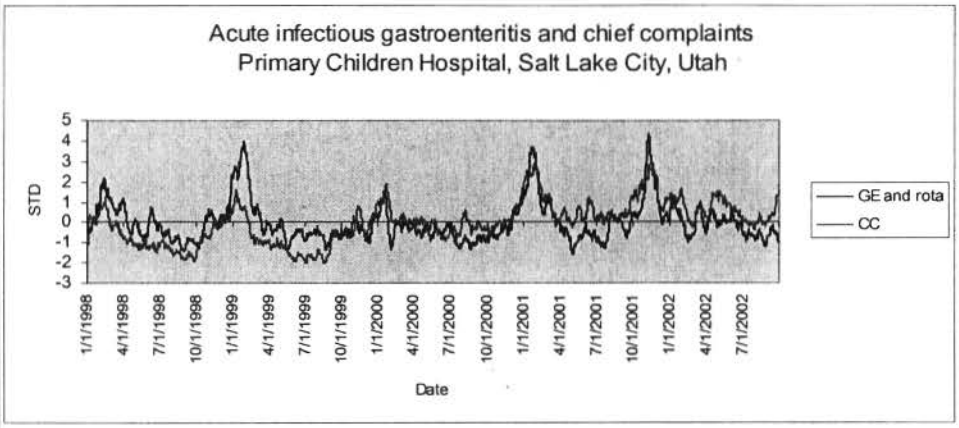
	ICD-9	Chief Complaint
Sensitivity	0.32	0.50
Specificity	0.99	0.92
PPV	0.37	0.09
NPV	0.99	0.99
AUC	N/A	0.75

Pitt-CMU

Proof 2: Text processing methods can support CASE detection

18

Long Story Short: Free text chief complaints are panning out



Proof 3: Text processing methods can support EPIDEMIC detection

Pitt

19

Leading Indicators: Detection Timeliness and Minimum Detectable Outbreak Effects

Indicator	Disease	Sampling %	Earliness	Lower bound estimate of detectability	Surrogate for:
OTC electrolyte sales*	Pediatric respiratory and diarrheal	95% sampling of six counties	2.3 weeks earlier (1.4-3.4)	30 unit sales over baseline (of 20 in Allegheny County, pop. 1.3M). With geographically localized outbreaks, possibly fewer	Cholera, salmonella
OTC respiratory equipment*	Respiratory Syncytial Virus	50% 6-county region	Precedes hospitalization XC 4d/ further studies needed	25 unit sales over baseline (of 35 in Allegheny County, pop. 1.3M). With geographically localized outbreaks, possibly fewer	Anthrax, plague
OTC cough and cold*	Influenza	50% 6-county region	Precedes ED registrations XC 3 days***	200 unit increase in daily sales (which are 600) in winter, 60 (of 200) in summer	Anthrax, plague
ED chief complaints (coded)*	Influenza	15% 1 county (population 1.3 M)	Precedes Pneumonia&Influenza deaths XC 2 weeks		Anthrax, plague
ED free text**	Diarrheal and respiratory	N.Applicable (but currently 70% 1 county 50% MSA)	Not measured/feasibility proven	Individual case detection sensitivity 30-50%. Additional analysis required to translate into lower bound estimate	Diarrheal and respiratory

*N.B. Measures are SINGLE measurements, so confidence intervals cannot be computed and results are highly provisional

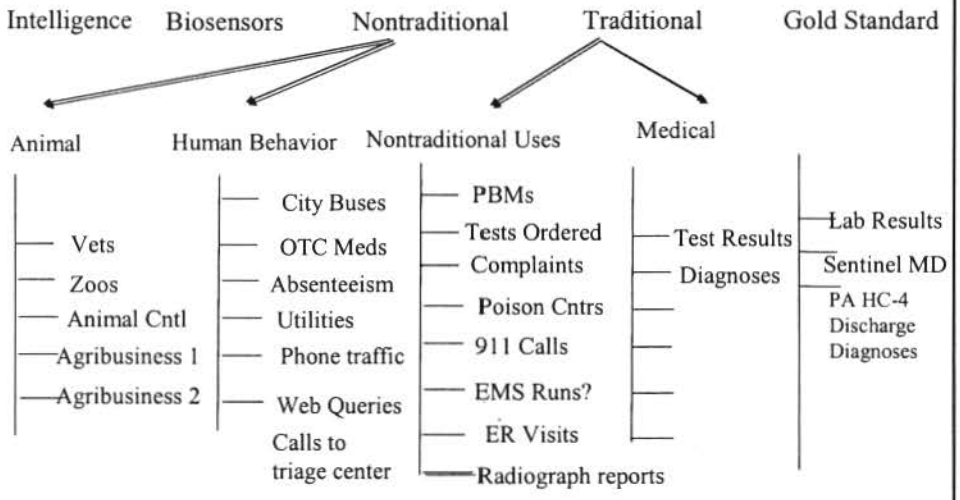
** Different analytic method using detection performance for individual cases.

*** By unpublished analysis of 1999-2000 outbreak

Pitt-CMU

20

Data Sources



Pitt-CMU

Legend: Red: we have; Blue: under contract; Black: we have investigated or are investigating

21

Notes on Data Sources

- Nurse call centers: one study with Influenza
- Poison call centers: not much evidence that they are useful for diarrhea or respiratory outbreaks, but there are so many threats
- Phone call volumes: still pursuing
- Agribusiness: goal for next year
- NLP of radiographs: ongoing
- Pittsburgh Bus System: ongoing analysis

Pitt-CMU

22

PLANS for FY03

- Quantitative Goals
- Transition Path for the Technology
- Ideas on Accomplishing Pathogen ID and Attack Characterization
- Other Funding Sources and Division of Labor (if appropriate)
- Further Evaluation/Validation Plans

Quantitative Goals

- DATA: Obtain 80%, five year data sets in the INDIANA-PENNSYLVANIA-UTAH testbed for at least two other types of data
- ALGORITHMS:
  - Cohort threats: Utilize spatial and other information to move up the time of detection to within 2 days of 'when-they-first-wake-up sick'
  - Smaller outbreaks (as small as one case)

Transition Path for the Technology

- ASP model: PA and Utah already, Delaware next. Indiana likely
- Click-license model:
  - RODS
  - CoCo (complaint coder)
  - WSARE
- For-the-asking
  - BioWar
  - Fast Spatial Scan
  - SecondStage

Other Funding Sources and Division of Labor (if appropriate)

- Other funding: Commonwealth of PA
- Division of labor:
  - Completely new:
    - Applied work developing early warning public health surveillance system in PA
    - Health system resident component
    - Clinical data work
  - Complementary: Basic research on data and algorithms
  - Highly synergistic: e.g., value of the Pennsylvania HC-4data set as a gold standard

Ideas on Accomplishing Pathogen ID and Attack Characterization

- There are important characteristics outbreaks that are difficult to characterize completely from routinely collected data alone
  - Pathogen
  - Source
  - Scope (people, animals)
  - Environmental scope
  - Route of transmission
  - Host factors
- SecondStage Version 1.0
  - XML and Java
  - core functions to support epidemic data collection and storage
  - Available now
- Version 2.0 Data collection about individuals driven by population data

Summary: Technical Approach

- What Are the Key New Ideas? How Are They Achieved?
  - Bayesian method for fusing individual observations when available with aggregate data (PANDA algorithm)
  - Multi center trials are required for evaluation
  - Brute force search over large variable spaces is critical (WSARE algorithm)
  - Agent-based simulations are a component in evaluation (Biowar simulator)
  - Rapid computer-assisted investigations are a critical component of a detection system to reduce false positives and adequately characterize an outbreak (SecondStage)

Summary: Objectives and Performance Goals

Key Metrics We Use

-Timeliness and false alarm rate

-Size of outbreak detected

Data Sources We Have

Chief complaints(2 states)
Cattle (1 feedlot)
911 (1 city)
OTC (3+ states)
Poison call (1 state)
Health dept (1 HD)
Nurse calls (2 urban regions)
Hospital admits (3 states)
Clinical data (3 states)

Performance Goals and Technical Milestones

DATA: Multiply number of data sources we are using by factor of 2 (see side bar). Expand data coverage sufficiently for N of 15 studies of chief complaints and at least 2 other types of data.

ALGORITHMS: Utilize historical models to reduce false positive rate by 50%

Detect outbreaks of 50% smaller size by using non spatial associations

Key deliverable: WSARE 3

Key deliverable: Multi attribute Bayesian version of Panda

EpiSPIRE Project Update

IBM T. J. Watson Research Center
Johns Hopkins University
SDI Inc.

1

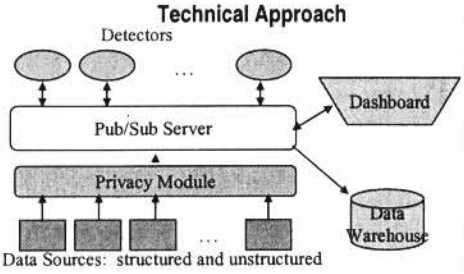
Team

- IBM T. J. Watson Research Center
  - Charu Aggarwal
  - Murray Campbell
  - Yuan-Chi Chang
  - Vijay Iyengar
  - Mahesh Joshi
  - Chung-Sheng Li (PI)
  - Ching-Yung Lin
  - Milind Naphade
  - Ed Pednault
  - Jeanette Rosenthal
  - John R. Smith
  - Belle Tseng
  - Min Wang
  - Joel Wolf
  - Kung-Lung Wu
  - Philip S. Yu
- Johns Hopkins University
  - Prof. Greg Glass (co-I)
  - Rebecca Freeman
- SDI
  - Andrew Kress

2

EpiSPIRE
IBM/JHU – PI:Chung-Sheng Li

Technical Approach



Objectives and Performance Goals

- Validate data sources against ground truth
  - Lag studies versus physician office visits
- Validate detection algorithms
  - Evaluate on real and synthetic data
- System for real-time access to heterogeneous data sources
- Leverage multi-modal data sources to improve precision of bio-surveillance

Accomplishments

- Comprehensive study of site bio-surveillance
  - Cataloging and initial evaluation of site-specific data sources, including active surveillance
  - Specialized site-specific detectors
  - Prototype surveillance system
- Initial evaluation of detector technologies
- Leverage environmental data for baseline

Programmatics

- Ongoing: Evaluation of non-traditional data sources. Evaluate in both site/non-site contexts.
- Ongoing: Development and testing of detection algorithms with real and synthetic data.
- 6/03: Deployable site surveillance infrastructure. Begin deployment to candidate sites/bases.
- 12/03: Model environmental effects on disease processes.

3

Technical Approach: Main Results

- Site-specific prototype bio-surveillance system
  - Comprehensive active and passive monitoring of site-specific non-traditional data
  - In conjunction with regional and environmental traditional and non-traditional data
  - Establish single and multi-channel detectors for early anomaly detection
- Implications:
  - Applicable to sites with a captive population (such as military sites)
- Validation: The prototype system has been validated by using
  - IBM T. J. Watson Research Center (located in Yorktown and Hawthorne, New York) as the surrogate site
  - Medical insurance claims (SDI) for Westchester County as "ground truth"
  - BioWar attack scenarios

4

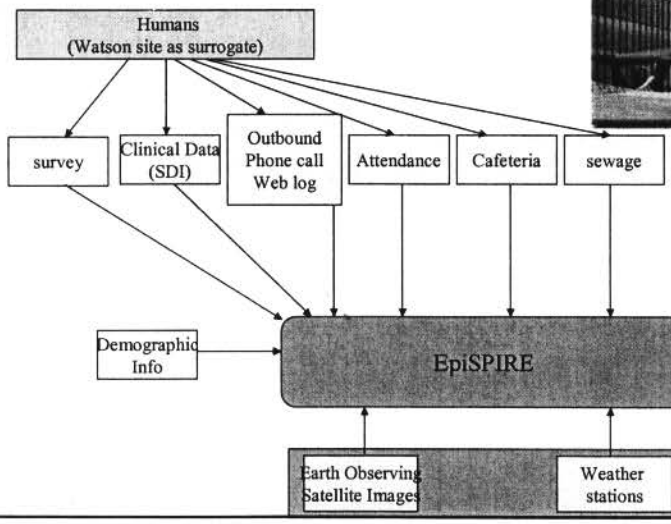
Technical Approach: New Ideas

- Methodology
  - Site-specific simultaneous active and passive surveillance
  - Use of real time environmental data in preparing the baseline of disease modeling
- Data gathering
  - Active surveillance (survey), outgoing phone, outgoing web logs, and attendance characterize worksite
  - Automatic car counting through surveillance video
- Advanced Analytics
  - Detectors: Proximity based, Holt-Winters, directional, differential, adaptive threshold, incremental clustering
  - Privacy: Extended k-anonymity model to allow more flexible anonymization

5

EpiSPIRE for Site Specific Bio-surveillance

•Comprehensive catalog of all site-specific data sources



6

Reverse Engineering of Human Behavior from Information Integration

Event

metaphor

concept

semantic

feature

raw

abstraction

# of sources

heterogeneity

Human behavior

Phenomenon

Capture

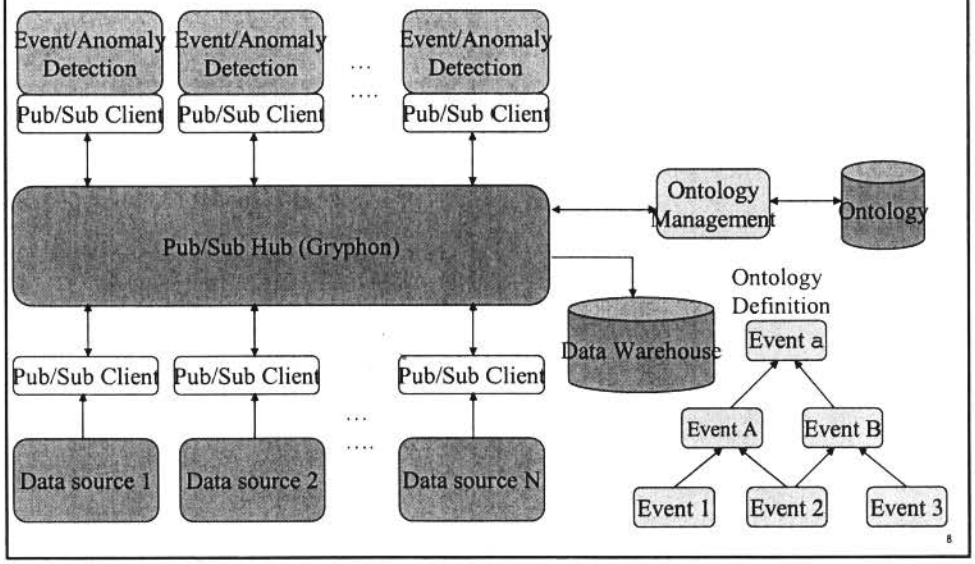
Structured (relational)

Semi-structured (XML, HTML)

Non-structured (Text, image, video, audio)

7

Technical Approach:
Event/Alert Generation – System View

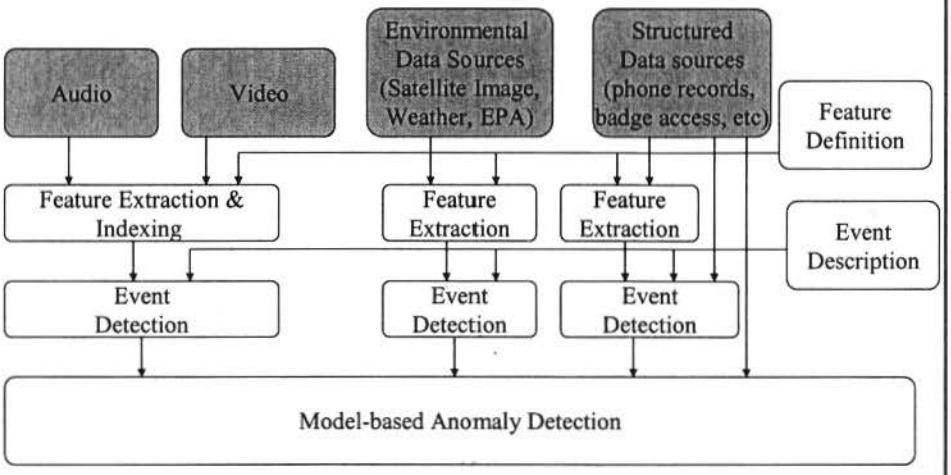


8

Technical Approach

Model-based Anomaly Detector

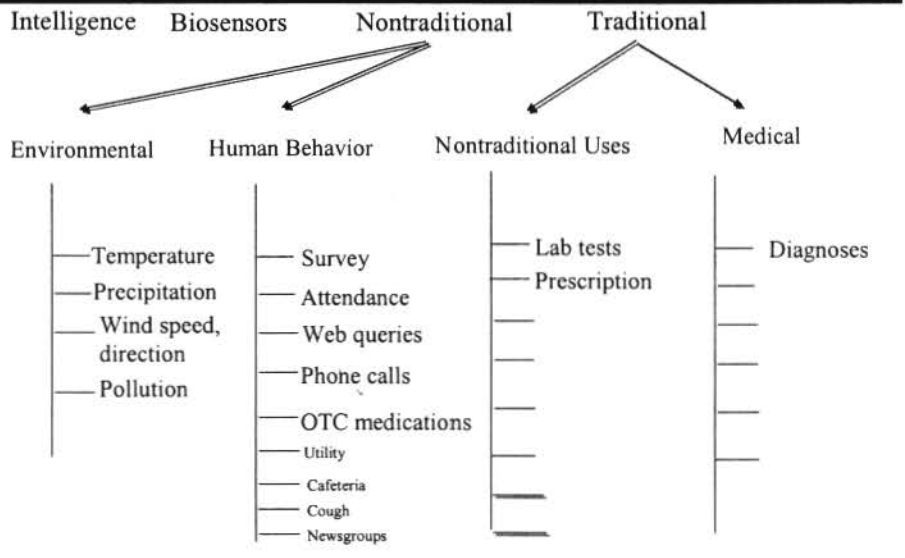
•Alert generation through monitoring of data sources



9

Technical Approach:

Non-traditional data sources



10

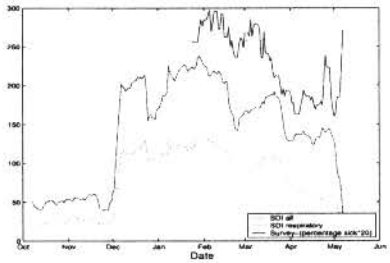
Sensitivity and Latency of Data Sources

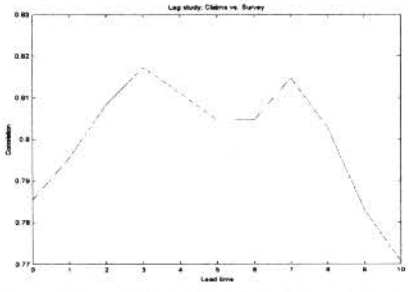
- Size of detectable outbreak at Watson (with 99% confidence): 4-10% of site population, depending on data source
- Leading early indicators for site-specific surveillance (validated during winter 2002):
  - Active surveillance (survey): 3 days
  - Phone records: 4 days

11

Nontraditional Data Sources:
Active Surveillance

- **Hypothesis**:
  - A disease outbreak at a worksite (or the surrounding community) can be detected through active surveillance
- **Methodology**:
  - Regularly survey employees at IBM Watson worksite, asking them to self-assess their health
  - If feeling ill, employees describe symptoms
- **Validation Strategy**:
  - Lag study between survey illness level and medical insurance claims (SDI) for surrounding area
  - Peak correlation of .82 suggests survey provides three days of lead time over doctor visits

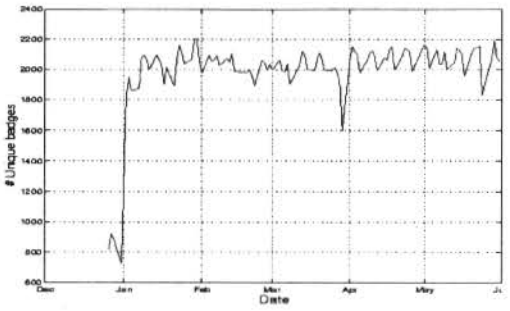




12

Nontraditional Data Sources:
Worksite Attendance

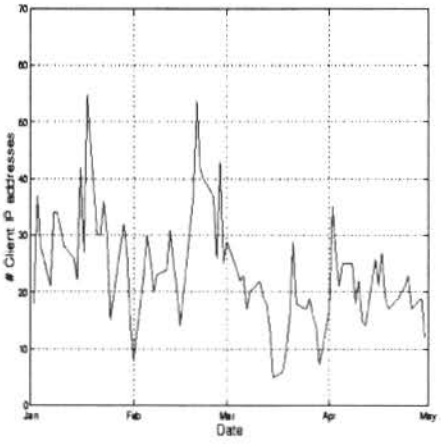
- **Hypothesis**:
  - A disease outbreak at a worksite (or the surrounding community) can be detected by monitoring worksite attendance
- **Methodology**:
  - Use badge entry system to monitor worksite attendance at IBM Watson site
- **Validation strategy**:
  - Lag study between worksite attendance level and medical insurance claims (SDI) for surrounding area found no significant correlation
  - Conclusion: Worksite attendance appears to be insensitive to mild illness



13

Nontraditional Data Sources:
Firewall Server Logs

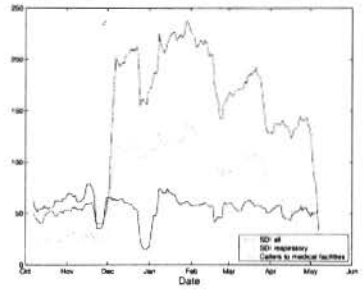
- **Hypothesis**:
  - Occurrences of disease outbreaks can be detected through the mining of the proxy server log traces
- **Methodology**:
  - Use proxy logs to monitor web accesses.
  - Identify a set of destination IP addresses which correspond to medical websites.
  - Additional specificity may be available with full URL logging.
- **Validation Strategy**:
  - Lag study between medically related web accesses and medical insurance claims (SDI) for surrounding area was inconclusive.

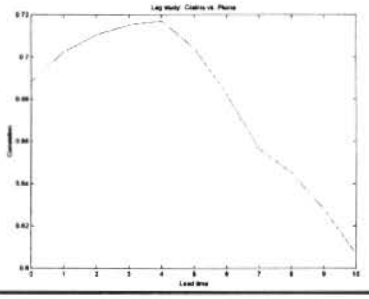


14

Nontraditional Data Sources:
Phone Records

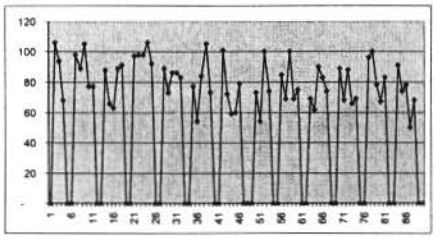
- **Hypothesis**:
  - Disease outbreaks will be detectable in the number of calls (callers) to medical facilities in the geographical area around impacted work sites.
- **Methodology**:
  - Monitor phone calls to medically related phone numbers from IBM Watson site.
- **Validation Strategy**:
  - Lag study between callers to medically related numbers and medical insurance claims (SDI) for surrounding area
  - Peak correlation of .72 suggests survey provides four days of lead time over doctor visits



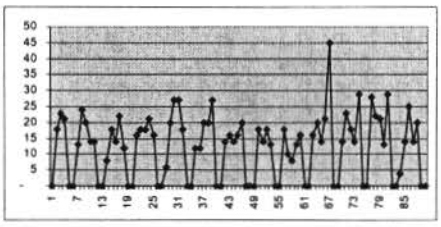


15

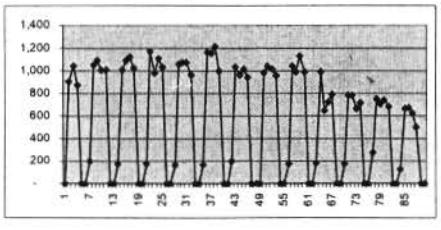
Nontraditional Data Sources:
Cafeteria Data (Jan-Mar, 2002)



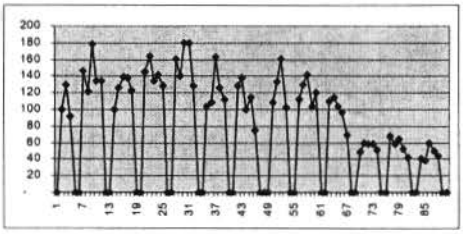
soup



Fountain juice



Total beverage



Fresh fruits

16

Nontraditional Data Sources:

Cafeteria Data (Jan-Mar, 2002)

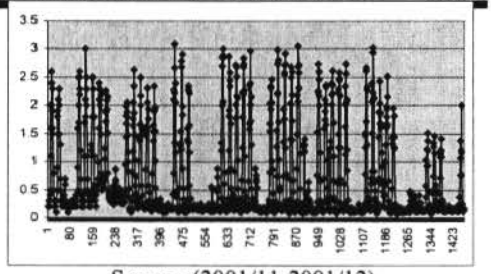
•correlation is weaker than for our other data sources (.66).
•We consider the lag study to be inconclusive, as the largest peak in correlation is found at 19 days, which is beyond what we believe to be reasonable.

17

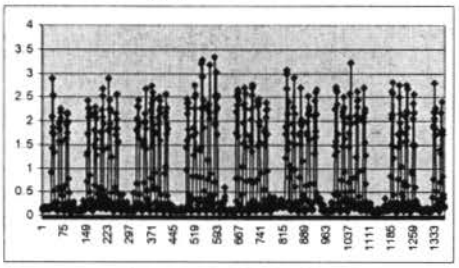
Nontraditional Data Sources:

Utility/Cafeteria Data

- **Hypothesis**: Occurrences of disease outbreaks can be detected through the mining of the utility and cafeteria transactions data
- **Methodology**:
  - We have obtained access to the utility data of Watson
    - Electricity
    - Water
    - Sewage generation
    - Cafeteria transactions
- **Validation Strategy**:
  - The data will be correlated with medical insurance claims (SDI)



Sewage (2001/11-2001/12)



Sewage (2002/01-2001/02)

18

Nontraditional Data Sources: Surveillance Audio

- Hypothesis: using audio data collected at public gathering to predict the well being of the population and the potential disease outbreak
- Methodology:
  - We have collected audio data at public gathering locations including lobby, cafeteria, and seminar
- Validation strategy:
  - Although we have verified the technical feasibility of cough detection, its usefulness limited by:
    - Very low frequency of coughing
    - Difficulty in distinguishing individuals

19

Nontraditional Data Sources:

Video Surveillance Auto Monitoring

- **Hypothesis**: video sources can be used to augment the badge data (people leaving facility)
- **Methodology**
  - Collect streaming video of the webcam of vehicles leaving site
- **Validation Strategy**
  - we will validate the feasibility of extracting traffic volume from streaming video - **completed**

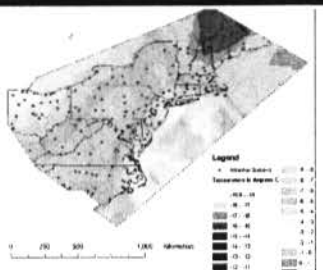



	Total Frames	Inbound (A/M)	Outbound (A/M)	Precision
Clip 004	2417	106/95	106/100	**91.3%**
Clip 007	3856	33/32	45/44	**97.4%**
Clip 021	5460	143/143	62/65	**98.6%**
Total	11733	282/270	213/209	**96.7%**

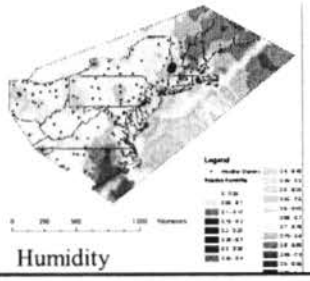
20

Nontraditional Data Sources:
Environmental Model

- **Objective**:
  - To establish the impact of environmental variables on disease
- **Environmental data**:
  - ongoing progress of collecting satellite images,
  - EPA readings of air pollution, and
  - weather station measurements of the northeast region (including DC, Maryland, and NY state)
- **Validation plan**
  - We are validating against medical insurance claims (SDI)



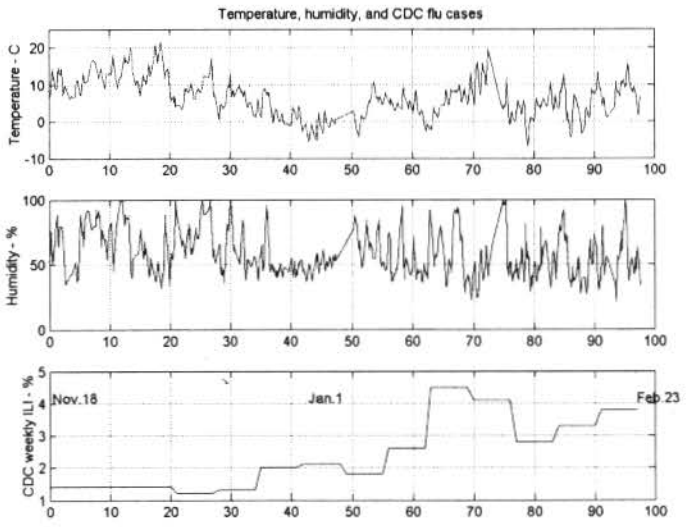
Temperature



Humidity

21

Nontraditional Data Sources:
Environmental Data and Flu Cases



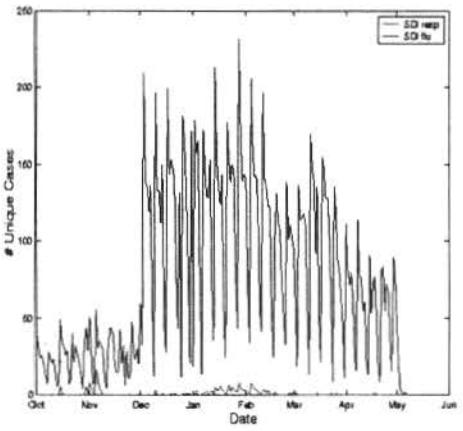
Weather data retrieved from period Nov.18, 2001 to Feb. 23, 2002, LGA, New York

22

Traditional and Nontraditional Data Sources
SDI

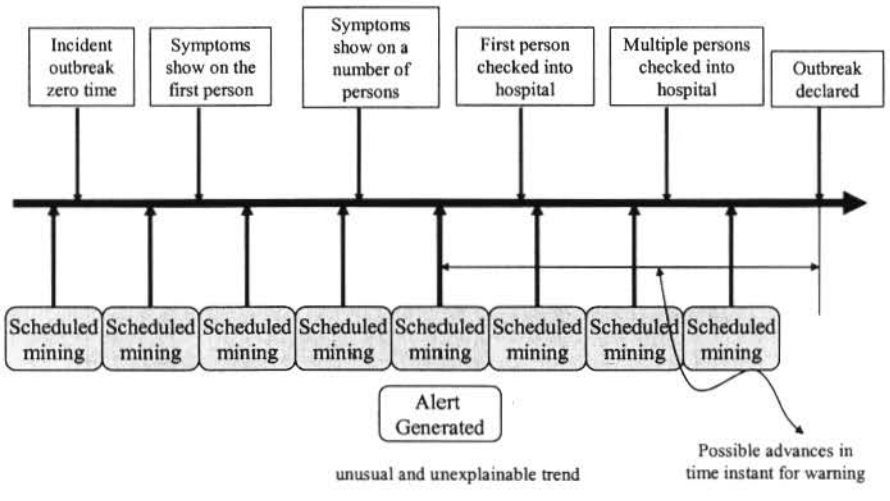
- August 2001 through May 2002
- ICD9 codes 001-139 (Infectious diseases) and 460-519 (Respiratory)
- NY, Pittsburgh, DC areas
- Includes anonymized patient identifier, service date, process date, up to 4 ICD9 codes, age, gender, zipcode
- 1.2 million claims
- SDI data covers nationwide back to 1998
- Other
  - OTC medications
    - National coverage, store level
    - 1 day lag
  - Lab tests ordered
    - National coverage, by zip code
    - 1-3 day lag
  - Prescription medications
    - National coverage, by zip code
  - Near-real-time possible

Westchester County



23

Time line for advanced detection of anomaly



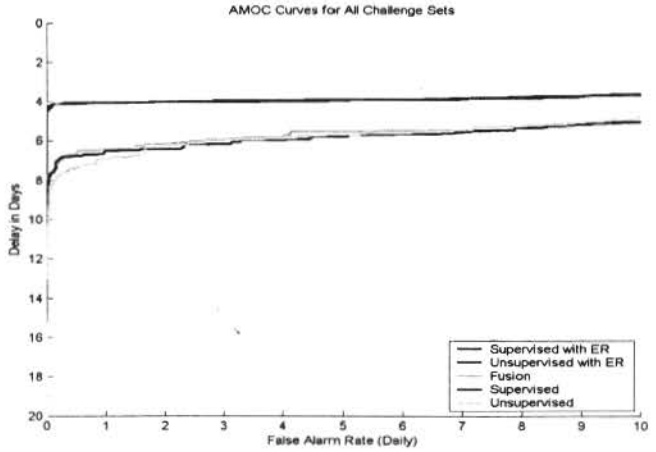
24

Detectors

- **Combination signature detector:** Builds model of multi-channel attack signature
  - Supervised approach
- **Adaptive Threshold**: Forecast normal behavior, alarm on sufficient deviation
  - Unsupervised approach
- **Site detectors**: Detect departures from normal tracking behavior between sites
  - Unsupervised
  - Outperformed our general detectors on BioWar site attacks
- **Fusion detector**: Combines two or more detectors into a single alarm
  - Combined combination signature and adaptive threshold detectors in August Challenge: slightly outperformed both
- **Stadium attack detector**: Detect signature of attack that disperses throughout an area
  - Unsupervised
  - Not used in August Challenge

25

AMOC Curves for August Challenge



26

Privacy Protection Techniques

- IRB for active surveillance
- Data sharing agreement with SDI as well as data providers for SDI
- Anonymization of survey, phone and proxy log
- A new privacy module has been developed and used
  - Extended k-anonymity model based techniques
    - Usage based metrics: Tailor information loss due to anonymization
      - e.g., anonymized table used for predictive modeling
    - Allow more flexible generalizations
      - Intervals for numeric attributes, multi-level generalizations using taxonomy tree
    - Aggressive exploration of solution space using genetic algorithm framework
      - Minimize loss metric while satisfying anonymity constraint
  - Applied prototype to survey
    - k=10 (specified at survey subscription)
    - Retained 4 attributes
      - Age (<46, rest), Children (y/n), flu shot (y/n), work place (haw/ykt)
      - 28 out of 387 suppressed
  - Applied prototype to SDI data with k=1000
    - Our flexible anonymization produced abstractions adapted to the characteristics of patients in various zip codes and age groups
      - e.g., Zip code abstractions 0070xx (had 10464 patients), 04451x (had 2100 patients)

27

Objectives and Performance Goals

- Key Progress Measures
  - Data Sources: Which data sources show promise in anticipating disease outbreaks?
    - We have used lag studies versus ground truth (physician office visits).
  - Detection algorithms: Which algorithms can detect an attack early with minimal false alarms?
    - Synthetic data (BioWar): attack detection, localization
    - Real data: evaluate non-attack behavior of detectors
    - Combined real/synthetic data

28

Objectives and Performance Goals:
Key deliverables

- **System:**
  - **EpiSPIRE**
    - Prototype anomaly detection system
    - **Availability**: end of year
- **Data products:**
  - **Environmental data** for the entire northeast region of US
    - High spatial and temporal resolution temperature and moisture
    - **Availability**: now
- **Infrastructure Modules:**
  - **Detectors**:
    - Single and multi-channel anomaly detectors
    - **Availability**: now
  - **Privacy module**
    - **Availability**: now
  - **Application Roaming Framework**:
    - **Availability**: now
  - **Multiple Abstraction Level Mining (MALM)** of time series:
    - **Availability**: now
  - **Video car counting module**:
    - **Availability**: now
  - **Information exchange**:
    - **Availability**: now

29

Accomplishments

- Comprehensive study of site bio-surveillance
  - Catalog and initial evaluation of site-specific data sources
  - Specialized site-specific detectors
  - Prototype site surveillance system
- Detectors developed
  - Initial evaluation using synthetic and real data
- Privacy protection system
  - Extended k-anonymity technique
    - Described in: V. Iyengar, "Transforming Data to Satisfy Privacy Constraints", in KDD 2002
- High resolution real-time environmental data made available for baseline modeling

30

Accomplishments:
Lessons Learned & Transition Opportunities

- Full support from the parent organization is of paramount importance – and fortunately we do have full support from IBM
- It is easy to underestimate the waiting period for going through IRB
  - It took us 3 full months from the initial requests to get approval
- Establish vehicle(s) for sharing technologies with other projects could take a very long time
  - It took us 6 full months to establish JSA with JHU/APL
- Bridging data on paper and electronic data is painful
  - It took me a whole weekend to enter the cafeteria data manually
- Collaboration with local health facilities is more difficult than we think
- Who has Committed to Using it in the Future?
  - JHU/APL ESSENCE II:
    - environmental data,
    - shape based detectors (MALM),
    - web event detector
  - Mellon-Pitt
    - Environmental data
    - Video based car counter
  - Veridian
    - Privacy module
    - Cough counter
- What Other Opportunities are there ?
  - Profile detector
  - Differential detector
  - Environmental data
  - Privacy module

31

Programmatics

- Project Plan
  - Ongoing characterization of nontraditional data sources versus gold standards
    - Evaluate promising data sources in non-site contexts
  - Ongoing development and evaluation of detection algorithms
    - Evaluate on real and synthetic data
  - Ongoing development and testing of site surveillance prototype
  - 6/03: Deployable site surveillance infrastructure
    - Begin deployment to candidate sites or bases
  - 12/03: Develop model of environmental effects on disease processes

32

Programmatics:
Additional requirements, Evolution, and Collaboration

- **Additional data sources** (optional)
  - Animal data of metro NY (working with Dr. McNamara)
- **Ontology (both qualitative and quantitative)**
  - For defining temporal, spatial, and spatio-temporal "trends" or "patterns"
- **Project Evolution**
  - Focus on site-specific surveillance
    - IBM Watson site is selected to emulate captive population scenarios such as Pentagon, military bases and other military facilities
  - Active surveillance being added as part of our project since the kickoff meeting
- **Collaboration with other**
  - JHU/APL:
    - joint study agreement to delivering technologies for validation
    - Jointly define event/alert data structure for communication
  - Veridian:
    - Cough counting
  - Mellon-Pitt
    - Plan to share video car counting tool (through alphaworks)
- Collaboration with other Testbeds
  - NASA ESIP
  - CDC
  - IBM Sphinx: **Sp**atial **H**ealth **I**nformation e**X**change
- **Other potential funding source(s)**
  - NIH

33

ESSENCE II Status Summary

DARPA Bio Surveillance PI Meeting
3 December 2002

Joe Lombardo

The ESSENCE II Team

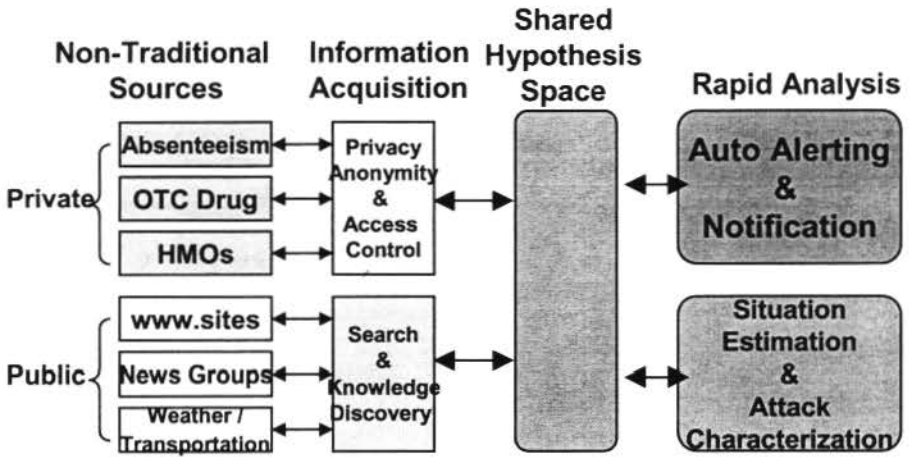
DoD-GEIS, JHUAPL, CMU, Kaiser

Babin,Steven M.
Brown, Kendall
Burkom,Howard S.
Casper, Jack
Cutchis,Protagoras N.
Elbert, Yevgeniy
Florio,Elaine N.
Foster,Virginia B.
Happel Lewis,Sheryl L.
Henry, Jade
Kelley, Pat
Lombardo,Joe S.
Loschen,Wayne A.
Mansfield, Jay
Magruder,Steven F.
Marsden-Haug, Nicola
Maurer,Scott G.
Nelson, Kevin
Pavlin, Julie
Polk, Christina
Sari,James W.
Schumacher, Camille
Sniegoski,Carol A.
Snyder, Mark
Sweeney, Latanya
Witt, Clara
White,David M.
Wojcik,Richard A.

The ESSENCE II Bio Surveillance System

PI: Joe Lombardo, JHUAPL CO-PI: LTC Julie Pavlin DoD_GEIS

Technical Approach



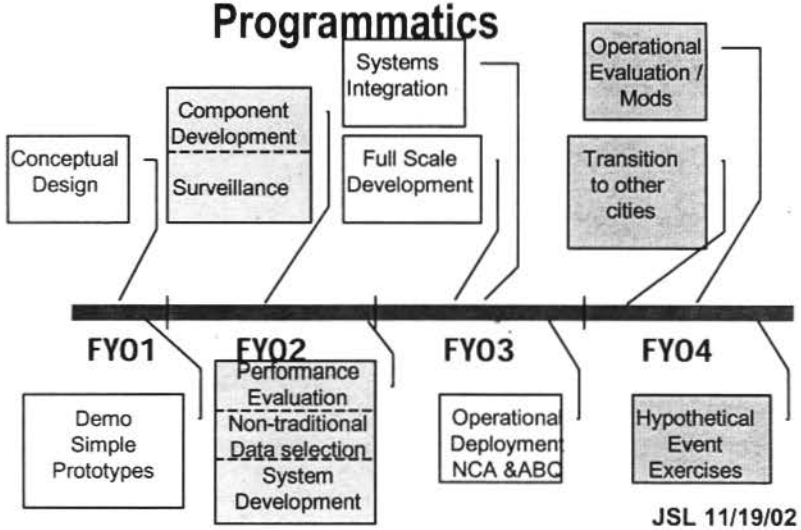
Objectives and Performance Goals

- Provide the capability to identify abnormal health state due to release of pathogen days before traditional methods
- 24/7 autonomous surveillance, early detection & warning
- Integrate across military & civilian non-traditional data sources to identify abnormal health conditions
- Provides quantitative measures needed for a rapid response for both military and civilian public health officers

Accomplishments

- Operational deployment to global military health officers & local health departments
- Work with local health departments and health care agencies to use the technology for local exercises.
- Spatial / temporal outbreak detector development
- Automation of the collection and archiving processes
- Begin the process of transitioning the DARPA technology to the DTRA and PEOCBD programs.

Programmatics



ESSENCE II Team Data Sources

Traditional	Non-Traditional (clinical)	Non-Traditional
Lab results	**Lab test requests**	Absenteeism
Nursing home	Emergency room	Over the counter meds
	Civ. / Mil. electronic claims	**Poison center calls**
	Prescription medications Mil./**Civ.**	**Metro rider ship**
	Nurse hotline Mil. / Civ.	**TV viewing**
	Electronic vital signs	**Health care websites**
	Animal Health Vet. Hosp./Clinics	**Video surveillance**
	ME (Unexplained Deaths)	**Wildlife Centers**
		Animal Auctions

Gold Standard

Acquired daily and used for surveillance

Acquired, but not used in daily surveillance

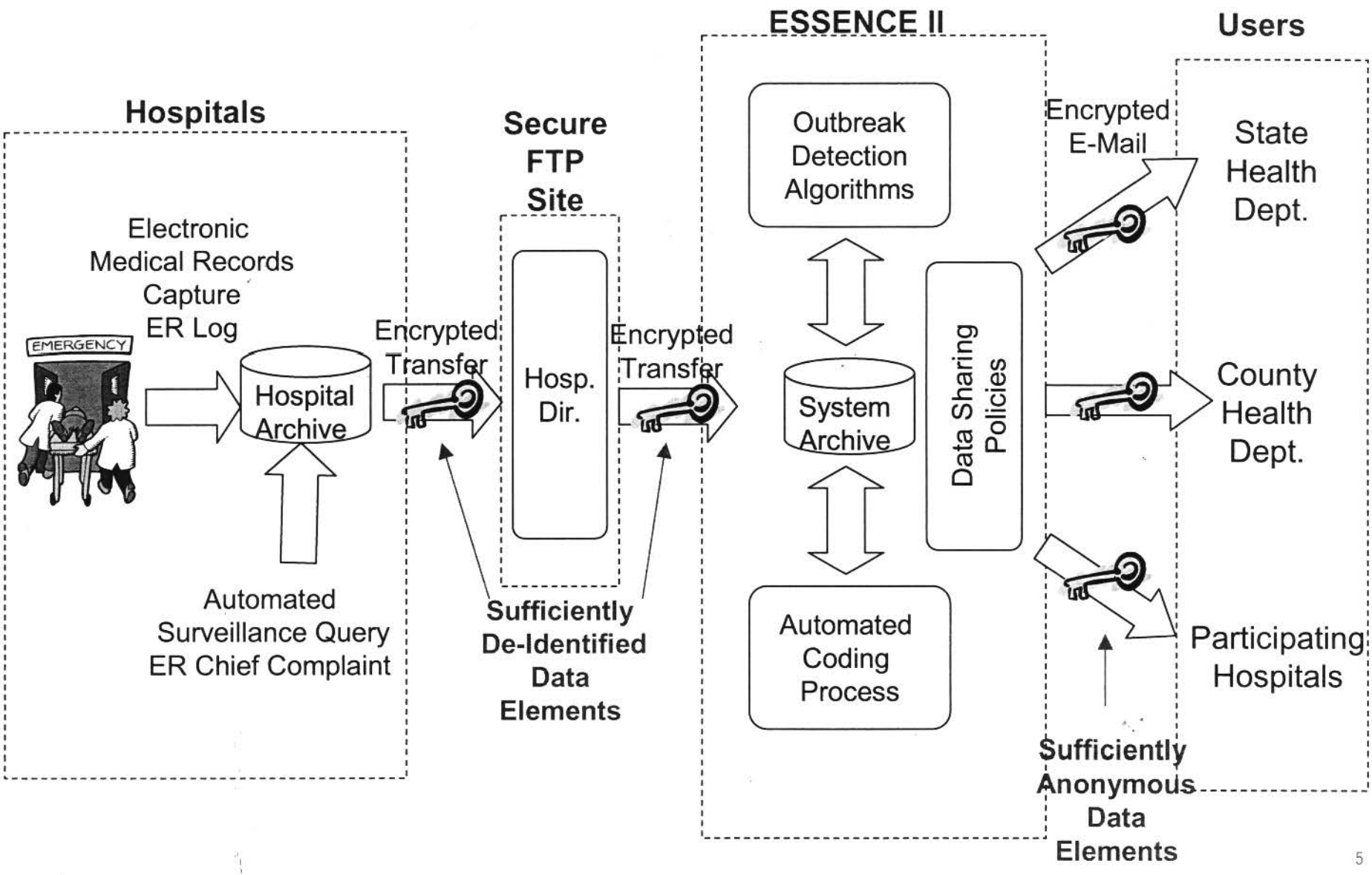
Previously acquired and analyzed, but should be revisited

Acquired but not analyzed

Analyzed, but not currently used for surveillance

Trying to acquire

Automated Data Collection Requirement



Alternative National Data Sources Evaluation

1) **Surveillance Data Inc.** (Prescription claims, Physician Office Claims)

2) **Quintiles Transnational** (Prescription claims, Physician office claims, ER Data)

3) **WebMD** (Prescription claims, Physician office claims, click stream data etc.)

4) **HMO's** (hotline calls, PO claims, Hospital Claims, Prescription claims, Labs, Imaging, Click stream, Patient demographics & history)

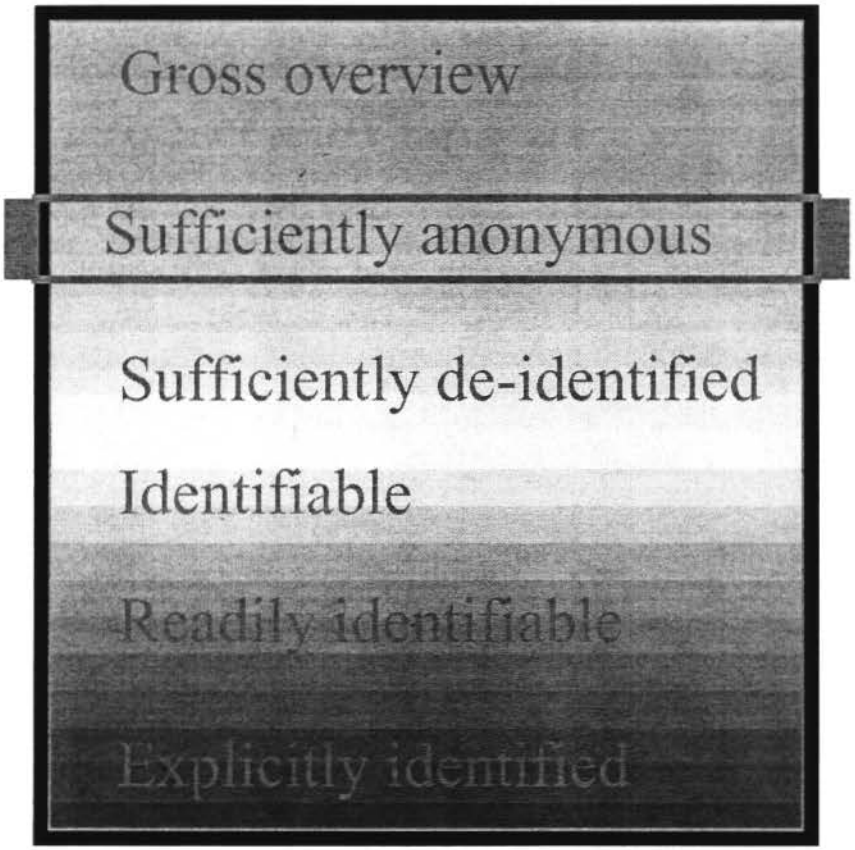
ISSUES:

Value of national data sources for non-traditional health indicators for real-time Surveillance?

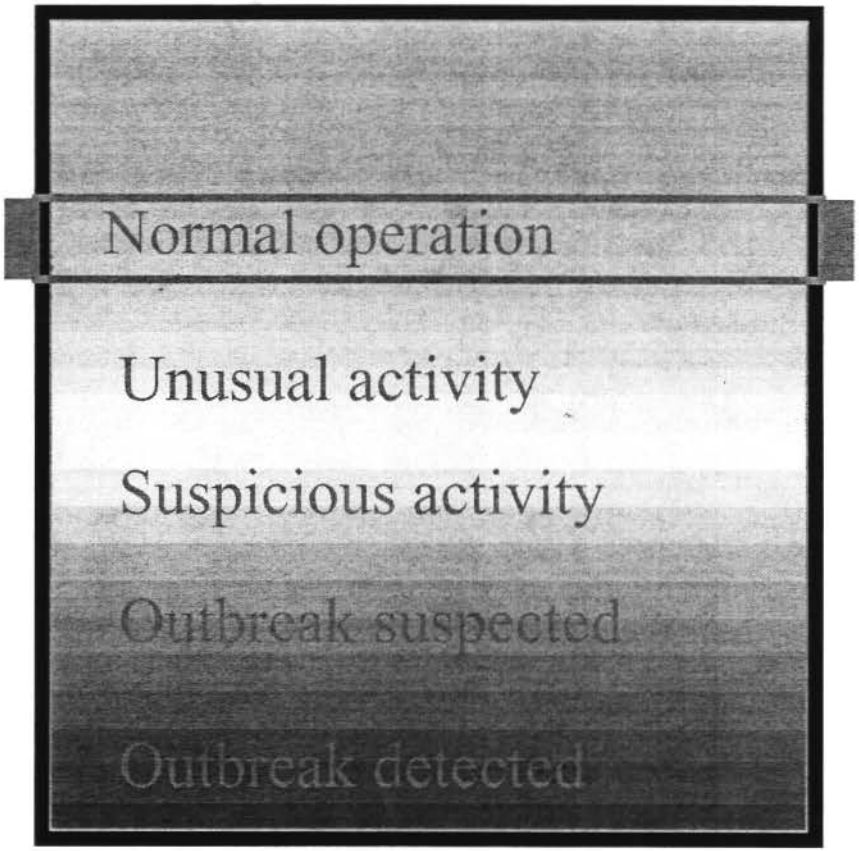
a) Timeliness
b) Coverage for jurisdiction of interest
c) Privacy and dual coverage
d) Contractual arrangements to obtain continuous data stream

Privacy Addressed by CMU's Reasonable Cause Model

Levels of identifiability matching detection status



Datafly Identifiability 0..1



Detection Status 0..1

ESSENCE II Algorithm Approaches

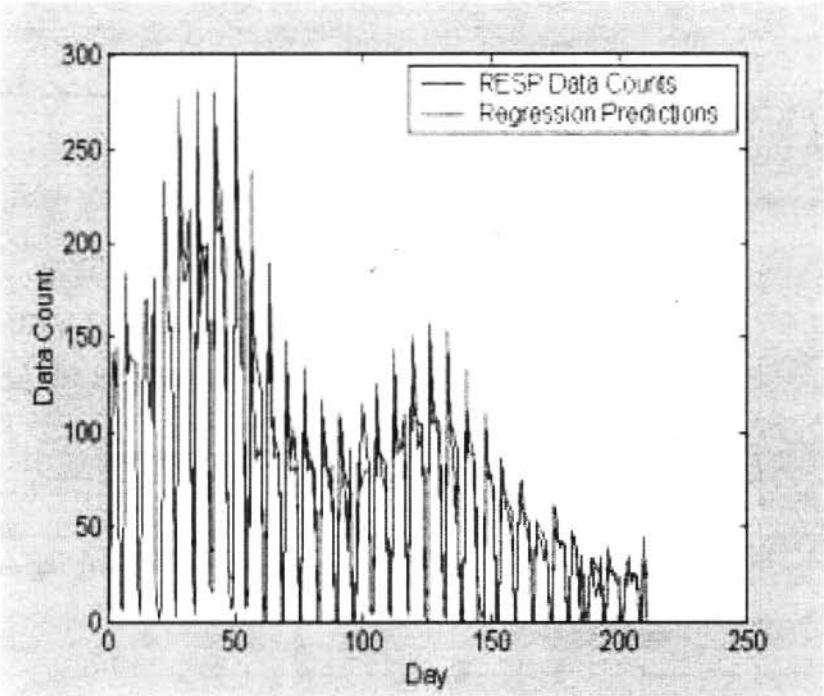
Algorithms are broken down by temporal & spatial features:

	Temporal	Spatial
Automated currently in use	Exponential Weighted Moving Average (EWMA) Auto Regressive	Spatial Odds Ratio
Manual applied	EWMA & AR Early Aberration Reporting System (EARS)	Sat Scan - single source - multiple source
In development	Multivariate - regression - process control	Improved source integration Spatial source estimation models

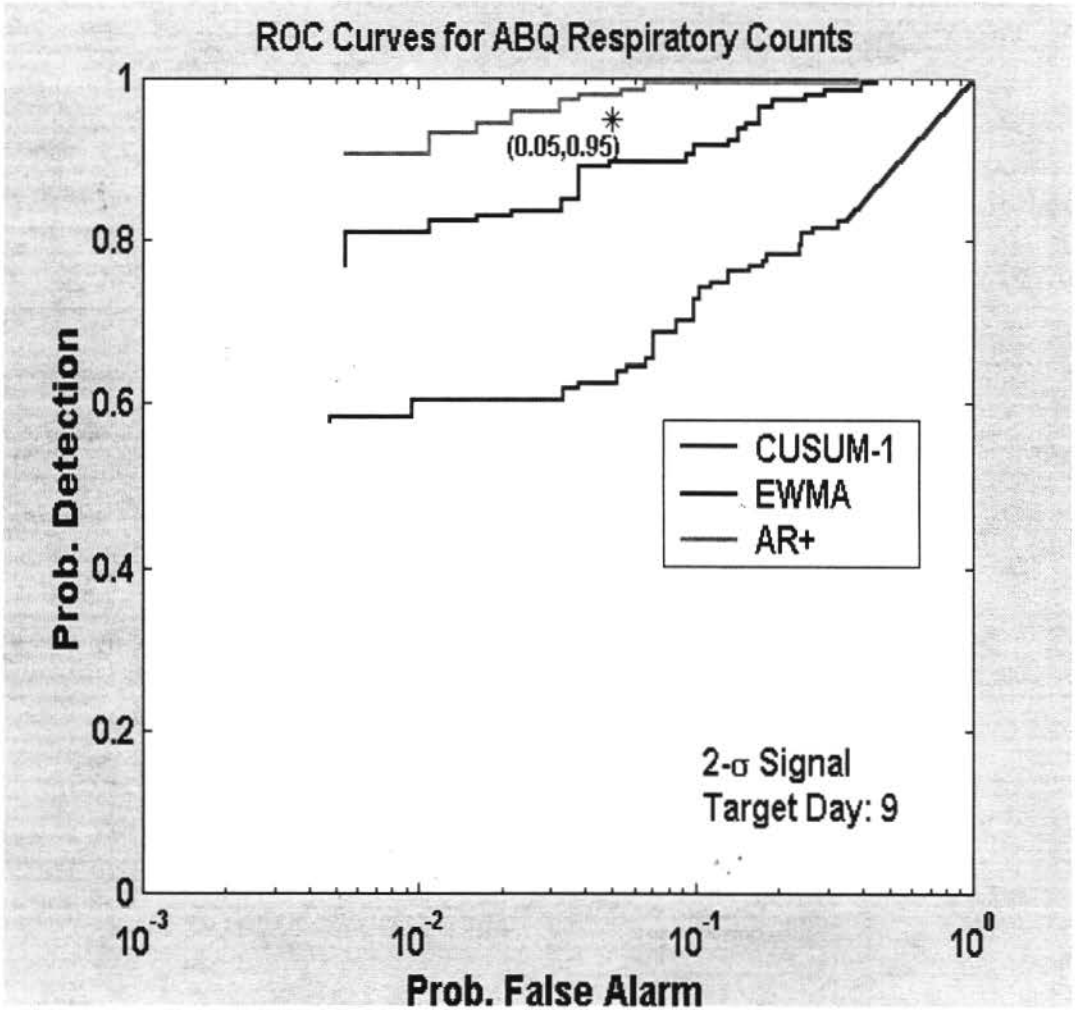
Note: <u>Algorithms are currently tuned to data streams from the NCA & ABQ</u>

Temporal Performance Against ABQ Data

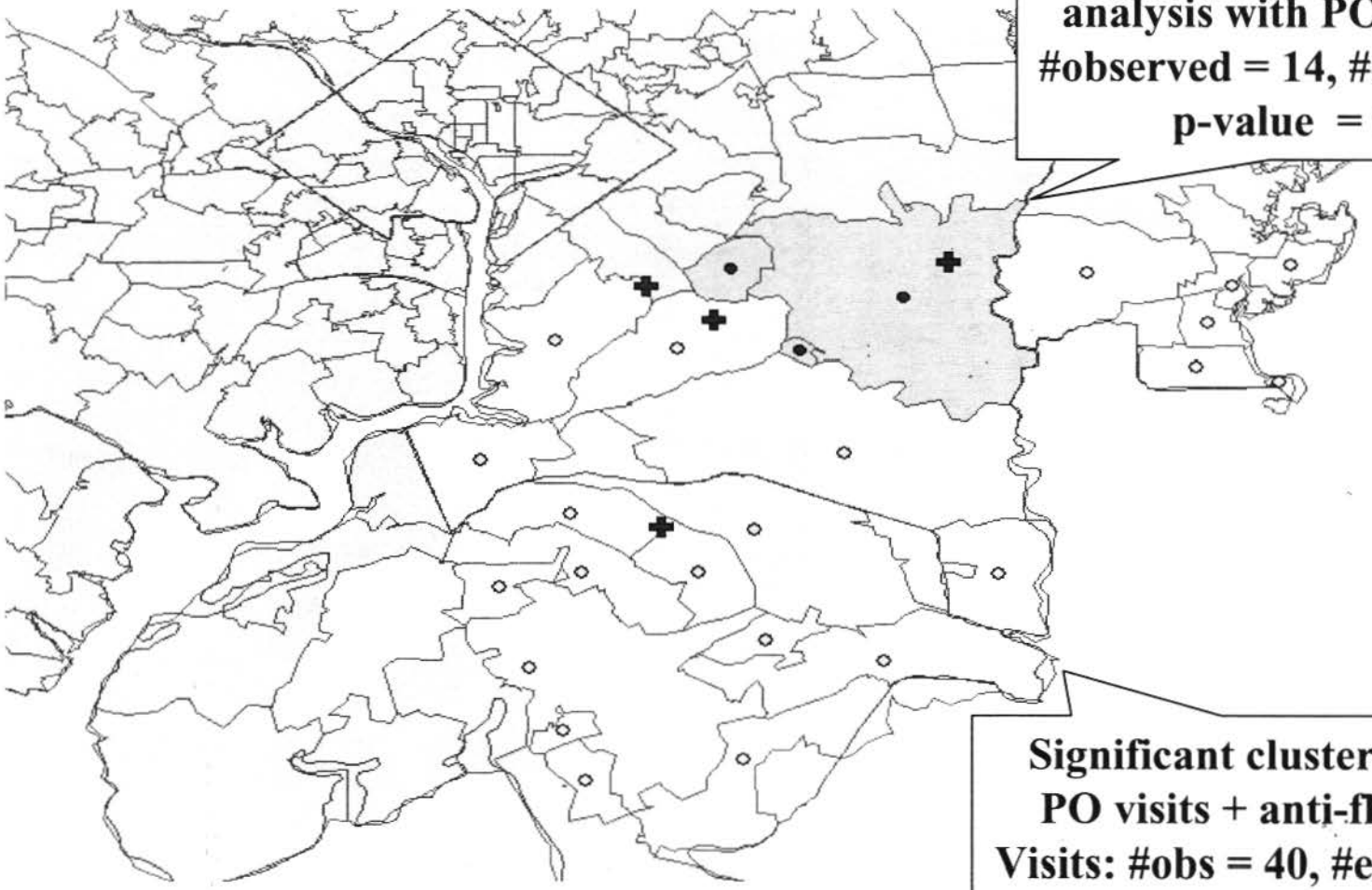
Data



Algorithm Performance



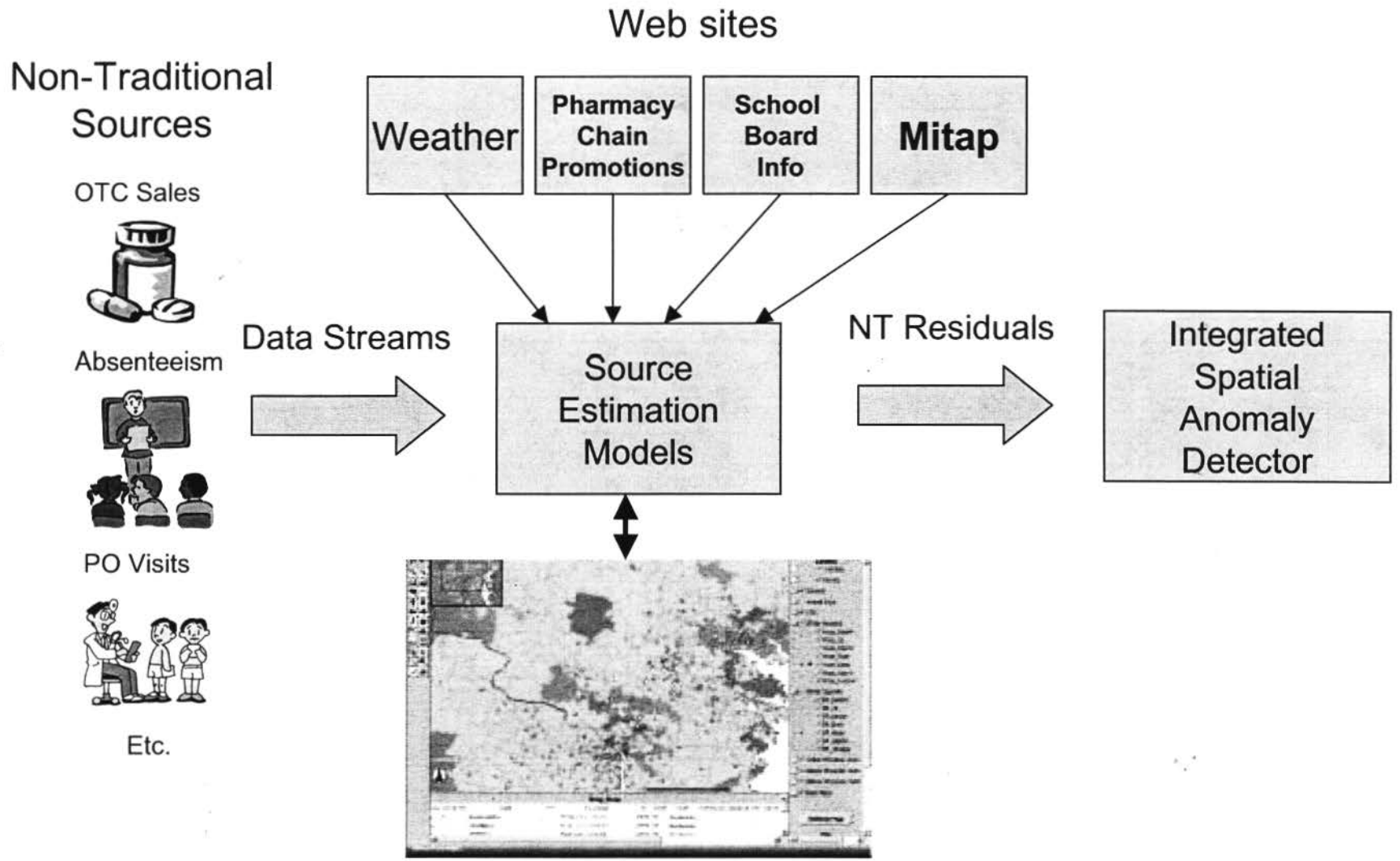
Improved Cluster Recognition Using Multiple Sources



Non-significant cluster in analysis with PO visits only
#observed = 14, #expected = 4
p-value = 0.093

Significant cluster found:
PO visits + anti-flu sales
Visits: #obs = 40, #exp = 25.5
Sales: #obs = 60, #exp = 34.4
p-value = 0.001

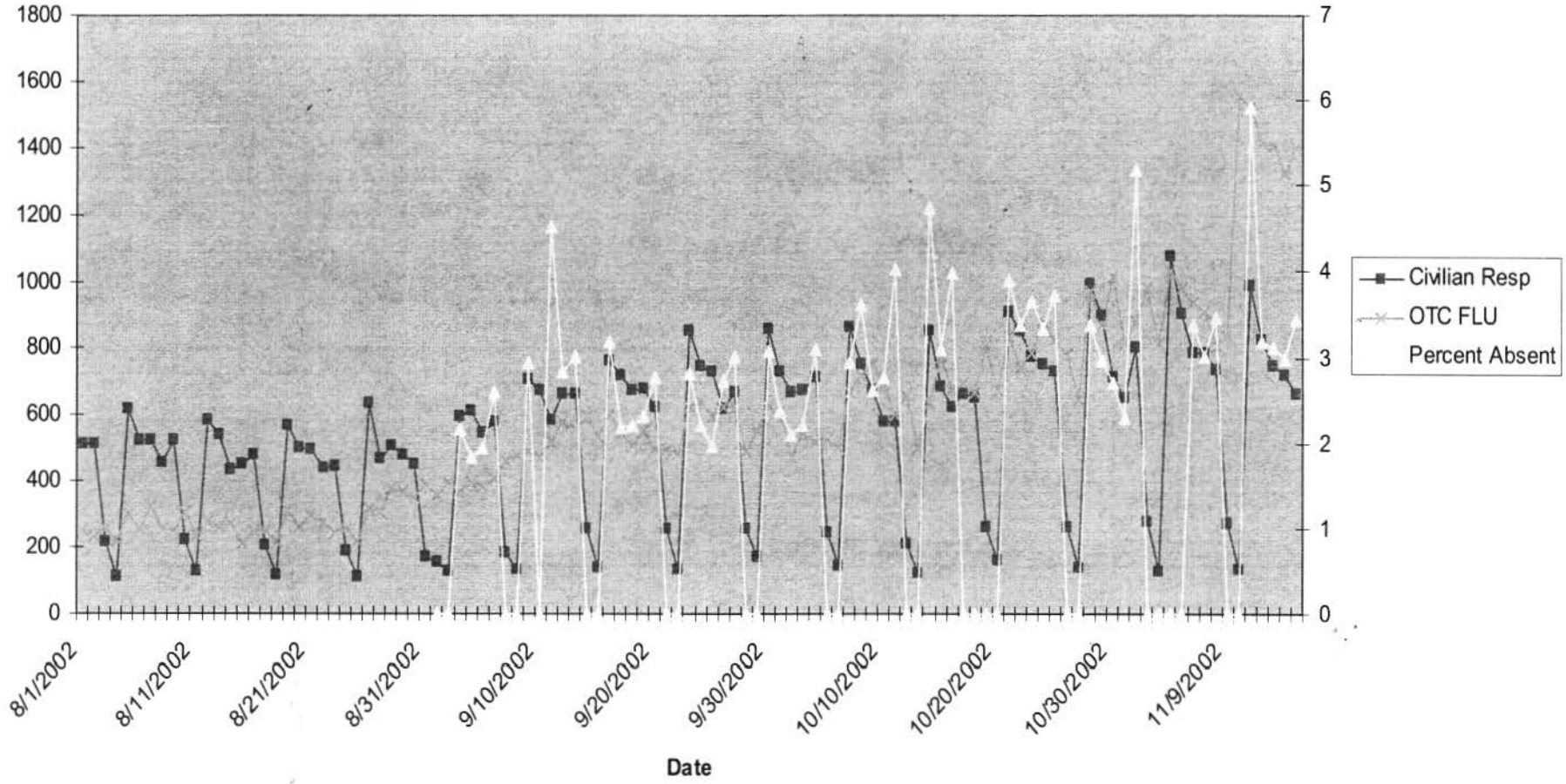
Situation Estimation Approach



ESSENCE II Situation Assessment View

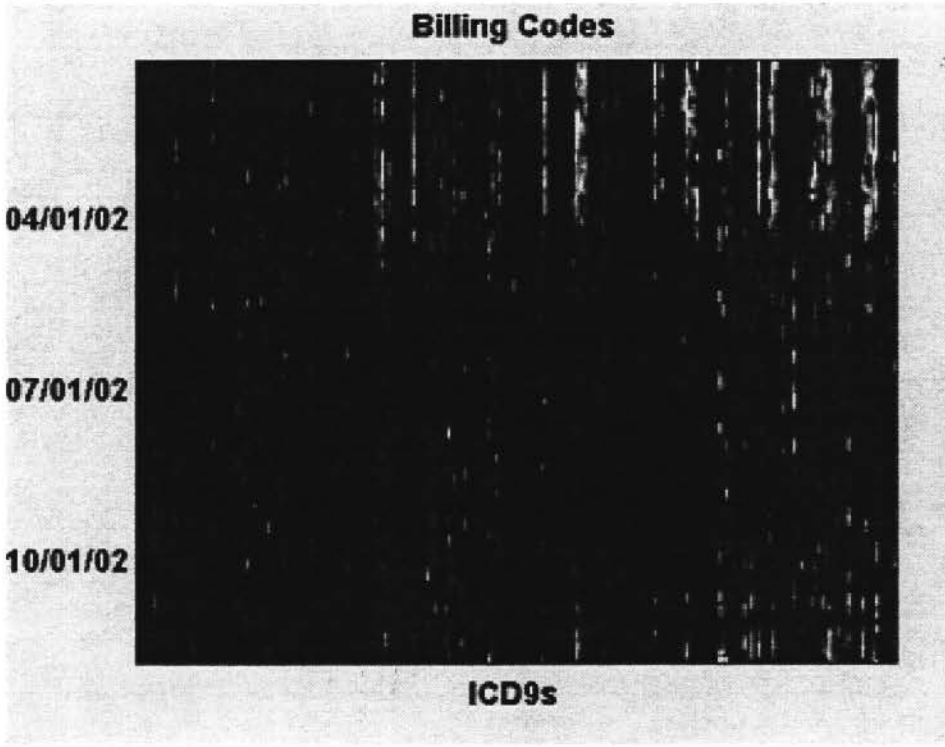
Comparison of Data Source Activity

Civilian Respiratory Claims, OTC FLU sales, and School Absenteeism
8/1/2002 - 11/15/2002

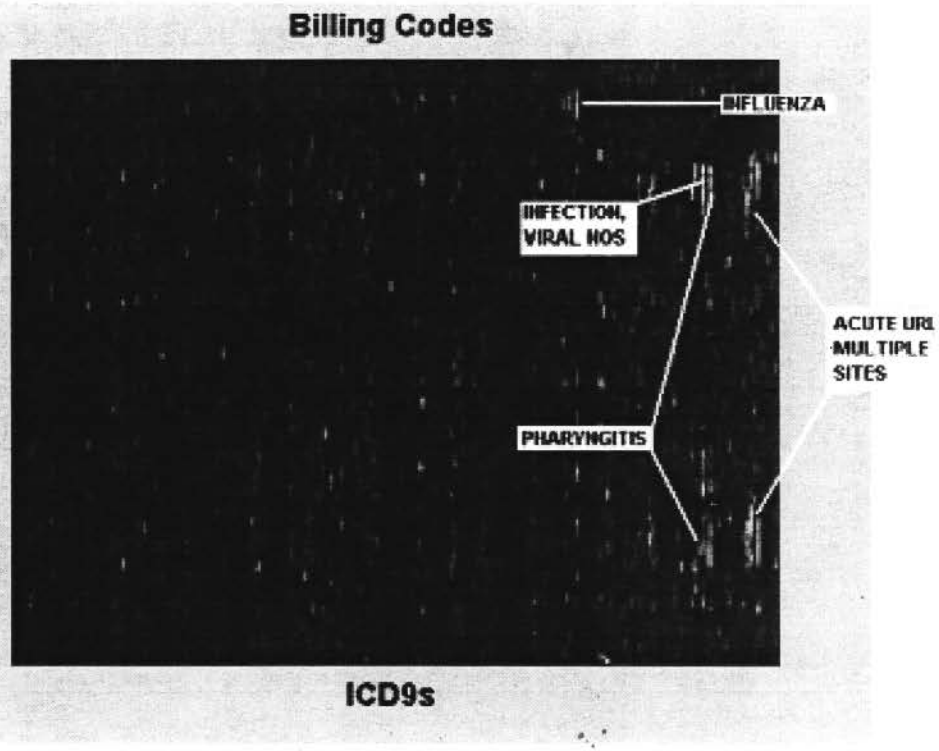


Seasonal Trend Removal
PO Visits / ICD9 Codes

Standard Normalization



Seasonal Trend Removed



Attack Characterization Approach

Rapid Investigation Toolset

GIDEON

Transport / Dispersion Models

Environmental Disease Mapping

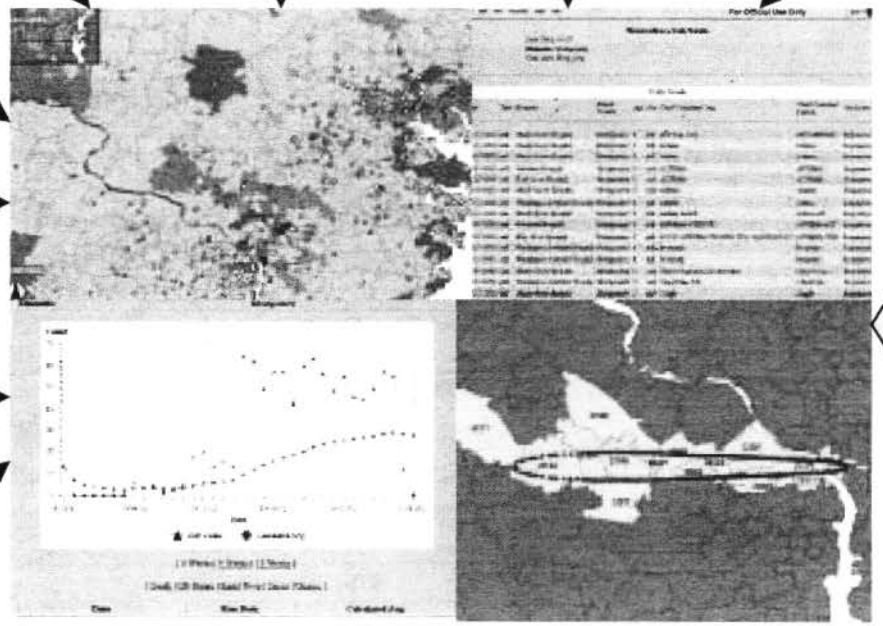
BioWars SIM

Situation Assessment

Raw Data Content

Disease Cluster Information

Disease ID Templates



Epidemiologist's Experience & Judgment

Raw & Processed NT Data

Integrated View of Situation Predictions

Completed
In development
Supplied by others

Performance Issues

What is working and what is not?

Working:

1) Collaboration with DoD-GEIS & local health departments
2) Interest in transitioning technology to other DoD programs & CDC
3) Developing an understanding of how to conduct surveillance with non-traditional data sources.

What is not:

1) It is difficult to perform development and conduct surveillance at the same time.
2) BioWars as an evaluation tool is not working well

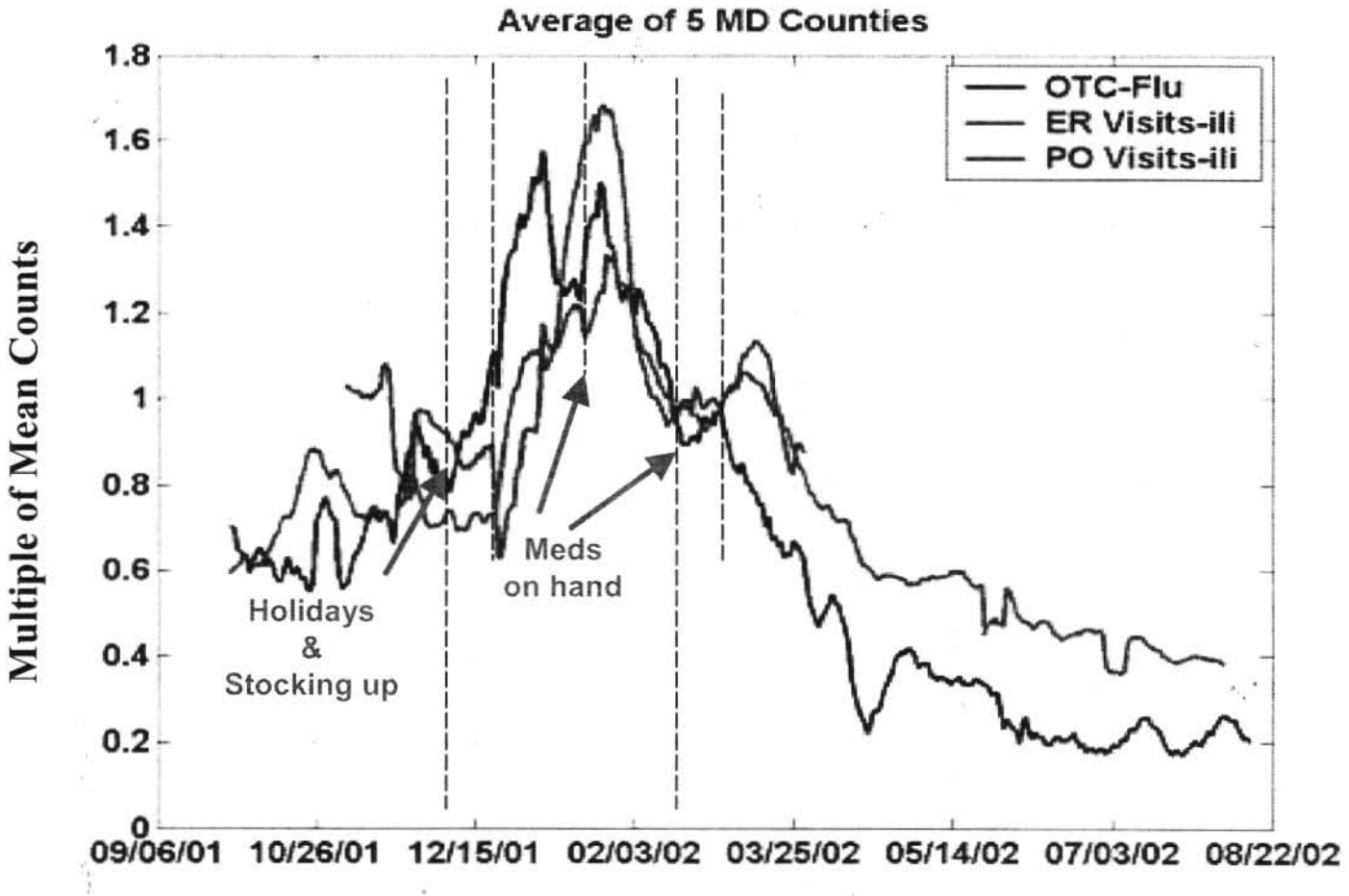
Needs Improvement:

1) Collaboration with other investigators both on the ESSENCE II team and on other teams.
2) Capture and continuity of data sources

Performance Issues

Non-Traditional Data: How Much Earlier?

Function of the length of time between disease onset and acute phase!



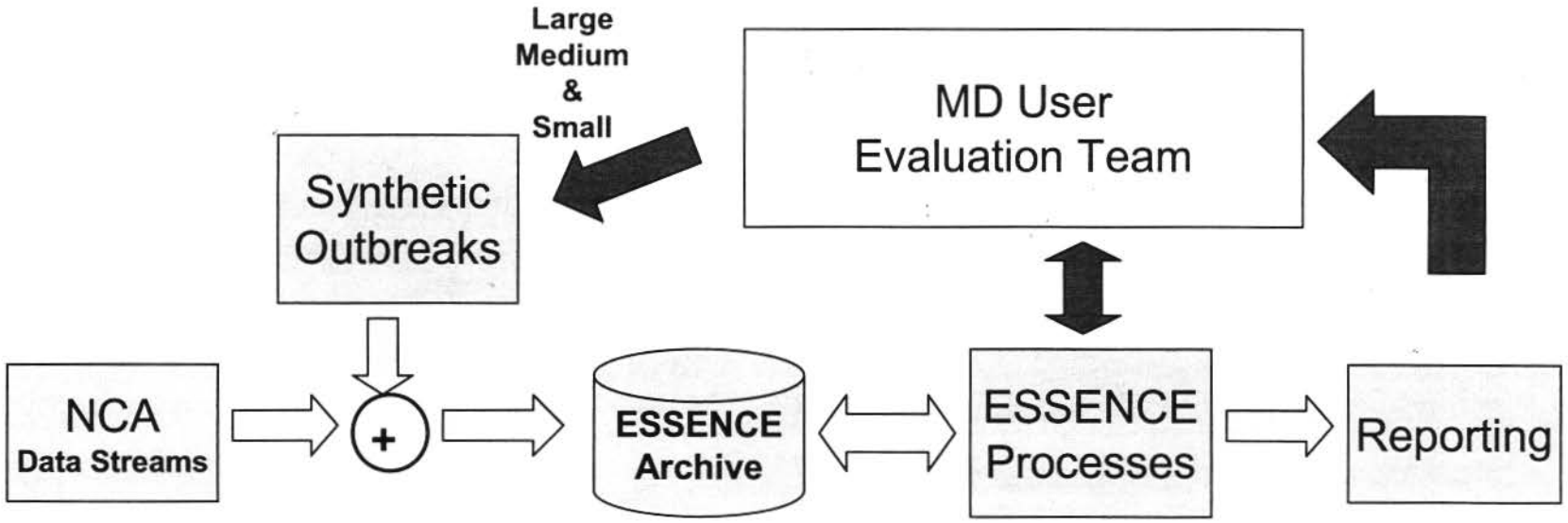
ESSENCE II Accomplishments FY02

What Has Been Accomplished So Far ?

1) Performing surveillance for the NCR & world wide using technology developed under this program.
2) Gained significant support and collaboration with the local public health jurisdictions.
3) Assessed the value of different non-traditional data sources as early Indicators of disease.
4) Surveyed prevailing laws and liability issues inherent in sharing non-traditional data sources and developed anonymity tools to reduce privacy risks and liability.
5) Have demonstrated the ability to identify naturally occurring outbreaks using the DARPA technology.
6) Have acquired new data indicator sources including the development of seeds for a national animal surveillance network.
7) Developed new visualization and drill down user interface for military surveillance.
8) Developed and implemented improved temporal & spatial detectors.
9) Made significant progress in automating the collection and archiving of daily indicator data.

Further Evaluation Plans

Maryland Health Department Blind Evaluation



FY03 Plans

1) Acquire automated HIPAA approved data sources in real-time
2) Determine performance of existing algorithms against BioWars data
3) Analyze existing sources to develop models for the SE module
4) Determine the impact of real-time capture on existing system processes
5) Continue the development and integration of temporal & spatial detectors to improve the alerting time and false alarm rates
6) Assemble the components needed for evaluation of SE and attack characterization modules
7) Continue the integration of components into a deliverable ESSENCE II system for NCA users
8) Continue daily surveillance of NCA data
9) Continue to measure the performance of system components
10) Work with all of the health departments in the NCA to develop a collaborative surveillance environment.

Transition Plans for ESSENCE II

1) Move ESSENCE II Technology into ESSENCE I

2) Deliver basic ESSENCE system to Maryland Health Dept.

3) Provide ESSENCE Technology to DC Health Dept. & support their surveillance activities.

4) Work with MWCOG to develop CONOPS for integrated NCA surveillance.

5) Develop data sharing policies suitable for Northern VA. PH Depts.

6) Work with the Northern Virginia Hospital Alliance

7) Support the DTRA BDI initiatives

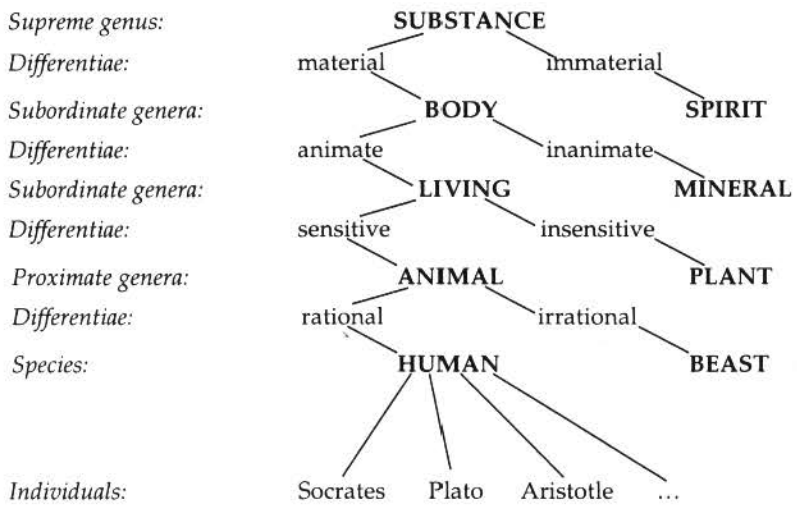
8) Support PEOCBD sensor integration activities

Bio-ALIRT Data Ontology

Mark A. Musen
Zach Pincus
Stanford University School of Medicine
December 4, 2002

1

Ontologies are nothing new



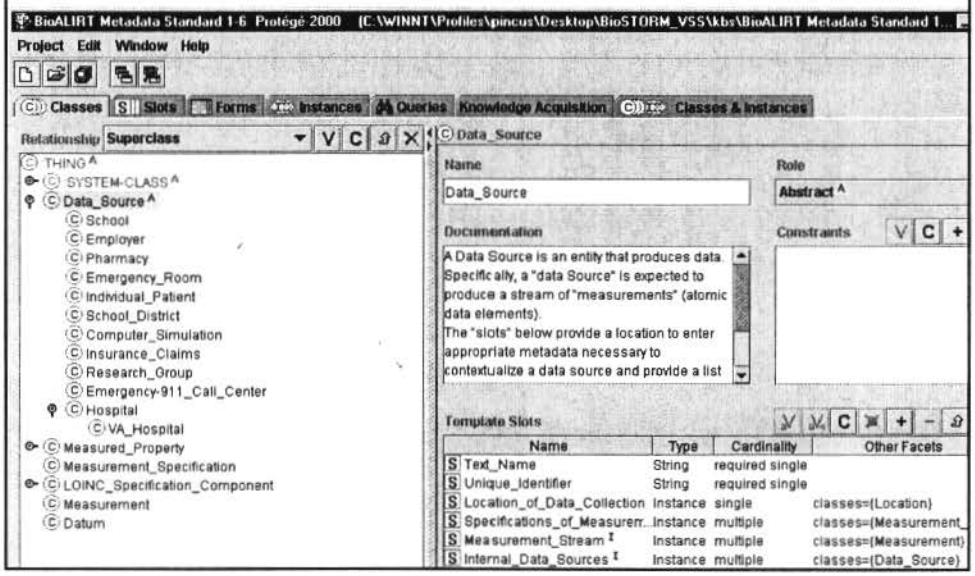
2

An ontology

- Provides a domain of discourse for characterizing some application area
- Enumerates concepts, attributes of concepts, and relationships among concepts, thus defining a structure for the application area
- Defines constraints on relationships among concepts
- Is a *conceptual entity*, that often is encoded in a frame-based or object-oriented representation

3

Bio-ALIRT Ontology



Potential purposes of the Bio-ALIRT ontology:

- Aid Bio-ALIRT program in enumerating and categorizing data sources available from individual projects
- Aid individual projects in building surveillance systems that access diverse data sources in a straightforward, uniform way
- Allow diverse surveillance systems and diverse data sources to interoperate

5

Major Bio-ALIRT ontology elements

- Data sources: Entities that supply data (e.g., pharmacies, schools)
- Measurement streams: Kinds of data streams (e.g., OTC sales, absenteeism rates) available from a given data source
- Measurements: Specific data from a measurement stream
- Measurement specifications: LOINC metadata terms for describing measurements

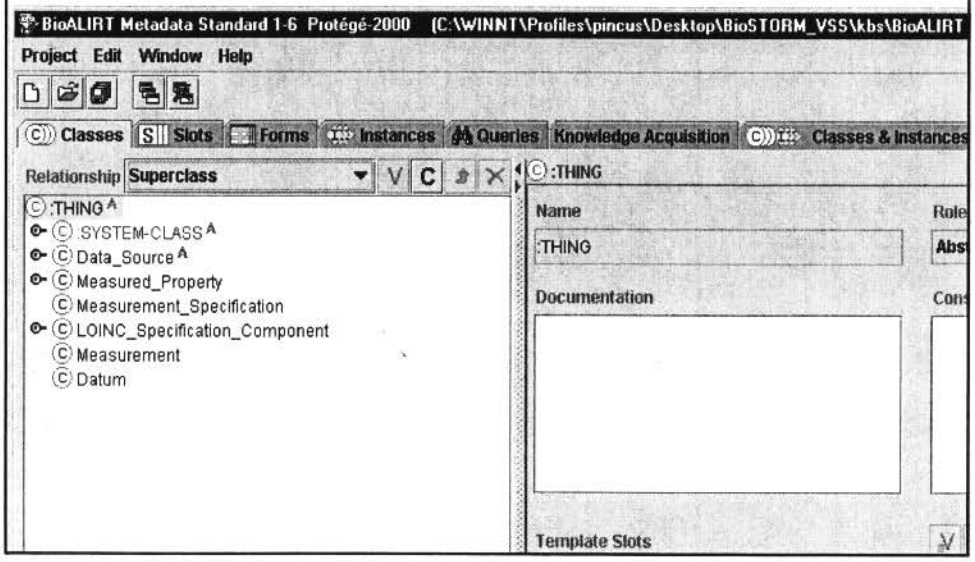
6

An important distinction

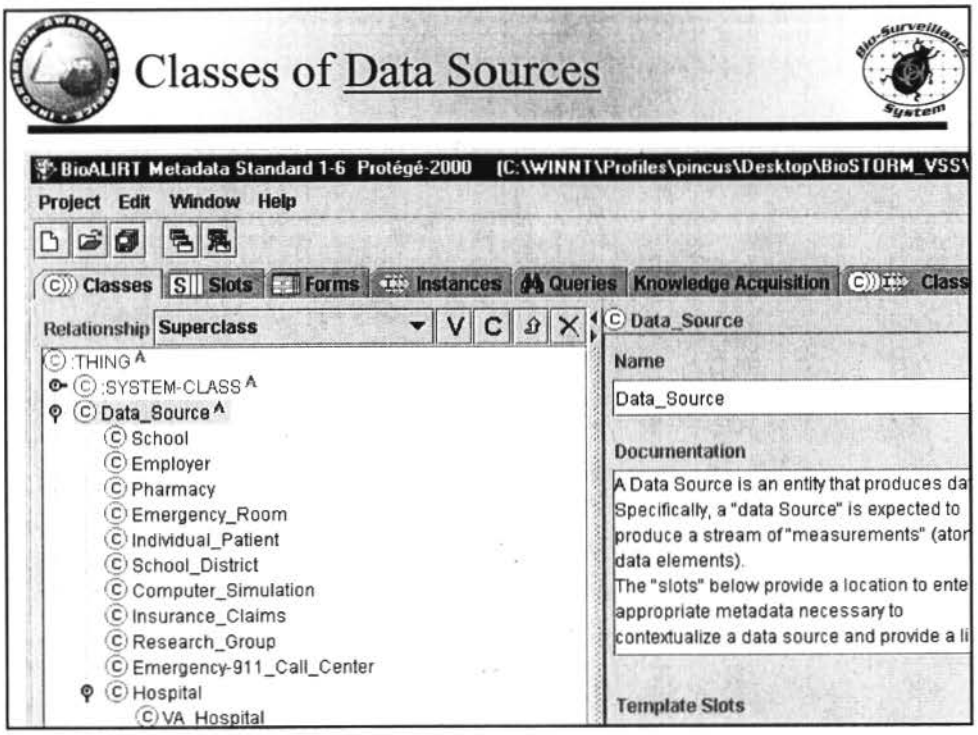
- The **ontology** defines generic properties of data common to all data sources and data streams within the program
- **Instances** of concepts of the ontology define project-specific data sources, data streams, and actual data

7

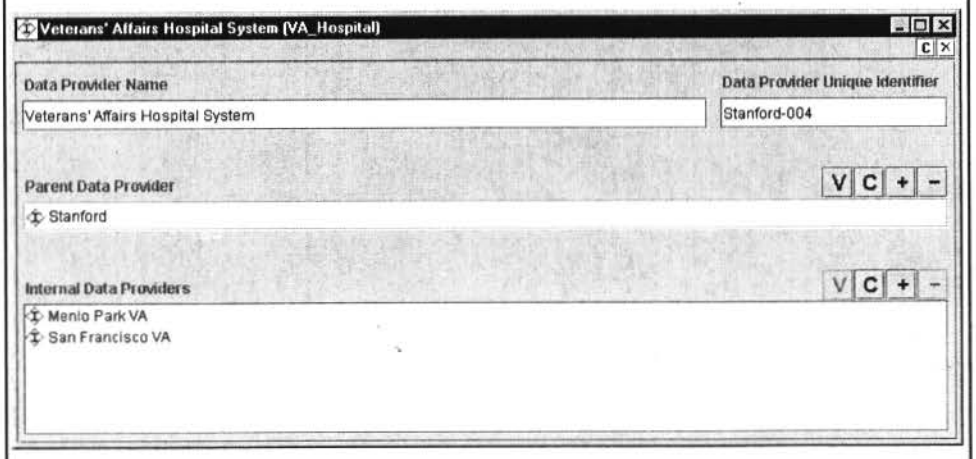
Bio-ALIRT Data Sources Ontology in Protégé-2000



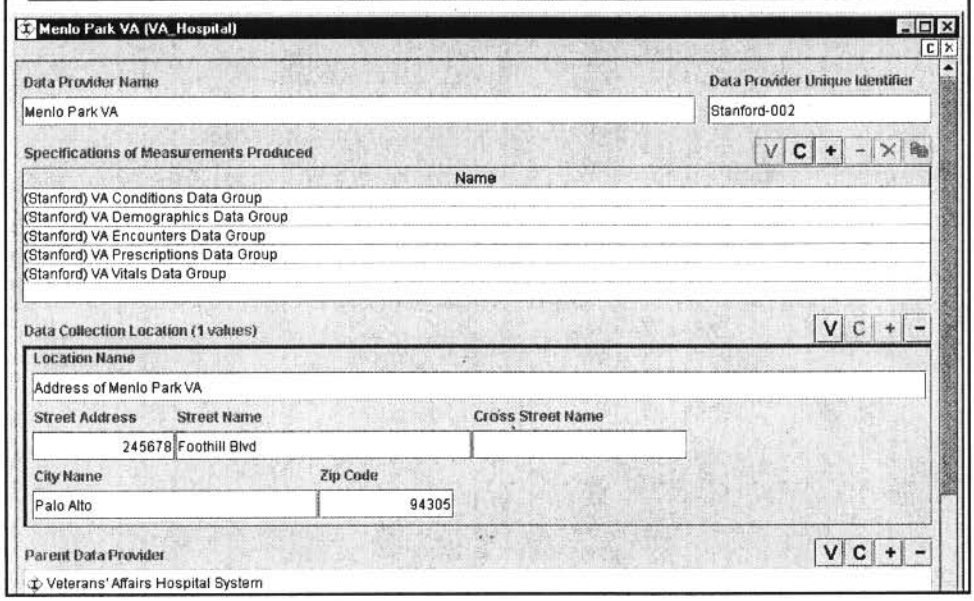
Classes of Data Sources



An Instance of a Data Source



Each Data Source is associated with a number of Measurement Streams



An instance of a Measurement Stream

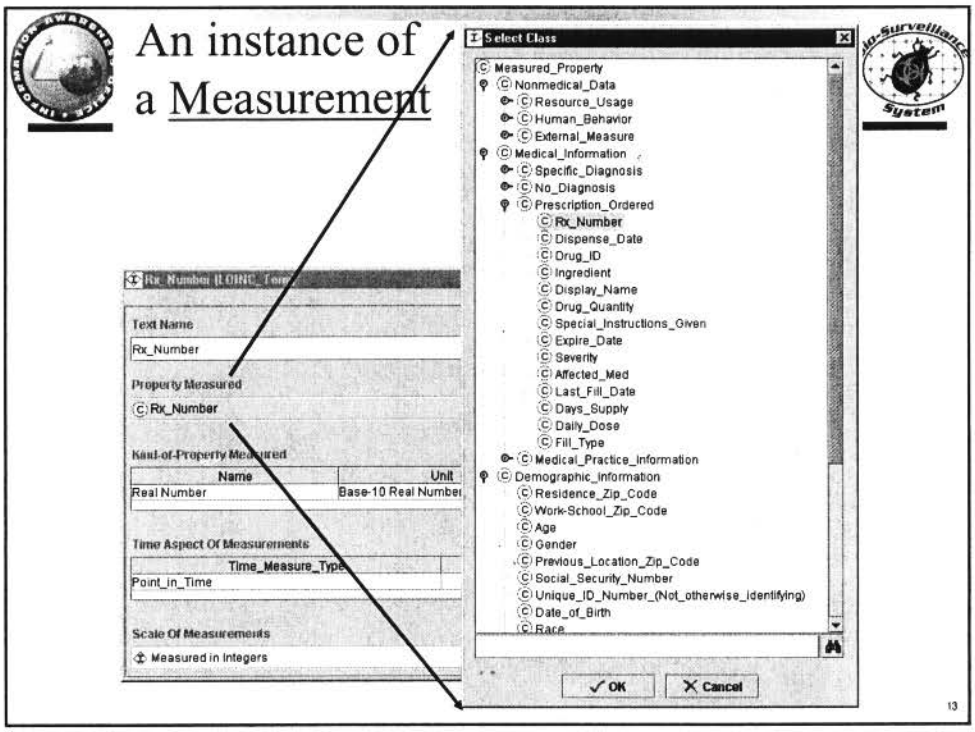
(Stanford) VA Prescriptions Data Group (Measurement_Specification)

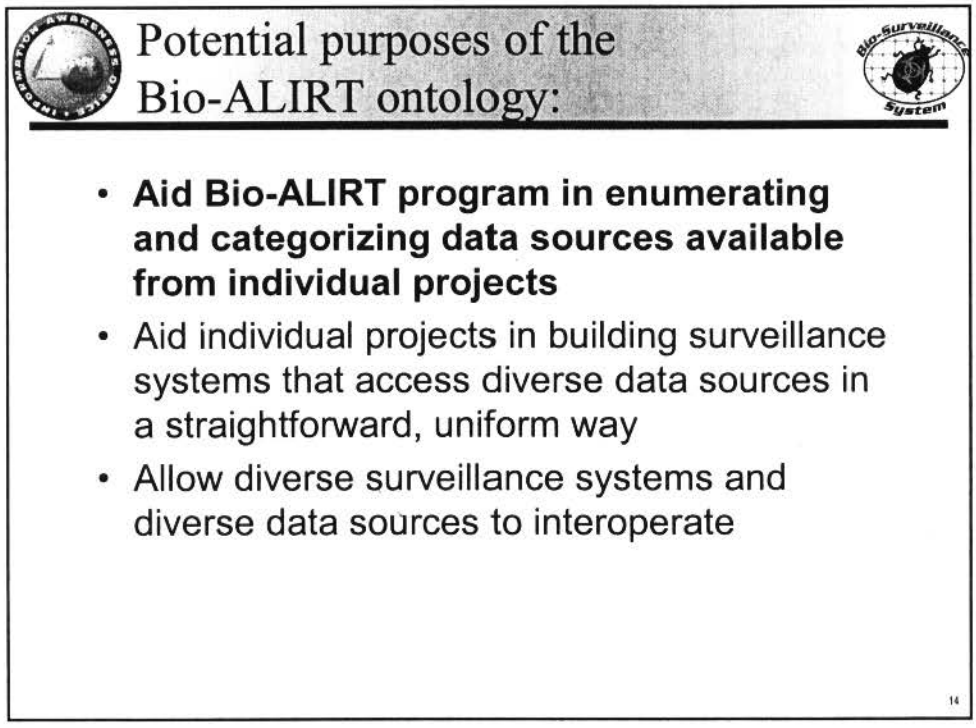
Measurement Specification Name

(Stanford) VA Prescriptions Data Group

LOINC Term(s)

Name	Property_Measured	Kind-of-Property_Measu...	Time_Aspect_of_Measu...	Scale_of_Measurements
Rx_Number	Rx_Number	Real Number	Point_in_Time	Measured in Integers
Dispensed on Date	Dispense_Date	Date	Point_in_Time	Narrative Measurement
Drug ID	Drug_ID	Categorical Measure	Point_in_Time	Measured in Integers
Ingredient	Ingredient	Text	Point_in_Time	Narrative Measurement
Display Name	Display_Name	Text	Point_in_Time	Narrative Measurement
Quantity of Drug	Drug_Quantity	Count of Integers	Point_in_Time	Measured in Integers
Special Instructions Given	Special_Instructions_Give	Text	Point_in_Time	Narrative Measurement
Division Issuing Prescript	Division	Text	Point_in_Time	Narrative Measurement
Rx Expires on on Date	Expire_Date	Date	Point_in_Time	Narrative Measurement
Severity	Severity	Text	Point_in_Time	Narrative Measurement
Affected Medications	Affected_Med	Text	Point_in_Time	Narrative Measurement
Last Refill Date	Last_Fill_Date	Date	Point_in_Time	Narrative Measurement
Days Supplied	Days_Supply	Count of Integers	Point_in_Time	Measured in Integers
Daily Dose	Daily_Dose	Real Number	Point_in_Time	Measured in Real Numbe
Station Number	Station	Raw Integer	Point_in_Time	Measured in Integers
Fill Type	Fill_Type	Categorical Measure	Point_in_Time	Rx Fill Types
Provider Name	Provider_Name	Text	Point_in_Time	Narrative Measurement


An instance of a Measurement
Select Class
Measured_Property
Nonmedical_Data
Resource_Usage
Human_Behavior
External_Measure
Medical_Information
Specific_Diagnosis
No_Diagnosis
Prescription_Ordered
Rx_Number
Dispense_Date
Drug_ID
Ingredient
Display_Name
Drug_Quantity
Special_Instructions_Given
Expire_Date
Severity
Affected_Med
Last_Fill_Date
Days_Supply
Daily_Dose
Fill_Type
Medical_Practice_Information
Demographic_Information
Residence_Zip_Code
Work-School_Zip_Code
Age
Gender
Previous_Location_Zip_Code
Social_Security_Number
Unique_ID_Number_(Not_otherwise_identifying)
Date_of_Birth
Race
OK
Cancel
Text Name
Rx_Number
Property Measured
Rx_Number
Kind-of-Property Measured
Name
Unit
Real Number
Base-10 Real Number
Time Aspect Of Measurements
Time_Measure_Type
Point_in_Time
Scale Of Measurements
Measured in Integers
13


Potential purposes of the Bio-ALIRT ontology:
• Aid Bio-ALIRT program in enumerating and categorizing data sources available from individual projects
• Aid individual projects in building surveillance systems that access diverse data sources in a straightforward, uniform way
• Allow diverse surveillance systems and diverse data sources to interoperate
14

Advantages to the Bio-ALIRT program

- Common location for enumerating available data sources in machine-processable form
- Common language for describing nontraditional data sources, for which no standards currently exist
- Extensible framework for incorporation of new data descriptions as new surveillance targets are identified
- Use of metadata consistent with LOINC standard
- Ability to translate conceptual ontology contents into a variety of syntaxes (e.g., NEDDS)

15

Potential purposes of the Bio-ALIRT ontology:

- Aid Bio-ALIRT program in enumerating and categorizing data sources available from individual projects
- **Aid individual projects in building surveillance systems that access diverse data sources in a straightforward, uniform way**
- Allow diverse surveillance systems and diverse data sources to interoperate

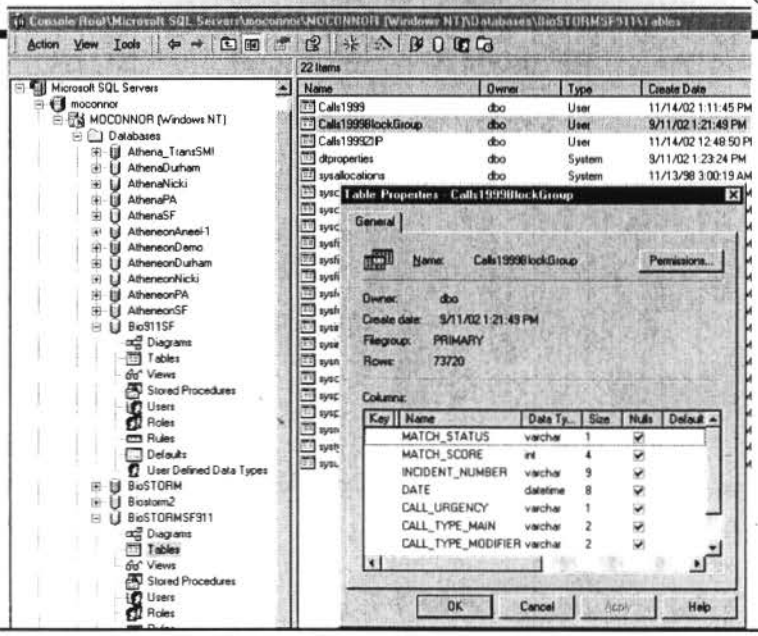
16

Ontology serves as a Data Broker

- Ontology offers a consistent interface to diverse data
  - Data are accessed by analytic problem solvers using general-purpose routines that rely on the ontology to map generic queries onto particular data resources
  - Surveillance algorithms need to know nothing about original, raw format of data
- Data sources and data flow can be managed on an abstract level

17

911 data in relational database



18

Ontology defines how data should be accessed from the database

(Stanford) Emergency-911 Call Data Group (Measurement_Specification)

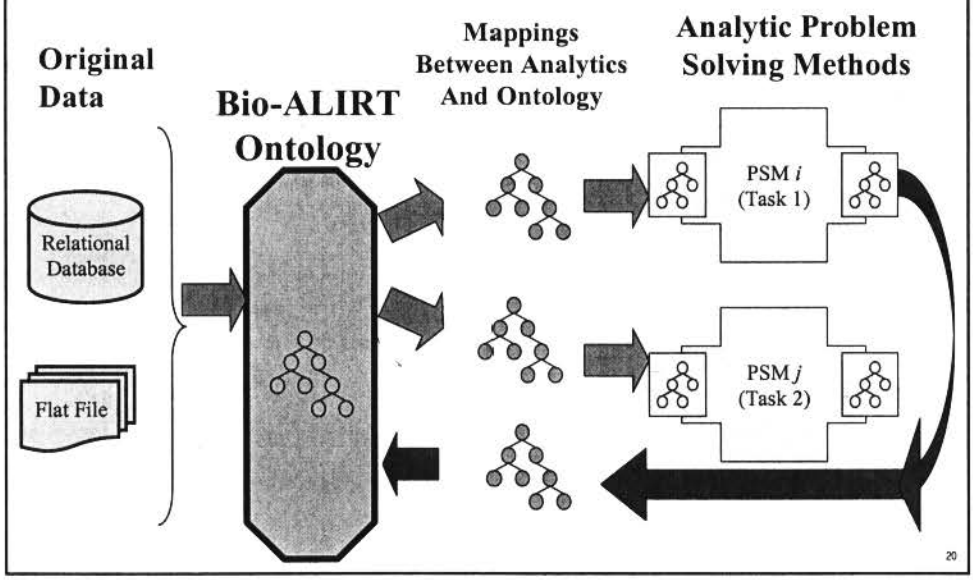
Measurement Specification Name

(Stanford) Emergency-911 Call Data Group

LOINC Term(s)

Name	Property_Measured	Kind-of-Property_Measured	Time_Aspect_of_Mea
Status	Status	Categorical Measure	Point_in_Time
Score	Score	Percent	Point_in_Time
Side	Side	Categorical Measure	Point_in_Time
GIS X-coordinate	GIS_X_Coordinate	Real Number	Point_in_Time
GIS Y-coordinate	GIS_Y_Coordinate	Real Number	Point_in_Time
Incident ID Code	Incident_ID	Text	Point_in_Time
Date	Date_of_Visit	Date	Point_in_Time
Site	Hospital_Address	Text	Point_in_Time
Area	Area_Coding	Real Number	Point_in_Time
Census Block Group	Census_Block_Group	Categorical Measure	Point_in_Time
Type	CBD_Call_Type	Categorical Measure	Point_in_Time

Data Broker Approach



Use of the ontology as a mediator makes our system extensible

- When new data come along, analytic problem solvers do not require any additional programming! Only new mappings are required.
- When new analytic problem solvers are implemented, access to existing data sources is no big deal. At most, new mappings are required.

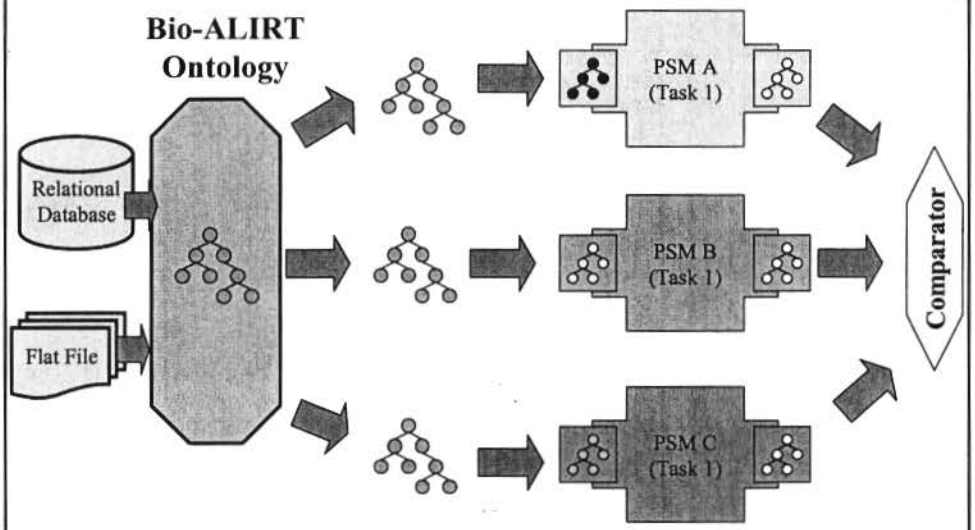
21

Potential purposes of the Bio-ALIRT ontology:

- Aid Bio-ALIRT program in enumerating and categorizing data sources available from individual projects
- Aid individual projects in building surveillance systems that access diverse data sources in a straightforward, uniform way
- **Allow diverse surveillance systems and diverse data sources to interoperate**

22

Evaluating Alternative Analytic Problem-Solving Methods



PSM = Problem-Solving Method

23

Possible approach for next year's challenge problem

- Use ontology to mediate between analytic methods and data streams
  - For each group, compare performance of that group's analytic method(s) on some standard set of data streams
  - For each group, compare performance of that group's analytic method(s) on real-world data streams from the other groups
- Use this approach *within groups* to assess alternative analytic methods

24

Next steps for the ontology

- Stanford will be making another round of revisions to the ontology in the next few weeks based on the group's feedback
- All program personnel should continue to make suggestions to Stanford for extensions necessary to accommodate existing and forthcoming data streams
- Projects should use the ontology to catalog their particular data sources and data streams

25

Potential purposes of the Bio-ALIRT ontology:

- Aid Bio-ALIRT program in enumerating and categorizing data sources available from individual projects
- Aid individual projects in building surveillance systems that access diverse data sources in a straightforward, uniform way
- Allow diverse surveillance systems and diverse data sources to interoperate

26

1 of 5

Dimensionality Reduction of Multiple Time Series Data

Ron Bloom
Veridian, Los Angeles

1

Rationale for Grouping

- Make multivariate model building and prediction more tractable (reduce number of model parameters).
- Improve statistics in low count rate series.
- Simplify interpretation of input/output using aggregate quantities.
- Reduce confounding due to promotions and shelf re-organization

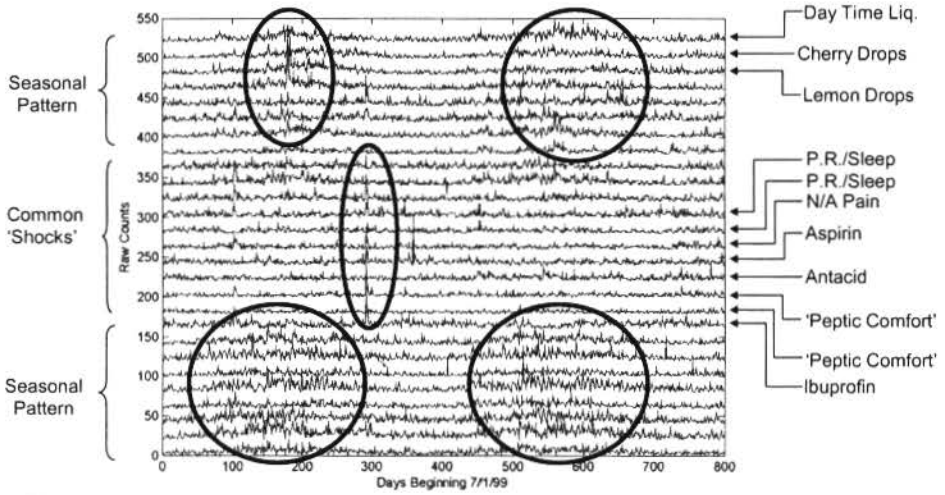
2

Combination methods examined

- Rational grouping according to spectral features (common weekly cycles).
- Hierarchical linear aggregation based on correlation matrix (common seasonal/annual behavior).
- Principal Component Analysis (PCA) (common monthly trends).
- Aggregation according to common short time-scale behavior.

3

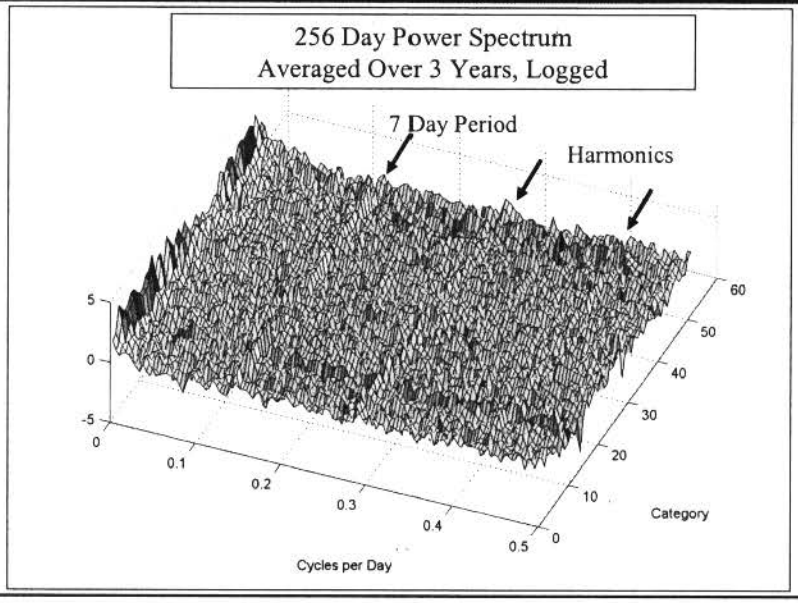
A Candidate For Dimensionality Reduction: 51 OTC Drug Types at Norfolk Area Exchange



Data also contain a weekly cycle buried in these fluctuations

4

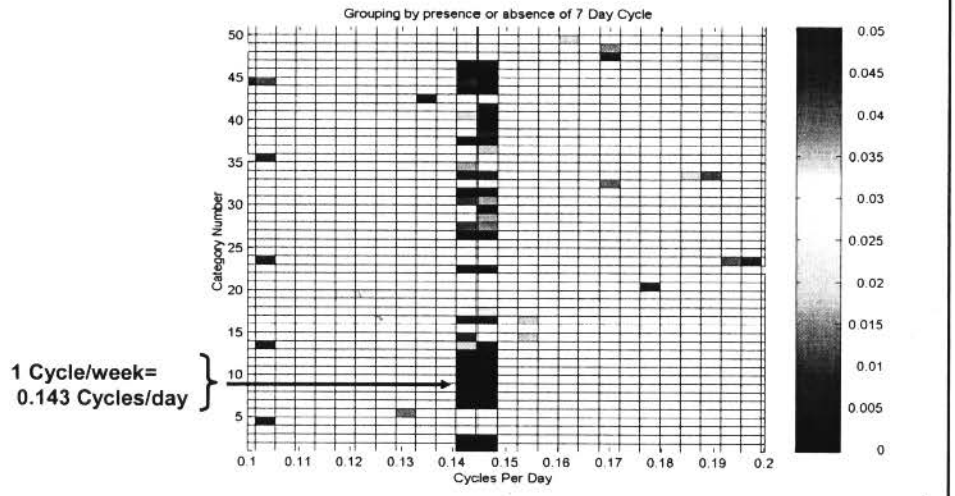
Method 1: Aggregation by Presence of Weekly Cycle



5

Power Spectra Clearly Resolves Presence/Absence of Weekly Cycle

Colored cells show significant power at the 5% level by local F-ratio test.



6

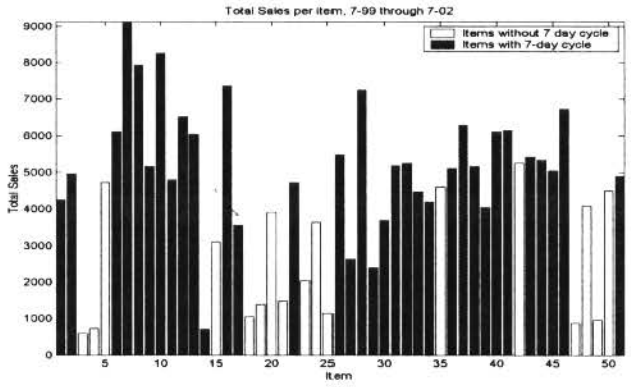
OTC items grouped according to 7 day cycle in purchase volume

Category:	7 Day Cycle Present?	
	Yes	No
1) Pain/Fever Only	7	2
2) #1+ Flu SS	5	0
3) Cough Drop/Throat Only	14	1
4) #3+ Flu SS	0	1
5) Antihistamine/Sinus/Decong.	4	3
6) Stomach/GI	5	3
7) Skin/Rash/Itch	0	5
8) Child	0	1

7

Other Influences on Weekly Cycles? Some evidence for the following:

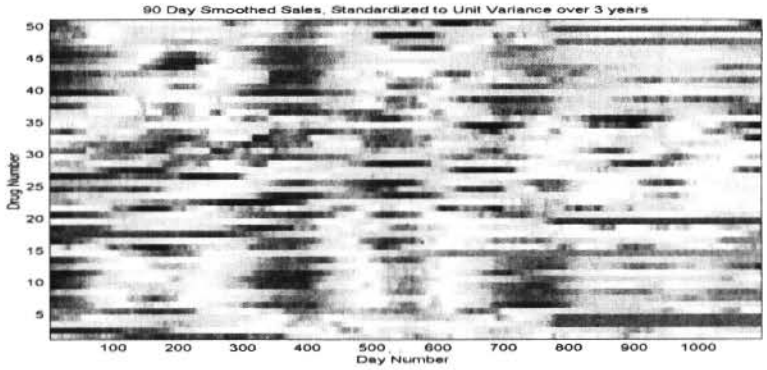
1. Generic drugs: more likely to show 7 day cycles.
2. Large package sizes: less likely to show 7 day cycle.
3. Total Sales: low volume items less likely to show cycle (see graph).



8

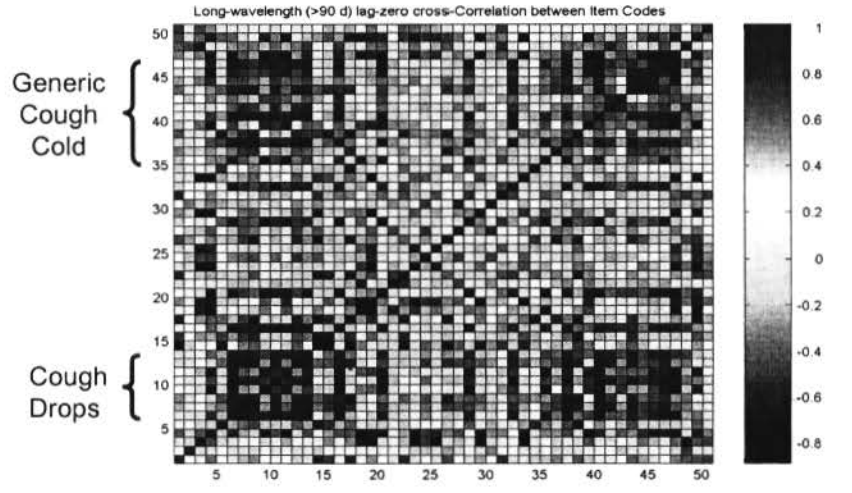
Method 2: Seasonal Trend Grouping Using Hierarchical Aggregation Algorithm

1. Low-pass filtered data used to calculate correlations with wavelengths longer than 90 days.
2. Correlation value used to merge individual time-series in recursive clustering algorithm.
3. 7 distinct groupings obtained from 51 Norfolk OTC drug categories



9

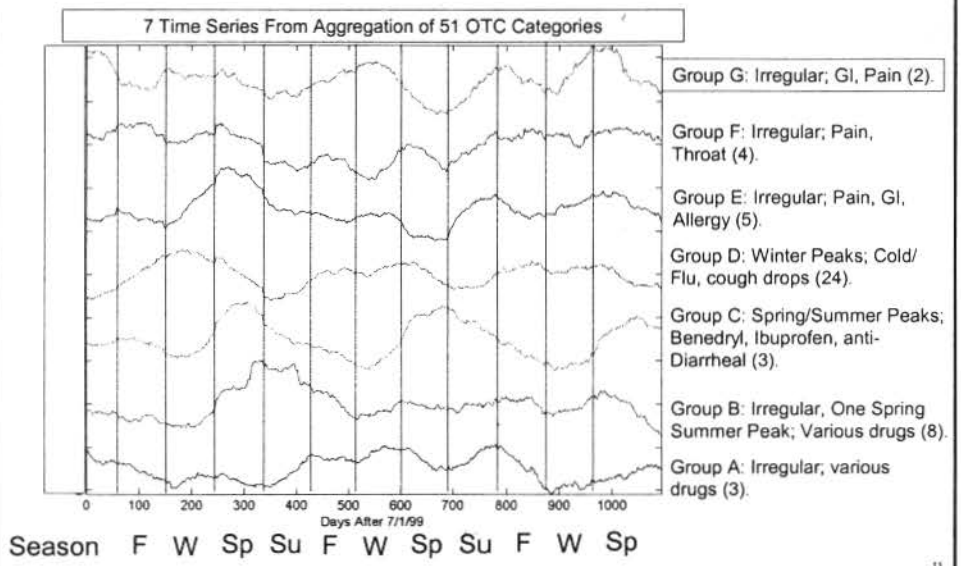
Long Term Cross Correlation Structure for 51 OTC Drug Categories



Full matrix used to initialize recursive clustering algorithm

10

Dimensionality Reduced from 51 to 7 on the Basis of Seasonal Correlation



11

Grouping by Seasonal Trends

Drugs With Winter-Time Peaks

A	
1	MOTRIN TABLETS 24'S 00001707
15	C/S IMODIUM AD CAPLET 6 CT 00008568
17	TYLENOL PM CAPLET 24S 00011926
22	TUMS ASST FLAVORS 3 ROLL PACK 00013905
31	NEX ANTACID TAB FLAVORED 150CT 00023455
B	
2	MOTRIN BLISTER PK 8CT 00001769
3	BANDAID PLUS W/ MYCITRACIN .5Z 00001811
4	CALADRYL LOTION 6oz 00001827
19	CORTIZONE 10 CREAM 1.3/1 OZ 00013006
26	ALEVE TAB 24CT 00022918
32	NEX THIN-COAT ASPIRIN 100 TABS 00023612
47	BENADRYL ITCH RELIEF STCK 14ml 00074711
51	EXCEDRIN MIGRANE CAPLETS 24 CT 00194410
C	
5	C/S BENADRYL ALGY/CAPSLS 25 mg 00001905
28	NEX IBUPROFEN TAB 200MG 50 CT. 00023312
35	NEX ANTI-DIARRHEAL CAPLETS 6'S 00023535
D	
6	HALLS COUGH DROPS BAGICECUBES 00001935
7	C/S HALLS COUGHDROPS HONEY LEM 00001942
8	C/S HALLS COUGH DROPS BAG 30's 00001949
9	HALLS COUGH DROPS HONEY LEMON 00001957
10	HALLS COUGH DROPS STK CHERRY 00001966
11	HALLS COUGH DRP STK ICE BLUE 00001977
12	HALLS COUGH DROP STK MENTLYPTS 00001986
13	HALLS VITAMIN C DROPS 00001997
14	ALKA SELZER NIGHT TIME 20S 00008129
15	C/S IMODIUM AD CAPLET 8 CT 00008568
16	C/S HALLS COUGHDROP BAG CHERRY 00008604
18	PEDIACARE CG4-CD LD LIQ 4 OZ 00011940
20	C/S ROBITUSSIN DM SYRUP 4 oz. 00013442
36	NEX MAXSTR N/A SINUS 24 CAPS 00023550
37	NEX NITE TIME COLD MEDCN 10 OZ 00023580
38	NEX LONG LAST NASAL SPRAY 1 OZ 00024430
39	NEX TUSSIN DM COUGH SYRUP 8 OZ 00024546
40	NEX NITE TIME COLD CHERRY 10OZ 00024566
41	NEX SUPHEDRIN DECONG TABS 24'S 00024619
42	NEX APHEDRID TABS 24'S 00024668
43	NEX MENTHOL EUCALYP DROPS 40'S 00041400
44	NEX HONEY/LEM COUGH DROPS 40'S 00041420

12

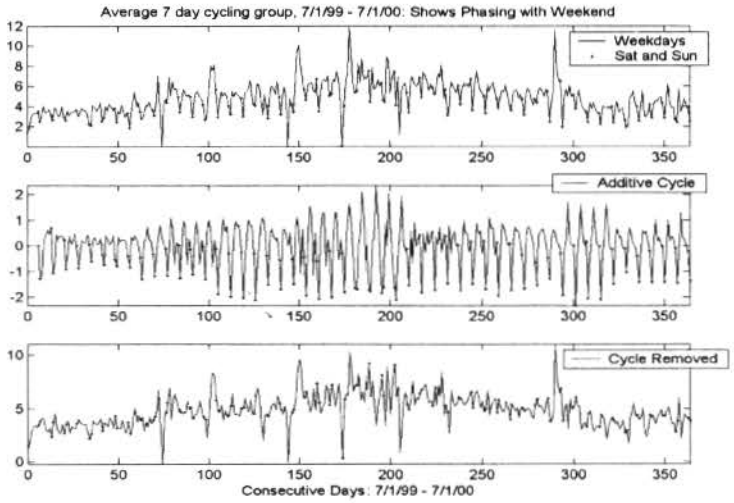
Method 3: Principal Component Analysis (PCA)

- Diagonalization of covariance matrix for multiple time series
- Yields uncorrelated linear combinations of the raw data and associated variances (the eigenvalues of the covariance matrix).
- If a small number of linear combinations dominate the overall variance in the data, these are taken to be the new data with reduced dimensionality.

13

PCA of Month-Scale Variations

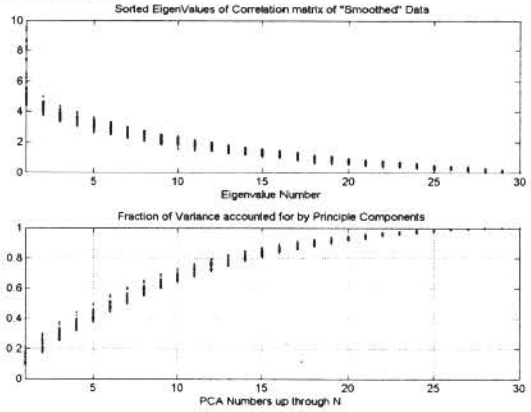
Step 1: Remove weekly variation (where applicable)



14

PCA of Month-Scale Variations (continued)

Step 2: Find eigenvalues of (singular) correlation matrix of consecutive 30 day data segments

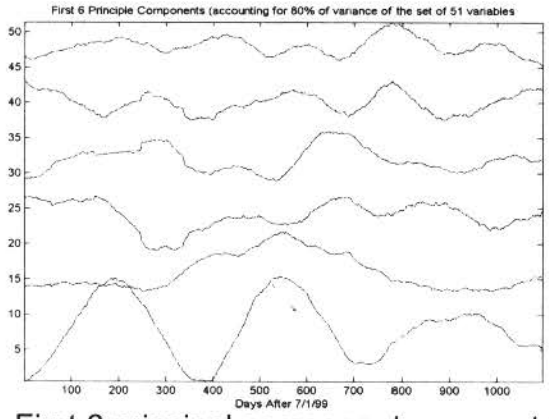


PCA analysis of does not reveal any dominant eigenvalues
No groupings are evident for month-scale phenomenon

15

PCA analysis on longer scales again shows seasonal grouping

PCA analysis of 3 years of 51 dimensional OTC data with variation on timescales < 90 days removed



PC 6
PC 5
PC 4
PC 3
PC 2
PC 1: Winter cold remedies

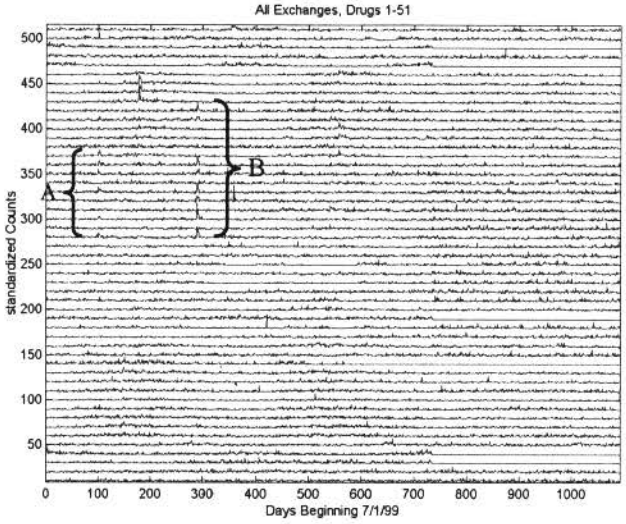
First 6 principal components account for 90% of the variation in the data.

16

Method 4: Aggregation according to common response on short timescale

Standardization to unit variance: highlights local *transients*

Brackets show examples of transients common to multiple series



Aggregation according to common response on short timescale (continued)

OTC Category:	Transient	
	A	B
1) Pain/Fever Only	2	3
2) #1+ Flu SS	0	1
3) Cough Drop/Throat Only	0	0
4) #3+ Flu SS	0	0
5) Antihistamine/Sinus/Decong.	1	2
6) Stomach/GI	2	2
7) Skin/Rash/Itch	0	0
8) Child	0	0

Transients related to naval deployments?

18

Conclusions

- Specific statistical methods can identify drugs with and without particular temporal sales patterns
  - Power spectral analysis can identify 7-day trends
  - Hierarchical aggregation algorithms can identify seasonal trends
- Certain temporal trends were significant, while others were not
  - Significant: 7-day, seasonal
  - No significance: Monthly
  - Potentially significant: Short-term spikes

19

Future Direction

- Group OTC drugs using expert knowledge
- Analyze temporal patterns within and between groups using the methods described here.

20

10906

Are Knowledge-Based Approaches E x t e n s i b l e?

Mark A. Musen
Stanford University

1

How extensible are software architectures that support

- Incorporation of alternative, modular, analytic problem solvers
- Incorporation of control knowledge for deploying problem solvers using an explicit knowledge base
- Potential for parallellization of problem solvers
- Access to arbitrary, real-world data sources via a mediating ontology

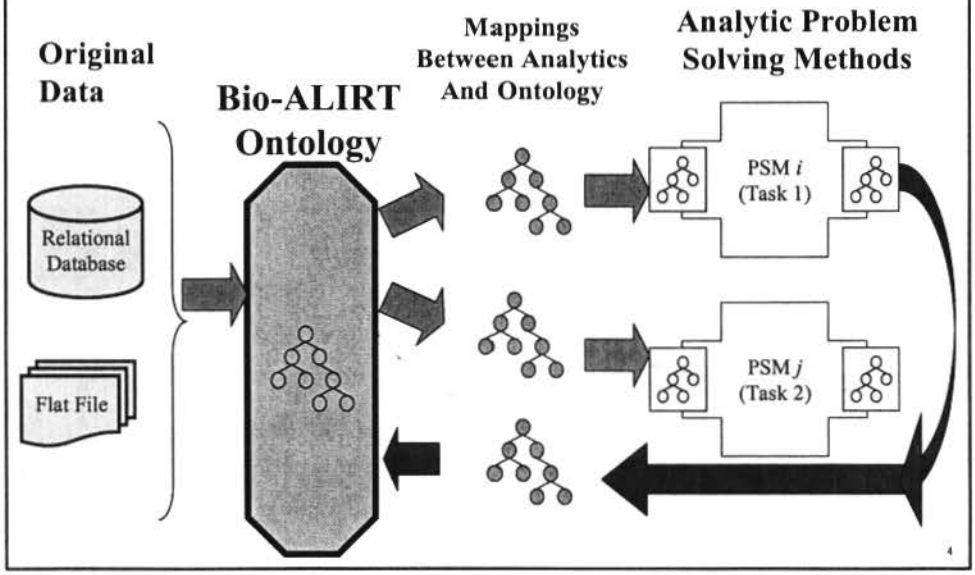
2

Modular Problem Solving Architecture

- Allows intergration of both knowledge-based and statistical problem solvers
- Facilitates application of best available method to the data
  - We can declare in a knowledge base (or determine dynamically) which methods should be used for which data sources (e.g., EWMA vs. CUSUM)
- Enables controlled evaluation of different methods for a specific task in a larger architecture
  - We can identify influence of forecasting methods for individual data streams on overall system performance
- Takes advantage of common data interfaces, minimizing software development

3

Modular Problem Solving



4

How extensible are software architectures that support

- Incorporation of alternative, modular, analytic problem solvers
- **Incorporation of control knowledge for deploying problem solvers using an explicit knowledge base**
- Potential for parallellization of problem solvers
- Access to arbitrary, real-world data sources via a mediating ontology

5

Bio-STORM provides a control structure for modular problem solvers

- Integrates domain knowledge, data streams, and analytic problem solvers into a coherent run-time system
- Provides dynamic control structure to:
  - Add and remove analytic problem solvers and data streams
  - Reconfigure data flow and problem-solver invocation

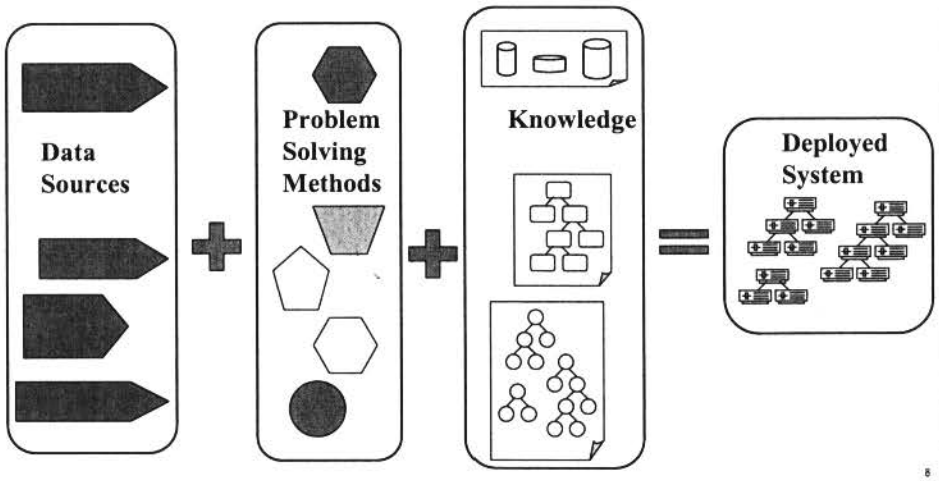
6

Extensibility in choice and deployment of problem solvers

- Each problem solver is designed to access a common ontology describing properties of the data
- An explicit knowledge base stores
  - The capabilities of each problem solver
  - The kinds of data on which the problem solver might operate
  - The kinds of data that each problem solver may output
- The result is a flexible, extensible architecture

7

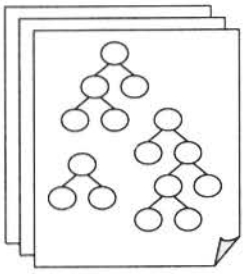
Modular data sources, problem solving methods (PSMs), and KBs



8

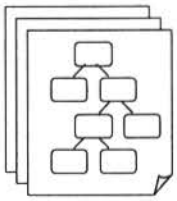
Current BioSTORM knowledge bases

PSM Knowledge Bases



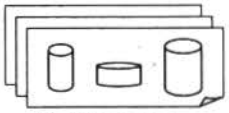
- Spatio-Temporal Abstraction
- Kalman Filtering
- Fusion
- Normalization

Domain Knowledge Bases



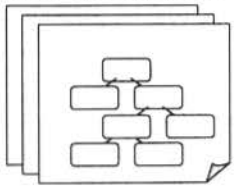
- Syndrome Definitions
- Population Modeling
- Disease Modeling

Data Source Knowledge Bases



- 911 Data Sources
- Clinical Data Sources
- Absenteeism

Control Knowledge Bases



- Data Sources Interfaces
- PSM Interfaces
- PSM Configuration

9

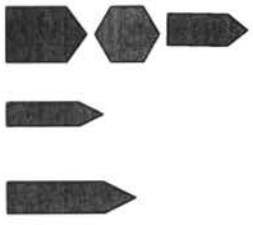
Knowledge-Based Run-Time Deployment

Control Knowledge Bases

Data Source Knowledge Bases

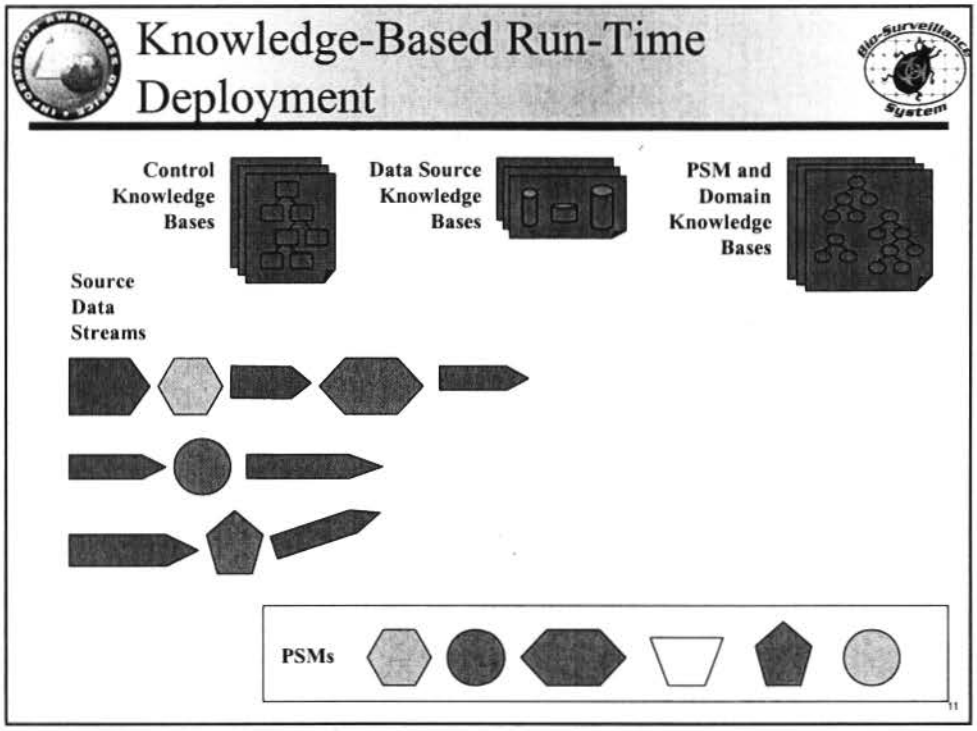
PSM and Domain Knowledge Bases

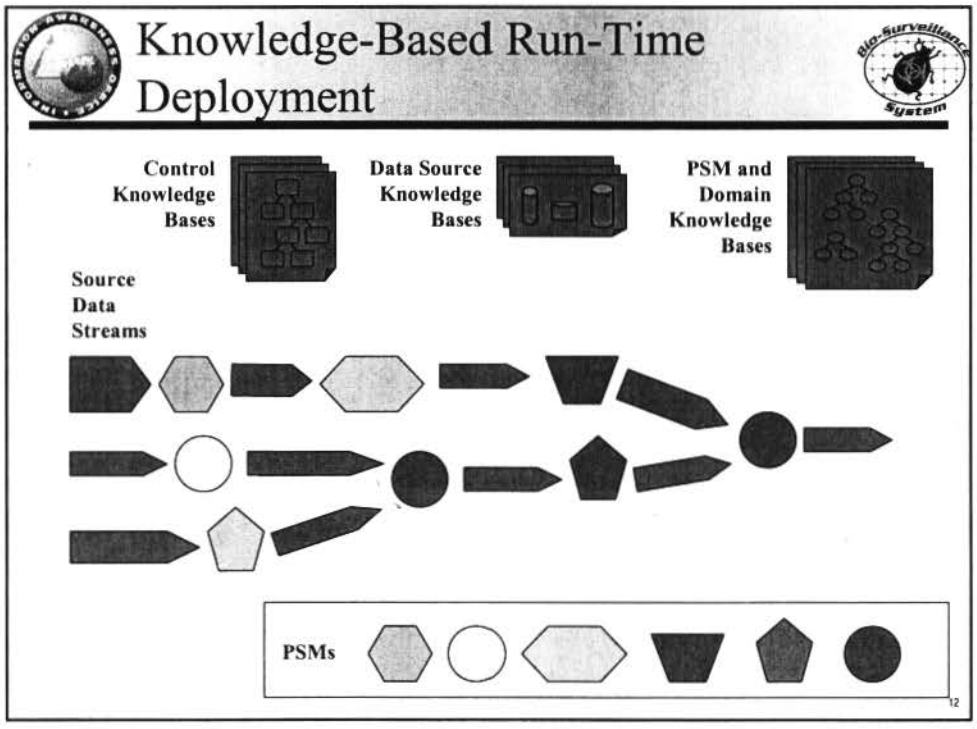
Source Data Streams



PSMs

10


Knowledge-Based Run-Time Deployment
Control Knowledge Bases
Data Source Knowledge Bases
PSM and Domain Knowledge Bases
Source Data Streams
PSMs
11

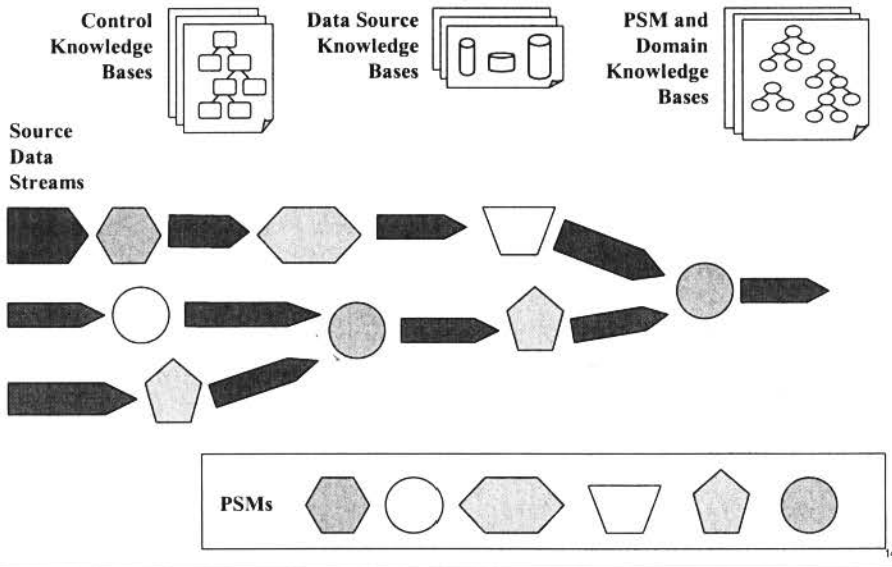

Knowledge-Based Run-Time Deployment
Control Knowledge Bases
Data Source Knowledge Bases
PSM and Domain Knowledge Bases
Source Data Streams
PSMs
12

How extensible are software architectures that support

- Incorporation of alternative, modular, analytic problem solvers
- Incorporation of control knowledge for deploying problem solvers using an explicit knowledge base
- **Potential for parallellization of problem solvers**
- Access to arbitrary, real-world data sources via a mediating ontology

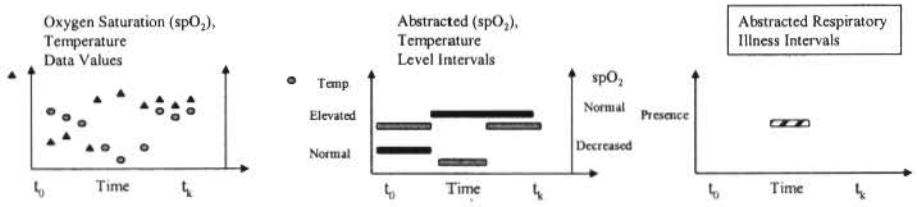
13

Each problem solver contributes a "chunk" to the overall solution



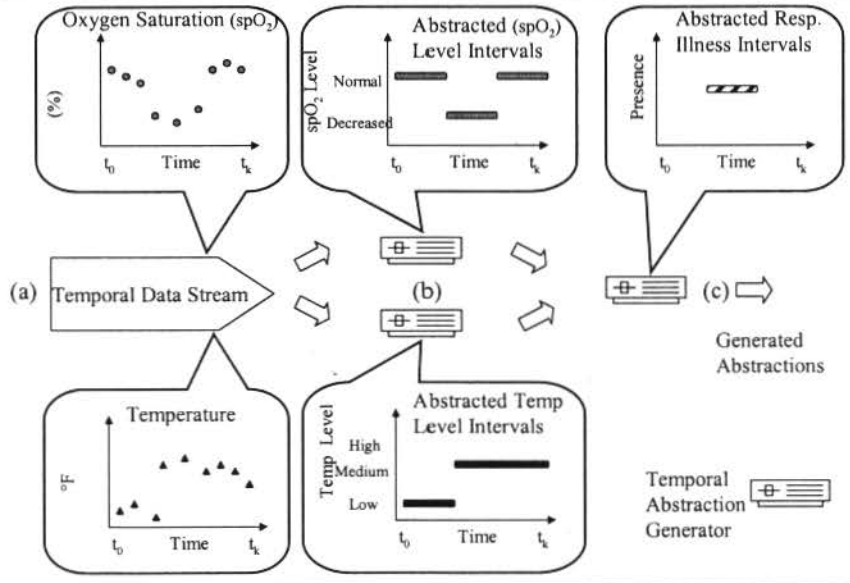
14

Example: Parallelized Respiratory Syndrome Abstraction



15

Example: Parallelized Respiratory Syndrome Abstraction



16

Knowledge-Based Parallel Processing Architecture for PSMs

- Each abstraction processor can be responsible for a single abstraction
- Other configurations possible, however:
  - Single processor per abstraction type
  - Different processors for different parts of the timeline
  - Single processor per spatial region
  - Single processor per spatial region, per timeline portion

17

Real-time surveillance of thousands of data streams requires will require

- Supercomputer clusters
- Grid computing
- "Bio-ALIRT at home"

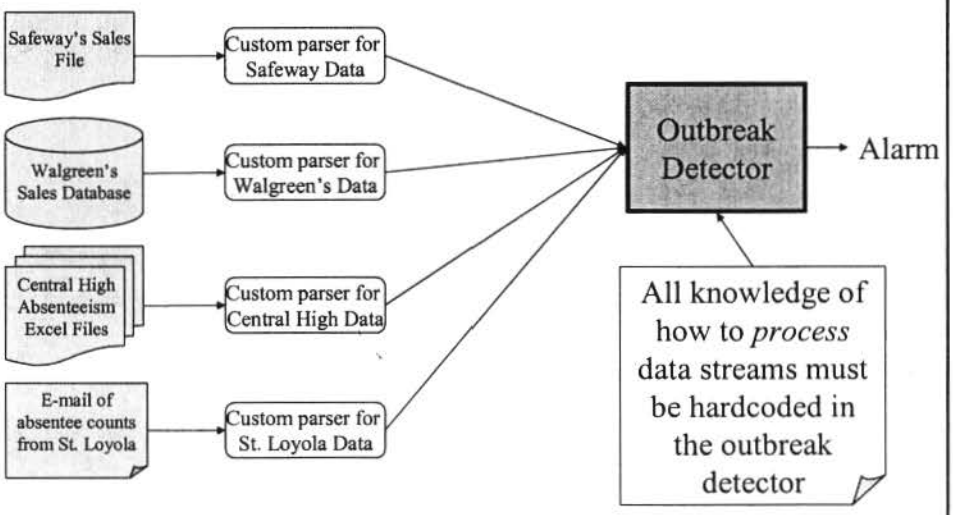
18

How extensible are software architectures that support

- Incorporation of alternative, modular, analytic problem solvers
- Incorporation of control knowledge for deploying problem solvers using an explicit knowledge base
- Potential for parallellization of problem solvers
- **Access to arbitrary, real-world data sources via a mediating ontology?**

19

Without a mediating ontology ....



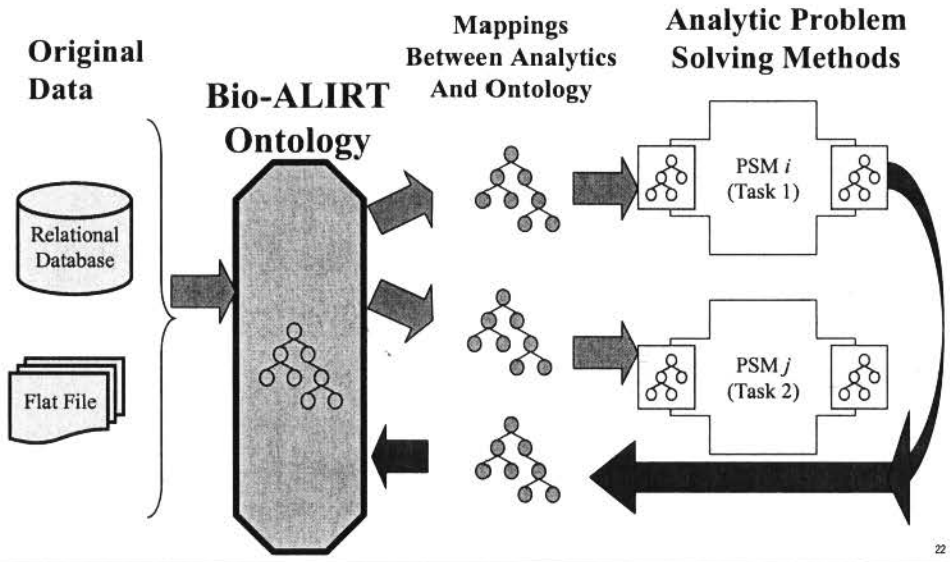
20

Without explicit knowledge

- Adding new data sources is time-consuming
  - Developers need to write a parser to deal with each new data source
  - Outbreak detector code needs to be changed to deal with each each new data source
- Specific data analysis methods are written for specific data
  - No easy way to test *all analysis methods* on a given set of data
  - No easy way to test *all data* on one method

21

Data Broker Approach



22

Key Points of the BioSTORM Architecture

- Extensible mediating ontology for providing computational access to data sources
- Flexible knowledge-driven problem-solving architecture
- Scalable problem solving approach

23

How extensible are software architectures that support

- Incorporation of alternative, modular, analytic problem solvers
- Incorporation of control knowledge for deploying problem solvers using an explicit knowledge base
- Potential for parallellization of problem solvers
- Access to arbitrary, real-world data sources via a mediating ontology

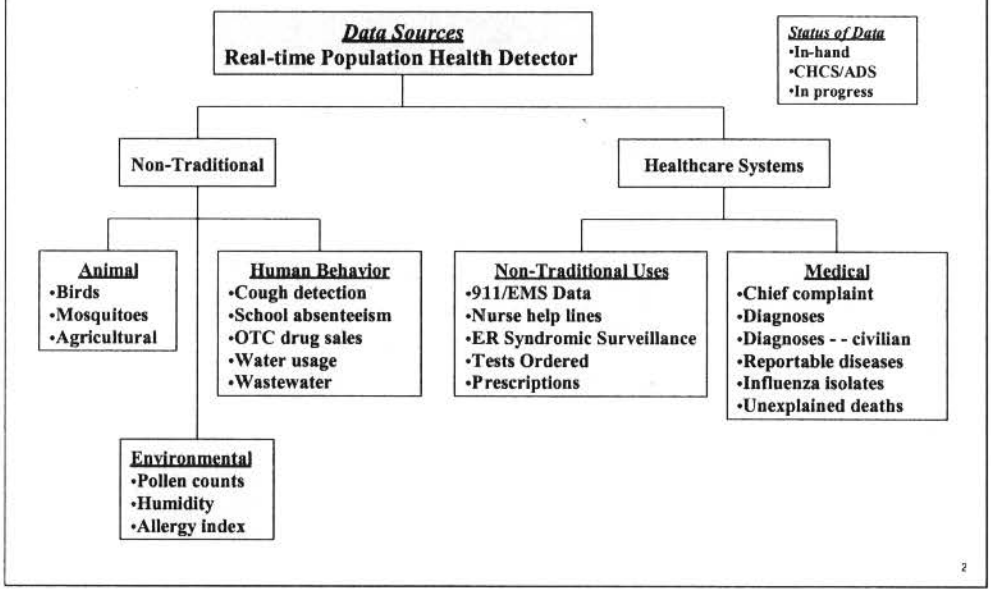
24

NT Data

Outline

- Overview
- 911 Call Data
- Wastewater Data
- Cough Detection
- School Absenteeism
- OTC Data

What Are the Non-traditional Data Sources?



NT Data

Outline

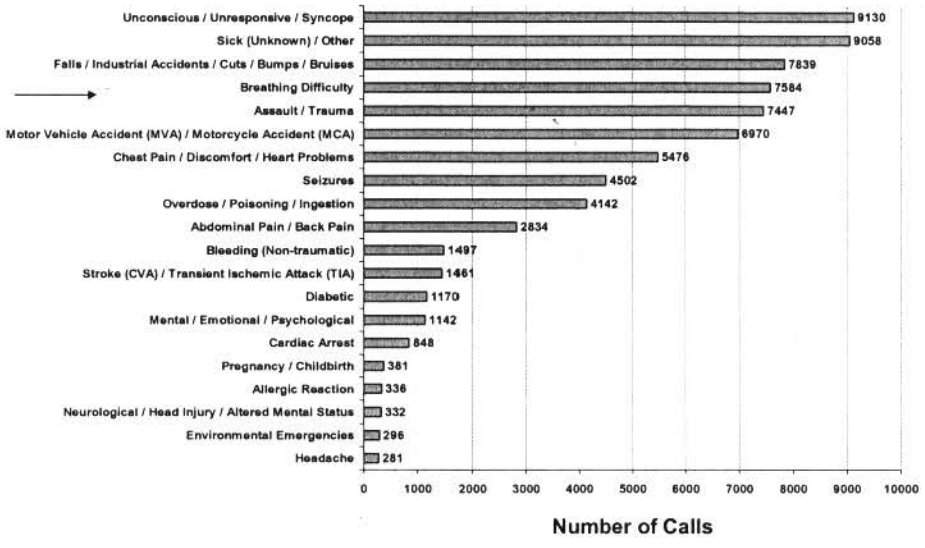
- Overview
- 911 Call Data
- Wastewater Data
- Cough Detection
- School Absenteeism
- OTC Data

3

Overview of 911 Data

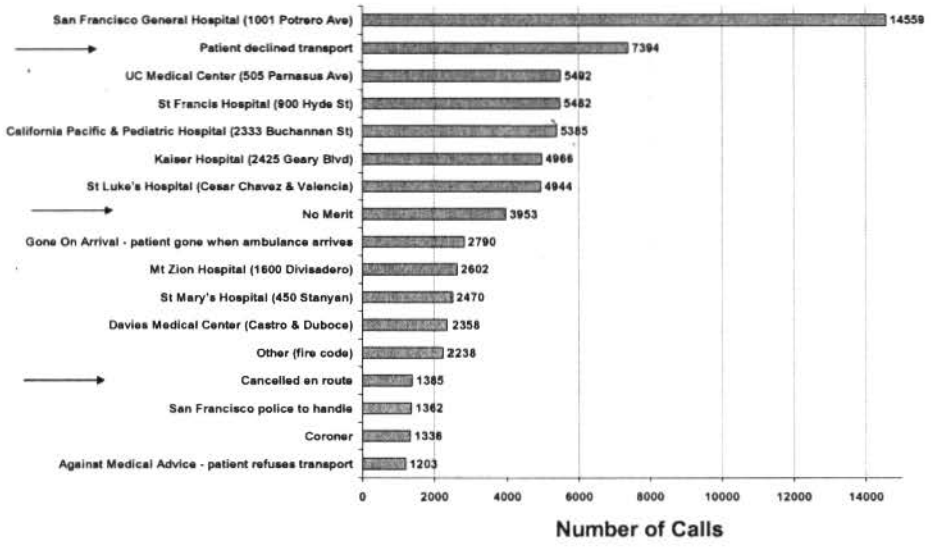
- All calls in the city use a common dispatch center
- We have records for 1999 and are in the process of obtaining data for subsequent years
  - Potentially available in real-time, possible lag (hours) for paramedic to verify or enter codes.
- Each record is for a call to 911
  - We focus on the subset requiring an EMS response
- Fields for each call include
  - **Date** (currently to the day, time may be available)
  - **Location** (street address – geo-coded to at least Block Group)
  - **Urgency** (3 levels)
  - **Reason for call** (two part code: 26 main codes, and ~10 secondary code for each main code)
  - **Disposition** (hospital patient taken to, cancelled call, ...)

Annual Number of 911 Calls by Main Call Code



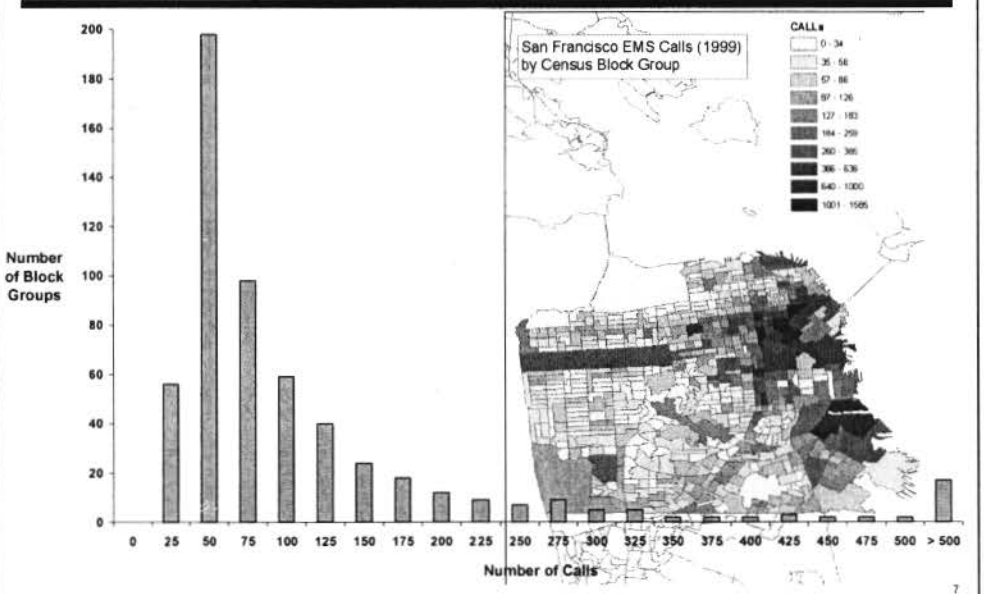
5

Annual Number of 911 Calls by Disposition

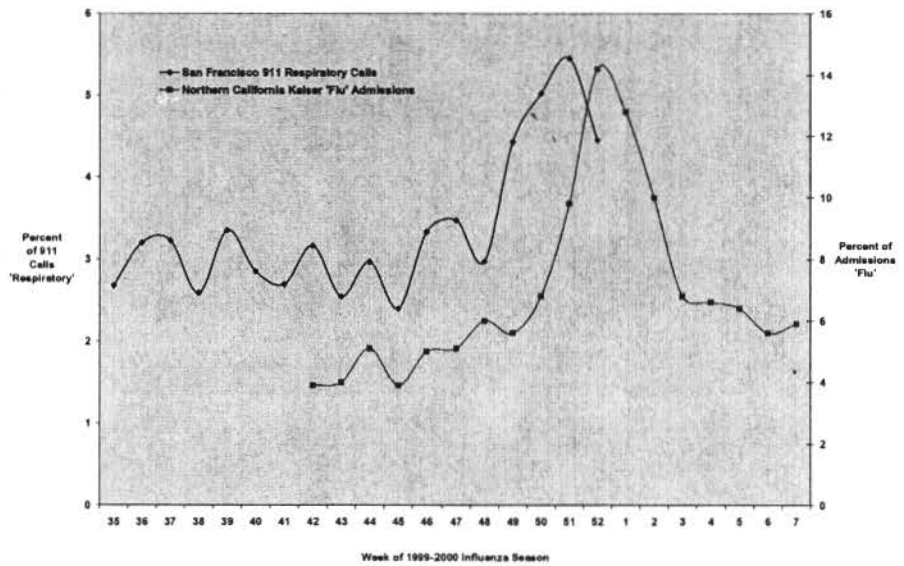


6

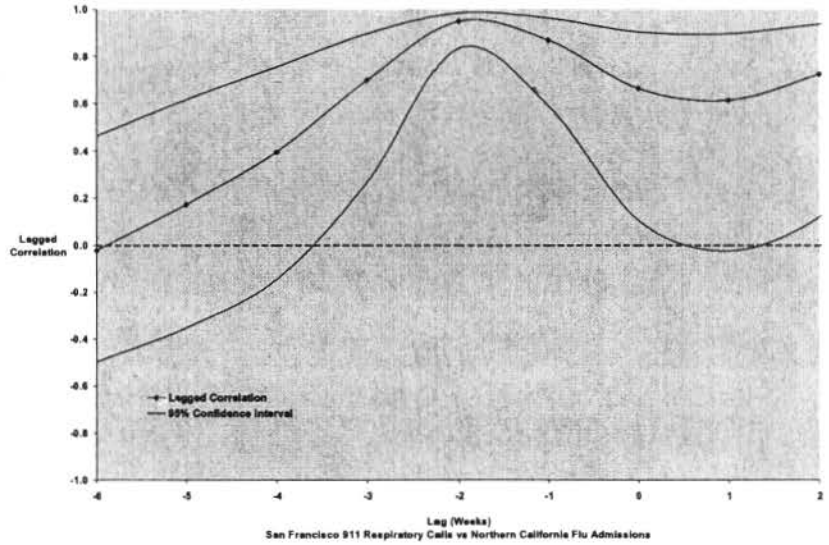
Spatial Distribution of Annual 911 Calls by Census Block Group



Respiratory 911 Calls and 'Flu' Admissions



Lagged Correlation Between 911 Calls and 'Flu' Admissions



9

Observations on Relationship Between 911 Calls and 'Flu' Admissions

- Peak correlation between series occurs when 911 calls are moved forward ~2 weeks relative to admissions
- Epidemiologically plausible lead time if
  - Respiratory 911 calls may not result in transport or immediate admission
  - Later stages of illness and public awareness of epidemic increase likelihood of hospital admission for "flu"
- Suggests 911 calls may be an early indicator of large scale respiratory illness in population
- Caveat – this is only one year of data!

10

Next Steps in Analysis

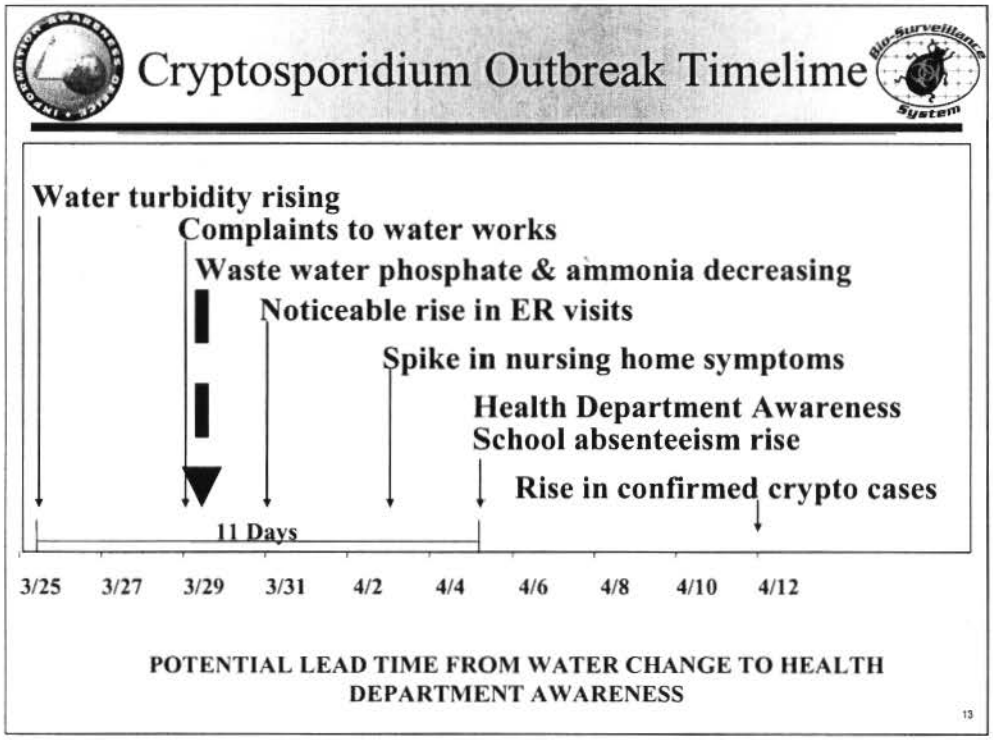
- Repeat analysis with additional years of data and daily time resolution
- Examine other groupings of call codes and urgency
- Analyze time series by spatial sub-regions
- Validate diagnostic codes with chart audit

NT Data

Outline

- Overview
- 911 Call Data
- Wastewater Data
  - Test Case -- Metropolitan Milwaukee Sewerage District Data during 1993 Cryptosporidium Outbreak (~400,000 people infected)
- Cough Detection
- School Absenteeism
- OTC Data

12



March 25-27th – Water turbidity on rise waste water phosphate & ammonia decreasing

March 29th - Complaints to water works

March 31st - Noticeable rise in ER visits

April 3rd – Spike in nursing home symptoms

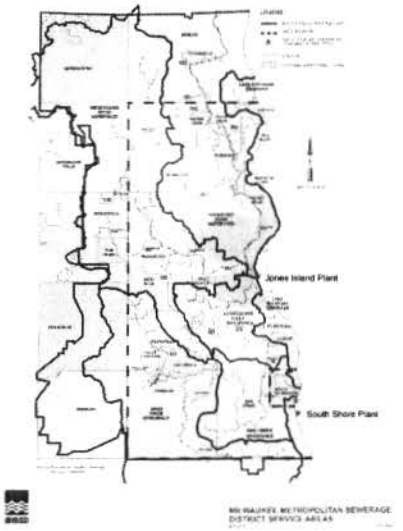
April 5th - School absenteeism rise

April 5th - Health Department Awareness

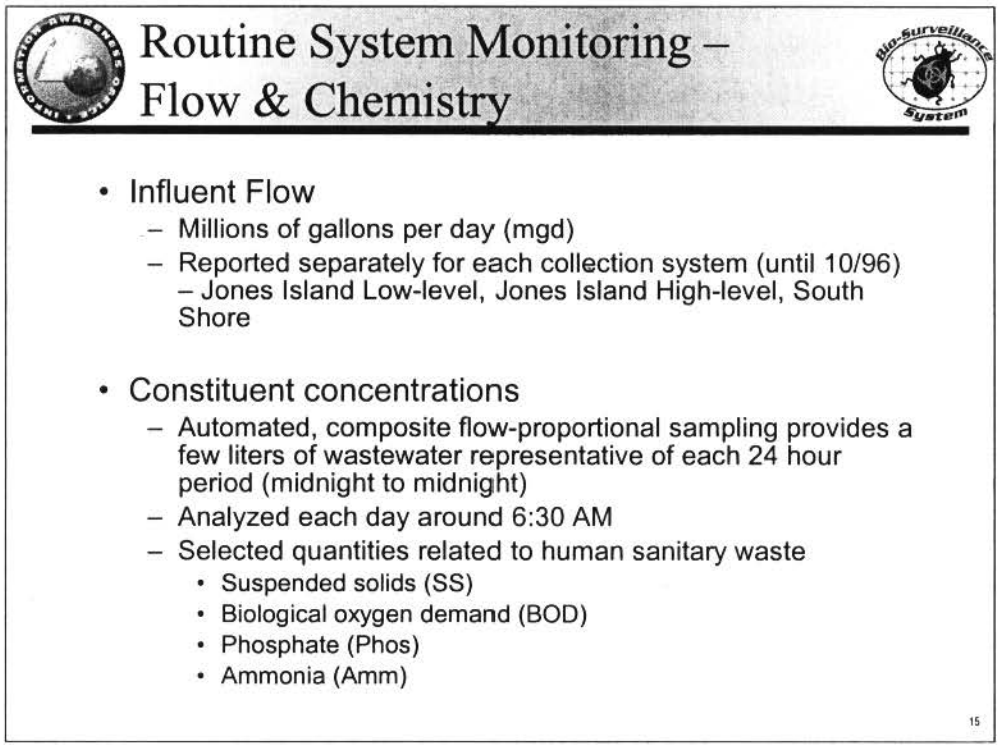
April 12th Rise in confirmed crypto cases

Metropolitan Milwaukee Sewerage District (MMSD)

- **Two Waste Water Treatment Plants**
  - Older facility is Jones Island
    - historic port, central location
  - Newer facility is South Shore
- **Jones Island**
  - "Low-level" collection system
    - urban core (subsystem 7)
  - "High-level" collection system
    - urban area (subsystems 5 & 6)
- **South Shore**
  - Serves additional parts of Milwaukee and outlying municipalities (Subsystems 1-4)
  - Diversion from Jones Island conveyance lines



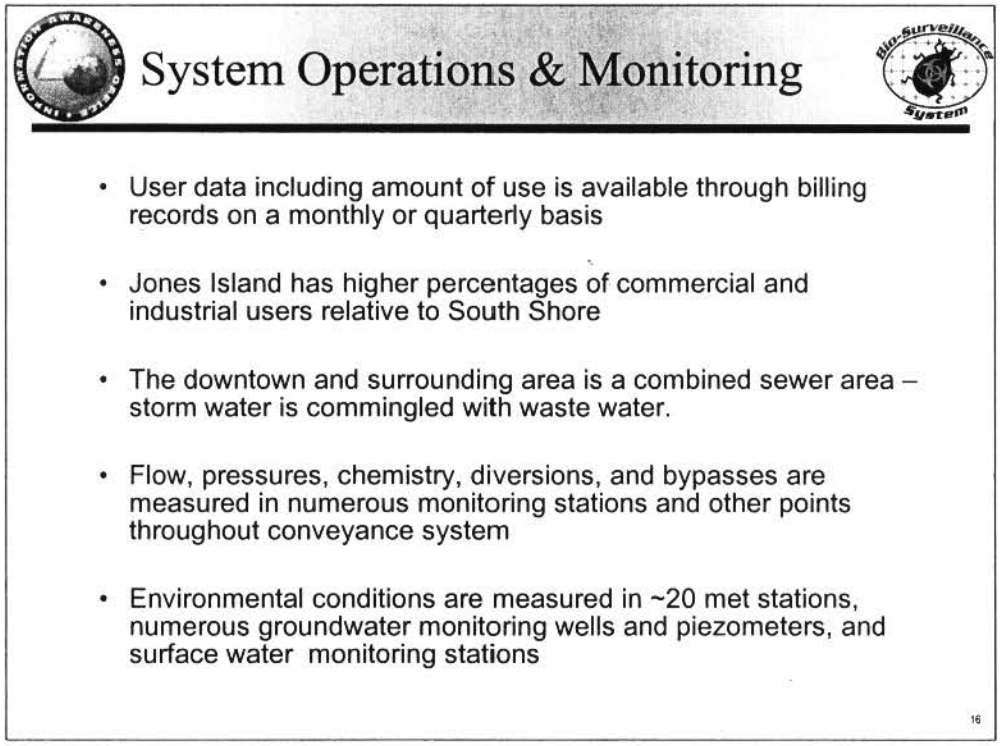
MMSD treats sewer for all Milwaukee except for South Milwaukee.



Influent flow before any filtering or treatment or compositing from different flow lines.

About 10 parameters usually analyzed.

Low-level, high-level, & ISS are all into Jones Island Plant. Started operating ISS about 4-5 months after outbreak but it took about 8 years to design & build. From 1/1/90 to 10/11/96 separate analyses were performed on influent from low-level & high-level. To cut costs, after 10/11/96 the high-level, low-level and ISS samples were composited and only one set of analyses run for Jones Island.



This slide shows the complexity of the system & the abundance of data sources to evaluate this complexity.

Larger users have monthly use measured; smaller users have quarterly use measured or estimated. Billing address and sewer service area are supplied for 1330 clients. Also some users have BOD and TSS measurements at inlet.

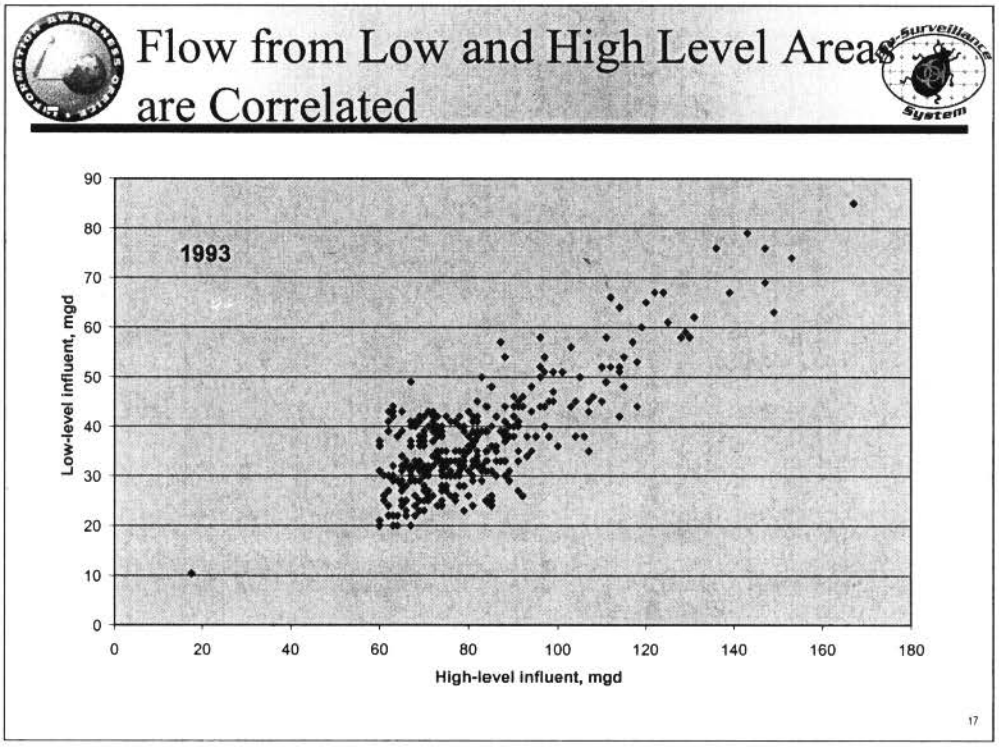
Jones Island serves the more industrial area around downtown.

Other points in conveyance system are classified as diversion chambers, approaches, junction chambers, open channels, flumes, tunnel, bypass, river measuring points, pumping stations, cutbacks, normal pipe in system and others.

Bypasses are direct releases of raw or partially treated sewage to rivers. Diversion is generally from conveyances that go to South Shore rerouted to Jones Island.

Wells & piezometers are used to measure the water table elevations. Samples for chemistry are also collected from wells. Flow rate or river stage (height of water) and samples are taken at surface water stations. I believe these stations are on Milwaukee, Menominee, Kinnickinnic Rivers, Lake Michigan, and some creeks.

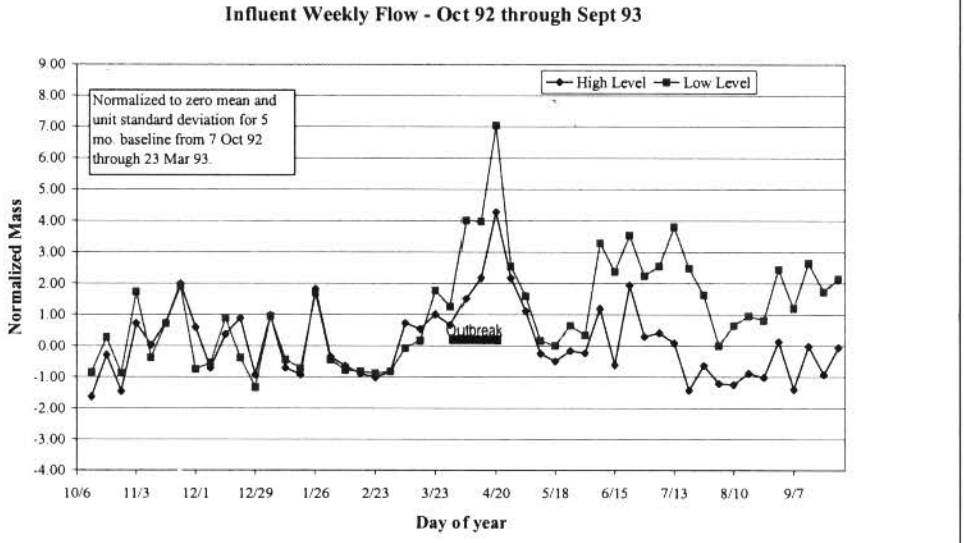
Met stations measure mostly precipitation possibly on hourly basis, maybe continuously. Report this to city and state agency, Dept. of Natural Resources.



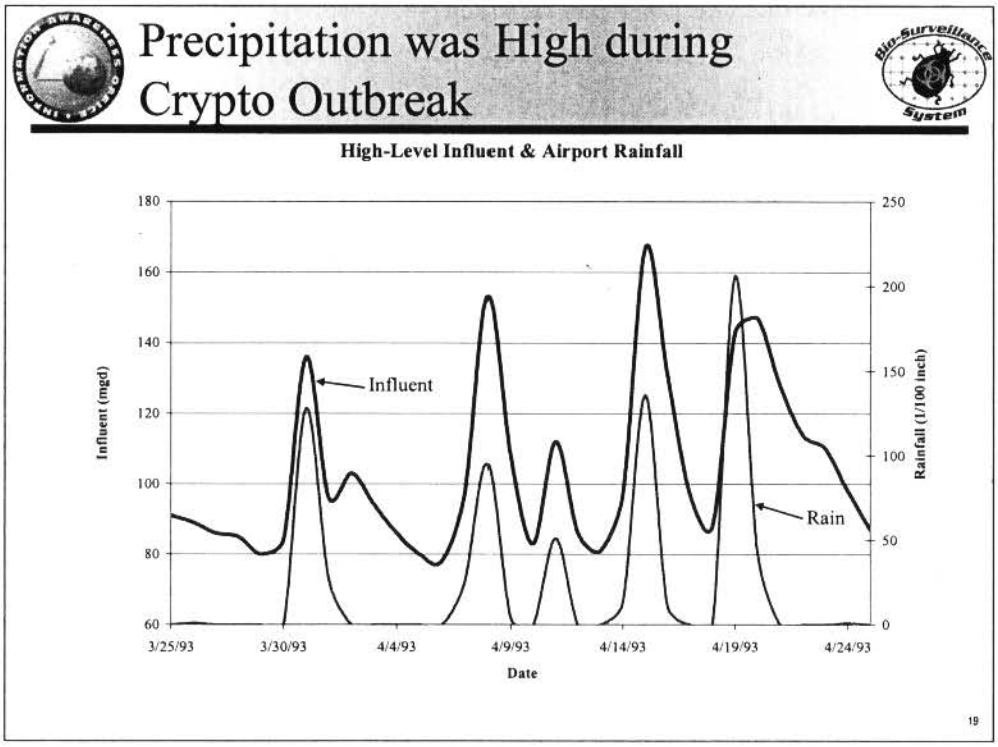
We started with Jones Island data because it was consistent through this period & flow rate data were available electronically. Flow data were not available electronically for South Shore.

Low-flow rate, mgd is really flow rate for low-level influent. Same for high.

Flow Increased During Outbreak



Initial look indicates that flow is 1.5 to 7 SDs (low-level system) and 0.8 to 4+ (high-level-system) higher than the mean for the 4 weeks during outbreak.

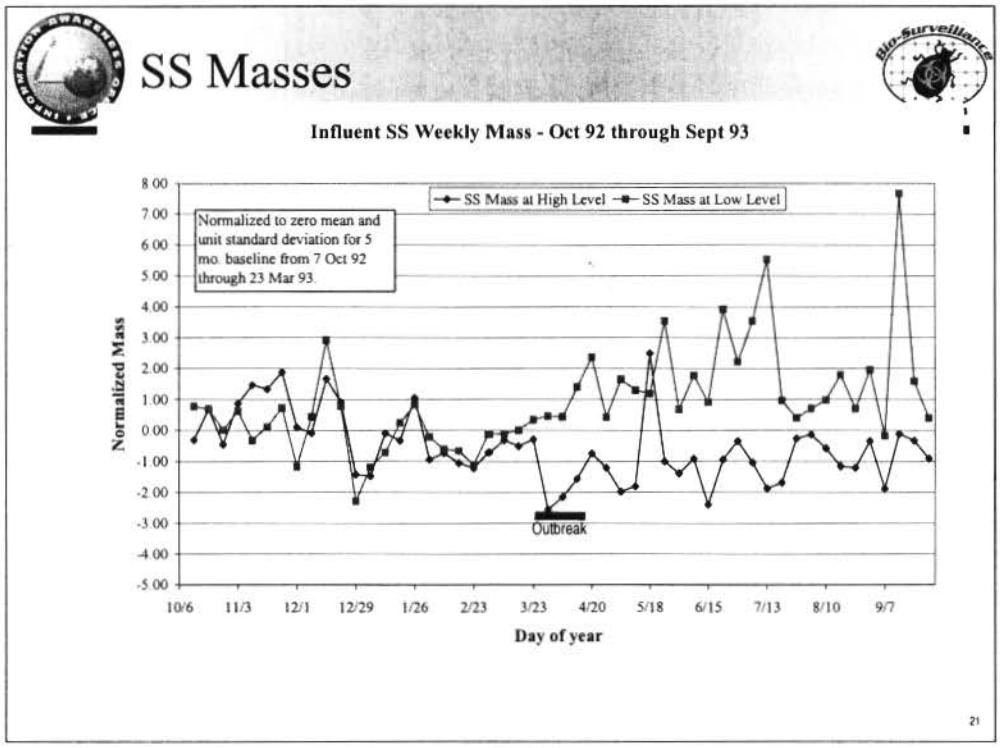


Example of major confounding factor is precipitation. Shows clear affect of rainfall on influent to the Jones island high-level feed.

Examine Load or "Mass" Associated With Measured Concentrations

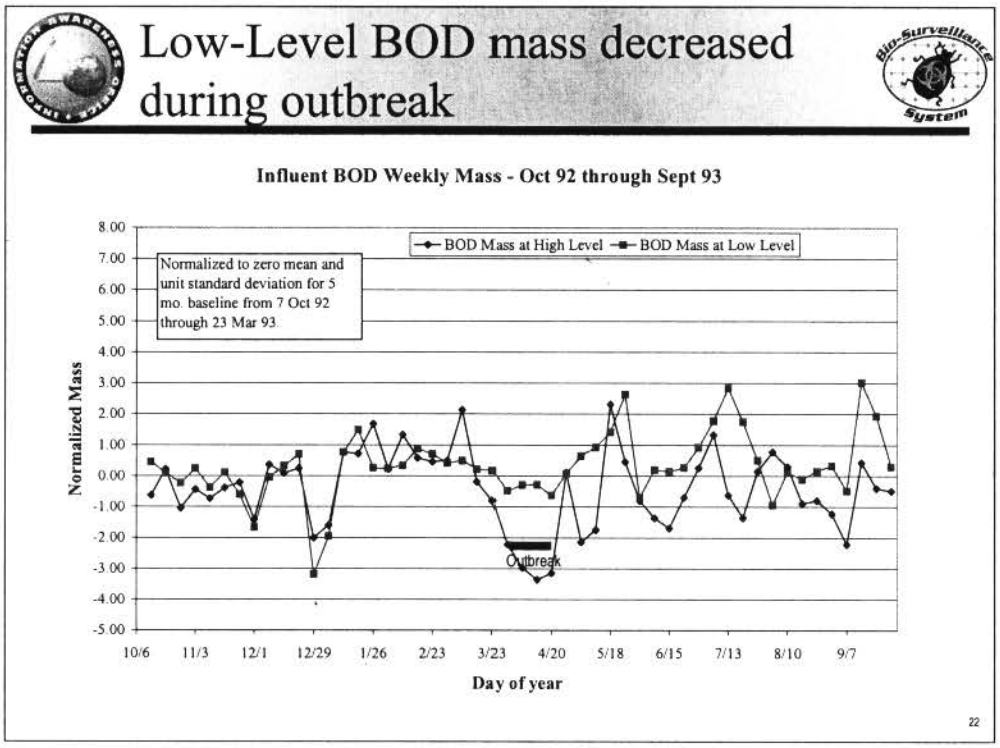
- Multiply flow rate times measured concentration for:
  - Suspended solids (SS)
  - Biological oxygen demand (BOD)
  - Phosphate (Phos)
  - Ammonia (Amm)
- Separate calculations for high level and low level influent

20



Suspended Solids: High-Level mass is .8 to 2.5 SDs lower, low-level .4 to 2.4 higher during outbreak

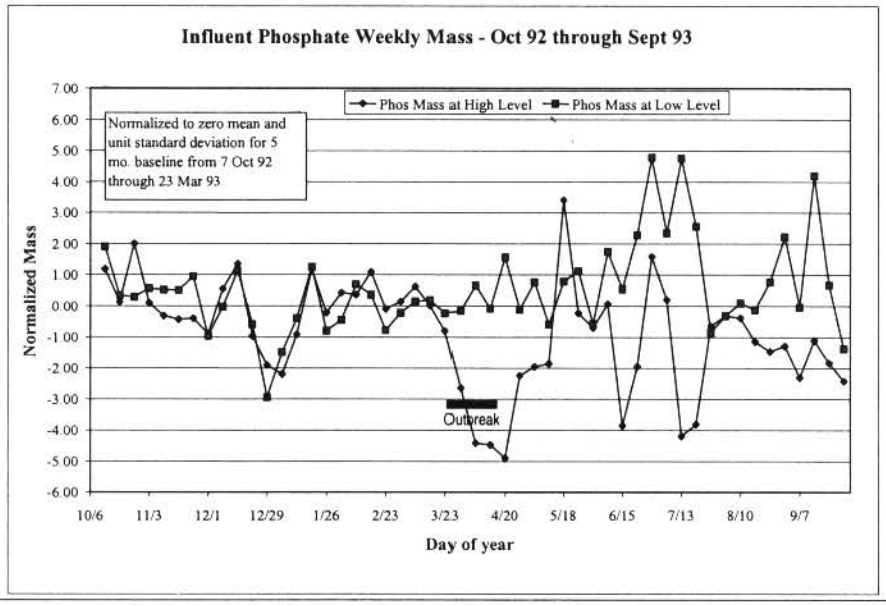
Avg. suspended solid concentration * avg flow for that week.



Biological Oxygen Demand: High-Level mass is 2 to 3 SDs lower, low-level .2 to higher .8 lower during outbreak

More storm water in low-level than high-level.

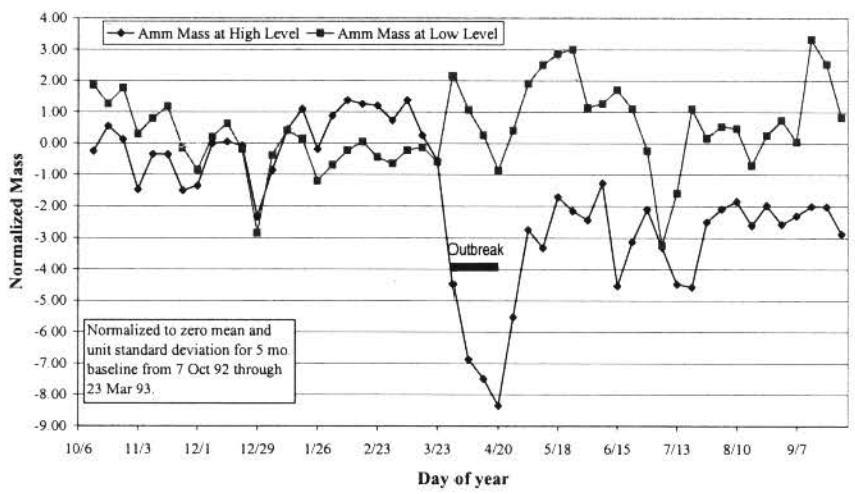
High-Level Phosphate Mass Decreased During Outbreak



Note: The analyte is phosphate, not phosphorus

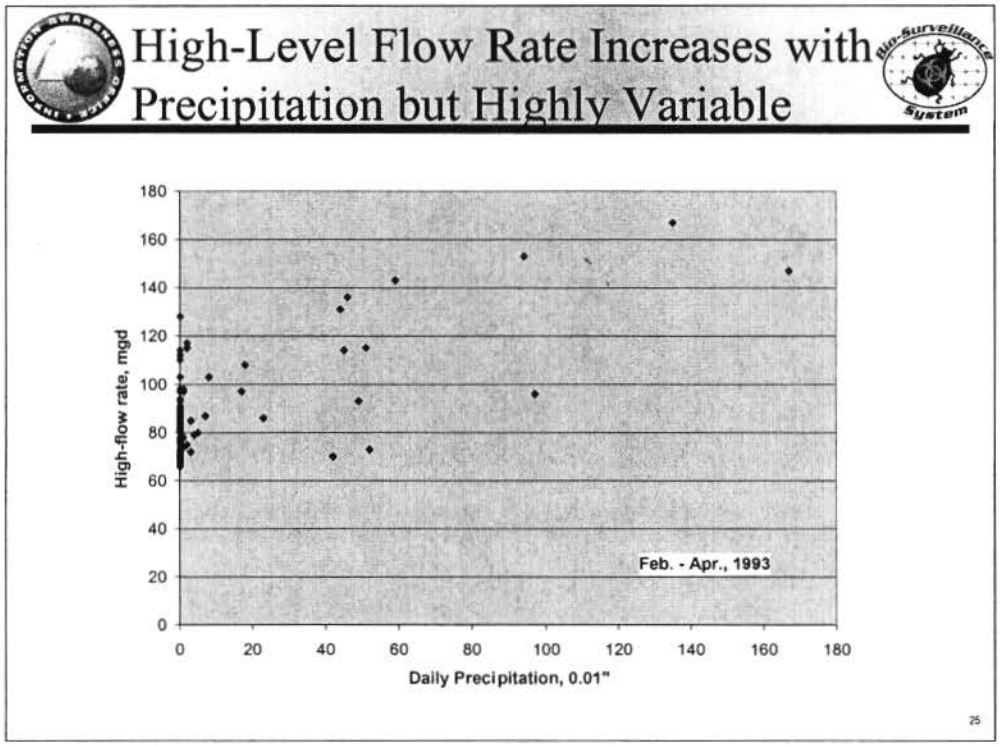
Phosphate: High-Level mass is 2.2 to 5 SDs lower, low-level is slightly lower to 1.5 higher during outbreak

Low-Level Ammonia Mass Decreased During Outbreak

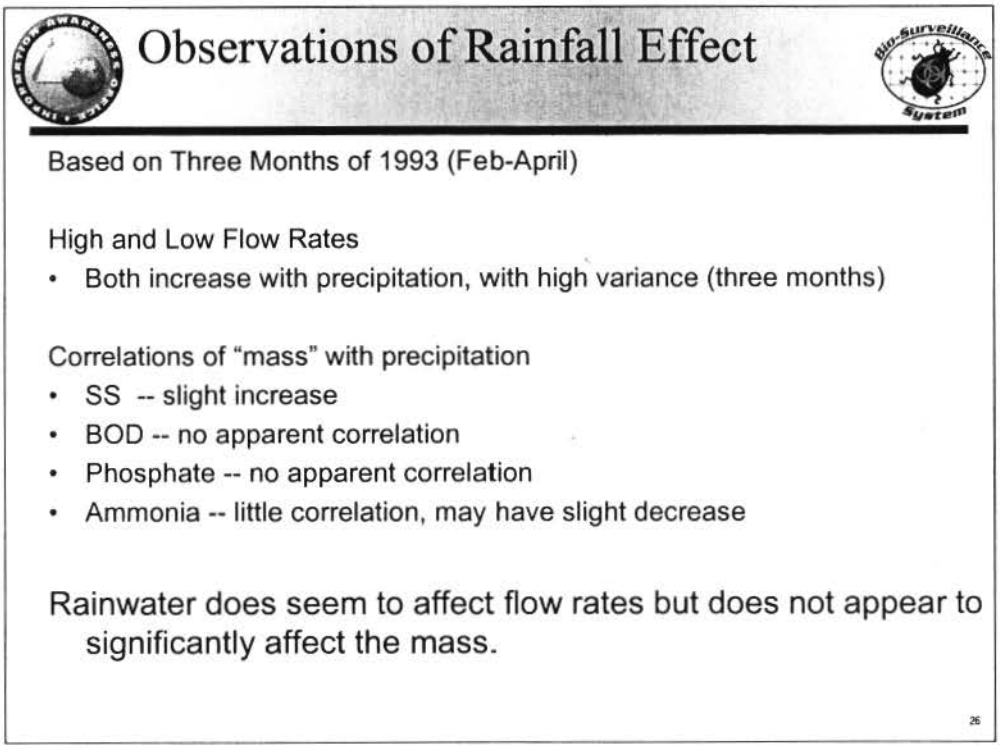


24

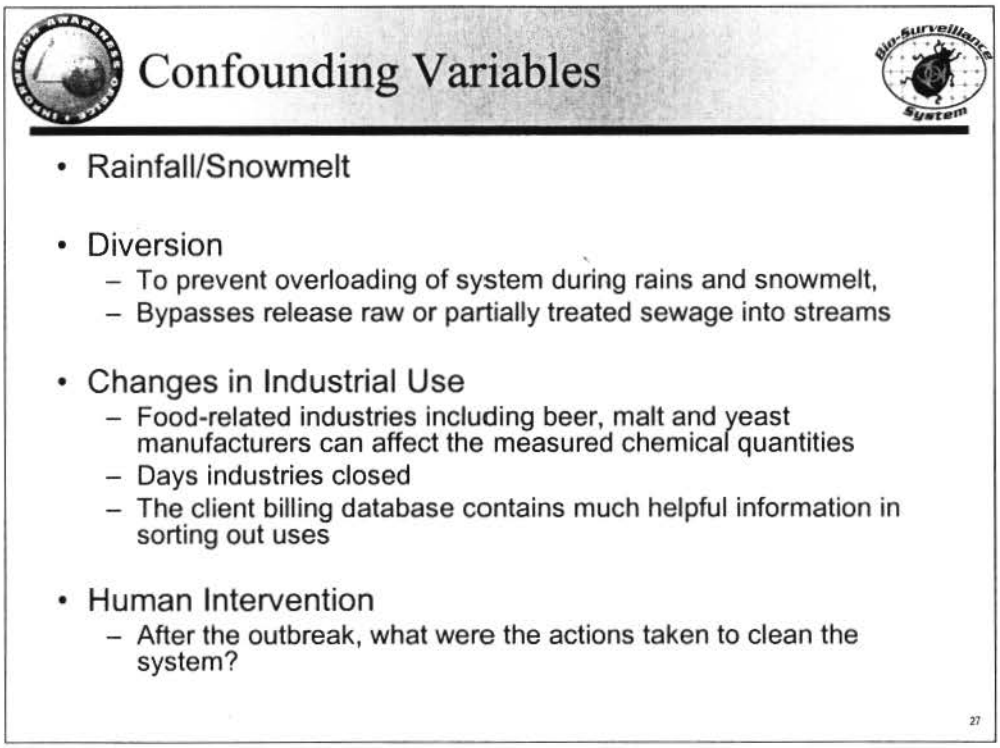
Ammonia: High-Level mass is 5.5 to more than 9 SDs lower, low-level 1 higher to 1 lower during outbreak



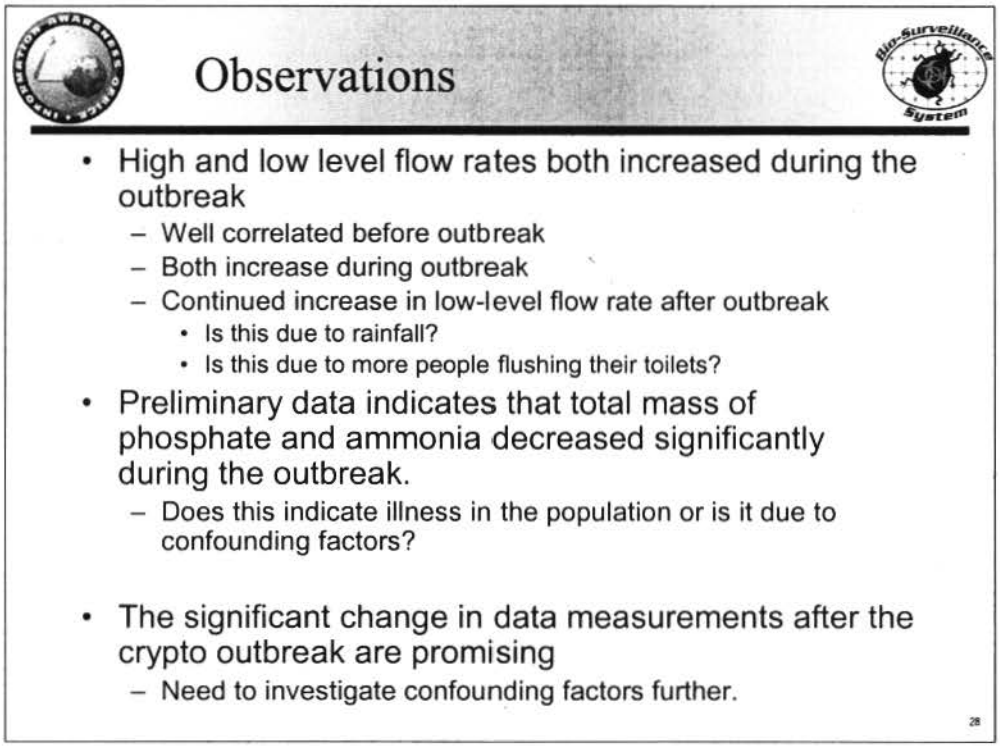
High precipitation affected flow rates but with some variability (not straight line fit). And shows that rainfall does not greatly affect total mass loadings of chemical parameters.



"Use of mass" because the mass is the flow rate x concentration rainfall should not have major affect on it.



Even now some bypasses into rivers or Lake Michigan occur but the frequency has been reduced greatly since the ISS has been used for overflow. The bypass frequency averaged about 40-60 year before the ISS and 2-10 since the ISS.

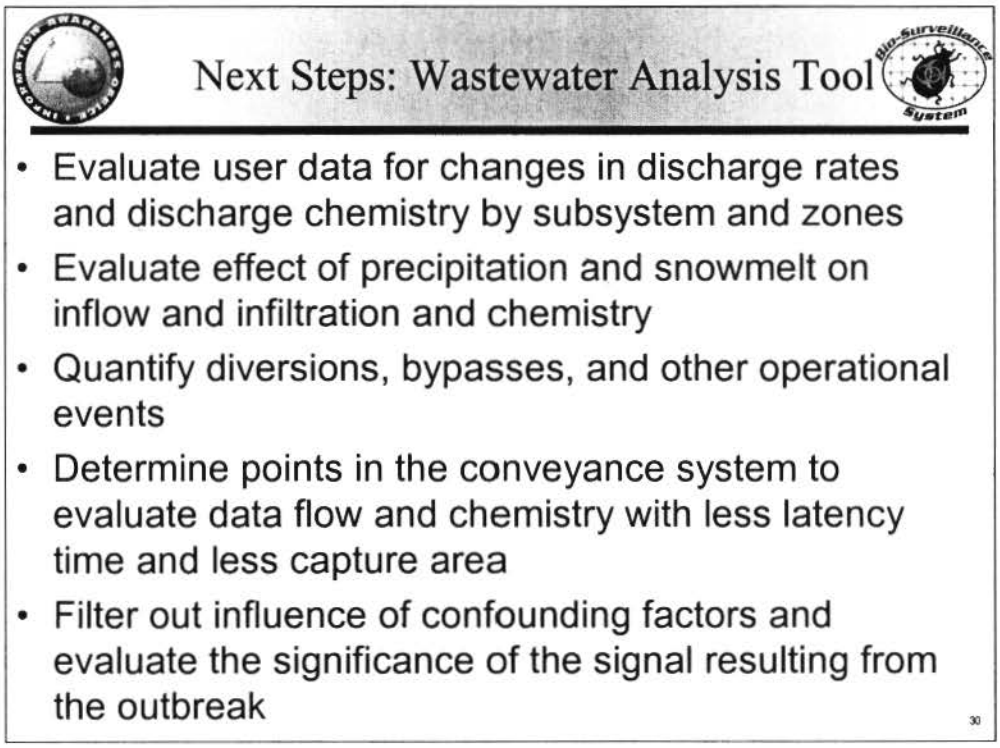


Proceeding to Objective 2 hinges on the answer to objective 1 being positive.

Next Steps

- **Model the system to determine if the decrease was due to outbreak or other factors.**
  - Determine if there is an increase of flow rate due to factors other than rainfall.
  - Attempt to localize rainfall during outbreak
  - Analyze diversion effects on data
  - Analyze the geographical distribution of the crypto outbreak
    - Were people in high-level areas mostly affected?
  - Research relationship of SS, BOD, phosphate and ammonia to human health.

29



Key factors are:

1)inputs to system (users and their changes)

2)Environmental – primarily precip & snowmelt but also includes streamflow, lake level, groundwater level, and may include temp, sunlight, & wind speed and direction

3) Operational – may include changes in how bypasses and diversions are determined, what & how data are collected thru conveyance system, changes due to environmental compliance and public policies.

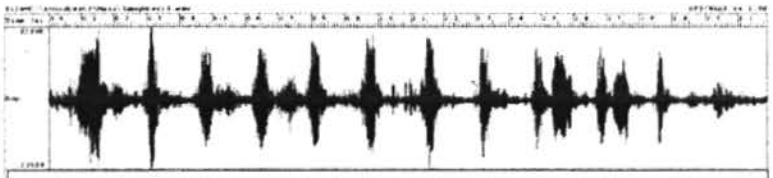
NT Data

Outline

- Overview
- 911 Call Data
- Wastewater Data
- Cough Detection
- School Absenteeism
- OTC Data

31

Overview of Cough Detection



- Software obtained from Audible Magic, Inc.
- Computes sound "vectors" from list of audio files that represent features to differentiate sounds.

32

Cough Detection Method

1. Break audio stream into 25-40 msec "frames".
2. Compute Fourier Transform for each frame.
3. Calculate feature set
   - "Loudness" -- Square root of signal energy from Fourier Transform.
   - Vector of Mel-Filtered Cepstral Coefficients (MFCCs) and derivatives.
4. Fit distribution of each feature and its derivative to a Gaussian mixture model.
5. Use parameters of Gaussian mixture model to calculate "distance" between audio file and class of interest.

33

Sound Effects Database

File Category	Number of Files	Comments
Applause	15	Generally 1-3 seconds long, includes a variety of different audience sizes.
Laughter	13	0.5-2.5 seconds long. Includes single person laughing and multiple persons laughing.
Cough (1)	51	Cough sounds from sound effects database, various lengths.
Cough (2)	35	Cough sounds from sound effects database, edited to eliminate extended coughing sounds, generally less than 1 second long.
Cough (3)	17	Cough sounds recorded at Veridian, edited to include one attack only
Cough (4)	7	Coughs recorded during a meeting, includes background sounds.
Door Slamming	15	From a sound effects database. Generally less than 1 second duration
Glass Breakng	14	From a sound effects database.
Phone Rings	17	From a sound effects database. Includes cell phone rings and conventional phone rings.

34

Sound Effects Classes Confusion Matrix

The sound effects database was used to create sound classes.

Results summarized in confusion matrix. The value *m,n* (row, column) is the probability that a file from category *m* is classified as class *n*.

	Cough (2)	Laughter	Applause	Door Slam	Phone	Glass
Cough (2)	0.5	0.5	0.0	0.0	0.0	0.0
Laughter	0.62	0.38	0.0	0.0	0.0	0.0
Applause	0.0	0.0	1.0	0.0	0.0	0.0
Door Slam	0.0	1.0	0.0	0.0	0.0	0.0
Phone	0.58	0.0	0.08	0.0	0.34	0.0
Glass	0.0	0.22	0.0	0.0	0.0	0.78

35

Cough Detection Follow-on

Work with Audible Magic

- resolve issues of mis-classification
- tune software
- add real-time processing.

Follow-on studies include:

- What is the ideal number of samples for training set?
- What is the optimum number of classes for classification?
- What are the effects of sampling rate on classification?

36

NT Data

Outline

- Overview
- 911 Call Data
- Wastewater Data
- Cough Detection
- School Absenteeism
- OTC Data

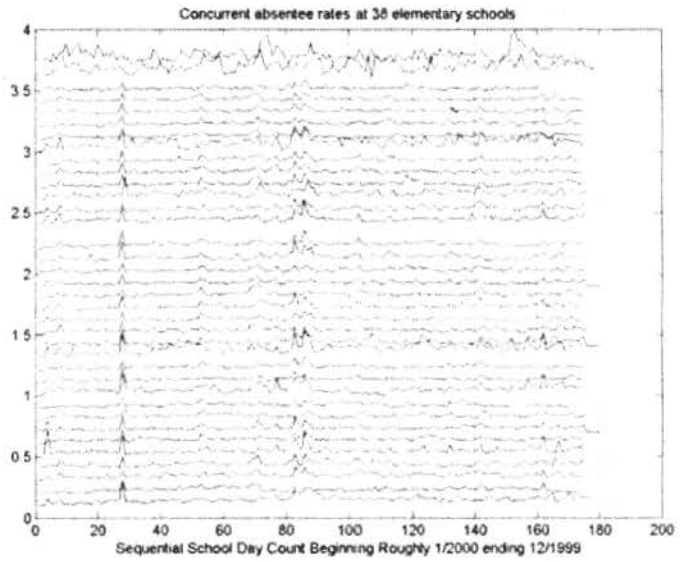
37

School Absenteeism

- 3 years historical data from Norfolk Public Schools
  - 6 high schools
  - 8 middle schools
  - 37 Elementary Schools.
- Daily attendance sorted by school and grade level
- Enrollment data, by school and grade level
- Data quality issues
- We are in the process of making arrangements for receiving live data feeds from Norfolk Public Schools.

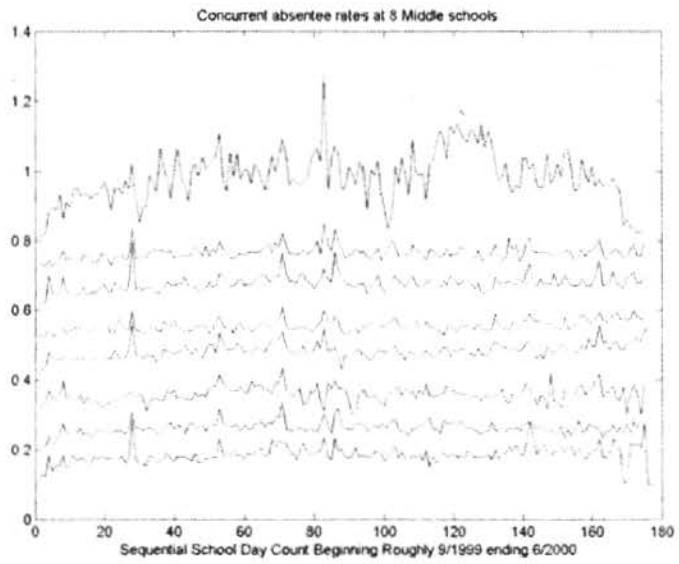
38

Concurrent Elementary School Absentee Rates



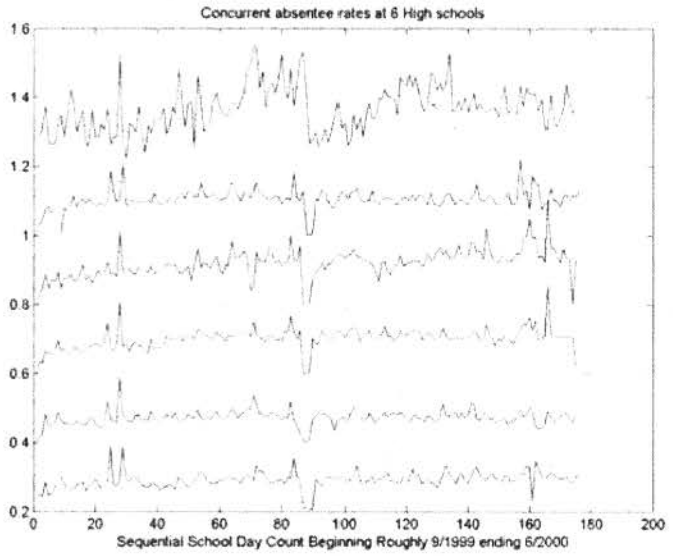
39

Concurrent Middle School Absentee Rates



40

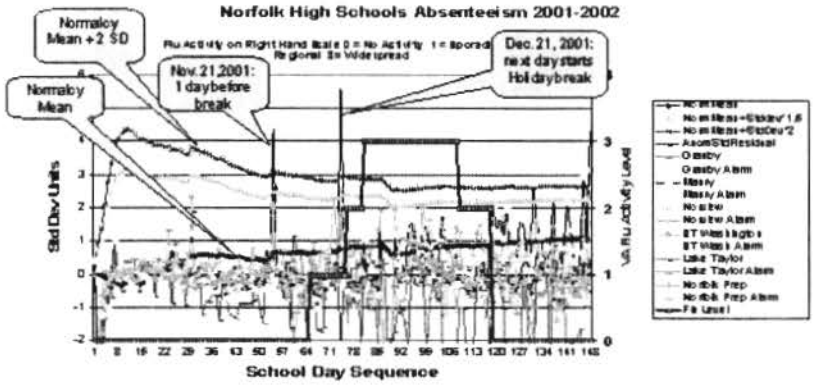
Concurrent High School Absentee Rates



41

Kalman Filter Results for High Schools

High Schools Percent Absenteeism Kalman Filter Results: Explainable Alarm Conditions



42

NT Data

Outline

- Overview
- 911 Call Data
- Wastewater Data
- Cough Detection
- School Absenteeism
- OTC Data

43

OTC

- Obtained three years of historical data (since July 1999(from 11 Naval Base Exchanges in the Hampton Roads area reporting sales of 55 drug types.
- Currently obtaining real-time daily data feeds via email.
- Currently pursuing MOU with Army Air Force Exchanges
  - data warehouse located in Dallas, Texas.

44

OTC Groupings

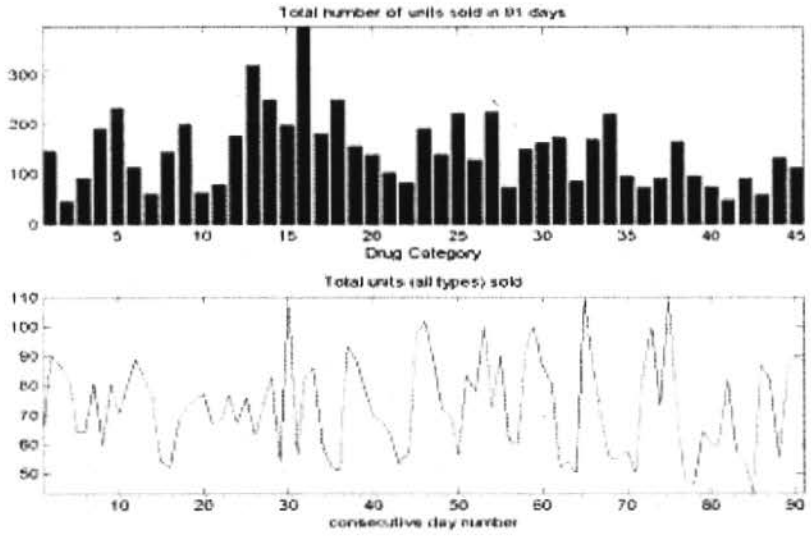
- We compressed the 99-00, 00-01, 01-02 Norfolk OTC sales data into 8 rational subgroups of products to make it easier for all to see what is happening.

The categories are

1: Pain Reliever, Antipyretic only.
2: Pain Reliever, antipyretic in combination with antihistamines, decongestants, and/or other flu/cold symptom relief drugs.
3: Cough syrup / drops / sore throat spray only
4: (3) combined with antihistamine, decongestant, and/or other flu/cold symptom relief compounds.
5: antihistamine/decongestant, antihistamine or decongestant, tablets
6: stomach / GI remedies
7: Skin
8: Other

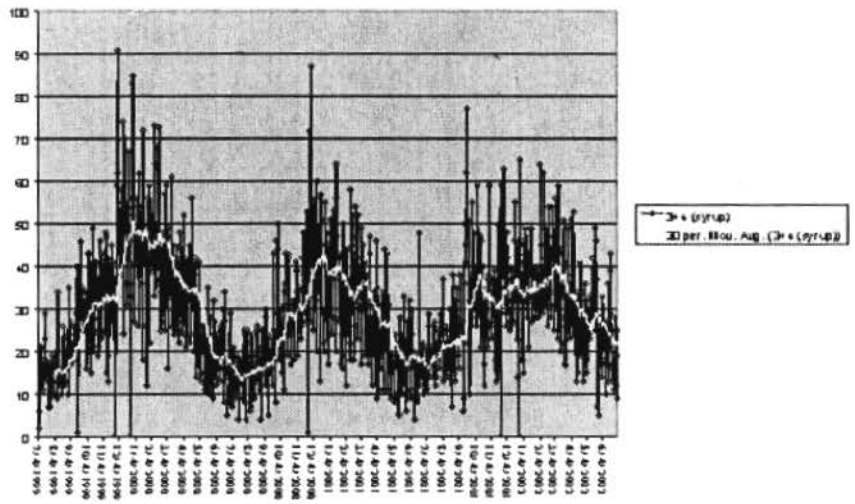
45

OTC Marginal Totals



'99-'02 Norfolk OTC
Syrup (Cough, Throat, Cold/Flu Ss)

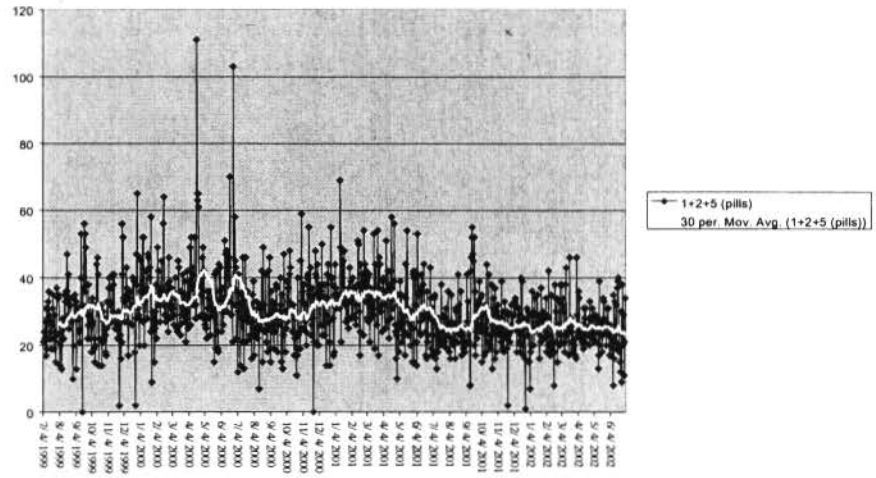
OTC Liquid Strongly Peaks with Winter Season



47

'99-'02 Norfolk OTC: Pills (Fever, Pain, Cold, Flu Ss)

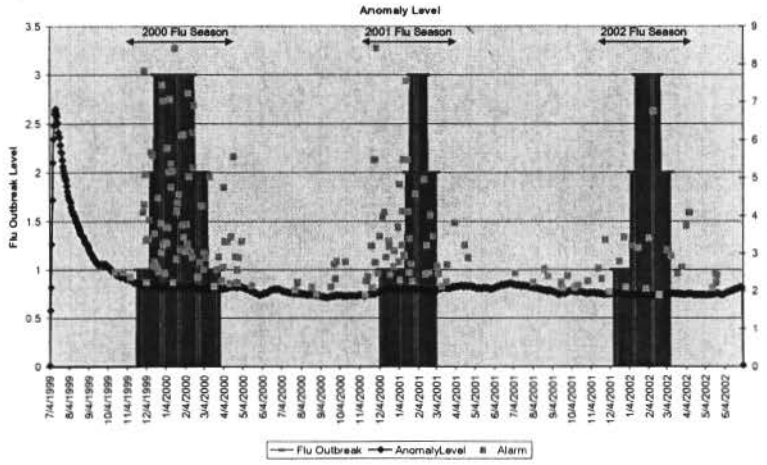
No clear yearly cycling in OTC "pills"



48

Cough Sales Anomalies at Norfolk Px Pharmacy

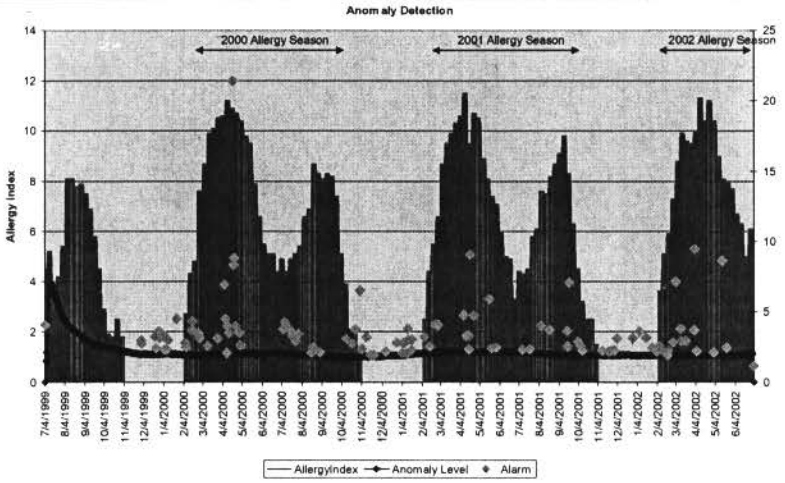
Cough Sales Seem to Follow Flu Season



49

Antihistamine Sale Anomalies at Norfolk Px Pharmacy

Antihistamine Sales Somewhat Follow Allergy Index

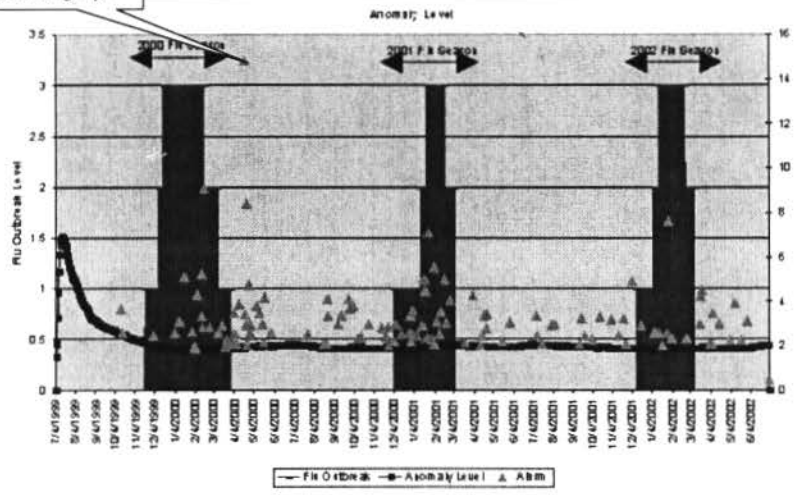


50

GI Sale Anomalies at Norfolk Px Pharmacy

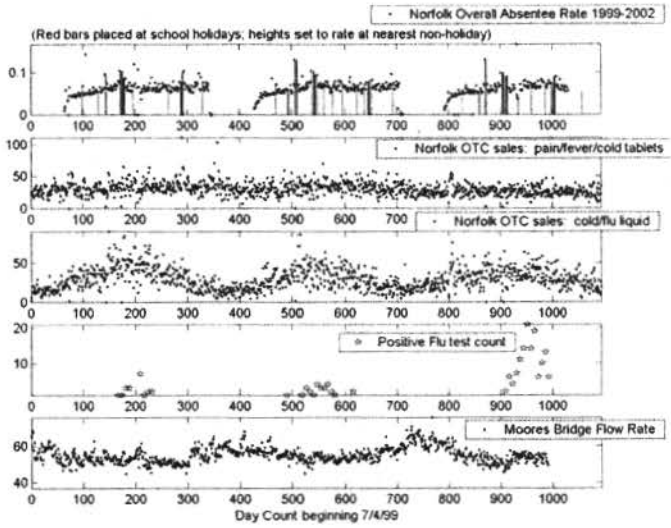
April 19, 2000 right before deployment of 2 Battlegroups

GI Sales Somewhat Follow Flu Season



51

Summary plots of three years of NT data collected for Hampton Roads.



52

OTC Conclusions

- Most of the cold/cough/flu sales data increased around the flu/allergy season.
- There was one noticeable exception on April 19th, 2000.
  - An unexpected surge in sales of cough, cold, antihistamine, aspirin and stomach drugs at the main Post Exchange at the Navy Base.
  - Could this be explained by the deployment of 2 battle groups from the naval base in the following days?
- Significant difference between liquids vs. tablets
- Changes in the differential purchasing patterns under normal conditions for different drugs may hold the key to detection of the unusual.

53

What's Next

- More comparison against gold-standards
  - Work with Data Evaluation group
- Systematic categorization and removal of trends
- Systematic capture of effects from known confounding factors
- Systematic analysis of individual data source contributions to system

1 of 8

PROJECT UPDATE

Veridian-Stanford Team
Dr. Gene E. McClellan

3nd PI Meeting
3-5 December 2002

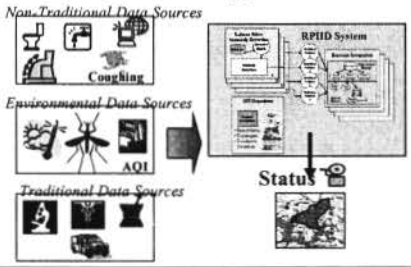
SMI

VERIDIAN

1

Real-time Population Health Detector (RPHD)
Veridian-Stanford/McClellan

Technical Approach



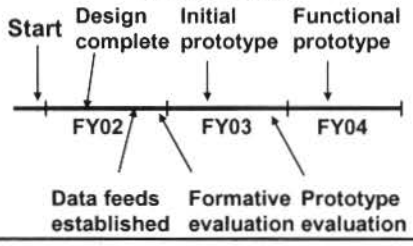
Objectives and Performance Goals

- Achieve early detection of outbreaks by use of behavioral, pre-diagnostic clinical, and animal data
- Improve detection and characterization of disease outbreaks through fusion of behavioral and medical indicators
- Create a durable software system that facilitates monitoring of data in multiple domains
- Improve ROC curves and timeliness of detection with non-traditional data sources and methods

Accomplishments

- Implemented and tested a prototype RPHD that integrates statistical detection algorithms with knowledge-based analysis
- Installed initial RPHD component at Naval Medical Center, Portsmouth
- Signed MOU and acquired historical and real-time, non-traditional data from City of Norfolk
- Obtained and analyzed non-traditional data, including wastewater data from Milwaukee

Programmatics



2

Technical Approach

What Is the Most Important and Revolutionary Result or Implication of This Project?

- Integration of statistical and knowledge-based techniques in an extensible monitoring system
  - Complementary strengths
    - Quantitative, predictive
    - Qualitative, explanatory
  - Effective utilization of broader range of data types
    - Standardized metadata
  - Explicit representation of knowledge for
    - Evidence combination
    - Disease characterization

3

Technical Approach

What Are the Key New Ideas? How Are They Achieved?

- *Anomaly detection* based on automated forecasting with dynamic models
  - Build robust and adaptive background models in Kalman filter detectors
  - Provide maximum likelihood estimation of model parameters to optimize performance
- Optimal grouping (dimensionality reduction) through covariance cluster analysis
- Spatial decomposition, manipulation, abstraction to enhance specificity and sensitivity of detection
- Parallel/distributed anomaly detection to ensure scalability and user acceptance

4

Technical Approach

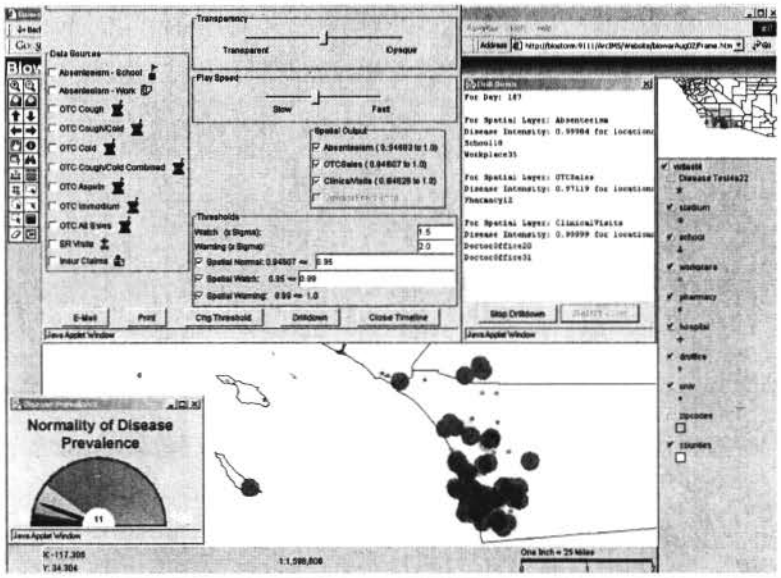
What Successes Have You Had This Year That Demonstrate the Feasibility, Validity, and Power of Your Approach?

- Have fused data from multiple sources through Bayesian belief network to monitor probability that disease prevalence is normal
- Successful anomaly detection on BioWar simulated data
- Accomplished spatial correlation of anomalies, visualization with a browser-based GIS, and drill-down to source-based data displays

Based on these successes, we have an end-to-end prototype of the RPHD system for monitoring population health.

5

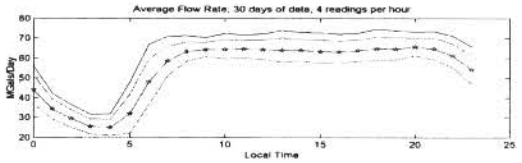
RPHD User Interface Showing Alert Condition Based on BioWar Scenario



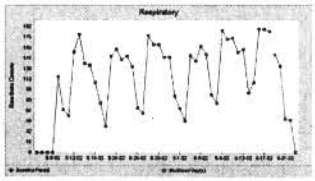
6

Examples of Attack Scenarios That Can Be Detected

- Overnight affliction of roughly 1/3 of population with gastrointestinal illness
  - Based on analysis of Norfolk water supply data

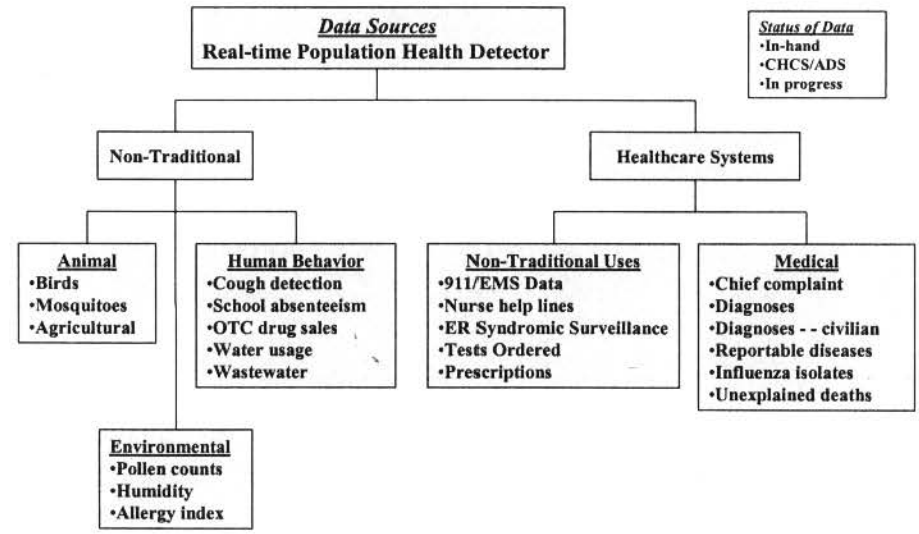


- Outbreak of about 20 cases of respiratory illness within the population served by the Naval Medical Center Portsmouth and reporting clinics
  - Cases on same day
  - False alarm less than monthly
  - More sensitive on weekends



7

What Are the Non-traditional Data Sources?



8

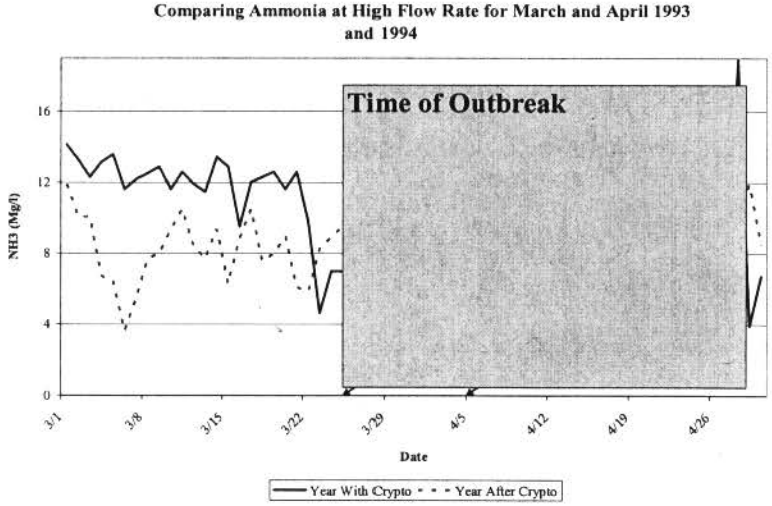
Technical Approach

What Are the Non-traditional Data Sources?

- *What Have You Selected, Acquired, and Characterized? How?*
  - Wastewater data -historical
  - Potable water usage -historical + e-mail @ 15 min int.
  - School absentee rates -historical
  - OTC drug sales -historical + e-mail @ 24 h int.
  - 911/EMS dispatch data -historical
- *How Much Faster Does Each One Serve As Leading Indicators of Outbreaks? How Do You Know?*
  - Cryptosporidium outbreak indicates possibility of detection by wastewater analysis as much as 11 days before historical recognition by public health officials and a day or so before it could be detected by clinical/ER surveillance
  - Initial evaluation of 911/EMS data show potential lead over hospital flu admissions by ~2 weeks consistent with NYC DOH analysis.

9

Preliminary analysis of Milwaukee wastewater measurements shows correlation with the cryptosporidium outbreak of 1993



10

Technical Approach

- Analysis of Milwaukee wastewater data
  - Shows significant decrease in wastewater concentrations associated with human sanitation during the cryptosporidium outbreak of 1993
    - biological oxygen demand
    - ammonia
    - phosphate
    - suspended solids
  - At least six days earlier than awareness by public health officials
  - Likely a day earlier than clinical/ER indications
  - Further analysis required to establish cause and effect

11

Technical Approach

What Advanced Analytics are being Used/Developed for Detection?

- Multiple source detection fusion by vector statistical inference on post-forecast residuals and their covariance components
- Bayesian processing of pre-diagnostic data
- Knowledge-based spatial reasoning
- Cough detection by audio spectral feature clustering

Characterization?

- Exploring combination of statistical analysis with spatial and temporal reasoning for disease characterization
- NBC CREST, which simulates biological weapon attacks and their medical consequences, will be used as a guide to attack and weapon characterization

12

Technical Approach - - Privacy

Are New Privacy Protection Techniques Being Developed? How Will They Be Used ?

- Privacy protection will be assured by:
  - Utilization of privacy protection technology available from others
  - Developing a distributed architecture for RPHD that will work with the most sensitive patient data inside the originating institution
    - Takes advantage of a distributed network of local experts/custodians responsible for and familiar with each data source
    - Will export only abstracted/aggregated data to central public health site; but traceable

13

Objectives and Performance Goals

What Is the Key Measure of This Project's Promise?

- Early, cost-effective detection of disease outbreaks

What Are You Measuring to Evaluate Your Progress?

- Timeliness of detection relative to traditional detection
- Performance against best available simulation data
- User acceptability
- Ease of accommodating new data sources in architecture

How Do You Know Your Results Are Valid?

- Careful analysis of data against known events
  - Outbreaks (flu, cryptosporidiosis)
  - Control for confounding factors (e.g., school absenteeism increase due to holidays, OTC increase due to ship deployment)
- Peer review

What Internal Evaluations Will Be Performed in Addition to the Common August Project Evaluation?

- Data-specific evaluations will be performed as new sources are added
- Will use real background data for sensitivity/specificity analyses in FY03

14

Objectives and Performance Goals

What Are the Key Performance Goals?

- Prove capabilities of a locally managed system providing
  - real-time analysis of non-traditional data providing early alerts
  - disease characterization
  - linkage to decision-support and emergency response systems
- Provide quantitative evaluation of alternative problem solving methodologies and contributions of individual data streams
- Demonstrate a measurable decrease in time from disease outbreak to appropriate response

What Are the Key Technical Milestones ?

- Move from simulated data to real-time analysis of real-world data
- Achieve adaptive aberration detection with Kalman filters
- Incorporate the Protégé ontology to support data management among systems
- Incorporate spatial reasoning with graphical display of results

15

Objectives and Performance Goals

What Are the Tradeoffs Between Sensitivity, Specificity, and Spatial/temporal Coverage/granularity?

- Must decrease granularity (spatial resolution) to increase sensitivity for larger attacks
  - 911 data analyzed by zip code rather than by census block
  - However, this decreases sensitivity to focused attacks
- Must group individual items to increase sensitivity
  - Aggregation of OTC drug categories
  - May reduce ability to characterize illness

16

Objectives and Performance Goals - I

What Are the Key Deliverables and When Will They Be Available for Integration And/or Transition?

Technology Product	Description	Potential Customers	Date of Availability
RPHD	Real-time Population Health Detector	Naval Medical Center, Portsmouth Hampton Roads Regional Epidemiologist	February 2004 functional prototype December 2004 fully deployed

17

Objectives and Performance Goals - II

What Are the Key Deliverables and When Will They Be Available for Integration And/or Transition?

Technology Product	Description	Potential Customers	Date of Availability
Cough Detector	A component of RPHD for the detection the sounds of coughing in a domestic, business, or classroom environment using frame-based sound classification.	Any Bio-Surveillance Data Collection System	April 2003
Water Usage Analysis Tool	A component of RPHD to analyze water usage data as an indicator of population health.	Any Bio-Surveillance Data Collection System	May 2003
Wastewater Analysis Tool	A component of RPHD to analyze wastewater data as an indicator of population health.	Any Bio-Surveillance Data Collection System	June 2003
Ontologies	Ontologies authored in Protégé 2000 for standard data formats such as NEDSS Logistical Data Model (NLDM).	Bio-ALIRT Program participants	Dec 2002

18

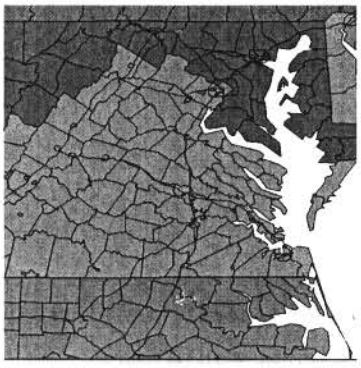
Accomplishments

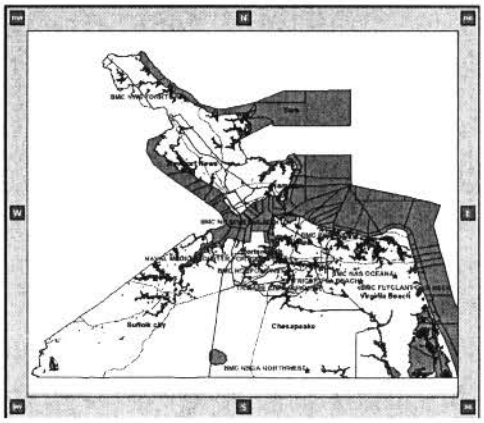
What Is the Most Important Indicator That This Project's Promise Might Be Achieved?

- RPHD has proven it's ability to find real anomalies in real, non-traditional data
  - School absenteeism connected with real events
  - OTC spikes potentially correlated with ship deployments
  - Significant changes in Milwaukee wastewater chemistry temporally associated with the cryptosporidium outbreak of 1993
  - BASIS detection of gastrointestinal cluster in NMCP data

19

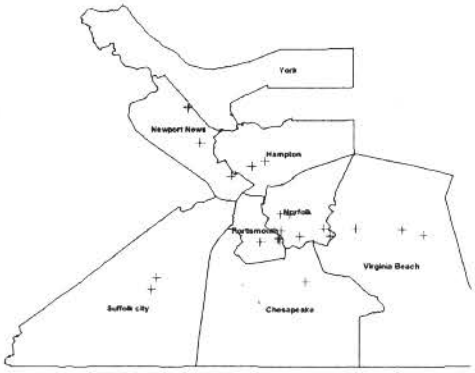
Reporting Clinics for NMCP





20

Hospitals in Hampton Roads



21

Accomplishments

- Focused cooperation of both military and civilian officials in Hampton Roads
  - Established network of military and civilian relationships for transitioning RPHD to the Hampton Roads community
  - MOUs with the Naval Medical Center, Portsmouth (NMCP) and the City of Norfolk
  - Extensive contacts with civilian hospital and public health personnel
  - Discussions with Navy Public Works Center Norfolk indicate potential for collecting non-traditional data specific to Navy residential facilities

22

Accomplishments

What Has Been Your Key Accomplishment So Far?

- Creation of an end-to-end RPHD prototype and installation a functional component at NMCP

What Obstacles Have Been Encountered?

- Arrangement of MOUs and agreements
- Uncertain mechanisms for military/civilian public health cooperation
- Lack of automation infrastructure among data providers

Where Has Progress Exceeded Expectations?

- Uniformly positive expressions of support and cooperation from both military and civilian leaders, medical, public health personnel in Hampton Roads

What Insights Have Been Gained?

- Must promote an effective military-civilian relationship relative to public health
- Data idiosyncrasies drive modeling approach, source by source
- Intelligent grouping of data sources requires a combination of statistical and knowledge-based techniques

23

Programmatics

Describe How Your Project Is Collaborating With Others.

- Designed and implemented a data-sharing ontology
  - standardized nomenclature
  - provides mapping for non-traditional data sources
  - Will support sharing of data for standardized testing of algorithms
- Supported Development of CMU's BioWar
  - Provided background material on models of bio-agent-induced disease to CMU
  - Tested early BioWar code
  - Intensive analysis of BioWar output including medical, spatial, data integrity aspects
  - Provided constructive feedback on BioWar
    - Expert medical critique of clinical and epidemiological parameters
- Acquired CMU's WSARE
- Investigating applicability of IBM approach to campus monitoring via outgoing health-related phone calls
- Interchange with IBM on cough detection technology

24

Accomplishments

Summary of the 02 Algorithm Evaluation

- Individual AMOC curves for each attack size show consistent behavior over a wide range of false alarm rates.

	5% False Alarm Probability		10% False Alarm Probability	
	Average Score	Average delay (Days)	Average Score	Average Delay (Days)
Small	0.746	5.1	0.746	5.1
Inside	0.754	4.9	0.788	4.2
Medium	0.786	4.3	0.794	4.1
Large	0.822	3.6	0.822	3.6

RPHD performed best on Large Attacks. Performance decreased with attack size.

25

Accomplishments

Summary of the 02 Data Evaluation

- Significant non-traditional sources of data have been arranged in Hampton Roads
  - School absenteeism
  - OTC drug sales
  - Potable water usage
  - Wastewater flow and characteristics
- Preliminary analyses performed with RPHD
- Potential for significant impact of non-traditional data illustrated with Milwaukee wastewater data covering the cryptosporidiosis outbreak of 1993

26

Accomplishments

What Specific Transition Opportunities Have Been Identified And/or Exploited?

Who is Using Your Product Now?

- RPHD component is operational at the Naval Medical Center, Portsmouth

Who has Committed to Using it in the Future?

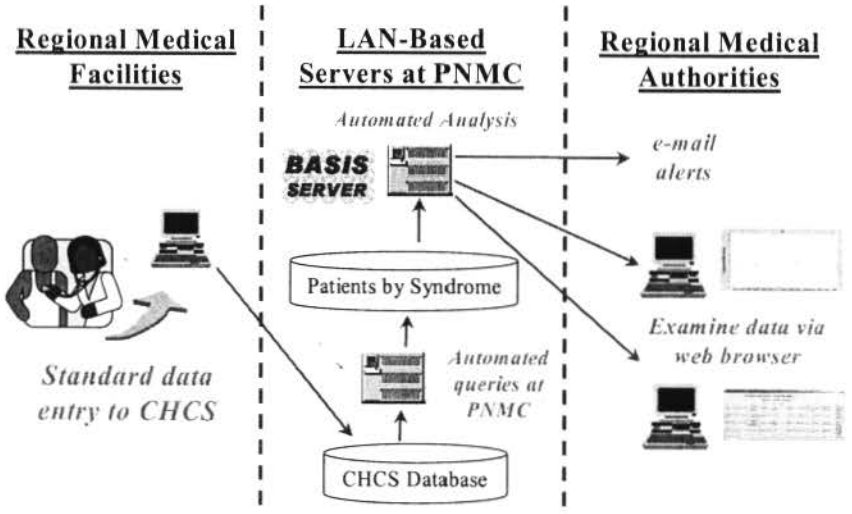
- Hampton Roads Regional Office of the Virginia Dept. of Health
- City of Norfolk

What Other Opportunities are there ?

- Additional jurisdictions in Hampton Roads area
- Sentara and other civilian hospital systems
- Eastern Virginia Medical School
- Army and Air Force Hospitals
- Metropolitan Medical Response System
- Naval Health Research Center

27

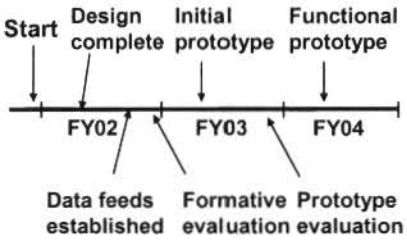
BASIS Automated Alerts for PNMC



28

Programmatics

- Identify Current Project Plan, Including Key Milestones for Evaluations, Deliverables, and Transitions



29

RPHD Tasking FY03

- Non-traditional Population Health Indicator Data
  - Evaluate IBM's closed-campus phone monitoring for application at Norfolk Navy Base
  - Evaluate evidence for cryptosporidiosis outbreak via Milwaukee wastewater analysis
  - Analyze water usage to detect large-scale outbreak of gastrointestinal illness
- Advanced Analytical Techniques
  - Improve normalcy models to account for cyclic variations, environmental influences, etc.
  - Incorporate dynamic parameter fitting and cluster analysis to optimize Kalman filters
  - Exploit richness in data via knowledge-based processing techniques
  - Develop advanced spatial reasoning and spatial decomposition/recomposition methods
  - Augment data fusion through vector statistical inference on post-forecast residuals and their covariance components
  - Prototype methods for disease characterization
- RPHD System Development
  - Implement end-to-end system that automates processing and analysis of "live" data
  - Utilize Protégé ontology to access data at run time
  - Develop web-based "light" client to present geographic results
  - Test the RPHD system with simulated and real-world data
  - Apply privacy protection methods from IBM and CMU
- Disease Surveillance Activities
  - Expand disease surveillance at Naval Medical Center Portsmouth (NMCP)
  - Test prototype audio cough detection in military classrooms
- Transition to Operational Use
  - RPHD components at NMCP, Air Force and Army hospitals in Hampton Roads
  - RPHD components within city public health departments and private sector medical facilities in Hampton Roads

30

RPHD Tasking FY04

- Non-traditional Population Health Indicator Data
  - Determine nationwide applicability of non-traditional data for regional disease surveillance
  - Enhance knowledge bases of data sources and their properties
  - Implement operational methods for routine acquisition and protection of data
- Advanced Analytical Techniques
  - Finalize adaptive, multivariate forecasting models underlying aberration detection
  - Develop additional knowledge-based problem solvers in response to FY03 experience
  - Characterize and encode behavior of all problem-solvers in a knowledge base, allowing RASTA control structure to improve overall system performance
- RPHD System Development
  - Refine all RPHD components based on FY03 demos to ensure robust, deployable system
  - Develop fault tolerant servers with data backup and recovery solutions
  - Ensure easy installation and software maintenance
  - Create user manuals and documentation
- Disease Surveillance Activities
  - Conduct unified military and civilian disease surveillance for Hampton Roads
  - Support user training in Hampton Roads
- Transition to Operational Use
  - Finalize transition to Hampton Roads Regional Public Health and Metropolitan Medical Response System (MMRS)
  - Develop additional Navy applications via Naval Health Research Center, San Diego

31

Programmatics

What Else is Needed?

Data Sources to be Acquired, Characterized, and Analyzed ?

- Nurse help-line data
- Data involving known outbreaks for performance assessment

Other

- Data de-identification software
- Literature and/or research results predicting behavioral observables based on illness
- Evidence-based case definition of syndromes for disease surveillance

32

Programmatics

What, if anything, Has Changed From Your Original Plan? Why ?

- Greater integration of Stanford and Veridian components in composite software architecture
- Driven toward more distributed processing by privacy concerns
- Because of the programmatic focus on BioWar, have spent little effort developing simulation

33

Programmatics

Identify Plans for Internal Evaluations

- Data-specific evaluations will be performed as new sources are added
- Will use real background data for sensitivity/specificity analyses in FY03

Identify Other Funding Sources, Division of Labor, and Deliverables

- N/A

34

Programmatics

Describe Your Relationship with Customers/Testbeds

- Naval Medical Center, Portsmouth (NMCP)
  - Transition partner defined by MOU between AFRL Rome and NMCP
  - Data supplier and user of Bio-Surveillance technology
  - Dominant military healthcare provider in Hampton Roads
- City of Norfolk
  - Transition partner defined by MOU between Veridian and City of Norfolk
  - Data supplier and user of Bio-Surveillance technology
- Metropolitan Medical Response System (MMRS) of Hampton Roads
  - Provides multi-jurisdiction coordination
  - Primary responsibility for large scale medical emergency
- NHRC
  - BASIS detection techniques implemented in NHRC's MDSS
  - Likely transition partner beyond Hampton Roads

35

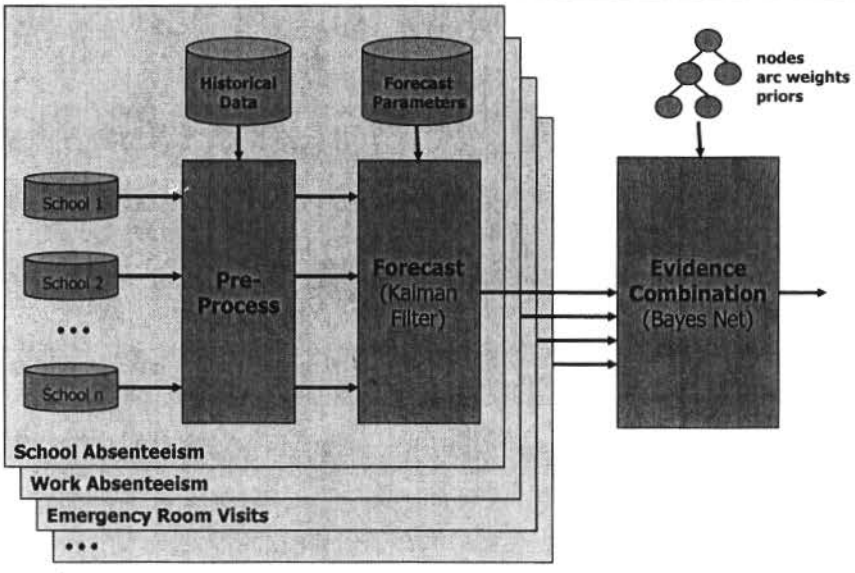
1.A.9

RPHD Analysis Techniques

December 2002

1

RPHD Component Interaction



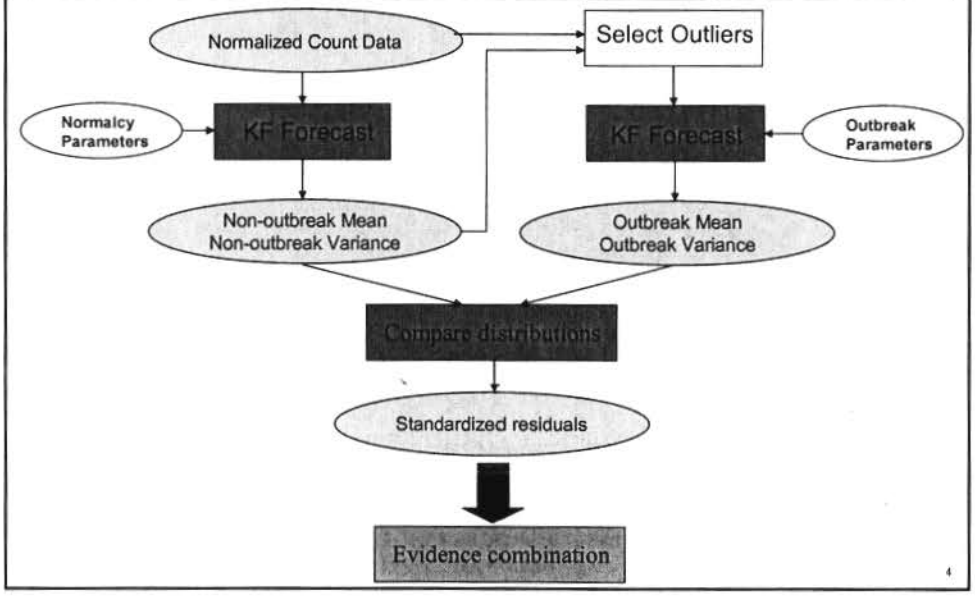
2

Forecast Component: Kalman Filter Internals

- State vector (internal representation) of data:
  - Aggregate mean
    - Weighted sum for each data type (school absenteeism, work absenteeism, OTC cough sales, etc.)
  - Local slope
    - Not exposed, used in forecast updates only
- 'Structural Model'
  - Double random walk (aggregate mean and local slope)
  - Additive noise term
- 'Normal' and 'Outbreak' Parameters
  - Parameters encode knowledge of trends and short term variability in data

3

Forecast at Each Time Step



4

Filter Outputs For Each Data Type

- Day number k
- Normalcy model aggregate mean: m_k^n
- Normalcy model standard deviation: σ_k^n
- Outbreak model aggregate mean: m_k^d
- Forecast standard residual at each location
- Output to next stage:

$$v_k = (m_k^d - m_k^n) / \sigma_k^n$$

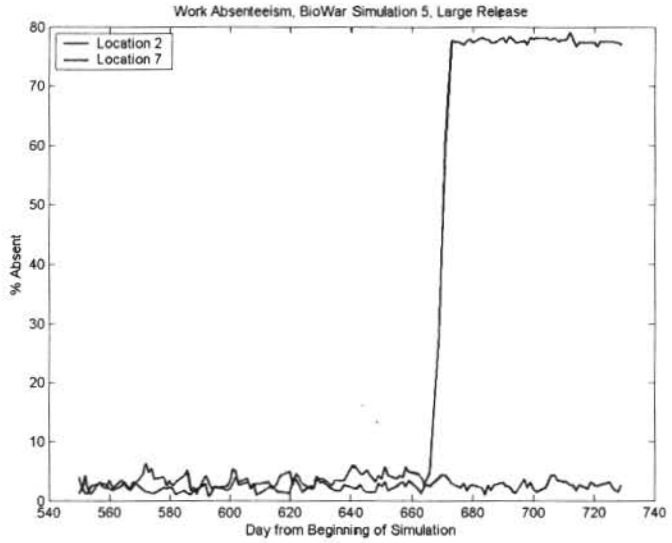
5

Examples: BioWar Data

- Set 5, challenge problem data
- Large attack, commencing at day 663
- 12 workplace locations affected

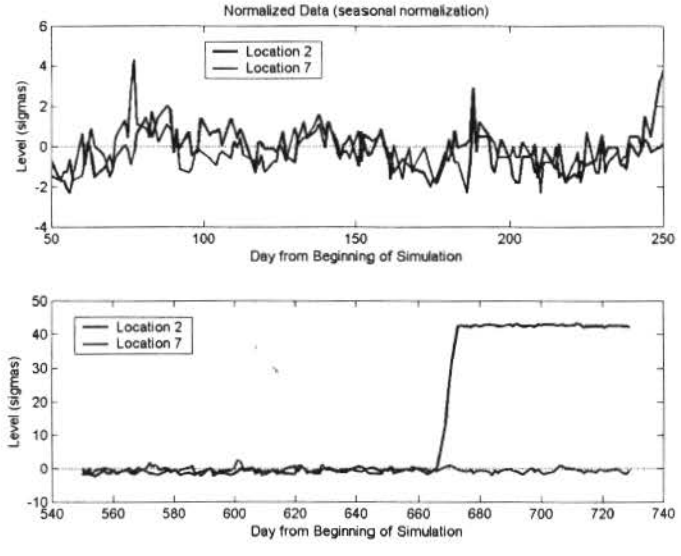
6

BioWar Simulation Set 5
Workplace Absentee Data



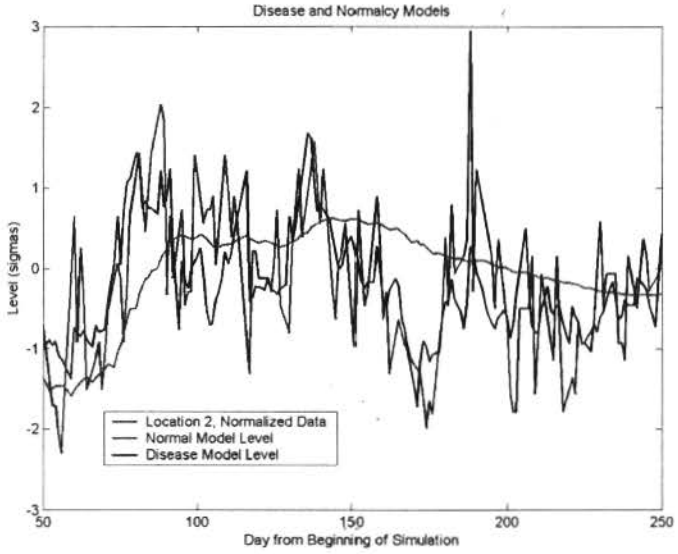
7

BioWar Set 5
Preprocessed Data



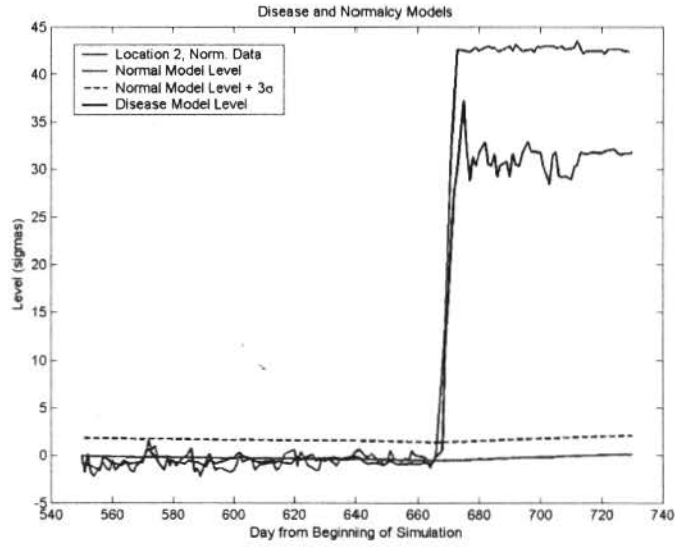
8

BioWar Set 5: 'Normal' and 'Outbreak' Model Levels (pre-attack)



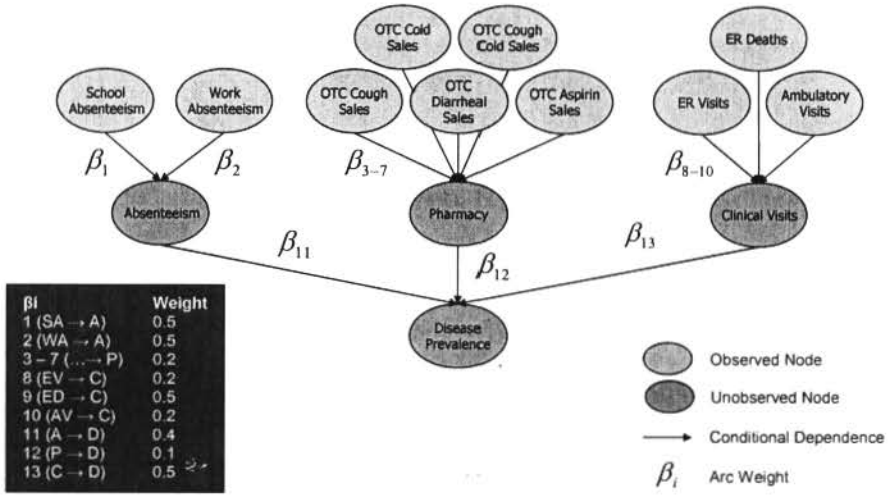
9

BioWar Set 5: 'Normal' and 'Outbreak' Model Levels (attack at day 663)



10

Belief Network Structure



βi	Weight
1 (SA → A)	0.5
2 (WA → A)	0.5
3 – 7 (… → P)	0.2
8 (EV → C)	0.2
9 (ED → C)	0.5
10 (AV → C)	0.2
11 (A → D)	0.4
12 (P → D)	0.1
13 (C → D)	0.5

11

Implementation of Belief Net

- Construct graph
  - Nodes have Normal (Gaussian) distribution
  - Arcs specify dependence
  - Graphical structure facilitates knowledge acquisition and representation
- Calculate precision matrix from graph and invert to determine covariance matrix

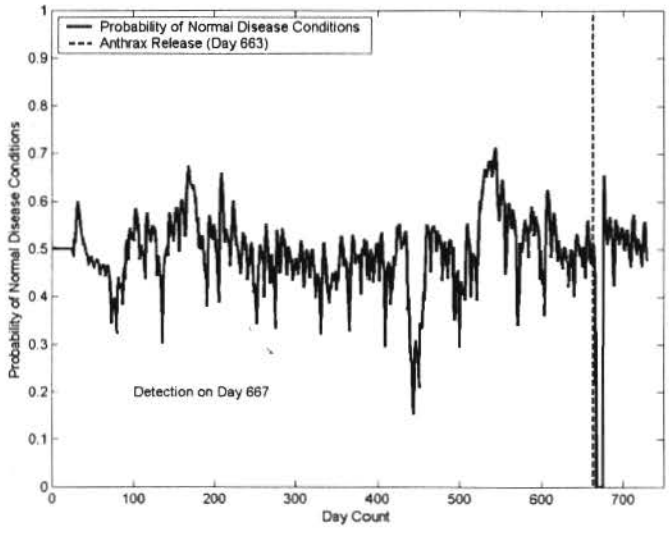
12

Implementation of Belief Net

- At each time point
  - Incorporate 'aggregate' evidence and calculate posterior distributions on unobserved nodes with linear time algorithm
  - Calculate probability of seeing value as large as posterior mean, assuming it is drawn from the expected distribution

13

BioWar Set 5: Output of BN



14

System Performance

Average number of days from attack to detection as a function of false positive rate from AMOC analysis.

False Positive Rates	Avg. # of Days Overall	Avg. # of Days For Attacks Inside Buildings	Avg. # of Days For Small Attacks	Avg. # of Days for Medium Attacks	Avg. # of Days for Large Attacks
0.1 %	4.46	4.91	5.07	4.28	3.72
1.0%	4.22	4.25	5.07	4.12	3.56
3.0%	4.18	3.75	4.76	4.08	3.2
5.0%	4.06	3.75	4.76	4.04	3.2
10.0%	3.66	3.66	4.23	4.00	2.68

15

Transition from Simulated to Real Data (Problems and Solutions)

Specification of Kalman filter model parameters

- Simulated data
  - Based on variance of background data
  - Required some tuning 'by hand' for different data sets
- Real data
  - Model fit to 'normalcy' data using bootstrap techniques to obtain initial parameter estimates and refined using maximum likelihood estimation
  - Good convergence requires use of more simple structural model (random walk on level+additive noise)

16

Transition from Simulated to Real Data (continued)

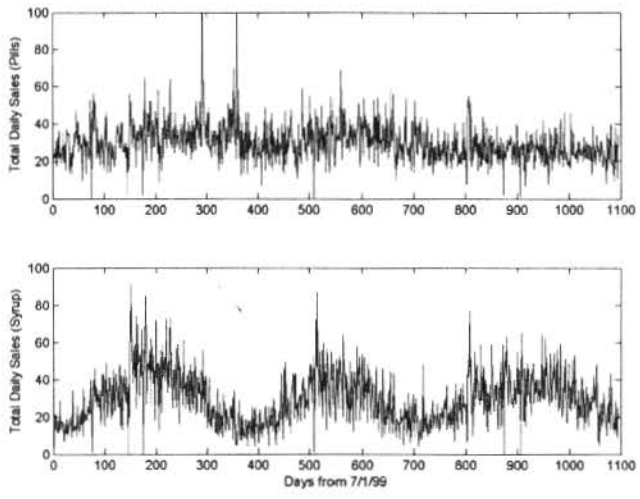
Dimensionality Reduction (OTC Drug Sales)

- Simulated Data:
  - Not an issue, OTC Sales already in 5 groupings
- Real Data:
  - 51 drugs in first analysis from Norfolk exchanges
  - Differences in packaging, count, dose, etc.
  - Simple 'rational' groupings led to unexpected results
  - Techniques developed to group drugs by periodicities at various wavelengths (using Principal Component Analysis and Hierarchical Aggregation based on correlation measure)

17

Grouping OTC Drugs: An Example

Grouping of OTC Sinus/Cold/Flu Remedy by 'Pills' and 'Syrup'



18

Current Work

- Preprocessing
  - Replace seasonal averaging with models which remove all known trends
- Adaptive estimation of Kalman Filter parameters
  - Replace 'training set' with all data determined 'normal'
- Periodic re-evaluation of data groupings
  - Respond to deletions and additions to drug types sold

19

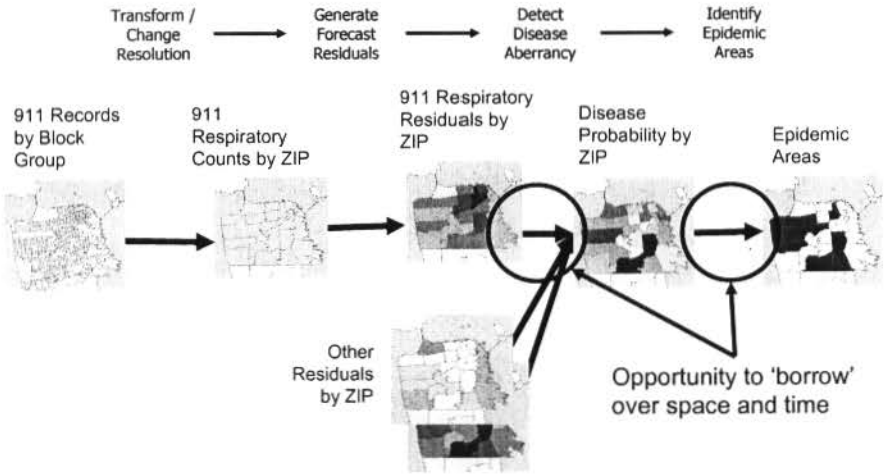
Current Work (continued)

Spatial Processing

- Decompose data to common spatial units
  - Combine evidence at unit level then identify aberrant regions with global operators
  - Identify aberrant regions for each data source with global operators then combine aberrant regions
- Extend evidence combination to incorporate temporal lags and employ other aggregation operators (e.g., max)

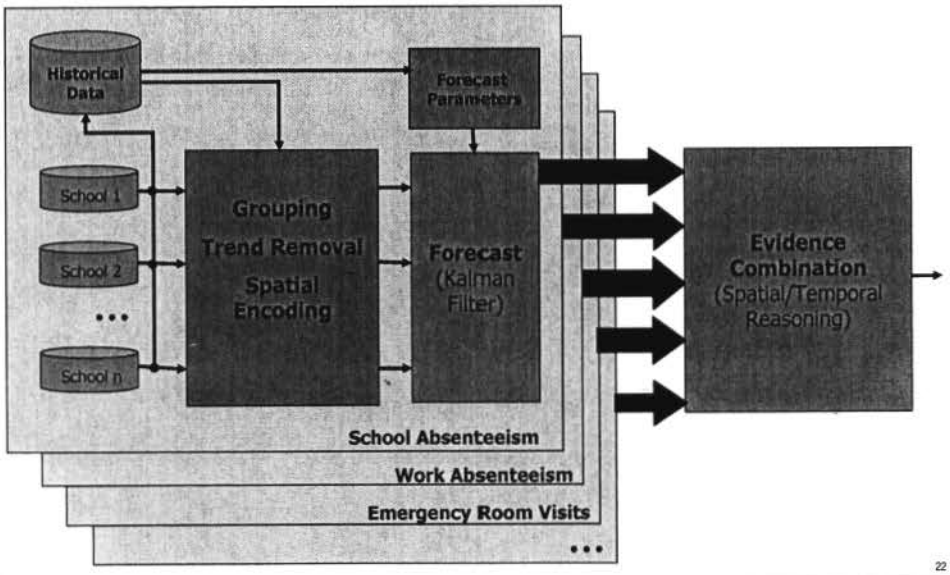
20

Spatial Reasoning Approach



21

Updated Component Interaction



22

TITLE PAGE

Briefing Materials concerning TIA BIOALERT 2

FOLDER 1.b-i TOC 14

MEETINGS

- 1.b Aug Challenge Algorism Evaluation Results
    - Potomac Institute for Policy Studies
- 1.c Bio-Surveillance Evaluation
    - Potomac Institute for Policy Studies
- 1.d DARPA BSS Program Data Evaluation
    - Johns Hopkins U Applied Physics Lab
    - WRAIR
    - IBM
    - Verdian-Stanford
    - Pitt-CMU
- 1.e First Annual Bio-ALIRT Evaluation
    - Potomac Institute for Policy Studies
- 1.f Next Annual Bio-ALIRT Testing
    - Potomac Institute for Policy Studies

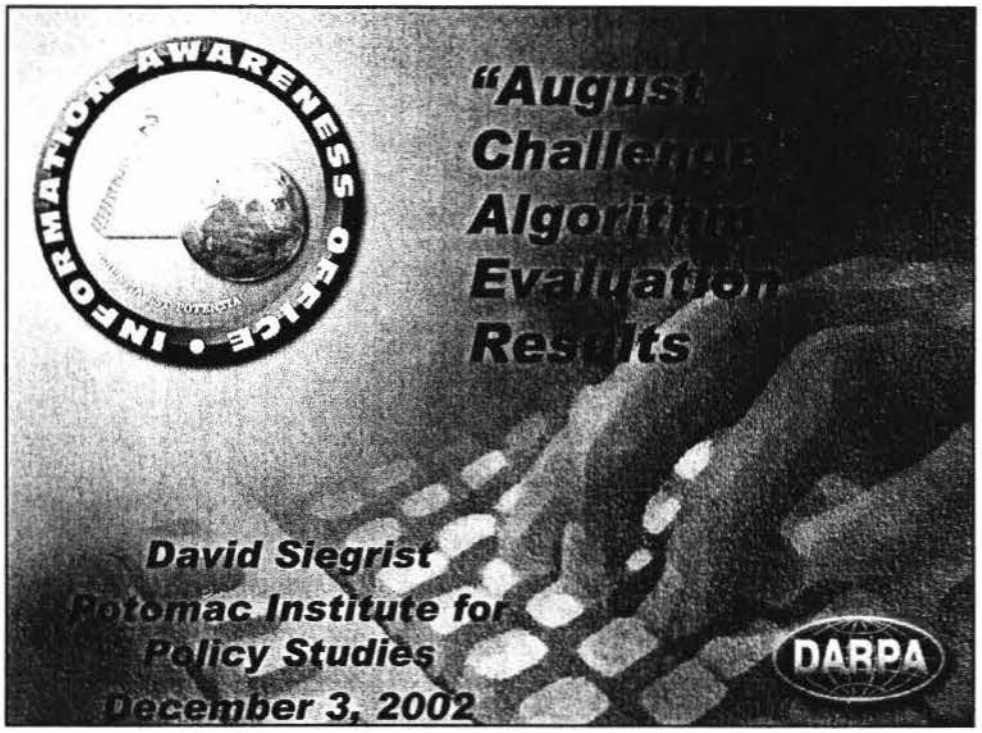
FOLDER 1.b-i TOC 14 – Cont'd – Page 2

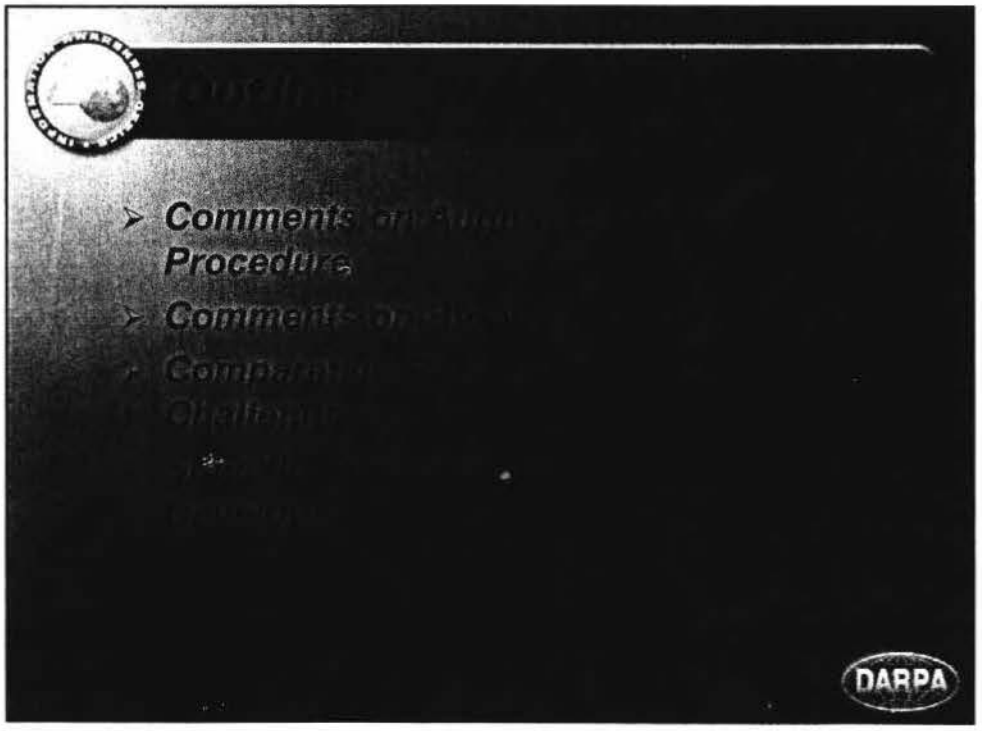
- 1.g RPHD Tasking by FY (input-all team members)
- 1.h Bio-ALIRT: Bio-Event Advanced Leading Indicator Recognition Technology

  Principal Investigators' Meeting 3-5 Dec 02
- 1.i Bio-ALIRT: Bio-Event Advanced Leading Indicator Recognition Technology – ACTION ITEMS

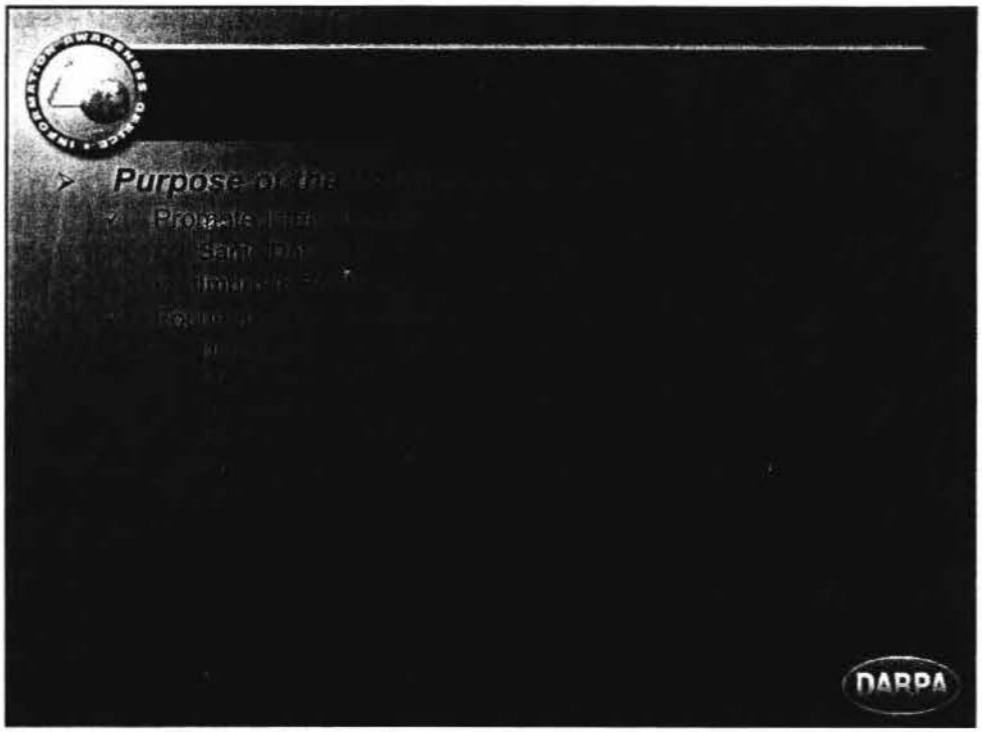
[redacted]

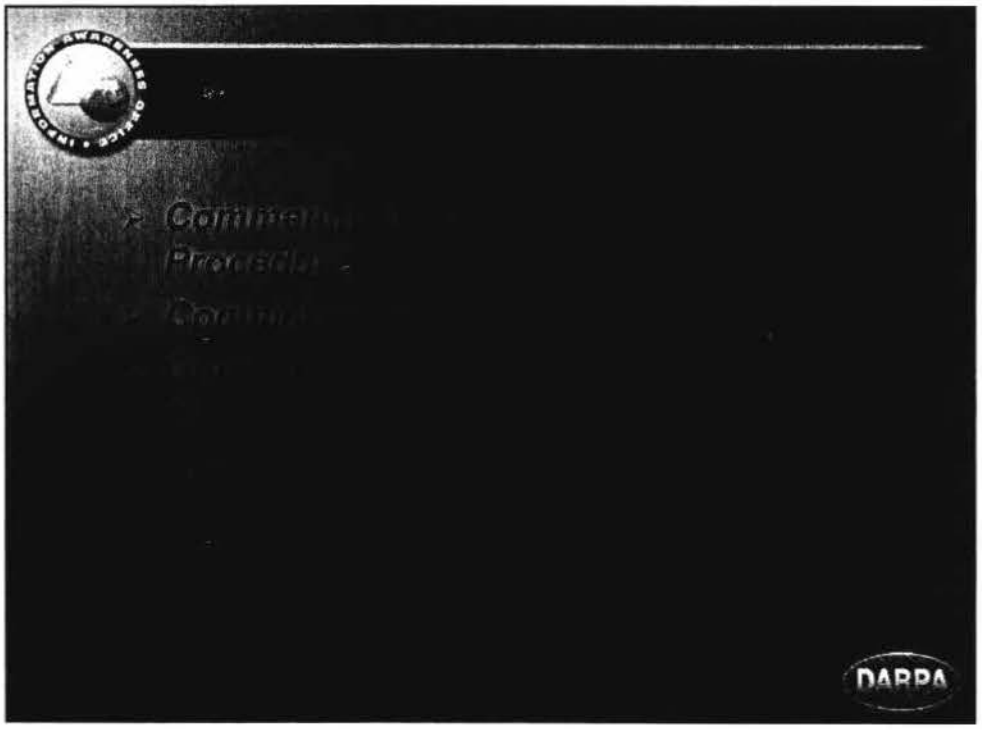
to Bo

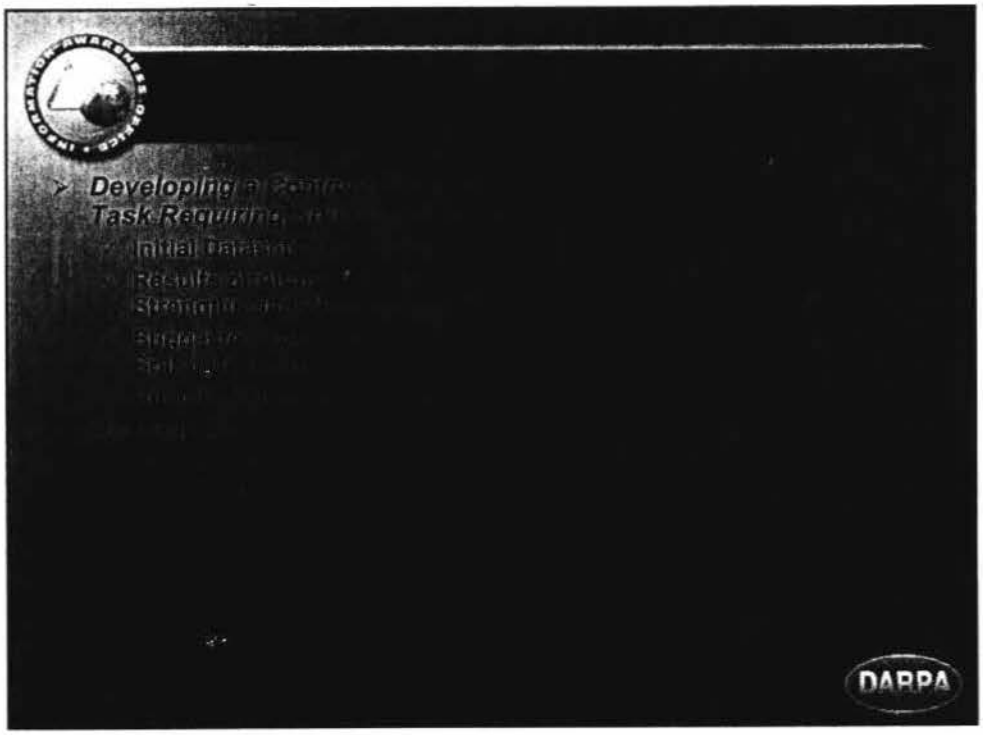


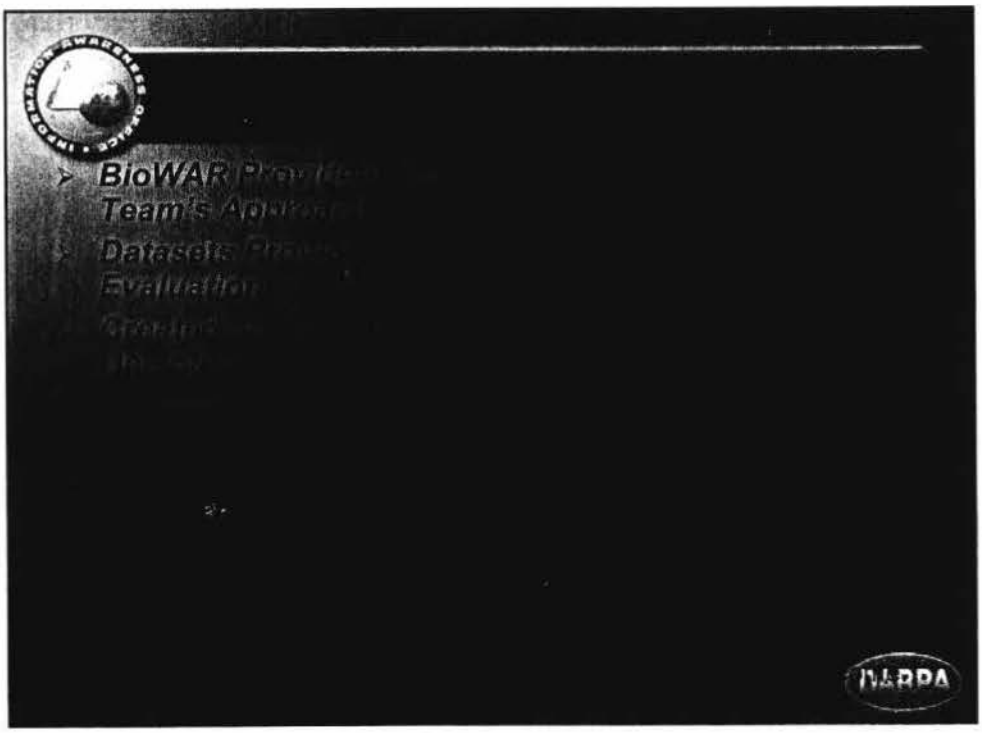


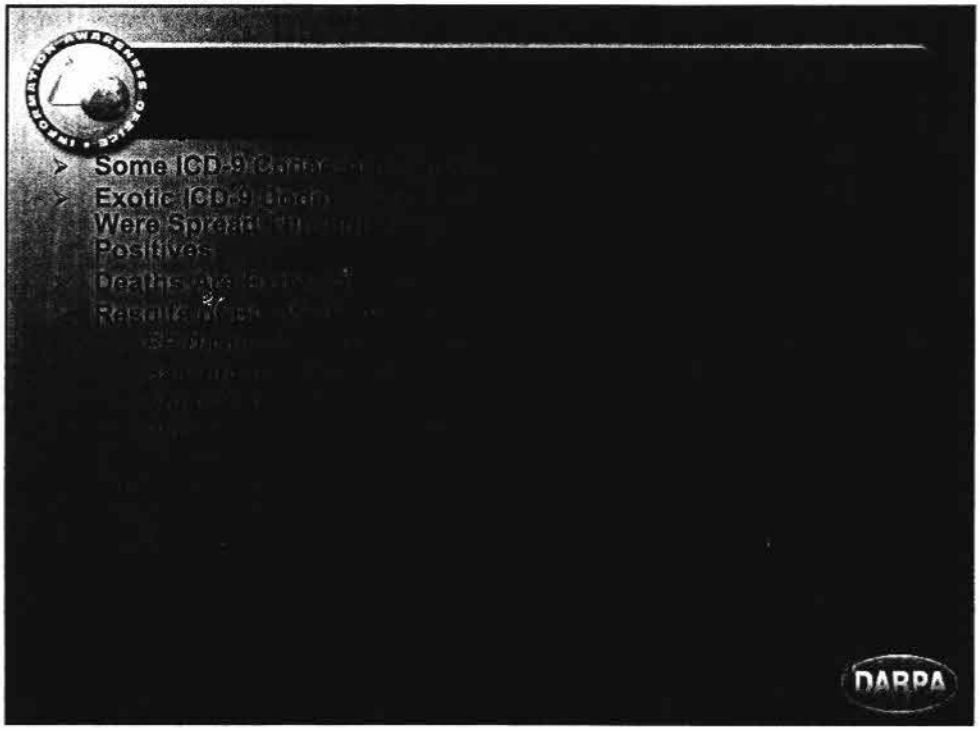
BEST COPY AVAILABLE

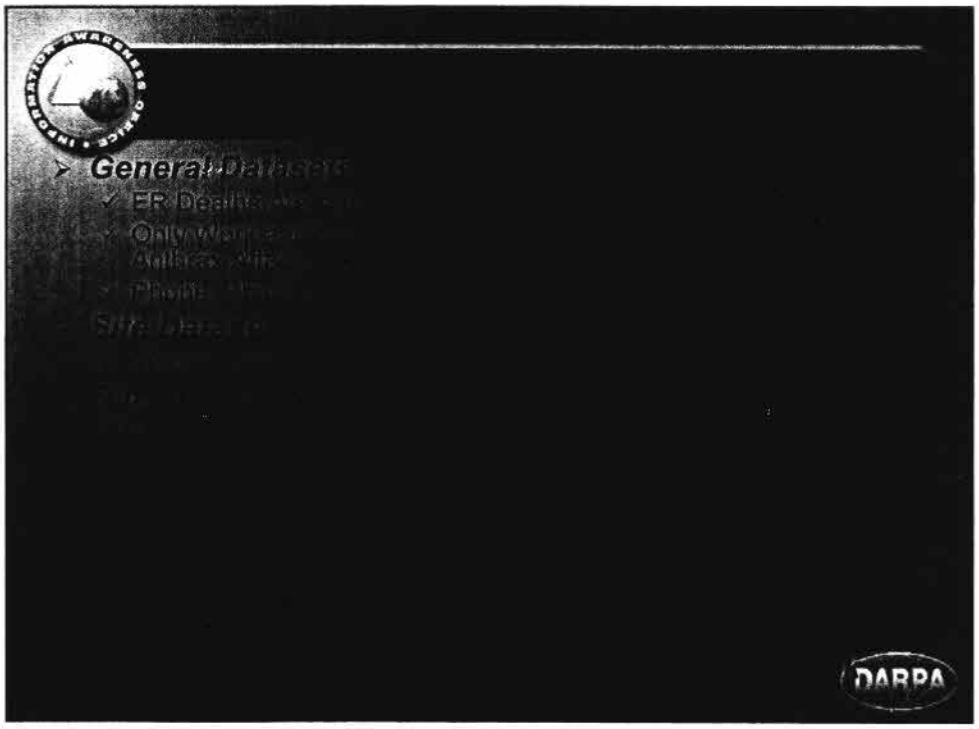

DARPA

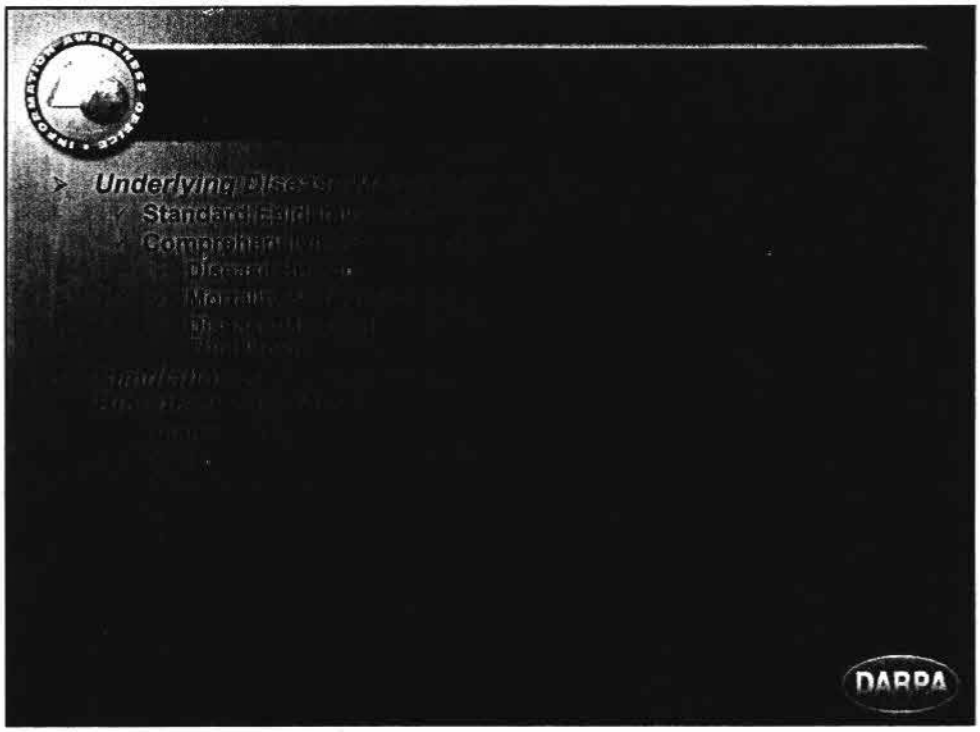

DARPA

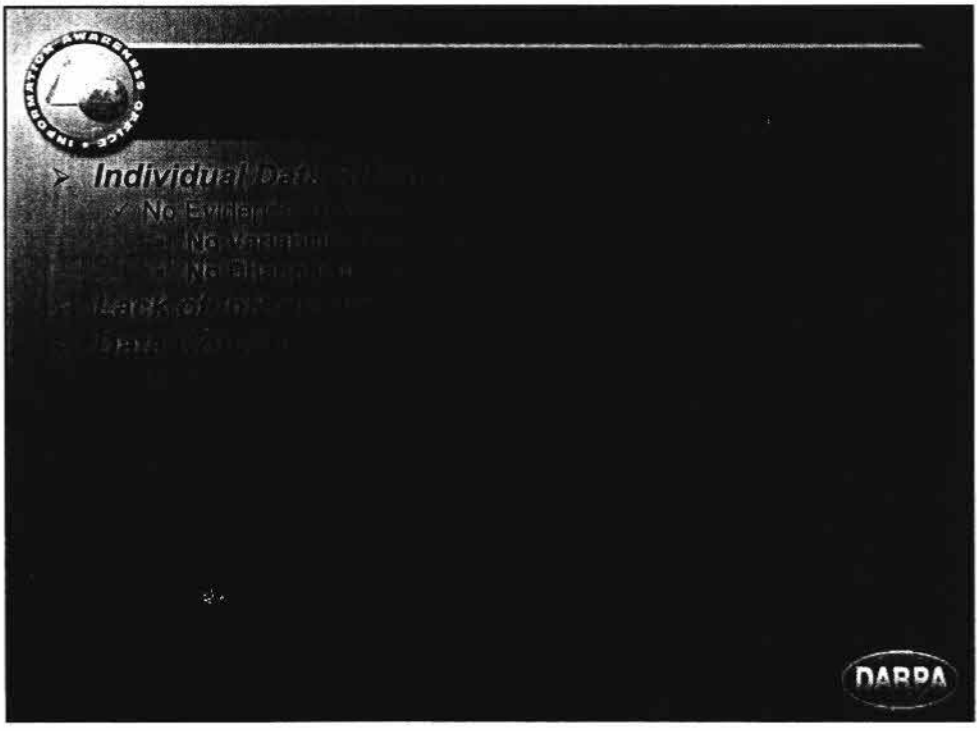

DARPA

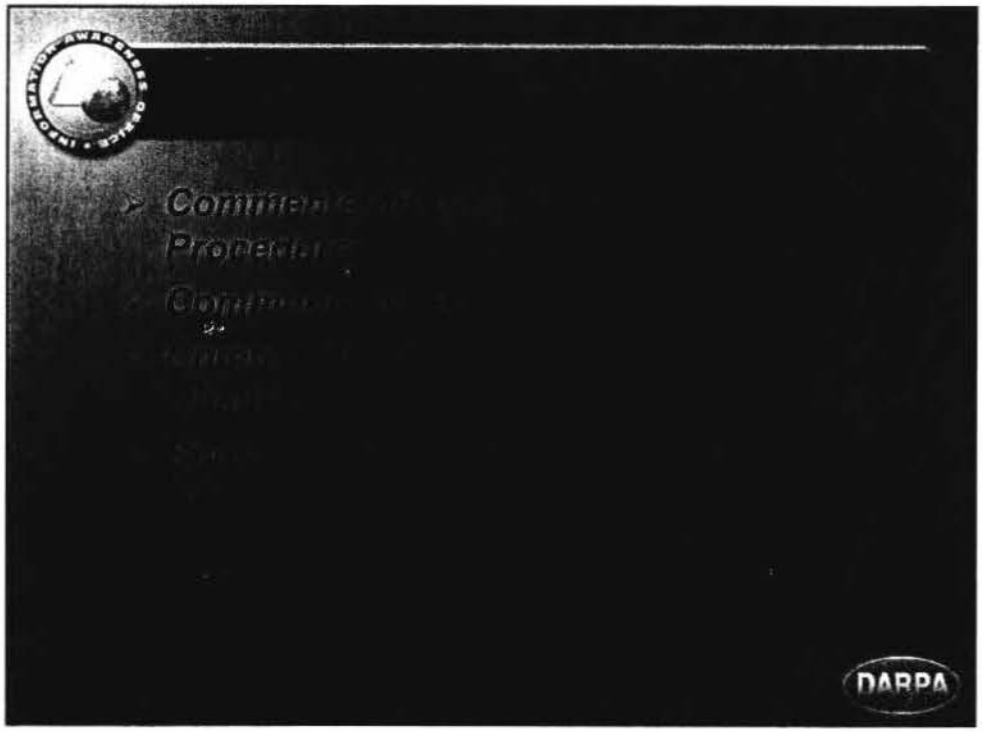

DARPA

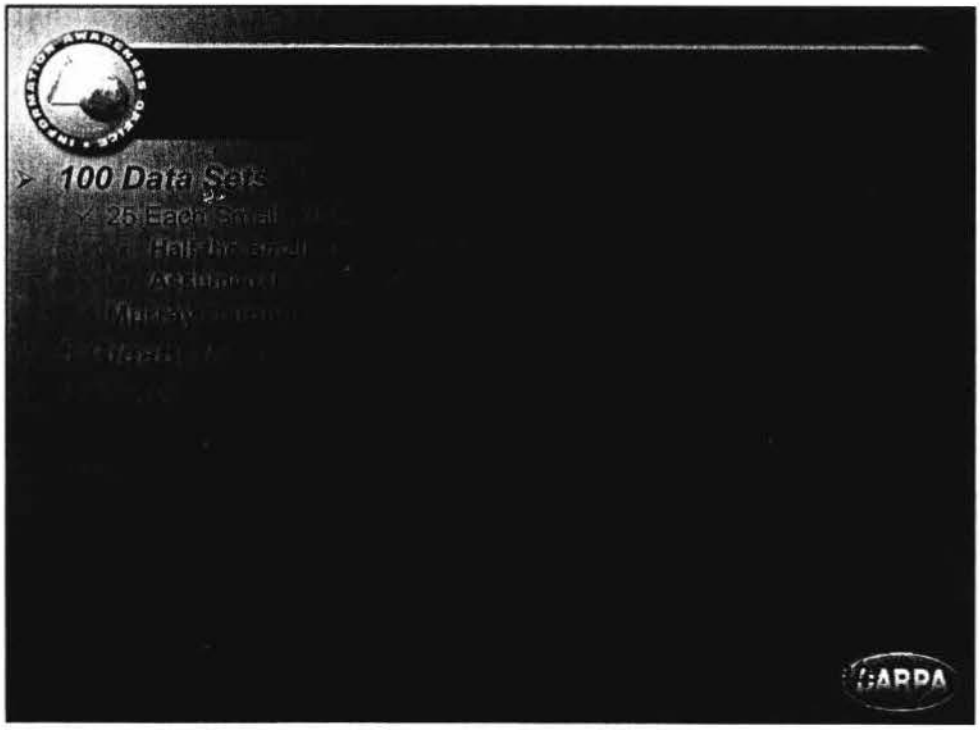

DARPA


DARPA


DARPA


DARPA


DARPA


DARPA

[redacted]

12

10

"August Challenge" Algorithm Evaluation Results

David Siegrist, James Richardson
Potomac Institute for Policy Studies
November 14, 2002

Draft 1.0 for Review

Outline

- Comments on August Challenge Procedure
- Comments on Bio-WAR Datasets
- Comparative Results of August Challenge
- Specific Tests Performed by Each Developer

Draft 1.0 for Review

Outline

- Comments on August Challenge Procedure
- Comments on Bio-WAR Datasets
- Comparative Results of August Challenge
- Specific Tests Performed by Each Developer

Bio-Surveillance August Challenge

- Purpose of the "August Challenge"
  - Promote Integration of Program, Common Measure
    - Same Data Sets Used by All Groups
    - Improve BioWAR with Teams' Inputs
  - Focus on Common Data Set
    - Developers: Results Provide Foundation to Better Understand the Capabilities & Limitations of Their Own System
    - Simulators: Common Results Between Developers Shows Strengths of BioWAR Simulation & Directions for Growth
  - Develop a Common Set of Metrics to Assess Each Developer's Progress
  - Results Will Provide a Framework for Discussing the Needs of Year 2 Research
  - Disclaimer
    - Results Are on an Individual Basis – Not Competitive
    - Interpretations May Differ Even Using the Same Metrics

Outline

- Comments on August Challenge Procedure
- Comments on Bio-War Datasets
- Comparative Results of August Challenge
- Specific Tests Performed by Each Developer

BioWAR Focus on Spiral Development

- Developing a Common Set of Data for All Developers Is a Difficult Task Requiring Spiral Development
  - Initial Datasets Test General Concepts
  - Results of August Challenge Will Demonstrate BioWAR Data/System Strengths and Weaknesses to Test Developer's Approaches
  - Suggested Improvements in BioWAR Used As Basis for Continued Spiral Development
  - Results Help Developers Refine Approaches for Next Developmental Phase
- Benefits of Simulated Data to Developer's Research
  - Provides Repeatable Data Sets for Experiments and Consistent Performance Assessment
  - Can Be Generated to Include a Range of Conditions (e.g.., Small, Medium, Large Attacks)
  - Includes "Ground Truth" (I.e., An Understanding of the Attack/release)
  - Common Metrics Help to Evaluate Progress
- Simulated Data Is a Possible Stepping Stone for Algorithm Development

Strengths of BioWAR

- BioWAR Provided Foundation for Testing Each Team's Approach
- Datasets Provided a Common Framework for Evaluation
- Created an Atmosphere of Team Interaction and Deciding Common Procedures
  - Potential for Internal Infrastructure Development
    - Cross-validation Exercises Among Team Members
- Future Direction:
  - A Compilation of Real and Simulated Data Is Necessary When:
    - Real Data Cannot Provide Realistic Datasets of a Bio-attack
    - The Means to Evaluate Results Are Confounded With Unmeasurable Conditions (e.g., Flu, Allergies, Social Conditions)
    - Limitations on the Use of Historical Data Can Be Overcome.

Specific Comments on BioWAR
(Pitt / CMU)

- Some ICD-9 Codes Are Counterintuitive After the First Deaths
- Exotic ICD-9 Codes That Could Give Preliminary Clues to a Bio-agent Attack Were Spread Through Normal Data Leading to Higher False Positives
- Deaths Are Extremely Rare, Except During an Attack
- Results of Bio-WAR May Not Be Representative of Real World Data
  - ER Deaths Are a Prime Indicator
  - Background ER Deaths Were Generally Non-existent
  - Non-ER Data Does Not Seem to Give Indications Ahead of ER Deaths
  - Work Absentees Give Evidence of an Attack After ER Deaths Occurred
- Some Developer's Techniques Were Not Testable Since Bio-attacks Were Not Detectable in Advance Within Datasets Other Than ER
- Small Building Attacks Required Modification to Developer's Approaches
  - No Deaths Occurred Due To:
    - Anthrax Delivery on Weekends Had No Effect on Mondays; No Workers on Weekends
    - ICD-9, Absentee, OTC Datasets Gave No Corroborating Anthrax Indications

Specific Comments on BioWAR (IBM)

- General Datasets
  - ER Deaths Are Almost Entirely Noise Free
  - Only Work and School Absentees Were Sufficient to Indicate Anthrax Attacks
  - Phone, Pharmacy OTC Did Not Provide Appropriate Indications
- Site Datasets
  - Ignores ER Data
- Datasets Do Not Allow Testing of Meaningful Distinctions Between Developer's Approaches
  - Not Yet Ready to Provide Basis for Developing Insights Into Real World Bio-attacks
  - Many Relationships Between Data Types Seemed Unrealistic
  - Behavioral Models Between Data Types Were Cumbersome
    - Fixes for One Problem Within BioWAR Caused Unrealistic Conditions to Appear in Other Parts of the Simulation
- One Solution: Use Real Non-attack Data and Supplement With BioWAR Conditions of an Attack

Specific Comments on BioWAR (Veridian)

- Underlying Disease Models Not Validated by Domain Experts
  - Standard Epidemiological Parameters Were Not Made Available
  - Comprehensive List of Underlying Assumptions Not Available
    - Disease Spread
    - Mortality & Symptomatology
    - Disease Mappings to Symptoms, ICD-9 Codes, Illness Behavior Thresholds
- Simulation Disease Models Were Not Representative of Actual Bio-attack Conditions
  - Deaths Occur Only Due to Anthrax
  - ER Deaths and Insurance Claims After Deaths Contained Unlikely ICD-9 Codes (E.g. Viral Gastroenteritis)
  - Normal Models Exclusive of Background Anomalies (E.g. Flu) Are Necessary
  - Strong School and Work Absentee Counts Were Prevalent up to a Year After an Attack
- Input Assumptions, Files and Data That Described Their Approach Were Not Available

Specific Comments on BioWAR
(Johns Hopkins APL)

- Individual Data Streams Have Unrealistic Background
  - No Evidence That Models Interact Among Outside Factors
    - No Variability Due to Weather, Changes in Disease Strain
    - No Changes in Work Absenteeism Due to Vacations, etc
- Lack of Information on Spatial Dispersion
- Data Conditions Unrealistic
  - ICD-9 Code Deficiencies
  - Diagnoses Too Uniform
  - Unnatural School Enrollment Totals
  - Infection Distributions Not Reflective of Common Diagnoses
- Disproportionate Time Spent on Improving Simulator to a First-order Quality
  - Time Spent Could Have Been Spent on Authenticating Approaches on Real Data
- Exercise Demonstrated a "One Size Fits All" Approach to Attack Simulation

Outline

- Comments on August-Challenge Procedure
- Comments on BioWAR Datasets
- Comparative Results of August Challenge
- Specific Tests Performed by Each Developer

[redacted]

I.D

DARPA BSS Program Data Evaluation Comments from the Working Group

3 December 2002

1

Outline

- Types of data sources that have been evaluated
- For each
  - What validation has been performed?
  - How timely are they?
  - How sensitive and specific are they?
  - How difficult are they to obtain?
  - Do they make epidemiologic sense?
  - Are they worthwhile pursuing?
  - If so, how should they be approached data-wise?
  - What other studies need to be done?
- What other sources should be investigated?

2

Data Sources

- Outpatient visits
  - ER vs. all
  - ICD9 codes vs. text
- OTC sales
  - Drugs
  - Equipment
  - Electrolytes
- 911 calls
- Prescriptions
- Absenteeism
  - School
  - Work
- Work-based health surveys
- Medical website hits
- Calls to medical offices
- Poison control calls
- Animals
  - Domestic
  - Wildlife
  - Agribusiness
- Sewage
  - Flow
  - Contents

3

Outpatient visits

- JHU/APL and WRAIR
  - HMO data warehouses of ICD9 data in DC area
  - Military treatment facilities (MTFs) ICD9 data worldwide
  - ICD9s and CC text strings from ERs in DC area
- IBM
  - Insurance claims
  - Used for validation of other sources
- Verdian-Stanford
  - Norfolk area MTFs ICD9 codes
- Mellon-Pitt
  - ED registration in PA and UT
    - Both ICD9s and CC free text

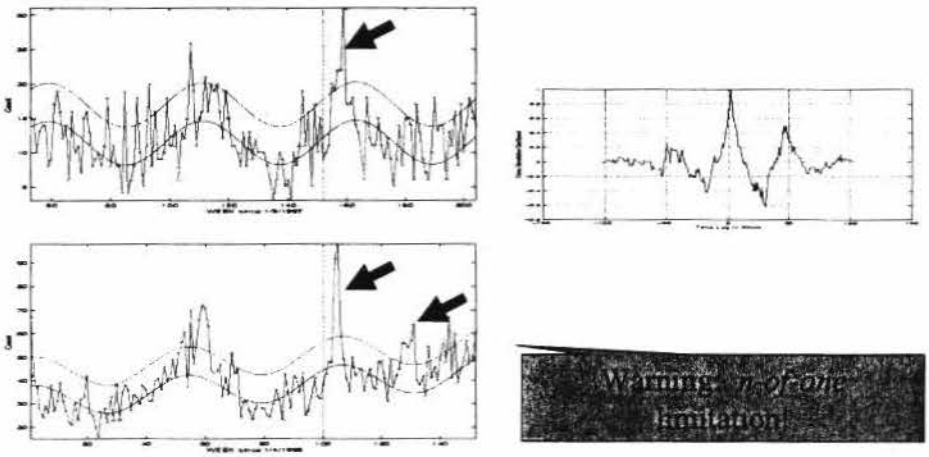
4

Issues

- We know these are good data sources and are predictive
- How timely? Varies, can be fast, may cost more
- How good is the NLP?
- What is the difference between ER visits and other physician office visits?
- Is it better to look at counts or % of totals?

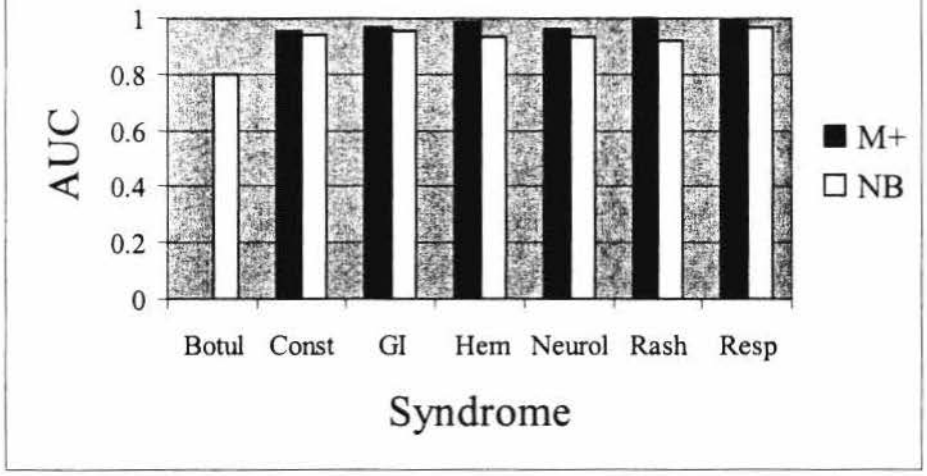
5

MP: ED Chief Complaints (ICD-9 coded) 2 Weeks Earlier than Pneumonia and Flu Deaths



Tsui, Wagner et al. *JAMIA Symposium Supplement* 2001, and to appear JAMIA 2002

MP: Accuracy of Natural Language Parsing of Chief Complaints into Syndrome Groups



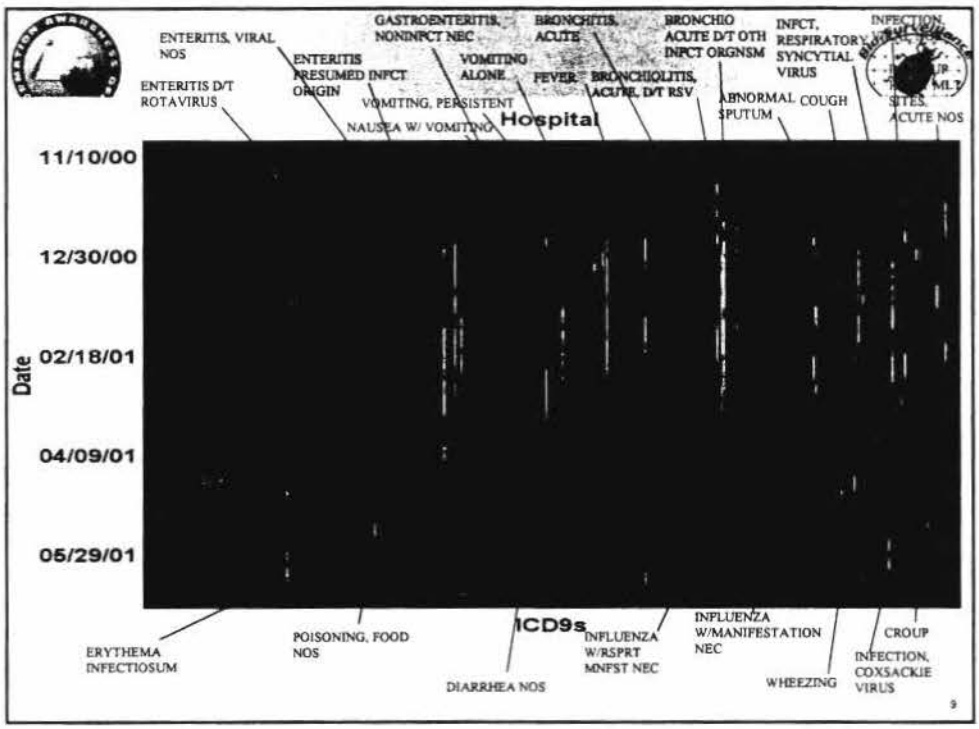
* Legend: AUC, area under the ROC curve; M+, a probabilistic NLP algorithms; NB, Naïve Bayes, a second probabilistic NLP algorithm. N.B. There were no Botulinic test cases for M+

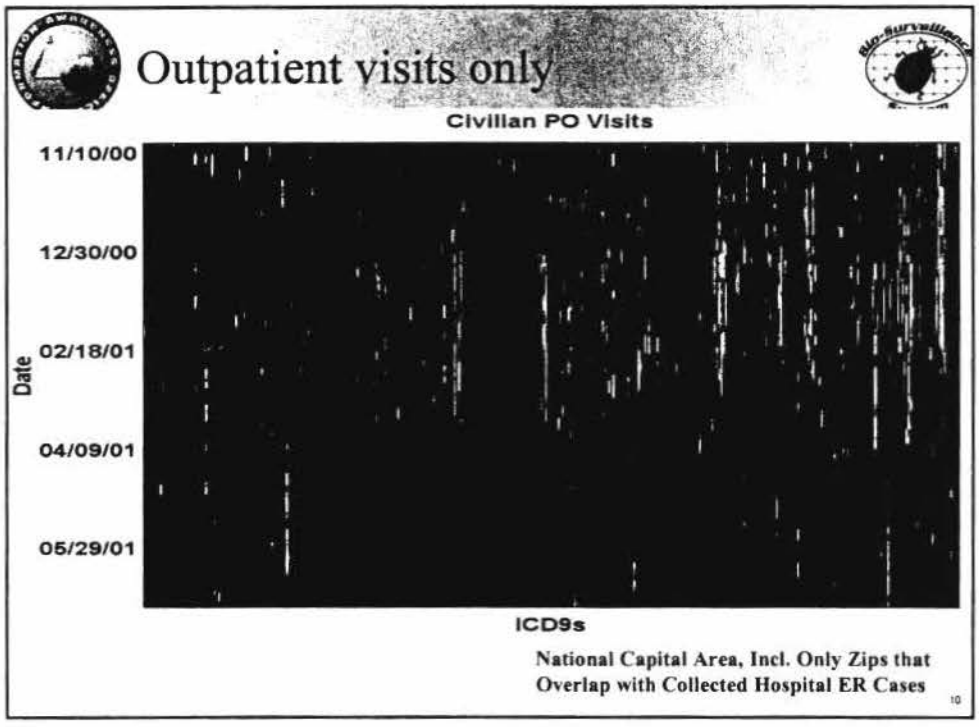
7

MP: Accuracy of NLP for Detecting Cases of Acute Infectious GI Disorder fm Free-text Chief Complaints

	ICD-9	Chief Complaint
Sensitivity	**0.32**	**0.50**
Specificity	**0.99**	**0.92**
PPV	**0.37**	**0.09**
NPV	**0.99**	**0.99**
AUC	**N/A**	**0.75**

8


ENTERITIS, VIRAL NOS
ENTERITIS D/T ROTAVIRUS
ENTERITIS PRESUMED INFCT ORIGIN
GASTROENTERITIS, NONINFCT NEC
VOMITING ALONE
VOMITING, PERSISTENT
NAUSEA W/ VOMITING
BRONCHITIS, ACUTE
FEVER
BRONCHIOLITIS, ACUTE, D/T RSV
BRONCHO ACUTE D/T OTH INFCT ORGNSM
ABNORMAL SPUTUM
INFCT, RESPIRATORY SYNCYTIAL VIRUS
COUGH
SITES, ACUTE NOS
Hospital
11/10/00
12/30/00
02/18/01
04/09/01
05/29/01
Date
ERYTHEMA INFECTIOSUM
POISONING, FOOD NOS
DIARRHEA NOS
INFLUENZA W/RSPRT MNFST NEC
INFLUENZA W/MANIFESTATION NEC
WHEEZING
CROUP
INFECTION, COXSACKIE VIRUS
ICD9s
9


Outpatient visits only
Civilian PO Visits
11/10/00
12/30/00
02/18/01
04/09/01
05/29/01
Date
ICD9s
National Capital Area, Incl. Only Zips that Overlap with Collected Hospital ER Cases
10

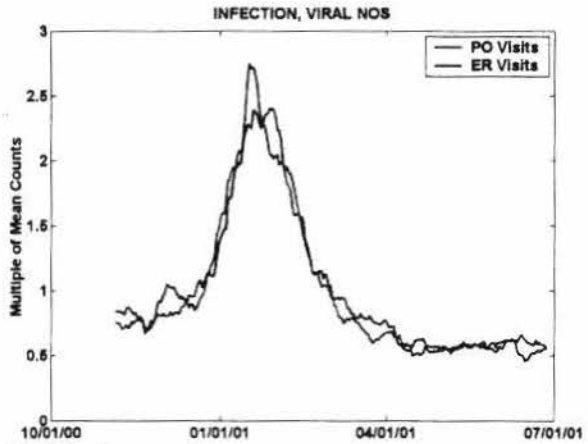
APL: Observed ER- Civilian Outpatient Visit Correlation by Syndrome Group

Syndrome	FLU-Season	Non-FLU Season
URI	0.9	(Insufficient Data)
LRI	0.7-0.9	.7
LGI	0.7-0.9	<.5
Fever	<.6	.6

Data Include 2 FLU Seasons, 3 Regions

11

APL: ER vs. Office Visits



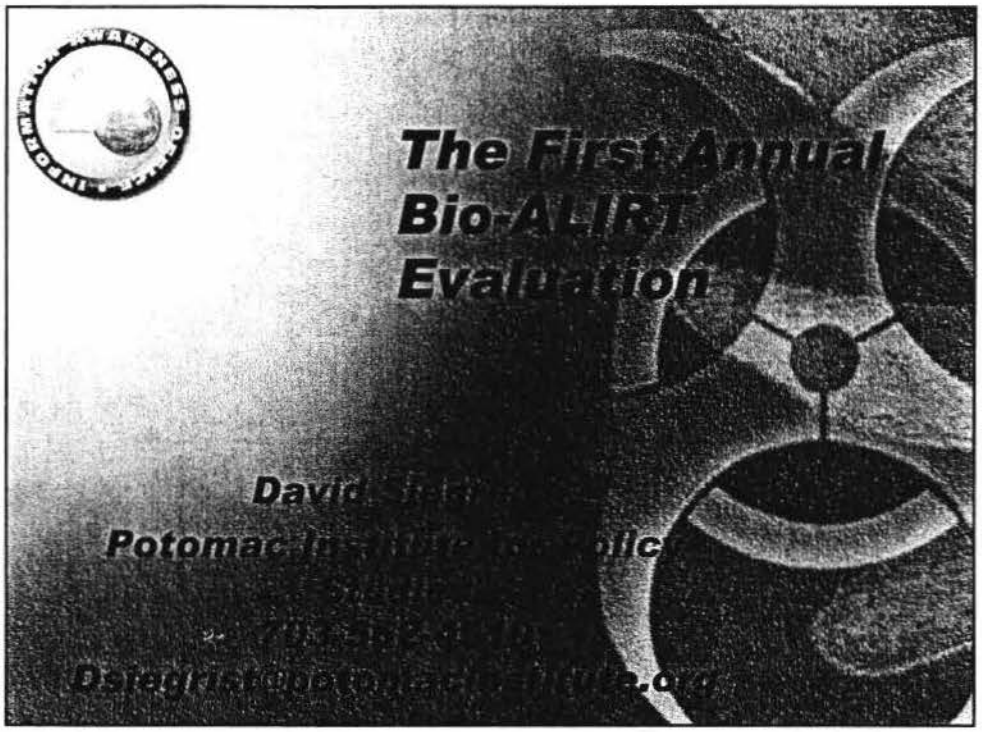
Mean Weekly Counts:

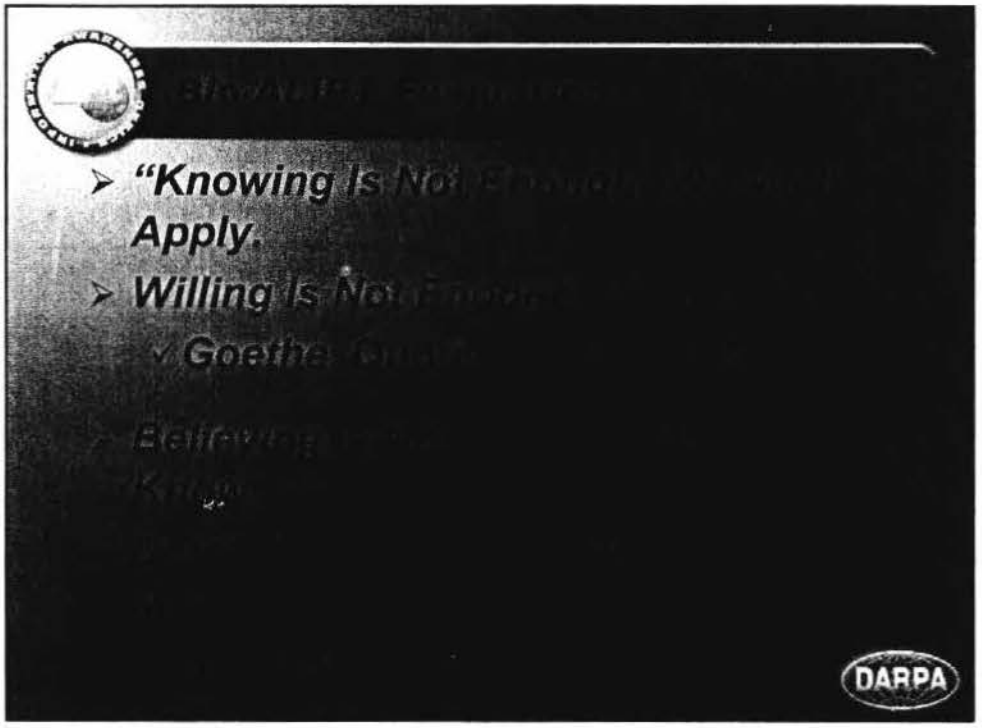
PO Visits: 1001 ER Visits: 492
Correlation: 0.98 Lag of PO Visits: 1 day

12

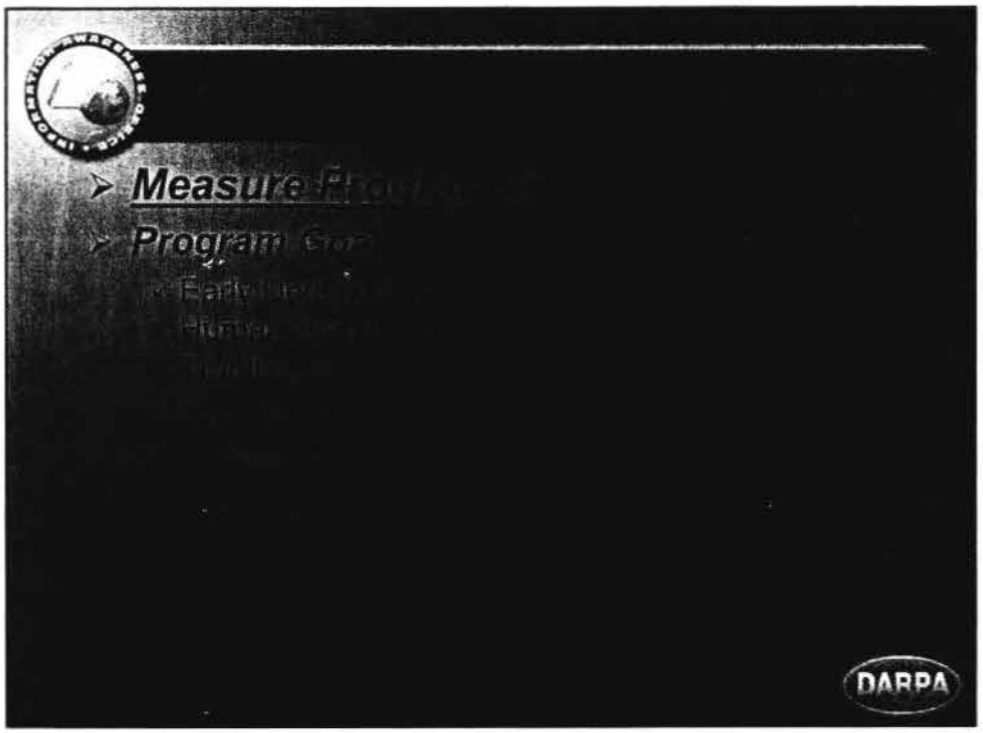
[redacted]

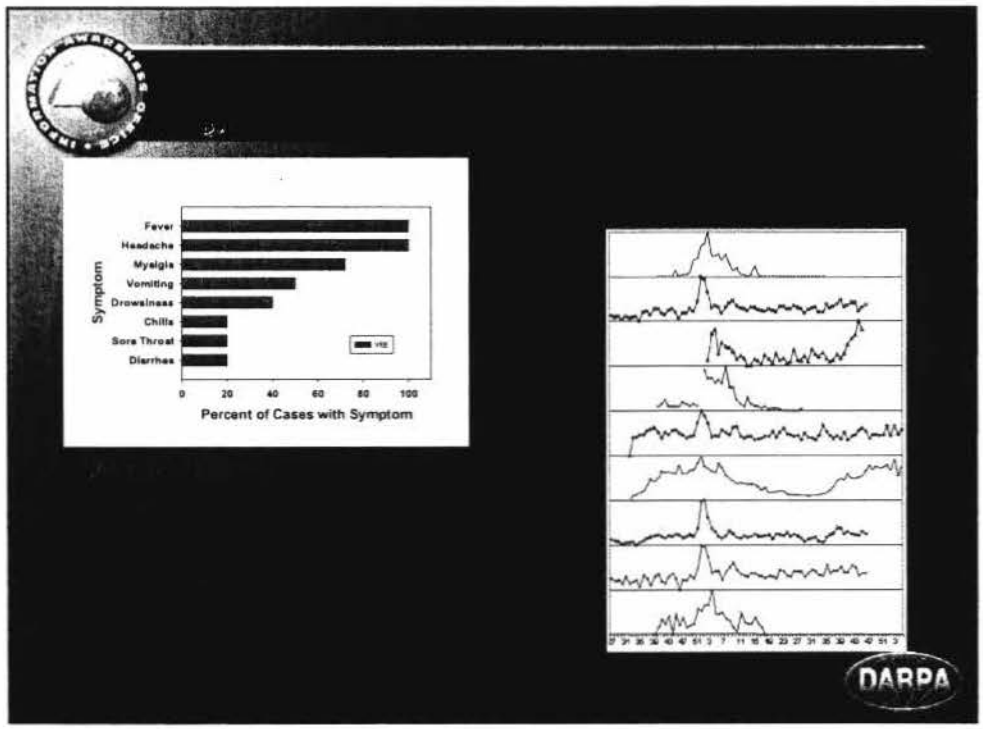
1 of E

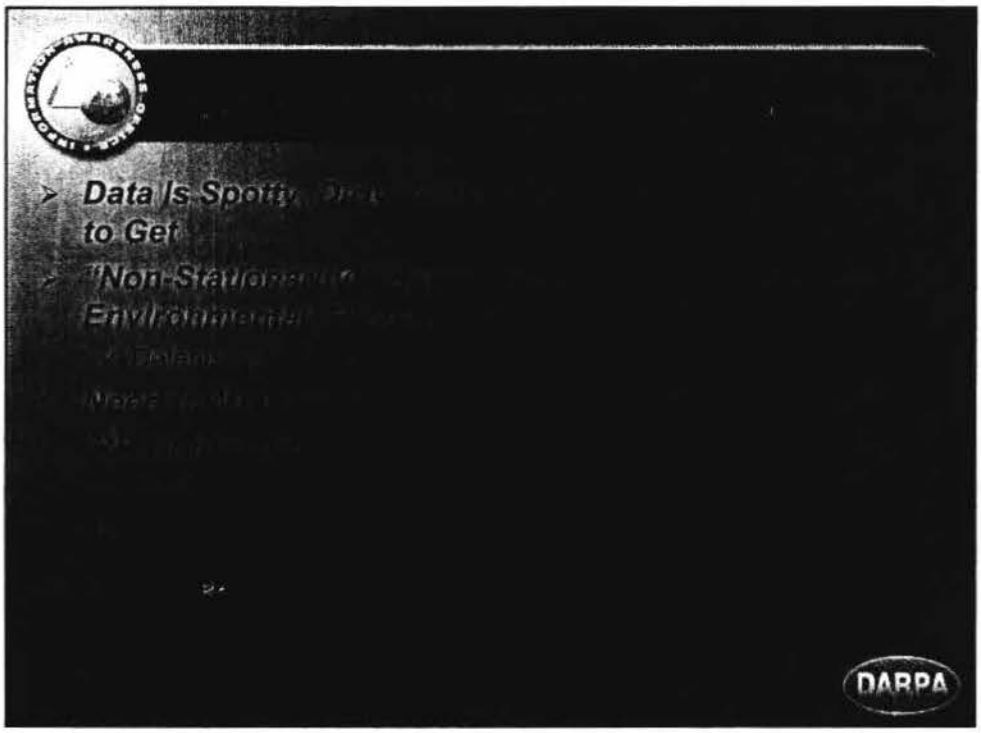




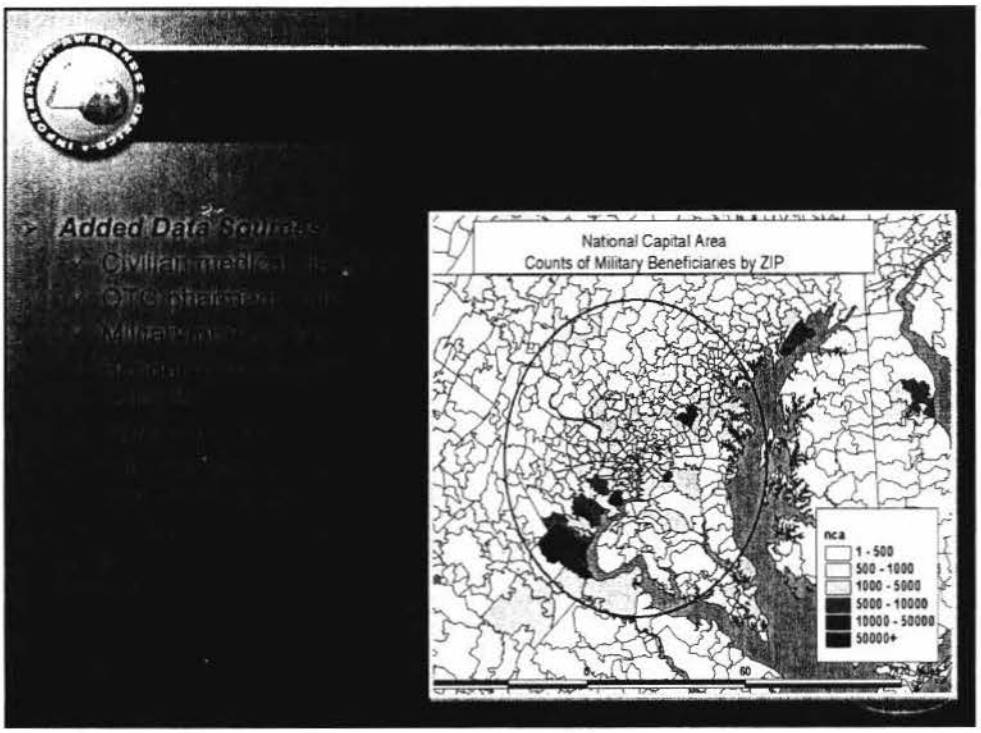
BEST COPY AVAILABLE

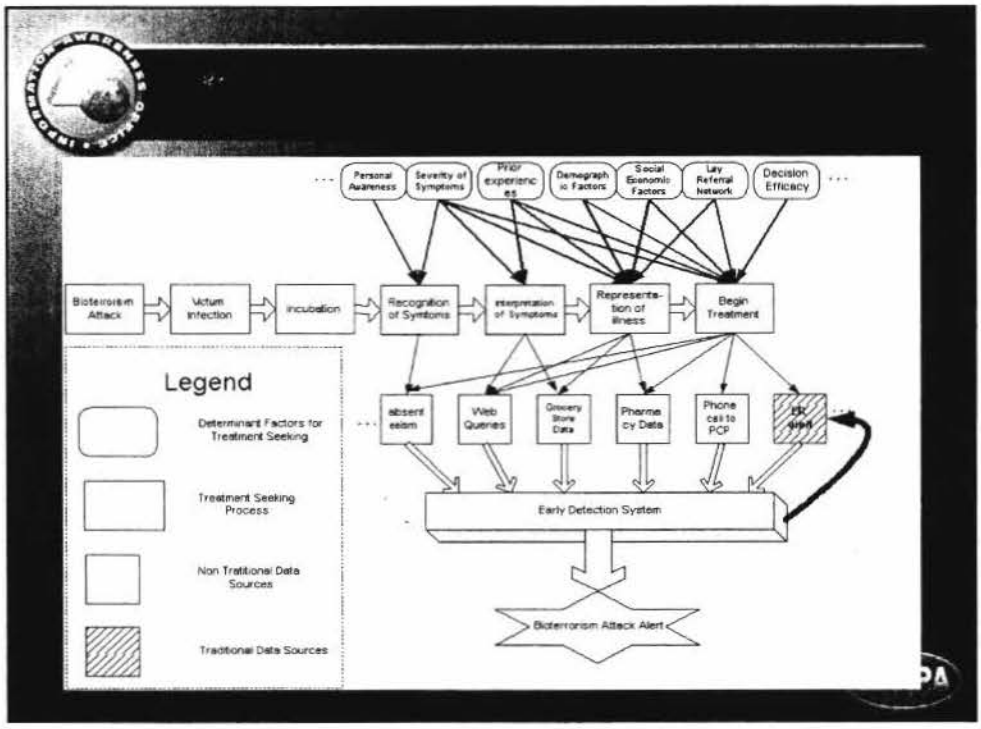

Measure
Program
DARPA

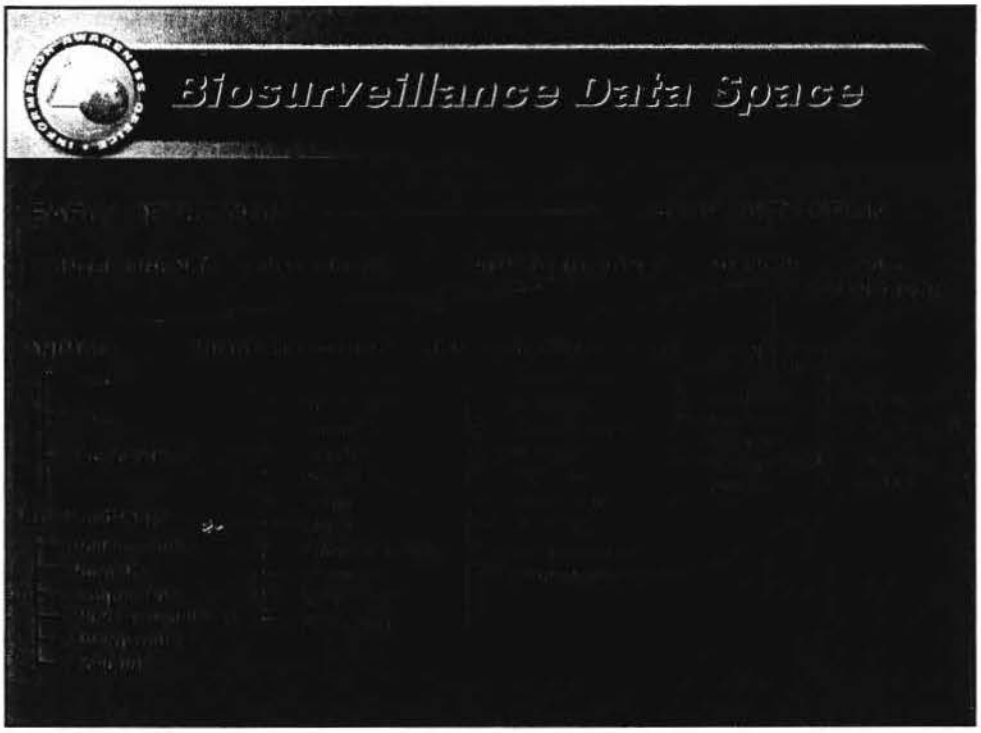

Fever
Headache
Myalgia
Vomiting
Drowsiness
Chills
Sore Throat
Diarrhea
Symptom
0
20
40
60
80
100
Percent of Cases with Symptom
DARPA


DARPA


DARPA


Added Data Sources
National Capital Area
Counts of Military Beneficiaries by ZIP
nca
1 - 500
500 - 1000
1000 - 5000
5000 - 10000
10000 - 50000
50000+


Personal Awareness
Severity of Symptoms
Prior experienc es
Demograph ic Factors
Social Economic Factors
Lay Referral Network
Decision Efficacy
Bioterrorism Attack
Victim Infection
Incubation
Recognition of Symptoms
Interpretation of Symptoms
Representa- tion of Illness
Begin Treatment
Legend
Determinant Factors for Treatment Seeking
Treatment Seeking Process
Non Traditional Data Sources
Traditional Data Sources
absent eeism
Web Queries
Grocery Store Data
Pharma cy Data
Phone call to PCP
Early Detection System
Bioterrorism Attack Alert



100 Data Sources Developed

Non-Traditional Data Sources

Data Source	Organization	Issues	Utility	Frequency of Collection	Latency	Temporal Resolution	Geographic localization / granularity	Status	IRB?	Personal ID
IBM										
Health survey: Regularly poll a population, asking individuals to self assess their health	IBM Watson Research Lab	Survey must be organized on a site/regional basis. Modeling missing responses. Time to establish baseline.	High, particularly in cases of heightened alert. Provides some specificity if symptoms are queried when ill. Sensitivity controllable by sample size.	Real-time	Symptoms showing on a number of persons in survey area. Evidence that survey precedes doctor visits by 3 days	Second	Site/Regional. Can be adjusted as needed	Initial survey complete. 400 individuals out of 2200 participated. were polled weekly via email	Yes at Watson, approved Jan 2002	Yes, but we anonymized
Site attendance: Electronically recorded from badge entry	IBM Watson Research Lab	Not all sites record attendance electronically. Affected by external factors.	Low for mild illness. Moderate for serious illness	Currently daily, potentially real-time	Symptoms showing on a number of persons at site	Second	Site	Data collection ongoing	Not required, but Watson HR, Legal approval needed	Yes, but we anonymized
Proxy server logs: Outgoing accesses to medically-related websites	IBM Watson Research Lab	Not all sites use a proxy server. Identifying medically related accesses.	Moderate if applicable and sufficient volume. May provide some specificity if URL is tracked	Currently monthly, potentially real-time	Symptoms showing on a number of persons at site	Second	Site	Four month sample collected	Not required, but Watson HR, Legal approval needed	Yes, but we anonymized
Phone calls: Outgoing calls to medically-related numbers	IBM Watson Research Lab	Identifying and maintaining list of medically-related phone numbers	Moderate if applicable and sufficient volume	Currently daily, potentially real-time	Symptoms showing on a number of persons at site. Evidence that med. rel. phone calls precede doctor visits by 3-4 days	Second	Site	Data collection ongoing	Not required, but Watson HR, Legal approval needed	Yes, but we anonymized

Data Source

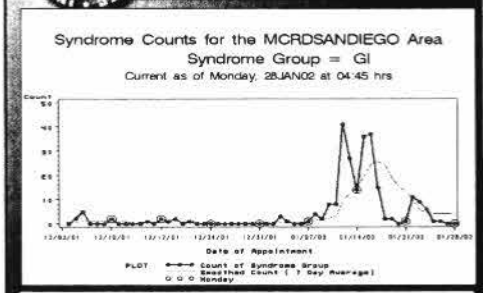
Survey Data

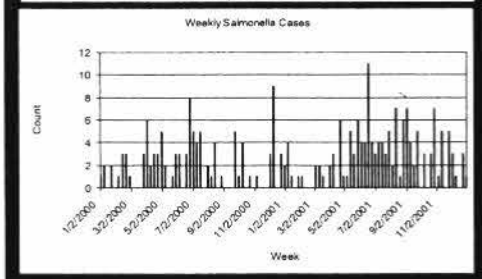
Medical Data

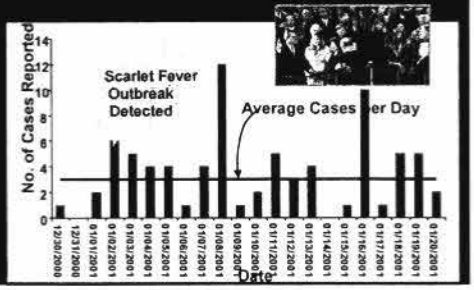
DARPA

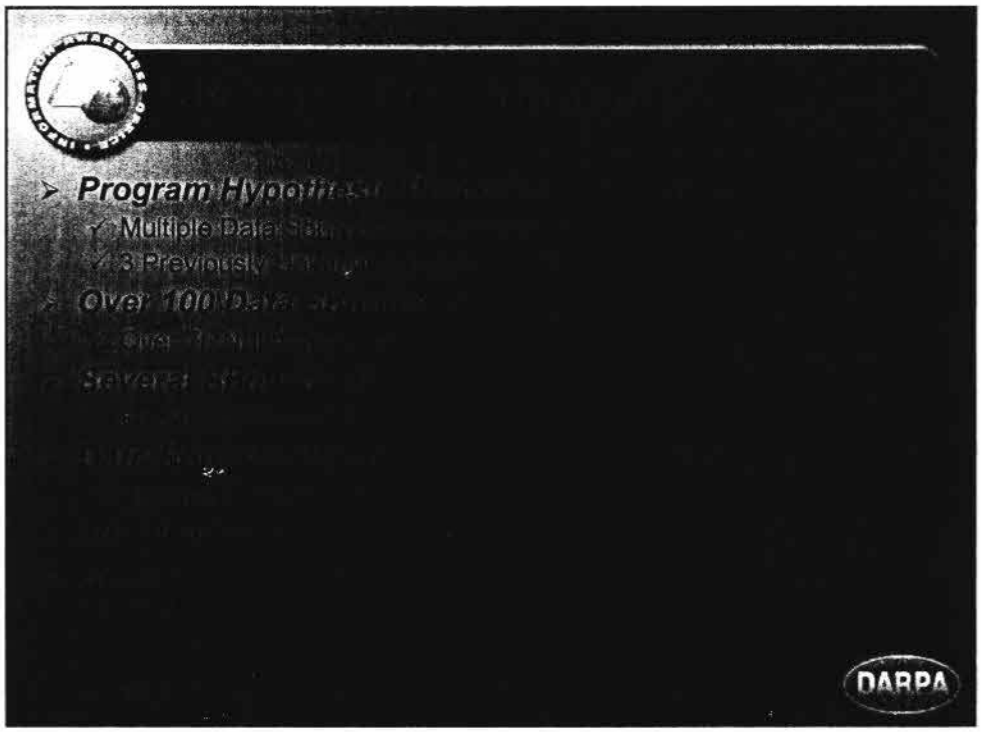
Three Previously Unknown Outbreaks Detected

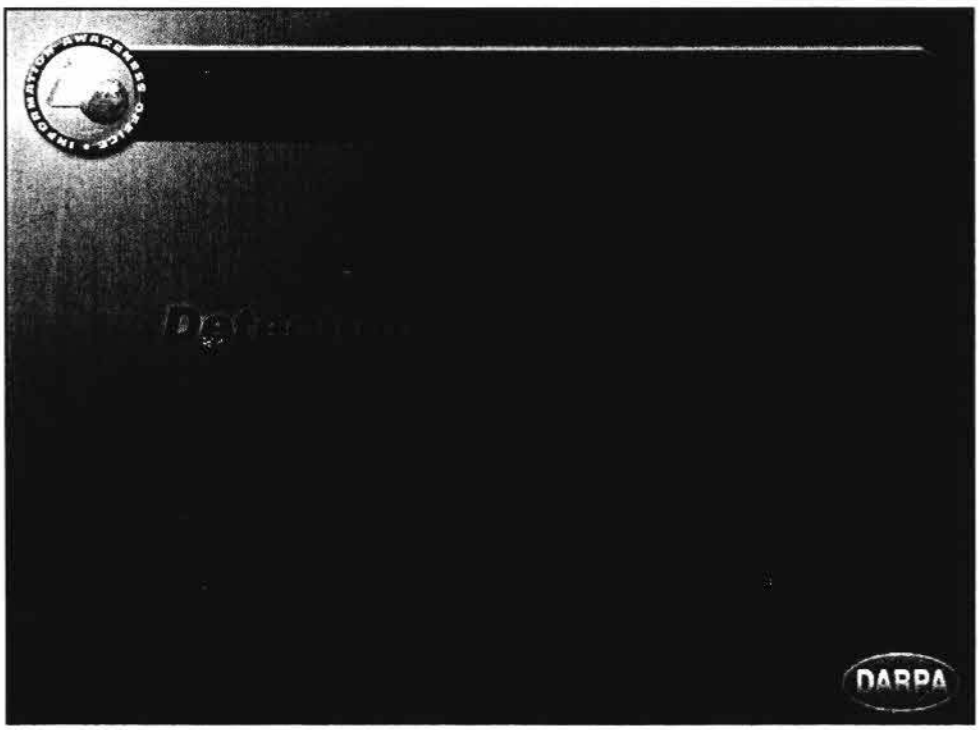
- **ESSENCE II Detected Scarlet Fever in NCR Using HMO Data**
- **GEIS Detected Norwalk Virus in MCRDSD Using ICD-9 Data**
- **PittMellon Detected Salmonella in Pittsburgh Using Lab Result Data**



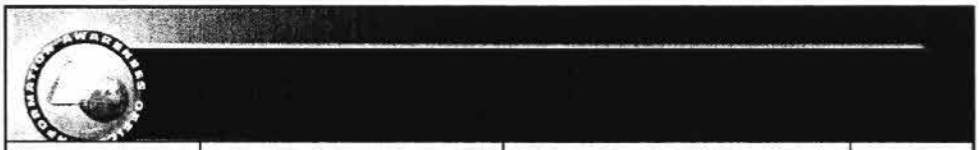





DARPA

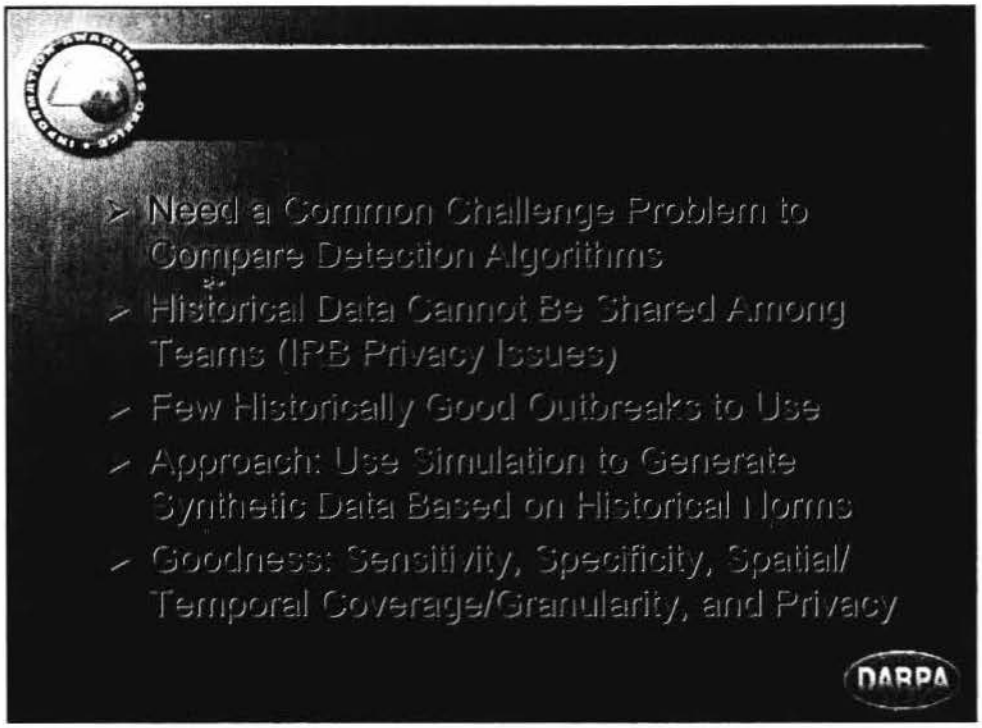

DARPA

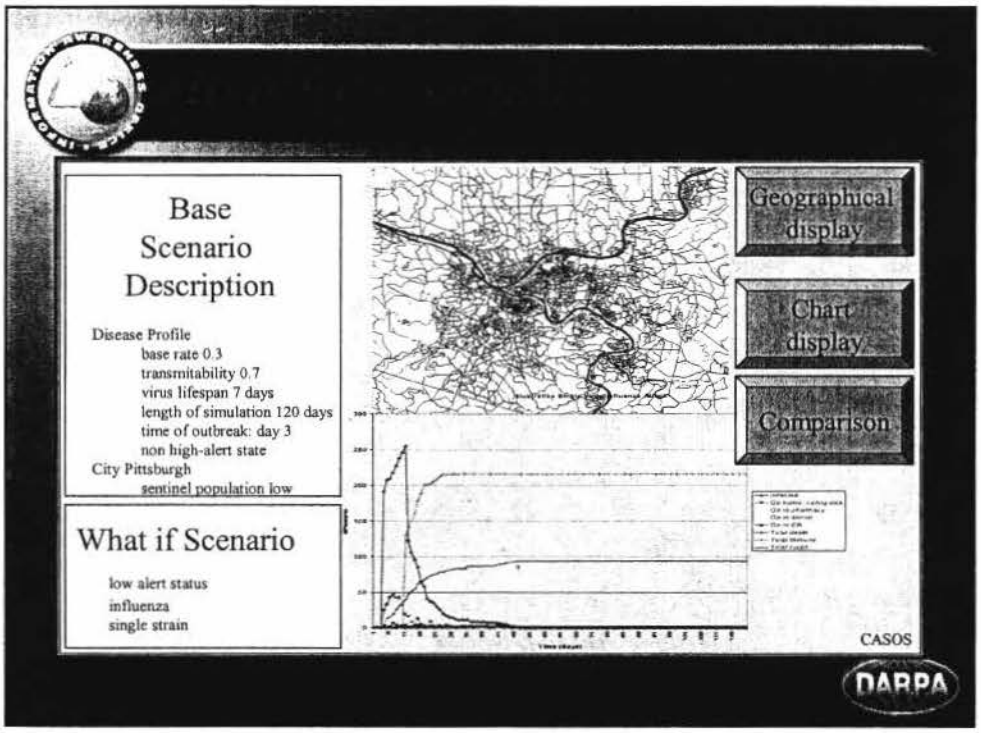
Fusion Method	Advantages	Disadvantages	Any Data type?	Teams tried it at all
Veridian/Stanford				
General Methods			Types: Behavioral – B; Non-Traditional Medical – NTM; Traditional Medical – TM	
Full Vector Time Series Forecasting: Vector State-Space Structural Models or Vector ARIMA (both using Kalman Filter recursive schema)	Most Natural "theoretical" Characterization of serially correlated heterogeneous data	Many parameters (full covariance matrices for structural model disturbance processes), Parameter estimation hard, requires large datasets	B, NTM, TM	No
Hybrid: Univariate forecasts on structural time-series models applied to multiple data streams (using Kalman Filter) with multivariate inference on post-forecast residuals	Can apply Classical Multivariate inferential tests for extremity and/or shift in covariance structure of vector gaussian I.I.D residuals	Post-forecast residual covariance matrices must be estimated from short streams of possibly "gappy" data; *this is true of any multivariate inferential technique*	B, NTM, TM	
Principle Components analysis on covariance structure of post-forecast residuals: Used as a form of dynamic cluster analysis	Reduction of dimensionality of multiple data source post-forecast residuals; Monitors correlated clusters of sources; this is the most natural technique to track changes in relationships between multiple sources;	Principle components are not invariant with data-scaling. Well-known problem of choosing the most "natural" units or non-dimensionalization when dealing with heterogeneous data types	B, NTM, TM	
Space, time, and space-time Cluster Analysis	Covariance structure of vector forecast residuals can be evaluated in light of special, *stronger*, "null" hypotheses that stand upon reasonable (e.g. Poisson or locally constant Poisson models) statistical models for the dispersion of these indicators in	Requires some validated structural model parameterizing joint statistical distributions, of various demographic observables as function of distance between sites	B, NTM, TM	

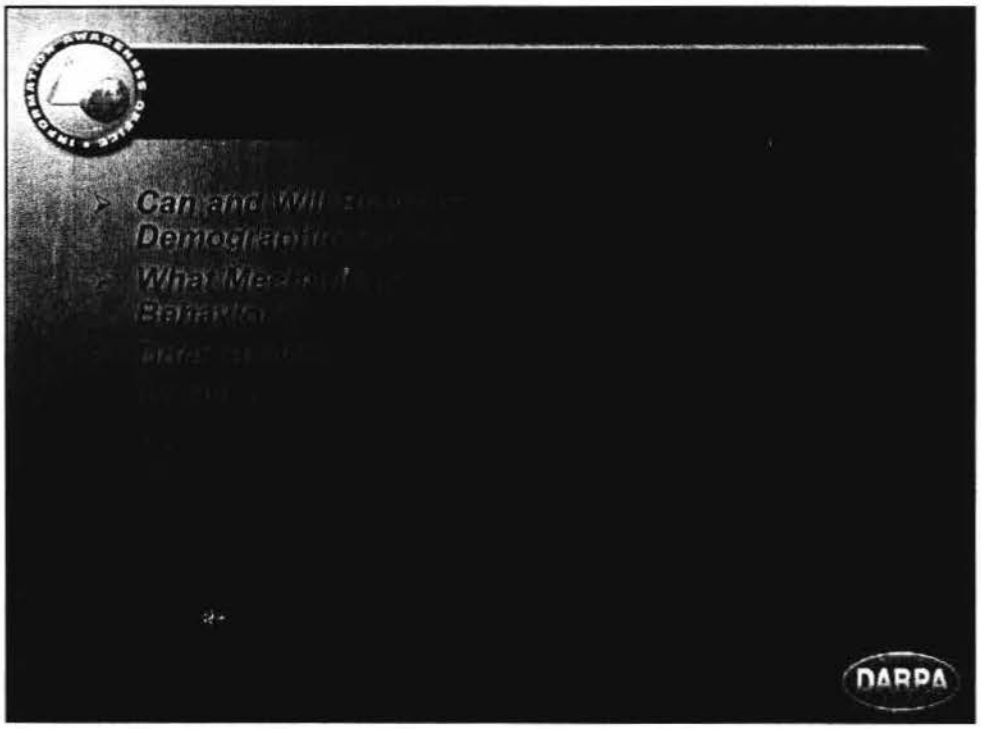


Detection Method	Advantages	Disadvantages	Tried and used
Veridian/Stanford			
Structural Time Series Model-Based forecasting Using Kalman Filter Recursion	Most general Schema for modelling non-stationary univariate (or multivariate) random processes; All linear models (e.g. ARIMA, SARIMA, etc can be cast as unobserved component state-space forms)	Unobserved component variances are free parameters which must be fit by maximum likelihood estimation on "training set". The physical meaning of these decomposition terms may be obscure; and hence impossible to estimate independently	
Control Chart (C-Chart)	Simple to implement; No technical obscurities; Forecasting system to assess deviation of sequence of counts from conditions of stability; assumption of Poisson generating process as "null" hypothesis; Poisson model can be updated "locally"	Simple explicit non-recursive development of state estimate from finite historical baseline; requires access to recent historical data set; Strong ("Poisson") distributional assumption on data (counts of rare events in large field of opportunity).	Veridian
CuSUM	Can be used to assess the alleged lack of serial correlation in the sequence forecast residuals: an indication of structural change; i.e. of "model lack of fit"	Not as powerful an indicator of structural change as certain forms of "switching" structural models.	
Randomization			
2x2 for Time Series	Track relative frequencies among multiple categories epoch by epoch; Adaptable to various categorical data types; Non-dimensional; Tolerates incomplete baseline and partial reporting	-	Veridian
2xN for Spatial Change	Compare relative apportioning of events (e.g. clinic visits) among various (N) categories epoch by epoch; this is a test for abrupt change in apportionment. Adaptable to various data types; Tolerates incomplete baseline and partial reporting	change not spatially localized – spatial structure of the apportionment of events is implicit by virtue of fixed assignments of the N categories to specific site locations.	
Spatial Scan Statistic	locates significant clustering in both space and time	computationally intensive	


Selected Developmental Algorithms
PRODUCTS
WSARE v1, 2, 3
DARPA


Need a Common Challenge Problem to Compare Detection Algorithms
Historical Data Cannot Be Shared Among Teams (IRB Privacy Issues)
Few Historically Good Outbreaks to Use
Approach: Use Simulation to Generate Synthetic Data Based on Historical Norms
Goodness: Sensitivity, Specificity, Spatial/Temporal Coverage/Granularity, and Privacy
DARPA


Base
Scenario
Description
Disease Profile
base rate 0.3
transmitability 0.7
virus lifespan 7 days
length of simulation 120 days
time of outbreak: day 3
non high-alert state
City Pittsburgh
sentinel population low
What if Scenario
low alert status
influenza
single strain
Geographical display
Chart display
Comparison
CASOS
DARPA

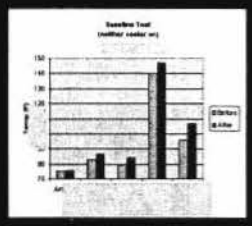

DARPA

Detection Evaluation Progression
Primary Program Goal: Early Detection

- **Aug, 2002 Evaluation.**
  - ✓ 1. Start with Early Detection of Large Area Aerosol Anthrax Release
    - Canonical Scenario with Highest Impact, Potential for Detection
    - Stresses Anomaly Detection and Fusion Algorithms
    - Use Largely Simulated Data Superimposed on Historical Background Rates
  - ✓ 2. Additional Evaluation: Smaller Size Attack Stressing Granularity
  - ✓ Non-Traditional Data Tested as Well
    - Needs Show Empirical Relation to Disease & Realistic Simulation Before With and W/O Evaluation
  - ✓ Sensitivity vs Specificity to be Measured on Multiple Runs

DARPA

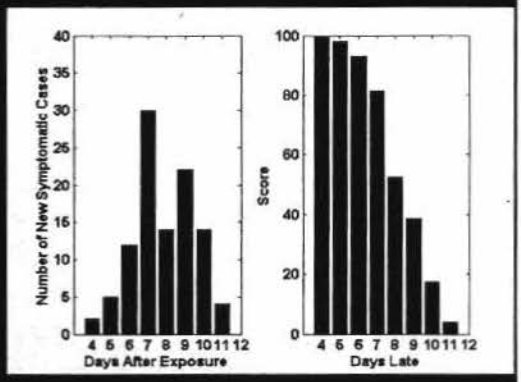
- ***Common Anthrax Scenario with Significant Aerosol Source***
  - ✓ Data Includes "Near Misses" as Well as Actual Outbreaks. FPs Detract Significantly.
- ***In Evaluation, Performers Make the "Call," Asserting Outbreak***
  - ✓ In Operation, Earlier "Hint" Will Be Used to Signal Epidemiologist
  - ✓ Outbreak Determination "Call" Is Then Assessed for Accuracy and Timeliness
    - [illegible]



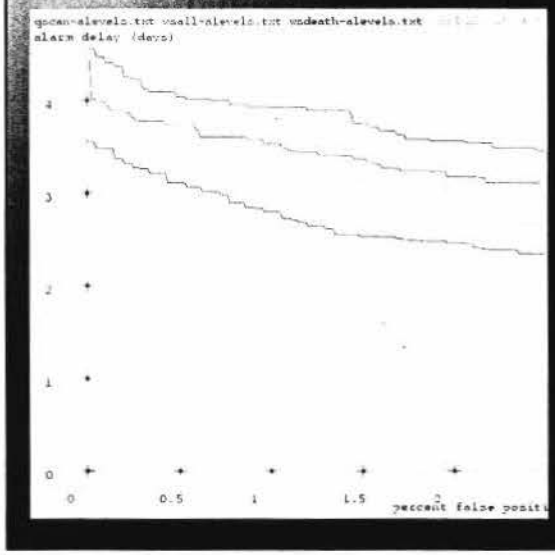
DARPA

Beyond ROC: More Specific Measures

- From Knowledge
- Data Mining
- Operating Char
- (Fawcett &
- Repeat

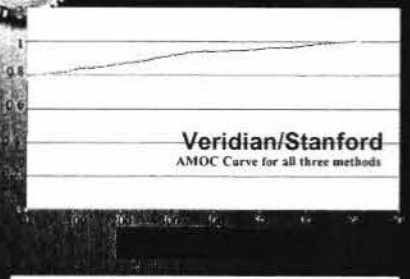


DARPA

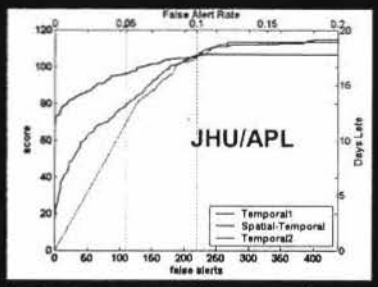


DARPA

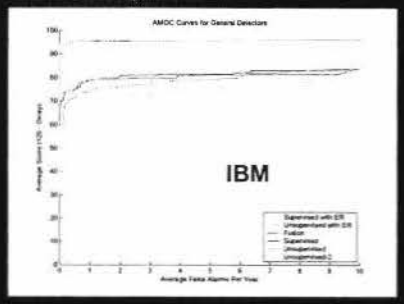
Algorithm Challenge Summaries



Veridian/Stanford
AMOC Curve for all three methods



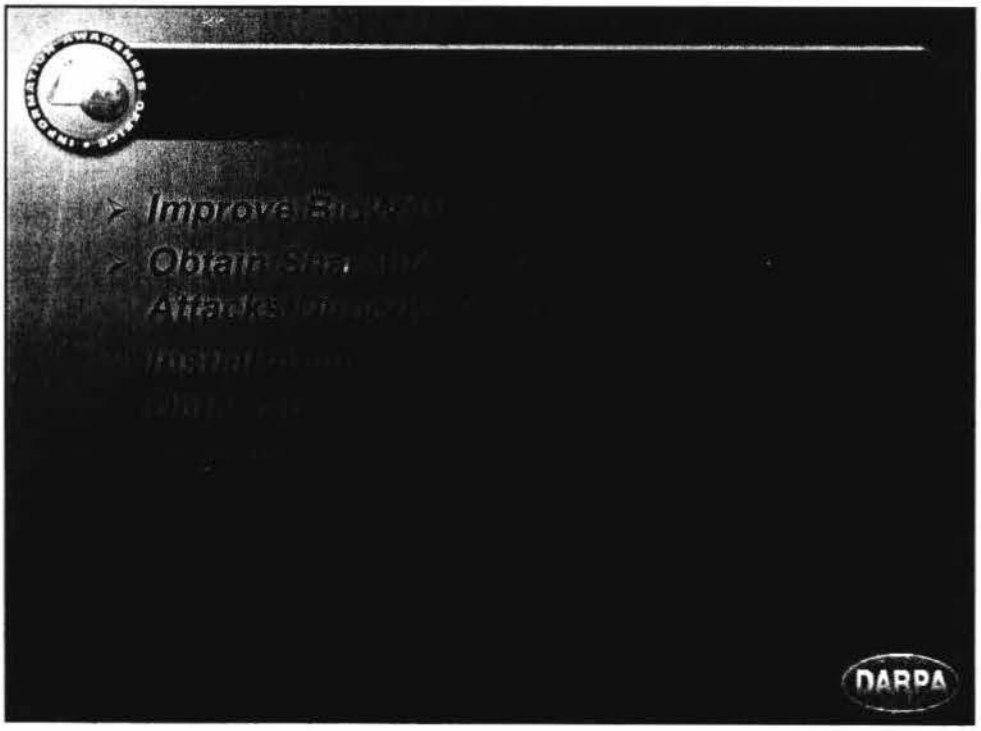
JHU/APL



IBM

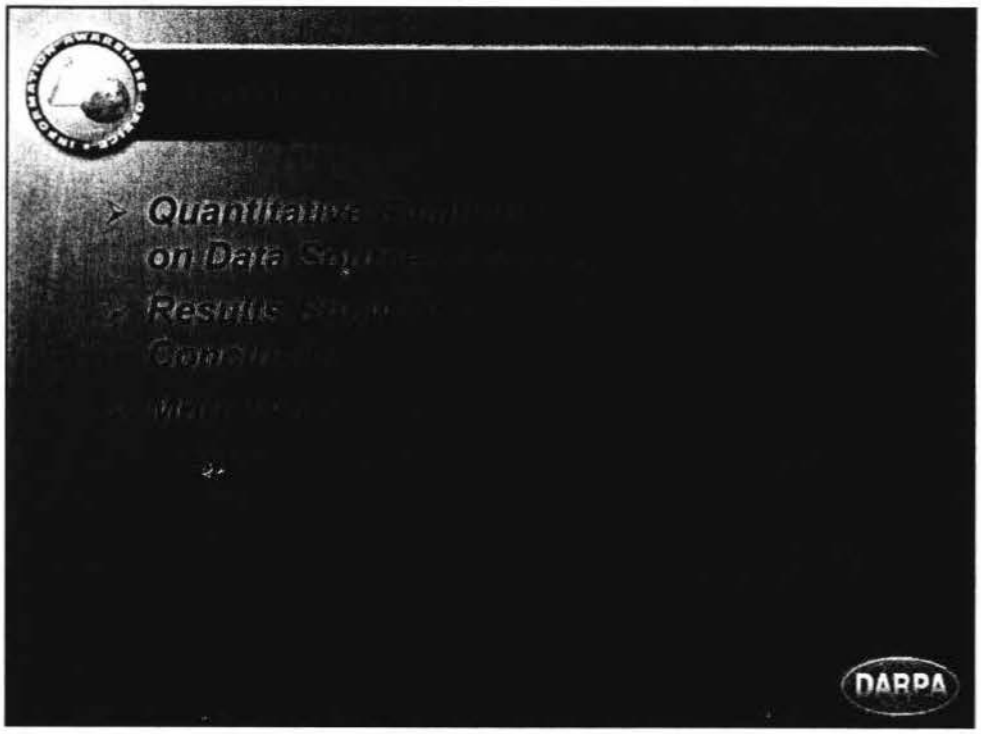
Pitt/Mellon

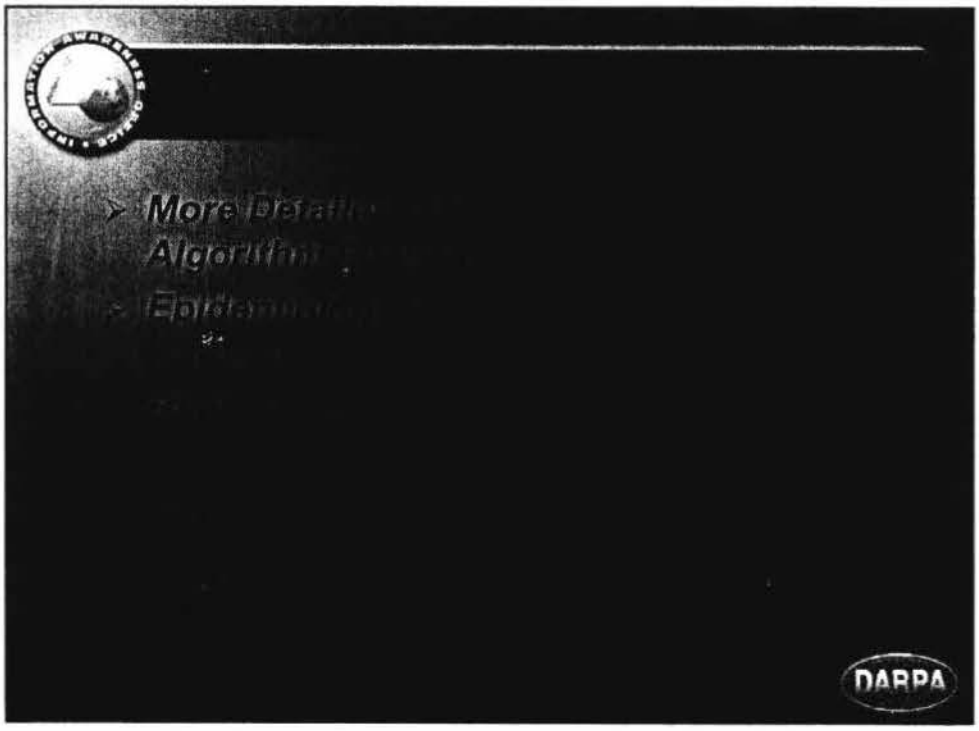
DARPA



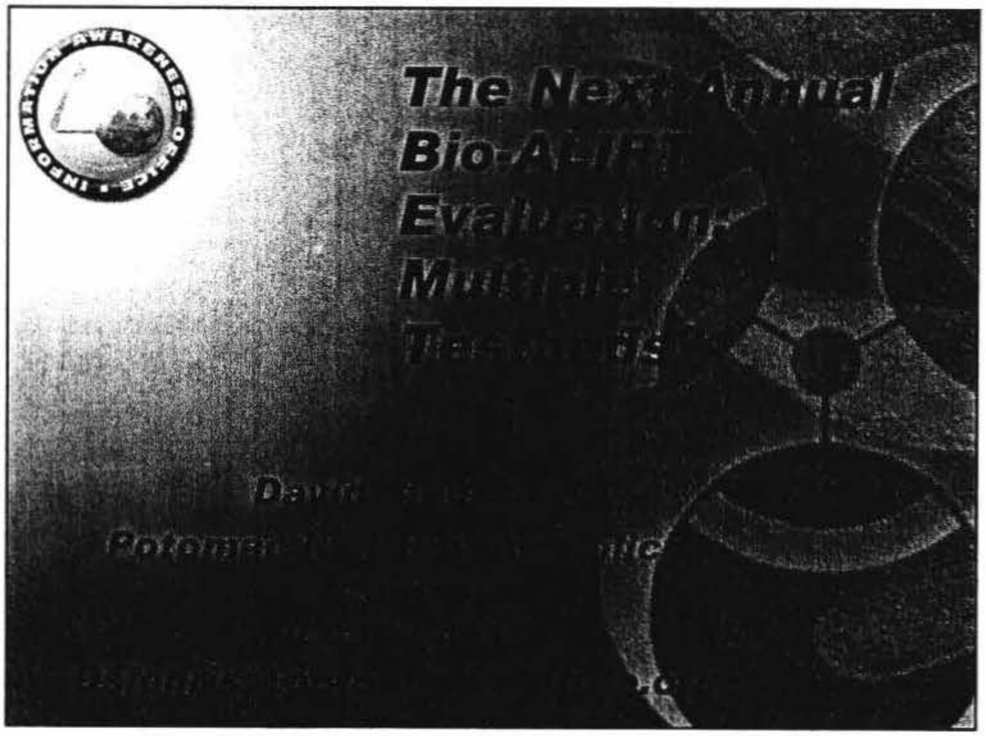
- Improve
- Obtain
- Attacks

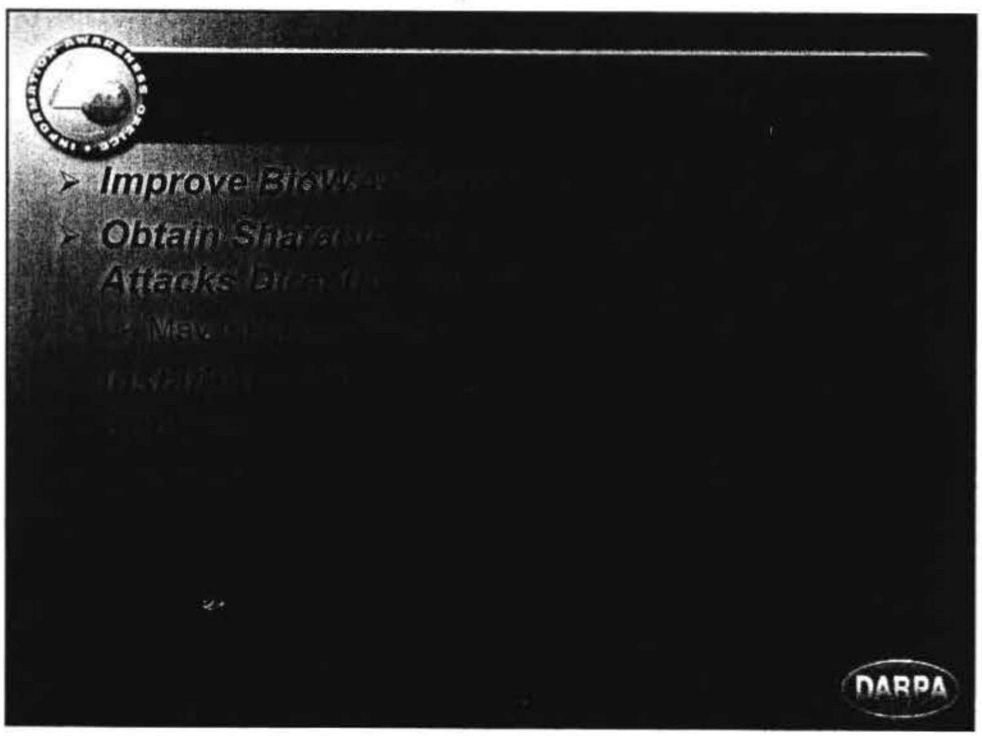
DARPA

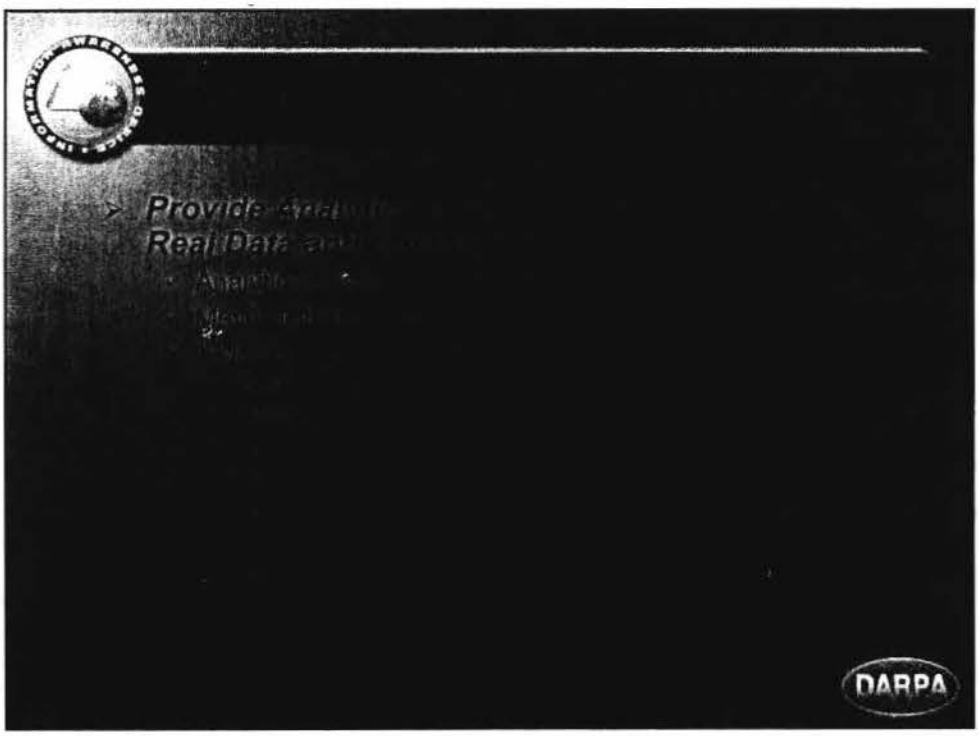

DARPA

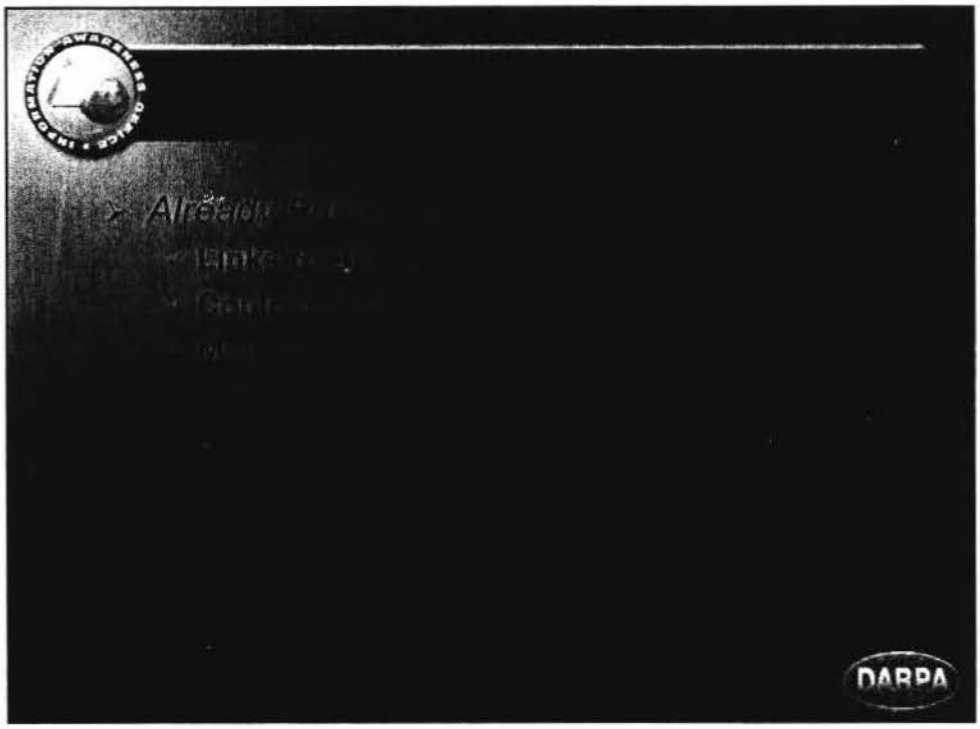

DARPA

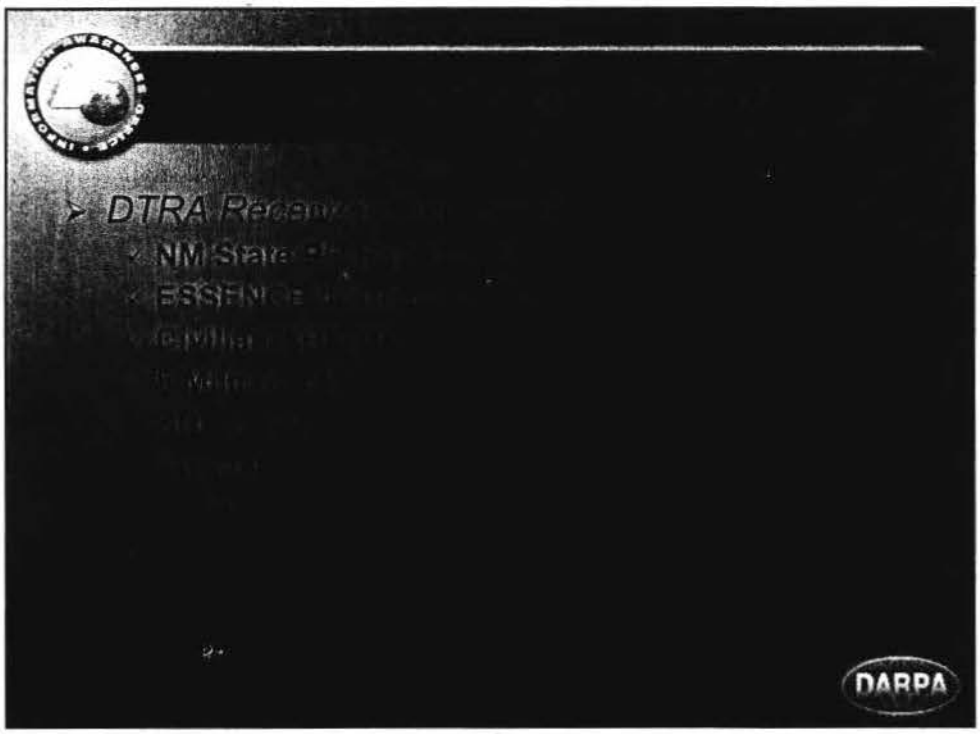
1. F 7

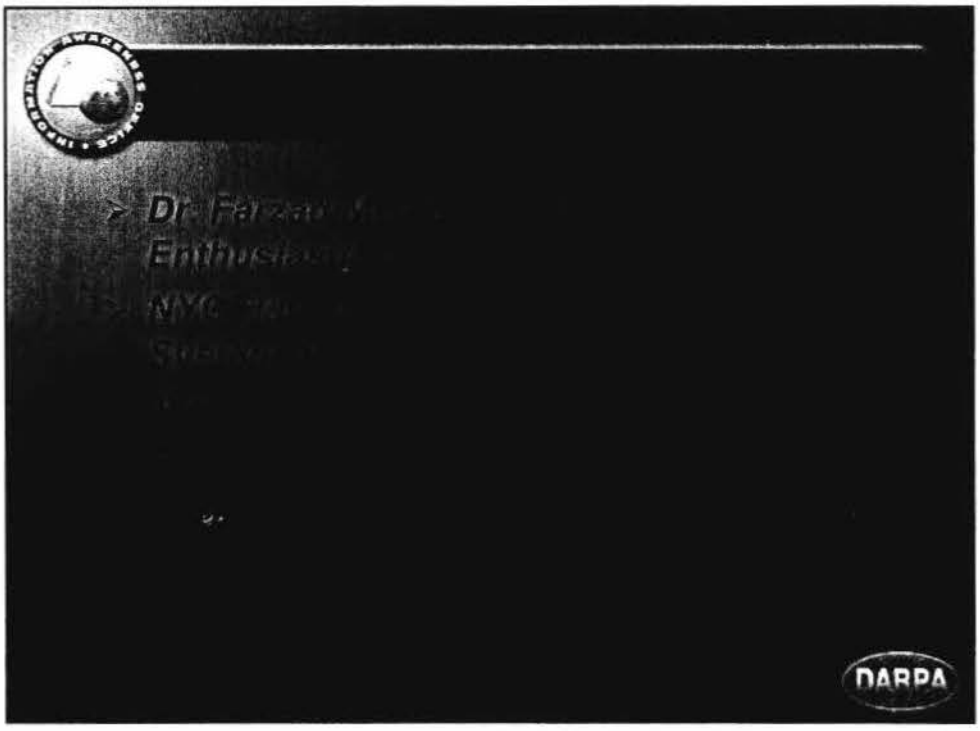

The Next Annual
INFORMATION AWARENESS OFFICE

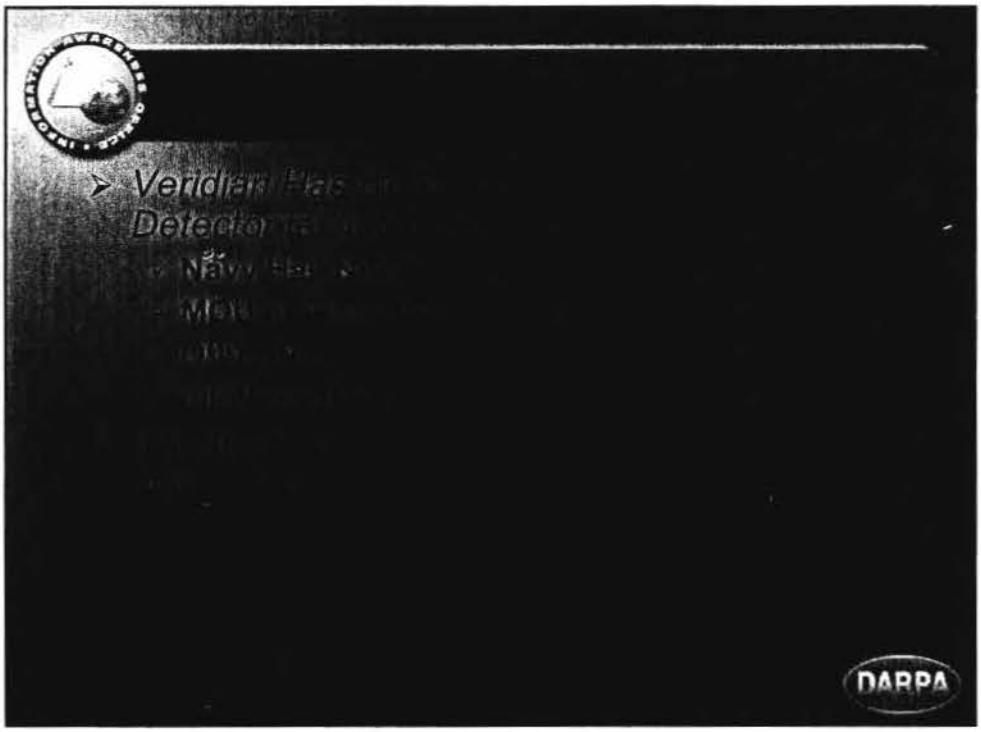

INFORMATION AWARENESS OFFICE
Improve
Obtain
Attacks
DARPA

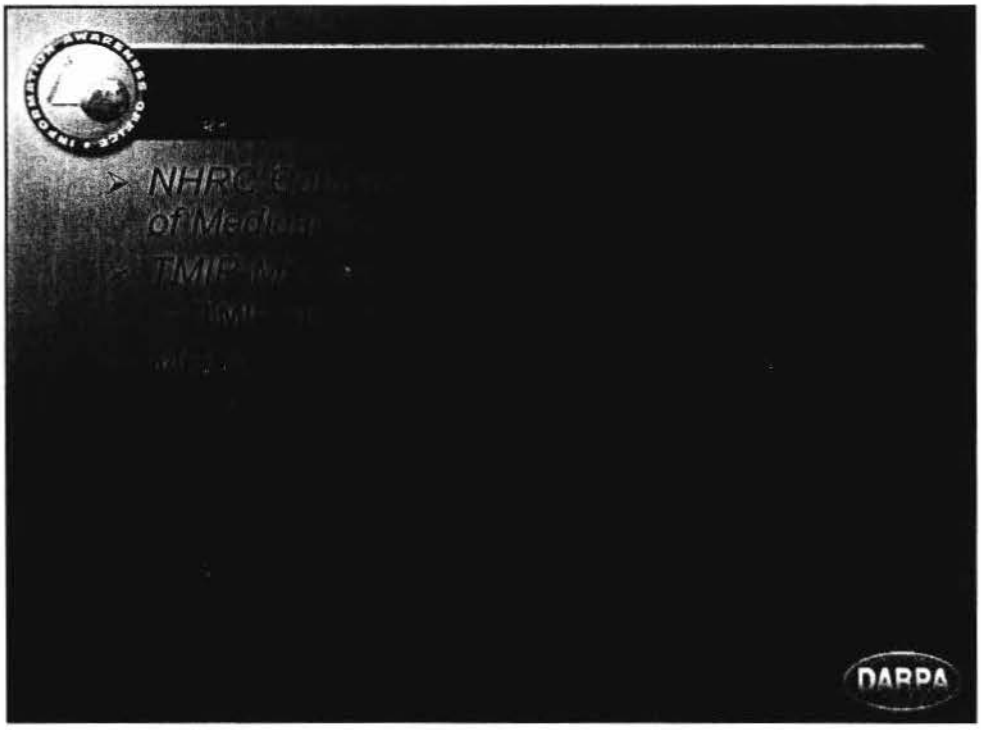

DARPA

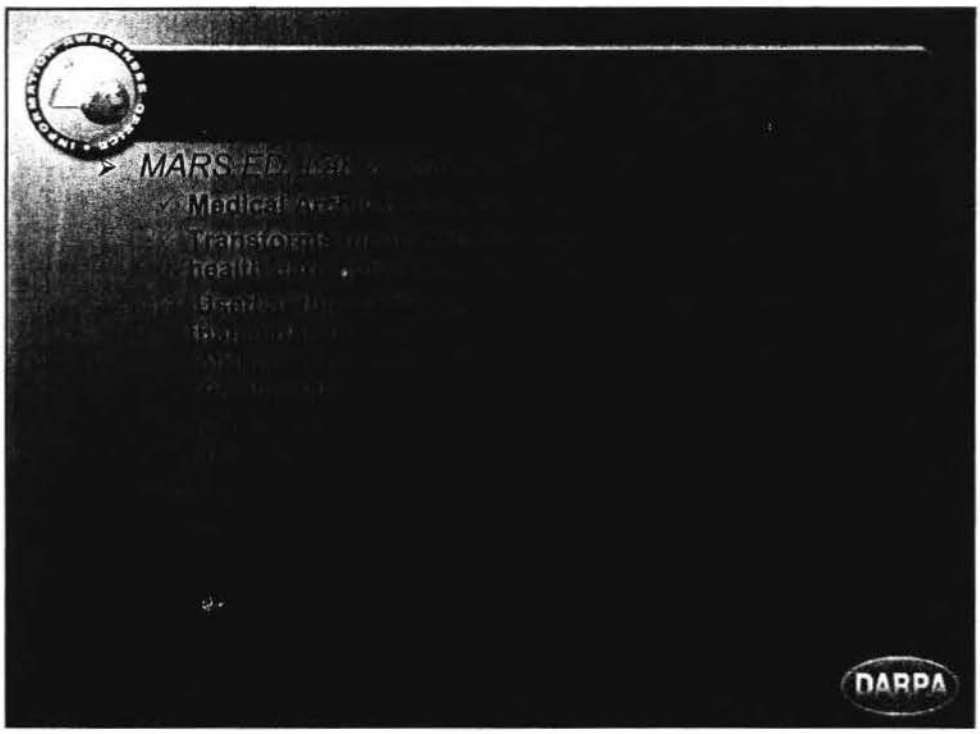

DARPA

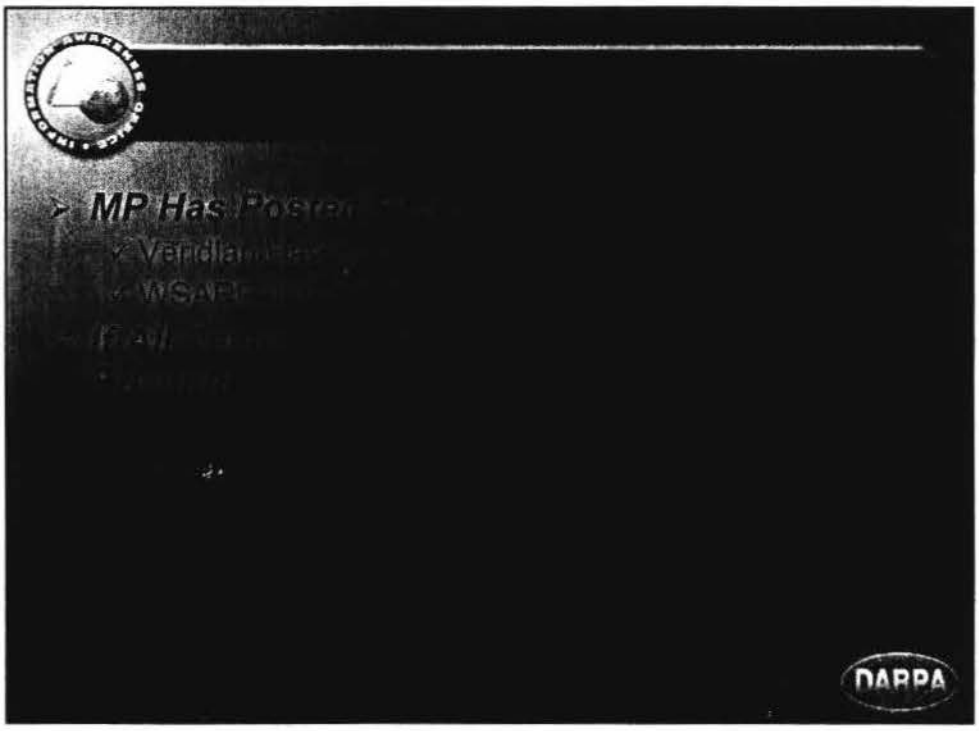

DARPA

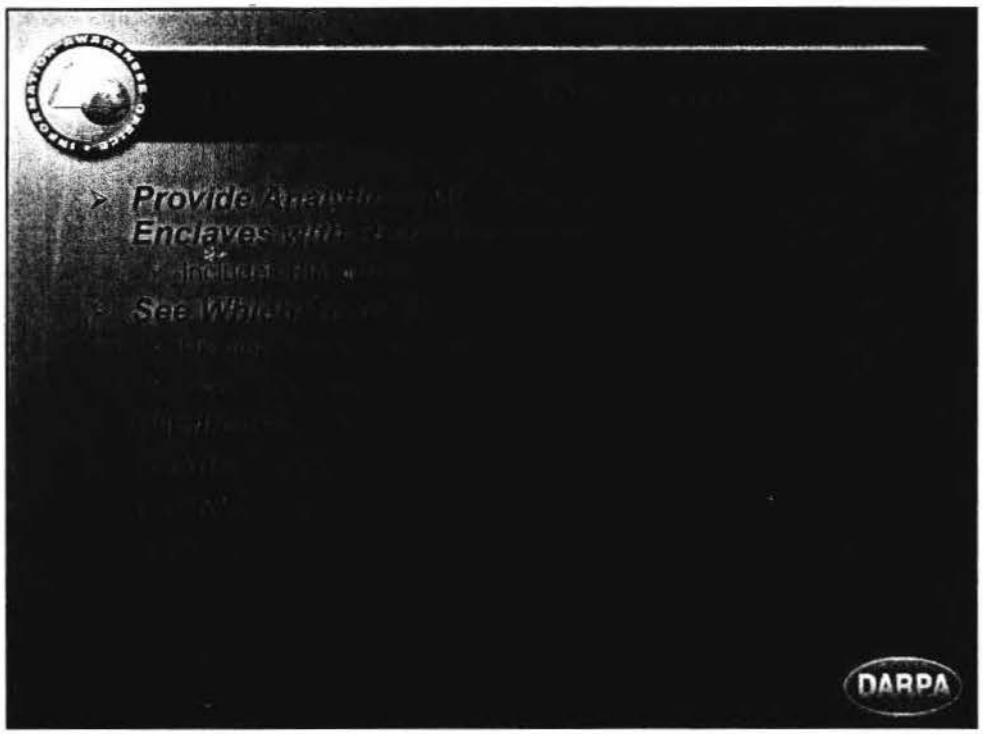

DARPA


DARPA

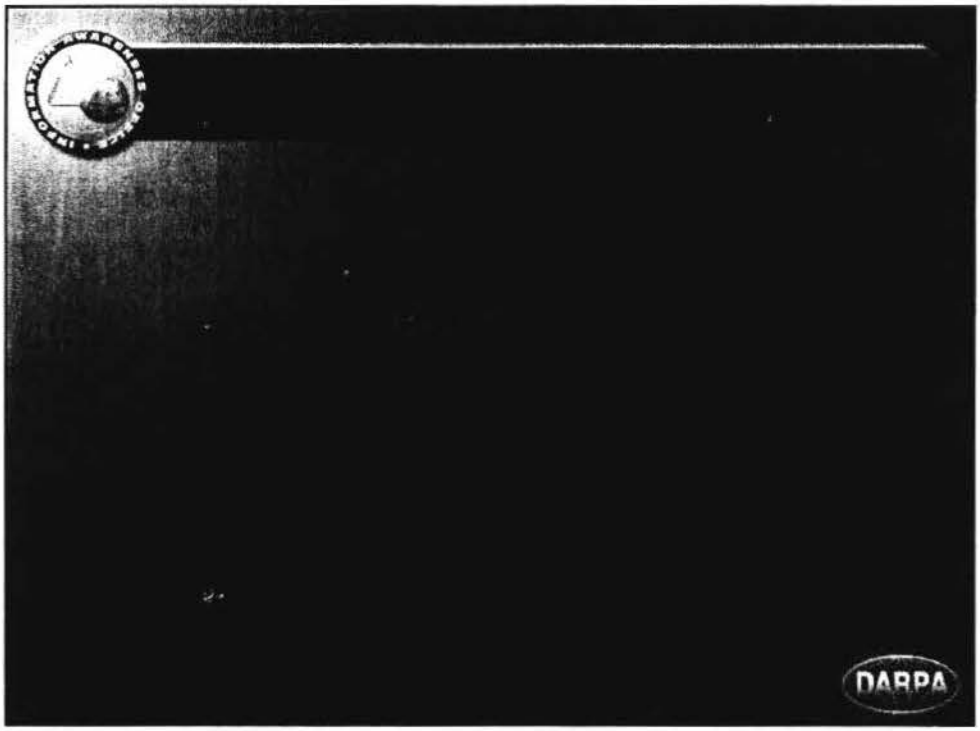

DARPA

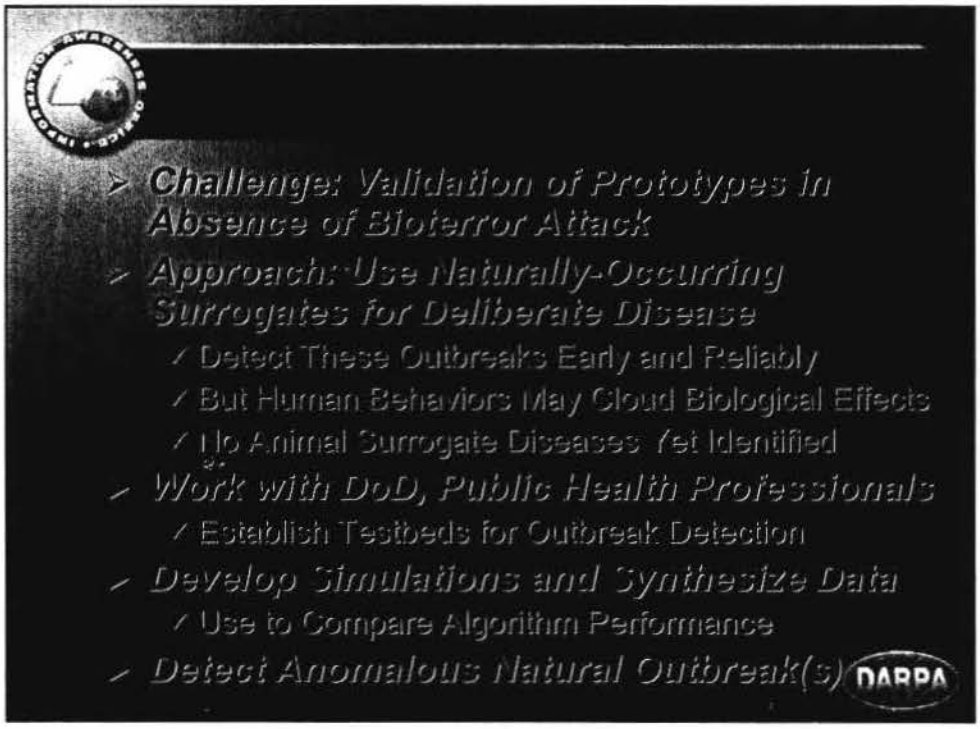

DARPA


DARPA


DARPA


DARPA


DARPA


Challenge: Validation of Prototypes in Absence of Bioterror Attack
Approach: Use Naturally-Occurring Surrogates for Deliberate Disease
Detect These Outbreaks Early and Reliably
But Human Behaviors May Cloud Biological Effects
No Animal Surrogate Diseases Yet Identified
Work with DoD, Public Health Professionals
Establish Testbeds for Outbreak Detection
Develop Simulations and Synthesize Data
Use to Compare Algorithm Performance
Detect Anomalous Natural Outbreak(s)
DARPA

16 G

RPHD Tasking FY01-02

- Non-traditional Population Health Indicator Data
  - Successfully negotiated MOU and acquired historical and real-time, non-traditional data from City of Norfolk
  - Obtained and analyzed non-traditional data [including school absentee rates, OTC drug sales, 911/EMS dispatches, potable water usage, and wastewater composition (Milwaukee)]
  - Analyzed San Francisco 911/EMS data
  - Developed and populated Protégé ontology to catalog and share data across projects
- Advanced Analytical Techniques
  - Integrated statistical and knowledge-based techniques in a flexible monitoring system
  - Cough detection by audio spectral feature clustering
  - Adapted Kalman filters for aberration detection via statistical inference on post-forecast residuals
- RPHD System Development
  - Implemented and tested a prototype RPHD that integrates statistical detection algorithms with knowledge-based analysis
  - Successfully detected outbreaks in BioWar simulated data
  - Visualized spatial correlation of anomalies with a browser-based GIS with drill-down capability to source-based data displays
- Disease Surveillance Activities
  - Supported syndromic surveillance at Naval Medical Center, Portsmouth (NMCP)
- Transition to Operational Use
  - Negotiated MOU, installed initial RPHD component (BASIS) at NMCP
  - Transitioned BASIS technology into MDSS system at the Naval Health Research Center (NHRC)

RPHD Tasking FY03

- Non-traditional Population Health Indicator Data
  - Evaluate IBM's closed-campus phone monitoring for application at Norfolk Navy Base
  - Evaluate evidence for cryptosporidiosis outbreak via Milwaukee wastewater analysis
  - Analyze water usage to detect large-scale outbreak of gastrointestinal illness
- Advanced Analytical Techniques
  - Improve normalcy models to account for cyclic variations, environmental influences, etc.
  - Incorporate dynamic parameter fitting and cluster analysis to optimize Kalman filters
  - Exploit richness in data via knowledge-based processing techniques
  - Develop advanced spatial reasoning and spatial decomposition/recomposition methods
  - Augment data fusion through vector statistical inference on post-forecast residuals and their covariance components
  - Prototype methods for disease characterization
- RPHD System Development
  - Implement end-to-end system that automates processing and analysis of "live" data
  - Utilize Protégé ontology to access data at run time
  - Develop web-based "light" client to present geographic results
  - Test the RPHD system with simulated and real-world data
  - Apply privacy protection methods from IBM and CMU
- Disease Surveillance Activities
  - Expand disease surveillance at Naval Medical Center Portsmouth (NMCP)
  - Test prototype audio cough detection in military classrooms
- Transition to Operational Use
  - RPHD components at NMCP, Air Force and Army hospitals in Hampton Roads
  - RPHD components within city public health departments and private sector medical facilities in Hampton Roads

RPHD Tasking FY04

- Non-traditional Population Health Indicator Data
  - Determine nationwide applicability of non-traditional data for regional disease surveillance
  - Enhance knowledge bases of data sources and their properties
  - Implement operational methods for routine acquisition and protection of data
- Advanced Analytical Techniques
  - Finalize adaptive, multivariate forecasting models underlying aberration detection
  - Develop additional knowledge-based problem solvers in response to FY03 experience
  - Characterize and encode behavior of all problem-solvers in a knowledge base, allowing RASTA control structure to improve overall system performance
- RPHD System Development
  - Refine all RPHD components based on FY03 demos to ensure robust, deployable system
  - Develop fault tolerant servers with data backup and recovery solutions
  - Ensure easy installation and software maintenance
  - Create user manuals and documentation
- Disease Surveillance Activities
  - Conduct unified military and civilian disease surveillance for Hampton Roads
  - Support user training in Hampton Roads
- Transition to Operational Use
  - Finalize transition to Hampton Roads Regional Public Health and Metropolitan Medical Response System (MMRS)
  - Develop additional Navy applications via Naval Health Research Center, San Diego

Pitt Performance Goals

- 2001-2002
  - Data:
    - Pursued scores of data types; obtained 10
    - First statistically valid measurements of timeliness
  - Algorithms: Developed initial versions of WSARE and PANDA
  - Performance goals: identified methods for 2 week speed up relative to hospital admissions for pediatric
- 2003/4
  - Data:
    - Obtain 20 more types of data.
    - Expand depth and breadth for data we have sufficient to make statistically valid measurements for OTC, chief complaint, transportation, 911 call center and utility data
    - Develop new approaches to data screening based on linkage
  - Algorithms: Develop second versions of WSARE, PANDA; complete Bayesian aerosol detector
  - Performance goals: approach one day latency from time sick individuals first awaken with symptoms to detection.

Pitt/CMU Tasks FY01-02

- Developing Detection Algorithms:
  - WSARE,
  - PANDA,
  - Bayesian aerosol,
  - Fast Scan Statistic,
  - NLP,
  - SecondStage
- Characterize non traditional data:
- Biowar Development and use for evaluation

Pitt/CMU Tasks FY03

- ***Developing Detection Algorithms:***
  - ***WSARE,***
  - ***PANDA,***
  - ***Bayesian aerosol,***
  - ***Fast Scan Statistic***
- ***Characterize non traditional data: Gather 25 More Data Sources***
- ***Biowar to be Enhanced for Higher Fidelity***
- ***Characterize non traditional data: Use Gold Standard data to Validate Nontraditional Sources***
- ***"Large Scale Data Evaluations"***

Pitt/CMU Tasks FY04

- ***Developing Detection Algorithms:***
  - ***WSARE,***
  - ***PANDA,***
  - ***Bayesian aerosol,***
  - ***Fast Scan Statistic***
- ***Characterize non traditional data: Gather 25 More Data Sources***
- ***Biowar to be Enhanced for Higher Fidelity***
- ***Characterize non traditional data: Use Gold Standard data to Validate Nontraditional Sources***
- ***Transition Technologies and Data Characterizations***

ESSENCE II Tasking FY01-02

Non-Traditional Health Indicator Data

- Acquire data sources (absenteeism, over-the-counter & prescription medications, electronic claims, emergency room chief complaint, animal health indicators, nurse hotline, HMO web click stream, video cough, etc.)
- Analyze sources to determine the strength and timeliness of the indicator

Advanced Analytical Techniques

- Select optimum temporal and spatial detection techniques for data in the NCA
- Implement detection techniques against live NCA data streams
- Participate in BioWars evaluation detection approaches

ESSENCE System Development

- Develop & implement automated data acquisition, text parsing & archiving software
- Implement privacy algorithms on data source server
- Develop & implement web based worldwide military disease surveillance capability
- Develop civilian NCA prototype & demonstrate during county exercise

Disease Surveillance Activities

- Standup worldwide military system
- Web sites to distribute NCA surveillance information to local health departments
- Daily analysis of NCA data to identify clusters of disease
- Standup NCA disease surveillance operations center

Transition to Operational Use

- Primary transition into DoD-GEIS
- Transition of components to DTRA, PEOCBD, & CDC

ESSENCE II Tasking FY03

Non-Traditional Health Indicator Data

- Acquire additional data sources (national animal data, 911 calls, national poison center, etc.)
- Socio-economic models for selected indicators
- Seasonal models

Advanced Analytical Techniques

- HMO data mining
- Spatial data integration
- Automation of detection techniques

ESSENCE System Development

- Privacy algorithms on civilian data and for NCA data sharing
- Situation assessment algorithms & text processing
- NCA military / civilian cross-jurisdictional user interfaces
- ArcIMS based ESSENCE I upgrade
- ESSENCE II integration and testing

Disease Surveillance Activities

- Continued daily NCA surveillance to identify clusters of disease
- Continued military surveillance activities

Transition to Operational Use

- Investigate transition to NHRC for San Diego area
- Transition of components to DTRA, PEOCBD, & CDC

ESSENCE II Tasking FY04

Non-Traditional Health Indicator Data

- Apply knowledge gained in the NCA to other locations in CONUS
- Explore rapid model development and background assessment techniques
- Acquire national sources of data

Advanced Analytical Techniques

- Refine detection methods based on operational use
- Develop more generalized techniques independent of location

ESSENCE System Development

- Implementation of ESSENCE in a operational military setting
- Implementation of ESSENCE II in a civilian setting with collaboration of the local health departments.
- Complete the integration and testing of the two ESSENCE versions

Disease Surveillance Activities

- Support the military surveillance users in upgrades
- Train the civilian users of ERSSENCE II in NCA
- Support upgrades to the civilian version of ESSENCE

Transition to Operational Use

- Complete transition of ESSENCE to military and civilian users

IBM/JHU/SDI Tasks: 2001-2002

- Data source evaluation
  - Most promising: Worksite survey, worksite medical phone calls, worksite medical web accesses
  - Baseline: environmental, HMO claim from SDI
  - TBD: Cafeteria, worksite attendance, utility
  - Not considered promising: Cough counting, newsgroups
- Anomaly detection algorithms
  - Site detectors,
  - supervised/unsupervised general detectors
  - Disease models based on environmental data (JHU)
- Surveillance system prototype
  - Prototype with live data feeds, data archiving, detectors operational at Watson site by YE 2002
- Transfer
  - Video car counting, interpolated environmental data, MALM, privacy prototype are available for transfer

IBM/JHU/SDI Tasks: 2003

- Data source evaluation
  - Ongoing characterization/evaluation of existing/new data sources
  - Improved methodology for data source evaluation
  - Explore the extension of selected site-based data sources to a regional basis, e.g., NY City
  - Environmental effects on disease processes
- Further development and testing of detectors
  - Improved testing methodology
  - Hybrid site/regional detectors
- Surveillance system
  - Incorporate additional data sources, detectors, environmental data
- Transfer
  - Employ site surveillance system at other sites, e.g., military bases
  - Make detector technologies, data source insights available to systems contractors

1.H

25

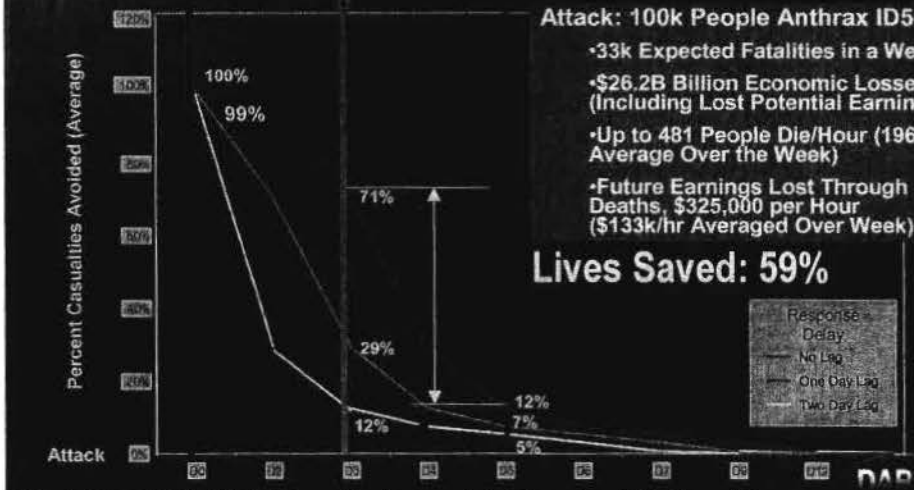
Bio-ALIRT:
Bio-Event Advanced Leading Indicator Recognition Technology BioSurveillance Program

Principal Investigators' Meeting
3-5 December 2002
Arlington, VA

Ted Senator
DARPA/IAO
703-696-2231
tsenator@darpa.mil

Motivation

BioSensors | BioSurveillance | Clinical Diagnostics

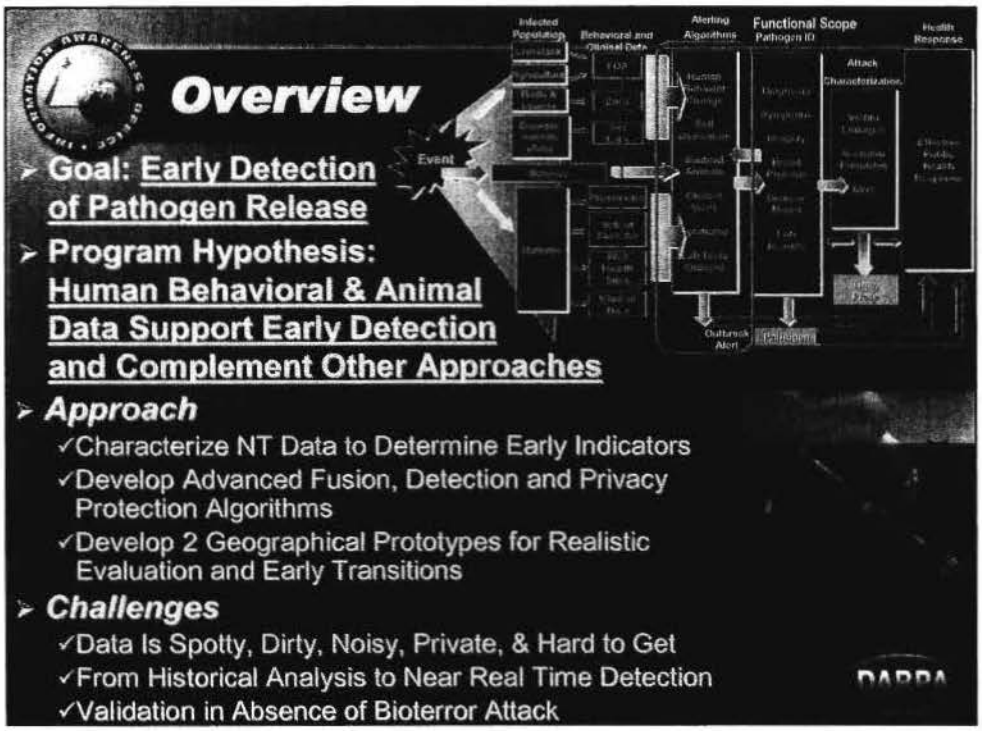


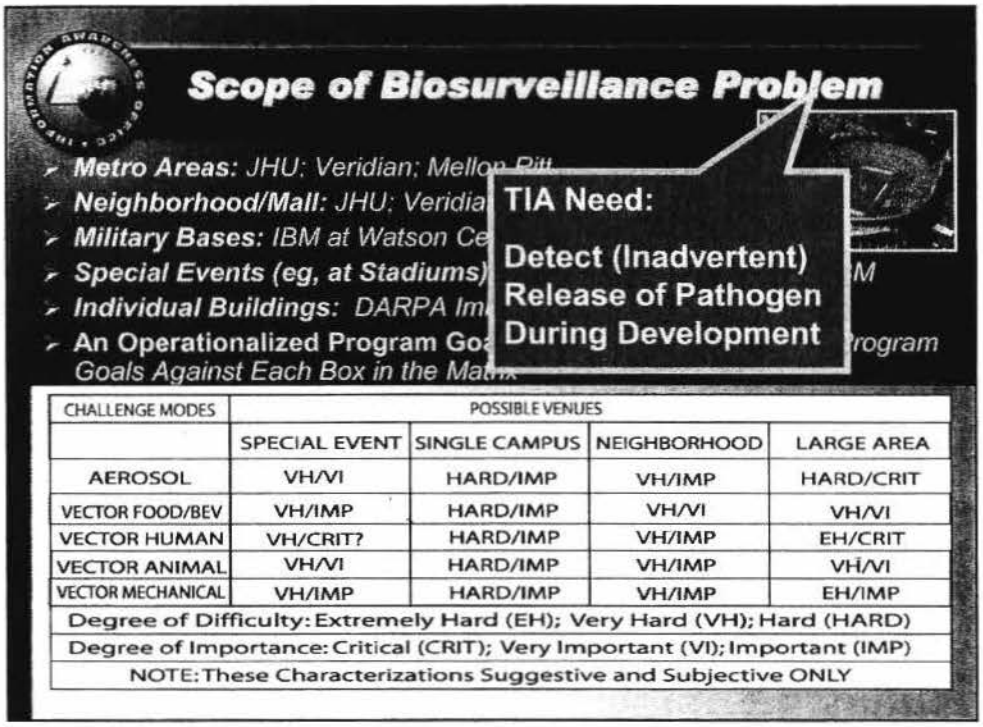
Attack: 100k People Anthrax ID50

- 33k Expected Fatalities in a Week
- $26.2B Billion Economic Losses (Including Lost Potential Earnings)
- Up to 481 People Die/Hour (196/hr Average Over the Week)
- Future Earnings Lost Through Deaths, $325,000 per Hour ($133k/hr Averaged Over Week)

Lives Saved: 59%

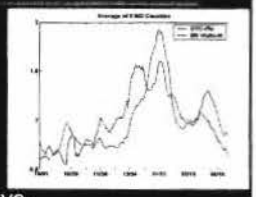
Chart from OSD; Model from Kaufman, et al, *Emerging Infectious Diseases*



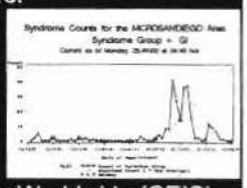


CHALLENGE MODES	POSSIBLE VENUES			
	SPECIAL EVENT	SINGLE CAMPUS	NEIGHBORHOOD	LARGE AREA
AEROSOL	VH/VI	HARD/IMP	VH/IMP	HARD/CRIT
VECTOR FOOD/BEV	VH/IMP	HARD/IMP	VH/VI	VH/VI
VECTOR HUMAN	VH/CRIT?	HARD/IMP	VH/IMP	EH/CRIT
VECTOR ANIMAL	VH/VI	HARD/IMP	VH/IMP	VH/VI
VECTOR MECHANICAL	VH/IMP	HARD/IMP	VH/IMP	EH/IMP
Degree of Difficulty: Extremely Hard (EH); Very Hard (VH); Hard (HARD)				
Degree of Importance: Critical (CRIT); Very Important (VI); Important (IMP)				
NOTE: These Characterizations Suggestive and Subjective ONLY				

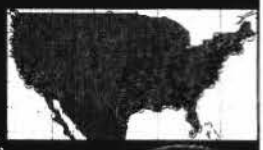
Things Going Well



- **DATA:**
  - ✓ Over 100 Secured and Analyzed. 75+ Nontraditional.
  - ✓ 4 Leading Indicators Identified. OTC Leads Peak by 5 Days.
- **ALGORITHMS:**
  - ✓ Developing 10+ Selected Techniques
  - ✓ Evaluating Using BioWAR Simulation with Synthetic Data
  - ✓ Detects in 3.54 days with 1% FPR w/ Simulated Data

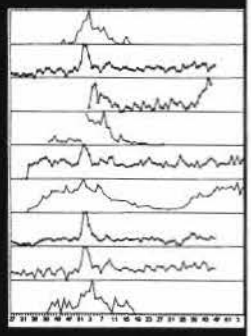


- **PROTOTYPES:**
  - ✓ Currently Monitoring Health of All Non-Deployed US Forces Worldwide (GEIS); the National Capital Region (Essence II); and Portsmouth Naval Hospital (RPHD)
  - ✓ Detected 3 Previously Unknown Disease Outbreaks
  - ✓ Favorably Reviewed by President and Governor Ridge
  - ✓ Inaugural & Olympics Surveillance Support. NYAM
  - ✓ "Proof-of-Concept" for DoD BD (DTRA and PEO/CBD)
    - DARPA Biosurveillance of NCR w/PEO Announced by DoD.
    - ESSENCE II Used as Model by BDI; Transition NT Data in 04



DARPA

Things Needing Improvement



- ***More/Better Nontraditional Data Sources for Early Detection***
  - ✓ E.g., Animals!
- ***Better Evaluation of Data Sources***
  - ✓ Temporal Correlations
  - ✓ Geographic Validations
  - ✓ More Reliable and Better Understood
- ***Intra-Program Cooperation***
- ***Common Realistic Objective Evaluation***
  - ✓ Fewer Artificialities
- ***Transition Planning***
  - ✓ Total Information Awareness Program
  - ✓ DoD CBD Initiative, DHS, DHHS/CDC

DARPA

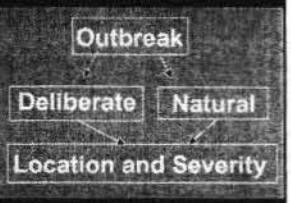
Status Update

- ***DIRO Program Review***
  - ✓ Program Continuation for FY03
  - ✓ Name Change to Bio-ALIRT
  - ✓ Go/NoGo Criteria
  - ✓ Non-Traditional Data Questions
- ***Financial Review***
  - ✓ All Participants to Continue in FY03
  - ✓ Reallocated $1.2M from FY03 to FY04
- ***(Minor) Course Corrections***
  - ✓ Less Systematization
    - Proliferation of Syndromic Surveillance Systems
    - DoD CBD Initiative; DHS, DHHS/CDC ?
  - ✓ More Data Characterization, Algorithm Development
  - ✓ TIA-Scenarios Added
  - ✓ "DARPA Inside" Credit

DARPA

Quantitative Goals: Go/NoGo

- ***FY02***
  - ✓ **Acquire and Assess 100 Data Sources √**
  - ✓ **Develop Robust Detection & Fusion Algorithms √**
  - ✓ **Develop Functioning Prototypes√**
  - ✓ **Detect Previously Unknown Outbreaks √**
- ***FY03***
  - ✓ **Test 4 Leading Data Types in Other Regions/Methods**
  - ✓ **Acquire and Assess 30 Additional Data Sources**
  - ✓ **Develop Dynamic Privacy Protection Techniques**
  - ✓ **Detect Outbreak One Day Earlier**
  - ✓ **Transition 3 Products**
- ***FY04***
  - ✓ **Publish Leading Early Indicators (Peer Reviewed)**
  - ✓ **Detect Outbreak an Additional Day Earlier**
  - ✓ **Validate System by Predicting Local Flu Outbreaks**
  - ✓ **Transition Prototype(s) to DoD/Military Components**
  - ✓ **Incorporate NT Data into TIA**



DARPA

Meeting Purpose and Goals

- ***Enable Collective Program Progress***
  - ✓ Working Meeting for the PI's
  - ✓ Productive Interactions Between Participants
- ***Facilitate Transitions and Cooperation***

- ***Goals***
  - ✓ Document Year 1 Progress and (re)Establish Year 2 Plans
  - ✓ Understand Utility of Different Non-Traditional Data Sources
  - ✓ Determine Evaluation Approach for 03
  - ✓ Specify Additional Data Sources and Algorithms
  - ✓ Define TIA Scenario and Uses
  - ✓ Update Transition Products & Relationships
  - ✓ Refresh Privacy Understanding and Actions

DARPA

Backups

DARPA

Bio-ALIRT:
Bio-Event Advanced Leading Indicator Recognition Technology BioSurveillance Program

Principal Investigators' Meeting

3-4 December 2002

Arlington, VA

Action Items and Wrap-Up

Ted Senator
DARPA/IAO
703-696-2231
tsenator@darpa.mil

DARPA

IMPORTANT REMINDER

Information Presented and Discussed at this Meeting:

- *Is NOT (necessarily) Approved for Public Release*
- *Belongs to the Presenter or the Government*
- *MAY Be Subject to Export Control Regulations*
- *Is Sensitive*

- *This Information is Shared within Bio-ALIRT, IAO/TIA, DARPA, DoD, and the Government for Official Purposes*
  - ✓ Maximal Sharing is Encouraged Within these Bounds
  - ✓ Treat Others' Information as You Would Want Them to Treat Yours

- *It is Your Responsibility to Treat Information Properly. If you Are Unsure about any Particular Use, Please Ask Someone Who Knows FIRST*

DARPA

Thanks

- To Potomac Institute and Schafer Corp. for Hosting the Meeting
- To Bob Linn, Angela Clemons, Carole Kelliher, and Aileen Born for Registration, Food, Assistance, etc.
- Participants for Prepared, Informed, Thoughtful, Enthusiastic Discussions
- All for Flexibility of Schedule

DARPA

Observations

- HHS, CDC, NYC Extremely Positive
  - ✓ "This is what I had wanted the NY meeting to be. 'This is fun"
  - ✓ How to Build on non-DoD Connections?
- We're in the Lead -- By Far -- in Medical/Syndromic Surveillance
- But Demand for Deployed/Operational Systems may/will still Out-Run Us

DARPA

Some Good Ideas We Heard

- Opening Up Evaluation to non-Participants
- Survey of Behavior by WRAIR through Kaiser
- Repeat IBM Active Surveillance Experiment at IBM Watson and at JHU/APL, Veridian, U Pitt
- Common Color-Coding (and Attributes) for Data Charts -- and Compare/Contrast Charts for Other Aspects (e.g., Detection Algorithms) Too
- Try Auto-Correlation of Absence Data
- Add VA Coverage (5M) to MTF Coverage (4M)
- Five Categories of Animals: Zoo, Farm, Pets, Working, Wild
- Temporal Sliders on Geographical Displays

DARPA

Some Things We Learned

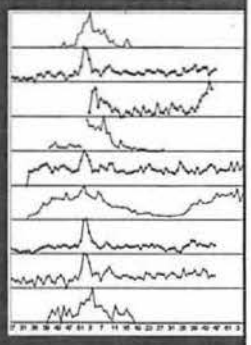
- Good Data Sources are useful for both Detection and Investigation
- Maximizing Correlation Coefficient as Function of Lead/Lag Time is NOT a Good Measure of Earliness of Detectability
- Animals are not very Cooperative
- Sales Promotions Affect Share, not Category (can they also affect timing ?)
- Sample Density: PO>ER; Sickness: ER>PO
- Call Center Industry has Centralized Nationally
- Passerines are Perching Songbirds
- Judith is not a Lawyer but She Can Read

DARPA

Things Needing Improvement

- *More/Better Nontraditional Data Sources for Early Detection*
  - ✓ Animals!
- *Better Evaluation of Data Sources*
  - ✓ Temporal Correlations
  - ✓ Geographic Validations
  - ✓ More Reliable and Better Understood
- *Intra-Program Cooperation*
- *Common Realistic Objective Evaluation*
  - ✓ Fewer Artificialities
- *Transition Planning*
  - ✓ Total Information Awareness Program
  - ✓ DoD CBD Initiative, DHS, DHHS/CDC

More Outbreaks
N >> 1
Multiple Locations
Common Data



DARPA

Methodology

- *Clearly Identify Hypotheses Being Tested*
  - ✓ Attack Scenario
  - ✓ Data Source and/or Algorithm
  - ✓ Concept of Operations
- *Present the Message Clearly*
  - ✓ Annotated Data Slides
  - ✓ Highlight Technical Advances
- *Measure Progress*
- *Program-Wide Summaries: Data, Algorithms, etc.*
- *Standardize*
  - ✓ Data Descriptions & Data Sets
  - ✓ Outbreaks/Cases
  - ✓ Statistics
  - ✓ Evaluation Results
- *Give DARPA Credit in Software, Publications, etc.*

DARPA

Approach

- **Year 1: Looked Where Light Was Brightest**
  - ✓ Characterized Data Sources
  - ✓ Multiple n=1 studies
  - ✓ Ad-Hoc
  - ✓ Groups Worked Separately
- **Year 2: Need to Look Where Keys are Likely to Be**
  - ✓ Evaluate Data Sources
  - ✓ Need n=N studies
  - ✓ Systematic
  - ✓ Groups Need to Work Together
- **It's Still All About the Data**

DARPA

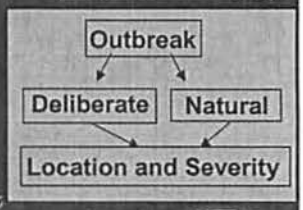
Status Update: Reminder

- **DIRO Program Review**
  - ✓ Program Continuation for FY03
  - ✓ Name Change to Bio-ALIRT
  - ✓ Go/NoGo Criteria
  - ✓ Non-Traditional Data Questions
- **Financial Review**
  - ✓ All Participants to Continue in FY02
  - ✓ Reallocated $1.2M from FY03 to FY02
- **(Minor) Course Corrections**
  - ✓ Less Systematization
    - Proliferation of Syndromic Surveillance Systems
    - DoD CBD Initiative, DHS, DHHS/CDC ?
  - ✓ More Data Characterization, Algorithm Development
  - ✓ TIA-Scenarios Added
  - ✓ DARPA Insider Credit

DARPA

Quantitative Goals: Go/No Go

- **FY02**
  - ✓ Acquire and Assess 100 Data Sources √
  - ✓ Develop Robust Detection & Fusion Algorithms √
  - ✓ Develop Functioning Prototypes √
  - ✓ Detect Previously Unknown Outbreaks √
- **FY03**
  - ✓ Test 4 Leading Data Types in Other Regions/Methods
  - ✓ Acquire and Assess 30 Additional Data Sources
  - ✓ Develop Dynamic Privacy Protection Techniques
  - ✓ Detect Outbreak One Day Earlier
  - ✓ Transition 3 Products
- **FY04**
  - ✓ Publish Leading Early Indicators (Peer Reviewed)
  - ✓ Detect Outbreak an Additional Day Earlier
  - ✓ Validate System by Predicting Local Flu Outbreaks
  - ✓ Transition Prototype(s) to DoD/Military Components
  - ✓ Incorporate NT Data into TIA



DARPA

Meeting Purpose and Goals

- **Enable Collective Program Progress**
  - ✓ Working Meeting for the PI's
  - ✓ Productive Interactions Between Participants
- **Facilitate Transitions and Cooperation**
- **Goals**
  - ✓ Document Year 1 Progress and (re)Establish Year 2 Plans
  - ✓ Understand Utility of Different Non-Traditional Data Sources
  - ✓ Determine Evaluation Approach for 03
  - ✓ Specify Additional Data Sources and Algorithms
  - ✓ Define TIA Scenario and Uses
  - ✓ Update Transition Products & Relationships
  - ✓ Refresh Privacy Understanding and Actions

DARPA

Action Items

Item	Who	Due
Provide (Revised) Slides to Rey Wong	All	12/5
Post Slides on Program Web Page	SETA	12/10
Define TIA Scenario	Siegrist	1/31
Post DARPA Logos & Acknowledgement Statement	SETA	12/6
Review FY03 Tasks, $	PM/RLs	12/13
Schedule for Y2 Evaluation Plan	Siegrist	12/6
Y2 Evaluation Plan	Siegrist	1/31
Schedule/Format Ongoing Interactions (Conference Calls, Working Group Mtgs, Groove Spaces, etc.)		
Frameworks for Common Data, etc, Charts	Siegrist	1/31
Ideas for Follow-On Program	All	6/30
Bio-Defense TRIAD Ideas	S&S	?

DARPA

Wrap-Up from Kickoff Mtg

- Four Outstanding Interdisciplinary Team
- Need to Re-Emphasize Use of Non-Traditional Data Sources for Early Detection
- Refocus on Military Needs, Data, Transitions, etc.
- Need to Specify Metrics and Experiments, especially with regards to 12-18 month Program Level Go/NoGo Decision
- Key Metric is How Much Earlier Detection Can Occur (at specified performance levels)
- Collaboration Between Teams Enables Best System
  - Sharing of Data, Insights, etc.

DARPA

Action Items from Kickoff

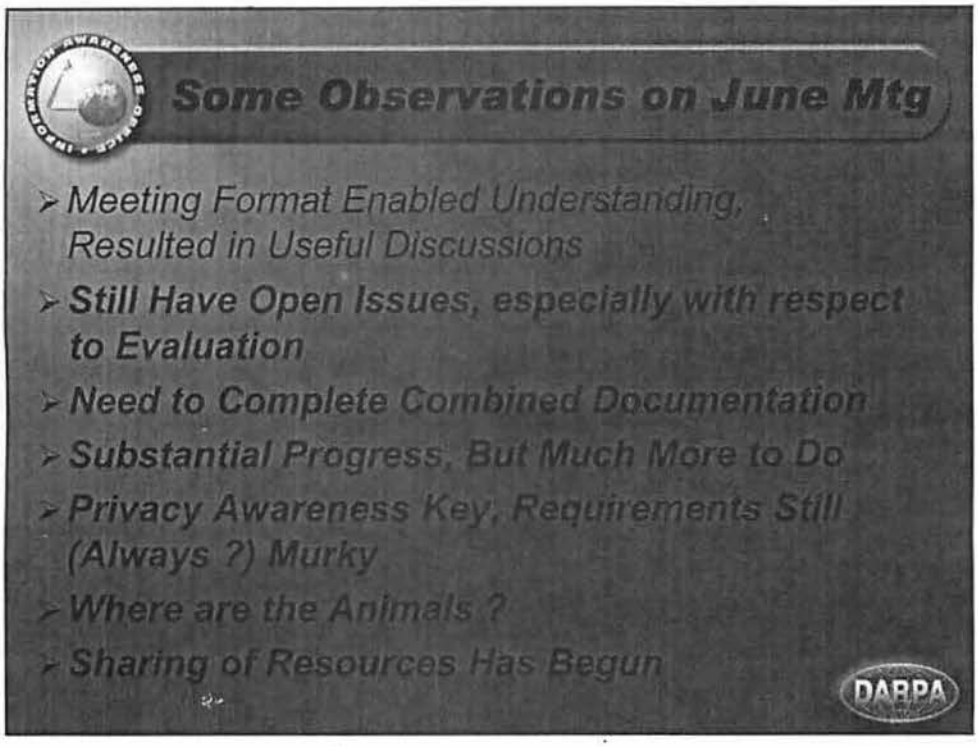
- *Tie Data Sources to Attack Scenarios/ Number of People Exposed/Contagious* — *Two Scenarios Done in August Challenge*
- *Revise Research Plans to Incorporate Annual Evals* — *DONE*
- *Determine How to Measure (Combined) Sentinel Value(s) of Data Sources: Identify and Acquire Other Potentially Useful Data Sources – All teams* — *Ongoing*
- *Specify and Schedule Technology Developer Contributions to Prototype System Developers* — *Ongoing*
- *Data: Historical, Almanac* — *Ongoing*
  - ✓ Petting Zoos, VA, National Marketers, Web, Public Records, Deaths
- *Develop Common Metrics and Evaluation Plans* — *Eval DONE*
  - ✓ Framework by Next Week, Detailed within 6 Months
- *Meeting Material to Schafer (for web site, CD)* — *DONE*

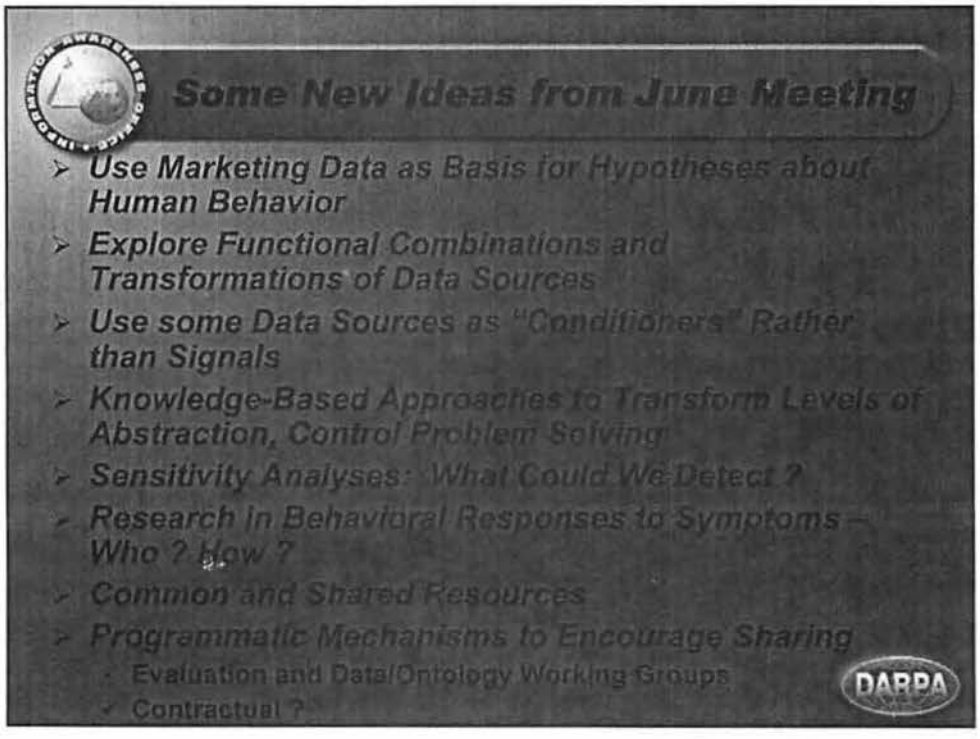
DARPA

June Meeting Objectives

- *Update Common Understanding of Current Program Goals, Objectives, Constraints, etc. – Good*
  - ✓ Enable Earlier Detection Using Non-Traditional Data
  - ✓ "Prove the Point" and Transition Knowledge and Technology
- *Develop Program-Wide View/Understanding of Data Sources, Technical Functions, Privacy, Evaluation, etc. –*
  Understanding of Issues but no Single Solutions, Need to Complete Documentation
- *Ratify Common Evaluation Plan; Develop Data Map; Agree on Data Standards; etc. – Progress, but far from Complete*
- *Identify Additional Opportunities for Collaboration Among Program Participants – Good Ideas, How to Share DataFly, BioWar?*

DARPA


Some Observations on June Mtg
Meeting Format Enabled Understanding, Resulted in Useful Discussions
Still Have Open Issues, especially with respect to Evaluation
Need to Complete Combined Documentation
Substantial Progress, But Much More to Do
Privacy Awareness Key, Requirements Still (Always ?) Murky
Where are the Animals ?
Sharing of Resources Has Begun
DARPA


Some New Ideas from June Meeting
Use Marketing Data as Basis for Hypotheses about Human Behavior
Explore Functional Combinations and Transformations of Data Sources
Use some Data Sources as "Conditioners" Rather than Signals
Knowledge-Based Approaches to Transform Levels of Abstraction, Control Problem Solving
Sensitivity Analyses: What Could We Detect ?
Research in Behavioral Responses to Symptoms – Who ? How ?
Common and Shared Resources
Programmatic Mechanisms to Encourage Sharing
Evaluation and Data/Ontology Working Groups
Contractual ?
DARPA

BioSurveillance Program June Mtg How Are We Doing ?

- *NON-TRADITIONAL Data Sources – Behind Where We Should Be*
- *Advanced Analytics (Detection and Fusion) – Good Ideas, But How to Evaluate/Select ? How to Combine Knowledge-Based and Numerical Approaches?*
- *Privacy – Awareness of Needs and Some New Technology*
- *Evaluation – Surfaced Detailed Issues and Ideas, Lots More to Do Real Soon*
- *Transition – Good Progress on Internal Sharing, Good Start to some External Customers*

DARPA

Big Issues from June PI Meeting

- *Acquisition, Characterization, Analysis, and Use of NON-TRADITIONAL DATA*
- *How to Evaluate ?*
  - ✓ Algorithms and Data
  - ✓ Simulated and Natural
- *It's All About the Data !*
  - ✓ Algorithm Work is about Choosing Right Variables
  - ✓ Architecture Involves Transforming Abstraction Levels

DARPA

Ideas for Next (December) Meeting

- *Early-Mid November, Away from DC Area ?*
  - ✓ After FY03 Budget and Funding Decisions
  - ✓ "Final Report" on FY01/02 Contracts
- *Agenda Items*
  - ✓ Status/Accomplishments Updates
  - ✓ Goals/Plans for Contract Continuations
  - ✓ Results of Year 1 Experiment/Evaluation
  - ✓ Transition Discussions with Partners
  - ✓ Demonstrations ?
- *DC Area Technology Demonstration Day ? Or ad-hoc ?*

DARPA

Action Items from June PI Mtg

Item	Who	Due
Provide (Revised) Slides to Rey Wong - Include Specific/Quantitative Performance Goals	All	6/10
Produce Proceedings – Post on Web Site	SETA	6/12
Provide Technical Documents/Links for Web Site	All	6/14
Review MiTAP to determine utility and use	All	6/30
Complete Program-Wide Session Products	Siegrist/PIs	6/14
Named Components for Transitions (internal & external; availability, status)		
Complete Characterization of Attack Scenarios	Siegrist	6/30
Formalize Privacy Policies and Procedures	All	6/30
Establish Transition Agreements	PM	Ongoing
Identify (External) Transition Opportunites	All	Ongoing
Evaluation Planning and Execution Initial Schedule and Plan	Siegrist + Wkg Gp	6/19
Continue Excellent Technical Work always	All	

DARPA

TITLE PAGE

FOLDER TOC 15

E-MAILs CONCERNING TIA

Loose-leaf collection – some multiple sender / recipients (copies to each)

- 15.A DARPA (e-mail)
- 15.A.1 DARPA (Catelog)
- 15.B Potomac Institute
- 15.C Syntek Technologies (IAO)
- 15.D Veridan
- 15.E Pittsburg U
- 15.F Knoltex
- 15.H Carnegie Mellon U

ddyer

From: jpoindexter
Sent: Friday, May 31, 2002 9:00 PM
To: tarmour; ddyer
Subject: RE: IBM Privacy Institute

Doug,

Why don't you see if you can make contact with this group and get more details on what they are doing.

-----Original Message-----
From: tarmour
Sent: Friday, May 31, 2002 6:36 PM
To: John Poindexter (E-mail); Douglas Dyer (E-mail)
Subject: IBM Privacy Institute

Just stumbled on this.

-t

Established in November 2001, the IBM Privacy Institute is an organization within IBM Research to promote and advance research in privacy and data protection technology. The institute's goal is to develop the necessary technologies for enterprises that enable the transition from today's privacy-unaware or even privacy-intrusive ways of doing e-business to privacy-enabling ways. As the industry's first technology research institute devoted to privacy, it will lead IBM's participation within the scientific and technical communities in this important field.

The Institute has created a global research program to develop new privacy-enhancing services and technologies. It will address such issues as privacy in e-business, pervasive and mobile computing, knowledge management, and intrusion detection. See the examples below for more details.

http://www.research.ibm.com/privacy/projects.html#watermark

> Doug
>
> -----Original Message-----
> From: jpoindexter
> Sent: Sunday, May 19, 2002 11:24 AM
> To: tarmour; 'David Gunning'; ddyer
> Cc: tsenator; rpopp
> Subject: RE: Dr. Latanya Sweeney's Home Page
>
>
> Tom, Ted and Doug,
>
> She does sound interesting.
>
> Ted,
>
> Does she have funding through EELD to investigate privacy protection
> technologies?
>
> Doug,
>
> Are there any proposals in this area?
>
> Let's think about her for the CSTB study.
>
> -----Original Message-----
> From: tarmour
> Sent: Saturday, May 18, 2002 3:53 PM
> To: 'David Gunning'; jpoindexter; ddyer
> Cc: tsenator
> Subject: RE: Dr. Latanya Sweeney's Home Page
>
>
> Thanks, Dave.
>
> John and Doug-
>
> You may already know of this person. She appears to be involved in Ted's
> Biosurveillance program...according to her cv,
> http://sweeney.heinz.cmu.edu/cv.html,
> "My work specific to this project centers on rendering data sufficiently
> anonymous data in real-time and of course, demonstrating the ability to
use
> such data for detection purposes."
>
> If she's not already involved in either of our privacy studies, she would
be
> a good candidate. We may even want to fund her to do work for us.
>
> Here's how she describes her main interest:
>
> "Applying these pursuits to the task of maintaining privacy and
> confidentiality in a globally networked, technically empowered society has
> allowed me to explore a new area of computer science, which I term
> computational disclosure control (also known as "data anonymity"). This is
> the study of computational techniques for discovering and controlling
> inferences that can be drawn from disclosed data such that the identity of
> individuals and other entities contained in the released data cannot be
> recognized even though the data remain practically useful."
>
> BTW...most of her work seems to have been on medical data, inspired,
> perhaps, by HIPAA.
>
> And she makes this statement:
>
> "I also work on policy issues. Maintaining privacy and confidentiality in
an

> electronic setting requires a symbiotic relationship between technology
and
> policy."
>
> Indeed!
>
> She runs the "Laboratory for International Data Privacy" at CMU
> http://sos.heinz.cmu.edu/dataprivacy/index.html
> whose mission is described thusly:
>
> "The overall objective of the Laboratory for International Data Privacy
> (LIDAP) at Carnegie Mellon University is to provide intellectual
leadership
> to society in shaping the evolving relationship between technology and the
> legal right to or public expectation of privacy in the collection and
> sharing of data. The LIDAP is unique and necessary because its
> interdisciplinary perspective within an academic setting provides a
vantage
> point for examining data privacy problems across different kinds of
> applications and data, including financial, economic, criminal, medical
and
> genetic information, to name a few, in various socio-political settings.
The
> LIDAP constructs balanced, integrated solutions across boundaries, weaving
> technology and policy together."
>
> -t
>
> > -----Original Message-----
> > From: David Gunning [mailto:david@gunning.ws]
> > Sent: Wednesday, May 15, 2002 9:43 PM
> > To: tarmour
> > Subject: Dr. Latanya Sweeney's Home Page
> >
> >
> > Tom,
> >
> > Here is a link to Latanya Sweeney's home page. She is the
> > CMU researcher on
> > privacy and may have something interesting you could use in your talk.
> >
> > http://sweeney.heinz.cmu.edu/
> >
> > Dave
> >

ddyer

From: rpopp
Sent: Sunday, May 26, 2002 8:59 PM
To: ddyer; jpoindexter
Subject: RE: Recommendations on Acxiom

Doug, did you broach w/ Acxiom the subject of estimated costs for performing #1 & #2??

-----Original Message-----
From: ddyer
Sent: Tuesday, May 21, 2002 11:41 AM
To: jpoindexter; rpopp
Subject: Recommendations on Acxiom

John, Bob,

I've spoken about these recommendations to both of you separately, but to recap:

Acxiom is the nation's largest commercial data warehose company ($1B/year) with customers like Citibank, Walmart, and other companies whose names you know. They have a history of treating privacyissues fairly and they don't advertise at all. As a result they haven't been hurt as much as ChoicePoint, Seisint, etc by privacy concerns and press inquiries. Essentially, Acxiom buys or otherwise acquires transaction information, uses a key, proprietary technology they call Abilitec that links these transactions uniquely with a person/address pair (UUIDs for both), and then projects from this database to provide datasets to their customers. Abilitec is important because it is about 99% accurate despite differences in spellings and errors in data and because of high performance (hundreds of millions of these links/second on current hardware). Customer data sets Acxiom creates can be relational databases or in some other structured format. Acxiom also hosts supercomputers that enable their customers to do analysis and data mining. One example of why this data mining pays off is to determine who to send pre-approved credit card applications to. They also do mass mailings and even host a few high-performance web sites.

Coverage claimed: Acxiom spends about $50M for data on US data and covers more than 80% of the population. They have 80% coverage in the UK and have some coverage in Australia, Canada, and Germany. They are interested in expanding when it makes sense financially and do continual analysis to assess costs and payoffs (so they know many of the data sources that exist around the world---and have incentives to keep up to date).

I think we can win with Acxiom in four ways:

1) Engage Acxiom in conjunction with the Rand study to identify all the relevant databases (without knowing Rand's knowledge base, I'm guessing that Acxiom has more current knowledge about commercial databases).

2) Have Acxiom provide us with a statistical data set (using UUIDs rather than any identity or address information) for use in the TIA critical experiment (I don't know if we have a name for this yet, but it's the one which involves discovering the red-team signature, discerning bad behavior from odd or normal behavior). We can use this real, large, but private data set to accelerate our critical experiment.

3) Acxiom's Jennifer Barrett is a lawyer and chief privacy officer. She's testified before Congress and offered to provide help. One of the key suggestions she made is that people will object to Big Brother, wide-coverage databases, but they don't object to use of relevant data for specific purposes that we can all agree on. Rather than getting all the data for any purpose, we should start with the goal, tracking terrorists to avoid attacks, and then identify the data needed (although we can't define all of this, we can say that our templates and models of terrorists are good places to start). Already, this guidance has shaped my thinking.

4) Ultimately, the US may need huge databases of commercial transactions that cover the world or certain areas outside the US. This information provides economic utility, and thus provides two reasons why foreign countries would be interested. Acxiom could build this mega-scale database.

At any rate, there is little or no chance that Acxiom or any other commercial data warehouse/mining company is likely to advance the state of the art. Any innovation DARPA might pay for would be quickly intertwined with proprietary technologies. This doesn't reduce the security/economic utility of a system like the one described in (4).

Cheers,

Doug

Lt Col Doug Dyer, PhD
DARPA/IAO
3701 N. Fairfax Drive, Arlington VA
Voice: 703-696-7442

Re: JE Moving to Office of Force Transformation?

By John M. Poindexter on 1/5/02 9:24 PM

Mike,

This is a very thoughtful contribution that I hope everybody will study and be prepared to discuss how we can implement your objectives here. I agree it is not too early to start thinking about this. It may be too early to start talking about it. I had posted a note some weeks ago that we should define the end game which is beyond DARPA so that we all have a vision of where we are going.

Bran Ferren's USA Vision

By John M. Poindexter on 2/10/02 1:41 PM

I've gotten permission from Bran Ferren to distribute his paper on Universal Situation Assessment which I have uploaded here USA White Paper Draft (UV1.1).doc. He says I can distribute as I please. He was going to clean it up a bit, but he is satisfied with it.

Highly recommend reading to get a slightly broader vision than the present focus of TIA. I endorse his ultimate vision -- maybe that's the end game for TIA.

He has some good technology ideas in one of the appendices. We should be sure to capture these ideas in the BAA.

Catalog of Transactional Data Bases

By Thomas Armour/DARPA-IAO on 3/5/02 11:25 AM

-----Original Message-----

From: jpoindexter

Sent: Saturday, March 02, 2002 4:58 PM

To: ddyer; tarmour(contr-iao); tsenator; Brian Sharkey (E-mail); Art Fritzson (E-mail)

Cc: rpopp

Subject: Catalog of Transactional Data Bases

All,

I've been thinking it would be very useful and probably a requirement to have a catalog of all available data bases that contain transactional data that might be relevant to our problem domain. I have in mind classified, unclassified, government and commercial (both internal and for public access). This would be a monumental effort -- also somewhat sensitive.

There may be pieces of this already inventoried someplace.

I'm thinking it would include description of data base, who owns it, what is the coverage, type of data base, schema, frequency of update, access control, etc. The first step would be to scope the problem, develop a framework for doing review, identify a comprehensive list of relevant transaction categories, and collect information on a hundred or so data bases. This would not be just domestic, but try to scope the world-wide problem.

This would be extremely valuable for Doug's work on Genisys, provide ideas for Ted's work, and beginning to define a long range plan.

The question is what would be a good organization to do the initial work on this. I've been thinking about RAND, but not firmed up on this.

Any thoughts and recommendations on this would be appreciated.

Doug's Response

By Thomas Armour/DARPA-IAO on 3/5/02 11:26 AM

Monday, 3/4

I think is required.

Jim Hendler and others just completed a database review for the Air Force---and there were tons of "just Air Force" databases, so this is a very large task, but as you say, some summary information must exist. We can't expect to get the entire answer, but a partial answer will be needed---our goal is to cover everything...

I don't have any better answer about who could do this task.

Doug

Brian's Response

By Thomas Armour/DARPA-IAO on 3/5/02 11:27 AM

Monday, 3/4

John,

I think this is a very good idea and RAND is a reasonable choice given that

you may not wish to directly engage with firms like Gartner and rather have

that connection made via a Think Tank that can get access to Gartner and

others. Also, depending on progress made on the policy side, it is possible

that there will be "partial" relaxation of restrictions as the government

probes the laws and gets a better understanding of the consequences of these

legal decisions. A firm like RAND could help with both the data collection

and the strategy for probing the relaxation of policy as a function of the

value and sensitivity of the transaction data.

Ted's Response (Study TOR)

By Thomas Armour/DARPA-IAO on 3/5/02 11:28 AM

Tu, 3/5

Colleagues, Here is a first draft Terms of Reference for the database cataloging effort that John proposed. Comments, suggestions, etc. are solicited and welcome. -- Ted

Terms of Reference.doc

Comments on TOR

By Thomas Armour/DARPA-IAO on 3/5/02 11:32 AM

1. Use policies might include classification (if any).

2. I like the idea of dividing the job up into domains that would then constitute chapters in the mother-of-all book of databases. It occurs to me, however, that defining these domains is itself a pretty big job, and should perhaps be Task #1 in the study. I presume we want these domains to be mutually exclusive, yet together spanning the entire space. But it may not be possible (now that I think about it), probably impossible to make them mutually exclusive. So will need to know when we have identified the same database in several domains and only do the detailed analysis once. Need a good global index, I guess.

Also...another way to parse the space might be use...somewhat different from industry. So instance, these guys use transitional data to fulfill goverment regulatory requirements:

http://www.economic-consulting.com/practiceareas/TransferPricing/transacdatab.html

3. This is probably a fine point not worth doing anything about...but just to be clear. The goal of Total Information Awareness is to identify and be able to exploit everything. (Isn't it?) Clearly this must be done in stages, and will not be complete in the lifetime of the TIA program. But that is the goal. So purpose of this study is a) to identify the low-hanging fruit and most

valuable databases for early inclusion and b) to produce a (strategic) plan that would define both the technical capabilities necessary to exploit *all* of the various database types, and a notional schedule to inclusion of databases in the program. IMHO, anyway!

4. As to the performer...Rand seems like a good choice. Other possibilities: IDA. One of the DoE labs...Sandia, PNNL, Livermore especially (since they do intel work).

Re: Comments on TOR

By John M. Poindexter on 3/5/02 8:21 PM

Good comments.

Mike's response

By Mike Szymanski on 3/6/02 12:18 AM

Why not go to someone in this specific line of business?

Seisint is a global information management and technology company. Seisint's technology solutions enable rapid access to, manipulation and analysis of large-scale data stores that may contain billions of records. Seisint currently houses one of the worlds largest US-centric data repositories (on individual and business) developed from both public and commercial sources.

Re: Mike's response

By John M. Poindexter on 3/7/02 10:51 PM

Mike,

I think it is useful to have an FFRDC do the initial scoping work because of the sensitivity.

Re: Mike's response

By J. Brian Sharkey on 3/8/02 6:29 AM

John,

Just to capture what we discussed yesterday, in addition to the transaction data types, we sjhould capture whether or not there are "electronic" or other methods for automatically collecting the information ans the "plan for a plan" should include a recommendation of firms/organizations who we might hire to collect the various data activities as the policy issues become resolved.

Brian

Items of particular5 interest

By Thomas Armour/DARPA-IAO on 3/12/02 5:04 PM

Note item 5i in the TOR:

i. Recommend appropriate intelligence support organization. Consider innovative ways to accomplish the intelligence function without generating a large manpower bill.

And 10b:

b. Plan for permanent interagency presence and interagency coordination mechanisms (e.g., joint interagency coordination groups). Access to fused intelligence, law enforcement information, and the ability to coordinate operations with non-DOD agencies will be particularly important.

Re: Items of particular5 interest

By John M. Poindexter on 3/12/02 7:30 PM

Yes, I noticed those parts. I think Mike McConnell should make early contact with the staff elements working on this to tell them about TIA. Maybe we can have an impact on their early thinking.

Art, please bring this to Mike's attention.

Re: Items of particular5 interest

By Art Fritzson/Booz Allen Hamilton on 3/15/02 1:28 PM

Will do

GeoSpatial and Temporal Tracker

By John M. Poindexter on 3/24/02 3:39 PM

I have thought for sometime that we will need a geospatial display of information in TIA. The technology described below might be a short-term or long-term solution.

ArcView Tracking Analyst Extension.url

(quoted from email)

Dr. Poindexter

We have been working on FCS with Gary Sauer and he mentioned that you may be

interested in some the geospatial solutions that we have developed/are

developing. We have done a number of things with terrain analysis for all

the services along with having developed a commercial product that ESRI,

Inc. (www.esri.com) sells called Tracking Analyst

(http://www.esri.com/software/arcview/extensions/trackingext.html older

version- new one coming out within next two months). TA will track several

hundred entities and display them in a map viewer (more specifically ESRI's

ArcMap) in real time while also having a playback feature that allows for a

variety of historical analyses. We would be happy to meet with you and give

you a demonstration of our products/capabilities at your convenience.

Bill Foshay

Discussion with John Hanke, CEO of Keyhole, on EarthViewer

By John M. Poindexter on 4/13/02 9:08 AM

John,

Very nice article. It is close to my vision. I'm sorry I won't be able to make the 16 April meeting. I will definitely want to get with you at a later date. Please let me know when you will be in town again.

For sometime now I've been searching for a better way of visualizing the InfoSpace beyond screens full of text. A linkage between GeoSpace and InfoSpace seems like a natural way of taking advantage of human cognition since we live in the physical world. I've thought of a briefing for a senior decision maker that starts with a view of Earth from 100 miles where you can see the hot spots. As you fly down more and more detail becomes visible. As long as the information you are looking at has a strong geospatial component, the earth remains the background. At some point the scene morphs into another space such as one representing the social networks of terrorist organizations for example. Then morphs into a space where relationships and inferences can be seen in some sort of visual diagrams. At the bottom you finally can drill into the text documents that provide the fine grain detail -- maybe somewhat along the lines of Inxight's hyerbolic tree or vanishing wall where time is the prominent dimension.

To some extent this has a little flavor of Gelertner's MirrorWorlds.

Some questions our technical people will have will be how to link in new information layers, how flexible is this in a dynamic situation, can the EarthView window be made bigger (I suppose this takes a faster connection and processors), is there any work going on with NIMA. Also it would be good to get a couple more copies of EarthViewer for evaluation.

-----Original Message-----

From: John Hanke [mailto:john@keyholecorp.com]

Sent: Saturday, April 13, 2002 2:02 AM

To: jpoindexter

Cc: Brian Sharkey (E-mail)

Subject: RE: EarthViewer

john,

looks like the meeting is on for tuesday. if there is interest, i hope to catch you at a future date. fyi, we'll be in the NYT circuits section this coming thursday. also, this just hit in "the bulletin" (published w/ newsweek) in australia. very enthusiastic.

http://bulletin.ninemsn.com.au/bulletin/EdDesk.nsf/printing/340A80E83D9E2DC8CA256B8E000B3CA6

E-BULLETIN

EARTHVIEWER'S ROCKET SCIENCE

You're flying at warp speed with a 3D view of not only the planet but even your own street. This isn't science fiction, Joshua Gliddon discovers, but high-tech reality.

Hurtling out of the clouds at what must be Mach 5, features on the ground - fields, roads, a house - lose their fuzziness as you plunge groundwards. What looks like a Martian canal turns out to be a bomb-cratered runway. Welcome to Kandahar, Afghanistan. The war on terror is underway, but you're no paratrooper. Kandahar has been rendered in high-resolution three-dimension by a program called EarthViewer.

Geographical Information Systems (GIS) and high-res satellite imagery have been around for years. The thawing of the Cold War liberated terabytes of imagery used by the Americans and the Soviets to track each others' troops and nukes, and these have been used by governments and business to track everything from population increases to wayward icebergs.

There's just one problem. The images are flat, static and boring. Keyhole, the parent of EarthViewer founded by a group of ex-Silicon Graphics employees, marries this high-res imagery with topographical data and pushes it all through a super-powerful, console games-inspired 3D engine.

If you thought that seeing pictures of the Big Blue Marble from space was life changing, wait until you fly, in real time, across the globe before plunging down into your country, your city, your town and even your street before screaming to a halt above your home. No one else does what EarthViewer can do in real time, which is why it's so amazing.

"EarthViewer works because you don't need to have all the data in a local database," says Keyhole CEO John Hanke. "And we're able to use a powerful games engine to render all the data."

You get the sense that Hanke wants to understate the true potential of EarthViewer. After all, it's a small company, headquartered in California, with less than 20 employees on board. And in one sense it's wise he does understate the potential. After all, coverage of the US and Afghanistan isn't the entire Earth. Getting to that level means gaining access to the many databases of high resolution images that are around, but not publically available. Australians may have to wait a while until we can drill down into Sydney. But it will happen in the next two or three years, says Hanke. "A

full resolution map of the Earth would need about six petabytes of storage. That's completely feasible as long as we can get the data."

Sony has injected some capital, and a couple of deals with real estate boards have been inked. Expanding the coverage is simply a matter of sourcing the images, says Hanke, 36. "Once we've got them we can just plug them into our software and it's there."

So where did Keyhole get the idea for a 3D world from? Hanke admits that pictures of Earth are nothing new, and pays homage to the early cartographers who simply wanted to build a more accurate model of the Earth.

We've got an intuitive understanding that the world is a big place, but it's not until we spin the globe and plonk down a finger - "I'll go there" - that we start to think about the pure logistics of movement through space.

Looking at the globe gives us a better understanding of ourselves.

Think of EarthViewer, then, as tomorrow's globe, but with infinitely high resolution. And one that can be toted around in a laptop or beamed across the (coming) wireless internet into a handheld computer or tablet device.

"There's also precedent for what we've done in science fiction," Hanke says. "If you read Neal Stephenson's Snow Crash, which was written 10 or 12 years ago, he talks about a program that's similar to EarthViewer."

Sun Microsystems' chief scientist John Gage used EarthViewer as the backdrop to his presentation at February's World **Congress** on Information Technology in Adelaide. At a later, private presentation, he ran an automated tour of Afghanistan and the US using EarthViewer. Attendees were transfixed as they came hurtling out of the clouds towards the bomb-cratered runway.

Gage later qualified his enthusiasm for the software, saying that Sun had

nothing to do with the company, but that he thought such a global view of the Earth was unprecedented. "It's what you can do with a little imagination and good software skills," he said. He should have added "and with good, high-resolution imagery, too", because without the pictures EarthViewer is nothing.

Computer interfaces are mired in the old Windows-Mouse-Pointer model that was developed nearly three decades ago at Xerox's Palo Alto Research Centre. Apple ripped it off PARC, and Microsoft ripped it off Apple. The problem is that the old graphical user interface is starting to creak and tear apart with the sheer volume of information that today's computers can deal with. When PARC researchers hit upon the idea of using a desktop as a metaphor for organising work, computers could store a couple of hundred text documents at best.

Now we're able to store the entire Australian National Library on a network drive. Google provides a command-line driven window onto the world wide web. It can also search pictures and news groups, but wading through hundreds of documents before you get to the one you're after is a hit-and-miss affair.

Organising information spatially could be the answer for certain types of information. EarthViewer shouldn't just be thought of as a fancy, games-engine driven globe. It's actually a sophisticated geographical interface. "Any information that's spatial, or related to spatial data, can be linked into our software and presented using EarthViewer," says Hanke. "If we go beyond the literal stuff, like real-estate applications, then we're able to start looking at how wealth is distributed, at the links between crime data and geography, and we can represent that using colours

rather than simple button icons."

He's still underselling it. EarthViewer is a teaching aid and communication device. The runway at Kandahar could be linked to Afghan demographic information, which could then provide a link to regional geopolitical data. Scenarios could be played out, old imperialist boundaries redrawn. Throw in some artificial intelligence like that used in games like The Sims, and you've got a virtual, simulated planet. Add weather data, global-warming information and other environmental factors and you could have the entire Earth in your pocket. It's all possible, and it's almost here. You may never look at the Earth the same way again.

> From: jpoindexter [mailto:jpoindexter@darpa.mil]

> Sent: Monday, April 08, 2002 6:09 PM

> To: 'John Hanke'

> Cc: Brian Sharkey (E-mail)

> Subject: RE: EarthViewer

>

>

> John,

>

> Got it installed on my home machine which has a better video

> adapter. What

> a difference. It is really impressive.

>

> Brian Sharkey heads a team that is advising me on technology concepts and

> system architecture. They arrange for demos on interesting technology and

> are next door to DARPA. They may be prepared to host a demo on 16 or 17

> April. I am tied up both days, but I don't think I need to see

> it after my

> own viewing of the application. They would invite the relevant program

> managers from my office. Please coordinate with Brian. If we

> could plan it

> further in advance I could probably make it.

>

> Brian,

>

> This is a very interesting geospatial viewing application which may have

> enormous potential for our purposes in TIA. It has the capability of

> showing layers of data with geospatial coordinates. Bill Crowell, former

> dep dir NSA, brought it to my attention.

>

TIA System Architecture: TIA System Design Description DRAFT Version 1.0 Posted

By Bradley R. Bebee on 4/30/02 11:43 PM

A draft version of the TIA System Description Description DRAFT Version 1.0 (tia-sys-arch-2002-04-30.doc) has been added to the Files tool. This draft is under revision and will be updated frequently. One important upcoming addition is a list of conceptual operations for each TIA technology component to help in specifying potential uses.

The current draft is in Word format. In the near feature, it will be converted to XML and be available for browsing on the TIA System Portal.

Comments and feedback are welcomed...

Thanks, --Brad

Please Review ASAP

By John M. Poindexter on 5/1/02 8:09 AM

Brad and Greg,

Great job. We need to prepare a PPT briefing on this with the hilites for my use with various groups and DIRO.

IAO Staff,

Please review this as soon as possible -- especially to have context for BAA proposal evaluations.

Word of Caution

By John M. Poindexter on 5/1/02 9:57 PM

We need to be alert to fuctional areas that are not covered by current programs and be willing to propose and implement new programs where needed to achieve the functionality we feel is important.

Re: Please Review ASAP - Preliminary

By Greg Mack/Syntek on 5/1/02 1:06 PM

Pl ease keep in mind that this is preliminary. The overall framework is going to change soemwhat to make it consistent with some of the standards (e.g. ISO's & ITU-T's Reference Model for Open Distributed Processing, and the Joint Technical Architecture). We have also met

with BAH this morning to synthesize it with their Core Services Model. Comments are appreciated.

Re: Please Review ASAP - Preliminary

By John M. Poindexter on 5/1/02 7:33 PM

I know this is preliminary, but I want BAA proposal reviewer to have the system context in mind.

It is a living document and our blueprint -- a very important document.

Hybrid Collaboarative Environment

By John M. Poindexter on 5/2/02 1:30 PM

I would like for the TIA Sys Architecture document to eventually capture our ideas for the collaborative environment.

Re: Hybrid Collaboarative Environment

By Art Fritzson/Booz Allen Hamilton on 5/2/02 1:55 PM

We've characterized the System View as containing two transport platforms - the collaborative platform and the center based services architecture (to support application integration). I don't think we've quite got the mapping right between the Operations Model and the System Model but, for the time being, the System View (of the System Model)provides a placeholder for elaboration. We've uploaded our initial cut at this and the core services (the Services View of the Operations Model) into the TIA SDD Development shared space. We've also uploaded a description of the Grid Prototype work there.

Re: Hybrid Collaboarative Environment

By John M. Poindexter on 5/2/02 8:27 PM

Art,

Ok. We are working on a cartoon type diagram that captures more accurately the hybrid environment and will provide that for inclusion when we finish it.

An interesting confounding factor -- identity theft

By Greg Mack/Syntek on 5/13/02 1:19 PM

[excerpt from NY Times] "Tens of thousands of stolen credit-card numbers are being offered for sale each week on the Internet in a handful of thriving, membership-only cyberbazaars, operated largely by residents of the former Soviet Union, who have become central players in credit-card and identity theft. ... a credit-card reseller's conference in Odessa, Ukraine, at the end of this month. ... The Internet sites of the online marketplaces are mostly known only to their participants though that number can run as high as 2,000 registered users. "

http://story.news.yahoo.com/news?tmpl=story&ncid=68&e=2&cid=68&u=/nyt/20020513/ts_nyt/credit_card_theft_thrives_online_as_global_market_losses_grow

---

Re: An interesting confounding factor -- identity theft

By John M. Poindexter on 5/15/02 7:54 PM

This is an important theme for us. Identity theft is a very important problem that makes positive identification essential which in terms helps the solution to our problems.

Doug, please pass along this idea to ISAT study.

The Battle Must Be Fought Overseas

By Thomas Armour/DARPA-IAO on 5/28/02 8:30 AM

Interesting op-ed today in the NY Times by a former CIA officer. Among other things, he opines that Americans won't have the stomach to accept signficantly increased security...and invasions of privacy...domestically:

"While **Congress** and the administration hope the F.B.I. can become more adept in pre-empting terrorism on American soil, enhancing that ability would probably mean using extensive ethnic and religious profiling that could come to dominate numerous aspects of American life. It is a decent bet that many, if not most, Americans would have little stomach for a real war on terrorism on the home front."

He's also not optimistic about the CIA getting it done:

"Unfortunately, it is doubtful that any real change will happen in the C.I.A. without outside pressure."

http://www.nytimes.com/2002/05/28/opinion/28GERE.html?todaysheadlines

Version 1.0 Integrated TIA System Description Document Posted

By Bradley R. Bebee on 7/7/02 6:15 PM

All,

I have posted Version 1.0 of the SDD (tia-sdd-2002-07-08.pdf). This document has been formatted and edited for release and distribution.

Key changes to this version include:

- Integrated release between the Business Model (HAI) and the Systems Model (BAH)

- Applied DARPA IAO Document Template

- Updated Information View to include TIA Domain Object and Semantic Architecture concepts

- Added the initial version of the Business Application Services View

Thanks, --Brad

Re: Version 1.0 Integrated TIA System Description Document Posted

By John M. Poindexter on 7/8/02 6:54 AM

Brad et al,

Thanks.

Privacy Officer Is Possibility at Security Department

By Thomas Armour/DARPA-IAO on 7/10/02 8:46 AM

WASHINGTON, July 9 - Under Congressional pressure, the Bush administration said today that it was open to the idea of installing a chief privacy officer in a new Department of Homeland Security to make sure it weighed issues of confidentiality and the secure handling of personal information.

http://www.nytimes.com/2002/07/10/national/10PRIV.html?todaysheadlines

---

Re: Privacy Officer Is Possibility at Security Department

By Ted Senator/DARPA-IAO on 7/10/02 12:31 PM

Perhaps we should have a Privacy Officer for TIA, officially designated as part of our team.

Re: Privacy Officer Is Possibility at Security Department

By John M. Poindexter on 7/10/02 6:10 PM

We are working on a liason officer from Justice and of course Mike McConnell and his staff are providing expertise on current and future policy.

IAO In the News

*By Thomas Armour/DARPA-**IAO** on 9/3/02 7:03 PM*

Despite the scarrrrrry kicker:

"You're not paranoid, you are being watched: As tracking tools improve, true privacy may be lost"

it's reasonably balanced. We're mentioned here:

"One of many technology projects begun or accelerated after Sept. 11, an effort headed by former National Security Adviser John Poindexter, is trying to find out.

"We're still a long way from a society where government tracks and records our every move. Poindexter's work at the Defense Advanced Research Projects Agency is just that - research. His group is working with simulations, deferring decisions on what databases to include and how governments would get industry's cooperation. He acknowledges the privacy concerns and says policy-makers are beginning to discuss how to address them."

http://www.msnbc.com/news/800957.asp?0si&cp1=1&cp1=1

For Review: TIA Technical Note: Rapid Inclusion in the TIA System Experimentation Process

By Bradley R. Bebee on 11/14/02 3:12 PM

Admiral Poindexter,

I've uploaded a TIA Technical Note, Rapid Inclusion in the TIA System Experimentation Process, for your review (tia-standards-se.doc). In addition, I've created a directory specifically for TIA Tech Notes.

Thanks, --Brad

This document provides an overview of the system-wise steps to experimentation and a core set of recommendations for Software Engineering practices to facilitate the use of technology components within the TIA System Experimentation Processes. It is intended as an overview for Principal Investigators, Technical Leads, and Software Developers.

The goal of this document is to answer the questions:

1. How can our technology component be explored as part of the TIA Technology Exploration and Experimentation processes?

2. How should we build our technology component to facilitate integration into the TIA System?

This document is the first in a series of four documents aimed providing information to technology component developers:

1. Rapid Inclusion in the TIA System Experimentation Process (HAI/BAH)
    a. This document.
2. Interfacing with TIA Core Services: Programmers Guide (BAH/HAI)
    a. What are the TIA Core Services interfaces?
    b. How can we build and test our component using the Core Services?
3. Developing TIA Applications on the Groove Platform (BAH/HAI)
    a. What are the TIA recommendations for building collaborative applications on the Groove Platform?
    b. How should Groovy components use the Core Services?
4. Preparing for TIA Transition and Hardening (BAH)
    a. How is a component selected for transition and what is the transition process?
    b. How are intellectual property rights protected for proprietary and patented components during transition?

Re: For Review: TIA Technical Note: Rapid Inclusion in the TIA System Experimentation Process

By John M. Poindexter on 11/14/02 7:08 PM

Brad et al,

This looks like a really great approach. We will discuss this first version of the first of the series with the PM's at the offsite next week. Well done.

Re: For Review: TIA Technical Note: Rapid Inclusion in the TIA System Experimentation Process

By John M. Poindexter on 11/14/02 7:09 PM

Leanne,

Please make 25 color copies of tia-standards-se.doc and bring to offsite. Thanks.

Press Guidance

By John M. Poindexter on 11/14/02 9:22 PM

Here is the media guidance being used by DARPA PA and OSD. The first one is the one being used by SecDef. The other 2 are held in DARPA PA and may be used. Media Guidance

It would be helpful to either refer media requests to DARPA PA (Jan Walker) or at least coordinate with her before any company statements are made.

NY Times Article on use of private data by government, mostly law enforcement

By Ted Senator/DARPA-IAO on 12/10/02 8:21 AM

though we get a mention too. see New Tools for Domestic Spying, and Qualms.htm

as usual, one critique is that data mining techniques don't work well enough and that many commercial companies are selling snake oil. i agree with these critiques -- that's why we need darpa r&d, and why there is a danger in just deploying what was invented for other purposes (marketing) and companies want to sell.

Atlantic article suggests tide is turning (attached)

By Mike Szymanski on 12/12/02 1:04 AM

Just when I had given up on intelligent life in the media universe.....

More good news (attached; btw: nice pic under pressure)

By Mike Szymanski on 12/12/02 1:47 AM

Re: More good news (attached; btw: nice pic under pressure)

By Ted Senator/DARPA-IAO on 12/12/02 7:18 AM

two good articles that set the record straight. if they each had another paragraph or two about how we're developing privacy technologies too they'd be a little more complete.

Canada too

By Ted Senator/DARPA-IAO on 12/13/02 1:23 PM

they also recognize the need to collect and use different types of info than before. Canada Plan to Gather Travel Data Criticized.htm last paragraph is particularly interesting -- they collect more than they need because it's easier (i.e., less expensive) for private sector to send in everything than to do filtering (by attribute or by entity). At FinCEN, we saw the same thing -- the banks complained on the one hand that there were too many useless filings, and on the other, when we (**Congress**, actually) changed the criteria to require less, still sent in everything. I think a technical challenge for **IAO** is to develop the technology that makes it possible to INEXPENSIVELY filter data at the source to only those records and attributes of potential interest, and to purge the records/attributes from the collected data that don't turn out to be worthwhile.

Re: Canada too

By John M. Poindexter on 12/13/02 8:11 PM

My concept is that you leave all of the data where it is and just do queries on the distributed data bases.

Re: Canada too

By Ted Senator/DARPA-IAO on 12/13/02 8:54 PM

John, I think your concept works for what I've called secondary sources -- which are those that get queried based on partially instantiated patterns (the degenerate form of which is a subject based query). But the partially instantiated patterns have to come from somewhere -- i.e., the primary sources -- which are analyzed continuously with entity extraction, consolidation,

linkage, etc. These primary sources would be databases the Government already has access to in the intelligence and law enforcement communities. So your concept works fine -- all the Government needs is the primary sources and the query results from the secondary sources.

Clarification of Barbara Simons role and detailed discussion from Doug Dyer

By Mike Szymanski on 12/24/02 11:17 AM

-----Original Message-----

From: ddyer

Sent: Thursday, December 19, 2002 11:48 PM

To: 'Ed Lazowska'; Jack Thorpe; jpoindexter; bill.mark@sri.com; felten@cs.princeton.edu; Pat Lincoln; Rich Entlich; tarmour; knumrych; trohrkemper

Cc: Teresa Lunt (teresa.lunt@parc.com)

Subject: RE: Message to Markoff

Ed, I think this is OK. In fact, I think it's great. I don't understand why we're getting so beat up by the New York Times (and others). A year ago, they were beating us up for not connecting the dots. Now they're beating us up for trying to... and our efforts to ensure personal privacy make no difference. Where's the justice in that? I guess this is just how the news business works... you've got to sell it, so it has to be provocative.

Here are the facts as I see them.

Privacy protection:

* has been understood to be a central requirement from the beginning of the Total Information Awareness (TIA) program

* was part of our research agenda published in March 2001 in a broad area announcement and proposers information pamphlet

* is the topic of two important DARPA-sponsored studies and continues to foster new thinking in other studies we appreciate, but don't sponsor

* is the goal of the Genisys Privacy Protection program at DARPA which is just now getting underway, and

* has influenced many architectural decisions in the emerging TIA system prototype including the extensive use of synthetic data for technology development and benchmarking.

Prior to the ISAT study, Dr John Poindexter, his staff, and contracted researchers had identified anonymization and filtering as appropriate privacy protection technologies. The ISAT Security with Privacy study, suggested by ISAT members but immediately sponsored by Dr Poindexter, identified several additional privacy protection methods including selective revelation, which, in context of data

mining with anonymized transaction data, is a model of incremental release of additional information that is either authorized or not authorized based on evidence that accumulates through pattern matching. Only in the final stages of selective revelation, that is, only if enough evidence accumulates would an investigator or analyst be able to get access to a suspect's identity, and this would occur as it does today, under appropriate authority such as a judge issuing a subpoena. Many other useful technologies were identified by ISAT such as simplifying technologies for managing and effecting privacy protection policy and cryptographically protecting data and hardware. In the course of the study, DARPA suggested and study members endorsed the idea of "immutable audit," the notion that any inappropriate data mining activities conducted could be immediately captured by strong accounting systems that could not be turned off or altered. Technology aimed at providing immutable audit would make it difficult or impossible for abusers of the information system to avoid getting caught and punished---study members provided important insight into what it would take to create this technology. Subsequent to the ISAT study, DARPA developed ideas for "self-reporting data," that is important data that reports on its location and status periodically. Self-reporting data is useful for tracking important data even beyond the initial release from a database (ordinarily, the database is the only place where accounting takes place). For example, if an authorized requestor gets data from a database but then passes it directly to another person, self-reporting data would enable auditors to know who has accessed it. Self-reporting data makes use of and extends another well-known concept for tracking the origin and identity of data known as digital-watermarking. DARPA's architecture for TIA assumes a federation of existing commercial and government databases that have been integrated and simplified via query mediation (the Government does not want to create or maintain a huge database of its own). Access control is governed by existing laws and policies. To make this work, DARPA's TIA architecture also includes the idea of a simple appliance that sits on top of current databases and implements many of the privacy protection techniques identified using easy-to-understand business rules including those for role-based access control and immutable audit. DARPA's concept is that a trusted third party, rather than an organization's database administrator, control this privacy guard appliance, to ensure that it works as advertised, and that the appliance be cryptographically protected, not to hide what it is doing, but to protect it from outside tampering. PARC, which is under contract to DARPA in the Genisys Privacy Protection Program, will develop inference control technologies aimed at solving the problem of gaining unwarranted and sometimes unintended aggregate knowledge that implies identity when getting information from multiple sources. PARC will also create technologies that partially automate the audit process, making it much more likely that any abuse is identified and can be addressed. PARC also has the task of implementing the privacy guard appliance described above. This is undoubtedly an incomplete list and I invite corrections and additions, but these technologies seem to represent the potential of having much better assurances of personal privacy than we have in any database system today. To recap, we have:

* anonymization: with universally unique identifiers to keep transaction data separate from identity data; there are theoretical models to quantify and ensure the degree of anonymity for any particular data set

* filters: for removing data that isn't helpful or which is not authorized by policy; a related idea is to use software agents to expunge data if policy is changed to be more restrictive

* selective revelation: a method of getting more data as appropriate when evidence accumulates during data mining; when enough evidence accumulates to support a subpoena, identity or other personal information could be obtained so that suspects can be questioned directly and/or arrested; selective revelation is an important privacy safeguard if data mining becomes used widely in law enforcement (it is rare today)

* cryptographic methods: protect data and privacy protection hardware/software from unauthorized access and tampering

* immutable audit: the notion that any abuse of private information will be captured immediately with no possibility of cover-up

* self-reporting data: enables an extension of auditing beyond the normal logging point, the database

* TIA is a federation of existing databases integrated via query mediation, not a huge new Government database

* privacy guard appliance: helps integrate the federation and implements many of the privacy protection techniques identified; simple, transparent, swap-able, runs on easy-to-understand business rules, controlled by a trusted third-party, and tamper-resistant; the privacy guard implements at least role-based access control and immutable audit

* inference control: prevents aggregate information, obtained from multiple sources, from implying personal information that should not be released; distinguishes between query pathways do and don't imply too much information

* semi-automated audit: knowledge-based, enables deeper and more frequent audits than possible using only manual methods

I think we all believe these technologies are feasible and effective for enhancing personal privacy in the context of increased use of data mining for preventing terrorist attack. If we at DARPA knew of better ideas, we would be pursuing them. It seems as if a few people just can't believe these technical ideas will work or that the government can be trusted to implement them. Unfortunately, those same people seem to focus only on one problem, privacy---they don't have any solutions to the terrorism problem. We can only prevent terrorist attacks with information. Terrorist attacks kill people---even before 9/11, terrorist attacks on American targets throughout the last decade killed that three times as many as US military killed in action (including Desert Storm) and the potential for future attack makes even 9/11 seem minor in comparison. I can't accept the notion of a privacy vs. security trade-off in the long term, and only new privacy-protection technologies will improve the situation. I know we can make TIA work and enhance personal privacy---both goals are important.

The FOIA request mentioned by John Markoff's article included references to the ISAT study---it had to by law, because we were required to provide all information about any effort DARPA was involved with that impacted personal privacy---but the ISAT study wasn't a specific response the to FOIA request. As all of us know, the ISAT study invited participation by privacy advocates including EPIC who issued the FOIA request. The article implies the Government is hiding something or doing something wrong when in fact we've bent over backwards to be open with our efforts and invite reasoned debate, as Ed has kindly pointed out.

I want to sincerely thank those who contributed to DARPA's efforts by participating in this year's ISAT study, and I apologize for the bad press and any negativity you experience because of inaccurate reporting. Many of the media reports we've seen have been quite balanced in there analysis of TIA technology. Perhaps more importantly, Dr Poindexter's briefings to those in **Congress** and other Government entities have clarified the issues and, in the overwhelming majority of cases, agreement with and endorsement of our goals and approach in the development of TIA technology. I trust the American people---it's difficult to fool them for long---I'm optimistic that eventually, we'll get past the fears and innuendo that are popular in about half the TIA articles we see today.

Happy holidays, and my best wishes to all of you,

Doug Dyer

-----Original Message-----

From: Ed Lazowska [<mailto:lazowska@cs.washington.edu>]

Sent: Thursday, December 19, 2002 2:59 PM

To: Jack Thorpe; jpoindexter; bill.mark@sri.com; felten@cs.princeton.edu; Ed Lazowska; Pat Lincoln; Rich Entlich; ddyer; tarmour; knumrych; trohrkemper

Subject: Message to Markoff

I sent the attached message to Markoff. Wanted you to be aware of it. I hope this is OK. It concerns a detail, but one that I think is important under the circumstances.

-----Original Message-----

From: Ed Lazowska

Sent: Thursday, December 19, 2002 11:57 AM

To: John Markoff (markoff@nytimes.com)

Cc: Ed Lazowska

Subject: ISAT

John --

A factual matter that I think it's important to be aware of:

Barbara Simons is not a member of ISAT.

"ISAT studies" typically include both ISAT members and individuals who are not members of ISAT. (Just like National Academies CSTB studies include CSTB members and individuals who are not.) This was true of the Security and Privacy study. The fact that someone like Barbara -- a well-known privacy advocate -- was included in the study panel indicates ISAT's strong desire (and DARPA's strong desire, and Adm. Poindexter's strong desire) to obtain the broadest possible perspective and input on the issues.

However, a quote such as "It's laughable ..." should not be attributed to an ISAT member -- Barbara is not one.

This is a politically charged issue, so details such as this matter.

(There are also some minor things, such as the fact that ISAT meets multiple times each year, and studies are conducted over the course of the full year -- but these aren't of consequence.)

Thanks. Hope you have a good holiday season. Stop by when you're next up here.

Markoff Again

*By Thomas Armour/DARPA-**IAO** on 12/23/02 8:49 AM*

Actually, despite the errors and misunderstandings in this article, I think he is actually beginning to get a clue. Or maybe it's the co-author. :->

-t

---

http://www.nytimes.com/2002/12/23/technology/23PEEK.html?tntemail0

New York Times

December 23, 2002

Many Tools of Big Brother Are Up and Running

By JOHN MARKOFF and JOHN SCHWARTZ

In the Pentagon research effort to detect terrorism by electronically monitoring the civilian population, the most remarkable detail may be this: Most of the pieces of the system are already in place.

Because of the inroads the Internet and other digital network technologies have made into everyday life over the last decade, it is increasingly possible to amass Big Brother-like surveillance powers through Little Brother means. The basic components include everyday digital technologies like e-mail, online shopping and travel booking, A.T.M. systems, cellphone networks, electronic toll-collection systems and credit-card payment terminals.

In essence, the Pentagon's main job would be to spin strands of software technology that would weave these sources of data into a vast electronic dragnet.

Technologists say the types of computerized data sifting and pattern matching that might flag suspicious activities to government agencies and coordinate their surveillance are not much different from programs already in use by private companies. Such programs spot unusual credit card activity, for example, or let people at multiple locations collaborate on a project.

The civilian population, in other words, has willingly embraced the technical prerequisites for a national surveillance system that Pentagon planners are calling Total Information Awareness. The development has a certain historical resonance because it was the Pentagon's research agency that in the 1960's financed the technology that led directly to the modern Internet. Now the same agency - the Defense Advanced Research Projects Agency, or Darpa - is relying on commercial technology that has evolved from the network it pioneered.

The first generation of the Internet - called the Arpanet - consisted of electronic mail and file transfer software that connected people to people. The second generation connected people to databases and other information via the World Wide Web. Now a new generation of software connects computers directly to computers.

And that is the key to the Total Information Awareness project, which is overseen by John M. Poindexter, the former national security adviser under President Ronald Reagan. Dr. Poindexter was convicted in 1990

of a felony for his role in the Iran-contra affair, but that conviction was overturned by a federal appeals court because he had been granted immunity for his testimony before **Congress** about the case.

Although Dr. Poindexter's system has come under widespread criticism from **Congress** and civil liberties groups, a prototype is already in place and has been used in tests by military intelligence organizations.

Total Information Awareness could link for the first time such different electronic sources as video feeds from airport surveillance cameras, credit card transactions, airline reservations and telephone calling records. The data would be filtered through software that would constantly look for suspicious patterns of behavior.

The idea is for law enforcement or intelligence agencies to be alerted immediately to patterns in otherwise unremarkable sets of data that might indicate threats, allowing rapid reviews by human analysts. For example, a cluster of foreign visitors who all took flying lessons in separate parts of the country might not attract attention. Nor would it necessarily raise red flags if all those people reserved airline tickets for the same day. But a system that could detect both sets of actions might raise suspicions.

Some computer scientists wonder whether the system can work. "This wouldn't have been possible without the modern Internet, and even now it's a daunting task," said Dorothy Denning, a professor in the Department of Defense Analysis at the Naval Postgraduate School in Monterey, Calif. Part of the challenge, she said, is knowing what to look for. "Do we really know enough about the precursors to terrorist activity?" she said. "I don't think we're there yet."

The early version of the Total Information Awareness system employs a commercial software collaboration program called Groove. It was developed in 2000 by Ray Ozzie, a well-known software designer who is the inventor of Lotus Notes. Groove makes it possible for analysts at many different government agencies to share intelligence data instantly, and it links specialized programs that are designed to look for patterns of suspicious behavior.

Total Information Awareness also takes advantage of a simple and fundamental software technology called Extended Markup Language, or XML, that is at the heart of the third generation of Internet software. It was created by software designers at companies like Microsoft, Sun Microsystems and I.B.M., as well as independent Silicon Valley programmers.

The markup language allows data that has long been locked in isolated databases, known in the industry as silos, to be translated into a kind of universal language that can be read and used by many different systems. Information made compatible in this way can be shared among thousands, or even hundreds of thousands, of computers in ways that all of them can understand.

It is XML, a refinement of the Internet's original World Wide Web scheme, that has made it possible to consider welding thousands of databases together without centralizing the information. Computer scientists said that without such new third-generation Web technologies, it would have never been possible to conceive of the Total Information Awareness system, which is intended to ferret out the suspicious intentions of a handful of potential terrorists from the humdrum everyday electronic comings and goings of millions of average Americans.

Civil libertarians have questioned whether the government has the legal or constitutional grounds to conduct such electronic searches. And other critics have called it an outlandishly futuristic and ultimately unworkable scheme on technical grounds.

But on the latter point, technologists disagree. "It's well grounded in the best current theory about scalable systems," said Ramano Rao, chief technology officer at Inxight, a Sunnyvale, Calif., company that develops text-searching software. "It uses all the right buzzwords."

People close to the Pentagon's research program said Dr. Poindexter was acutely aware of the power and the invasiveness of his experimental surveillance system. In private conversations this summer, according to several Department of Defense contractors, he raised the possibility that the control of the Total Information Awareness system should be placed under the jurisdiction of an independent, nongovernmental organization like the Red Cross because of the potential for abuse.

Dr. Poindexter declined to be interviewed for this article. A Darpa spokeswoman, Jan Walker, wrote in an e-mail reply to questions that "we don't recall ever talking about" having a nongovernmental organization operate the Total Information Awareness program and that "we've not held any discussions with" such an organization.

The idea of using an independent organization to control a technology that has a high potential for abuse has been raised by previous administrations. An abortive plan to create a backdoor surveillance capability in encrypted communications, known as Clipper, was introduced by the Clinton administration in 1993. It called for keys to the code to be held by an organization independent of the F.B.I. and other law enforcement agencies.

Speaking of Dr. Poindexter, John Arquilla, an expert at the Naval Postgraduate School in Monterey on unconventional warfare, said, "The admiral is very concerned about the tension between security and civil liberties." He added that because of the changing nature of warfare and the threat of terrorism, the United States would be forced to make trade-offs between individuals' privacy and national security.

"In an age of terror wars, we have to learn the middle path to craft the security we need without incurring too great a cost on our civil liberties," he said.

Computer scientists who work with Darpa said that Dr. Poindexter was an enthusiastic backer of a Darpa-sponsored advisory group that had been initiated by a Microsoft researcher, Eric Horvitz, in October 2001 in the wake of the Sept. 11 terrorist attacks.

The group, which was composed of 41 computer scientists, policy experts and government officials, met three times to explore whether it was possible to employ sophisticated data-mining technologies against potential terrorist attacks while protecting individuals' privacy.

A number of the scientists proposed "black box" surveillance systems that would alert human intelligence analysts about suspicious patterns. Once the alerts were issued in such a system, they suggested, legal processes like those used for wiretapping could be employed.

But a number of the scientists and policy experts who attended the meetings were skeptical that technical safeguards would be adequate to ensure that such a system would not be abused.

The debate is a healthy one, said Don Upson, who is senior vice president of the government business unit of a software company in Fairfax, Va., webMethods, and the former secretary of technology for Virginia.

"I'm glad Darpa is doing this because somebody has to start defining what the rules are going to be" about how and when to use data, he said. "I believe we're headed down the path of setting the parameters of how we're going to use information."

(b)(6) [redacted]

From: Siegrist, David [Siegrist@PotomacInstitute.org]
Sent: Thursday, May 30, 2002 8:22 PM
To: 'Joe Lombardo (E-mail)'; 'Gene McClellan (E-mail)'; 'Chung-Sheng Li (E-mail)'; 'Mark Musen (E-mail)'; 'Mike Wagner (E-mail)'; 'Karen Cheng (E-mail)';
(b)(6) [redacted] 'Rich.Wojcik@jhuapl.edu'; 'GeneMc2@aol.com'; 'Liuzzi Raymond A Civ AFRL/IFTD'; 'mcam@us.ibm.com'; 'Latanya Sweeney'
Cc: rwong(contr-iao); tsenator; rlinn(contr-ixo)
Subject: Privacy Protections Work Product and Panel

Privacy Product Spec.xls

Assuring Privacy
Session Product: Table (see attached): Showing privacy technology for Developmental and Deployed systems with advantages and disadvantages. Also, an Issue List, including specification of alternative resolutions, and action items for resolution.
NOTE: See Sheet 2 of attachment as well. (Please advise me of suggested improvements on this draft format ASAP. Thanks.)

Prior Input: Identification of Active Privacy Protection development efforts that are part of PI overview brief.

Sample questions that may be addressed in the 4 team presentations:

Need for Privacy Protection in our DEVELOPMENTAL Program
List what each project is doing in its developmental project to assure privacy
Approach to any IRB's, NDA's, etc.

Need for Privacy Protection in a DEPLOYED Program
What Privacy technology development is being conducted
What assumptions are being made?
What is the Concept of Operations?
What is the tradeoff between sensitivity and privacy?
List what each project is doing in its planned deployed system to assure privacy
- List Advantages and Disadvantages of Each
- Identify Metrics and Ways to Measure Successful Privacy Protection

Are there a Set of Privacy methods that all the Projects should apply for Developmental and Deployed Projects?

List Any Issues Identified, with Alternatives for Resolution and Action Items

TEAM	Privacy Method (Developmental)	For Data Type	Advantages	Disadvantages
Veridian/Stanford				
JHU/APL				
IBM				
Mellon Pitt				

TEAM	Privacy Method (Deployed)	For Data Type	Advantages	Disadvantages

tsenator

From: Siegrist, David [Siegrist@PotomacInstitute.org]
Sent: Friday, May 17, 2002 4:50 PM
To: tsenator
Subject: RE: privacy speaker

We can wait till our meeting. I picked her in part because [redacted] (b)(6)
picked her for the ABQ conference on similar topic and she made sense. She is politically aware, having served in Bush 1, and has some publications in the area. I want someone who is a defender of civil rights, but not an obstructionist. A leading alternative is Professor Gostin of Georgetown, who was selected by CDC and Sloan to draft the model state law on bioterrorism. However, that is getting pilloried now as draconian, so he lost some of his civil rights credibility. Other professors of law with credentials in the field tend to focus on what happens after a bio attack, rather than what may be done prior. However, there are literally hoards of privacy lawyers, because no one really knows what HIPAA means. But they tend not to have the bioterrorism insight.

There may be others out there, some better, but she satisficed, and time is growing short for this particular meeting, but I thought privacy is a suitable talk for a plenary soon.

How's your meetings?

Dave

-----Original Message-----
From: tsenator [mailto:tsenator@darpa.mil]
Sent: Friday, May 17, 2002 4:39 PM
To: 'Siegrist@PotomacInstitute.org'
Subject: Re: privacy speaker

Can we wait to discuss on tuesday am? I agree with the general idea. who are the other candidates and why do you recommend her ?
Sent from my BlackBerry Handheld.

tsenator

From: Siegrist, David [Siegrist@PotomacInstitute.org]
Sent: Wednesday, May 29, 2002 1:31 PM
To: tsenator
Subject: FW: Latanya request for more time, fyi

-----Original Message-----
From: Latanya Sweeney [mailto:latanya@cs.cmu.edu]
Sent: Wednesday, May 29, 2002 1:28 PM
To: Siegrist, David
Subject: RE: Latest Agenda; PI Summary Format; Panel info; & Matrix

Hi,

Thanks for the response. If possible, 15 minutes on the public health law survey, 20 minutes on the reasonable cause model and 10-20 minutes on the scoring of data sources. Let me know what's possible. Thanks.

--Latanya

At 09:35 AM 5/29/2002 -0400, Siegrist, David wrote:
>Hi, Latonya

> On your major points, I think your research sounds fascinating. How
>many minutes for each one do you think you need?
>
> Cordially,
> David
>
>-----Original Message-----
>From: Latanya Sweeney [mailto:latanya@cs.cmu.edu]
>Sent: Wednesday, May 29, 2002 9:26 AM
>To: Siegrist, David
>Cc: 'Ted Senator (E-mail)'
>Subject: Re: Latest Agenda; PI Summary Format; Panel info; & Matrix
>
>
>
>Hi David,
>
>I think there are three major contributions my lab's efforts could make to
>the PI meeting but in order to accomplish such would require some
>modification to the agenda. I am humbly requesting you see if you can make
>some time for these matters. They are described below.
>
>The first major contribution concerns privacy. Under this DARPA effort, we
>had a group of 40 law students and 3 lawyers who surveyed public health
>laws and regulations in each of the 50 states and DC and answered about 60
>questions for each state. The questions centered on the collection and use
>of non-traditional data sources for bioterrorism surveillance; each kind of
>non-traditional data source was examined in each state. I would like to
>take a few minutes to report on these findings. Highlights include the
>variability found and some particularities with specific kinds of data
>sources and the nature of liability concerns.
>
>The second major contribution also concerns privacy, in particular, the
>development and use of computational tools that ensure privacy. I would

>like to introduce our reasonable cause model, which provides sufficiently
>anonymous data during normal surveillance. The data from each data holder
>is rendered sufficiently anonymous and then forwarded to the surveillance
>system. We can show that detection algorithms work well with such data.
>When the surveillance system detects suspicious activity, the anonymity of
>the data is lowered. If an explicit outbreak is detected, explicitly
>identified data is provided. Under our reasonable cause model, the level
>of anonymity of the data matches the surveillance system's need for the
>data. We can show that this system removes many of the liability and legal
>concerns surveillance systems face, and provides a solution that is
>minimally invasive to privacy. Our reasonable cause model is a
>computational solution patterned after the Probable Cause Predicate (which
>is used for 4th amendment protection against illegal search and seizures).
>This is a fascinating and potent solution to many of the liability and
>privacy problems faced by surveillance systems.
>
>
> If time permits, I would also like to report on some additional tools we
>(Tom Mitchell, Manuel Blum, myself, and our students) are developing that
>execute surveillance queries over distributed sources. In these systems
>surveillance questions get answered without the data themselves having to
>be transferred.
>

>The third and final contribution my lab could provide is a report card on
>non-traditional data sources. We took about 10 kinds of non-traditional
>data sources and answered the following kinds of questions on each: is
>there a signal; how early is the signal realized; who is providing the
>signal; and so on. The data sources include: web use data (correlated with
>health conditions); grocery data; video surveillance data (cough
>detection); death data; and so on. I think this report is a very important
>contribution because it makes the case for the use of these very
>non-traditional data sources.
>
>At present, our team will be talking about the data sources actually used
>in ESSENCE's real-time surveillance, but the discussion of these data
>sources utilized millions of concentrated records in multiple states in
>order to get the academic information on the quality of that kind of data,
>even though it may not be currently used in the DC area.
>
>If you could find some time in the schedule for me to address these
>matters, I feel quite strongly all the participants would benefit.
>
>--Latanya
>
>
>
>At 08:26 PM 5/28/2002 -0400, you wrote:
>>Attached are the latest agenda for the PI meeting, marked Rev 2, and the
>>latest format for the PI overview briefing, also marked Rev 2. PI's,
>>please send your charts to Rey and I by Monday AM.
>>
>>Panels. Ted and I are still working on the product specification for each
>>of the Panels, as well as the formats for presenters. Note that Ted wants
>>each team to identify its "champion" for the particular topic and panel
who
>>will express the views of the team in the facilitated discussion that
>>follows the series of 10 minute presentations. Please let me know who
>these
>>folks are by Friday. Note that in addition, a panel on Transition has
been
>>added.
>>
>>Data Source Matrix. You are each working on updating your team's data
>>source matrix, including creating the "data space" or "fishbone" chart by
>>populating the template previously forwarded to you for action. Please
>>submit your updated data source matrix and fishbone chart by COB Friday.

>In
>>addition, please add a field to the matrix "Data contains personal
>>identifiers: Yes or No." <<Agenda for BSS PI Meeting Rev 2.doc>> <<PI
>>Templates Rev 2.ppt>>
>>
>>   Thanks,
>>   David
>>
>>
>>
>

15C

tsenator

From: knumrych
Sent: Tuesday, April 09, 2002 3:33 PM
To: mszymanski(contr-iao)
Cc: bmeyrowitz; rpopp; tarmour(contr-iao); ddyer; pbarger(contr-iao); tsenator
Subject: RE: Round 2 BAA Q&As - draft

Mike,

This looks good.

Kristen

-----Original Message-----
From: mszymanski(contr-iao)
Sent: Tuesday, April 09, 2002 3:19 PM
To: knumrych
Cc: bmeyrowitz; rpopp; tarmour(contr-iao); ddyer; pbarger(contr-iao); tsenator
Subject: RE: Round 2 BAA Q&As - draft

Kristen,

Suggested rewrite for Q34:

34. Is the closing date 22 Apr 02? Can a Federal Lab submit a proposal and get funding to work on it? Do we need to have another partner (a university or private company) to submit a proposal?

Related question: Is it acceptable to DARPA for National Labs to participate on this BAA?

RESPONSE: The April 22, 2002 deadline for submitting proposals is a first round deadline. The BAA remains open for one year after the publishing date (March 21, 2002).

Certain Federal Labs may submit proposals. Proposals which are selected - from any source - will receive funding. For eligible Federal Labs, partnering with a university or private company is not required. From DFARS 235.017-1:

"DoD-sponsored FFRDCs that function primarily as research laboratories (C3I Laboratory operated by the Institute for Defense Analysis, Lincoln Laboratory operated by Massachusetts Institute of Technology, and Software Engineering Institute) may respond to solicitations and announcements for programs which promote research, development, demonstration, or transfer of technology (Section 217, Pub.L.103-337)."

DOE-sponsored FFRDCs may also respond to BAA 02-08. The full list of DOD- and DOE-sponsored FFRDCs is maintained by the National Science Foundation at the following URL: http://www.nsf.gov/sbe/srs/nsf02302/mastlist.htm#Topic1.

Mike

-----Original Message-----
From: knumrych
Sent: Tuesday, April 09, 2002 1:26 PM
To: knumrych; mszymanski(contr-iao); bmeyrowitz; rpopp; tarmour(contr-iao); ddyer; pbarger(contr-iao); tsenator
Subject: RE: Round 2 BAA Q&As - draft

Mike,

I chatted with Legal, and it's true that only the three FFRDCs listed in the FAR and the DOE Labs can compete under

the BAA. Other Government labs would have to be teamed with someone else as the prime or else submit proposals to us outside the BAA.

Kristen

-----Original Message-----
From: knumrych
Sent: Tuesday, April 09, 2002 12:22 PM
To: mszymanski(contr-iao); bmeyrowitz; rpopp; tarmour(contr-iao); ddyer; pbarger(contr-iao); tsenator
Subject: RE: Round 2 BAA Q&As - draft

Mike,

Regarding your response to question 32, does the BAA allow offerors to propose options for additional related efforts that could supplement their basic proposal?

Q34: I'm not certain all federal labs are allowed to respond to a BAA. It used to be that only DOE FFRDCs could respond. Barbara is not here to do to answer this question. I will try to poll the other COs to see if this is still the case or not.

Questions 44-48: I'm not sure these don't reveal the deliberative process of potential offerors to all other offerors. I'd like Barbara's opinion on this subset of responses before the answers are posted to the webpage.

Kristen

-----Original Message-----
From: mszymanski(contr-iao)
Sent: Monday, April 08, 2002 6:06 PM
To: bmeyrowitz; knumrych; rpopp; tarmour(contr-iao); ddyer; pbarger(contr-iao); tsenator
Subject: Round 2 BAA Q&As - draft
Importance: High

All,

Draft of Round 2 Q&As attached for review and comment. I'd like to get these up loaded to www.darpa.mil/iao by mid day. Quick turnaround appreciated.

<< File: Q&A 4-8-02-draft.doc >>
Many thanks,
Mike Szymanski
IAO SETA (Syntek Technologies)
703-248-1530 (office)
703-981-0880 (mobile)
703-807-0324 (fax)

tsenator

From: tsenator
Sent: Tuesday, April 16, 2002 9:58 PM
To: mszymanski(contr-iao)
Cc: jpoindexter
Subject: RE: Week # 3 BAA Q&As (draft)

better. we may want to also add a statement to the effect that DARPA will not release personal data to the public nor use it for any purpose other than verifying security clearances. given what our BAA is about, we i think we ought to be as emphatic as possible regarding how we will protect and use such information.

-----Original Message-----
From: mszymanski(contr-iao)
Sent: Tuesday, April 16, 2002 4:25 PM
To: tsenator
Subject: RE: Week # 3 BAA Q&As (draft)

Ted,

Thanks. I agree. How's this rewrite?

76. Can the Government prevent personal data (e.g., SSNs) from being released through FOIA, or if awarded might the proposal become public? Alternatively, might we be able to provide this information in the letter of transmittal?

RESPONSE: Personal data is exempt from FOIA. Trade secrets as well as commercial and financial information contained in a proposal submitted to DARPA are also exempt from FOIA. More information on FOIA exemptions is available at http://www.usdoj.gov/oip/foi-act.htm).

-----Original Message-----
From: tsenator
Sent: Monday, April 15, 2002 11:03 PM
To: mszymanski(contr-iao); rpopp; knumrych; bmeyrowitz; pbarger(contr-iao); ddyer; tarmour(contr-iao)
Subject: RE: Week # 3 BAA Q&As (draft)

I'd give a more direct answer to question 76 regarding release of personal information (i.e., SSN's) in response to FOIA requests. The draft provides the legal basis for rejecting such a request; I'd precede it with a blanket statement that under no circumstances would DARPA release personal information or proprietary information in response to a FOIA request. As written, it sounds like you need a lawyer to figure out what we'd do.

-----Original Message-----
From: mszymanski(contr-iao)
Sent: Monday, April 15, 2002 7:38 PM
To: rpopp; knumrych; bmeyrowitz; pbarger(contr-iao); tsenator; ddyer; tarmour(contr-iao)
Subject: Week # 3 BAA Q&As (draft)

All,

Draft Q&As for Week 3 posted in the files section of the TIA BAA space (also attached for easy reference). Please give a good look - there are a few touchy questions whose answers may need massaging.

<< File: Q&A 4-15-02-draft.doc >>
Thanks,
Mike Szymanski
IAO SETA (Syntek Technologies)
703-248-1530 (office)
703-981-0880 (mobile)
703-807-0324 (fax)

cc: via Groove

tsenator

From: Steven J. McKay [steven.mckay@veridian.com]
Sent: Tuesday, March 19, 2002 6:53 PM
To: Stephanie Strassel
Cc: William J. Koenig; Bob Schrag; Joshua Powers; Mike Pool; jkdavis@afterlife.ncsc.mil; clwayne@afterlife.ncsc.mil; mark.przybocki@nist.gov; tsenator; ccieri@ldc.upenn.edu; myl@ldc.upenn.edu; amitche0; Mark Liberman; lferro@mitre.org; day@linus.mitre.org; Alexis Mitchell; Elizabeth D. Liddy; Chinatsu Aone; Ralph Weischedel
Subject: Re: ACE/EELD data concerns

All,

Pending what I hear about the file format required for the Alembic Workbench, I can report the following based on what I did today. The option to sub-sample seemed the most reasonable path - Mark Liberman's substantive points, notwithstanding - and I've now determined that we have the files to support it. Specifically, I've saved off the following files per these sources:

Recommended Sources:
AFP - 81 files, 26,000 words
AP - 48 files, 19,400 words
CNN - 6 files, 2,900 words
NYTimes - 22 files, 22,400 words
Xinhua - 2 files, 500 words

With these files alone, we have 71,200 words. Mind you, these are sloppy word counts - ie, the files have not been pruned to included only substantive content. Per my discussion with Stephanie this morning, I understand that this will be ok. I also have the following additional files:

Non-copyright material (testimony, and USG-source documents):
4 files, 40,740 words

Other Sources:
LATimes - 13 files, 16,180 words
WashPost - 8 files, 14,400 words
WSJ - 1 file, 1000 words

All of the above files have been saved in .txt format. If its necessary to find additional files to augment those from the Recommended sources - we could reach outside the Contract Killing corpus and include files from the Al Qaeda/Terrorist archive that we're building. We have ample material in this document set, but it would probably be best to leave this for the following year's evaluation to stay more in-line with the EELD evaluation plans - just my two cents on this point. If it were useful to include a small portion, however, simply to beef up the word count from the Recommended sources, then this would seem reasonable to me.

These files are now ready to ship - again pending the responce on the file formats.

Going back to an earlier issue, the responces that I got from the EELD/EE contractors indicated to me that we could probably go ahead with re-issuing the document datasets, absent those that I've identified in this process (recall that the EELD/EE contractors had already recieved all of these documents - but it does not seem that much work has been accomplished against them, based on the replies that I recieved). Dissemination of the documents used in the training set (and later, the evaluation set), could then proceed as per LDC's established procedures. One more aspect of this process that might further mitigate any concern about the EE teams having seen these documents, is that the file names have been changed for most of the files proposed for this dataset.

This is our proposal to proceed. Comments are welcome.

Regards,

Steve

At 06:28 PM 03/15/2002 -0500, you wrote:
>Steve,
>
>I'd like to bring a few additional folks into this discussion, so I'll
>take a step back and summarize the issues as I understand them before
>proposing a couple of solutions. Please let me know if I've
>misrepresented any aspect of the EELD data.
>
>The LDC has been asked to complete ACE annotation (both EDT and RDC) on
>45K words of EELD material. The EELD data comes from FBIS and
>Lexis-Nexis; the collection contains both html files and text files.
>The data has not been catalogued in terms of its original document
>sources; it will require a substantial effort to retrieve this
>information for every file in the collection and for some cases may not
>be possible. I believe that a portion of the EELD data has been
>manually annotated (chronology writeups?) but I don't know what this
>annotation entails or the level of effort that was required to produce
>it.
>
>In terms of planning for ACE annotation of the data, 30K words of EELD
>data have been designated as part of the training/development set, and
>annotation of this data must be completed by June 1 for release to the
>research sites. We estimate that it will require approximately 100
>hours of annotation effort (including second passing and 5% dual
>annotation) for each of the two tasks, so 200 hours total. The two
>tasks cannot be done in parallel; EDT tagging must be completed before
>RDC tagging can begin. We also need to factor in the additional
>turnaround time required to do TIMEX tagging, which will be completed by
>Beth Sundheim's group (contingent upon their funding situation). TIMEX
>tagging makes the RDC annotation task easier, and although RDC
>annotation could be completed without the TIMEX tags, the tagging would
>then be slower than the estimates I've given here.
>
>At the PI meeting in late February, we established April 1 as a start
>date for annotation. LDC will devote a minimum of 25 hours/week to
>annotation of the EELD data (plus additional effort to RDC tag the ACE
>corpus data) in order to to have the data available for release by June
>1. As we established at the PI meeting, in order to stay on track, the
>annotation tool and guidelines revisions must be completed no later than
>the end of March, and we must also have ACE-formatted data in hand by
>then.
>
>The LDC raised two main concerns with respect to our annotation of the
>EELD
>data: data distribution and Intellectual Property Rights.
>
>As a general rule the LDC does not undertake annotation work for data
>which cannot be distributed to the larger research community. We
>believe that it is everyone's best interest to distribute data as widely
>as possible to enable research beyond the scope of the current
>evaluation. We must acquire IPR for all data that we distribute.
>
>Some of the EELD data comes from Lexis-Nexis. Lexis is a data reseller,
>and copyright information provided on their website indicates that all
>IPR arrangements must go directly to the source provider and that exact
>citations of the original source are required when making reference to
>the data. Our legal counsel has advised us that approaching Lexis for
>blanket rights to distribute up to 45K words of data without knowing its
>original source would be impossible; at minimum Lexis would need to know
>the original data source in order to determine the fees and royalties

>involved. As for FBIS, they are essentially a data redistributor; their
>website states that recipients of FBIS products must contact the
>original source or the copyright holder directly to obtain dissemination
>permission.
>
>Steve brought up the issue of Fair Use and whether the annotation and
>distribution of the EELD data could be covered by this. Because the
>data will not be limited to use by the LDC alone, we cannot assume
>UPenn's interpretation of Fair Use here. The LDC is not permitted to
>interpret Fair Use for third parties outside of UPenn, and distribution
>of the data by the LDC cannot claim Fair Use. All of these facts point
>to the need to acquire IPR for this data.
>
>I have two suggested courses of action. The first is to identify those
>sources within the EELD data set for which the LDC already holds IPR or
>could easily negotiate them. Our IPR team has compiled a list of
>English language sources that LDC has agreements with and are also used
>by Lexis Nexis. I'll split off a "recommended list" based on
>organizations that we have the most current, non-restricting agreements
>with:
>*Recommended - Agence France Press, Associated Press, CNN, NBC, NY
>Times,
>Xinhua
>*Others - ABC, Wall Street Journal, IRNA, NPR, LA Times, WA Post
>
>It should be possible to subsample the EELD collection for these
>particular sources and limit ACE annotation to those documents.
>
>If this proves to be impossible, we might consider instead working with
>data that has not been EELD-annotated, but which might contain the kind
>of topical focus that is necessary for EELD research. This might be
>achieved by looking to topically annotated material from the Topic
>Detection and Tracking corpora. We have annotated English broadcast and
>newswire material against 220 topics, some of which might be of interest
>for EELD. I can provide a list of topics if that would be useful. This
>is data for which the LDC already holds IPR; in fact, the original ACE
>data is drawn from the TDT corpora.
>
>In addition to these fundamental concerns, there are a few smaller data
>issues to consider. We need to ensure that the EELD data is in a format
>that will work with the Alembic Workbench annotation tool and that file
>naming and encoding is consistent before annotation can begin. It won't
>be possible for the LDC to take responsiblity for working out these data
>formatting issues given the number of other tasks we must complete
>before the April 1 start date.
>
>Please let me know if you have any questions or need any additional
>information. In order to justify moving ahead with annotation and the
>costs that will incur, we need to know that these data issues have been
>worked out or that there is a clear plan in place for working them out
>prior to April 1.
>
>Stephanie
>
>
>Steve McKay wrote:
>>
>> Stephanie,
>>
>> Two issues - first, does an agreement with Lexis-Nexis or FBIS cover any
>> source recieved from them? And do you have such an agreement with either
>> of these distributers? Next, have your folks determined that the Fair Use
>> exclusions do not cover the kind of research and development work that you
>> do? We've not gotten a legal opinion on this, but it does seem to fit our
>> effort.
>>
>> As for cataloguing the original document sources, no we haven't done that.

>> Its a simple matter for the .txt files that I've renamed, but the bulk of
>> the files are not named in this convention and it would be quite a chore to
>> gather such statistics, short of using an extraction program.
>>
>> As for the issue of broad dissemination to the level that you've described
>> - we'll certainly raise the issue with Ted, but I wonder if the copyright
>> issue won't be the more problematic one.
>>
>> Regards,
>>
>> Steve
>>
>> At 05:42 PM 03/13/2002 -0500, you wrote:
>> >Steve,
>> >
>> >Sorry for the delay in responding -- I don't usually work on IPR so I
>> >wanted to take some time to consult with our IPR team to make sure I
>> >understood all of the issues before getting back to you.
>> >
>> >I know that we have current or recent IPR relationships with some of the
>> >original data providers that showed up in the sample files you sent me
>> >-- for instance, IRNA and AFP were among the source data providers for
>> >the text files you sent, and we've had long-standing relationships with
>> >those sources.
>> >
>> >As for LDC membership, our members are commercial, government and
>> >educational institutions who are engaged in linguistic research,
>> >education and technology development. In order to access the data we
>> >distribute, members must sign an agreement stating that the data will be
>> >used only for these purposes. We do not place any restrictions on who
>> >can become an LDC member.
>> >
>> >We're eager to work out the issues of distribution and IPR as soon as
>> >possible since we're under a tight deadline. One thing that would be
>> >very helpful is a list of all of the source providers that are part of
>> >the archives you're using, and the amount of data (words or files) for
>> >each source. Is this information easily retrivable from the file names,
>> >or perhaps you've already extracted this information. In terms of IPR,
>> >it would probably be easiest for us to start with the sources like IRNA
>> >and AFP for which we have an existing relationship. In terms of data
>> >distribution, where should we go from here?
>> >
>> >Thanks for your help,
>> >Stephanie
>> >
>> >Steve McKay wrote:
>> >>
>> >> Stephanie,
>> >>
>> >> In case you missed these questions, I thought i'd follow up on two issues
>> >> highlighted from your earlier email.
>> >>
>> >> >Also, we're wondering about the source of the data and its intellectual
>> >> >property rights status. In the past, an assumed goal of the ACE program
>> >> >has been to make the data available to researchers outside of the local
>> >> >community once the evaluation has been completed. For the current ACE
>> >> >data, the LDC has negotiated IPR agreements with the data providers so
>> >> >that later distribution will be no problem. I'm not at all familiar
>> >> >with the nature of the EELD data, but if any of it is broadcast news or
>> >> >newswire/newspaper, it would seem sensible to select sources for which
>> >> >LDC already holds or has previously held an IPR agreement with the data
>> >> >provider, to facilliate later redistribution of resources.
>> >>
>> >> What is the status of FBIS and Lexis-Nexis articles in this regard?
>> >>
>> >> >We're also concerned about the EELD source material because as part of
>> >> >the University of Pennsylvania, and as part of the LDC's mission, we are

>> >> >limited when it comes to working with data that cannot be distributed
>> >> >more widely (i.e., to LDC's members), and are prohibited from working
>> >> >with classified material. As I said, I'm not sure what kind of data is
>> >> >contained in the EELD materials, but we should keep these points in mind
>> >> >when selecting material to be included in the ACE evaluation.
>> >>
>> >> There are no classified documents, but we have been viewing the EELD
>> >> document collections as somewhat sensitive, and limiting their
>distribution
>> >> per guidelines worked out with Ted Senator. I'd say that distribution to
>> >> ACE participants is clearly within those guidelines, however, I"m not
>clear
>> >> what our guidance would be on distribution to the wider LDC membership.
>> >> This is potentially an issue that we'd want to discuss with Ted.
>> >>
>> >> In this regard, who are the LDC members that the material might be
>> >> distributed to?
>> >>
>> >> Regards,
>> >>
>> >> Steve
>
>--
>Stephanie Strassel
>Annotation Group Manager * Phone: 215.898.9681
>Linguistic Data Consortium * Fax: 215.573.2175
>3615 Market Street * strassel@ldc.upenn.edu
>Philadelphia, PA 19104-2608 * http://www.ldc.upenn.edu

tsenator

From: Mark Liberman [myl@unagi.cis.upenn.edu]
Sent: Friday, March 15, 2002 9:11 PM
To: Stephanie Strassel
Cc: Steve McKay; William J. Koenig; Bob Schrag; Joshua Powers; Mike Pool; jkdavis@afterlife.ncsc.mil; clwayne@afterlife.ncsc.mil; mark.przybocki@nist.gov; tsenator; ccieri@ldc.upenn.edu; amitche0
Subject: Re: ACE/EELD data concerns

>The LDC has been asked to complete ACE annotation (both EDT and RDC) on
>45K words of EELD material. The EELD data comes from FBIS and
>Lexis-Nexis; the collection contains both html files and text files.
>The data has not been catalogued in terms of its original document
>sources; it will require a substantial effort to retrieve this
>information for every file in the collection and for some cases may not
>be possible. I believe that a portion of the EELD data has been
>manually annotated (chronology writeups?) but I don't know what this
>annotation entails or the level of effort that was required to produce
>it.

We went through exactly this situation before in the case of the TIPSTER "extraction" track.

For that track, a body of data (I think on mergers and acquisitions) had been compiled from DIALOG (a service similar to NEXIS), with the number of original copyright holders being several hundred. The person managing that aspect of the program had a subscription to DIALOG, and concluded that in order to distribute copies of the collection to N contractors, he simply needed to execute N identical searches, thus paying DIALOG N times the single-search fee. In fact he was mistaken: DIALOG had no legal ability to license him to re-distribute the material, no matter how many fees he paid.

According to what I was told, one of the publishers found out about this, somehow, and threatened to sue. This was felt to be a Bad Thing, since it was a foreign publisher and the U.S. was then strongly urging foreign governments to take a stronger line against IPR piracy. So we were asked to step in and clear up the IPR issues for this distribution. This turned out to be a very time-consuming and difficult process; we were able to mollify the particular complaining publishers, and arrange a formal contract with them; and the folks at DIALOG were extremely helpful; but the task was frankly almost impossible, and took many person-months to resolve even partially.

Note that the problems involved had nothing to do with distribution to the public at large, but just with distribution within the TIPSTER program proper. As in that case, re-distribution of NEXIS output even to ACE contractors is surely not licensed. Whether it could be construed as "fair use" is not clear, but let's just say that no publisher (including a not-for-profit publisher like a university press) would assume so for this volume of material.

In addition, it strikes me as short-sighted to put so much effort and money (not mainly the annotation effort, but rather the research effort to establish baseline analytic results) into a database that can't be used in the future. The language research community has a lot of evidence by now that foresight in such matters pays large dividends.

If it's decided to go ahead with this database, we'll certainly do our best to cope -- even if this involves IPR negotiations with hundreds of publishers -- but my frank advice would be to revise the plan

before it's too late, and to build IPR arrangements into the process from the start, by chosing a corpus for which IPR licensing is plausible task. This would entail keeping track of the sources, balancing the costs and benefits of multiplying the number of sources, and at least opening a discussion with each source by sending them a letter and giving them a couple of weeks to object, before incorporating their material into a corpus that will get any serious amount of annotation and analysis.

Regards,

Mark Liberman

tsenator

From: Amarnath Mukherjee [am@knoltex.com]
Sent: Friday, November 30, 2001 2:13 PM
To: tsenator
Subject: SBIR Topic Question

Ted,

Sorry I missed your call. In the event we are not able to synchronize today, this being the last legal day to talk for this SBIR, thought I'll send you my questions over email.

Essentially, what we are interested in is to understand are your relative priorities vis-a-vis:

- inconsistency checking, and
- input processing

For instance, are you primarily interested in the Inconsistency-Check Rules Engine
and Search Engine? Or do you need the capability to pre-process noisy unstructured data
as well?

Also for financial transactions, and other structured data, do you expect us to build adapters from scratch, or should we assume that the new software should
talk to existing adapters?

Your topic is very interesting.

Regards,
Amarnath
--
(408) 615-1611
am@knoltex.com

tsenator

From: Siegrist, David [Siegrist@PotomacInstitute.org]
Sent: Wednesday, May 29, 2002 1:31 PM
To: tsenator
Subject: FW: Latanya request for more time, fyi

-----Original Message-----
From: Latanya Sweeney [mailto:latanya@cs.cmu.edu]
Sent: Wednesday, May 29, 2002 1:28 PM
To: Siegrist, David
Subject: RE: Latest Agenda; PI Summary Format; Panel info; & Matrix

Hi,

Thanks for the response. If possible, 15 minutes on the public health law survey, 20 minutes on the reasonable cause model and 10-20 minutes on the scoring of data sources. Let me know what's possible. Thanks.

--Latanya

At 09:35 AM 5/29/2002 -0400, Siegrist, David wrote:
>Hi, Latonya

> On your major points, I think your research sounds fascinating. How
>many minutes for each one do you think you need?
>
> Cordially,
> David
>
>-----Original Message-----
>From: Latanya Sweeney [mailto:latanya@cs.cmu.edu]
>Sent: Wednesday, May 29, 2002 9:26 AM
>To: Siegrist, David
>Cc: 'Ted Senator (E-mail)'
>Subject: Re: Latest Agenda; PI Summary Format; Panel info; & Matrix
>
>
>
>Hi David,
>
>I think there are three major contributions my lab's efforts could make to
>the PI meeting but in order to accomplish such would require some
>modification to the agenda. I am humbly requesting you see if you can make
>some time for these matters. They are described below.
>
>The first major contribution concerns privacy. Under this DARPA effort, we
>had a group of 40 law students and 3 lawyers who surveyed public health
>laws and regulations in each of the 50 states and DC and answered about 60
>questions for each state. The questions centered on the collection and use
>of non-traditional data sources for bioterrorism surveillance; each kind of
>non-traditional data source was examined in each state. I would like to
>take a few minutes to report on these findings. Highlights include the
>variability found and some particularities with specific kinds of data
>sources and the nature of liability concerns.
>
>The second major contribution also concerns privacy, in particular, the
>development and use of computational tools that ensure privacy. I would

>like to introduce our reasonable cause model, which provides sufficiently
>anonymous data during normal surveillance. The data from each data holder
>is rendered sufficiently anonymous and then forwarded to the surveillance
>system. We can show that detection algorithms work well with such data.
>When the surveillance system detects suspicious activity, the anonymity of
>the data is lowered. If an explicit outbreak is detected, explicitly
>identified data is provided. Under our reasonable cause model, the level
>of anonymity of the data matches the surveillance system's need for the
>data. We can show that this system removes many of the liability and legal
>concerns surveillance systems face, and provides a solution that is
>minimally invasive to privacy. Our reasonable cause model is a
>computational solution patterned after the Probable Cause Predicate (which
>is used for 4th amendment protection against illegal search and seizures).
>This is a fascinating and potent solution to many of the liability and
>privacy problems faced by surveillance systems.
>
>
> If time permits, I would also like to report on some additional tools we
>(Tom Mitchell, Manuel Blum, myself, and our students) are developing that
>execute surveillance queries over distributed sources. In these systems
>surveillance questions get answered without the data themselves having to
>be transferred.
>

>The third and final contribution my lab could provide is a report card on
>non-traditional data sources. We took about 10 kinds of non-traditional
>data sources and answered the following kinds of questions on each: is
>there a signal; how early is the signal realized; who is providing the
>signal; and so on. The data sources include: web use data (correlated with
>health conditions); grocery data; video surveillance data (cough
>detection); death data; and so on. I think this report is a very important
>contribution because it makes the case for the use of these very
>non-traditional data sources.
>
>At present, our team will be talking about the data sources actually used
>in ESSENCE's real-time surveillance, but the discussion of these data
>sources utilized millions of concentrated records in multiple states in
>order to get the academic information on the quality of that kind of data,
>even though it may not be currently used in the DC area.
>
>If you could find some time in the schedule for me to address these
>matters, I feel quite strongly all the participants would benefit.
>
>--Latanya
>
>
>
>At 08:26 PM 5/28/2002 -0400, you wrote:
>>Attached are the latest agenda for the PI meeting, marked Rev 2, and the
>>latest format for the PI overview briefing, also marked Rev 2. PI's,
>>please send your charts to Rey and I by Monday AM.
>>
>>Panels. Ted and I are still working on the product specification for each
>>of the Panels, as well as the formats for presenters. Note that Ted wants
>>each team to identify its "champion" for the particular topic and panel
who
>>will express the views of the team in the facilitated discussion that
>>follows the series of 10 minute presentations. Please let me know who
>these
>>folks are by Friday. Note that in addition, a panel on Transition has
been
>>added.
>>
>>Data Source Matrix. You are each working on updating your team's data
>>source matrix, including creating the "data space" or "fishbone" chart by
>>populating the template previously forwarded to you for action. Please
>>submit your updated data source matrix and fishbone chart by COB Friday.

>In
>>addition, please add a field to the matrix "Data contains personal
>>identifiers: Yes or No." <<Agenda for BSS PI Meeting Rev 2.doc>> <<PI
>>Templates Rev 2.ppt>>
>>
>>   Thanks,
>>   David
>>
>>
>>
>

TITLE PAGE

Briefing Materials concerning TIA BIOALERT 1A

FOLDER 1.a.9. -1.a.17. TOC 4

KICKOFF BRIEFS

- 1.a.9 Epi-SPIRE: Bio-surveillance Prototype w/Model based Multi-Modal Mining-Environment Info
  - IBM T.J.Watson Research Center
  - Johns Hopkins U
- 1.a.10 Modeling
  - IBM T.J.Watson Research Center
  - Johns Hopkins U
- 1.a.11 BioSurveillance Validation
  - Potomac Institute for Policy Studies
- 1.a.12 Bio-Surveillance System – Advanced Medical Readiness (Kickoff Meeting 11-12 Oct 01)
  - Veridian-Stanford
  - University of New Mexico

- 1.a.12.a Simulation and Disease Models
  - Veridian
  - NATO (NBC Planning Guide)
- 1.a.13 Scalable Biosurveillance Systems: Focus Areas for Technical Contributions
  - Pitt-CMU
    - PITT-CMU Center for BioMed Info
  - BioMed Security Institute (MPC Corp.)
- 1.a.14 Modeling & Evaluating Response
  - Pitt-CMU
    - PITT-CMU Center for BioMed Info
  - BioMed Security Institute (MPC Corp.)
- 1.a.15 Bio-Surveillance System Data/Info Sharing

(b)(6)

- 1.a.17 Summary
  - DARPA

Epi-SPIRE: A Bio-surveillance Prototype using Model-based Multi-Modal Mining of Environmentally Related Information

Bio-surveillance Kickoff Meeting
October 11 – 12, 2001

PI: Dr. Chung-Sheng Li

IBM T. J. Watson Research Center
Johns Hopkins University

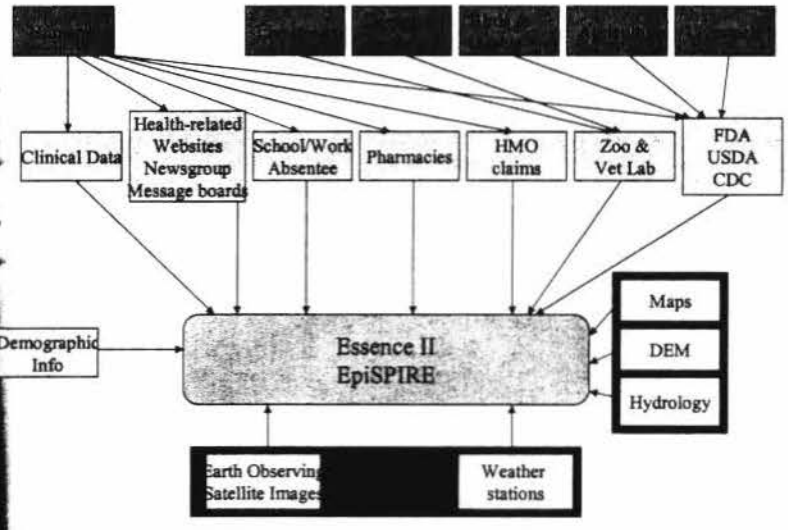
Team

- IBM T. J. Watson Research Center
  - Sankar Basu
  - Murray Campbell
  - Yuan-Chi Chang
  - Matthew Hill
  - Vijay Iyengar
  - Chung-Sheng Li (PI)
  - Milind Naphade
  - John R. Smith
- Johns Hopkins University
  - Prof. Greg Glass (co-I)
  - Rebecca Freeman

EpiSpire Highlights

- **Objectives**:
  - Provide real-time access of non-conventional & heterogeneous data sources
  - Extract models from these data sources for
    - Expected naturally occurring disease patterns
    - Suitability of conditions for various attack scenarios
    - Detecting anomalies in predicted vs. observed data
  - Refine models as more data become available
- **Project History**
  - 1994 NASA Digital Library Technology (DLT) Program "*Retrieval of Images by Means of Content*"
  - 1998 NASA Earth Science Information Partner (ESIP II) Program "*Progressive Mining of Satellite Images for Environmental and Public Health Applications*" (joint effort with JHU and Raytheon)

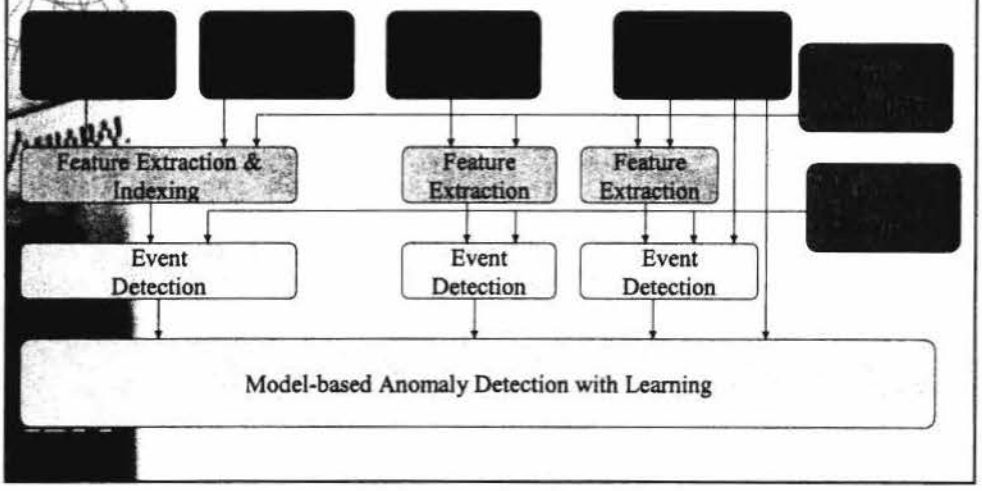
EpiSPIRE for Bio-surveillance in Essence II

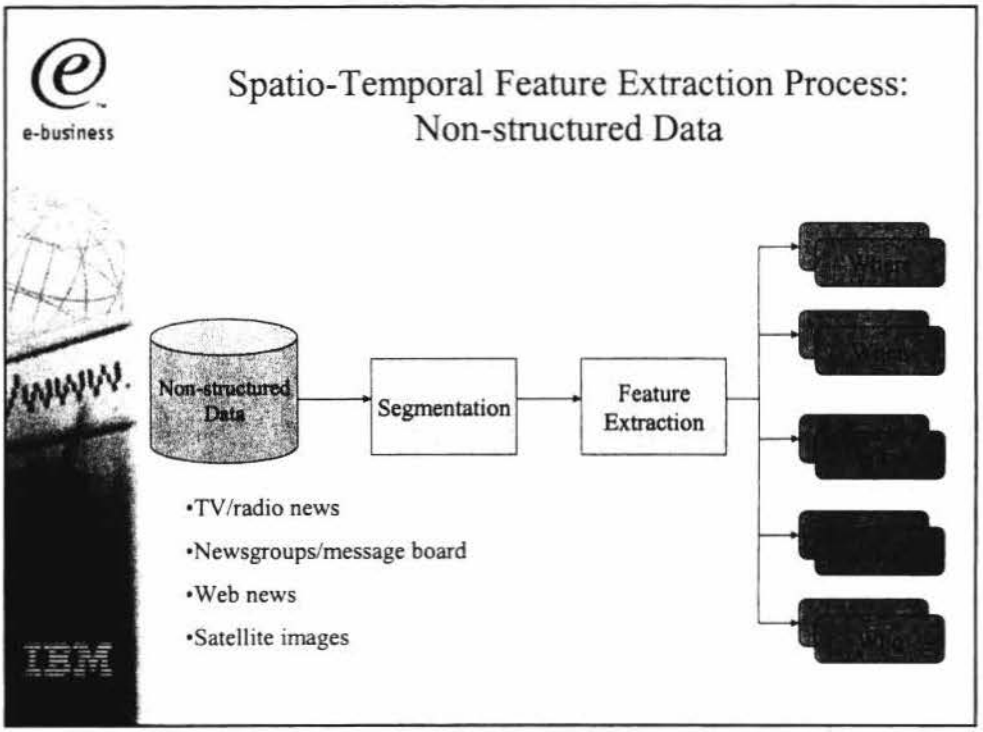


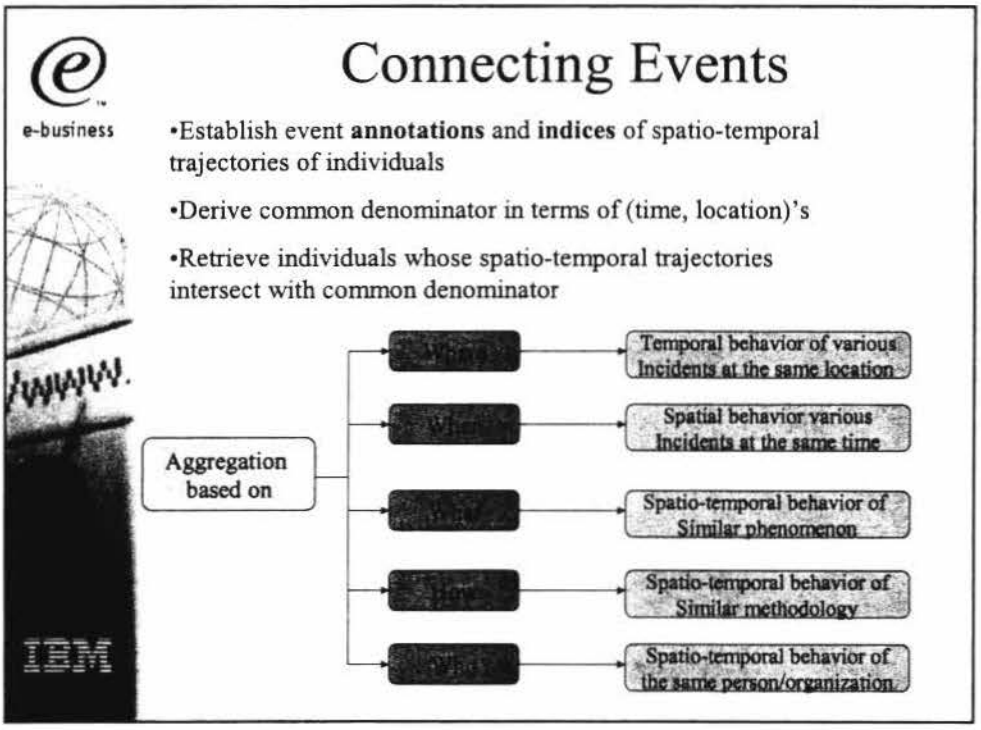
Focus of EpiSPIRE

- Anomaly detection through knowledge models of possible ongoing and future outbreaks through
  - Conventional data: active learning to leverage minimal ground truth
  - Non-conventional data: to establish the baseline (prediction for some vector-born diseases)
- Develop data warehouse containing event annotations & indices of the socio-economical, public health, and spatio-temporal behaviors of individuals, groups, and geographic regions
  - Determine common denominator from possible ongoing outbreak
  - Identify additional individuals or geographic regions where treatments & cures shall be applied from the data warehouse
- Validation
  - Data generation
    - Simulation/emulation to generate synthesized data
    - Perturbation of acquired data
  - Validation through simulation/emulation: both standalone and as part of Essence II

EpiSPIRE Framework



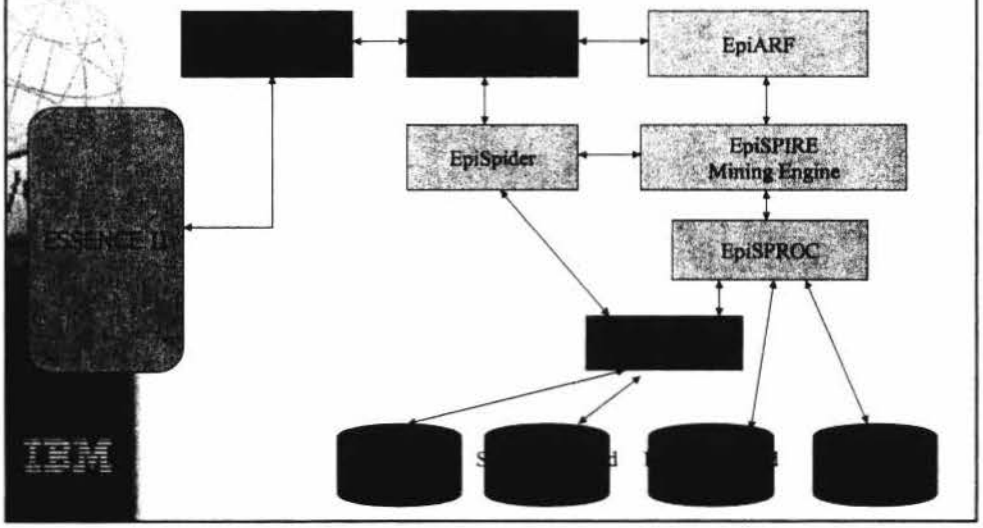

e-business
Spatio-Temporal Feature Extraction Process:
Non-structured Data
Non-structured Data
Segmentation
Feature Extraction
•TV/radio news
•Newsgroups/message board
•Web news
•Satellite images
IBM


e-business
Connecting Events
•Establish event **annotations** and **indices** of spatio-temporal trajectories of individuals
•Derive common denominator in terms of (time, location)'s
•Retrieve individuals whose spatio-temporal trajectories intersect with common denominator
Aggregation based on
Temporal behavior of various Incidents at the same location
Spatial behavior various Incidents at the same time
Spatio-temporal behavior of Similar phenomenon
Spatio-temporal behavior of Similar methodology
Spatio-temporal behavior of the same person/organization
IBM

e-business

Internal Architecture of Epi-SPIRE

Operational Aspects:



e-business

Plan & Milestones

- **FY01(Sept., 2001)**
  - Feature definition language/Event description language
  - Identify scenario(s) and data sources
  - Develop perturbation approach representing attack scenarios for data generation
- **FY02(Sept., 2002)**
  - MALM
  - Feature extraction modules
  - Event detection modules
  - Data generation, stretch goal - a simulator (BioCity – similar to SimCity) for synthesizing data
  - Develop and benchmark models for multi-channel time series data with perturbations
- **FY03(Sept., 2003)**
  - Epi Active Learning
  - Multi-modal EpiIMSE
  - Develop and benchmark models for multi-channel time series data with perturbations and non-structured data sources
- **FY04(Sept., 2004)**
  - Multi-modal EpiIMSE + federated search
  - System validation and benchmarking with both structured and non-structured data

IBM

e-business

Weather Data Feature Extractor and Event Detector/Search

October 11 – 12, 2001

Yuan-Chi Chang

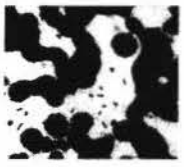
IBM

e-business

Weather and Bioagent Release

Right weather conditions create suitable release environment

- light, moderate wind (5 to 25 mph)
- evening or night
- temperature inversion
- temperature
- relative humidity
- precipitation



IBM

e-business

Near-real-time Weather Data

Continuous monitoring of environment variables

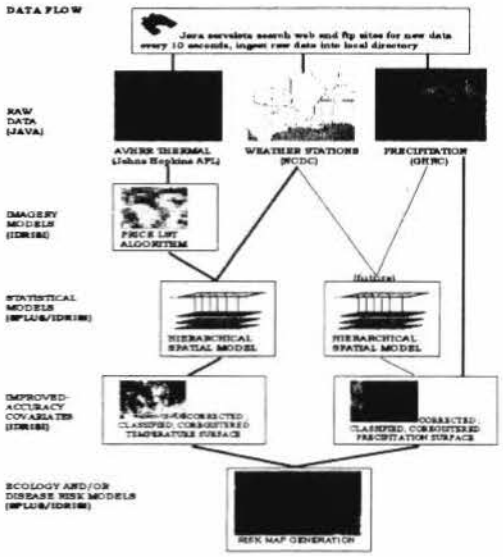
Air temperature
Relative humidity
Precipitation
Wind speed
Wind direction

Derived from station readings and satellite images for better resolution

Source: NOAA National Climate Data Center

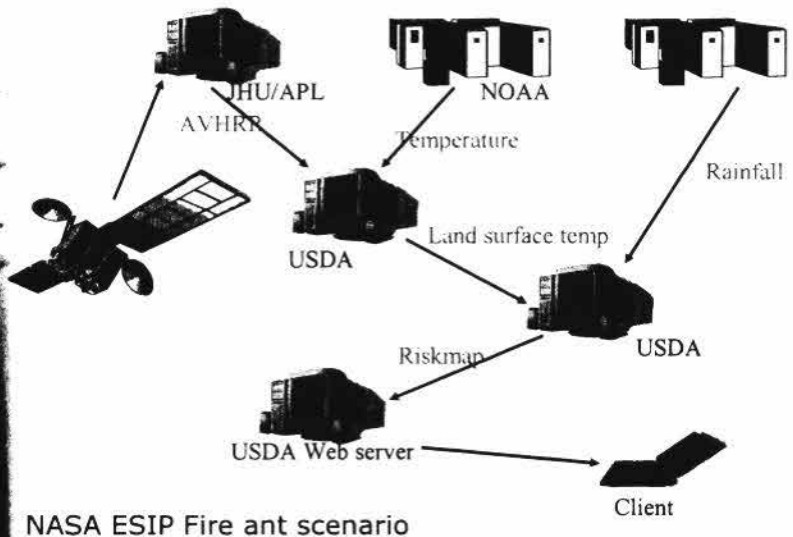
e-business

Monitor Weather - Data Fusion



NASA ESIP
Fire ant
scenario

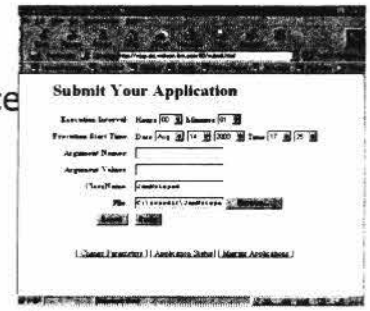
Monitor Weather - Data Flow



NASA ESIP Fire ant scenario

Feature Extractor Framework

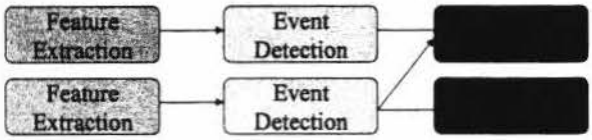
- Web data acquisition
  - Push/Pull
  - Web services (SOAP)
- Legacy software interface
  - IDRISI/Splus/Matlab
- Workflow driver
  - File system
  - Messaging
- Reliability
  - Service logging and monitoring
  - Email notification
  - Service roaming



e-business

Event Detector

- Event description language
  - input 1 (t0) > value 1 and input 2 (t0+t1) < value 2
  - slow rising followed by sharp falling (characterized by slope)
  - multi-input spatial-temporal series
  - parameterized/fuzzy description
- Event detector/trigger
  - configurable event description
  - multi-input, multi-detector hierarchy
  - asynchronous event triggering/propagation



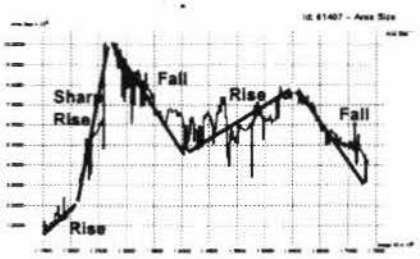
IBM

e-business

EpiMALM: Multiple Abstraction Level Mining of Time Series

- Expression of time-series behavior in terms of semantic labels at multiple abstraction levels
  - Content-based search based on time-series relatively unexplored
  - prior work on EKG, stock prices
- Extraction of time-series from imagery
- Extension of time-series language to permit expression of both intra- and inter-series and intra- and inter-object constraints (spatial and temporal)

- Example query:
  - Locate well logs sections which contain the pattern *sharply rising segment follow by flat segment follow by sharply falling segment*



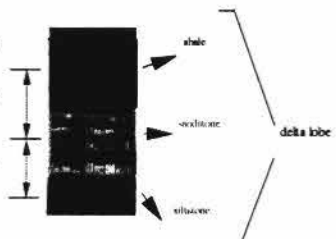
IBM

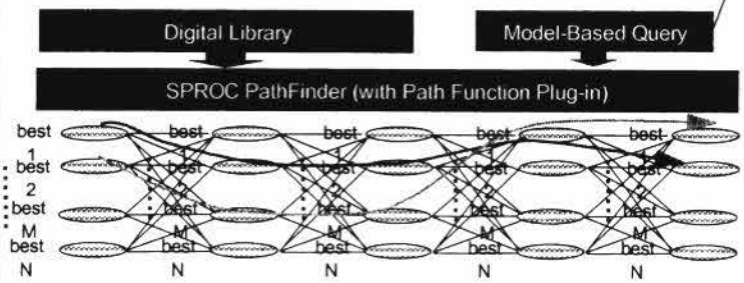
Event Search

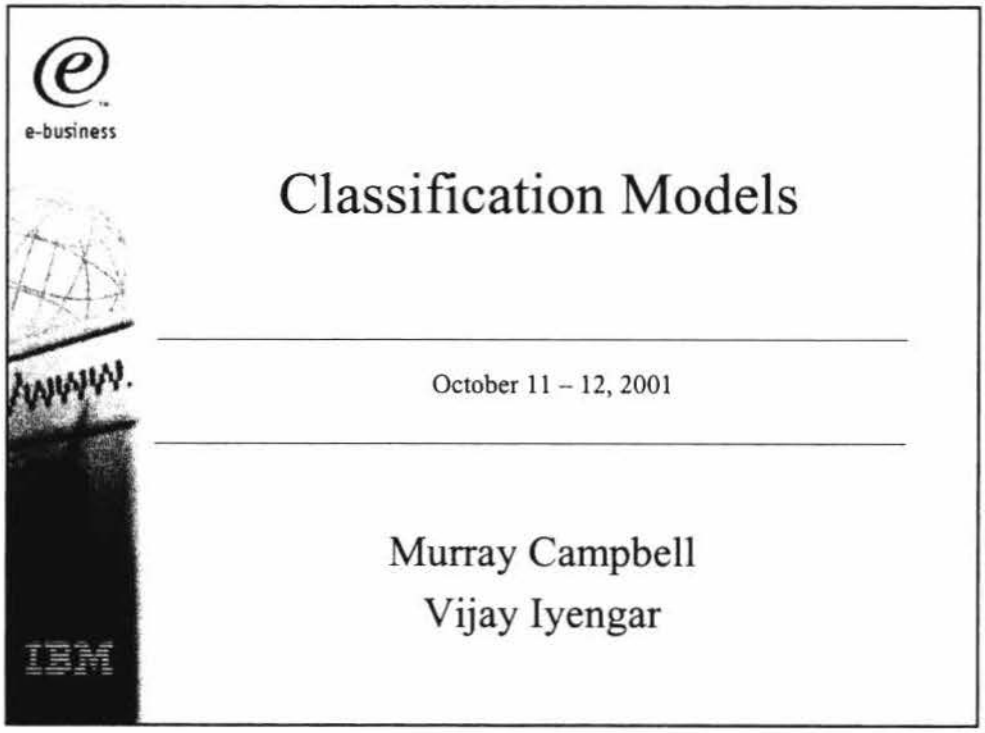
- Event indexing
  - index triggered events by magnitude, duration, spatial location, time stamp, etc.
  - possible approach: R-tree and variants
- Event search
  - multi-input correlated by magnitude, duration, spatial location, time stamp, etc.
  - crisp and fuzzy constraints
  - query description language: event description?
  - Epi-SPROC

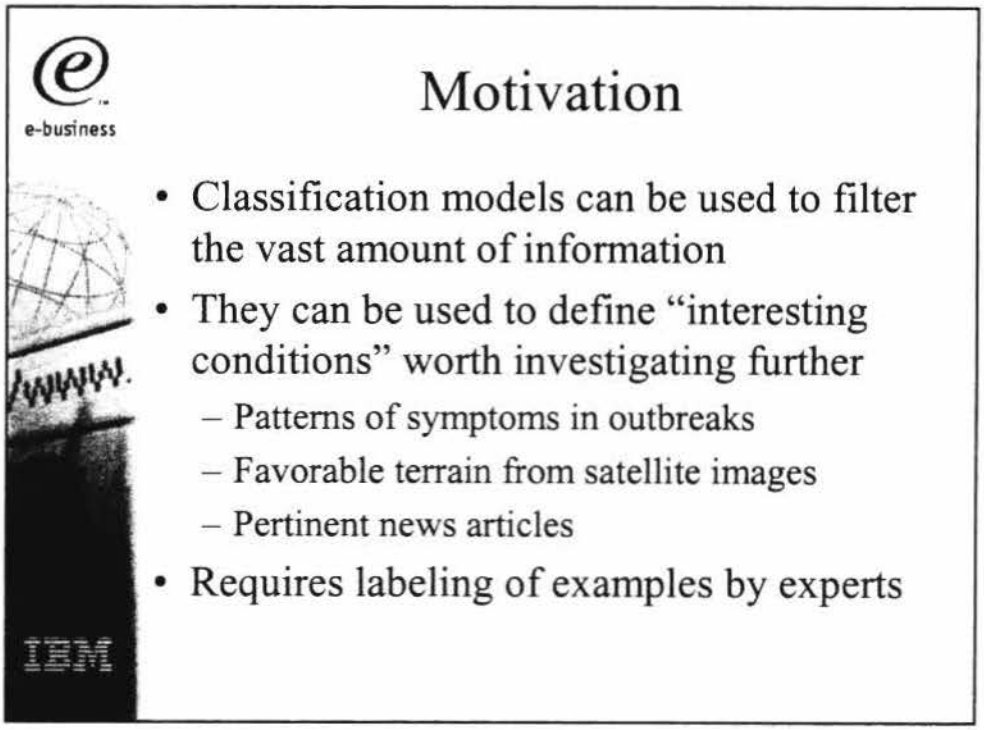
Epi-SPROC: Multi-modal Model-Based Bundle Search on Spatio-temporal Signatures

- Search for multiple items simultaneously
  - items with relationships and constraints
- Scenario - Oil/Gas Exploration
  - search combinations of rock samples






e-business
Classification Models
October 11 – 12, 2001
Murray Campbell
Vijay Iyengar
IBM


e-business
Motivation
• Classification models can be used to filter the vast amount of information
• They can be used to define "interesting conditions" worth investigating further
– Patterns of symptoms in outbreaks
– Favorable terrain from satellite images
– Pertinent news articles
• Requires labeling of examples by experts
IBM

e-business

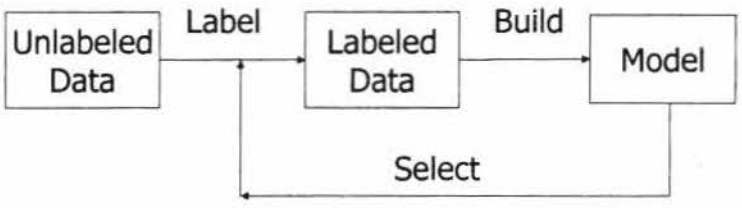
Active Learning

- Choose examples for labeling
  - "Maximize" incremental value
- Can result in significant reduction in the number of examples needed
  - Better than random or stratified sampling

IBM

e-business

Active Learning



IBM

e-business

Our approach

- Use ensembles of decision trees or rules
  - Accurate and robust
  - Handle heterogeneous attributes
- Applicable to various domains
  - Tested using benchmarks representing text, satellite images, user surveys
  - Required labeled data set reduced in size significantly: Factor of 2 to 10.

IBM

e-business

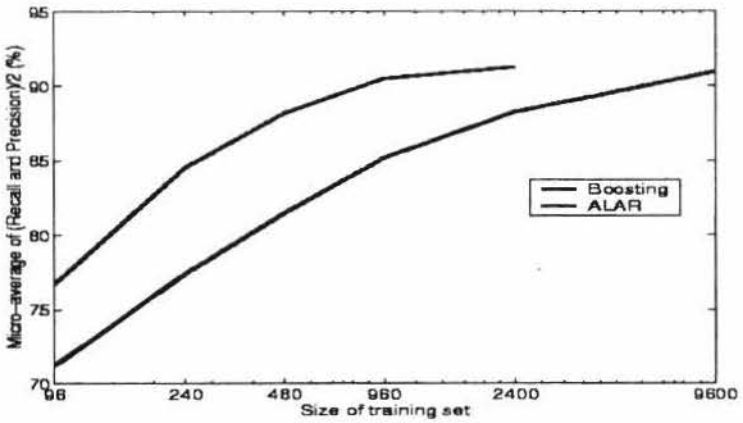
Method

1. Pick initial sample for labeling
   - Random or manual
2. Build ensemble model
3. Apply ensemble model to unlabeled data
4. Select for labeling examples with maximum uncertainty (with respect to majority vote)
5. Expert user labels selected examples
6. Repeat steps 2 to 5
7. Use final ensemble model in applications

IBM

e-business

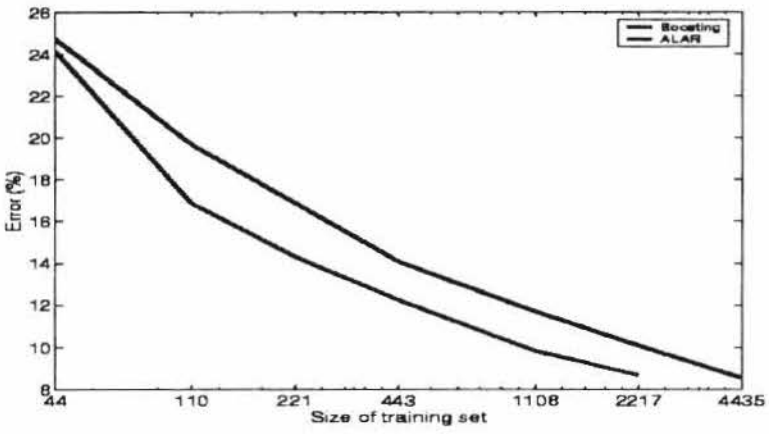
Experiments: Text Domain



IBM

e-business

Experiments: Satellite Image



IBM

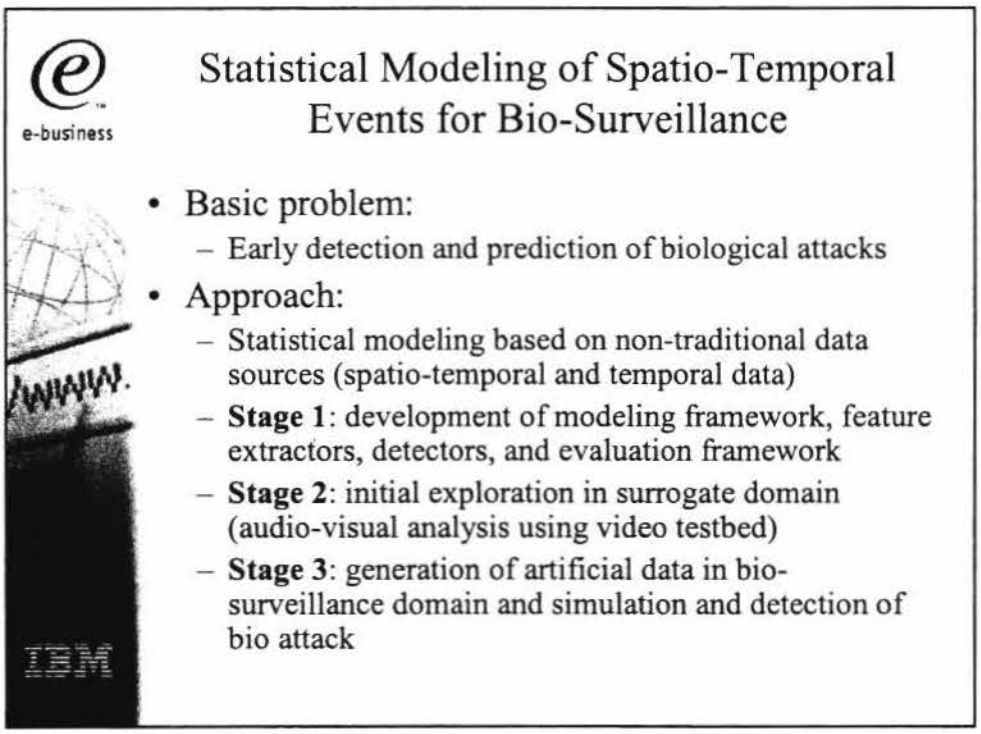
Status

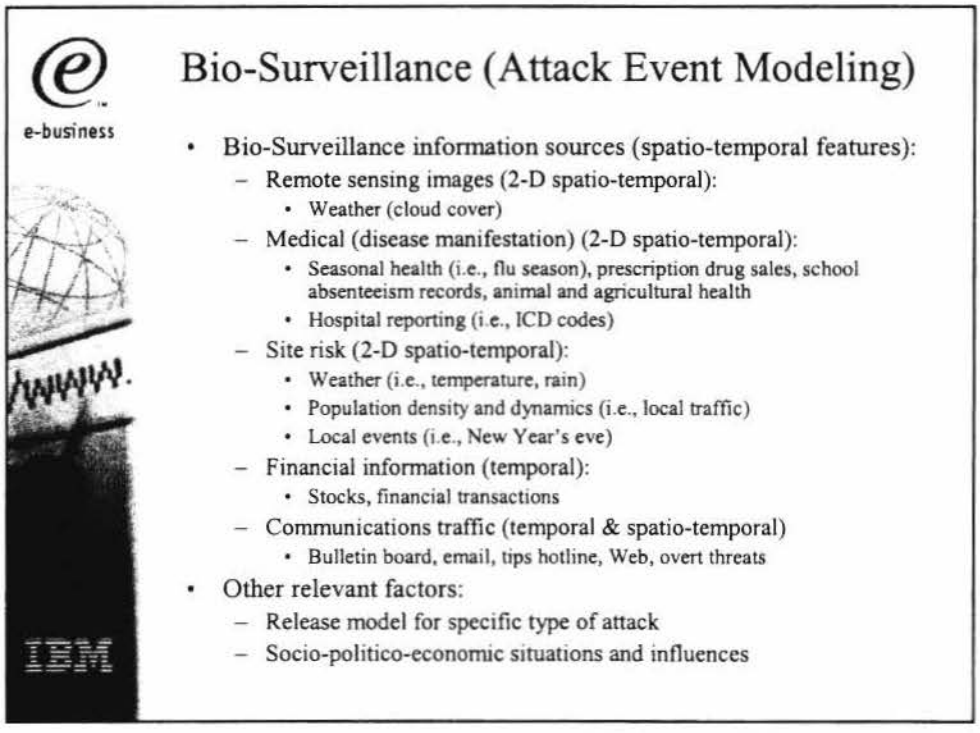
- Prototype implementation in progress
- Will be tested using public benchmarks
  - Accuracy and performance
- Interface to rest of system
  - Database
  - Other

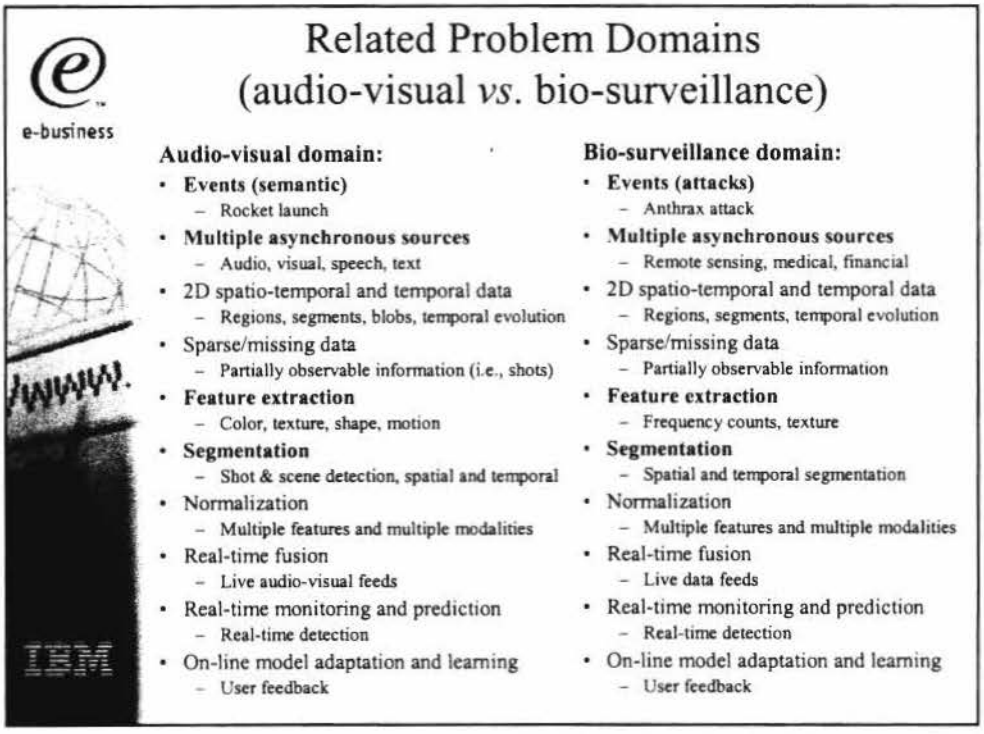
Statistical Modeling of Spatio-Temporal Events for Bio-Surveillance

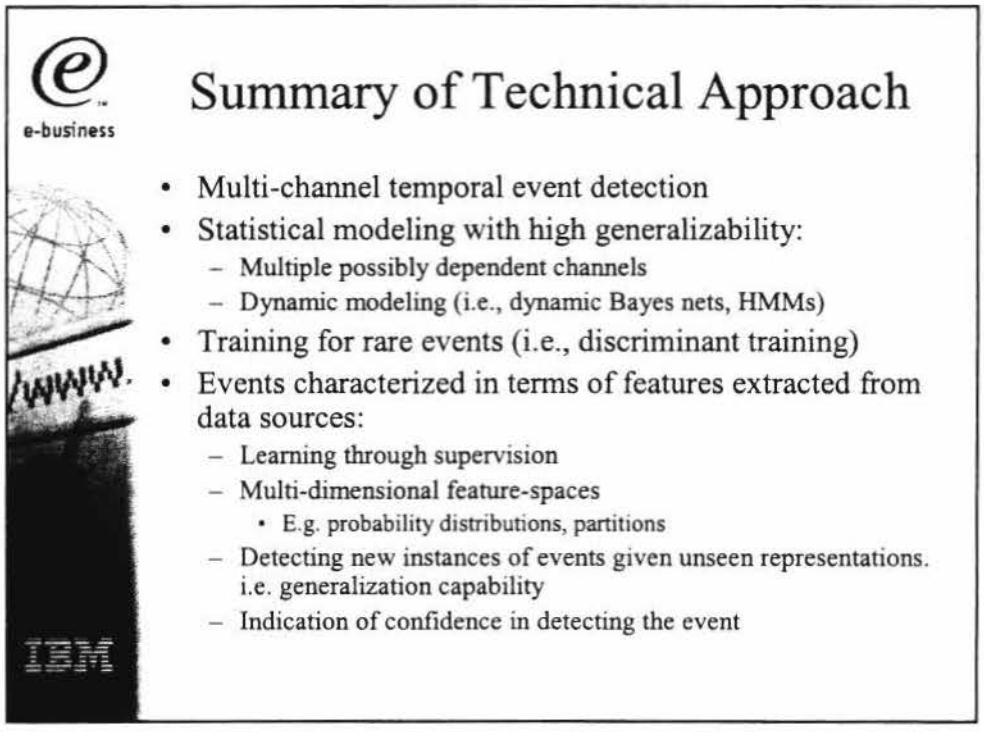
October 11 – 12, 2001

Sankar Basu, Yuan-Chi Chang, Chung-Sheng Li,
Milind Naphade, John R. Smith
IBM T. J. Watson Research Center
30 Saw Mill River Road
Hawthorne, NY 10532 USA
jrsmith@watson.ibm.com


e-business
Statistical Modeling of Spatio-Temporal Events for Bio-Surveillance
• Basic problem:
– Early detection and prediction of biological attacks
• Approach:
– Statistical modeling based on non-traditional data sources (spatio-temporal and temporal data)
– **Stage 1**: development of modeling framework, feature extractors, detectors, and evaluation framework
– **Stage 2**: initial exploration in surrogate domain (audio-visual analysis using video testbed)
– **Stage 3**: generation of artificial data in bio-surveillance domain and simulation and detection of bio attack
IBM


e-business
Bio-Surveillance (Attack Event Modeling)
• Bio-Surveillance information sources (spatio-temporal features):
– Remote sensing images (2-D spatio-temporal):
• Weather (cloud cover)
– Medical (disease manifestation) (2-D spatio-temporal):
• Seasonal health (i.e., flu season), prescription drug sales, school absenteeism records, animal and agricultural health
• Hospital reporting (i.e., ICD codes)
– Site risk (2-D spatio-temporal):
• Weather (i.e., temperature, rain)
• Population density and dynamics (i.e., local traffic)
• Local events (i.e., New Year's eve)
– Financial information (temporal):
• Stocks, financial transactions
– Communications traffic (temporal & spatio-temporal)
• Bulletin board, email, tips hotline, Web, overt threats
• Other relevant factors:
– Release model for specific type of attack
– Socio-politico-economic situations and influences
IBM


e-business
Related Problem Domains
(audio-visual vs. bio-surveillance)
Audio-visual domain:
• Events (semantic)
– Rocket launch
• Multiple asynchronous sources
– Audio, visual, speech, text
• 2D spatio-temporal and temporal data
– Regions, segments, blobs, temporal evolution
• Sparse/missing data
– Partially observable information (i.e., shots)
• Feature extraction
– Color, texture, shape, motion
• Segmentation
– Shot & scene detection, spatial and temporal
• Normalization
– Multiple features and multiple modalities
• Real-time fusion
– Live audio-visual feeds
• Real-time monitoring and prediction
– Real-time detection
• On-line model adaptation and learning
– User feedback
Bio-surveillance domain:
• Events (attacks)
– Anthrax attack
• Multiple asynchronous sources
– Remote sensing, medical, financial
• 2D spatio-temporal and temporal data
– Regions, segments, temporal evolution
• Sparse/missing data
– Partially observable information
• Feature extraction
– Frequency counts, texture
• Segmentation
– Spatial and temporal segmentation
• Normalization
– Multiple features and multiple modalities
• Real-time fusion
– Live data feeds
• Real-time monitoring and prediction
– Real-time detection
• On-line model adaptation and learning
– User feedback
IBM


e-business
Summary of Technical Approach
• Multi-channel temporal event detection
• Statistical modeling with high generalizability:
– Multiple possibly dependent channels
– Dynamic modeling (i.e., dynamic Bayes nets, HMMs)
• Training for rare events (i.e., discriminant training)
• Events characterized in terms of features extracted from data sources:
– Learning through supervision
– Multi-dimensional feature-spaces
• E.g. probability distributions, partitions
– Detecting new instances of events given unseen representations. i.e. generalization capability
– Indication of confidence in detecting the event
IBM

e-business

Specific Modeling Approaches

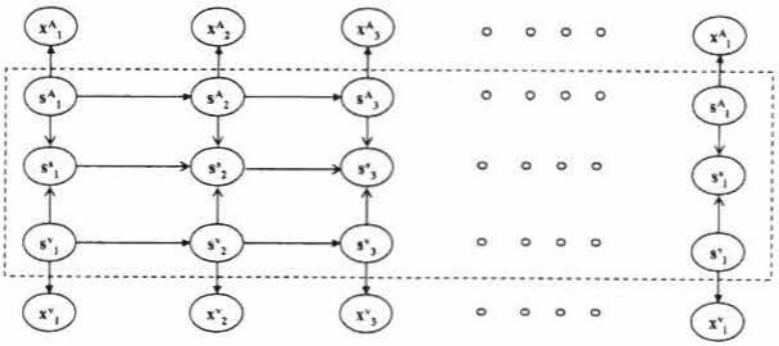
- Models:
  - Modeling of probability density/mass functions in spatio-temporal feature spaces.
    - Parametric models (i.e., Gaussian Mixtures)
    - Nonparametric models (i.e., Parzen windows)
    - Accounting for piece-wise stationary properties by switching between states (i.e., Markov chains, HMMs)
- Classification:
  - Estimation of optimal separation boundaries or partitions
    - Neural nets, support vector machines
  - Rule-based systems
    - Decision trees

IBM

e-business

Multi-modal Graphical Modeling

- Duration density input output Markov model:
  - Decoded state sequences treated as input sequences
  - Multimodal decisions considered the output sequence
  - Output sequence predicted based on the input sequences using explicit duration models, output



IBM

Model Fusion

- Combining results from multiple features and detectors (classifiers)
- Models for each feature type can be built separately (sufficient training)
- The classification results can be combined in the end in various ways
  - Averaging
  - Maximum
  - Product form

Next Steps

- Stage 1: Build modeling framework:
  - Development of feature extractors (spatio-temporal and temporal data) – texture, frequency counts
  - Development of modeling framework (static and dynamic modeling)
  - Development of detectors (training)
  - Development of evaluation framework (testbed)
- Stage 2: Evaluation in surrogate domains
  - Modeling of semantics events in video using audio-visual features (audio, visual, text)
  - Evaluation using standard testbeds (I.e., NIST TREC)
- Stage 3: Simulation and evaluation
  - Generation of artificial data in bio-surveillance domain
  - Simulation of bio attack and evaluation of detection accuracy

e-business

Bottom Line

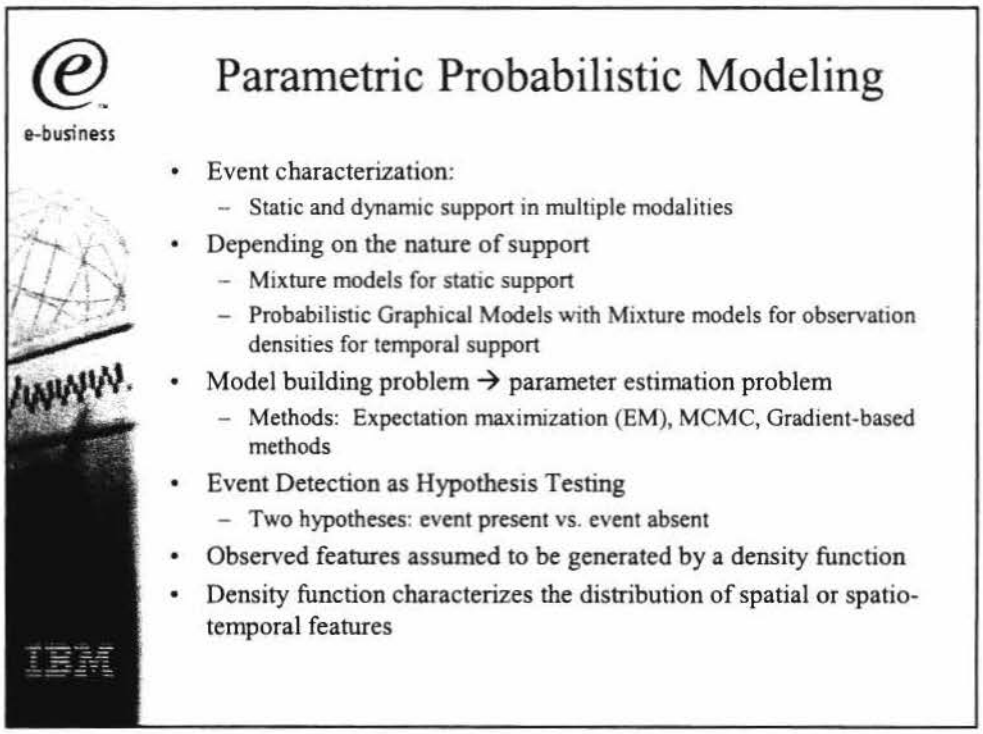
- Integrating **conventional** and **non-conventional** data modalities to improve the robustness of outbreak prediction
  - Feature extraction from non-conventional data sources (when, where, who, what how)
  - Event detection based on spatio-temporal trajectory
  - Data, information and knowledge fusion through model-based learning and search

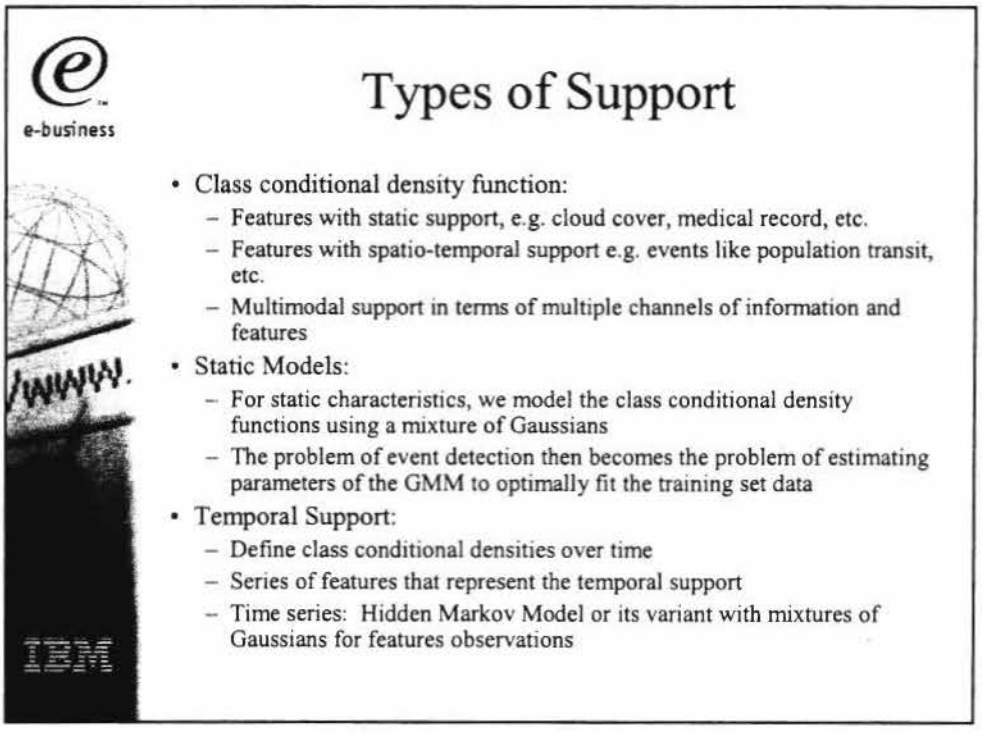
IBM

e-business

Backup

IBM


e-business
Parametric Probabilistic Modeling
• Event characterization:
– Static and dynamic support in multiple modalities
• Depending on the nature of support
– Mixture models for static support
– Probabilistic Graphical Models with Mixture models for observation densities for temporal support
• Model building problem → parameter estimation problem
– Methods: Expectation maximization (EM), MCMC, Gradient-based methods
• Event Detection as Hypothesis Testing
– Two hypotheses: event present vs. event absent
• Observed features assumed to be generated by a density function
• Density function characterizes the distribution of spatial or spatio-temporal features
IBM


e-business
Types of Support
• Class conditional density function:
– Features with static support, e.g. cloud cover, medical record, etc.
– Features with spatio-temporal support e.g. events like population transit, etc.
– Multimodal support in terms of multiple channels of information and features
• Static Models:
– For static characteristics, we model the class conditional density functions using a mixture of Gaussians
– The problem of event detection then becomes the problem of estimating parameters of the GMM to optimally fit the training set data
• Temporal Support:
– Define class conditional densities over time
– Series of features that represent the temporal support
– Time series: Hidden Markov Model or its variant with mixtures of Gaussians for features observations
IBM

e-business

Parameter Estimation and Detection

Parameter Estimation:

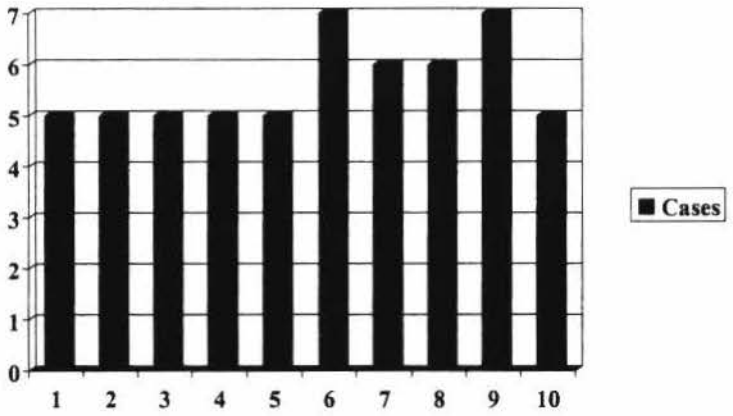
- Given a set of sufficient number of examples for each hypothesis, we can estimate the parameters of the GMM.
- Parameters
  - can be considered random leading to the problem of Bayesian
  - can be treated as deterministic but unknown.
- We treat parameters as deterministic but unknown and estimate them by maximizing the likelihood of observing the training data given the parameters.
- The maximum likelihood problem is set up using the EM algorithm

Detection:

- Once the models for both hypotheses have been estimated, we compute the log likelihood of the unknown features under each hypothesis and compute their difference.
- This serves as a measure of confidence in detecting the concept. The larger the confidence the greater the difference.
- Combining this with the prior on the occurrence of each hypothesis we can compute the probability of detection of the concept given the features.

e-business

Time Domain

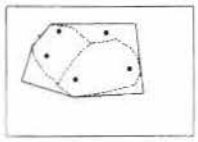


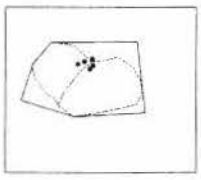
IBM

e-business

Spatial Patterns are Significant

- No difference from baseline counts
- Different spatial patterns may lead to different interpretation of event



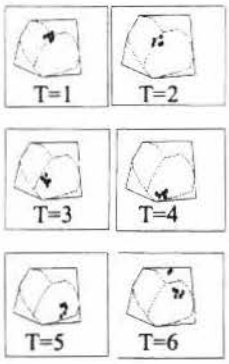


IBM

e-business

Evolution over time

- Overall count stable until T=6
- Counts/polygon somewhat more informative
- Spatial pattern indicative of active process

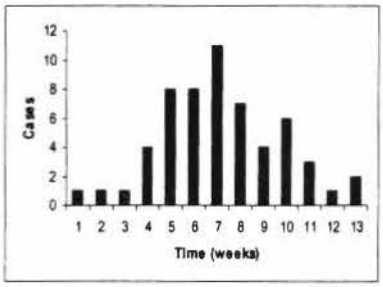


IBM

e-business

Shigella Outbreak Ft. Bragg, NC

- Outbreak of diarrheal disease among dependents and military personnel in and around Ft. Bragg, NC
  - McKee et al 2000 JID



IBM

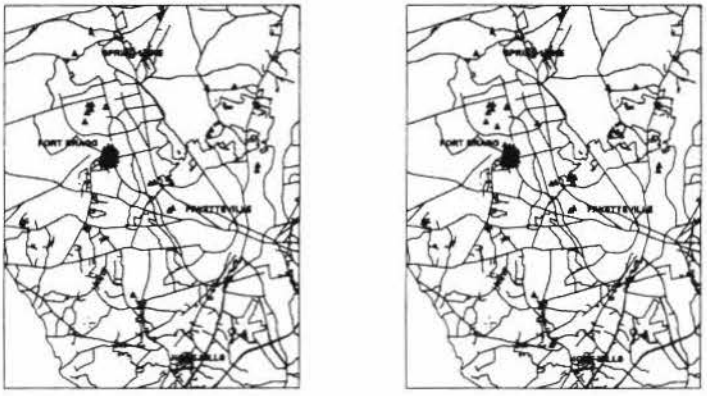
e-business

Cumulative Cases Week 1-4 (left) & 1-6 (right)

IBM

e-business

Cumulative Cases Week 1-8 (left) & 1-10 (right)



IBM

Risk Modeling

- Prevention
  - Infer permissive/suitable conditions for transmission
    - Targeted monitoring
    - Scenario evaluation -- how would it evolve in time/space – SEIR models – linked local models
  - Approach
    - Identify key environmental variables
    - Biological relationship of EV's to disease risk
    - Likely use --VBD's/zoonotic agents
- Earlier Detection
  - Identification of ongoing situation
    - Assist in identification of causal agent/source
    - Assist in action plan
      - How many cases, where, most likely individuals to query

Prevention

- Define environmental conditions that permit transmission of agent
  - (Issues of where, when, how long need environmental conditions)
- Model dynamics of risk
  - A priori established biological relationship between agent/vector/reservoir and environmental conditions
    - e.g. fire ants
  - Statistical relationship between environmental conditions and agent/vector/reservoir – past data about outbreak
    - e.g. hantavirus

Environmental Data

- Time scale is relative to disease progression in a case
- Time sensitive (variable relative to dz course)
  - Meteorological conditions
  - Transportation network load
  - Socio-behavioral responses to information
- (Relatively) time invariant (stable relative to dz course)
  - Land cover
  - Population size
  - Topography
  - Transportation network location
  - Socio-economic conditions

Difficulties

- Major problem is that expect heterogeneity in transmission/risk/behavior with environmental conditions
- Relevant data acquisition in NRT

e-business

Space-time Disease Model - General

- $Y_k(s,t)$ -- RV counts of disease observed -- dependent on pathogen, place and time
- $\lambda_k(s,t)$ -- Actual attack rate (rather than counts). $\lambda_k(s,t)$ is function of number of cases likely to occur. This measure is influenced by factors that vary over time (daily, seasonally and yearly) and space, ie.,
- $\log(\lambda_k(s,t)) = R(t) + U(s)$,
- R and U are the functions that describe how the log of attack rate at a particular time and place are influenced by the environmental variables.
- The temporal function, is modeled using the usual time series regression
- $R(t) = B_o + B_1X_1(t) + \ldots + B_pX_p(t) + e(t)$
- where $X_1(t), \ldots, X_p(t)$ are the climatic variables (and other rapidly changing variables, such as temperature, windspeed, and traffic flow) whose impact we want to evaluate at particular times, t.
- The parameters, $B_o, B_1, \ldots, B_p$, represent the strengths of those associations. A random error component, e(t), is incorporated and assumed to have a Gaussian distribution. However, **the model allows to autocorrelation in the errors, for temporally varying EV's.**

IBM

e-business

Space-time Disease Model - General II.

- The spatial component of the analysis U(s), is analyzed by creating a universal kriging model (Cressie 1990).
- $U(s) = a_o + a_1Z_1(s) + \ldots + a_qZ_q(s) + d(s)$,
- $Z_1(s), \ldots, Z_q(s)$ represent environmental variables, such as land cover, or elevation that are relatively unchanging over time but influence risk
- $a_o, a_1, \ldots a_q$ represent the strength of these associations and random error d(s) is incorporated. **The random error is structured so that it is correlated with distance between sampling sites (spatial correlation), rather than being independent** (Cressie 1990).
- Incorporating both spatially and temporally correlated random effects reduces mis-specifying the environmental effects on outcome and reduces smoothing effects.
- Overall model also fits into the framework of **generalized linear mixed models** (Breslow and Clayton 1993).
- Can infer the model parameters using **maximum likelihood and Monte Carlo Newton Raphson (MCNR) method** to estimate the model parameters (McCulloch 1997).

IBM

e-business

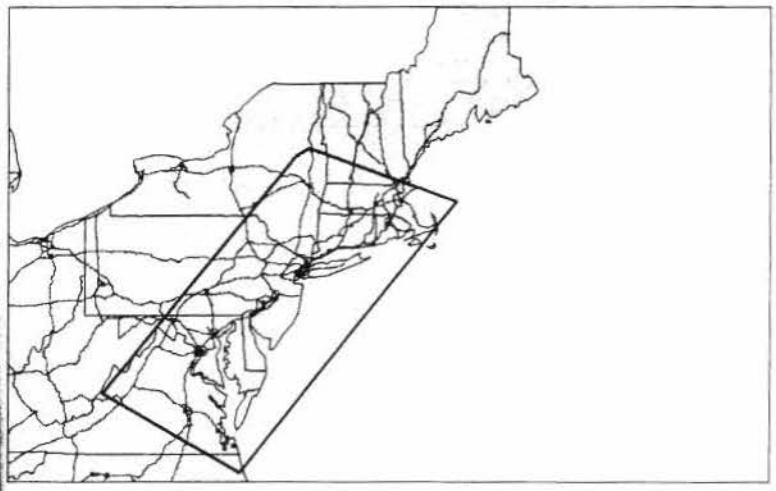
Analytical Outcome

- Identification of associated environmental variables with counts of cases/syndromic classes
  - If *B*'s are significant then variables at those times were relevant to cases – clue to source/agent e.g. *B*(wind direction 3 days prior to occurrence of cases)
  - If *a*'s are significant then event occurred over shorter time period than was sampled or not affected by measured time sensitive variables – e.g. anthrax in bldg vs wind direction

IBM

e-business

Target Area



IBM

e-business

Near-real-time Weather Da

Continuous monitoring of environment variables

Air temperature
Relative humidity
Precipitation
Wind speed
Wind direction

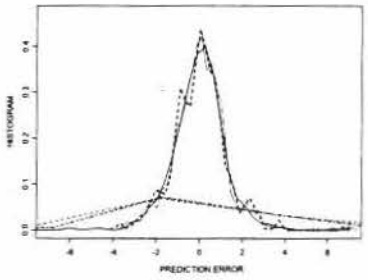
Derived from station readings and satellite images for better resolution

Source: NOAA National Climate Data Center

e-business

Improvement in Ground Temperature Estimation

- Merged satellite, ground station data
  - Errors center on zero
  - $\pm$ 1° C
  - Can be estimated under clouds
- Anticipate further improvement with additional covariates
  - NASA/IBM TJ Watson



Meteorological Data Sources

- Approximately 500 weather stations provide hourly feeds
- Estimate temperature, windspeed, wind direction, RH, precipitation 3x/day

Invariant Environmental Data

- Population distribution
  - Y2000 U.S. Census data
- Land cover -- MRLC
  - Influencing VBD/Zoonoses
  - Covariate for meteorological estimators
  - Modify population distribution estimates
- Topography
  - USGS DEM
- Transportation networks
  - U.S. DLG's

Time Sensitive Data

- Transportation network flow
  - Public transportation
    - Air travel -- 150,000 ppd from three Baltimore-Washington metro
      - 35,000 employees
      - 12,000 taxi dispatches/day
    - O-D matrix model of commercial air transport
      - Commercially available, daily summary of site/time of departure and arrival, plane type (upper bounds on passenger numbers).
    - O-D model for commercial travel train, light rail for region
  - Private transportation
    - Car travel
      - MDOT

Private Transportation Movement

- Traffic Speed Data
- This data will automatically refresh every five minutes.

Location	Average Speed	Last Reported
I - 83 @ Cold Bottom Rd South	61 MPH	10/10/2001 4:38:39 PM
I-270 @ Comus Rd North	62 MPH	10/10/2001 4:38:39 PM
I-270 @ Doctor Perry Rd North	50 MPH	10/10/2001 4:38:39 PM
I-270 @ I-370 North	65 MPH	10/10/2001 4:38:39 PM

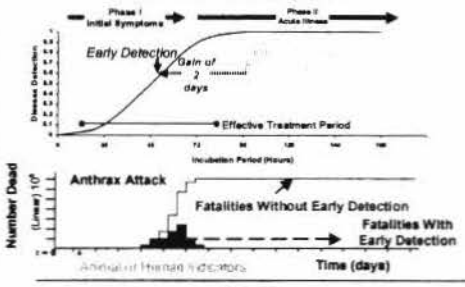
BioSurveillance Validation

David Siegrist
Potomac Institute for Policy Studies
August 22, 2001

Bio-Surveillance Validation and Transition

DARPA

Need for Early Detection

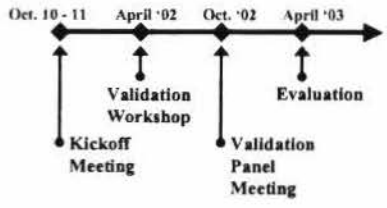


- Need to Validate New System w/ Little Data
  - Few Historical BW Data
  - Uses Non-Traditional Data Sources
  - Advanced Data Techniques
- Transition to Non-Traditional User
- Approach: Rigorous and Metrics Based
  - Use Influenza and Other Disease Analogues
  - ROC Curves for Sensitivity, Specificity
  - Panel of Experts Verifies Analysis
    - World Class Epidemiologists, BW Experts, Privacy and Other Specialists

Impact

- Validation
  - Building on Historical Base and Extending to Military Pathogens Increases Credibility to User
  - Builds Confidence in System Alerts
- Transition
  - Technology in Daily Use
    - Protects DoD Personnel
    - Documents Personnel Readiness

Schedule



Dave Siegrist
Potomac Institute for Policy Studies

Strawman Draft Scoring

• Use Non-Traditional Data Sources:	Pass/Fail
• Make Early Determination of Disease Outbreak:	20 Pts.
• Support Clustering (ID Disease in a Particular Area):	10 Pts.
• Support Clustering in Numerous Areas:	10 Pts.
• Disease Sensitivity (Identify Minor Perturbations):	10 Pts
• Disease Specificity (Differentiate Similar Diseases):	10 Pts.
• Minimal False Alarms:	10 Pts.
• DB Partition Supports Info Trading:	15 Pts.
• Technology Extensible to Other Regions	10 Pts.
• Advanced technology	5 Pts.
• Complete Personal Privacy Assurance:	0 Pts.
	100 Pts.
• Depersonalization without Anonymity:	Minus 20 Pts.
• Release Patient Information on Internet:	Minus 50 Pts.

- Violate HIPAA: Possible Prison Term and Fine
- False Negative: Possible Excess Morbidity and Mortality

Note: ROC Curve May Summarize Sensitivity/Specificity "Goodness"

Possible Validator Panelists

- Dr. David Franz, DVM, Southern Research, Former RIID
- Dr. Ken Alibek, PB Biosystems; Former *Biopreparat*
- Dr. Eric Noji or Dr. Tracee Treadwell, CDC
- Dr. Ken Hoffman, VA, Evaluator
- Dr. Everett Koop, Former Surgeon General
- Dr. Tracey McNamara, DVM, Bronx Zoo
- Dr. Bill Patrick, Former AMRIID
- Dr. Stephen Prior, Microbiologist, Potomac Institute
- Peter Lejeune, SMI and FEMA, Biosensors and Consequence Mgt
- Military Rep from JFCOM or Enviro Medicine?
- Sydney Connor, SAI Epidemiologist
- Jamie Pugh, Naval health Research Center, Epi Algorithms

Possible Validation Technique

- Independent Metrics-Based Rigorous Assessment
  - Potomac Institute Team
  - Analysis of Developer Assumptions & Data
  - Scoping of Algorithm Validity, Extensibility
  - Develop Life Cycle Validation Model w/Decision Points
  - No Test Cases and Software V&V at This Time
- Performers Develop Flu Season Onset Model with Own Data
  - Test it Against Historical Data. Predict Future Flu Season?
  - Develop Baseline Against Which to Normalize Possible Bioattack
  - Develop Credible Approach for Other Prodromes, Pathogens
- Program Briefing to Validator Panel
  - Potomac Synthesizes Views and Generates Report for PM

1.A.12

VERIDIAN SMI

Bio-Surveillance System Kickoff Meeting 11–12 October, 2001

A Bio-Surveillance System for Advanced Medical Readiness

Gene McClellan, PhD, Veridian
Mark A. Musen, MD, PhD, Stanford University
Martin Liggins, Veridian
Mildred Cho, PhD, Stanford University
Judith Brillman, MD, University of New Mexico

VERIDIAN SMI

Briefing Outline

- McClellan
  - Overview and Technical Approach
  - Data sources
  - Statistical processing ... BASIS
- Musen
  - Knowledge-based processing ... Bio-STORM
- Liggins
  - Multiple Hypothesis Fusion
  - Notification and Presentation
- Cho
  - Ethical Issues
- Brillman
  - Syndromes vs. Diagnoses
  - Collection of Data from Physicians
- McClellan
  - Simulation and Models
  - Milestones/schedule
  - Issues

VERIDIAN SMI

Overview and Technical Approach

Dr. Gene McClellan
Veridian

VERIDIAN SMI

The Problem

Traditional disease surveillance:

- Medical infrastructure is focused on individual patients.
- Surveillance by serendipity: an astute physician recognizes a patient's disease and orders tests to confirm the clinical indicators.
- Only after disease has been confirmed can a regional health response can be initiated.

For a covert terrorist event, the delays associated with this process will cause poor management of casualties.

New information technologies are needed to rapidly detect disease in the population.

VERIDIAN SMI

The Solution

- **Collect Behavioral and Medical Data**
  - Automatically and routinely acquire behavioral and medical data
  - Create relevant databases
  - Maintain medical privacy
- **Process and Analyze Data for Anomalies**
  - Integrate heterogeneous data
  - Create baseline normalcy models
  - Continuous Disease Hypotheses Analysis
  - Produce statistically significant indicators of population health
- **Effective Notification of Public Health Officials**
  - Inform action

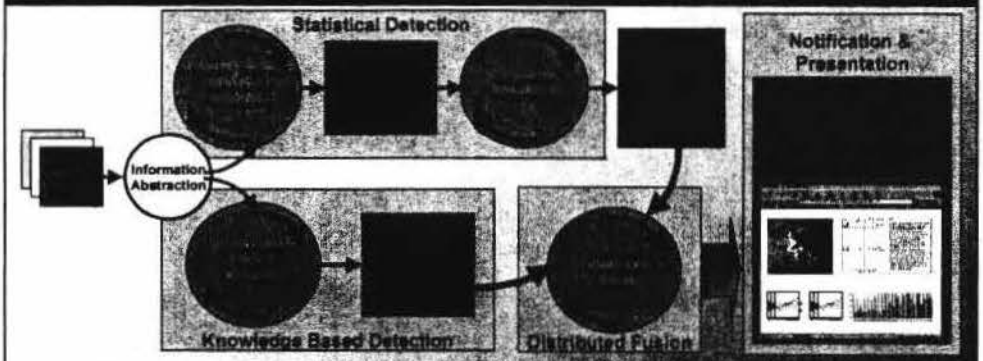
VERIDIAN SMI

Our Unique Contributions

- **Innovative ideas for early sign/symptom detection**
- **Integration of statistical and knowledge-based processing throughout the architecture**
- **Early generalization of system using disparate geographic areas**
  - San Francisco – ongoing relationship with public health system
  - Norfolk – initial focus on military data sources
- **Specific technology for creating data abstractions both within and across patient records**
- **Use of context to guide data interpretation**
- **Systematic incorporation of bioethical principles**

VERIDIAN SMI

Top Level Functional Flow



- **Disease tracking requires a tight coupling of technologies in order to [illegible] reduce false alarms and provide consistent timely updates**
  - Statistical detection links weakly typed indicators (e.g. [illegible]) for integrated assessment
  - Knowledge-based detection correlates existing medical [illegible] measures of illness
  - Multiple hypothesis fusion creates the best disease hypothesis

VERIDIAN SMI

Implementation

This Bio-Surveillance program will result in the development of:

- Autonomous detection of biological events
- Disease and epidemiological models to support detection, classification, and outbreak characterization
- An emulation capability to ensure that the detection and classification algorithms/models are robust in relevant scenarios
- A prototype system for the Norfolk area

VERIDIAN SMI

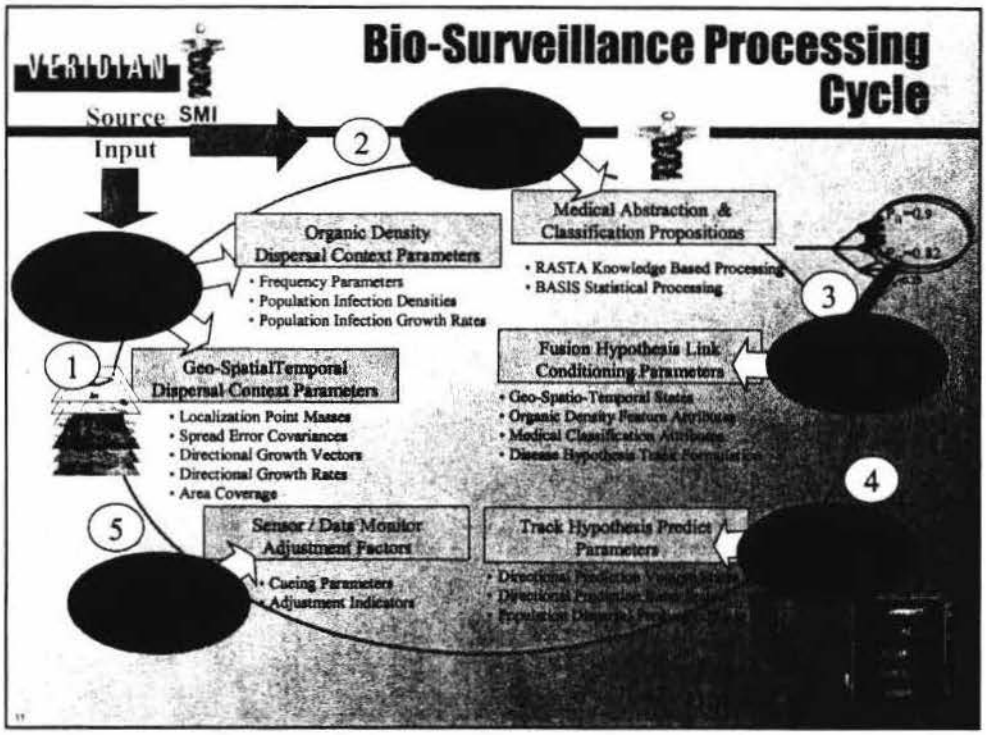
Benefits

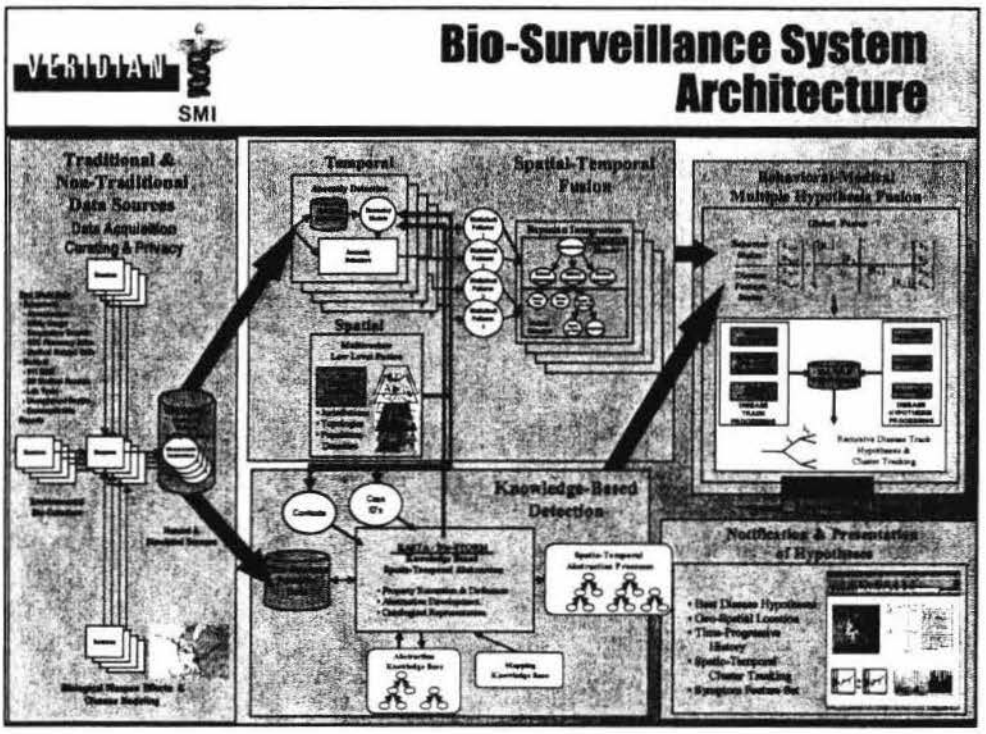
- **Allow rapid response by civilian and military public health officers**
  - Highly predictive health alerts to enable epidemiological investigation
  - Decrease morbidity/mortality of outbreaks
- **Improve disease surveillance**
  - Preparedness against naturally occurring disease
  - Preparedness against biological attack
- **Promotion of medical readiness of U.S. military forces through specific design features oriented toward military bases and the communities in which they reside**

VERIDIAN SMI

Bio-Surveillance Technologies

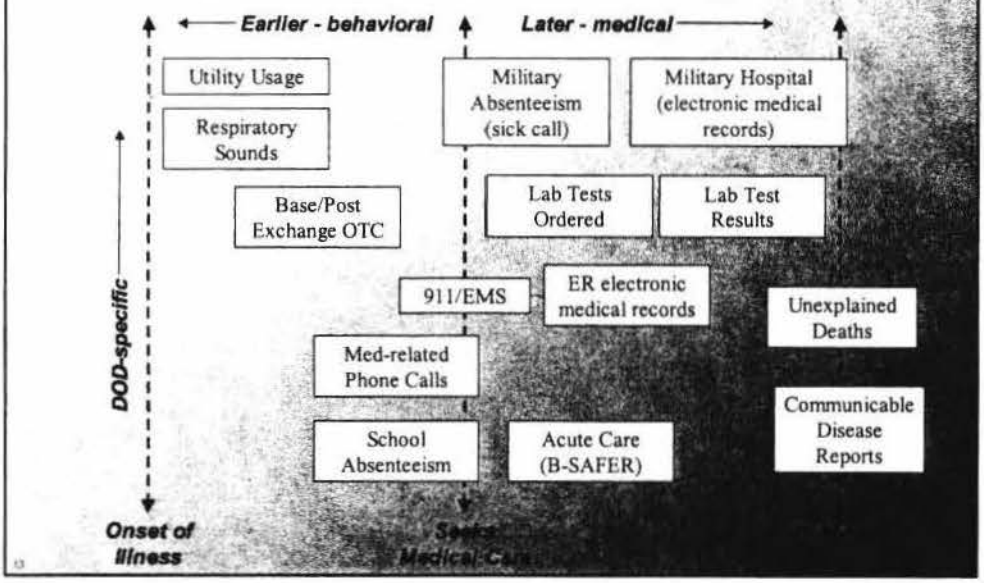
- **The success for a successful bio-surveillance system requires the tight coupling of a number of technologies**


VERIDIAN
SMI
Bio-Surveillance Processing Cycle
Source Input
2
Organic Density Dispersal Context Parameters
• Frequency Parameters
• Population Infection Densities
• Population Infection Growth Rates
Medical Abstraction & Classification Propositions
• RASTA Knowledge Based Processing
• BASIS Statistical Processing
3
1
Geo-SpatialTemporal Dispersal Context Parameters
• Localization Point Masses
• Spread Error Covariances
• Directional Growth Vectors
• Directional Growth Rates
• Area Coverage
Fusion Hypothesis Link Conditioning Parameters
• Geo-Spatio-Temporal States
• Organic Density Feature Attributes
• Medical Classification Attributes
4
5
Sensor / Data Monitor Adjustment Factors
• Cueing Parameters
• Adjustment Indicators
Track Hypothesis Predict Parameters


VERIDIAN
SMI
Bio-Surveillance System Architecture
Traditional & Non-Traditional Data Sources
Data Acquisition
Curating & Privacy
Temporal
Spatial-Temporal Fusion
Spatial
Behavioral-Medical Multiple Hypothesis Fusion
Knowledge-Based Detection
Notification & Presentation of Hypotheses

VERIDIAN SMI

Traditional and Non-Traditional Data Sources



VERIDIAN SMI

Initial Data Sources

- Establish Primary Data Sources
  - OTC Drug Sales √
  - 911/EMS √
  - Hospitals √
  - ER √
  - HMO
  - Military MTFs
- Data Source Logistics
  - POCs
  - Privacy Issues
  - Paperwork

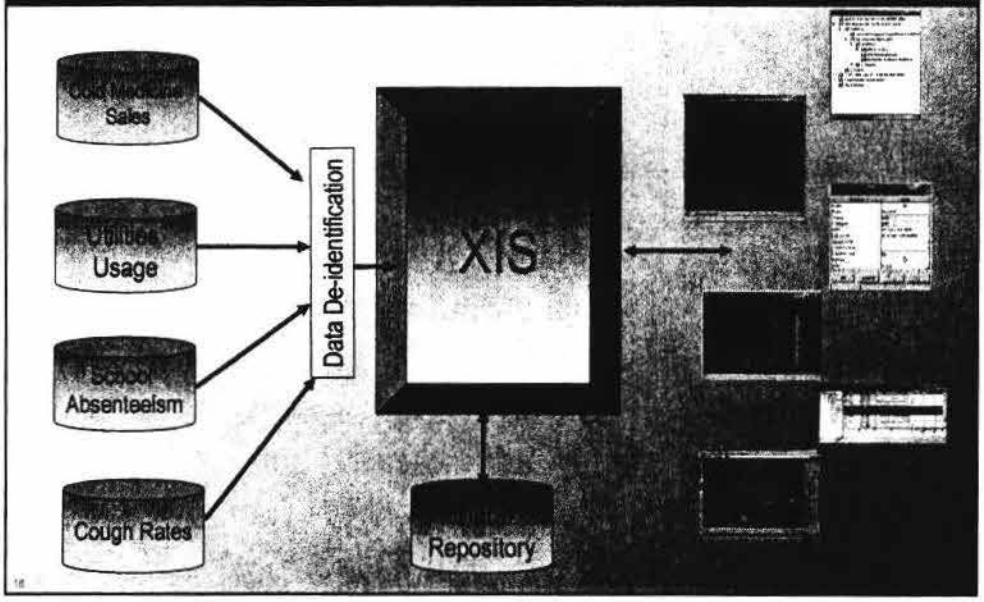
VERIDIAN SMI

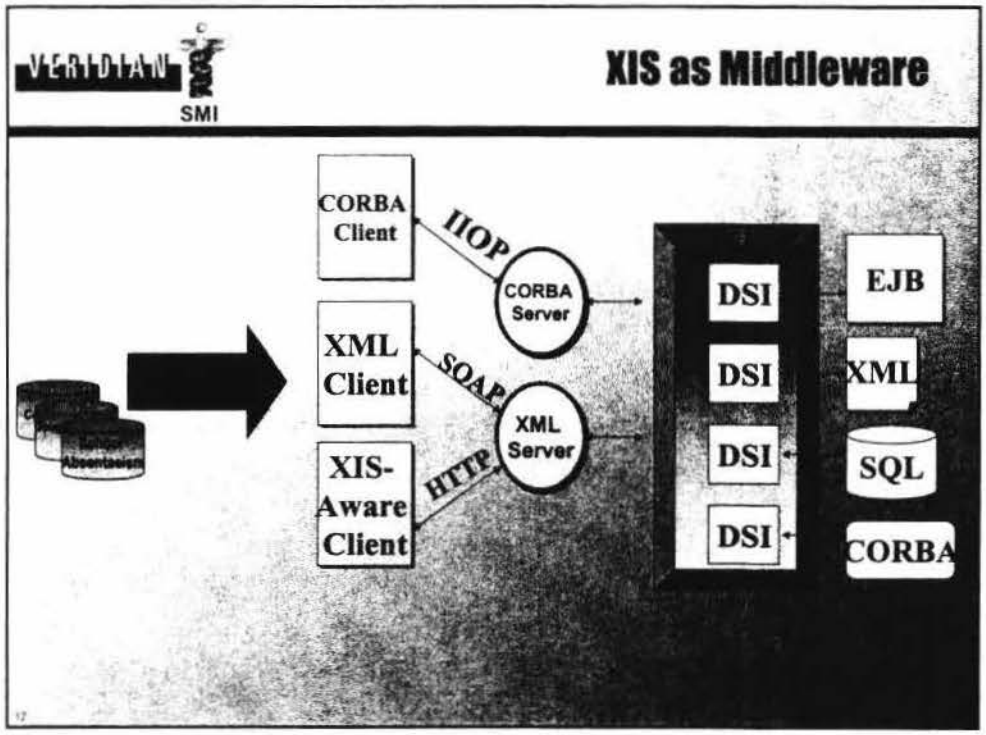
Data Processing - Goals

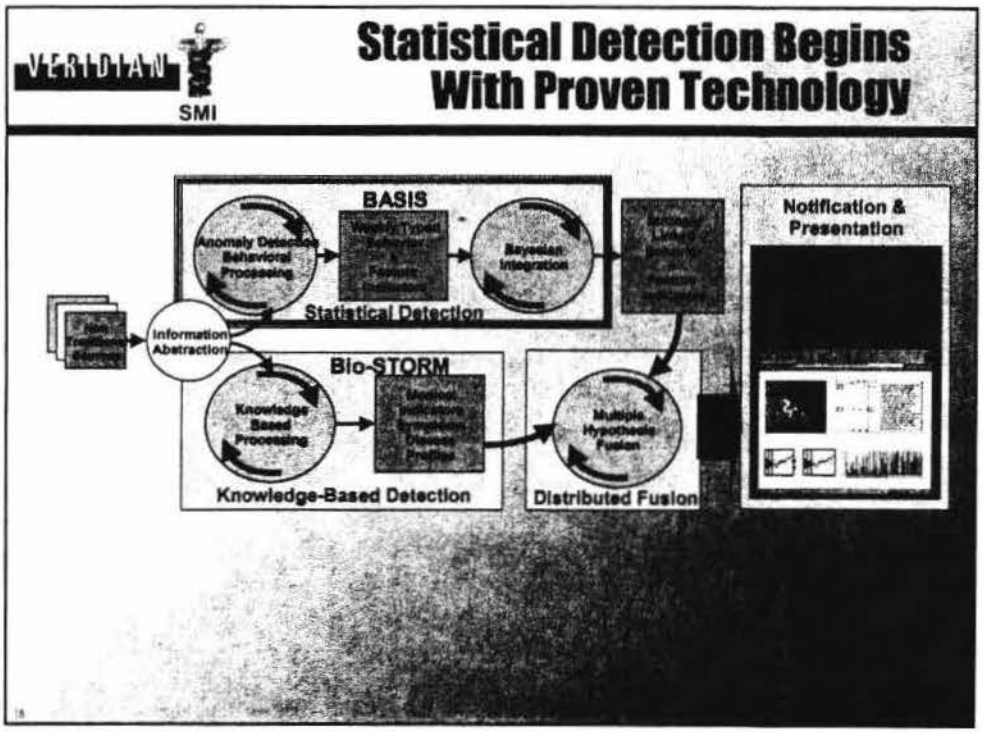
- Integrate Heterogeneous Data Sources in a flexible environment
  - Polexis XIS
- Create Data Ontologies
  - Protégé 2000
- Simulate data within emulation environment
- Establish timeline for transition from data simulation to real-world data

VERIDIAN SMI

Polexis's Extensible Information System (XIS)




VERIDIAN
SMI
XIS as Middleware
CORBA Client
IIOP
CORBA Server
XML Client
SOAP
XML Server
XIS-Aware Client
HTTP
DSI
DSI
DSI
DSI
EJB
XML
SQL
CORBA


VERIDIAN
SMI
Statistical Detection Begins With Proven Technology
BASIS
Anomaly Detection Behavioral Processing
Bayesian Integration
Statistical Detection
Information Abstraction
Bio-STORM
Knowledge Based Processing
Knowledge-Based Detection
Multiple Hypothesis Fusion
Distributed Fusion
Notification & Presentation

VERIDIAN SMI

Statistical Detection with BASIS

Biological Agent Symptomatology Information System

- Developed over the last three years for the Defense Advanced Research Projects Agency (DARPA)
- Medical surveillance tool for Biological Warfare Defense program
- Provide early indication of biological attack in the absence of bio-detectors
- Applications in both military and civilian arenas
- For maximum utility, must be useful in routine epidemiology

19

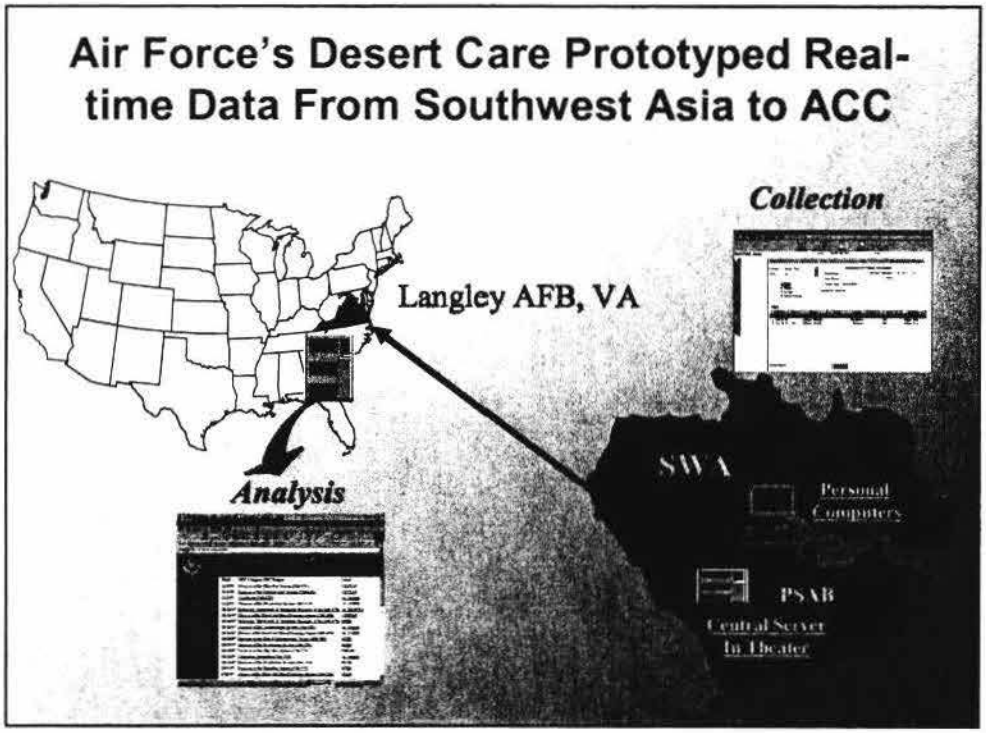
VERIDIAN SMI

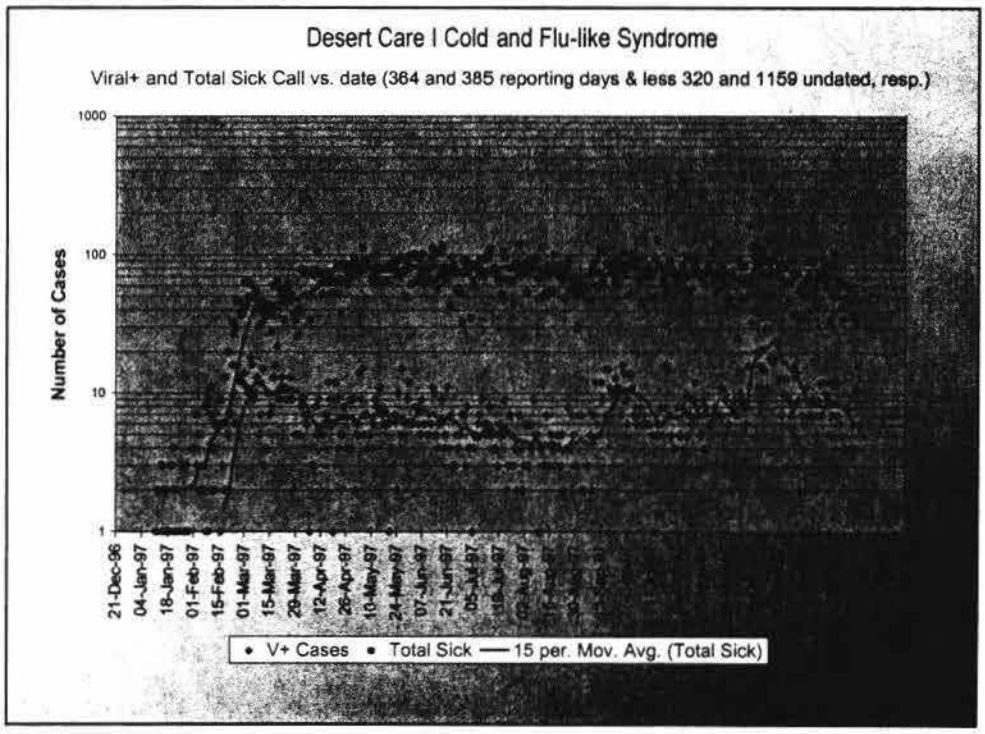
BASIS Provides Early Alerts

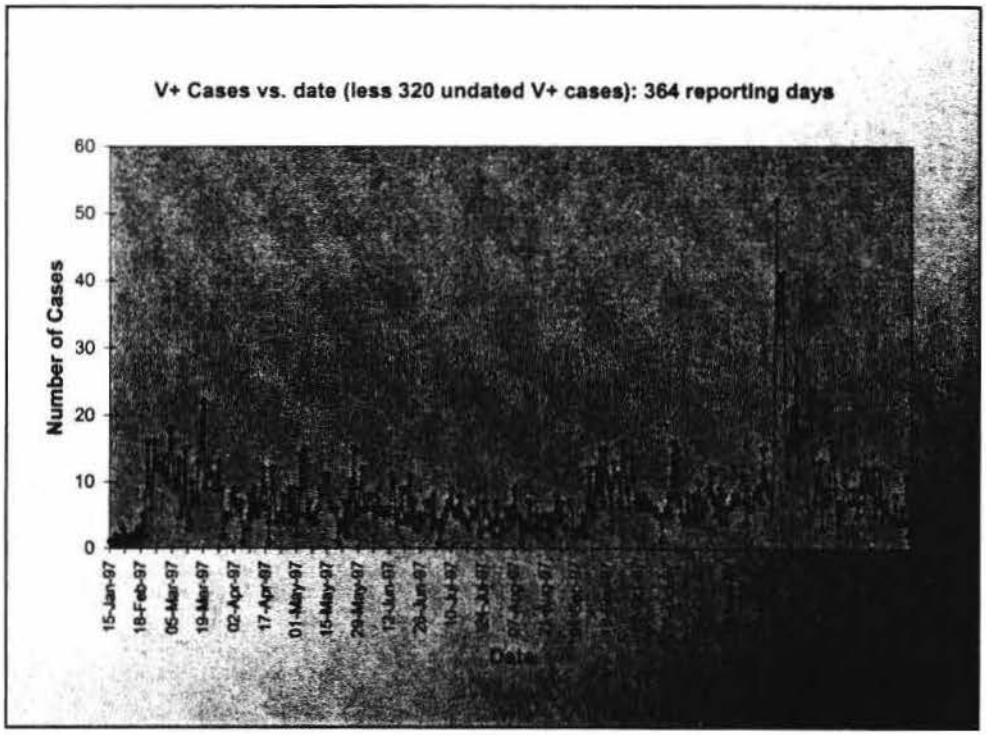
- Utilizes electronic medical records
- Statistical monitoring and comparison
  - C-chart
  - 2x2 contingency table
- Adjustable alert threshold

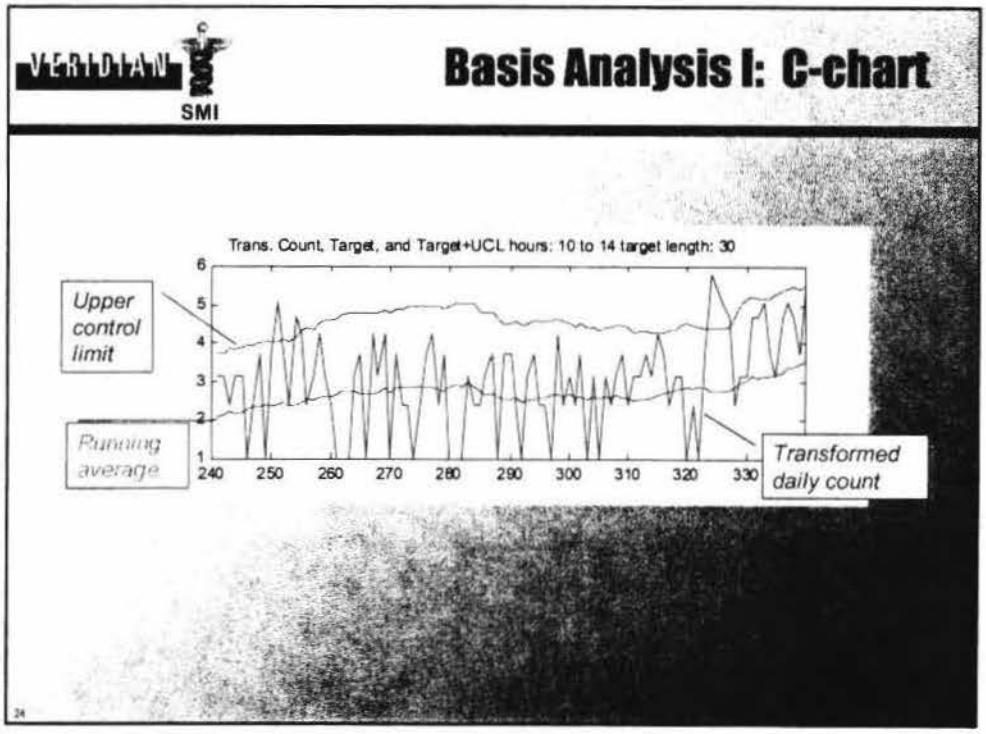
20

Air Force's Desert Care Prototyped Real-time Data From Southwest Asia to ACC



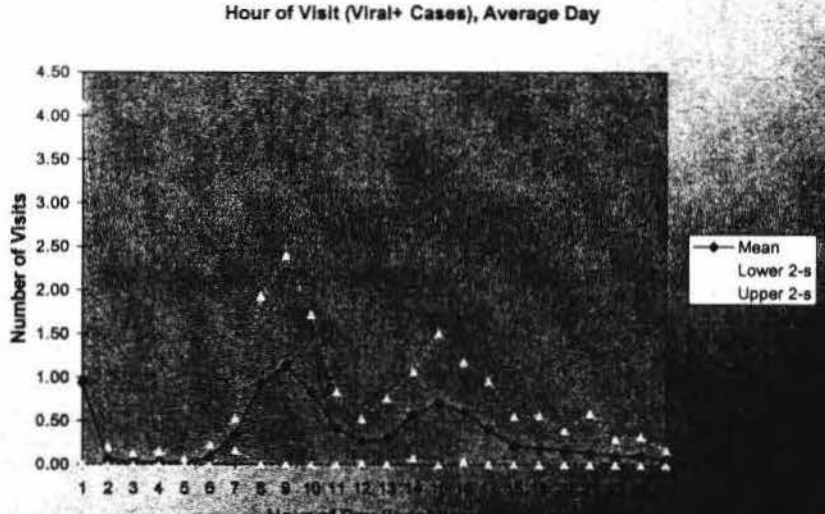



V+ Cases vs. date (less 320 undated V+ cases): 364 reporting days
Number of Cases
60
50
40
30
20
10
0
15-Jan-97
18-Feb-97
05-Mar-97
19-Mar-97
02-Apr-97
17-Apr-97
01-May-97
15-May-97
29-May-97
12-Jun-97
26-Jun-97
10-Jul-97
Date


VERIDIAN
SMI
Basis Analysis I: C-chart
Trans. Count, Target, and Target+UCL hours: 10 to 14 target length: 30
Upper control limit
Running average
Transformed daily count
6
5
4
3
2
1
240
250
260
270
280
290
300
310
320
330

VERIDIAN SMI

Diurnal Variation Is Significant



VERIDIAN SMI

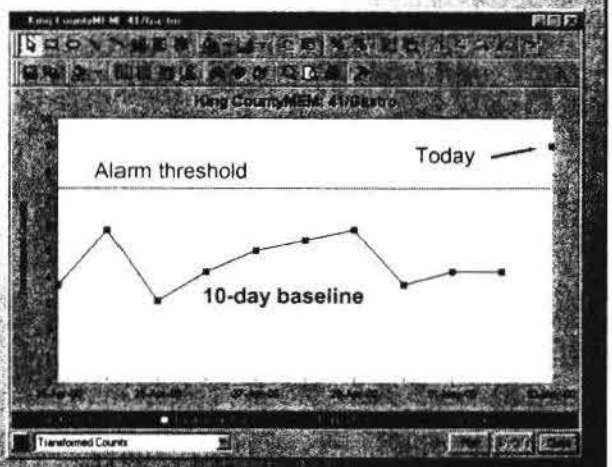
Basis Analysis II: 2x2 Contingency Table

- Accommodates imperfect baseline data
- Accommodates changing population
- Will alarm early in reporting period

VERIDIAN SMI

BASIS Monitors Time Series Data

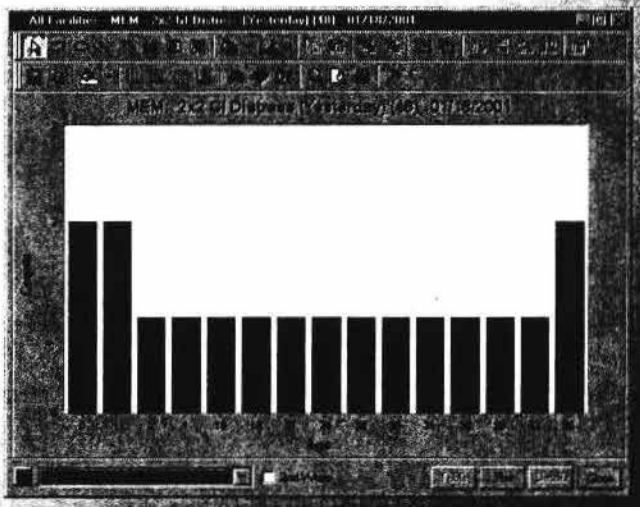
- Timeline analyses of
  - Signs/symptoms
  - Syndromes
  - Diagnoses
  - Any discrete observable
- Baseline
  - From database, or
  - Historical
- Monitors "today" and "yesterday"
- Supports routine epidemiology
  - Tailored timelines
  - Data drill-down



27

VERIDIAN SMI

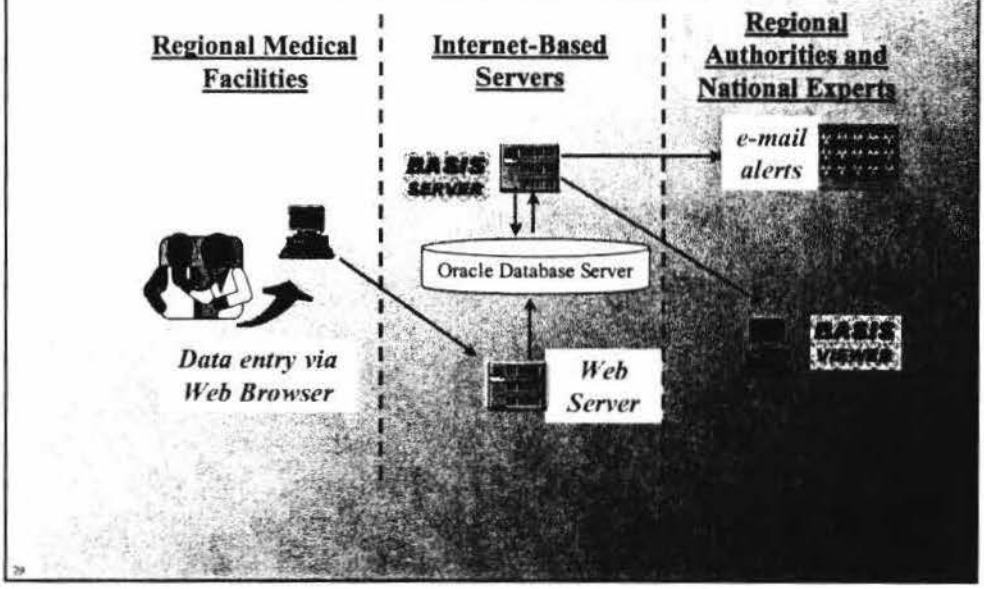
BASIS Provides Interactive Examination of Case Data



28

VERIDIAN SMI

BASIS Enables Automated Alarming



VERIDIAN SMI

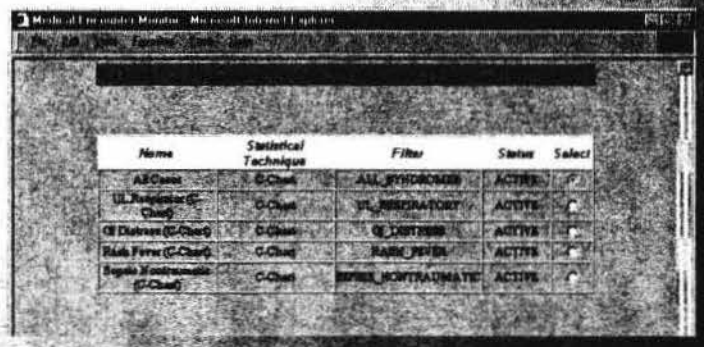
BASIS – Track Record

- Deployed successfully for syndromic surveillance during DARPA's support of CDC at
  - WTO Meeting, Seattle, Washington, November 1999
  - Republican Convention, Philadelphia, Pennsylvania, July 2000
  - Democratic Convention, Los Angeles, California, August 2000
  - Presidential Inauguration, Washington, DC, 2001
- Extensive experience with the mechanics of real-time functionality

VERIDIAN SMI

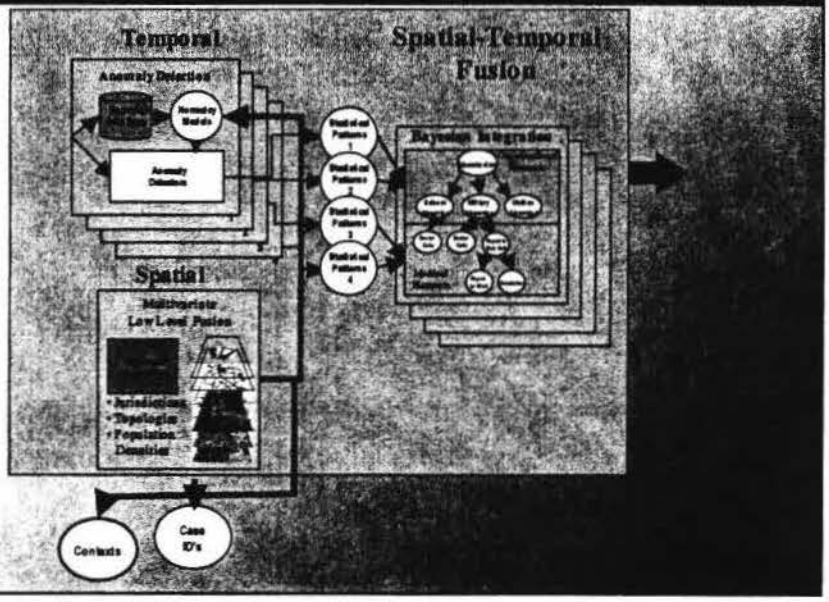
BASIS – Status

- In place at Univ. New Mexico to monitor RSVP data
- Can be altered to support most any data source
- Web version can be available to anyone on the internet with a browser and the BASIS URL



VERIDIAN SMI

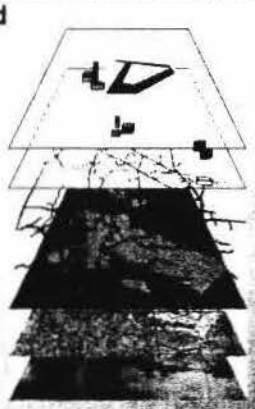
Details of Statistical Detection



VERIDIAN SMI

Layered Evaluation Aids in Associating Disease Spread

- Layering provides ideal means to associate disparate measurement primitives
- Convenient method to build natural and functional barriers to disease spread



Region of Interest

School Sites and Jurisdictions

Town / Base Road Infrastructure

Hydrology (Natural Disease Barriers)

Manmade Disease Barriers

Cartographic Maps

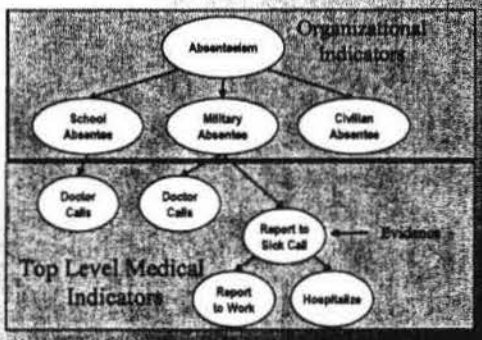
Disease Spread Probes

33

VERIDIAN SMI

Bayesian Integration Generates Strong Indicators

- Some weakly typed indicators can be linked directly
- Functions hand-in-hand with knowledge-based detection
- Feeds fusion system, which in turn provides most likely disease hypotheses



Bayesian Inference

34

VERIDIAN SMI

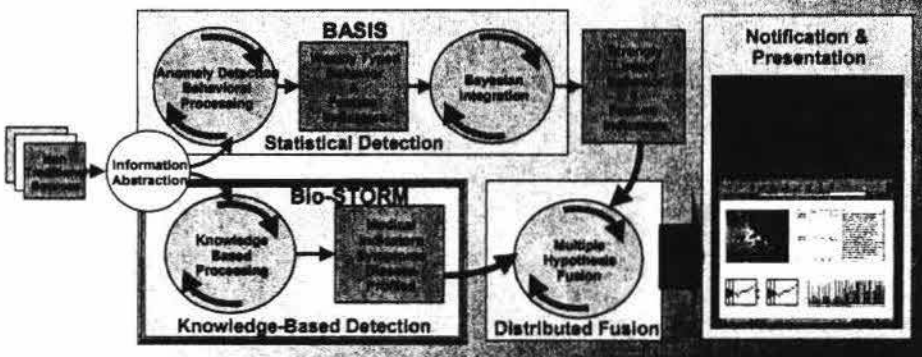
A Bio-Surveillance System for Advanced Medical Readiness: *Knowledge-Based Processing*

Dr. Mark Musen

Stanford Medical Informatics

VERIDIAN SMI

Integrating Knowledge-based Techniques with Statistical



Ideal data for automated surveillance would ...

- Be relevant to the problem of interest
- Have high specificity and sensitivity
- Be easily and rapidly collected
- Reflect a narrow window of space and time
- Respect privacy boundaries as much as possible

37

Data Available Through The Stanford Team

- Palo Alto Veterans Affairs Hospital
  - Complete medical record data
  - Inpatient and outpatient orders, lab results, and vital signs
  - Inpatient and outpatient pharmacy data
- San Francisco Department of Health
  - EMT dispatch data
  - ER diversion data
  - Reports of suspicious animal deaths

38

Data Available Through The Stanford Team, continued

- California Department of Health
  - Communicable disease reports
  - Unexplained deaths
  - Veterinary data (agricultural)
- San Francisco General Hospital Emergency Room
  - Nursing assessment and triage data
  - Orders, lab results, and vital signs
  - Condition on discharge and discharge diagnoses

Many relevant data are "qualitative"

- The occurrence of a political event that might incite bioterrorism
- Identification of an unusual micro-organism in the laboratory
- The suspicious death of an animal
- Reports of a new epidemic in an adjacent geographical region

VERIDIAN SMI

Qualitative data

- Provide context for interpretation of other data
- May require interpretation themselves
- Do not lend themselves to summary statistics
- Do not lend themselves to statistical comparisons
- Require domain knowledge for interpretation

41

VERIDIAN SMI

Even when data are numeric, what often matters are *patterns* in the data

"Has a patient experienced a second episode of moderate anemia that has lasted for more than 2 weeks"

42

VERIDIAN SMI

Typical medical-record data for blood counts

- 3 Feb HCT 38.3
- 10 March HCT 35.1
- 17 March HCT 33.4
- 21 March HCT 34.2
- 6 April HCT 28.1
- 10 April HCT 27.8
- 18 April HCT 29.5
- 28 April HCT 33.4
- 8 May HCT 35.5
- 14 May HCT 33.1
- 22 May HCT 29.5
- 26 May HCT 27.6
- 8 June HCT 30.5
- 14 June HCT 35.5

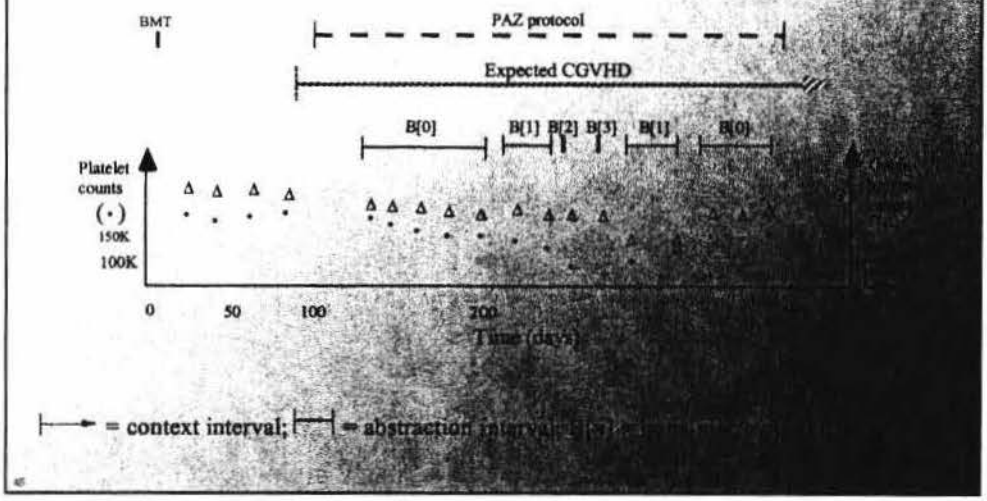
VERIDIAN SMI

Automated surveillance requires

- **The ability to process**
  - Numeric data
  - Qualitative data
  - Trends in data
  - Patterns in data
- **Knowledge of**
  - The data and how to interpret them
  - The implications of external events for interpreting the data

VERIDIAN SMI

Temporal Abstraction: Data from Bone-Marrow Transplantation



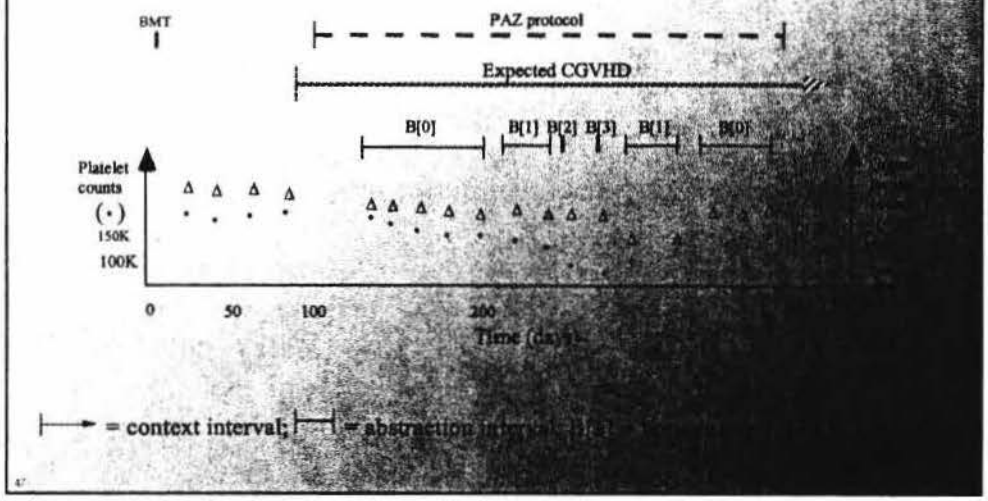
VERIDIAN SMI

The Knowledge-Based Temporal-Abstraction Method

- **Takes as input primary data and previously determined abstractions of those data**
- **Generates as output further abstractions of the input**
- **Requires an explicit knowledge base of**
  - The kinds of data that may be entered into the system and the properties of the data
  - How different contexts may affect data interpretation

VERIDIAN SMI

Temporal Abstraction: Data from Bone-Marrow Transplantation



VERIDIAN SMI

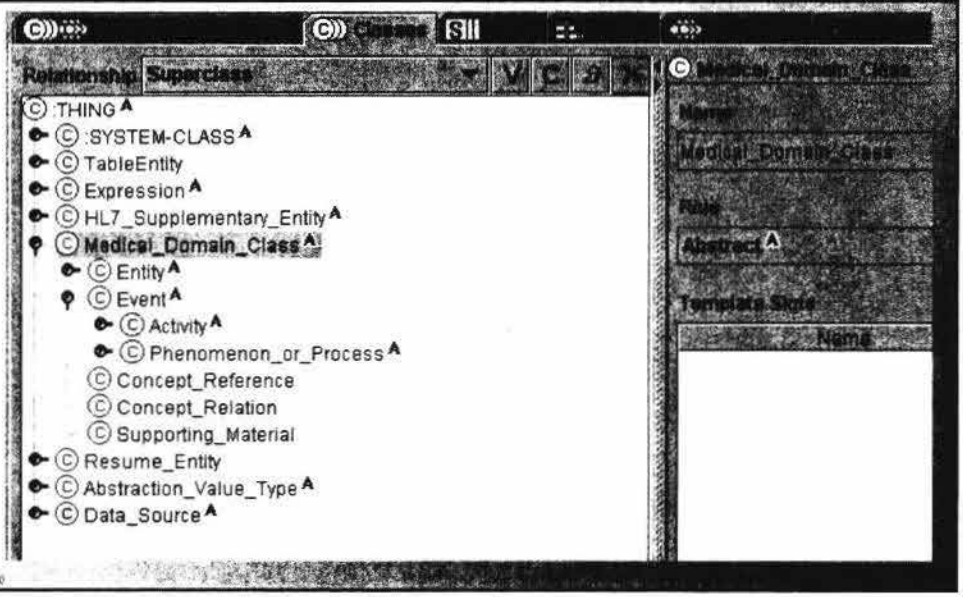
Types of temporal abstractions

- State abstractions (e.g., low, high)
- Gradient abstractions (e.g., increasing)
- Rate Abstractions (e.g., slow, fast)
- Pattern Abstractions
  - Linear patterns (e.g., crescendo)
  - Periodic patterns (e.g., diurnal)

Knowledge required for temporal abstraction

- **Structural knowledge**
  (e.g., definitional relationships among lab tests and clinical states)
- **Classification knowledge**
  (e.g., how numeric values map into qualitative ranges)
- **Temporal-semantic knowledge**
  (e.g., whether intervals are concatenable or downward hereditary)
- **Temporal-dynamic knowledge**
  (e.g., minimal values for a significant change, functions for persistence of a value over time)

Protégé-2000 provides tool support for building large KBs



VERIDIAN SMI

RÉSUMÉ (Yuval Shahar, 1992)

- Is a computer system that performs knowledge-based temporal abstractions on data of individual patients
- Used to identify automatically patterns in patient situations that predicate different courses of action required by clinical practice guidelines

VERIDIAN SMI

Past applications of the RESUME system for temporal abstraction

- Identification of patient situations that predicate therapy adjustments in
  - Clinical-practice guidelines
  - Clinical-trial protocols
- Monitoring glycemic control in diabetes
- Identification of abnormalities in children's growth

VERIDIAN SMI

We need knowledge-based temporal abstraction that can

- **Integrate data from multiple patients to yield population-based results**
  - Abstracting primary data *within* patient data sets
  - Abstracting population data *across* patient data sets
- **Adjust data-interpretation contexts in light of**
  - External world events
  - The presence of relevant data (or abstractions) known to the system
  - Tentative conclusions reached by other data-surveillance systems working in parallel
- **Abstract over space as well as time**
- **Execute quickly on huge data sets, taking advantage of modern supercomputer architectures**

53

VERIDIAN SMI

VERIDIAN SMI

Goals for the Stanford team

- Engineering of a robust, knowledge-based spatio-temporal abstraction system that can run on massively parallel supercomputers
- Design of abstraction methods that can operate within patient data sets and across patient data sets
- Design and construction of repositories that contain necessary temporal abstraction knowledge
- Acquisition and warehousing of data from San Francisco and Palo Alto
- Application of knowledge-based temporal abstraction to actual data to identify patterns in populations
- Experimentation in conjunction with Veridian to assess relative strengths and weaknesses of knowledge-based and statistical components of the overall architecture

58

VERIDIAN SMI

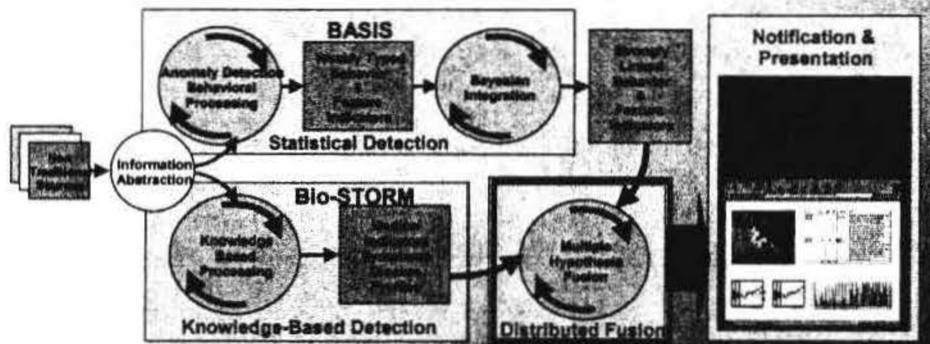
The Knowledge-Based Approach

- Allows all relevant clinical knowledge to be encoded directly in the computer in an explicit manner
- Incorporates qualitative data, contextual data, and explicit relationships among data sources into the problem solving
- May permit completely automated surveillance of clinical records and other data sources
- May complement purely statistical data monitoring in significant ways

56

Multiple Hypothesis Fusion Unifies Detection Processes

VERIDIAN SMI



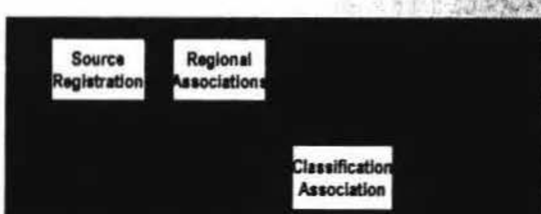
57

Multiple Hypothesis Fusion

VERIDIAN SMI

- Other Bayesian Knowledge Based Taxonomies cannot be directly linked
- Multiple Hypothesis Fusion correlates disparate Bayesian taxonomies to build disease tracking hypotheses
  - Tree structure builds representative hypotheses and tracks development of disease indicators
  - Disease state parameters
    - Kinematic state spread
    - Feature state contribution

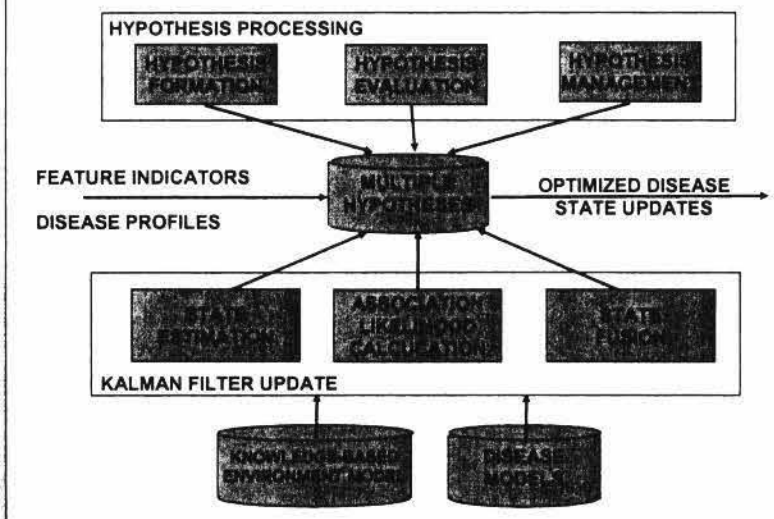
- Kinematic state
  - Point Source Locations (x,y)
  - Area Extent
  - Area Growth Rate
  - Member Set (Total # of members)
  - Member Density wrt Population Density
  - Member Growth Rate
  - Time Decay Factor
  - Corresponding Coverage Error
- Feature state (symptom contribution)
  - Average Duration/ [illegible]
  - Contagion Rate
  - Mobility Impact
  - [illegible]
  - [illegible]
- [illegible]
- [illegible]



58

VERIDIAN SMI

Information Fusion Depends on Tracking Multiple Hypotheses



59

VERIDIAN SMI

Multiple Hypothesis Fusion State Space Representation

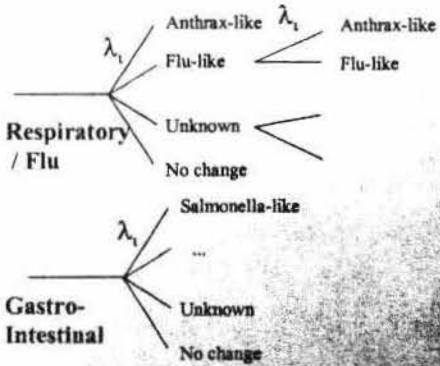
- Multiple hypothesis provide possible simultaneous solutions that can be evaluated over time
- Abstractions may require common data sources
  - Conditional relations (eg, position and velocity)
  - Implies common attributes must be placed in a state space approach to calculate conditionalities
  - Provides an effective means to measure commonality effects and errors
- Accommodates distance and time abstractions
- Handles continuous / discrete measurements
- Accommodates asynchronous update rates
- Provides for out-of-order measurements

60

VERIDIAN SMI

Multiple Hypotheses Example

Track Syndrome Hypotheses



Botulism-like
Rash with Fever
Meningitis-like
Unexpected Death

- **Multiple hypotheses allow rapid evaluation of disease relevance**
  - e.g. disease model discriminates between anthrax and influenza
- **Likelihoods quantitatively contribute to downselect of disease hypotheses**
- **Kalman Filter for each hypothesis**
  - Dynamic models [illegible] behavior for each [illegible]
  - Error performance [illegible] supported hypotheses
  - [illegible] over [illegible]

VERIDIAN SMI

Kalman Filtering Supports Multiple Hypothesis Fusion

- **Provides natural method of dynamic estimation (temporal) within geographical constraints (spatial)**
- **Takes advantage of developing models**
  - Can adapt to changing conditions/observations
  - Easy to add models as developed
- **Takes advantage of prior distribution of parameters**
- **Can incorporate outside influences on data sources**
- **Simpler than autoregressive moving average (ARMA) methods**
- **Able to handle incomplete datasets**
- **Models can be evaluated and altered over [illegible]**

VERIDIAN SMI

Kalman Filtering Performance

- **Features affecting KF performance**
  - Timing error – Reporting lags
  - Measurement error – Underreporting biases
  - Outliers – Misdiagnoses / Incorrect data recordings / Transmission
- **Kalman filter performs differently for each syndrome hypothesis and each disease type within a syndrome hypothesis**
- **Sample -- Respiratory - Flu (Representative Parameters)**
  - Time period of report process
  - Number of cases reported at time k
  - Number of absentees within region
  - Regional boundaries (Affects error covariance)
  - Known historical data (Model estimation)
  - Time variations of parameters outside data [illegible]

VERIDIAN SMI

Kalman Filter Sample Continued

- **System Model**
  - Transition matrix $F(k)$ determines how state variables evolve over time (k)
  - Incorporates model uncertainty $v(k)$
- **Observation model**
  - Measurement transformation matrix $H(k)$ relates measurement data $z(k)$ and state variables $x(k)$
  - Incorporates measurement uncertainty $w(k)$
- **Process model for steady state allows detectability**
  - Four allowable states (Markov States)
    - S1 – Normal state (no aberrations)
    - S2 – Sudden level changes (sustained)
    - S3 – Change in slope (Increase in reported rates)
    - S4 – Outliers

State Model

$$x(k+1) = F(k)x(k) + v(k)$$

Observation Model

$$z(k) = H(k)x(k) + w(k)$$

VERIDIAN SMI

Kalman Filter Sample, continued

- **State model**

$$x(k+1) = F(k)x(k) + v(k)$$

$$\begin{pmatrix} \mu(t+1) \\ \beta(t+1) \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \mu(t) \\ \beta(t) \end{pmatrix} + v(k)$$

  - Changes in level

$$\mu(t+1) = \mu(t) + \beta(t) + v_1(k)$$

  - Changes in slope

$$\beta(t+1) = \beta(t) + v_2(k)$$

- **Observation model**

$$z(k) = H(k)x(k) + w(k)$$

$$\begin{pmatrix} \text{Normalized} \\ \text{Case} \\ \text{Numbers} \end{pmatrix} = \begin{pmatrix} 1 & 0 \end{pmatrix} \begin{pmatrix} \mu(t) \\ \beta(t) \end{pmatrix} + \varepsilon(k)$$

  - Outliers

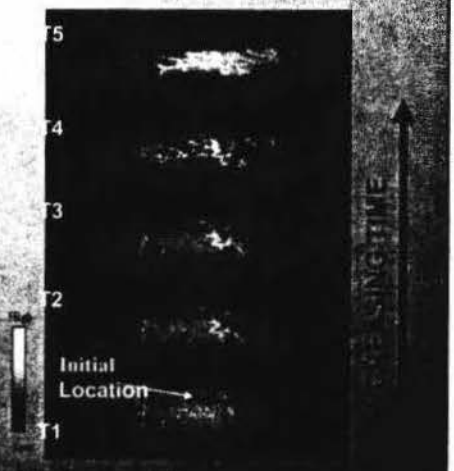
$$\mu(t+1) = \mu(t) + \varepsilon$$

- **Transition Probabilities determines state change**
  - s1 => s2 => s3 => s3

VERIDIAN SMI

Spatial Context Provides Normalization

- **Spatial context allows modification of uncertainty regions for Kalman filters**
- **Conditions**
  - Regional density of susceptible individuals
  - Jurisdictional boundaries
  - Military population densities
  - School boundaries
  - Natural barriers (e.g. rivers)
  - Movement arteries (roads, bridges...)



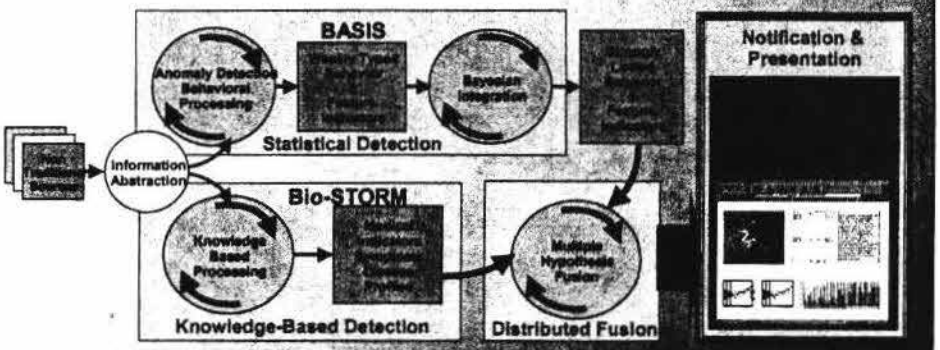
VERIDIAN SMI

Goals of Fusion Process

- **Spatial-Temporal Fusion**
  - Develop and evaluate non-traditional data sources within spatial-temporal fusion sub-system
  - Create the knowledge-based taxonomies to link correlated non-traditional data sources (minimize double counting)
  - Develop probability density surfaces to link non-traditional sources and spatial context (develop uncertainty regions)
- **Multiple Hypothesis Fusion**
  - Design and construct a rigorous multiple hypothesis framework to [illegible] results with Bio-STORM results
  - Develop syndrome-to-disease hypothesis tree structure
  - Create corresponding Kalman filter modules consistent with [illegible] to detect and analyze datasets
  - Evaluate disease hypotheses to assess relevance
- **System Test**
  - Experimentally test and evaluate system performance
  - Feed results back into system design / [illegible]

VERIDIAN SMI

Notification and Presentation



VERIDIAN SMI

Goals of Notification

The notification system will:

- Provide immediate and guaranteed alerting
- List favored hypotheses, ranked by confidence
- Identify "alerting" data sources, ranked by confidence
- Integrate into the military communication network
- Provide fail-safe capability (backup responders)
- Support secure, web-based reports

VERIDIAN SMI

Goals of the Presentation System

The presentation system will:

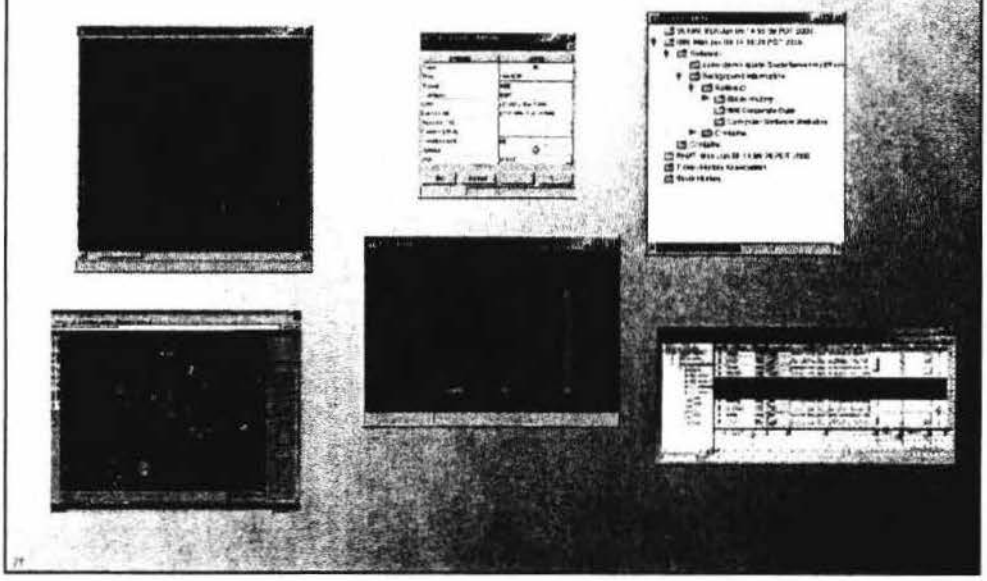
- Provide spatio-temporal display
- Provide a spatial focus with temporal dimension
- Provide a temporal focus with spatial dimension
- Decouple data sources from software to allow fast integration
- Allow plug-and-play of of disparate data sources

COTS tools will be used including:

- Polexis XIS (Display and data export capabilities)
- ESRI's ArcIMS web-based map services

VERIDIAN SMI

Presentation Via Familiar Interfaces



VERIDIAN SMI

Web-based Map Presentation

VERIDIAN SMI

A Bio-Surveillance System for Advanced Medical Readiness: *Ethical Issues*

Presenter: Mildred Cho, PhD
Stanford University Center for Biomedical Ethics

VERIDIAN SMI

Overview

- Ethical principles in research
- Regulatory parameters
- Specific ethical and regulatory issues in bio-surveillance and medical informatics research

VERIDIAN SMI

Ethical Principles

- **Privacy and confidentiality**
  - ♦ As values, *per se*
  - ♦ As evaluated in terms of harms and benefits
    - ♦ Identifiability of, and risks to, individuals
    - ♦ Identifiability of, and risks to, socially-defined groups
- **Voluntariness**
- **Informed consent**

VERIDIAN SMI

Regulatory parameters

- **Is it research?**
  - ◆ Research = a systematic study whose intent is to provide generalizable knowledge
- **Is the research on human subjects?**
  - ◆ Human subjects research = research on a living individual and
    - Data obtained through interaction or intervention
    - Data are identifiable (not completely anonymous)
  - ◆ If human subjects research, need Institutional Review Board (IRB) review

Regulatory parameters

- **IRB review**
  - Mandated by the "Common Rule" (32 CFR 219)
  - IRBs determine:
    - Whether research is exempt from review
    - Whether research qualifies for "expedited" review
    - Whether informed consent of human subjects is necessary
  - Common Rule applies to:
    - Research funded by, among others, DoD, DoE, VA, DHHS
    - Includes research conducted by civilian federal employees and military personnel

77

Regulatory parameters

- **If research is not exempt but more than "minimal risk", informed consent may be necessary**
  - Minimal risk = probability and magnitude of risks encountered in daily life or routine exams or tests
- **Areas of ambiguity in regulation:**
  - What is "identifiable"? (anonymous, coded, directly identifiable, otherwise identifiable?)
  - What is "minimal risk"?

78

Specific ethical issues

- **When is informed consent necessary?**
  - 10 USC 980: DoD-funded research on human subjects may not be conducted unless:
    - Advance informed consent of the subject obtained, or
    - If research intended to benefit subject, informed consent obtained from subject or legal representative
    - However, activities such as use of drugs not approved by FDA does not necessarily constitute research

Specific ethical issues

- **When is IRB review necessary?**
  - If human subjects are involved, except if:
    - Educational tests, surveys, observations of public behavior, unless data are linkable to subjects or could be damaging to subjects
    - Existing data or specimens if publicly available or if subjects cannot be identified
    - Public benefit or service programs

VERIDIAN SMI

Specific ethical issues

- **Military vs civilian populations**
  - No systematic difference between military and civilian populations in application of research regulations
  - No systematic difference between military and civilian researchers in application of research regulations
- **What kinds of data can be collected?**
  - "Private"=information gathered in a setting where an individual can reasonably expect that no observation or recording is taking place, or can reasonably expect that will not be made public (e.g., medical record)
  - How identifiable are individuals based on:
    - Individual data fields
    - Combined data fields

VERIDIAN SMI

Syndrome Surveillance in the Acute Care Setting:

The Intersection of Bio- Surveillance and Public Health

Judith C. Brillman MD
Associate Professor
Department of Emergency Medicine
University of New Mexico SOM

10/06/01 UNM ED
Pt NC, 2.5 year old female

- P: 225, thready
- BP: unobtainable
- RR: 30
- T: 40 C
- Sat on 100% Oxygen: 85%
- Unresponsive
- Floppy
- Skin mottled, no rash
- Neck supple

History

- 1 day of Fever, N/V
- Decreased mental status this morning
- Previously healthy
- IMM UTD

This Child is in Shock: Why?

- This is my corner of the world
- I know the symptoms and physical signs
- I never know the diagnosis or organism
- HELP ME.....!
  - Make a better empiric decision

VERIDIAN SMI

Surveillance

- Diagnosis vs. syndrome based
- Which is better for detecting unanticipated events?
  - Bio-terrorism
  - Emerging diseases
  - Naturally occurring epidemics

VERIDIAN SMI

Traditional Surveillance: Diagnosis Based

- Diseases to be surveyed must be defined in advance
  - Influenza
- Diagnostic criteria defined in advance
  - Positive z
- Reporting is incomplete

Limitations of Traditional Diagnostic Surveillance

Confessions of an ED Doc

Making a diagnosis results in lost information

- Limited ability to identify new diseases
  - "It looked like the flu"
- Limited ability to identify uncommon diseases
  - "I may have seen it, but I haven't diagnosed it."

Limitations of Traditional Diagnostic Surveillance

- Lab specimen must obtained:
  - "You can't find what you're not looking for"
- Microbiologic tests are slow:
  - "No wonder she died"
- Delays in identification of epidemiologic variance:
  - "Really you don't get a rash in influenza?"
- Inconsistent reporting:
  - "Why test for influenza, I'm not going to do anything about it"

Code Based Diagnostic Surveillance

- Monitor coded discharge diagnoses or chief complaints
  - High risk events
  - Acute care settings
    - ED's
    - Clinics
    - EMS
- Not lab or report based
- May miss syndromes that do not directly map into [illegible] code

Limitations of Non-Traditional Diagnostic Surveillance

- Limited ability to identify new diseases
  - Depends on "old" diagnoses
- Delays in identification of epidemiologic variance
- Limited ability to identify uncommon diseases
- Inaccuracy of chief complaints:
  - "I have the flu"

VERIDIAN SMI

Syndrome Surveillance

- Capture raw symptoms, signs & lab data
  - Prior to (mis) diagnosis
- Detect sentinel findings (rule based)
  - Wide mediastinum
  - "Chicken pox" in a previously infected host
- Anomaly detection: 2 tier system
  - Detect abnormal pattern of illness
  - Public health investigation for diagnosis
- Intention free: I don't have to look for anything in particular
- Early: No waiting for micro results

VERIDIAN SMI

Syndrome Surveillance: Limitations

- Must define symptoms in advance
  - What are most effective symptoms to survey?
- How will the data be collected?
  - New data, not pre-existing
- How can the data be analyzed?
- Requires comparative baseline data

VERIDIAN SMI

Syndrome Surveillance: Paradigm shift

- Paradigm shift: From diagnoses to symptoms and findings
  - From passive to active surveillance
- Venues appropriate for Syndromic Surveillance?
- Any venue in which diagnoses aren't known
  - Disaster situation
  - Military Operation
  - Pre-Hospital
  - Emergency Department

92

VERIDIAN SMI

B-SAFER
Bio-Chemical Surveillance, Analysis, Feedback, Evaluation & Response

- Working at the intersection of Bio-Terrorism Surveillance and Public Health
- Successor to RSVP
- Surveillance of:
  - Likely bio-terrorism agents
    - Infectious
    - Toxic
  - Toxidromes
  - Likely epidemic agents
  - Related to previous emerging diseases
  - Uncommon but dangerous entities
- Based on what data we need, rather than what we have

94

B-SAFER Partners in Crime

- **COLLABORATORS**
- Alan Zelicoff - SNL
- Sandra Zink, Dave Forslund - LANL
- New Mexico Department of Health
  - Gary Simpson
  - Edith Umland
  - Stuart Castle
  - Kevin Bersell

- **FUNDING**
- DOE -NN-CBNP
- New Mexico Department of Health
- LANL LDRD
- CDC

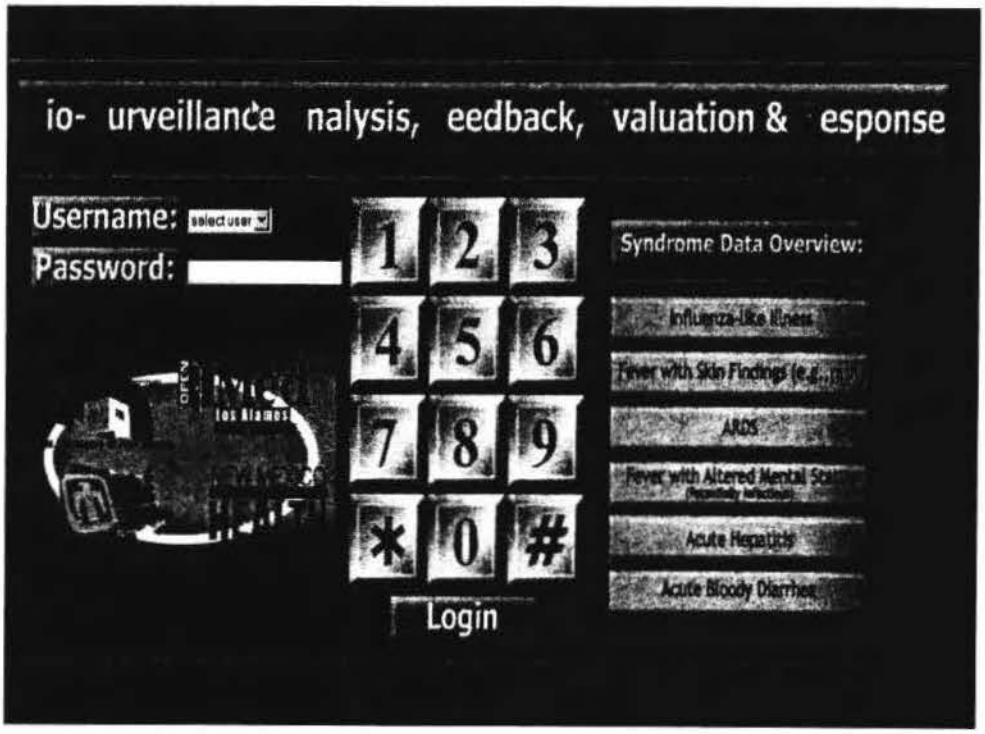
Provider Benefits: Feedback

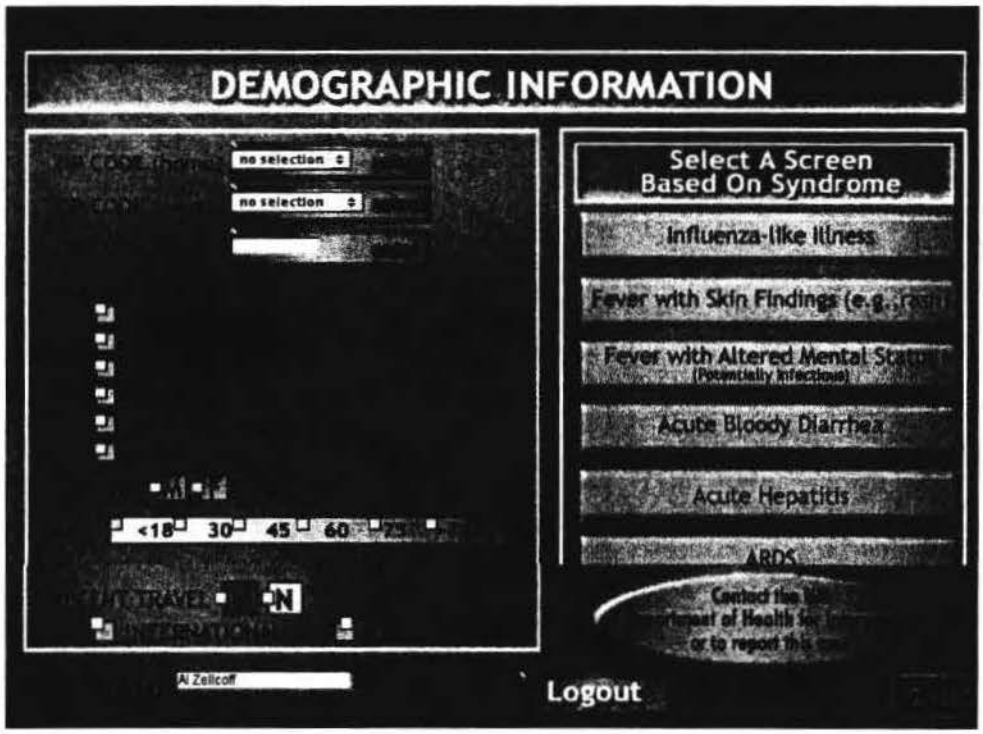
- "Near" real-time
- Syndrome occurrence for previous 30 days graphed
  - What's going around
- Map of events by zip code
- Text-based from NM DOH
- *Cultures: requests and results*
- *Differential diagnosis*
- *Disease facts*
- Links
  - NM Department of Health, CDC
  - *Gideon/Others*

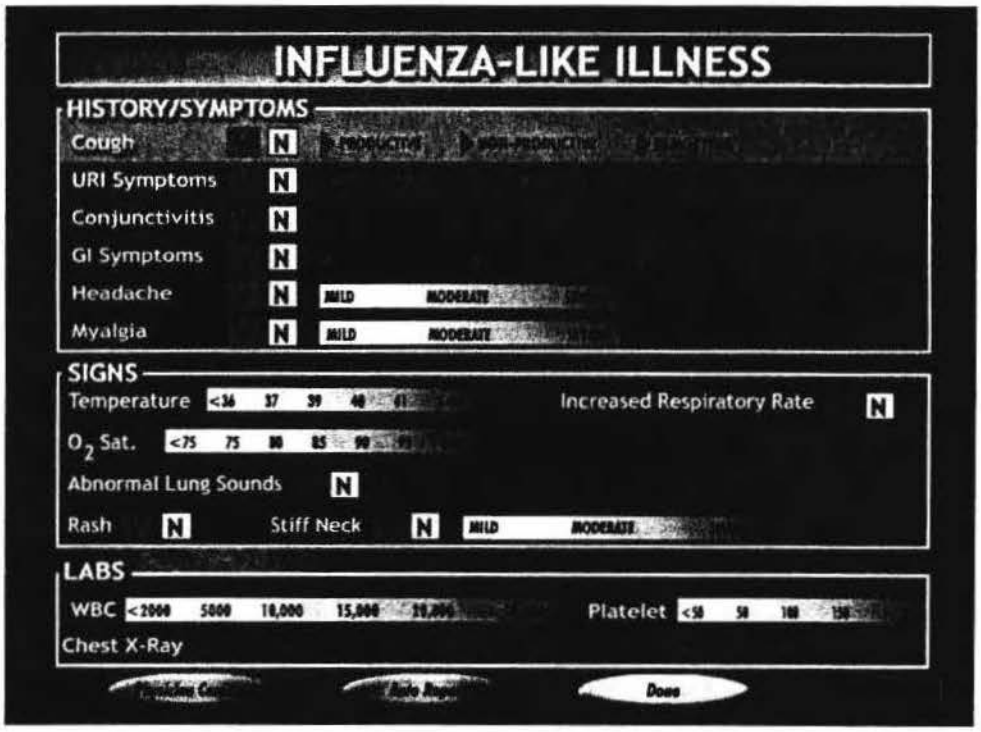
VERIDIAN SMI

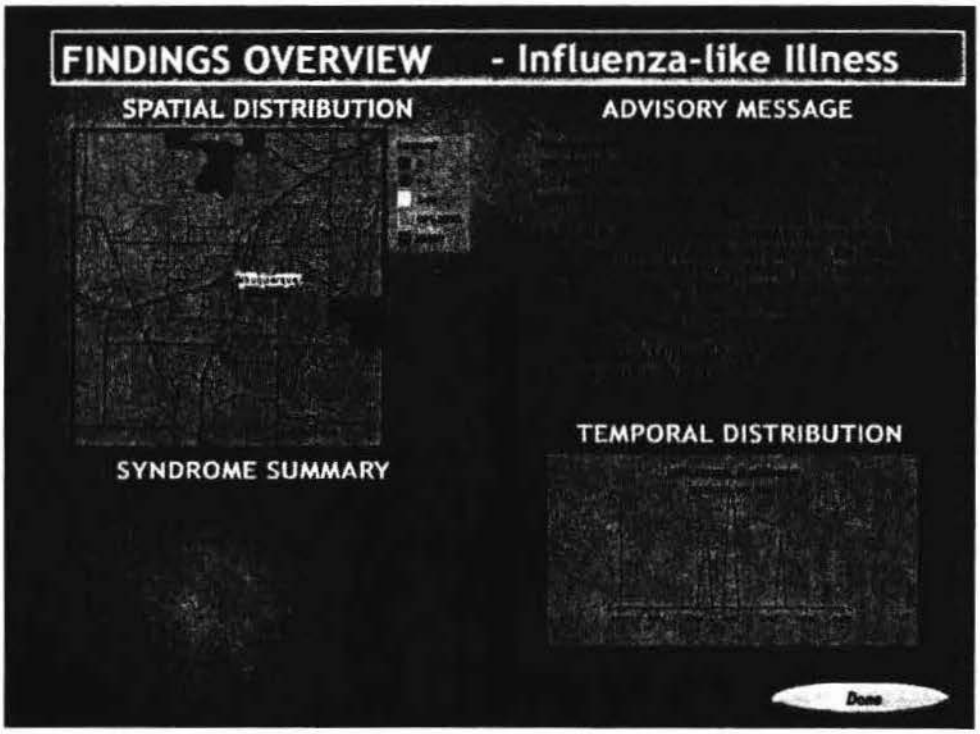
Reporting

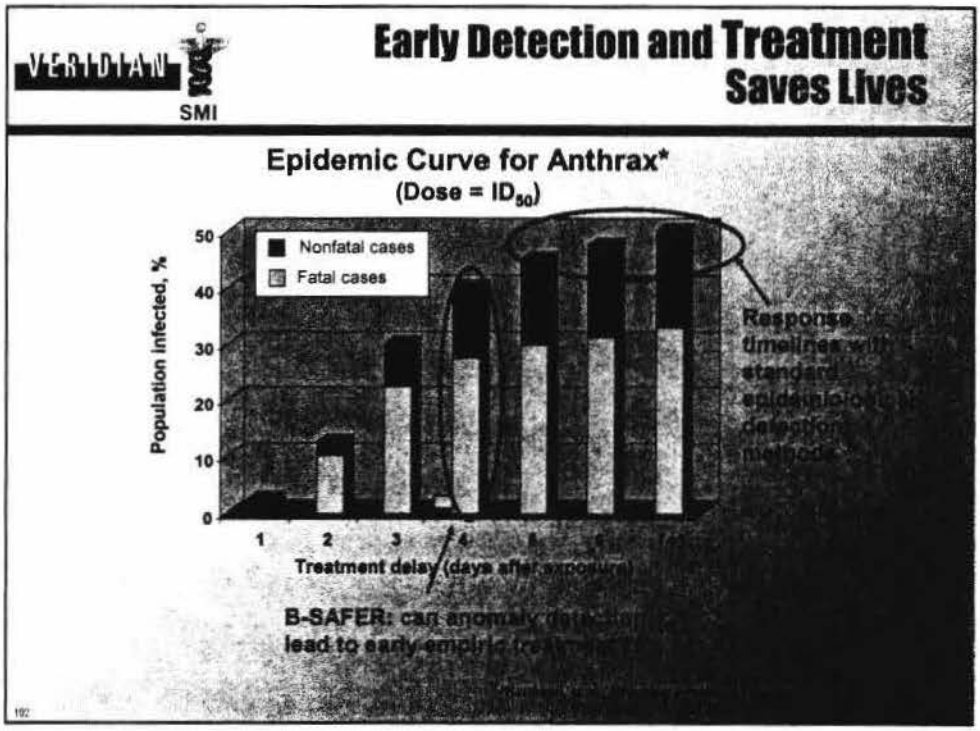
- Web based
- To: Public Health Department
  - *Other Entities: CDC, FEMA, etc.*
- "Near" real-time data flow
- Rule based automatic reports
  - *to E-mail pager*
- ***Reports based on anomaly detection***




DEMOGRAPHIC INFORMATION
no selection
no selection
<18
30
45
60
N
Select A Screen Based On Syndrome
Influenza-like Illness
Fever with Skin Findings (e.g.,
Fever with Altered Mental Status
(Potentially Infectious)
Acute Bloody Diarrhea
Acute Hepatitis
ARDS
Al Zelicoff
Logout


INFLUENZA-LIKE ILLNESS
HISTORY/SYMPTOMS
Cough
N
URI Symptoms
N
Conjunctivitis
N
GI Symptoms
N
Headache
N
MILD
MODERATE
Myalgia
N
MILD
MODERATE
SIGNS
Temperature
<36
37
39
40
Increased Respiratory Rate
N
O2 Sat.
<75
75
80
85
90
Abnormal Lung Sounds
N
Rash
N
Stiff Neck
N
MILD
MODERATE
LABS
WBC
<2000
5000
10,000
15,000
Platelet
<50
50
100
150
Chest X-Ray
Done


FINDINGS OVERVIEW - Influenza-like Illness
SPATIAL DISTRIBUTION
ADVISORY MESSAGE
SYNDROME SUMMARY
TEMPORAL DISTRIBUTION
Done


VERIDIAN
SMI
Early Detection and Treatment Saves Lives
Epidemic Curve for Anthrax*
(Dose = ID50)
Nonfatal cases
Fatal cases
Population Infected, %
0
10
20
30
40
50
1
2
3
4
Treatment delay (days after exposure)
B-SAFER: can anomaly
lead to early
192

VERIDIAN SMI

SIMULATION AND DISEASE MODELS

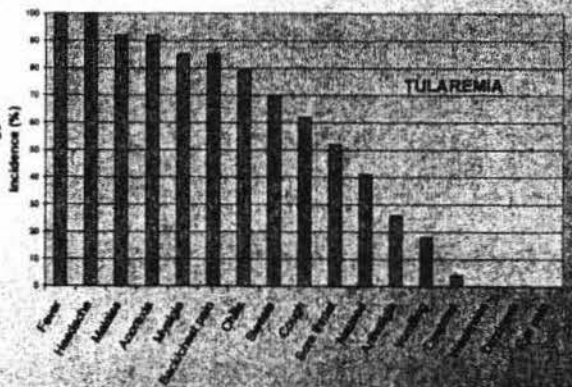
Dr. Gene McClellan
Veridian

VERIDIAN SMI

Build on Analysis of Real-World Data

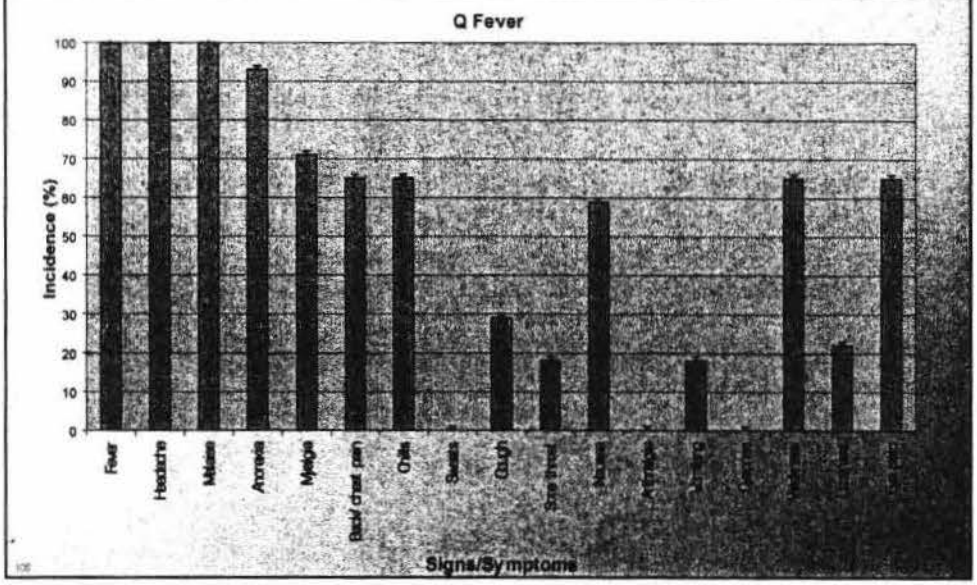
- Analysis of clinical pattern of diseases
- Incidence of signs and symptoms
- Supports population diagnosis

- Fort Detrick
- BW Vaccine trials
- Human data, '50s, '60s



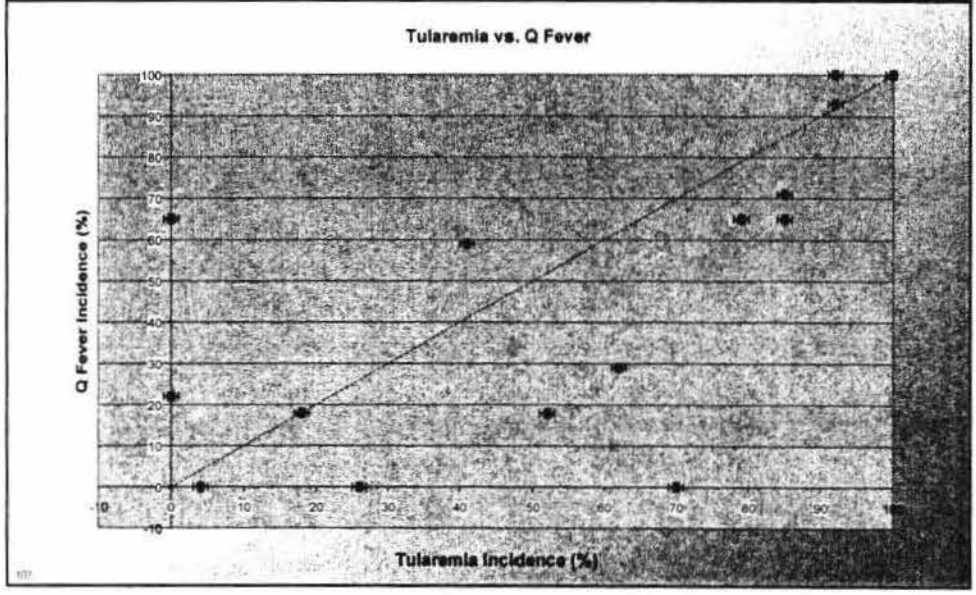
VERIDIAN SMI

Similar Data for Q Fever



VERIDIAN SMI

May Be Able to Differentiate Diseases within a Population



NATO NBC Medical Planning Guides

NATO Standarization Agreement (STANAG) 2476
Allied Medical Publication 8 (AMedP-8)
Medical Planning Guide for the Estimation of NBC Battle Casualties, Volume II (Biological)

- Representative threat agents and attacks
- Casualty tables providing time-phased morbidity and mortality
- Based on dose-dependent descriptions of human response
- Ratified February 2001

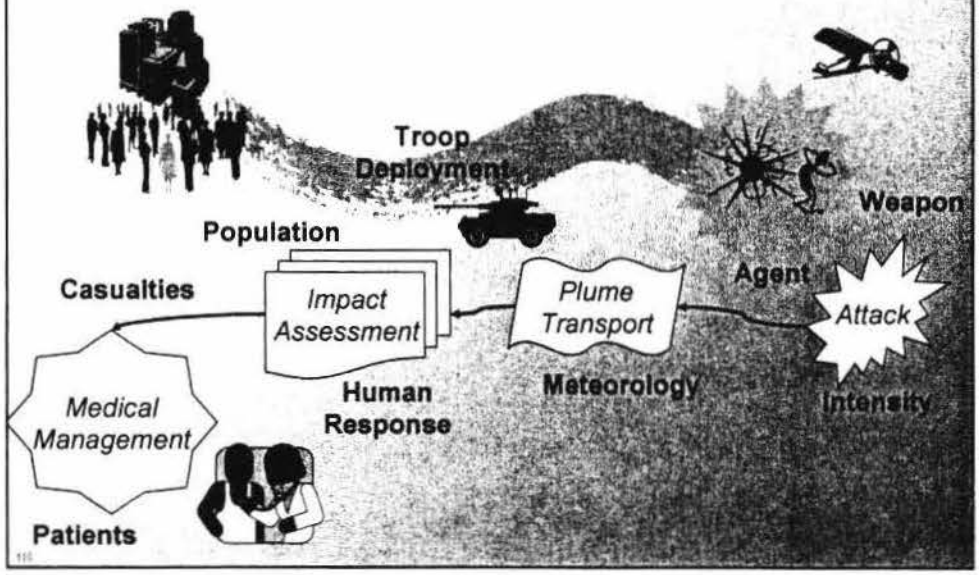
NATO Custodian:
Office of the Surgeon General, U.S. Army
NBC Defense Staff Officer:
LTC Debra Schnelle, MSC

AMedP-8 Example Table

Status of Unit Personnel by Time Period
Heavy Brigade, Forward Maneuver, Movement to Contact (4,042 Personnel)
Anthrax, 1 km Ground Line

Effectiveness Categories		Light attack									
		D2	D3	D4	D5	D6	D7	D10	D14	D21	D28
Casualties	Total	0	144	814	1078	930	874	373	345	344	0
Fatalities	Total	0	17	255	896	1457	1815	2126	2154	2155	2155
Capable	Total	4042	3881	2972	2068	1858	1683	1543	1543	1543	1543
Ill but capable	>25%-< 100%	49	1002	1150	493	109	10	0	0	0	0
Well	=100%	3993	2879	1822	1575	1848	1543	1543	1543	1543	1543
Convalescent		0	0	0	0	0	0	0	0	0	344
		Moderate attack									
		D2	D3	D4	D5	D6	D7	D10	D14	D21	D28
Casualties	Total	4	291	1449	1846	1193	844	532	487	496	0
Fatalities	Total	0	33	461	1470	2288	2875	2997	3038	3055	3055
Capable	Total	4038	3718	2132	917	681	529	513	513	513	513
Ill but capable	>25%-< 100%	113	1885	1480	400	87	10	0	0	0	0
Well	=100%	3925	1833	652	517	514	513	513	513	513	513
Convalescent		0	0	0	0	0	0	0	0	0	499
		Heavy attack									
		D2	D3	D4	D5	D6	D7	D10	D14	D21	D28
Casualties	Total	5	532	1681	1881	1148	778	474	436	434	0
Fatalities	Total	0	53	713	1829	2909	3016	3319	3359	3359	3359
Capable	Total	4037	3457	1648	536	287	250	249	249	249	249
Ill but capable	>25%-< 100%	230	2242	1335	285	38	1	0	0	0	0
Well	=100%	3807	1215	313	251	249	249	249	249	249	249
Convalescent		0	0	0	0	0	0	0	0	0	434

Overview of Casualty Assessment



Agents/Illnesses for AMedP-8 (Bio)

- **Ratification Draft (7/97 to 7/99)**
  - Tularemia, Q fever, SEB intoxication
  - Anthrax, Plague, VEE, Botulinum intoxication
- **1999/2000**
  - Smallpox
  - Brucellosis
  - Glanders
- **In Progress - - 2000/2001**
  - Hemorrhagic fever (Ebola, Marburg, ...)
  - Ricin intoxication
  - Mycotoxin intoxication

VERIDIAN SMI

Data Sources for Ratification Draft

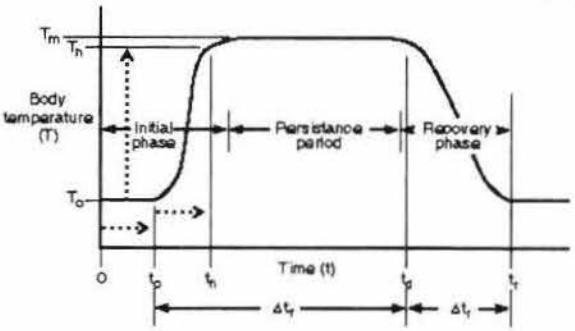
- **Experimental (human) data available**
  - Tularemia
  - Q fever
  - SEB intoxication

- **Clinical experience (natural/accidental exposures)**
  - Anthrax
  - Plague
  - VEE
  - Botulinum intoxication

- **Established modeling approach for subsequent work**

VERIDIAN SMI

Tularemia body temperature profile Use three statistical distributions -

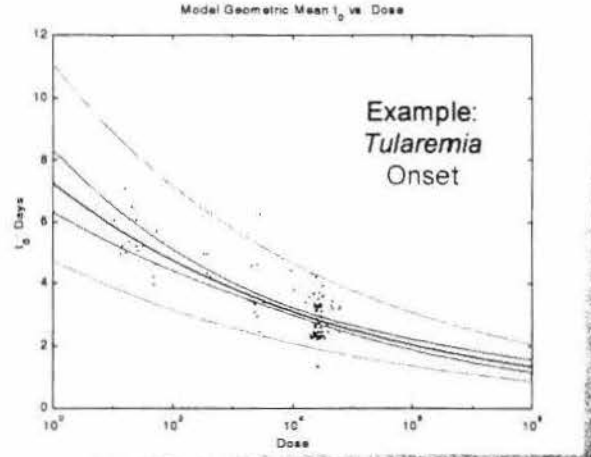


T_o = Normal body temperature (98.6°F)
T_h = Near-maximum body temperature
T_m = Maximum body temperature

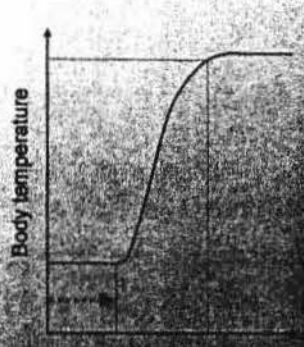
t_o = Temperature onset time
t_h = Near-maximum temperature time
t_d = Time temperature initiates decline
t_r = Time temperature returns to normal
Δt_f = Febrile period
Δt_r = Recovery period

VERIDIAN SMI

Human Data Supports Dose-Dependent Statistical Descriptions



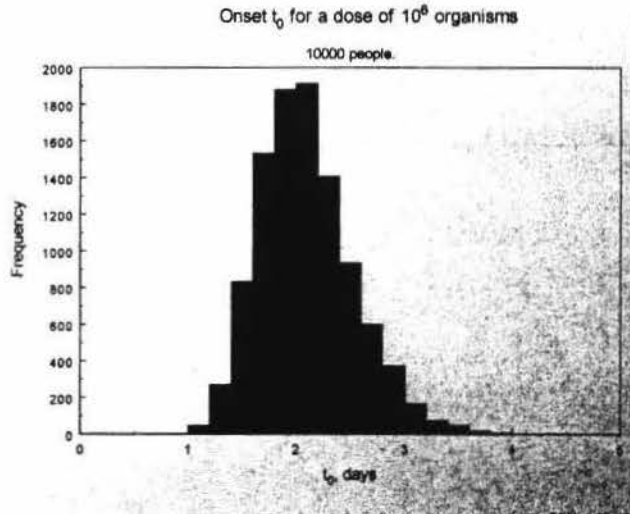
Vaccine trials from Fort Detrick archives



114

VERIDIAN SMI

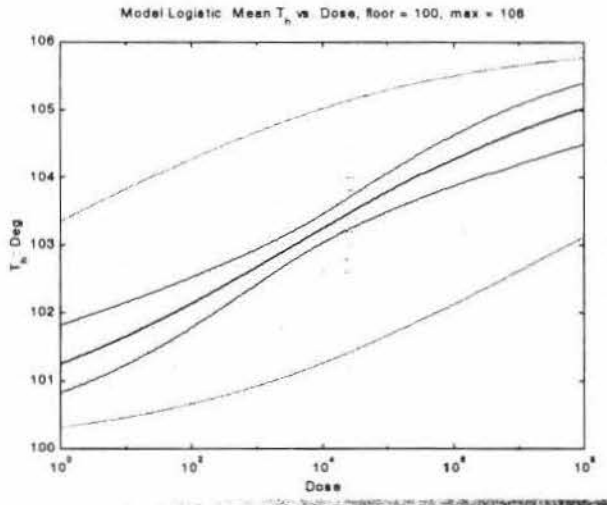
Onset Monte Carlo Simulation

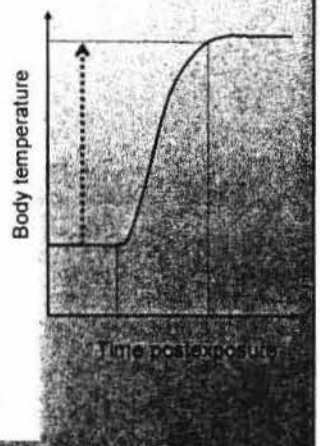


115

VERIDIAN SMI

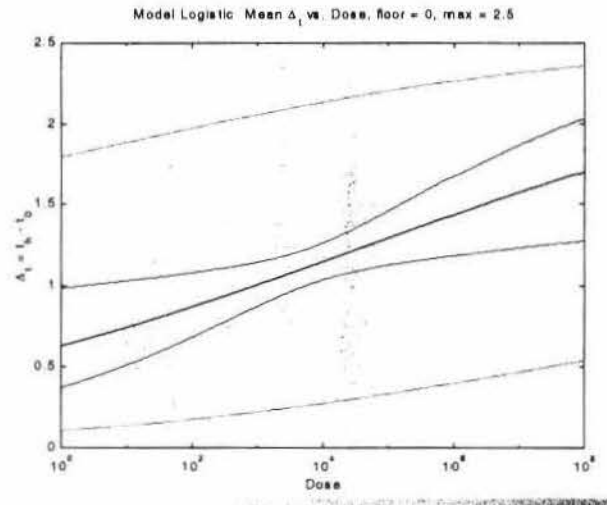
Fever Range is Limited; Use Logistic Model

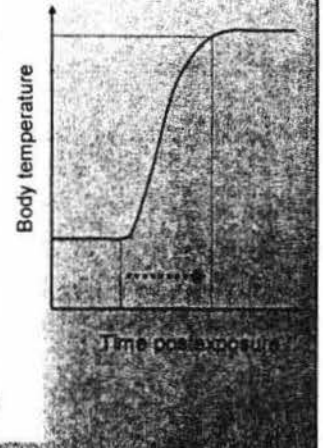


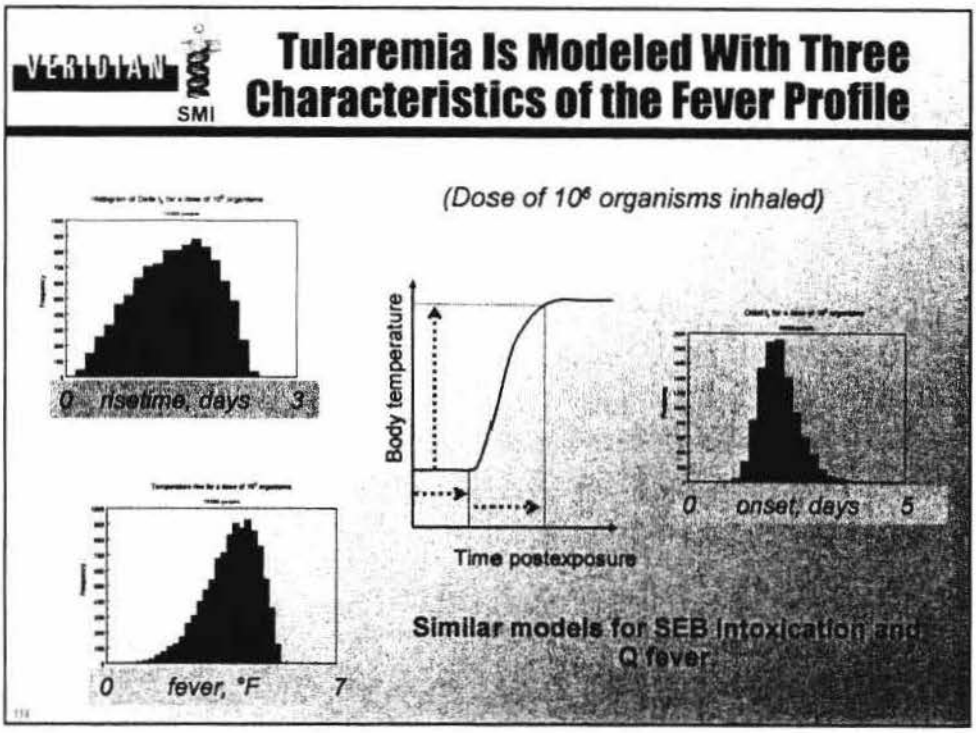


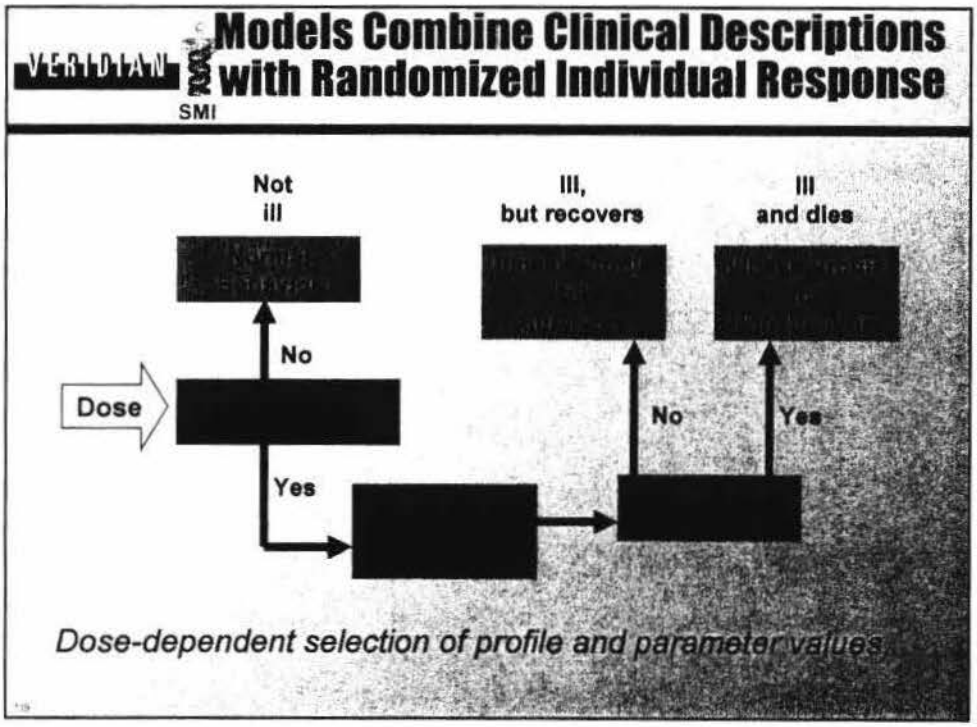
VERIDIAN SMI

Time-to-High-Fever ("Risetime"); Use Logistic Model, Also



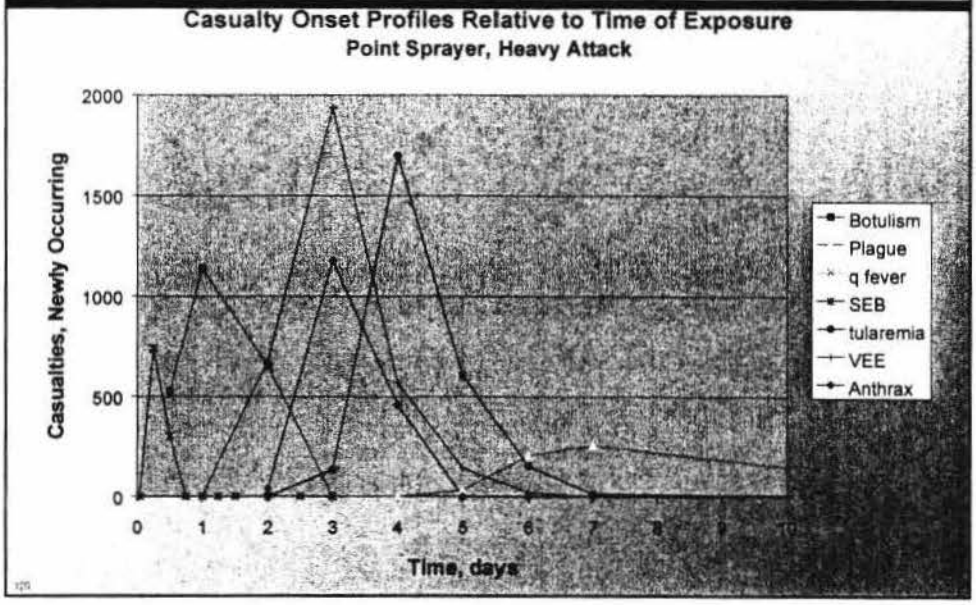



VERIDIAN
SMI
Tularemia Is Modeled With Three Characteristics of the Fever Profile
(Dose of 10^6 organisms inhaled)
0 risetime, days 3
Body temperature
Time postexposure
0 onset, days 5
0 fever, °F 7
Similar models for SEB intoxication and Q fever

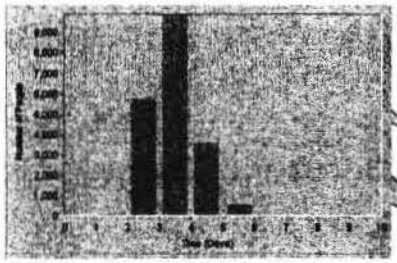

VERIDIAN
SMI
Models Combine Clinical Descriptions with Randomized Individual Response
Not ill
Ill, but recovers
Ill and dies
Dose
No
Yes
No
Yes
Dose-dependent selection of profile and parameter values

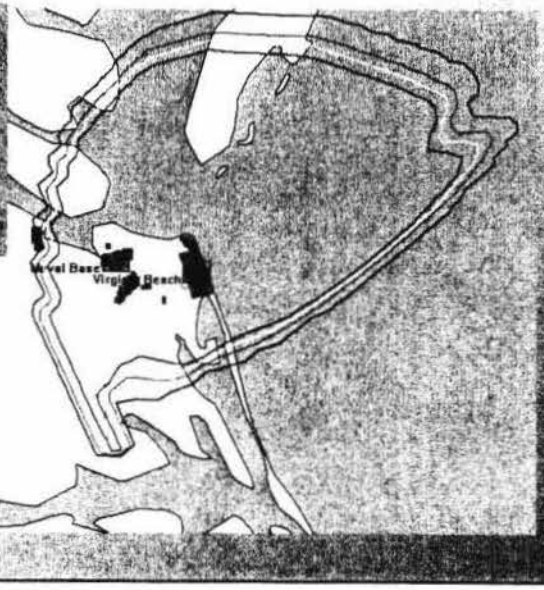
VERIDIAN SMI

Diseases Have Wide Variations in Onset After Attack



Casualty Estimation for Biological Attack





Sprayer Attack with
F. tularensis

VERIDIAN SMI

Evaluation

- **7 Measures of system performance [CDC, 2001]**
  1. **Simplicity** – develop parallel architecture in key areas, select most parsimonious solution
  2. **Flexibility** – maintain knowledge in ontologies, examine resources to transfer system between locations
  3. **Acceptability** – employ direct data feeds, monitor human effort required for reporting
  4. **Representativeness** – draw on multiple data sources for same time and area, identify 'gaps' in reporting between sources

122

VERIDIAN SMI

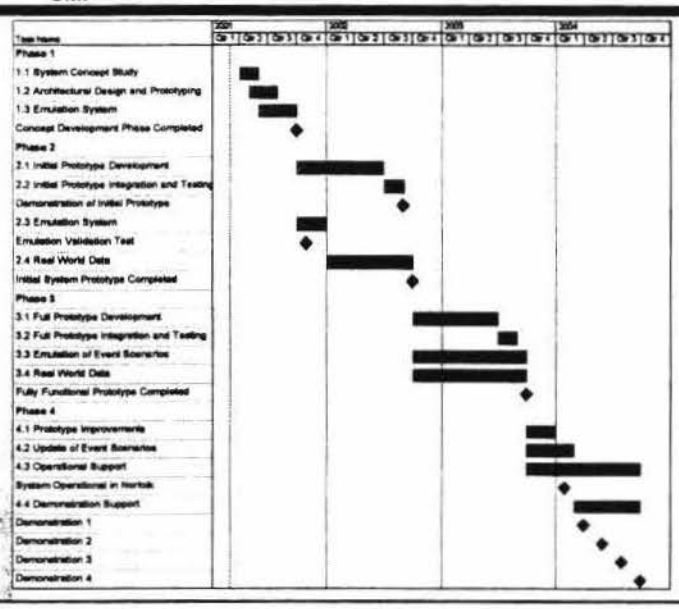
Evaluation, continued

- **Some measures require a 'gold standard'**
  5. **Sensitivity** (proportion of true cases or outbreaks detected)
  6. **Predictive value positive** (proportion of identified cases or outbreaks that are 'true')
  7. **Timeliness**
- **We will use data from known outbreaks with characteristics similar to bioterrorism syndromes**
  - onset of influenza season
  - other communicable disease outbreaks
- **We will use simulation data for additional testing**

123

VERIDIAN SMI

Bio-Surveillance System Project Schedule (original)



124

VERIDIAN SMI

Issues/Challenges

- **Obtaining Real-World Data**
  - Privacy Issues
    - Proper Consent
    - Proper Privacy Assurances or Accreditations
    - Paperwork
  - Logistics
    - Establishing POCs
    - Paperwork
    - Obtaining Real-Time Data Feeds
    - Technical infrastructure
- **Delayed Contract Start**

126

VERIDIAN SMI

Our Unique Contributions

- Innovative ideas for early sign/symptom detection
- Integration of statistical and knowledge-based processing throughout the architecture
- Early generalization of system using disparate geographic areas
  - San Francisco – ongoing relationship with public health system
  - Norfolk – initial focus on military data sources
- Specific technology for creating data abstractions both within and across patient records
- Use of context to guide data interpretation
- Systematic incorporation of bioethical principles

Scalable Biosurveillance Systems: Focus Areas for Technical Contributions

Mike Wagner MD PhD
Andrew Moore PhD
Kathleen Carley PhD

Biomedical Security Institute ("MPC" Corp.)
University of Pittsburgh and Carnegie Mellon University
Pittsburgh, PA
mmw@cbmi.upmc.edu

Project Team

LCE/CBMI
Wendy Chapman PhD
Greg Cooper MD PhD
Jeremy Espino MD
Doug Fridsma MD
Oleg Ivanov MD
Haobo Ma MD
Bob Olszewski PhD
Mark Roberts MD
Rich Tsui PhD
Mike Wagner MD PhD
Xiaoming Zeng MD

Heinz
Kathleen Carley PhD
George Duncan PhD
Rema Padman PhD

Robotics
Andrew Moore PhD
Jeff Schneider PhD

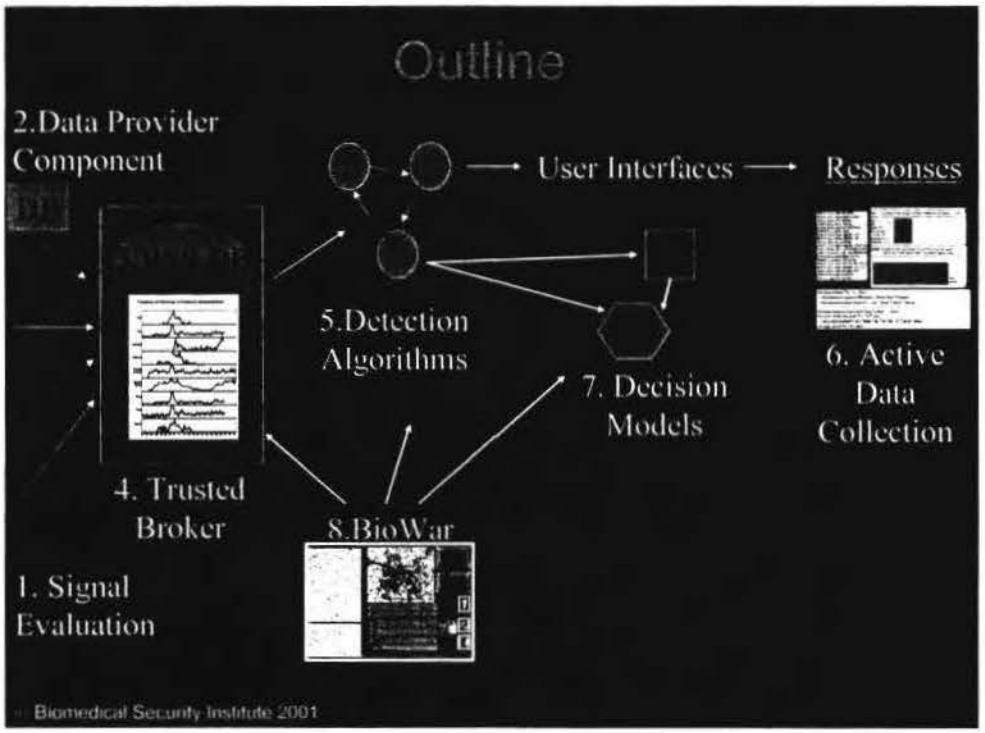
PSC
David Deerfield II PhD

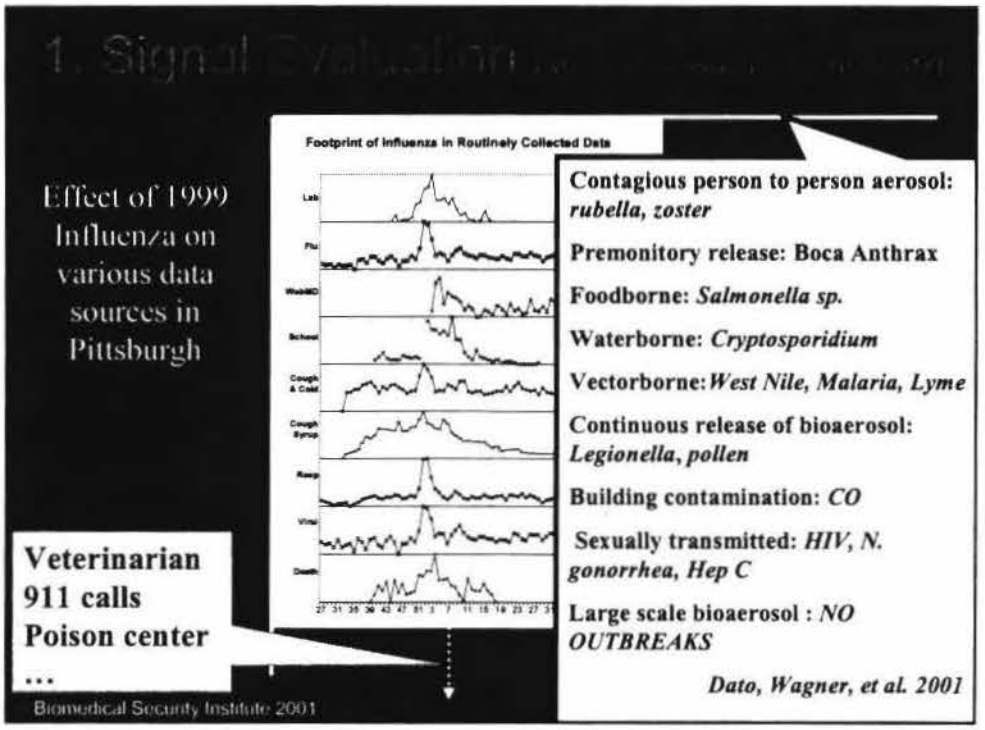
GSPH
Virginia Dato MD MPH
Ravi Sharma PhD

ACHD
Dave Piposzar MA

Pittsburgh Poison Center
Ed Krenzelok PharmD
Rita Mrvos RN

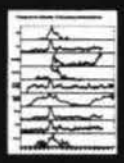
ED and Military WMD
Mike Allswede DO


Outline
2.Data Provider Component
User Interfaces
Responses
5.Detection Algorithms
7. Decision Models
6. Active Data Collection
4. Trusted Broker
8.BioWar
1. Signal Evaluation
Biomedical Security Institute 2001


1. Signal Evaluation
Effect of 1999 Influenza on various data sources in Pittsburgh
Footprint of Influenza in Routinely Collected Data
Veterinarian
911 calls
Poison center
...
Contagious person to person aerosol: rubella, zoster
Premonitory release: Boca Anthrax
Foodborne: Salmonella sp.
Waterborne: Cryptosporidium
Vectorborne: West Nile, Malaria, Lyme
Continuous release of bioaerosol: Legionella, pollen
Building contamination: CO
Sexually transmitted: HIV, N. gonorrhea, Hep C
Large scale bioaerosol : NO OUTBREAKS
Dato, Wagner, et al. 2001
Biomedical Security Institute 2001

2. Data-Provider-Resident Detection Component

Identifiable data | Case detectors | Access control



3. Biosurveillance Database

Need for Canonical Representation

- Heterogeneity in data sources⇒need for mappings
- DARPA project requires X validationamong teams 'Plug and play' modularity

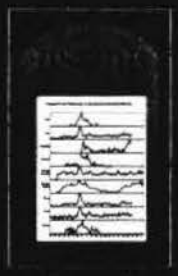
Case for Using NEDSS Standards

- They exist
- State, CDC, and local HD systems will use them

Our 'PHCDM' Database

- LOINC and Snomed
- Core Demographic Model
- NEDSS models for:
  - Aggregate ED data
  - Microbiology cultures
  - ED registration data
  - Radiology tests
  - Orders
- Out motto: "Whatever is fit for detection, we model"

4. Trusted Broker



Trusted Broker

- Trilateral agreements
  - Data contributor
  - University of Pittsburgh
  - Allegheny County Dept. of Health
- Governing board
- Research/not research
- External ethics advice (Ken Goodman, U. Miami)
- Recognition of need for transition customer

5. Spatial, Temporal and Demographics-based Detection

Andrew Moore (CMU)
Jeff Schneider (CMU)
Mike Wagner (Pitt)
Greg Cooper, Jeremy Espino, Bob Olszewski, Rich Tsui, Weng-Keen Wong
www.cs.cmu.edu/~awm
Center for Biomedical Informatics
University of Pittsburgh and Carnegie Mellon University

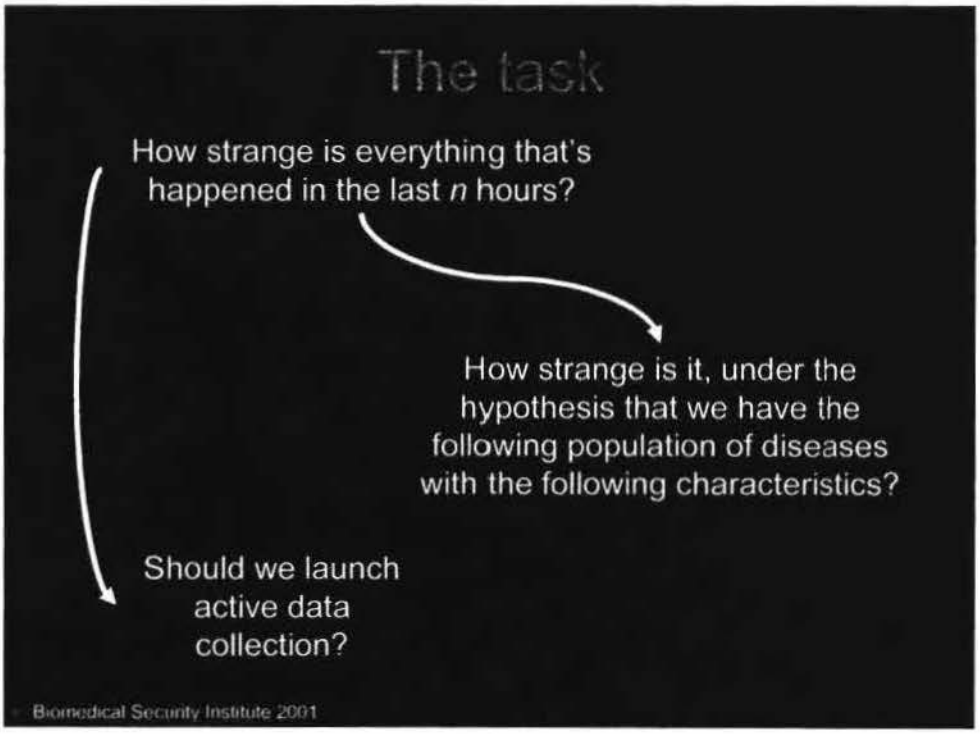
The task

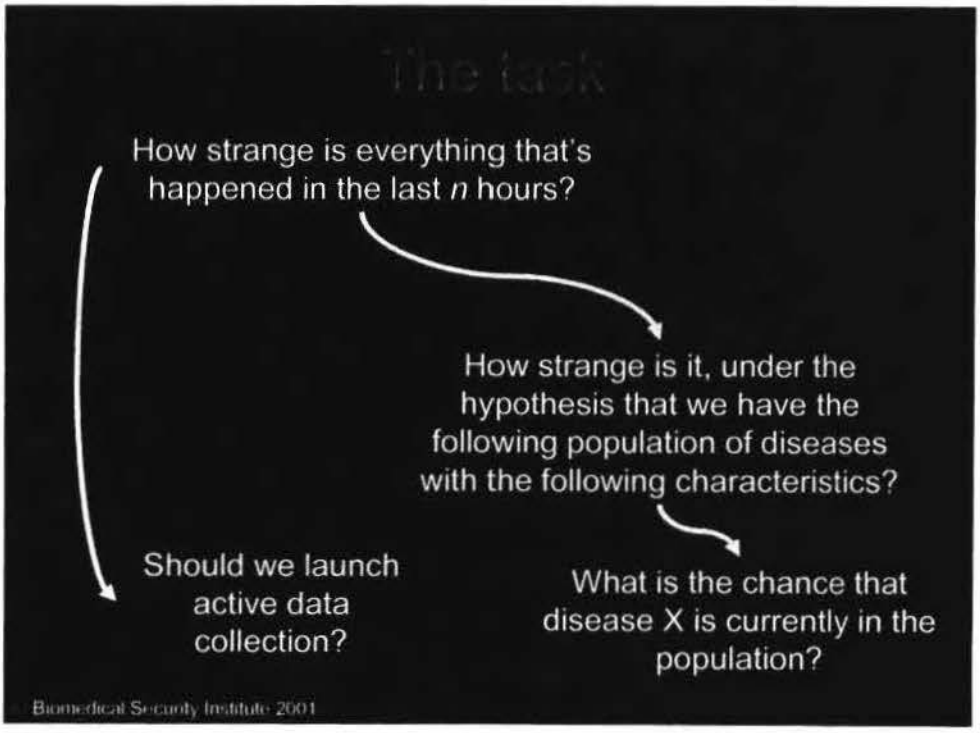
How strange is everything that's happened in the last *n* hours?

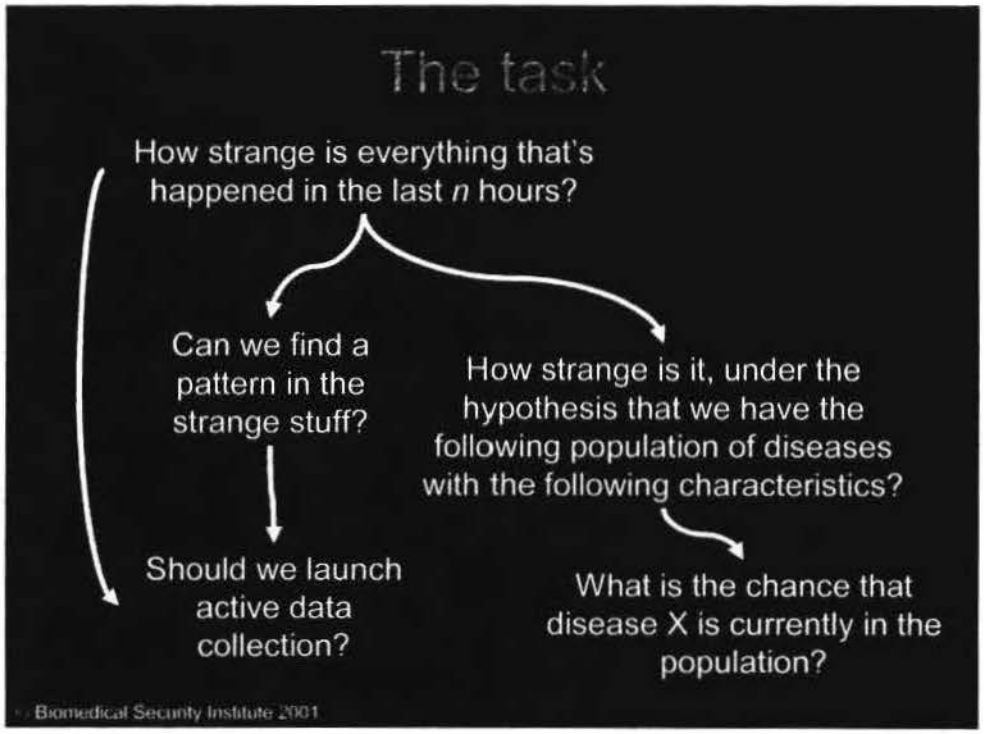
The task

How strange is everything that's happened in the last *n* hours?

Should we launch active data collection?


The task
How strange is everything that's happened in the last *n* hours?
How strange is it, under the hypothesis that we have the following population of diseases with the following characteristics?
Should we launch active data collection?
Biomedical Security Institute 2001


The task
How strange is everything that's happened in the last *n* hours?
How strange is it, under the hypothesis that we have the following population of diseases with the following characteristics?
Should we launch active data collection?
What is the chance that disease X is currently in the population?
Biomedical Security Institute 2001


The task
How strange is everything that's happened in the last *n* hours?
Can we find a pattern in the strange stuff?
Should we launch active data collection?
How strange is it, under the hypothesis that we have the following population of diseases with the following characteristics?
What is the chance that disease X is currently in the population?
Biomedical Security Institute 2001


The Task: Idea One
• **Look at the aggregate time series of various symptoms**
"It's strange that today we have gone two sigmas above usual respiratory complaint levels"
Biomedical Security Institute 2001

The Task: Idea Two

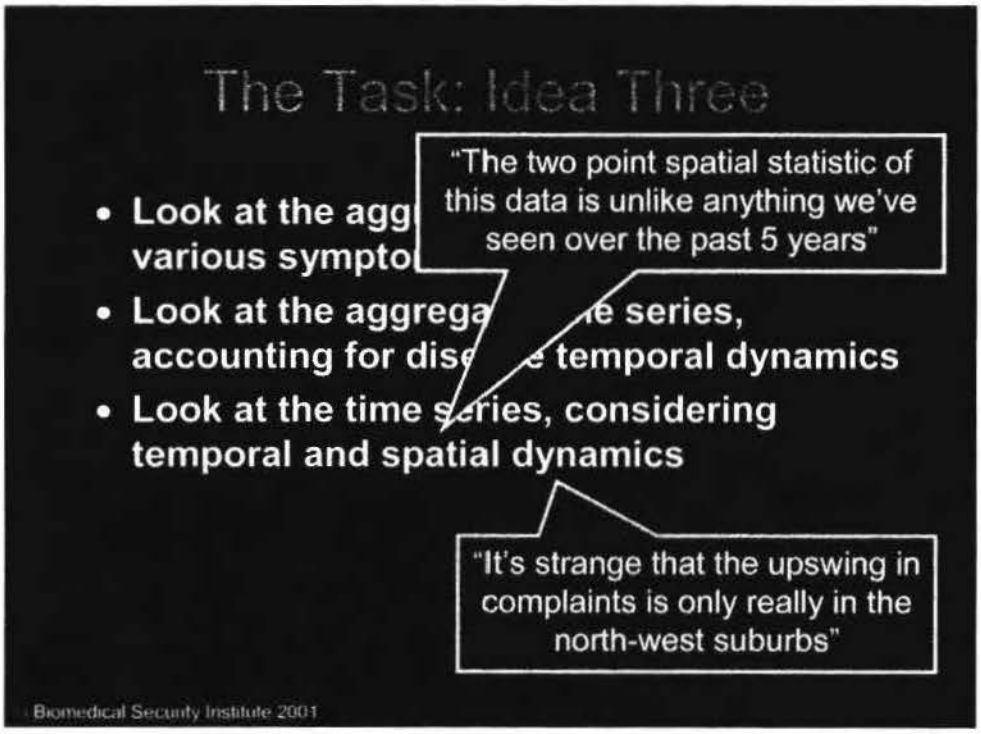
- **Look at the aggregate time series of various symptoms**
- **Look at the aggregate time series, accounting for disease temporal dynamics**

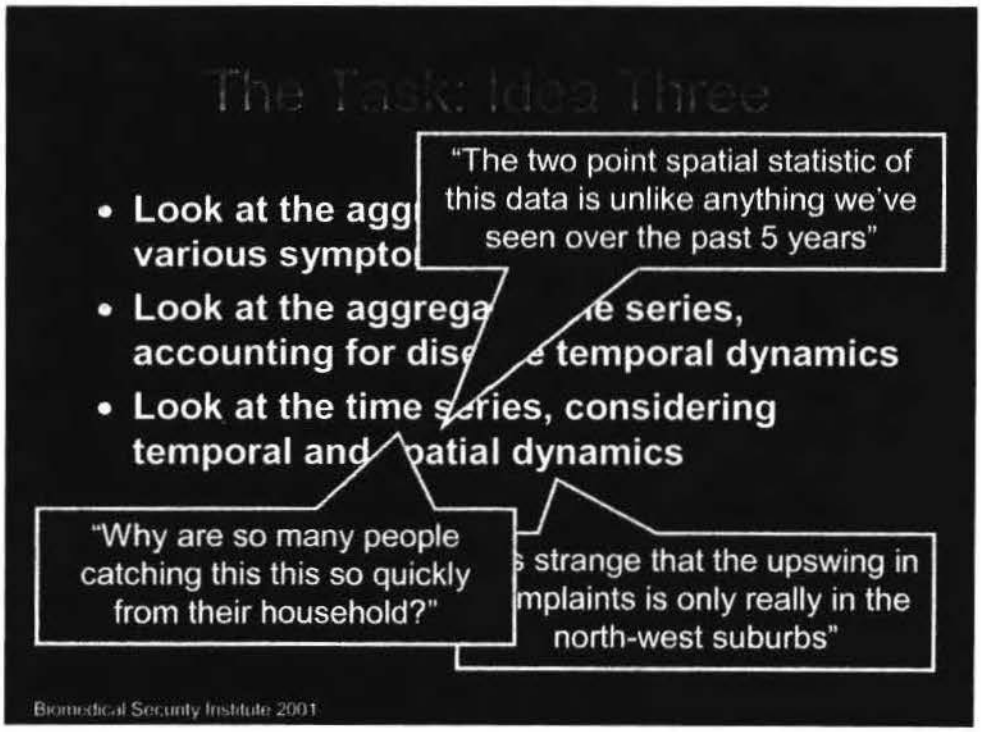
"The speed of the upswing in respiratory complaint levels is strange"

The Task: Idea Three

- **Look at the aggregate time series of various symptoms**
- **Look at the aggregate time series, accounting for disease temporal dynamics**
- **Look at the time series, considering temporal and spatial dynamics**

"It's strange that the upswing in complaints is only really in the north-west suburbs"


The Task: Idea Three
"The two point spatial statistic of this data is unlike anything we've seen over the past 5 years"
• Look at the agg
various sympto
• Look at the aggrega e series,
accounting for dis e temporal dynamics
• Look at the time series, considering
temporal and spatial dynamics
"It's strange that the upswing in complaints is only really in the north-west suburbs"
Biomedical Security Institute 2001


The Task: Idea Three
"The two point spatial statistic of this data is unlike anything we've seen over the past 5 years"
• Look at the agg
various sympto
• Look at the aggrega e series,
accounting for dis e temporal dynamics
• Look at the time series, considering
temporal and patial dynamics
"Why are so many people catching this this so quickly from their household?"
s strange that the upswing in mplaints is only really in the north-west suburbs"
Biomedical Security Institute 2001

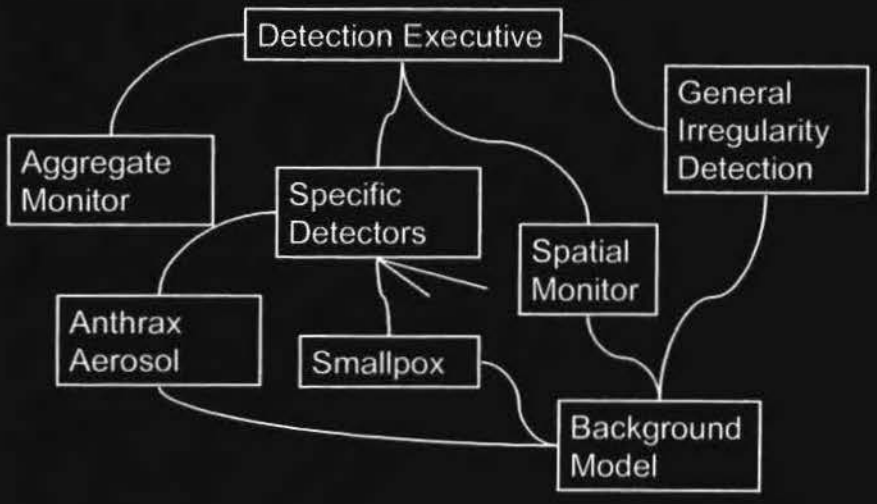
The Task: Idea Four

- **Look at the aggr**
  various symptor
- **Look at the aggr**
  accounting for d
- **Look at the time series, considerir**
  temporal and spatial dynamics
- **Look at the time series, considering temporal, spatial, and demographics**

"It would be a typical pattern of allergy-based complaints occurring in the suburbs, except I'm surprised it's so concentrated amongst young adults"

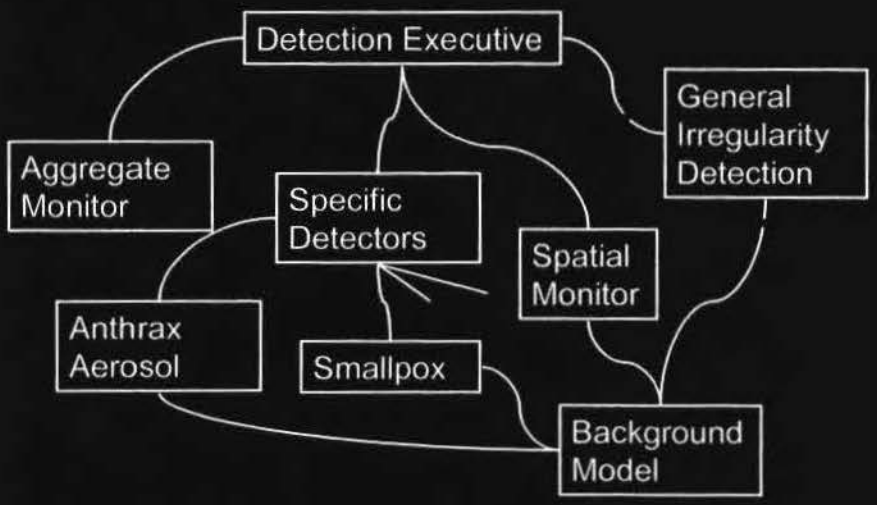
Biomedical Security Institute 2001

Sketch of complete BMSI detection approach



Biomedical Security Institute 2001

Sketch of complete BMSI detection approach



Prob-abilistic Anomalies

$$P(\text{Events of last } n \text{ hours} \mid \text{History})$$

$$= P(a_1, a_2, \ldots, a_M \mid \text{History})$$

$$= \prod_{j=1}^{M} P(a_j \mid \text{History prior to } a_j)$$

The Hidden Markov Model (HMM) approach...

$$= \prod_{j=1}^{M} P(a_j \mid \text{History prior to } a_j)$$

$$= \prod_{j=1}^{M} \sum_{q=1}^{NS} P(a_j \mid q) P(q \mid \text{History})$$

Biomedical Security Institute 2001

The AR approach...

$$= \prod_{j=1}^{M} P(a_j \mid \text{History prior to } a_j)$$

$$= \prod_{j=1}^{M} P(a_j \mid \text{Markov Blanket of recent history})$$

Biomedical Security Institute 2001

General Irregularity Detection

- Must cover a very wide space of possible irregularities
- Needs a very very careful evaluation of significance
- Should be able to explain detections to a non-statistician

Prototype System

- What features of ER cases are more common today than usual?

Prototype System

- What features of ER cases are more common today than usual?

Usual with respect to what?

Prototype System

- What features of ER cases are more common today than usual?

Usual with respect to what?

- The time of year?
- This day of the week?
- The current background of diseases?

Data Efficient Characterization of "Usual"

- HMM (Hidden Markov Model)
- Spatial HMM
- ANOVA (Analysis of Variance)

Will be implemented before September 2002

Data Efficient Characterization of "Usual"

- HMM (Hidden Markov Model)
- Spatial HMM
- ANOVA (Analysis of Variance)
  - If my classifier is allowed only access to time of year, day of week, and current background disease levels, how well will it model today's distribution of cases?
  - Now, how different is today's actual distribution of cases?

Data Efficient Characterization of "Usual"

- HMM (Hidden Markov Model)
- Spatial HMM
- ANOVA (Analysis of Variance)
  - If my classifier is allowed only access to time of year, day of week, and current background disease levels, how well will it model today's distribution o
  - Now, how dif of cases?

Will be implemented during Nov/Dec 2001

Concept Prototype Method

- What most distinguishes today's distribution of cases from…
- The distribution of cases 7 days ago and 14 days ago?

Concept Prototype Method

- Given 500 day's worth of ER cases at 15 hospitals…

Date	Cases
Thu 5/22/2000	C1, C2, C3, C4 …
Fri 5/23/2000	C1, C2, C3, C4 …
:	:
:	:
Sat 12/9/2000	C1, C2, C3, C4 …
Sun 12/10/2000	C1, C2, C3, C4 …
:	:
Sat 12/16/2000	C1, C2, C3, C4 …
:	:
Sat 12/23/2000	C1, C2, C3, C4 …
:	:
:	:
:	:
:	:
Fri 9/14/2001	C1, C2, C3, C4 …

Concept Prototype Method

- Given 500 day's worth of ER cases at 15 hospitals…
- For each day…
  - [illegible]

Date	Cases
Thu 5/22/2000	C1, C2, C3, C4 …
Fri 5/23/2000	C1, C2, C3, C4 …
:	:
:	:
Sat 12/9/2000	C1, C2, C3, C4 …
Sun 12/10/2000	C1, C2, C3, C4 …
:	:
Sat 12/16/2000	C1, C2, C3, C4 …
:	:
:	:
:	:
:	:
:	:
Fri 9/14/2001	C1, C2, C3, C4 …

Concept Prototype Method

- Given 500 day's worth of ER cases at 15 hospitals…
- For each day…
  - [illegible]
  - –
  - –

Date	Cases
Thu 5/22/2000	C1, C2, C3, C4 …
Fri 5/23/2000	C1, C2, C3, C4 …
:	:
:	:
Sun 12/10/2000	C1, C2, C3, C4 …
:	:
:	:
:	:
:	:
:	:
:	:
Fri 9/14/2001	C1, C2, C3, C4 …

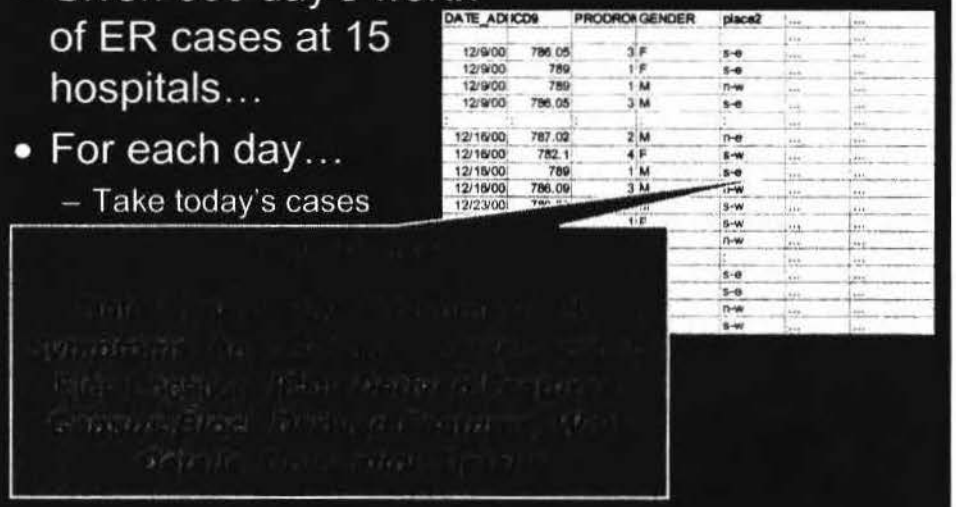
Concept Prototype Method

- Given 500 day's worth of ER cases at 15 hospitals…
- For each day…
  - Take today's cases
  - The cases one week ag
  - The cases two weeks ag
- Ask: "What's different about today?"

DATE_AD	ICD9	PRODRO	GENDER	place2	…	…
12/9/00	786.05	3	F	s-e	…	…
12/9/00	789	1	F	s-e	…	…
12/9/00	789	1	M	n-w	…	…
12/9/00	786.05	3	M	s-e	…	…
:	:	:	:	:	…	…
12/16/00	787.02	2	M	n-e	…	…
12/16/00	782.1	4	F	s-w	…	…
12/16/00	789	1	M	s-e	…	…
12/16/00	786.09	3	M	n-w	…	…
12/23/00	789.09	1	M	s-w	…	…
12/23/00	789.09	1	F	s-w	…	…
12/23/00	782.1	4	M	n-w	…	…
:	:	:	:	:	…	…
12/23/00	786.09	3	M	s-e	…	…
12/23/00	786.09	3	M	s-e	…	…
12/23/00	780.9	2	F	n-w	…	…
12/23/00	V40.9	7	M	s-w	…	…

Concept Prototype Method

- Given 500 day's worth of ER cases at 15 hospitals…
- For each day…
  - Take today's cases



Example

```
Sat 12-23-2000 (daynum 36882, dayindex 239)
SCORE =   2.720
8.76% ( 12/137) of today's cases have PRODROME=5
3.07% (  9/293) of   other cases have PRODROME=5
```

Searching for the best score…

- Try ICD9 = x for each value of x
- Try Gender=M, Gender=F
- Try CoarseRegion=NE, =NW, SE, SW..
- Try FineRegion=AA,AB,AC, … DD (4x4 Grid)
- Try Hospital=x, TimeofDay=x, Prodrome=X, …
- [In future… features of census blocks]

Overfitting Alert!

Example

```
Sat 12-23-2000 (daynum 36882, dayindex 239)
SCORE =   2.720  PVALUE =  0.286
8.76% ( 12/137) of today's cases have PRODROME=5
3.07% (  9/293) of   other cases have PRODROME=5
```

Example

```
Sat 12-23-2000 (daynum 36882, dayindex 239)
SCORE =    2.720   PVALUE =   0.286
8.76% ( 12/137) of today's c ses have PRODROME=5
3.07% (  9/293) of    othe   ses have PRODROME=5
```

If I had applied the same overfitting-prone procedure to the same kind of data in which there was no relation between case features and date, what's the chance I'd have found a ratio this strong?

Example

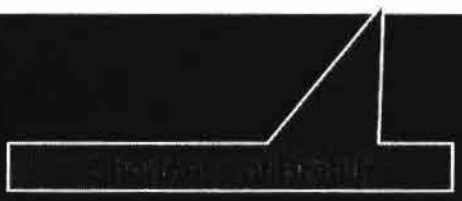
```
Sat 12-23-2000 (daynum 36882, dayindex 239)
SCORE =    2.720   PVALUE =   0.286
8.76% ( 12/137) of today's c ses have PRODROME=5
3.07% (  9/293) of    othe   ses have PRODROME=5
```

If I had applied the same overfitting-prone procedure to the same kind of data in which there was no relation between case features and date, what's the chance I'd have found a ratio this strong?

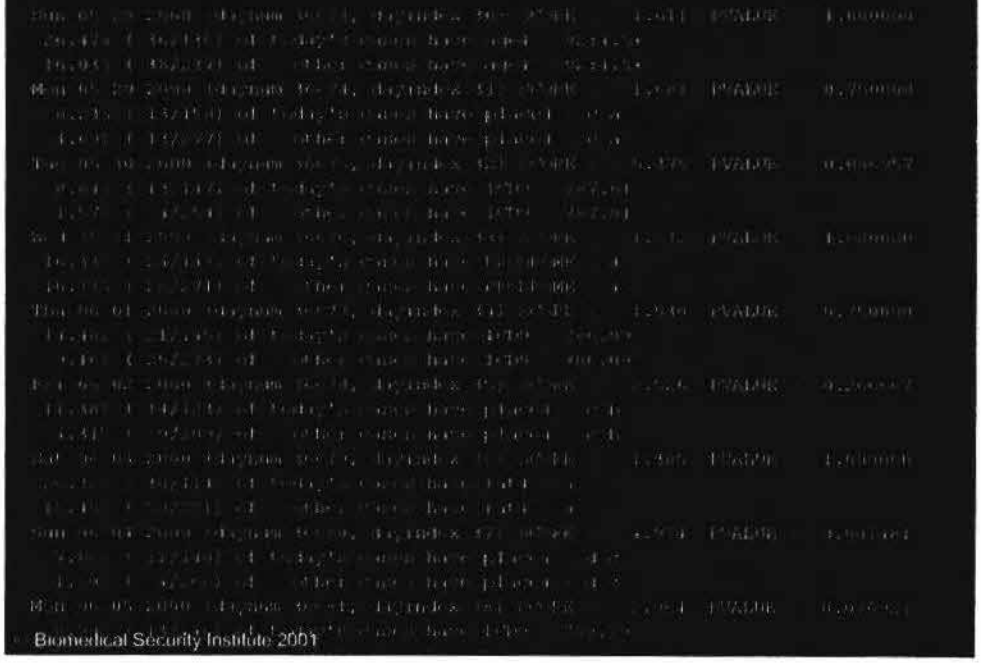
Biomedical Security Institute 2001

Interesting Example

```
Sat 01-06-2001 (daynum 36896, dayindex 253)
SCORE =   4.279  PVALUE =  0.009975
7.43% (11/148) of today's cases have ICD9=786.05
1.66% ( 5/302) of   other cases have ICD9=786.05
```



The first few of the results



Which of the 500 days have Irregularities?

- WARNING: A second overfitting opportunity.
- If we took 500 days of randomly generated data, then about 5 days would have a p-value below 0.01
- This can be solved with…
  - A Bonferroni correction
  - The FDR (False Discovery Rate) method [Benjamini & Hochberg, 1995, J. R. Stat Soc, 57 289]

Results: Dates that Survive FDR

FDR survivors (page 1 of 2)

Biomedical Security Institute 2001

FDR survivors (page 2 of 2)

Biomedical Security Institute 2001

Sanity check

- What happens if we generate a fake database in which we know that there can be no relation between date and case features?
- This can be achieved by shuffling all the dates in the database.
- The days detected by FDR are then…

Survived Randomized

It is good and important that far fewer survived on the randomized database

Searching for complex patterns

- What if we allow rules to have two components?

Complex Detections (1 of 2)

Complex Detections (2 of 2)

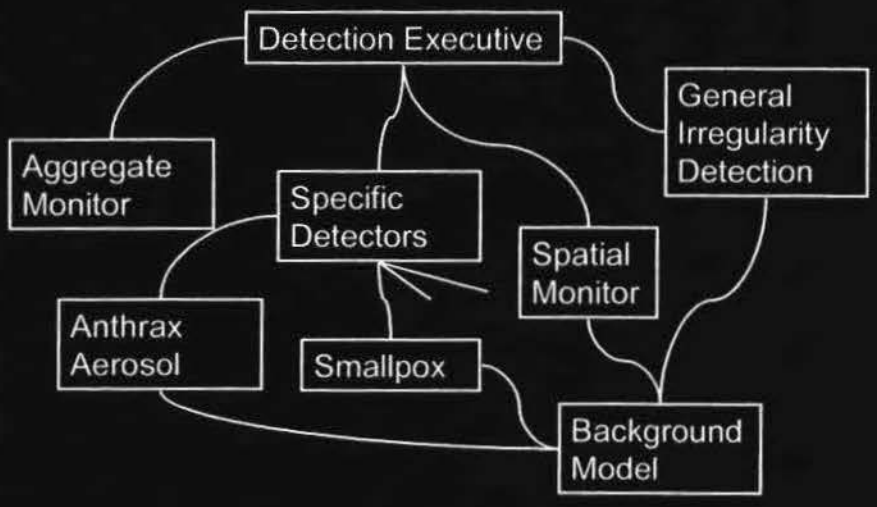
Complex detections on randomized data

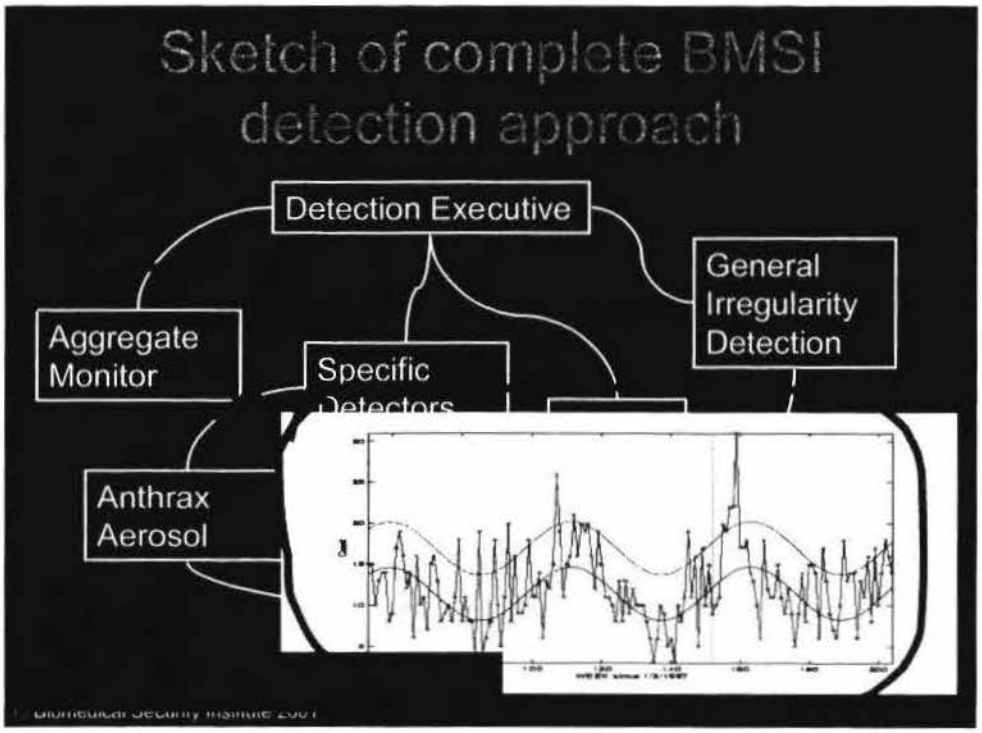
We're please that only one detection got through FDR on the randomized data.

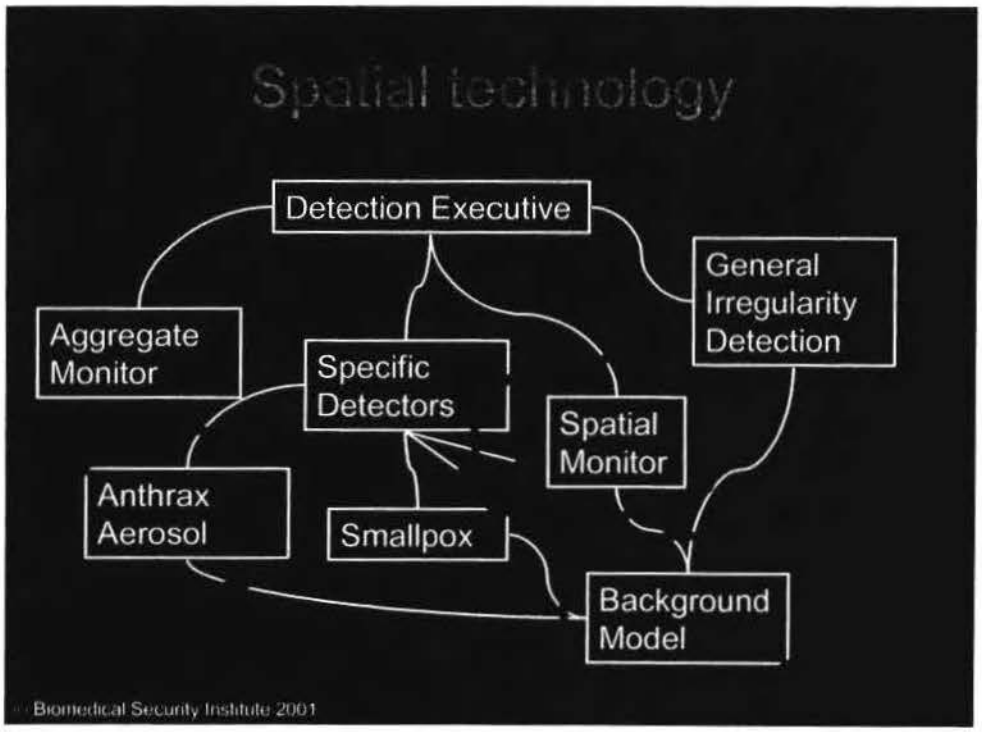
Next steps

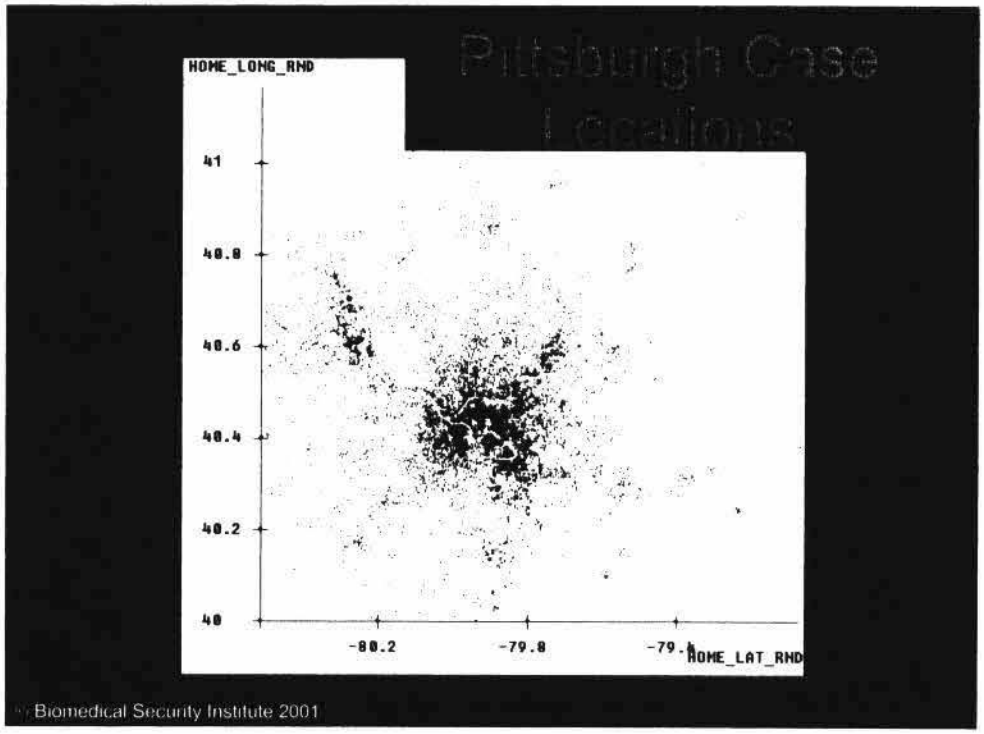
- Including non-ER data sources
- ANOVA-type scoring so we may compare against the full dataset instead of two other days
- Design of scores with greater statistical power
- Validation on simulation and real data

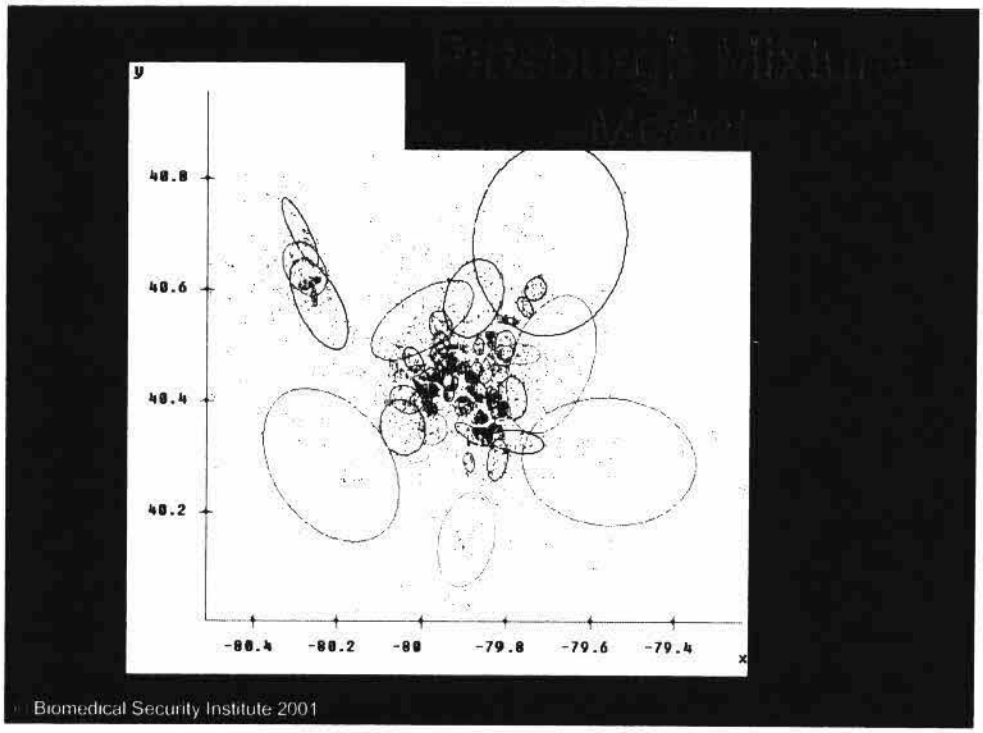
Sketch of complete BMSI detection approach

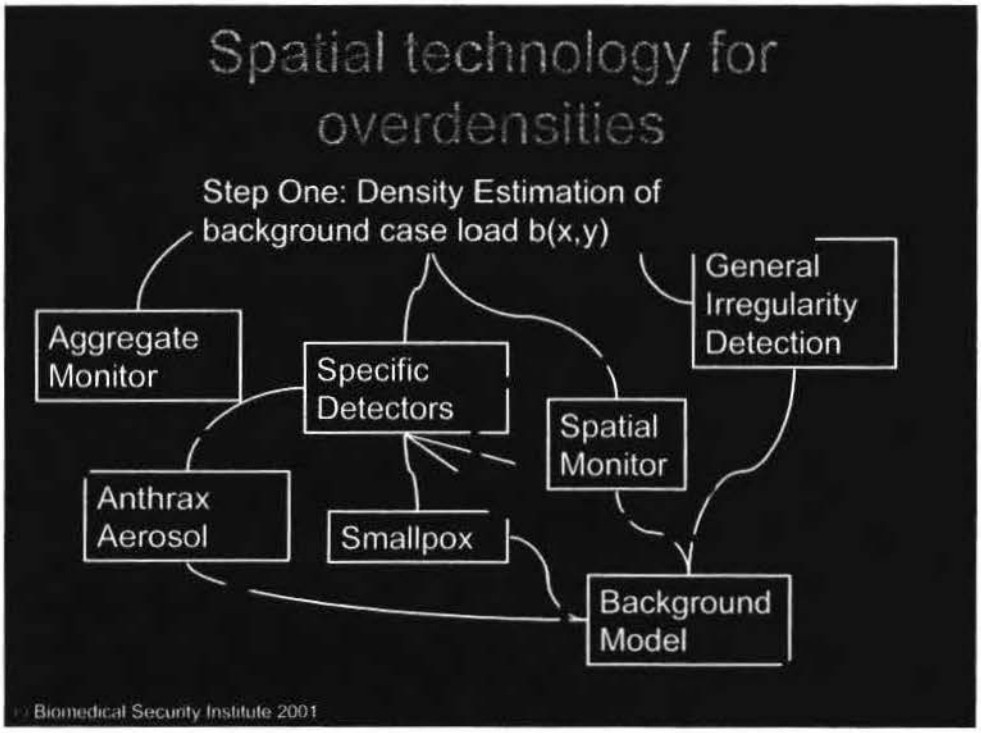


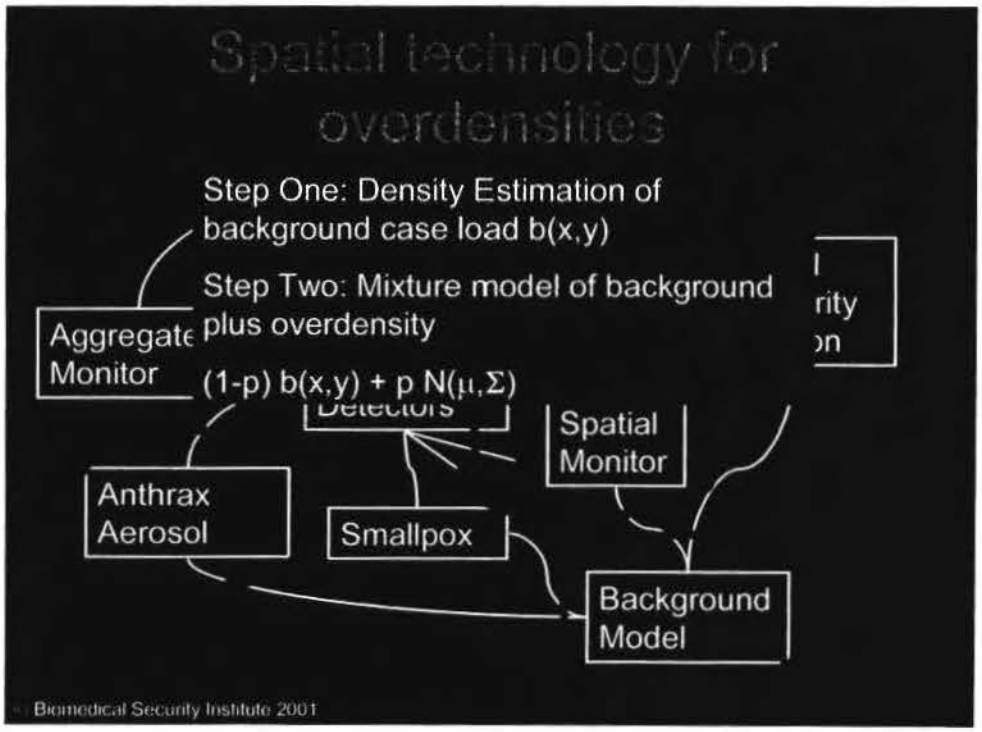

Sketch of complete BMSI detection approach
Detection Executive
General Irregularity Detection
Aggregate Monitor
Specific
Anthrax Aerosol


Spatial technology
Detection Executive
General Irregularity Detection
Aggregate Monitor
Specific Detectors
Spatial Monitor
Anthrax Aerosol
Smallpox
Background Model
Biomedical Security Institute 2001


Pittsburgh Case Locations
HOME_LONG_RND
41
40.8
40.6
40.4
40.2
40
-80.2
-79.8
-79.4
HOME_LAT_RND
Biomedical Security Institute 2001


Personnel Mixture Model
y
40.8
40.6
40.4
40.2
-80.4
-80.2
-80
-79.8
-79.6
-79.4
x
Biomedical Security Institute 2001


Spatial technology for overdensities
Step One: Density Estimation of background case load b(x,y)
General Irregularity Detection
Aggregate Monitor
Specific Detectors
Spatial Monitor
Anthrax Aerosol
Smallpox
Background Model
Biomedical Security Institute 2001


Spatial technology for overdensities
Step One: Density Estimation of background case load b(x,y)
Step Two: Mixture model of background plus overdensity
(1-p) b(x,y) + p N(μ,Σ)
Aggregate Monitor
Spatial Monitor
Anthrax Aerosol
Smallpox
Background Model
Biomedical Security Institute 2001

Spatial technology for overdensities

Step One: Density Estimation of background case load b(x,y)

Step Two: Mixture model of background plus overdensity

(1-p) b(x,y) + p N(μ,Σ)

Step Three: Mixture model of background plus overdensities

(1-p) b(x,y) + p_1 N(μ_1,Σ_1) + p_2 N(μ_2,Σ_2)...

Aggregate Monitor

Anthra Aerosc

rity on

Background Model

Biomedical Security Institute 2001

Spatial technology for overdensities

Step One: Density Estimation of background case load b(x,y)

Step Two: Mixture model of background plus overdensity

(1-p) b(x,y) + p N(μ,Σ)

Step Three: Mixture model of background plus overdensities

(1-p) b(x,y) + p_1 N(μ_1,Σ_1) + p_2 N(μ_2,Σ_2)...

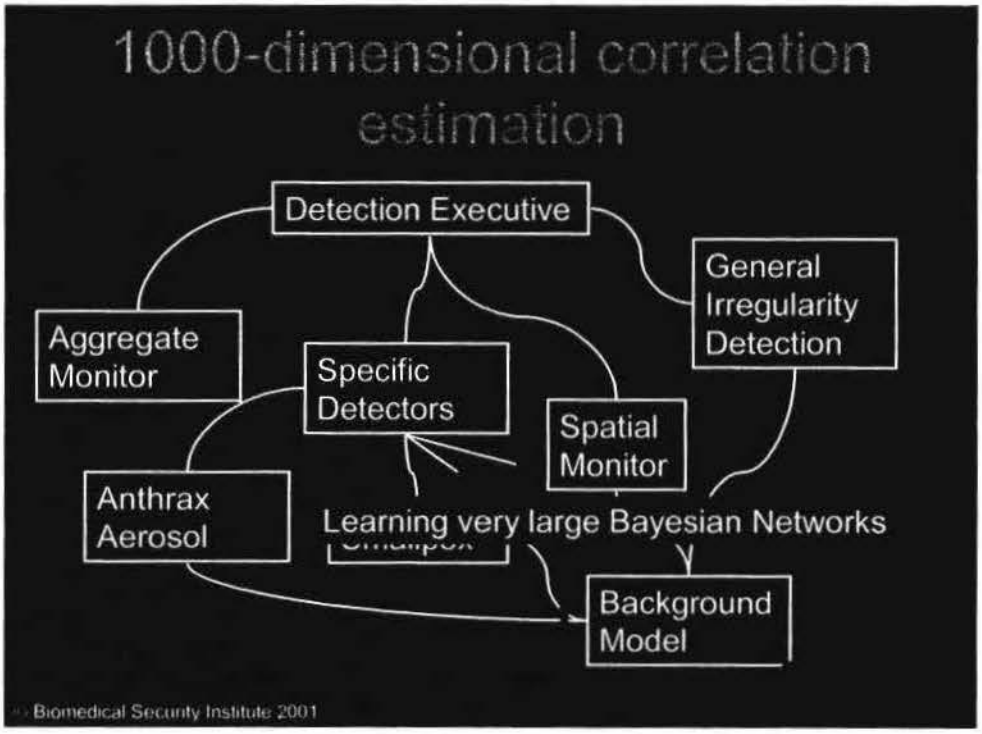
Aggregate Monitor

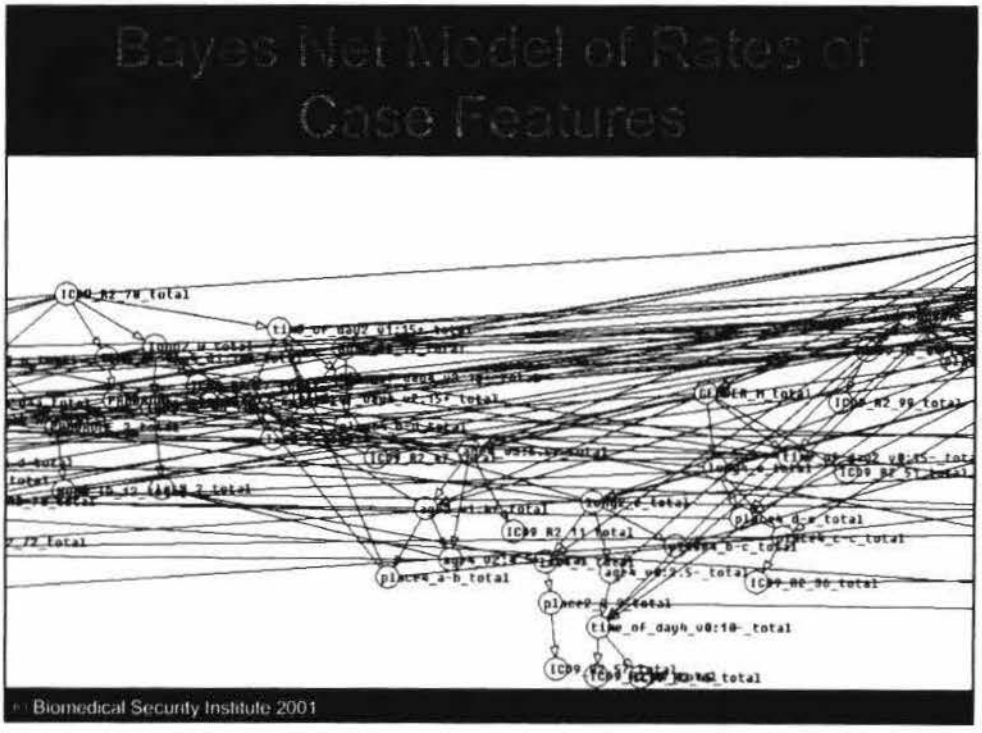
Anthra Aerosc

rity on

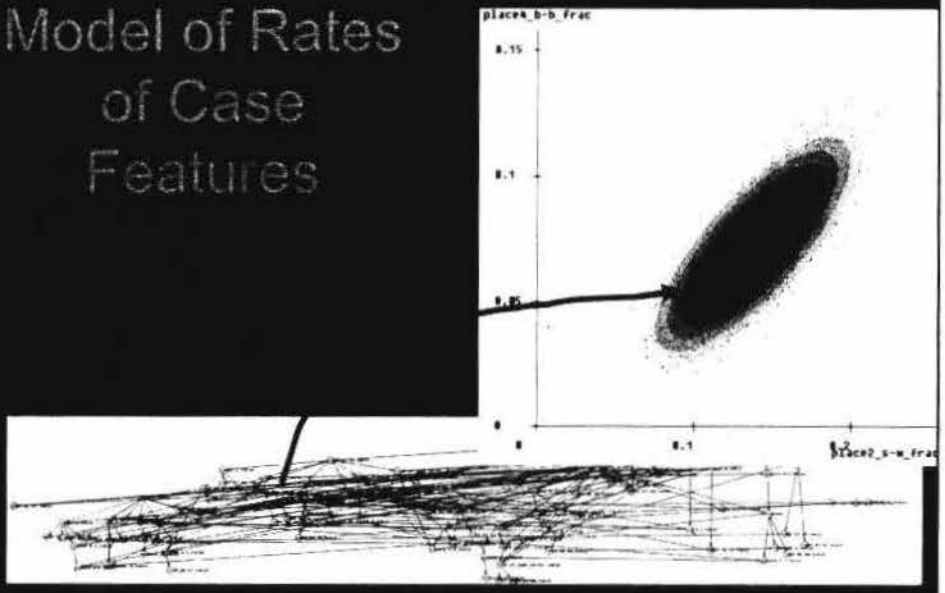
Then: Nonparametric. Inverse-problem for start time/place. Prodrome-specific.

Biomedical Security Institute 2001


1000-dimensional correlation estimation
Detection Executive
General Irregularity Detection
Aggregate Monitor
Specific Detectors
Spatial Monitor
Anthrax Aerosol
Learning very large Bayesian Networks
Background Model
Biomedical Security Institute 2001

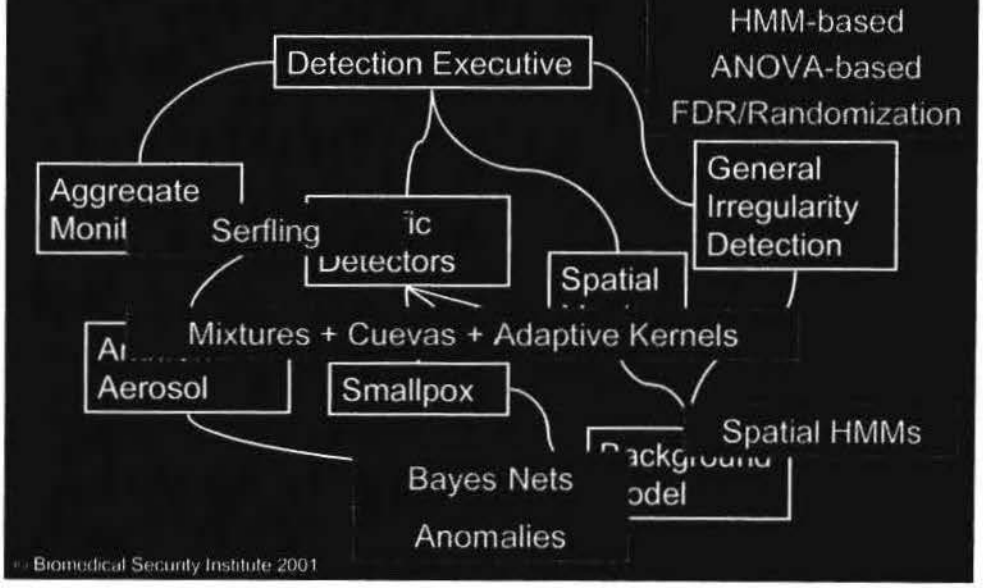

Bayes Net Model of Rates of Case Features
Biomedical Security Institute 2001

Bayes Net Model of Rates of Case Features

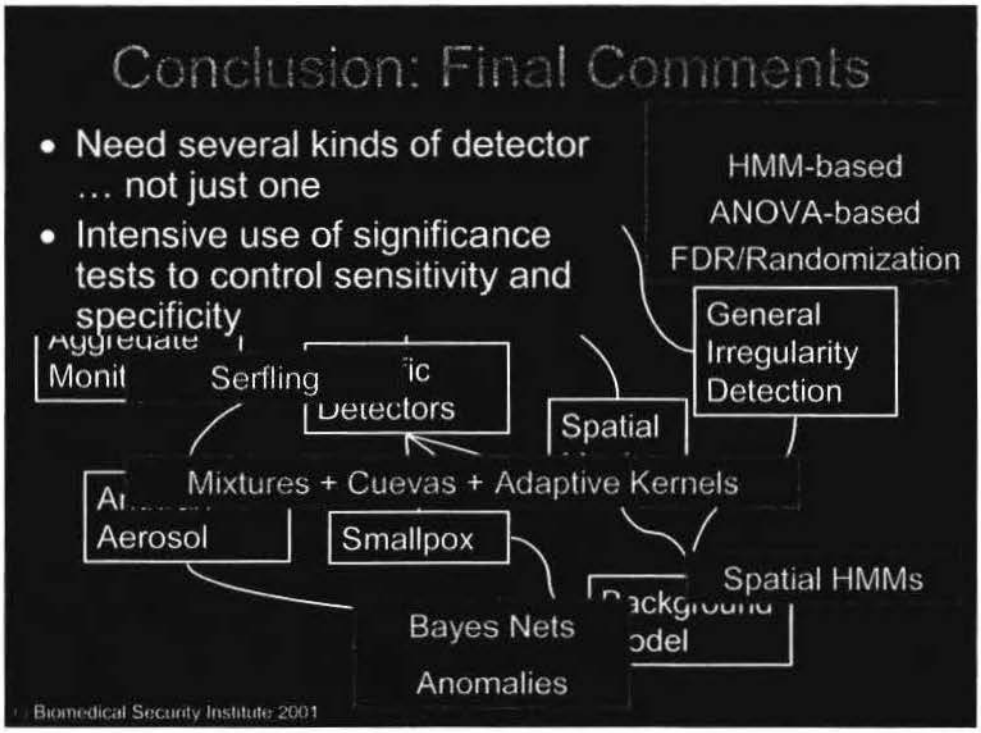


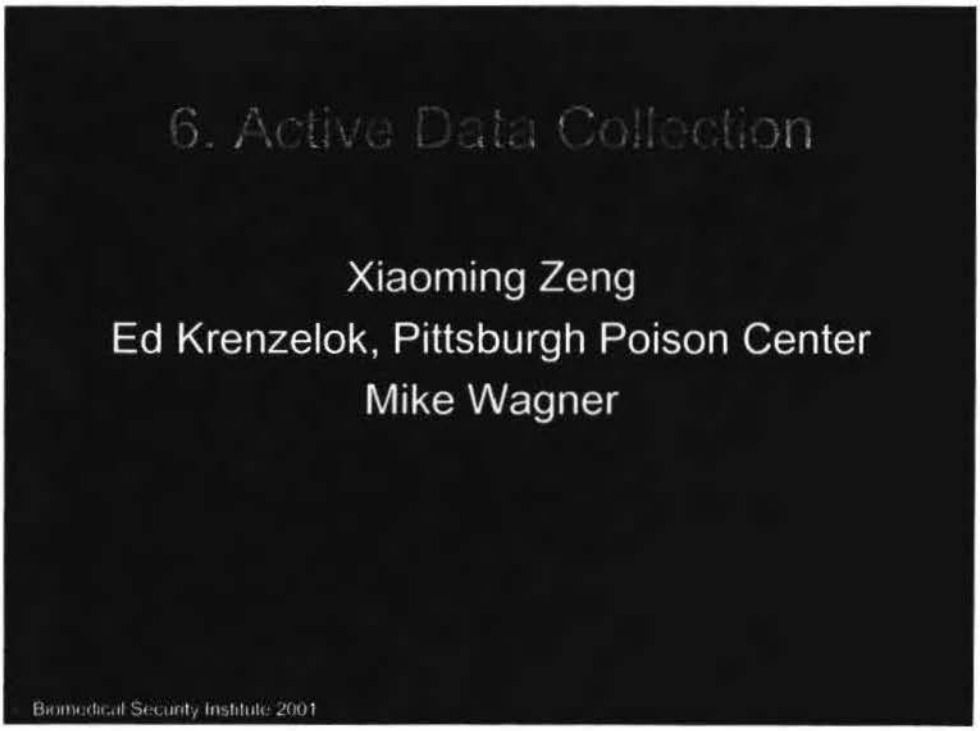
Biomedical Security Institute 2001

Conclusion: Final Comments

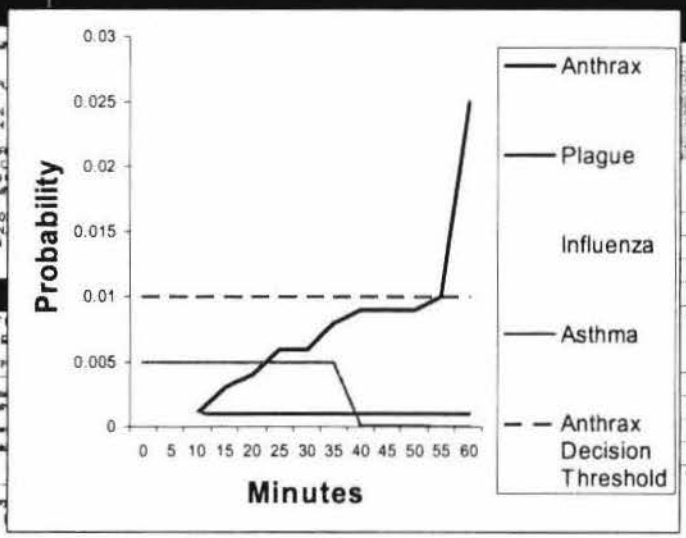


Biomedical Security Institute 2001


Conclusion: Final Comments
• Need several kinds of detector … not just one
• Intensive use of significance tests to control sensitivity and specificity
HMM-based
ANOVA-based
FDR/Randomization
General Irregularity Detection
Serfling
Spatial
Mixtures + Cuevas + Adaptive Kernels
Aerosol
Smallpox
Spatial HMMs
Bayes Nets
Anomalies
Biomedical Security Institute 2001


6. Active Data Collection
Xiaoming Zeng
Ed Krenzelok, Pittsburgh Poison Center
Mike Wagner
Biomedical Security Institute 2001

Linking Passive and Active Detection



Active Data Collection: Status

- Done
  - Respiratory and gastroenteritis forms in Java using branching logic of Texas Dept. of Health
  - Connection to database
- In Progress
  - Study of data availability and timeliness
  - Study of human factors, accuracy of encoding, economics, and capacity of Poison Centers
- Plans
  - Develop rash, hemorrhagic … forms
  - Develop detection algorithms that use these data
  - Value-of-information driven questioning
  - Test ENCODER, a semi-automatic encoding tool

7. Decision-Making Under Uncertainty: Early Response to Possible Large-Scale Anthrax Release

Scott Braithwaite
Doug Fridsma
Mark Roberts
Mike Wagner

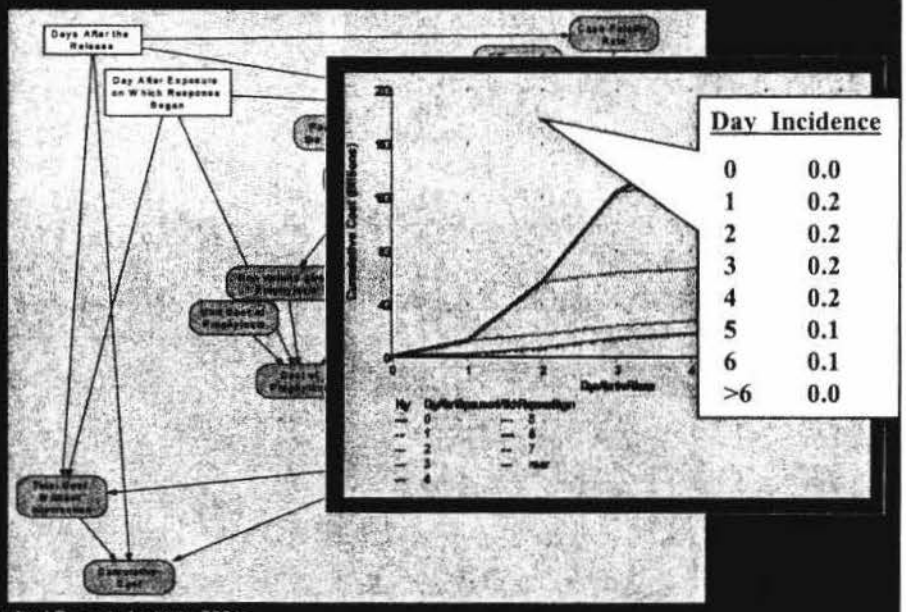
Starting Point: Economic Model of Anthrax Release (Kaufmann)

WHAT IF ...

- 100,000 people receive LD50 exposure
- Onset of illnesses occurs over days 1-7
- Mass treatment occurs on days 0, 1, 2, 3, 4, 5, 6 or 7... and it has 90% efficacy
- The costs are
  - Treatment of sick
  - Prophylaxis of healthy (exposed and unexposed)
  - Future earnings lost through deaths, valued at approximately $790,000 per

Kaufmann. The economic impact of a bioterrorist attack: Are prevention and postattack intervention programs justifiable *EID* 3(2):83-94, 1997.

Kaufmann Model (in Analytica)



Biomedical Security Institute 2001

Goal: A Decision Model

Detection System Output
P(anthrax, day 1, 100,000)=.001
P(anthrax, day 2, 100,000)=.01
P(anthrax, day 1, 50,000)=.00001
...

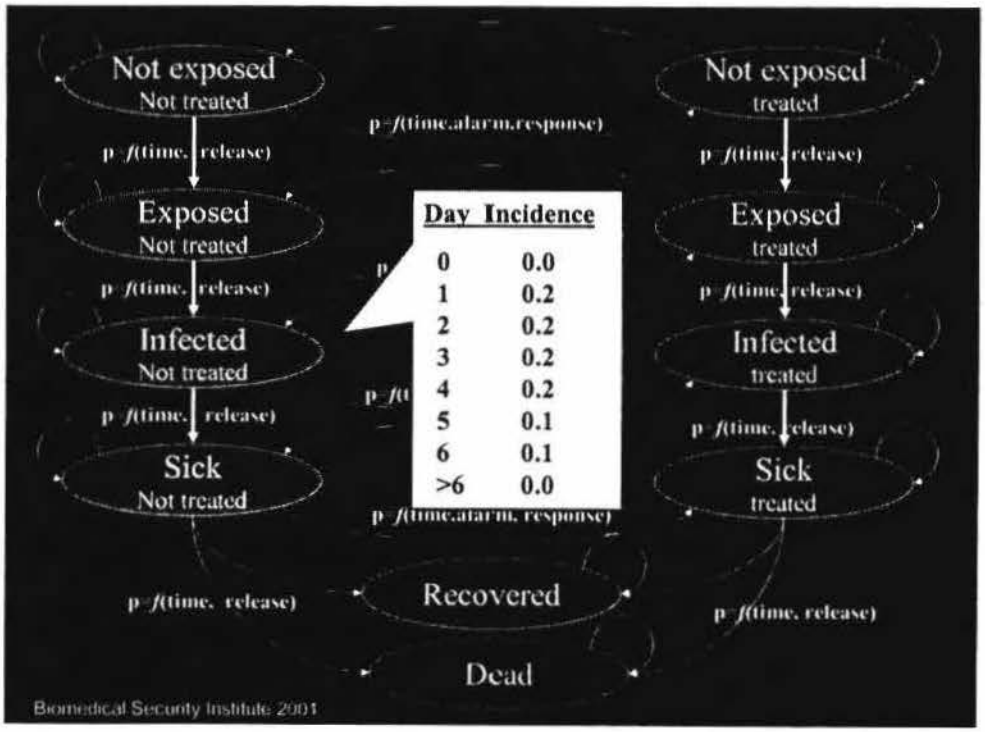
- U(#dead, resources consumed)
- Epidemic Model

State info at *t* — Decision at *t* — *t* + 1

Response Options
Wait an hour
Collect more information
Mobilize local pharmaceutical cache
Mobilize National Pharmaceutical Stockpile
...

Model 2: Markov Model of Anthrax Release

- Based on Kaufmann assumptions
- Time step still one day
- 30-day horizon



Comparison of Kaufmann and University of Pittsburgh Model

	Kaufmann Model	University of Pittsburgh Model
#Dead/100,00 assuming NO alarm	32,900	28,400
#Dead/100,000 assuming alarm and response day 0	7,000	7,600
Cost/100,000 assuming NO alarm ($million)	$26.2	$22.7
Cost/100,000 assuming alarm and response day 0) ($million)	$7.8	$6.0

Decision Models: Next Steps

- Add detection system characteristics
  - *WHAT IF the detection system has 0.01 specificity on day 1, and 0.1 on day 2?*
- Add response alternatives and costs
  - *What is the cost of a false alarm?*
  - *When the detection system says P(anthrax) =0.01, what is the optimal action?*
- Use BioWar and Episim simulators
  - *What if there are only X doses of Cipro?*

Summary and Additional Projects

DESCRIBED IN BRIEFING

1. Signal Evaluation
2. Data-Provider Component
3. Trusted Broker
4. Biosurveillance Data Model
5. Spatial temporal algorithms
6. Active data collection
7. Response decisions under uncertainty
8. Biowar simulator

ADDITIONAL PROJECTS

- Risk-Utility (RU) analysis
- Report on current capacity[1]
- Report on data availability[1]
- ICD-9: two papers[1,2]
- NLP[1,4]
- Grocery store framework[1,2]
- Review of literature on behaviors of sick[1]
- RODS[3]

1.a.14

BIOWAR
Modeling & Evaluating Response

Kathleen M. Carley

Professor of Sociology and Organizations: kathleen.carley@cmu.edu

Center for Computational Analysis of Social and Organizational Systems (CASOS)

Departments of Social and Decision Sciences, Engineering and Public Policy, and Heinz School of Public Policy and Management

Carnegie Mellon University

Team: Doug Fridsma – CBMI, BMSI, University of Pittsburgh; Benoit Morel – EPP, CASOS, Carnegie Mellon; David Deerfield – Pittsburgh Super Computer; Alex Yahja –EPP, CASOS, graduate student, Carnegie Mellon

Biomedical Security Institute

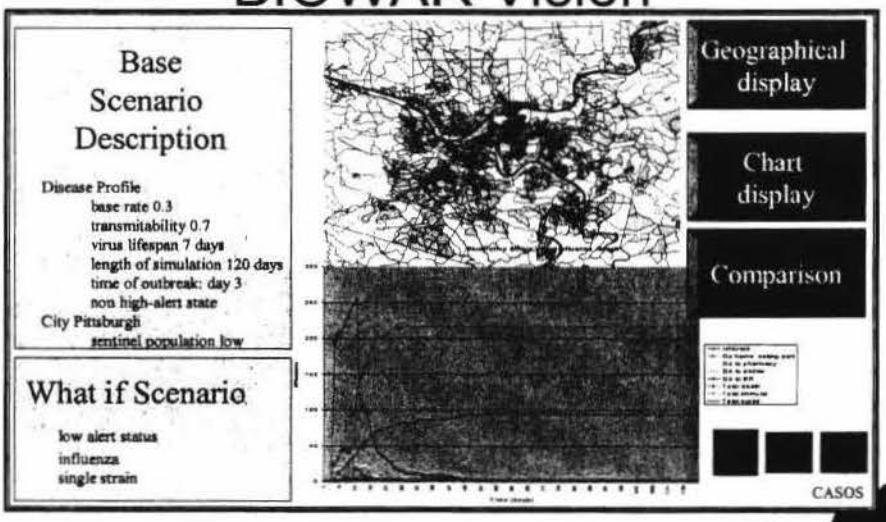
University of Pittsburgh and Carnegie Mellon

DARPA - Bio-Surveillance Program - Kickoff

BIOWAR

- Objective
  - Automated tools for evaluation of response policies, data efficacy, attack severity, and detection tools relating to weaponized biological attacks
- Tasks
  - Develop prototype computational model of responses to weaponized biological attacks at the city level
  - Generation of artificial data for early detection studies
  - Illustrations of use – "What IF"
  - Initial data integration and validation
- Approach
  - Combine network, epidemiological, and geographical components into adaptive multi-agent network model that can be used as a "what if" analyzer
- Progress
  - Initial alpha prototype model capable of generating high level general behavior

BIOWAR Vision



Why Use Computational Modeling and Analysis?

- *Ethical:* Cannot test response policies on real populations
- *Preparatory:* Can create hypothetical weapons with more potency than existing ones – Can examine wide range of scenarios
- *Cost effective:* Creating new technologies, procedures and legislation for data collection is expensive
- *Faster:* Real time evaluation of existing systems is too time consuming
- *Appropriate:* Complex non-linear dynamic system
- *Flexible:* Response to novel situations requires rapid evaluation of previously unexamined alternatives

Tasks

- Evaluation of existing computational models ☑
- Design and develop prototype system for examining weaponized biological attacks on city level populations
  - Combine epidemiological and communication network models ☑
  - Create disease models
  - Create realistic agent models ☑ prototype
  - Create symptom testing models
  - Be able to read/write to shared NEDSS database
  - Scale model to city size
  - Link to geographical location data model and presentation (Arcview)
- Initialize model with real-world data
  - Physical location, Census, Demographics, Social network, Cognitive biases
- Develop and illustrate "What If" capabilities
- Initial validation using influenza data

Limitations of Existing Models

- Epidemiological models assume uniformity of population - networks
- Social network / communication models ignore disease
- Existing agent-based models cognitively, socially and geographically unrealistic
- Lack of connection to real large scale data

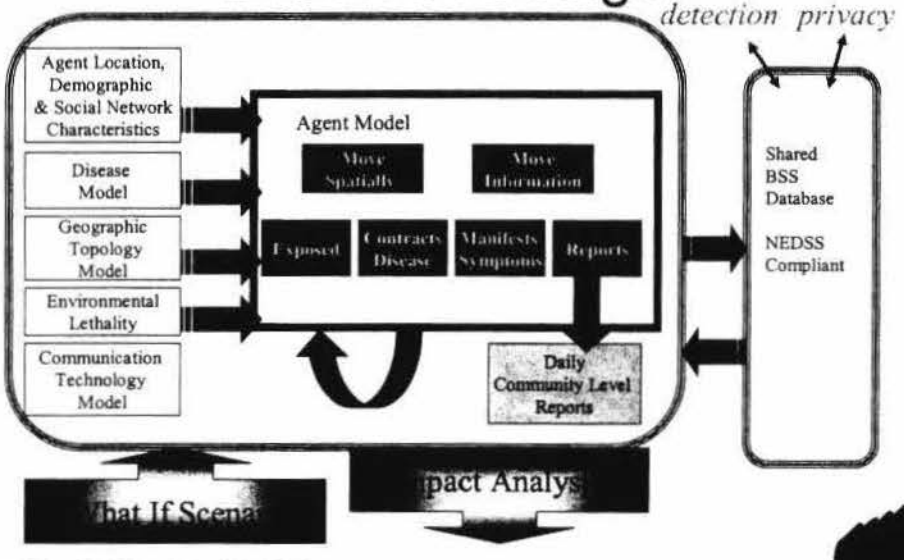
Challenges to Be Met

- Combine epidemiological, network and geographical location models
- Create cognitively and geographically realistic agent-based model
- Create a flexible enough system to explore a wide range of unanticipated scenarios
- Data integration and validation

Approach

- Multi-Agent Network Model
  - Cognitive realistic
  - Socially realistic - Embedded in to social, knowledge and task networks
  - Integrated with geographic model
  - Organizational/Unit network
  - Communication technologies
- Hybrid of many models
  - Spatial
  - Disease
  - Network
  - Epidemiological
  - Cost
  - Agent
- Examine
  - Existing standard diseases viral and non viral – influenza
  - Weaponized contagious – pneumonic plague, smallpox
  - Weaponized non-contagious – anthrax
- What If Analysis

BIOWAR Design



Parameterization & Initialization of BioWar

- **Integration – Utilizing Real Data for Parameterization**
  - Disease Data
    - Lethality, Type, Symptoms, Timing
  - Cognitive Data
  - Demographic Data
  - Geographic Data
  - Behavioral Data

Multi-source

Public

- **Data Source**
  - Disease Data
    - Archival, Medical Journals, Historical Accounts
  - Cognitive Data
    - Human experimental and field studies in cognitive science
  - Demographic Data
    - Census
    - GSS
  - Geographic Data
    - Maps, Census
  - Behavioral Data
    - Human experimental and field studies in sociology, anthropology, psychology

Verification & Validation

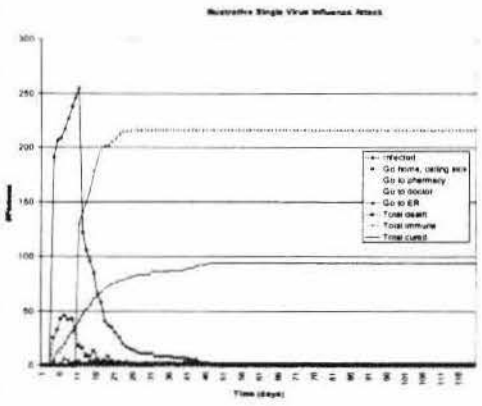
- **BioWar – Simulated Data**
  - General Behavior
    - Cross-sectional
    - Over time
  - Virtual Response Data
    - Absenteeism
    - ER visits
    - Pharmacy
    - Death rate
    - Web hits
    - Cost
- **Validation – Real Data**
  - General behavior
    - Herd immunity
  - Influenza
    - Grade School Absenteeism
    - ER reports
    - Pharmacy purchases
    - Death reports

Inform future data collection

More Options ⇄ *Samples, Incomplete*

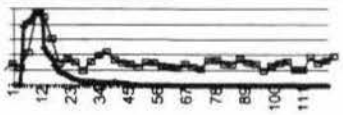
Validate and tune model

BioWar Alpha Prototype:
Illustrative Results Single Virus Influenza Attack

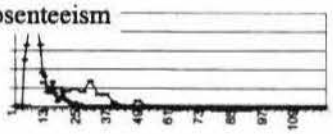


BioWar Actual

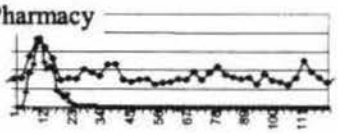
Infected Influenza



Absenteeism



Pharmacy

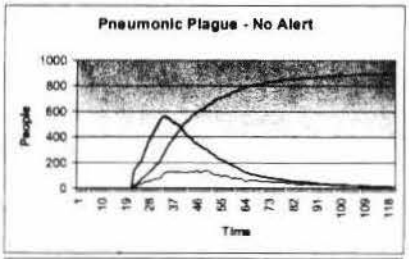


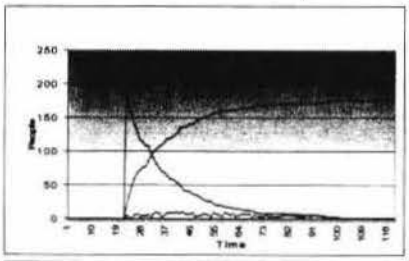
51 3 7 11 15 19 23 27 31 35

BioWar Alpha Prototype:
Illustrative Results of What IF Analysis

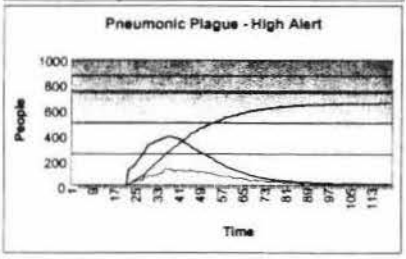
Infected
Death
Dr/ER

Pneumonic Plague - No Alert

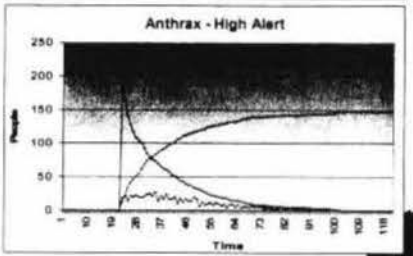




Pneumonic Plague - High Alert



Anthrax - High Alert



Summary of Plans

- Develop BIOWAR
- Combine network, epidemiological, geographical, disease, symptom, cost components into adaptive multi-agent what if analyzer
- Scale system to city level
- Illustrate use of BIOWAR
  - Evaluate possible early response policies
  - Evaluate relative efficacy of different early detection data sources and privacy policies
  - Evaluate relative severity of different types of attacks
- Generation of artificial data for early detection studies
  - Anthrax
  - Pneumonic Plague
  - Smallpox
  - Influenza
- Initial data integration and validation

1.a.15 5

Bio-Surveillance System Data/Information Sharing

(b)(6)

[redacted]

October 12, 2001

1

Outline

- Objective
- Teams Data Sources
- Petting Zoo
- Data Sharing
- How to get there?

2

Objectives

- **Data Sharing is highly desirable**
  - Allow the testing of prototype systems with different data sources beyond the individual team initial data
  - Help determine the data element having important values for the program
- **Develop a consistent methodologies for evaluating the alerts provided by the detection system**

3

Team Data Sources (1)

- **Use of existing team data sources**
  - National Capital Area data for GEIS from the 8 MTFs (JHU/APL Team)
  - Public Health Departments of the City and County of San Francisco (Veridian Team)
  - Other data sources (e.g., weather, CDC, and others)

4

Team Data Sources (2)

- **Identify other new data sources for collection**
  - Collect Behavior of Sick people
  - Arclight Systems
    - $R_x$ealTime

5

Petting Zoos (1)

Metro DC (DC, and suburban Maryland and Virginia)

- Old Maryland Farm 301 Watkins Park Drive, Upper Marlboro, MD 301-218-6770;goats, sheep, chickens, rabbits, peafowl, turkey, hog, donkey, cow, pony;
- Catoctin Wildlife Preserve 13019 Catoctin Furnace Rd, Thurmont, MD 301-271-3180; 300 exotic animals; Sept 9-5, Oct 10-5, Nov (temp > 45^0), closed Dec thru Feb
- Oxon Hill Farm 6411 Oxen Hill Rd, Oxon Hill, MD 301-839-1176; 8 AM – 4:30PM open year round; working farm with cows, chickens, farm animals

6

Petting Zoos (2)

- Leesburg Animal Park 19270 James Monroe Highway, Leesburg, VA (703)433-0002; Fri,Sat,Sun 10AM – 5PM, closed on Monday; Nov –10AM-4PM weather permitting), closed Dec to Mar; llamas, goats, sheep, donkeys, deer, antelope, squirrel monkeys, lemurs, zebra & more
- Reston Zoo 1228 Hunter Mill Rd, Vienna, VA 703-757-6222; 9-5 Weekdays, 9-6 Weekends; baby lambs, zebras, camels, giraffes, bison, antelope, ostrich, monkey, pony rides
- Kidwell Farm at Frying Pan Park 2709 West Ox Rd, Herndon, VA 703-437-9101; 10AM-6PM all year round; pigs, chickens, turkeys, cows, peacocks, goats, horses & more farm animals

Petting Zoos (3)

Metro Boston

- Brooksby Farm 38 Felton Street, Peaboby, MA 508-531-1631; open year round – 8:30-1PM T-F, 8:30-4PM weekend; chickens, geese, sheep, rabbits

San Francisco Bay Area

- Victoria's Fashion Stables 4193 Adobe Rd, Petaluma, CA 707-769-8820; mpz; rabbits, lambs, goats, dogs, cats, geese, horses; riding stables; pony rides
- All Creatures Great and Small Riverside, CA 909-684-2922; mpz;allcreatures@urs2.net; goats, sheeps, potbelly pigs, miniature horses, miniature donkeys, camels, llamas, water buffalo, alpacas, Tibetan yaks, bison, cattle

Data Sharing

- **Existing Data Sources**
  - Discuss the available data and what steps are needed to begin the process to share the information
  - Work out the details as to privacy issues, sharing arrangements of the data source, and other issues
- **New Data Sources**
  - **What data to collect, arrangement with the appropriate institutions**
- **Privacy Protection of the data is very important**
  - **What are the approaches**
  - **How to render the privacy protection to satisfy all concerns when sharing among other developers**

9

How

- **Cooperative efforts among the researchers/developers are needed**
- **Who will act as the focal point for this endeavor**
- **Development of a common data representations for all to use**

10

Meeting Objectives

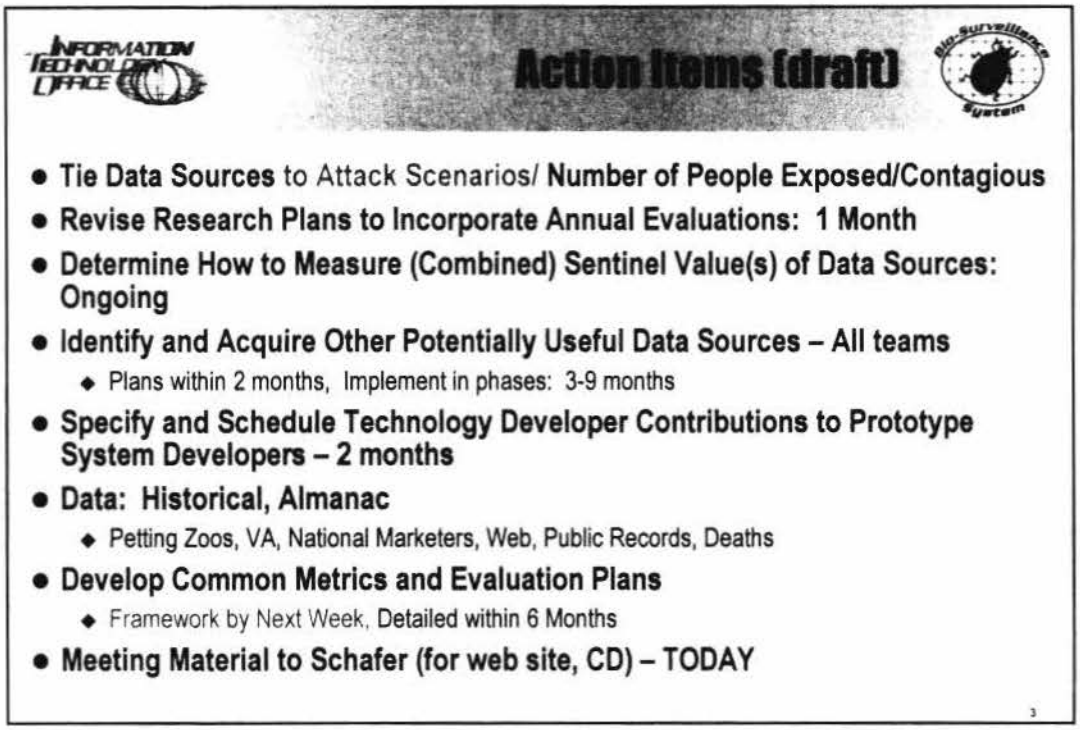
- Ensure Common Understanding of Program Goals, Objectives, Constraints, etc., among all participants and interested parties
- Share Information Between and Among Program Participants about Technical Approach, Program Plans, Validation Strategy, Data Sources, etc.
- Identify Opportunities for Collaboration Between Program Participants
- Provide Information to be used at DARPA Internal Program Review

1

Wrap-Up

- Four Outstanding Interdisciplinary Team
- Need to Re-Emphasize Use of Non-Traditional Data Sources for Early Detection
- Refocus on Military Needs, Data, Transitions, etc.
- Need to Specify Metrics and Experiments, especially with regards to 12-18 month Program Level Go/NoGo Decision
- Key Metric is How Much Earlier Detection Can Occur (at specified performance levels)
- Collaboration Between Teams Enables Best System
  - Sharing of Data, Insights, etc.

2


Action Items (draft)
• Tie Data Sources to Attack Scenarios/ Number of People Exposed/Contagious
• Revise Research Plans to Incorporate Annual Evaluations: 1 Month
• Determine How to Measure (Combined) Sentinel Value(s) of Data Sources: Ongoing
• Identify and Acquire Other Potentially Useful Data Sources – All teams
◆ Plans within 2 months, Implement in phases: 3-9 months
• Specify and Schedule Technology Developer Contributions to Prototype System Developers – 2 months
• Data: Historical, Almanac
◆ Petting Zoos, VA, National Marketers, Web, Public Records, Deaths
• Develop Common Metrics and Evaluation Plans
◆ Framework by Next Week, Detailed within 6 Months
• Meeting Material to Schafer (for web site, CD) – TODAY
3

TITLE PAGE

FOLDER 2.a.1-2.A.9 TOC 5

Briefing Materials concerning TIA BIOALERT 1A

DATA FUSION

- 2.a.1 Data Template
- 2.a.2 Fusion method comparison chart
- 2.a.3 Data Fusion for Biosurveillance
    - IBM T.J.Watson Research Center
    - Johns Hopkins University
    - SDI Inc.
- 2.a.4 Fusion method comparison chart-different data
- 2.a.5 Essence II Data Fusion: A Layered Approach
    - Johns Hopkins Applied Physics Lab
- 2.a.6 Pitt/CMU approaches to Fusion (June 2002)
    - Universities - Pittsburgh/Carnegie Mellon
- 2.a.7 Fusion method comparison chart

Fusion Method	Advantages	Disadvantages	Any Data Type?	Any Venue?	Any Infection Path?	Teams tried it at all?	Tried but little used	Tried and used	Under development
Matrices									
Blackboards									

Data Template

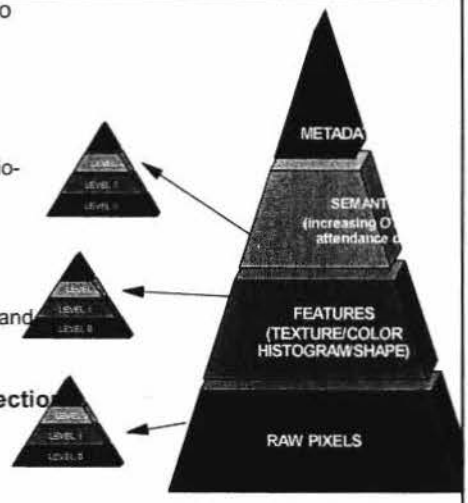
Fusion Method	Advantages	Disadvantages	Any Data Type?	Any Venue?	Any Infection Path?	Teams tried it at all?	Tried but little used	Tried and used	Under development
Spatial-Temporal Scan Statistic	avoids preselection bias	no classification or characterization	any with spatial and temporal identifiers	large or small scale attacks	slow to detect a small, widely dispersed attack			significant experience	
	handles mult. disparate data sources, spatial res.	must recalculate distribution for each run							
	avoids mult. testing bias								
	can adjust for late reporting								
	robust wrt data dropouts & catchups								
Time-domain matched filter	spatial-temporal: handles mult. variables in same spatial grid	covariance matrix requires >= 8 wks data for warmup	any with spatial and temporal identifiers	targeted method applicable on any scale	slow to detect a small, widely dispersed attack			used during 2001 Inaugural; several scenario tests	
	very early detection potential								
	different replicas to identify disease types								
Prewhitened energy detector	spatial-temporal: handles mult. variables in same spatial grid	covariance matrix requires >= 8 wks data for warmup	any with spatial and temporal identifiers	targeted method applicable on any scale	slow to detect a small, widely dispersed attack		test scenario use only		
	requires no disease information	no classification or characterization							
Logistic regression	gives log(odds) of outbreak with multiple data types	at least 4 wks data warmup required	need time series that is not sparse	not for site surveillance; more large scale					untried with real data
	contribution of each data source easily seen								

Data Fusion for Bio-surveillance

IBM T. J. Watson Research Center
Johns Hopkins University
SDI Inc.

Federating and Fusing Heterogeneous Data Sources

- **Spectrum of data continuum**: from data to metadata
  - Raw data
    - time series, audio
    - image, GIS data,
    - video
  - Features (through temporal, spatial, or spatio-temporal abstraction)
    - Time series (Badge, Phone, proxy, cafeteria)
    - GIS (SDI, environmental)
    - Audio
    - Video
  - Events /Alerts (additional temporal, spatial, and spatio-temporal annotation and abstraction)
  - Metadata
- **Spectrum of integration, fusion and detection continuum**
  - **Syntactic federation**: access and fuse incompatible schema and data models.
  - **Semantic federation**: access and fuse incompatible ontology (and PSM). e.g. definition of gold standard

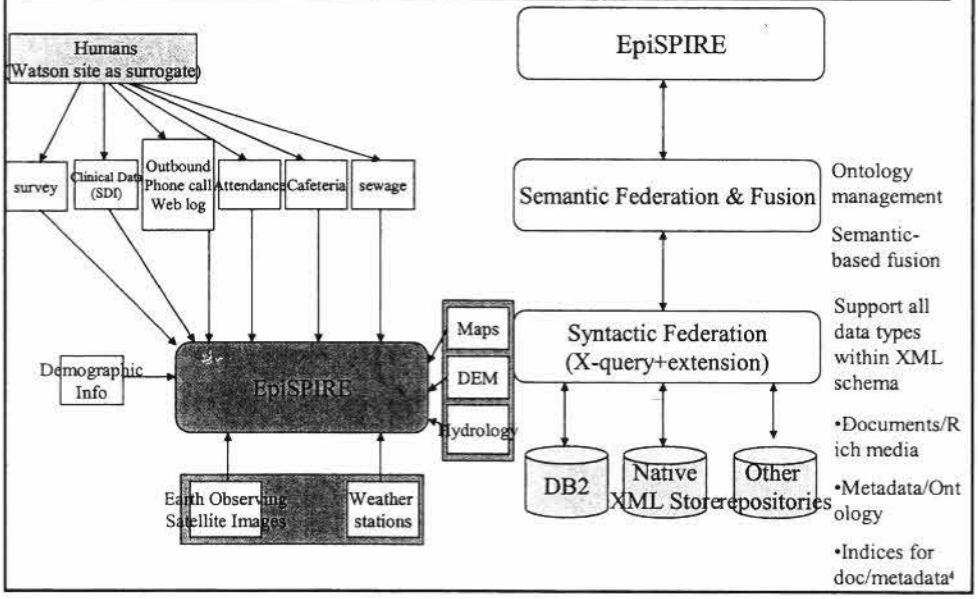


Scenarios of Data Fusion

- Different spatial and/or temporal resolution
  - Environmental data: from satellite images (from different instruments) and weather station
- Different modalities
  - Cough/sneeze counter: audio+video
- Different data sources
  - Phone+proxy
  - Phone+proxy+OTC

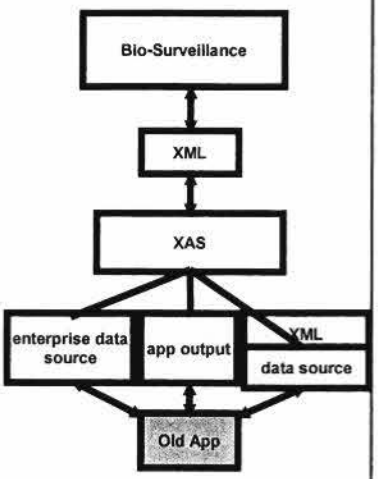
3

Semantic and Syntactic Federation: The Big Picture



Syntactic Federation: XAS Overview

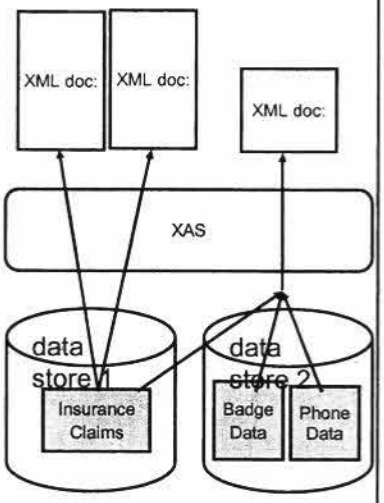
- Objective:
  - ƒ Enable XML based processing capability for enterprise e-business applications
  - ƒ While data remain in underlying enterprise data sources in native format
    - –DB2, IMS, VSAM, CICS, SAP, APP output
- How it works
  - ƒ Old applications continue to run as usual
  - ƒ New apps access data as XML docs
    - –don't care where data come from, how they are mapped
    - –docs may consist of logically related parts in multiple sources
    - –XML on demand
      - •no copying, no conversion preprocessing



5

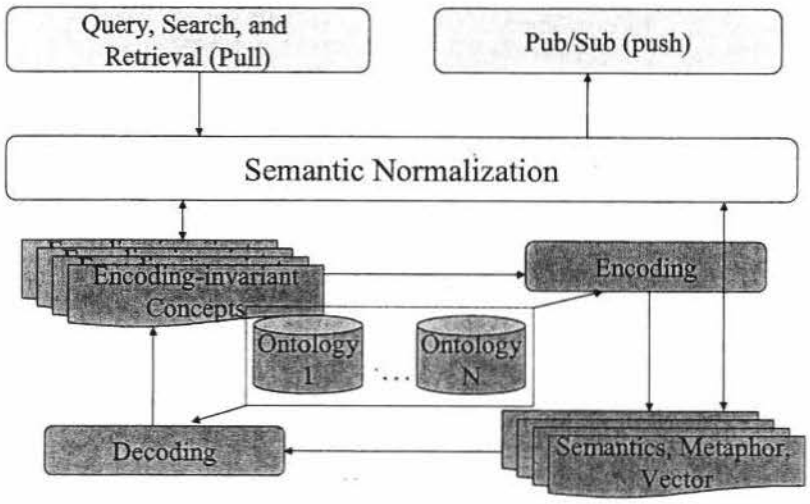
What is XAS: Key Features

- Representing XML documents according to what's useful, not how data are stored
- XML document is a view of underlying schema
  - ƒ Many to one mapping:
    - – present data in multiple different data sources as one XML document
    - ➔data integration
  - ƒ One to many mapping:
    - – A single piece of data can participate in many XML documents
    - ➔application integration
    - ➔personalization



6

Ontology Management: An approach for Semantic Federation



7

Near-real-time Weather Data

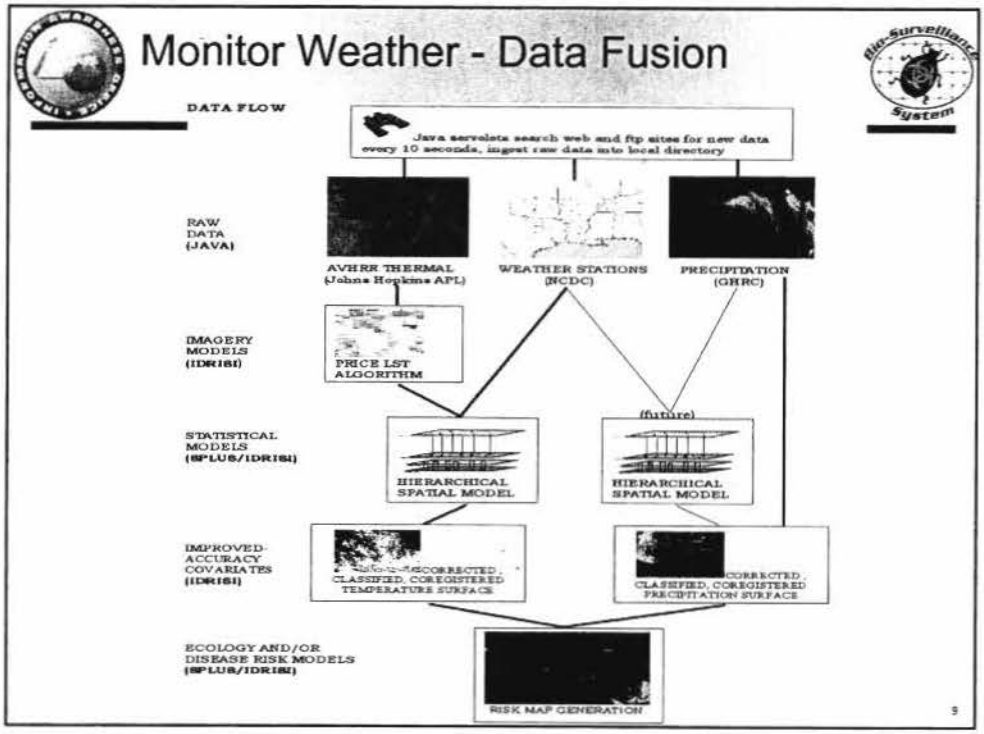
Continuous monitoring of environment variables

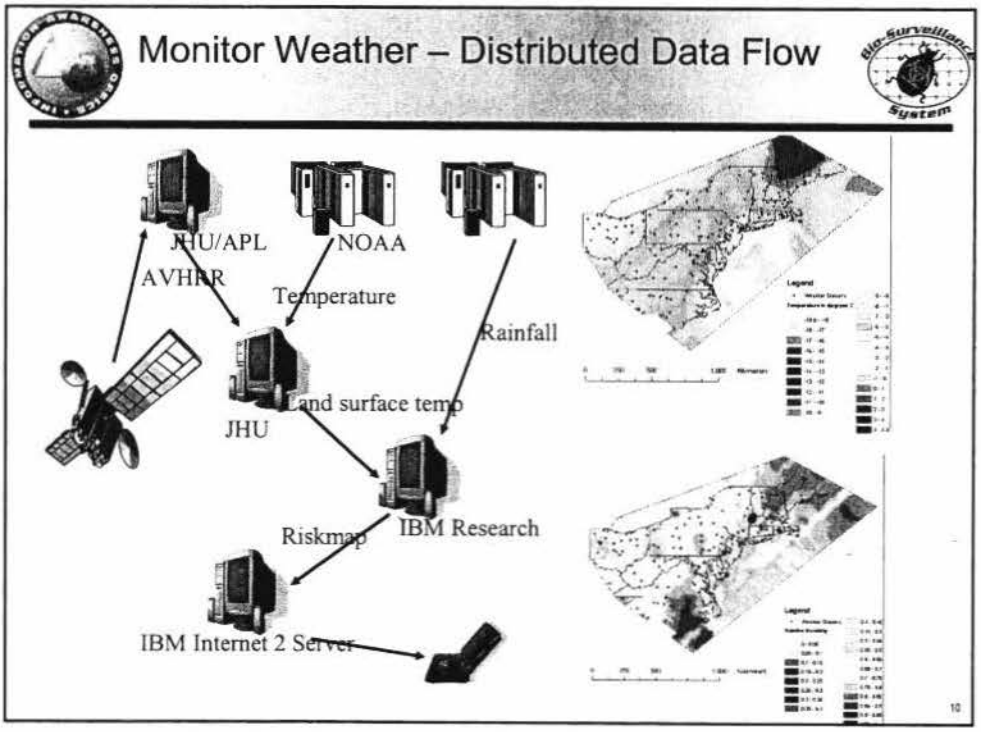
Air temperature
Relative humidity
Precipitation
Wind speed
Wind direction

Derived from station readings and satellite images for better resolution

Source: NOAA National Climate Data Center

8


Monitor Weather - Data Fusion
DATA FLOW
Java servelets search web and ftp sites for new data every 10 seconds, ingest raw data into local directory
RAW DATA (JAVA)
AVHRR THERMAL (Johns Hopkins APL)
WEATHER STATIONS (NCDC)
PRECIPITATION (GHRC)
IMAGERY MODELS (IDRISI)
PRICE LST ALGORITHM
STATISTICAL MODELS (SPLUS/IDRISI)
(future)
HIERARCHICAL SPATIAL MODEL
HIERARCHICAL SPATIAL MODEL
IMPROVED ACCURACY COVARIATES (IDRISI)
CLASSIFIED, COREGISTERED TEMPERATURE SURFACE
CLASSIFIED, COREGISTERED PRECIPITATION SURFACE
ECOLOGY AND/OR DISEASE RISK MODELS (SPLUS/IDRISI)
RISK MAP GENERATION
9

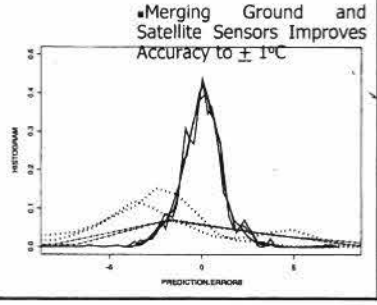

Monitor Weather – Distributed Data Flow
JHU/APL
NOAA
AVHRR
Temperature
Rainfall
Land surface temp
JHU
IBM Research
Riskmap
IBM Internet 2 Server
10

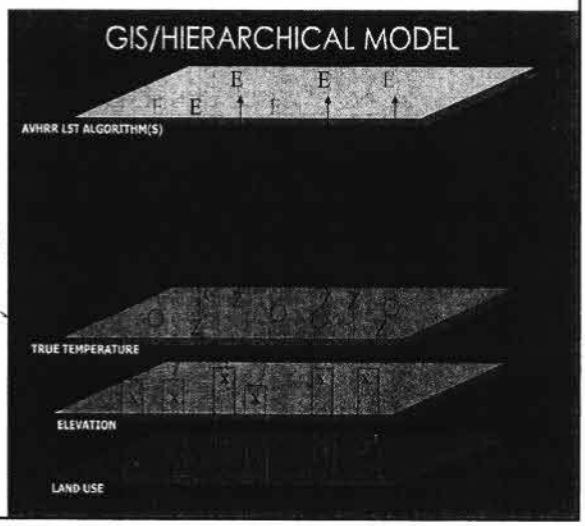
Ground surface temperature

- **Objective**:
  - Want to know surface temperature at high spatial resolution over large geographic region on a daily basis.
  - Applications
    - 'Urban heat islands & mortality'
    - Agriculture
- **Challenges**
  - Ground based sensors are sparsely distributed
  - Southern U.S. 1,000,000 sq km has 308 cooperative weather stations (hourly weather data)
- **Solutions**
  - Price – Reflectance from body is related to temperature. Satellites record reflected light as pass over objects –Split window algorithm
  - AVHRR sensor – 2X/day; 1 km resolution
    - Subproblem – relationship well known if perfect black body –emmisivity alters effect

Hierarchical modeling

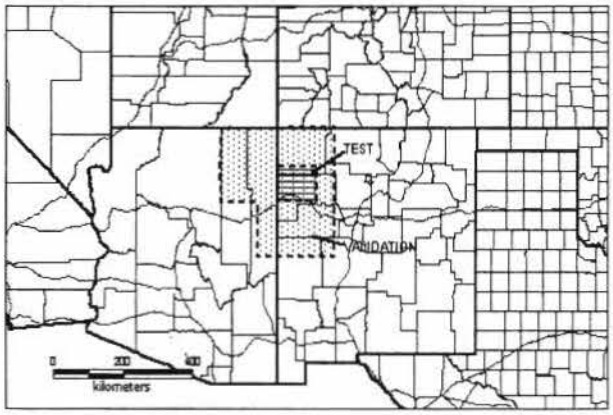
- Use weather station data to anchor estimates, characterize spatial correlation among sites and error propagation
- Can incorporate factors known/thought to affect temperature (elevation, land cover, etc as covariates)





Examples in Risk Map Generation for HPS

- 105,200 sq km
- Subdivide
  - Test
  - Validation
- Includes all cases from 1993outbreak

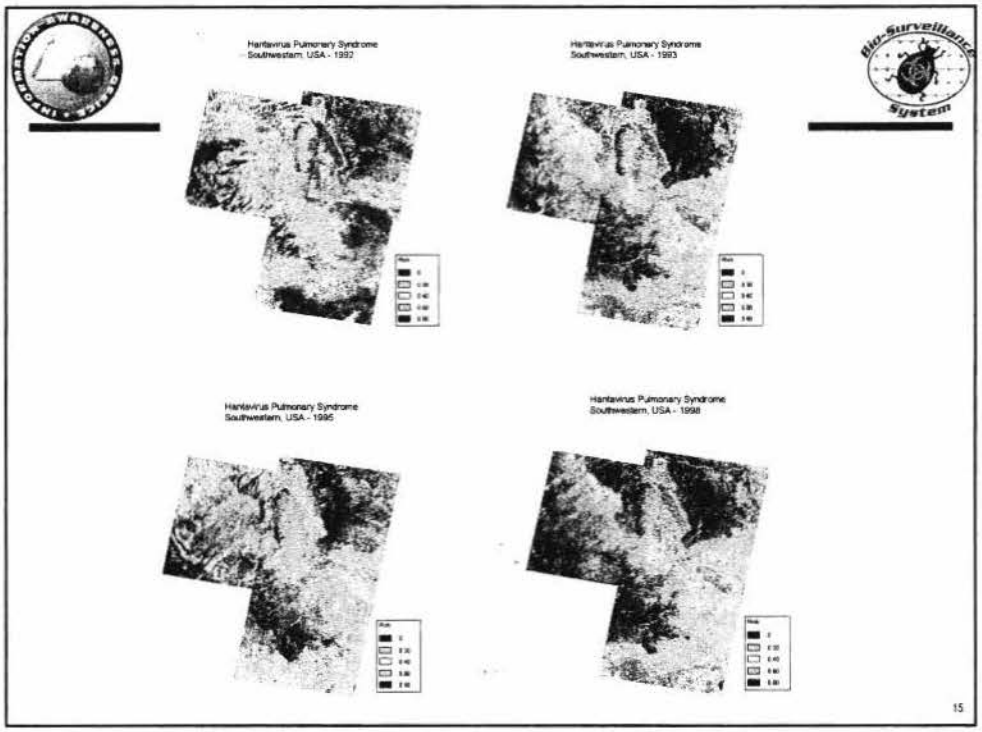


13

Epidemiologic Results Show Association of HPS Risk and Satellite Imagery & Geographic Consistency

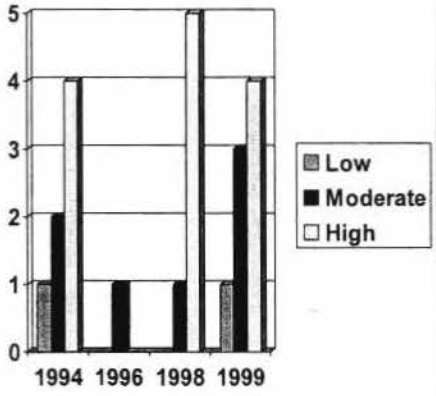
Risk Factor	Test Area OR (95% CI)	Remaining Area OR (95% CI)
TM 1	0.94 (0.89, 0.99)	0.95 (0.90, 0.99)
TM 5	1.14 (1.03, 1.26)	1.08 (1.01, 1.17)
TM 7	0.85 (0.73, 0.99)	0.90 (0.80, 1.00)
Elev >2094m	4.63 (0.88, 24.29)	2.56 (0.77, 8.43)

14



HPS Risk in Later Years is Consistent with Predicted Risk

- **The numbers of HPS cases were consistent with the extent of moderate/high risk**
- **Cases of HPS were primarily (91%) in predicted moderate – high risk areas**



16

Spatial analysis & modeling

- Use spatial distribution of environmental features to model exposures/risk
- GIS is traditionally weak in spatially explicit modeling (Geographers curse)
  - Query
  - Export
  - Spatial Statistics/Models
  - Import for visualization
- Solution requires integrating additional technologies
  - Remote sensing (satellite imagery)
  - Spatially explicit modeling (spatial covariance of observations)
    - Variance-covariance matrix of outcome has covariance between any two points modeled as a function of distance

17

Fusion/Detection Taxonomy

- *Probabilistic models*:
  - Clustering techniques (K-means), Graphical models (HMM,DBN)
- *Linear predictive techniques*:
  - AR/ARMA models, time-varying versions e.g., Kalman filters
- *Supervised or unsupervised* techniques may be used
- *Fusion Strategies*:
  - Low-level feature fusion
    - Concatenate the multi-modal channel features
  - High-level decision fusion
    - Combine detection based on each channel using operators such as "OR"ing "AND"ing etc.

18

Unsupervised Incremental Clustering

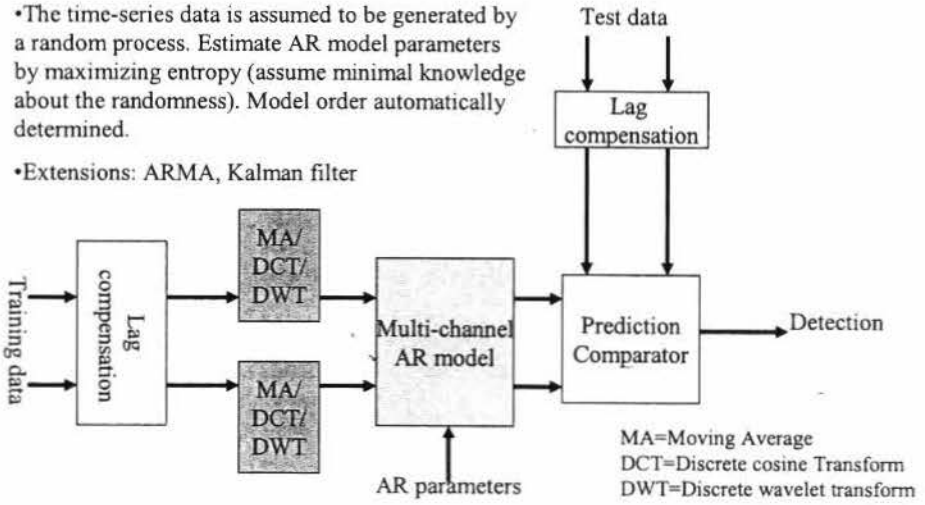
- A statistical algorithm (variant of *incremental k-means*).
  - Start with N clusters one of which is seeded with high values of source features.
  - Update clusters based on data source values observed till date.
  - Choose that cluster as abnormal which has largest norm (L1).
- Characteristics.
  - No use of prior or historical data of normal behavior or other outbreak behavior.
  - No online supervision for detecting abnormalities either.
- Extensions.
  - Markov chain based smoothing.
  - Historical normal data to develop robust "normal clusters."

19

Auto-regressive Prediction-Based Model

•The time-series data is assumed to be generated by a random process. Estimate AR model parameters by maximizing entropy (assume minimal knowledge about the randomness). Model order automatically determined.

•Extensions: ARMA, Kalman filter



20

Latency, Timeliness, and Spatial/Temporal Coverage

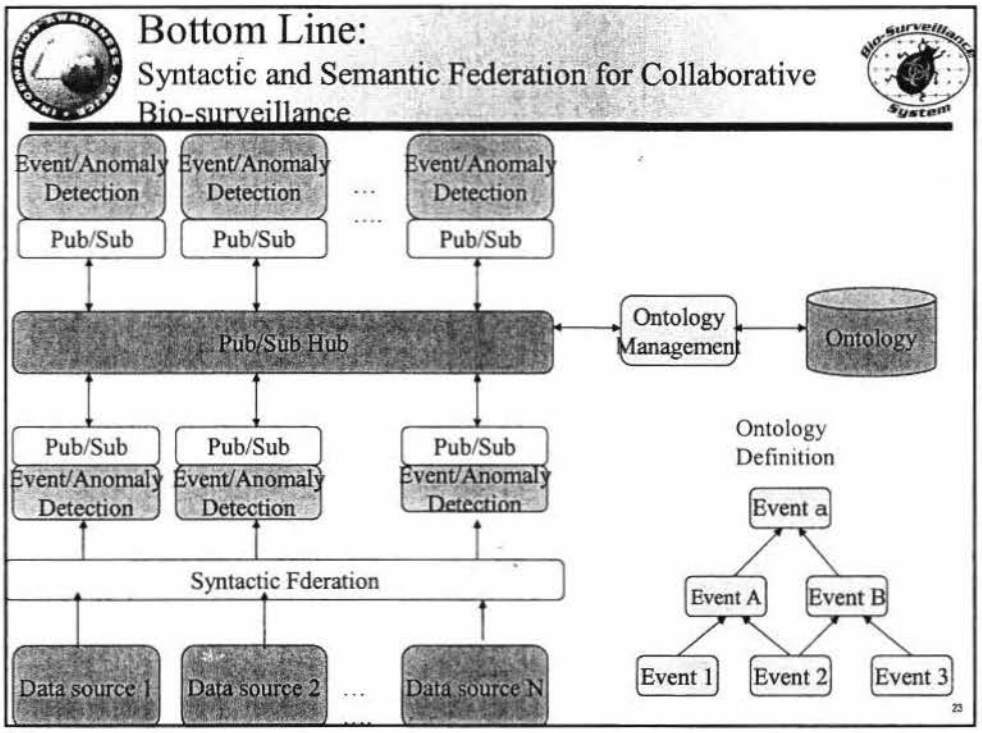
- Environmental data has improved temporal and spatial resolution as a result of fusing satellite images and weather station data
- Other tracking data set
  - Ideally, a data set will track the behavior of the population with full spatial and temporal coverage
  - However, due to the site limitation of the site specific data
    - Cafeteria beverage data does not have Monday coverage (Monday free) in additional to weekends
    - Survey data does not have weekend coverage
    - Badge, phone, proxy, and sewage has very little weekend coverage of the population
  - Fusion with regional data such as OTC, claims, etc. will be unavoidable to achieve full coverage

21

Objectives and Performance Goals: Evaluation Methodology

- ***In situ (in vivo):***
  - Validating behavior model for various event predictors, alert generator and active learning algorithms based on Watson data sources
  - Validate baseline respiratory model using Surveillance data
- ***In vitro:***
  - Validate performance of the anomaly detectors through the use of emulated attack scenarios superimposed on top of collected data
- Metric
  - Early determination of disease outbreak
    - Day(s) or hour(s) from alert generation to actual outbreak
  - Early determination of attack
    - Day(s) or hour(s) after emulated attack has occurred
  - Sensitivity/Specificity
    - Spatial resolution: location of the outbreak
    - Temporal resolution: instant of the outbreak
    - False alarm/miss ratio

22


Bottom Line:
Syntactic and Semantic Federation for Collaborative Bio-surveillance
Event/Anomaly Detection
Event/Anomaly Detection
Event/Anomaly Detection
Pub/Sub
Pub/Sub
Pub/Sub
Pub/Sub Hub
Ontology Management
Ontology
Pub/Sub
Pub/Sub
Pub/Sub
Event/Anomaly Detection
Event/Anomaly Detection
Event/Anomaly Detection
Ontology Definition
Event a
Event A
Event B
Event 1
Event 2
Event 3
Syntactic Fderation
Data source 1
Data source 2
Data source N
23

Fusion Method	Advantages	Disadvantages	Any Data Type?	Any Venue?	Any Infection Path?	Teams tried it at all?	Tried but little used	Tried and used	Under development
General Methods									
Auto-Regressive Moving Average (ARMA)	Well understood	Needs historical data to train	Yes	Yes	Yes				
Incremental temporal clustering	Unsupervised	False alarm rate may be high without some supervision	Yes	Yes	Yes				
Multi-dimensional demographic variation	Detects deviation from demographic profiles in naturally occuring disease outbreaks	Attack must have distinctive profile	Yes	Yes	Yes				
Fusion Techniques									
Syntactic Federation	Access and fuse incompatible schema and data models								
Semantic Federation	Access and fuse incompatible ontologies								

2.a.5

Essence II Data Fusion: A Layered Approach

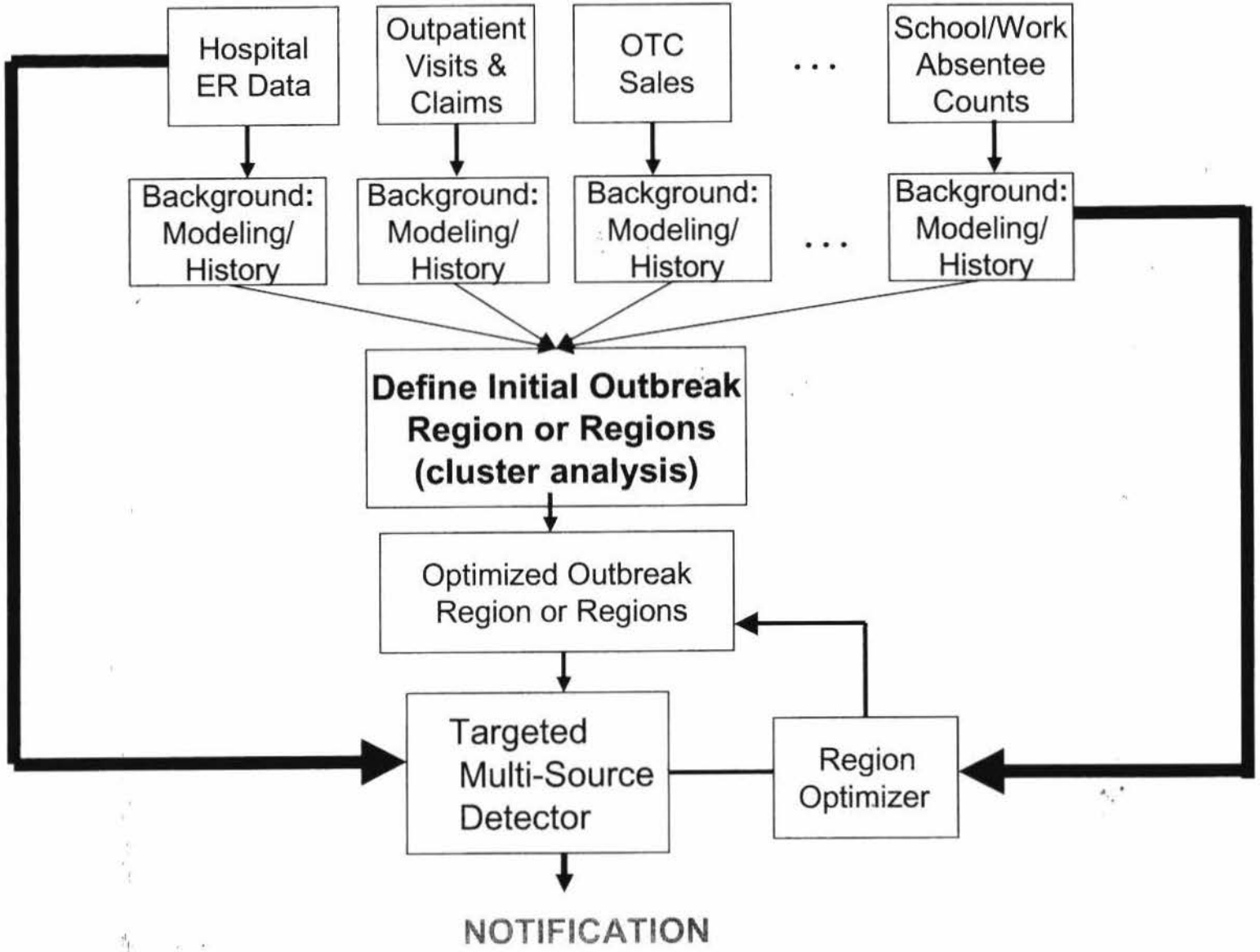
BSS PI Meeting

June 4-6, 2002

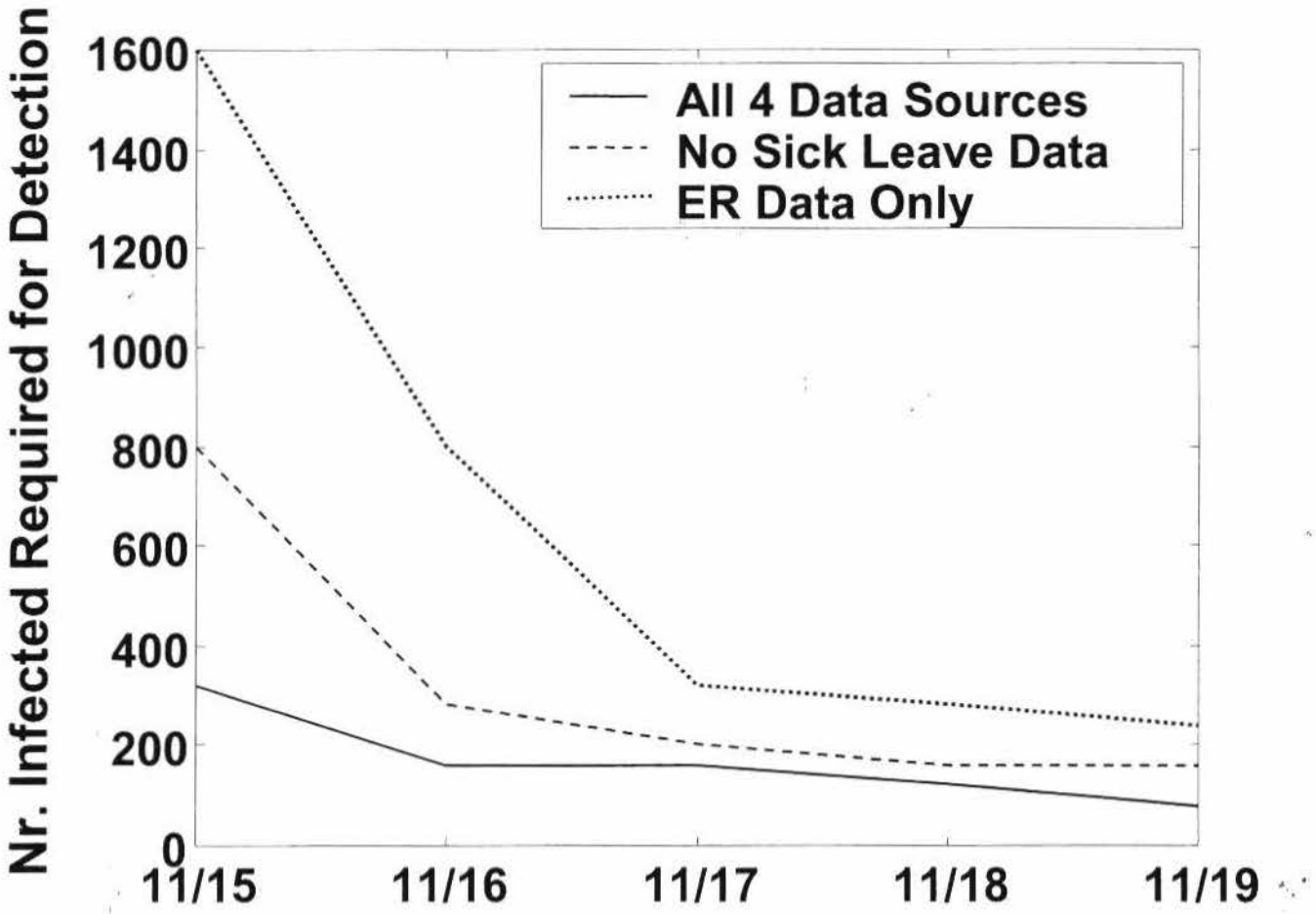
Howard S.Burkom

Johns Hopkins Applied Physics Laboratory

Outbreak Detection Using Multiple Data Sources



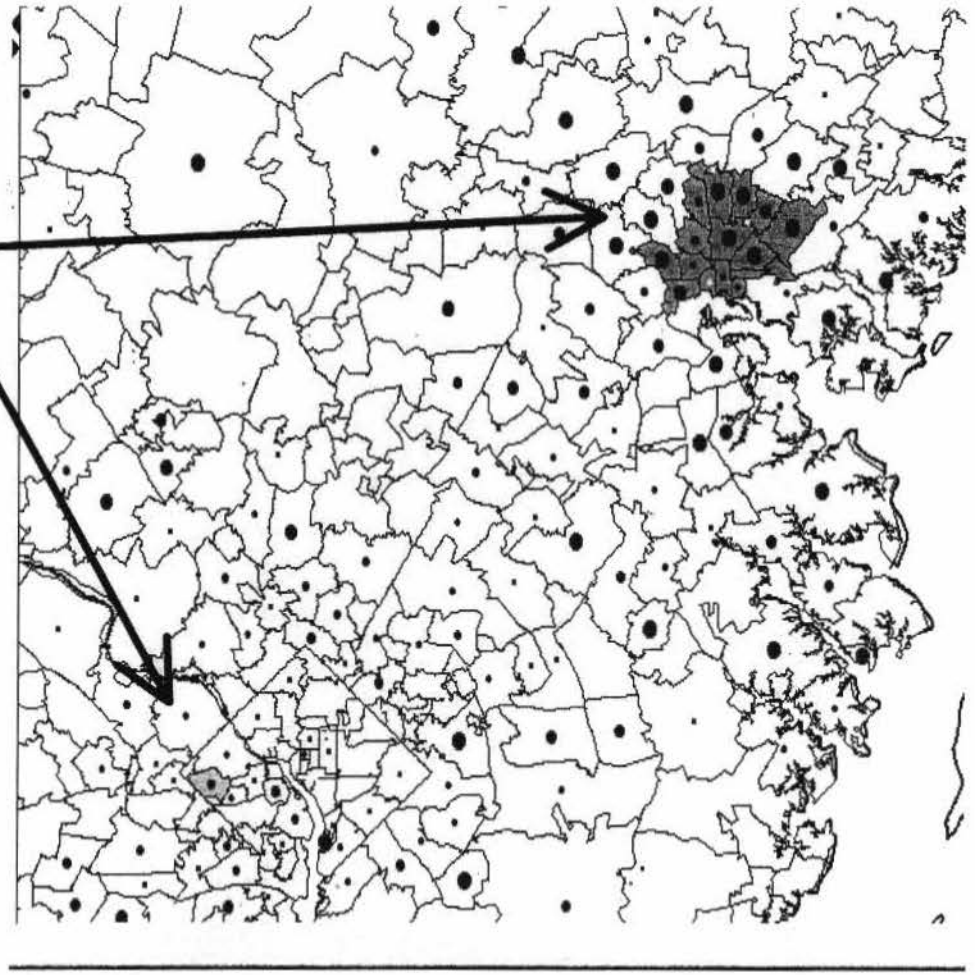
Scenario Results:
Detector Performance vs Nr. Infected



Clustering Based on Scan Statistics

- Finds clusters of maximum likelihood ratio (LR): those most significant given observed (O) and expected (E) spatial distribution
- Kulldorff formulation:

  $LR=(O/E)^{O}_{in}\,(O/E)^{O}_{out}$
- "Expected" count derived from modeling, past history, population, # eligible,…



Kulldorff Scan Statistic: Application to Biosurveillance Data

- ICD9-Coded data (military, civilian)
  - Critical choice of "outcome variable": code sets
  - Use of proportions to reduce late reporting prob.
- Over-the-counter sales
  - History, modeling work well, with guard band
- School absenteeism
  - Using AR modeling, regional averages
  - "Noisy" behavior, some data dubious

How to combine disparate scales, variances, spatial distribution?

Approach: Generalized Scan Statistic

- Simulation testbed with pluggable statistic
- Attempted so far:
  - "adjusted" Kulldorff (conventional epi use)
  - "stratified" Kulldorff
  - Odds ratio (Mantel-Haenszel adj. for multisource)
  - asymptotic χ^2
- Can complement other detection methods
- Applied to multiple, disparate data sources
- Promising early alerting results obtained from Monte Carlo simulations

Scan Statistic: Advantages

- Avoids preselection bias wrt location OR extent
- Modest increases in multiple sources can yield early detection
- Flexible: data streams may vary by:
  - Location (zipcode, hospital/provider site, patient/customer residence, school/store address)
  - Time windows used (cases, history, guardband)
  - Background estimation method: model, history, population, eligible customers
  - Representation of outcome variable
- Robust wrt data availability

To avoid overfitting!

For better spec./sens.

Data Fusion Scenario

- Authentic data for:
  - Civilian & military medical claims (by patient zip)
  - OTC antiflu sales (by store location)
  - Absenteeism: 2 counties (by school location)
- Outbreak assumptions: 500 symptomatic cases
  - Smallpox-like epicurve (~ 1 week prodromal period)
  - Site placed randomly in school counties: 100 trials
  - Spatial distribution of cases: gradual exponential falloff from site, with 75% of infected within 10 miles
  - data epicurves formed using presumed transfer rates and delays specific to data sources

Preliminary Scenario Results:
Adjusted Kulldorff Method

 # times/100 that simulated outbreak site in significant cluster:

Day after exposure	**5** **(5% of acute)**	**6** **(11%)**	**7** **(19%)**	**8** **(31%)**
ICD9 only	0	4	72	96
OTC only	23	37	58	94
School only	5	39	65	72
All 3	42	66	77	91

Scan Statistic: Issues

- Distribution computed empirically each day
  - Will be more computationally intensive as # sources and coverage grow
  - Expected efficiency from k-d tree approach supplied by J. Schneider & A. Moore of CMU
- Ideal location data difficult to obtain (would like work & residence sites for each patient/customer)
  - Spatial clusters are approximate
  - Can serve as launch point for methods requiring more resources or more unusual indicators
- Ongoing monitoring of data sources needed to maintain sensitivity/specificity

Targeted Data Fusion Approaches

- Time-Domain Matched Filter
  - Uses expected epicurve to form replica for data matching
  - Used in Inaugural 2001 surveillance, applied to several scenarios
  - ROC analysis => capability for 2-3 day alerting advantage
- Prewhitened Energy Detector
  - Based on model residuals "normalized" for different data sources with spatial covariance matrix
  - Not disease-model-dependent
  - Limited results from scenario testing
- Logistic Regression
  - Computed log(odds) of outbreak based on regression using residuals
  - Under development

Pitt/CMU approaches to Fusion: June 2002

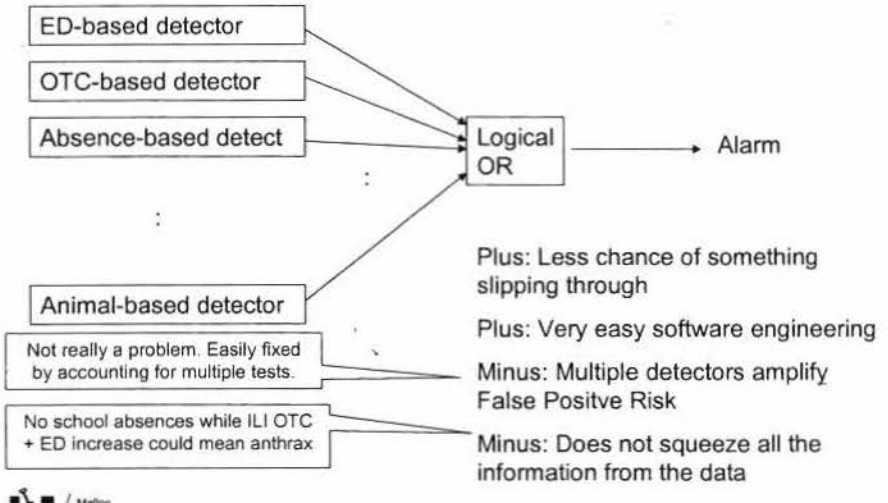
Center for Biomedical Informatics
University of Pittsburgh and Carnegie Mellon University

Greg Cooper (U Pitt, CBMI)
Jeremy Espino (U Pitt, CBMI)
Andrew Moore (CMU, Comp Sci)
Bob Olszewski (U Pitt, CBMI)
Jeff Schneider (CMU, Robotics)
Rich Tsui (U Pitt, CBMI)
Mike Wagner (U Pitt, CBMI)
Weng-Keen Wong (CMU Comp Sci)

Contact:
Mike Wagner
mmw@cbmi.upmc.edu

1

Idea One: A Zoo of detectors



2

Idea Two: Multivariate Kalman Filter

- Assume joint time dynamics for all signals are linear with Full-Covariance Gaussian noise

- Plus = Reasonable assumptions
- Minus = What if assumptions wrong?
- Minus = 10000 state variables if try to include all combinations of space / time /demographics

- Overall: We think this is a good idea though we haven't used this approach so far.

Pitt Mellon

3

Idea Three: (Dynamic) Bayesian Network

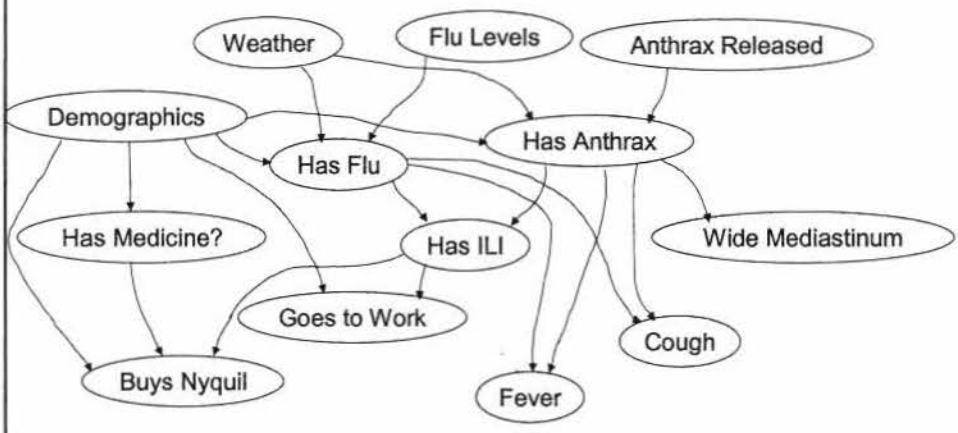
We are exploring several structures:

- Fusing Individual and Population Based (PANDA 3.0)
- Temporal-Spatial Aggregates (WSARE 3.0)

Pitt Mellon

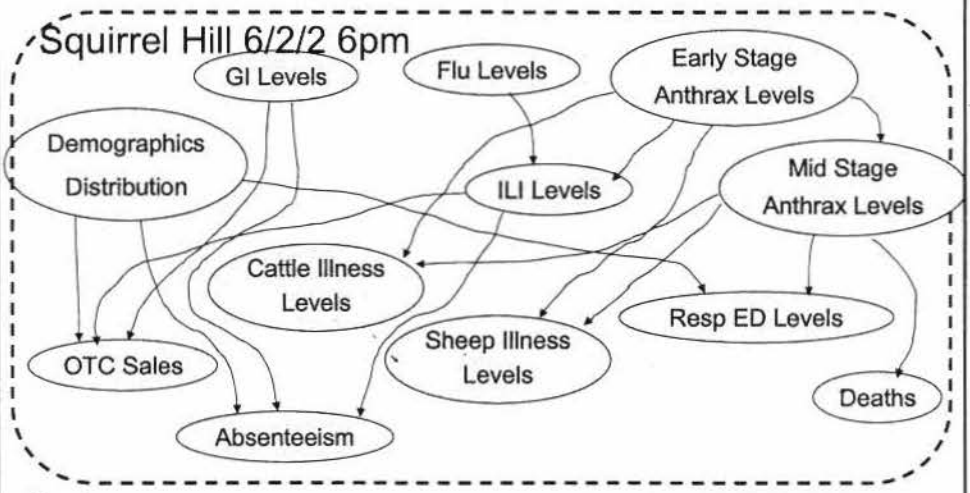
4

Patient-Based Example of Belief Net

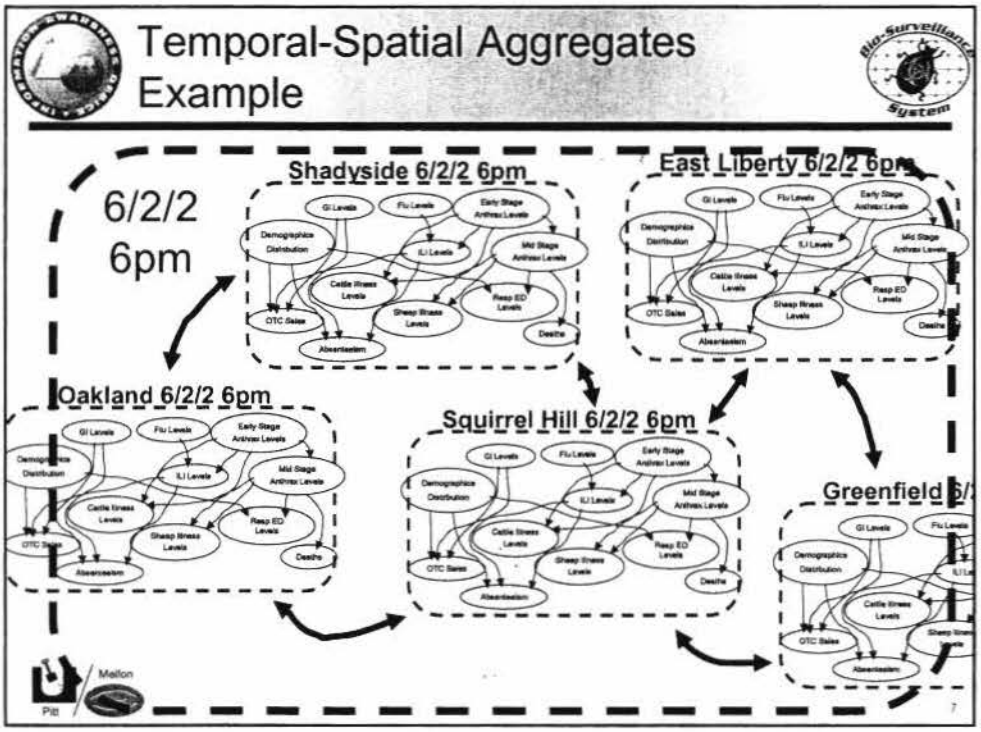


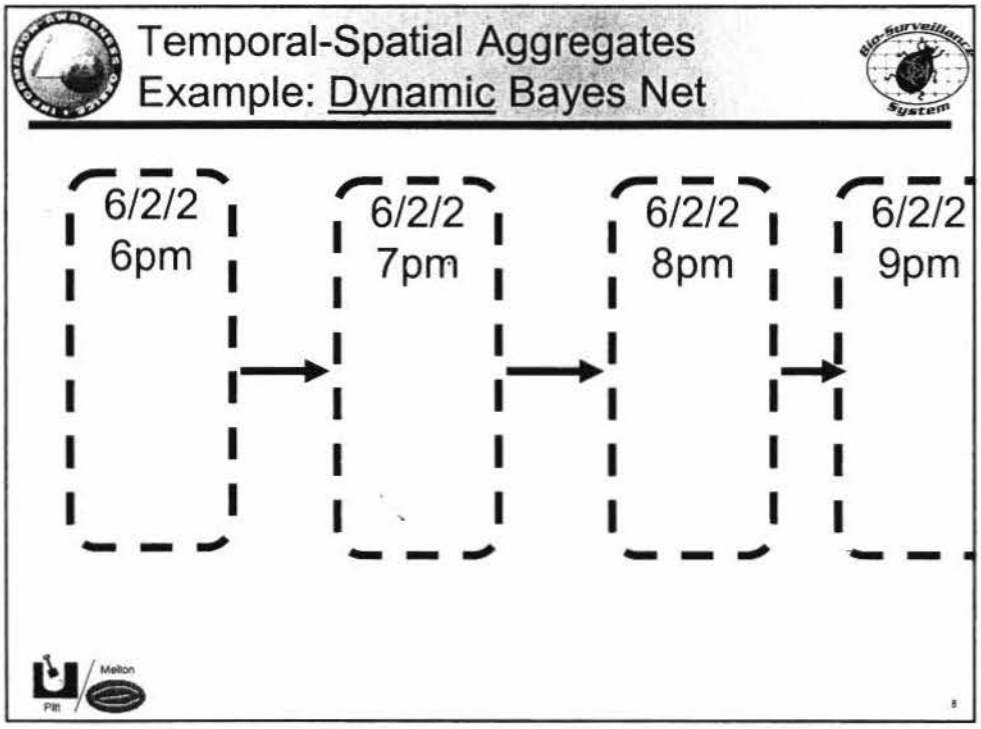
5

Temporal-Spatial Aggregates Example

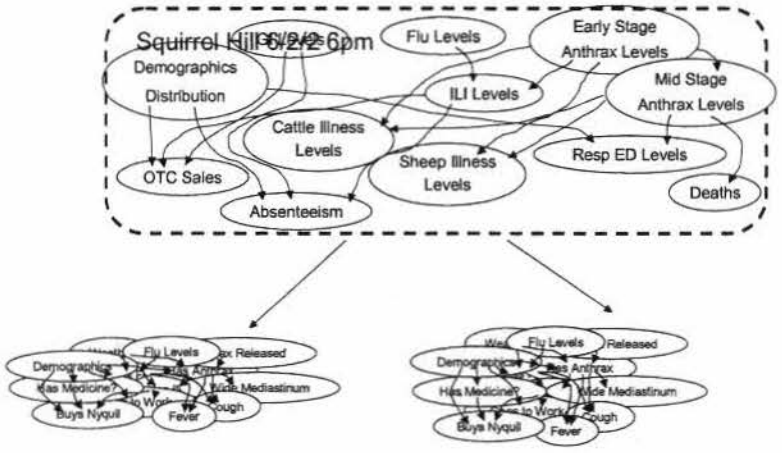


6


Temporal-Spatial Aggregates Example
6/2/2 6pm
Shadyside 6/2/2 6pm
East Liberty 6/2/2 6pm
Oakland 6/2/2 6pm
Squirrel Hill 6/2/2 6pm
Greenfield 6/
Mellon
Pitt


Temporal-Spatial Aggregates Example: Dynamic Bayes Net
6/2/2 6pm
6/2/2 7pm
6/2/2 8pm
6/2/2 9pm
Mellon
Pitt

Fusing Individual and Population Models



Patient 1 Patient 2

9

Answers to Specific Questions

1. How does the process prevent overfitting of the data: **Machine learning will involve leave-n-out cross validation**

2. How does the process handle sporadic missing data: **A strength of belief networks is handling missing data**

3. How does the process handle overconfidence (double counting): **It is possible to model all influences**

4. How does the system combine processing results from dissimilar data sets: **A strength of belief networks**

5. How does the system combine detection results from the non-traditional data sources and syndrome classification results from the more traditional data sources: **A strength of belief networks**

6. What is the technique to evaluate spatial correlation over a large number of data sources. **By including space as a variable index (i.e., spatial belief networks)**

7. How does the system temporally correlate widely distributed (but common) sources such as school absentee over a large number of schools and also take into account the spatial relationships between the data sources. **By including time as a variable index (i.e., dynamic belief networks)**

10

2007

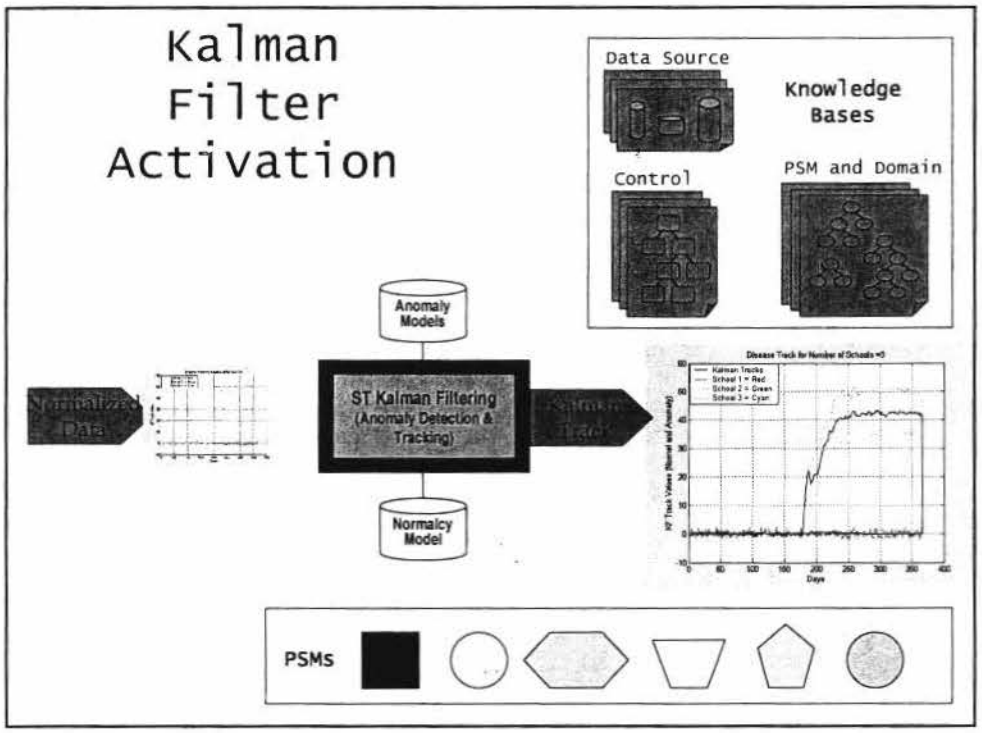
Fusion Method	Advantages	Disadvantages	Any Data Type?	Any Venue?	Any Infection Path?	Teams tried it at all?	Tried but little used	Tried and used	Under development
General Methods									
Statistical	Well defined math	Still have to decide on right method				Yes		Yes	Yes
Auto-Regressive Moving Average (ARMA)									
Stochastic Processing									
Matrices									
Temporal Correlators									
Spatial Correlators (Association)									
Temporal Clutering Methods									
Spatial Clustering Methods									
Knowledge Based Methods									
Blackboards		Not in this domain where it's all about the noise							
Expert Systems		Not in this domain where it's all about the noise							
Evidential Reasoning		Not in this domain where it's all about the noise							
Baysian Networks	Can be conceptually and proactically simple	Can require a massive event space				yes		yes	yes
Classification Clustering									
Fusion Techniques									
Kalman Filters									
Fusion of Similar Data									
Fusion of Dissimilar Data									
Centralized Fusion									
Distributed Fusion									
Anomaly Detectors									
Syndromic Classification Detectors	Simple	Simple				yes		yes	yes
Learning Techniques									
Neural Networks									
Hidden Markov Models	Good semantics for time, and fairly good for space	We like them for spatial fusion only				yes		no	yes
Optimization Techniques									
Maximum A-Posteriori									
Maximum Liklihood									
Least Squares									
Weighted Average									
Processing Techniques									
Batch Processing									
Sequential Processing									
Measures of Performance									

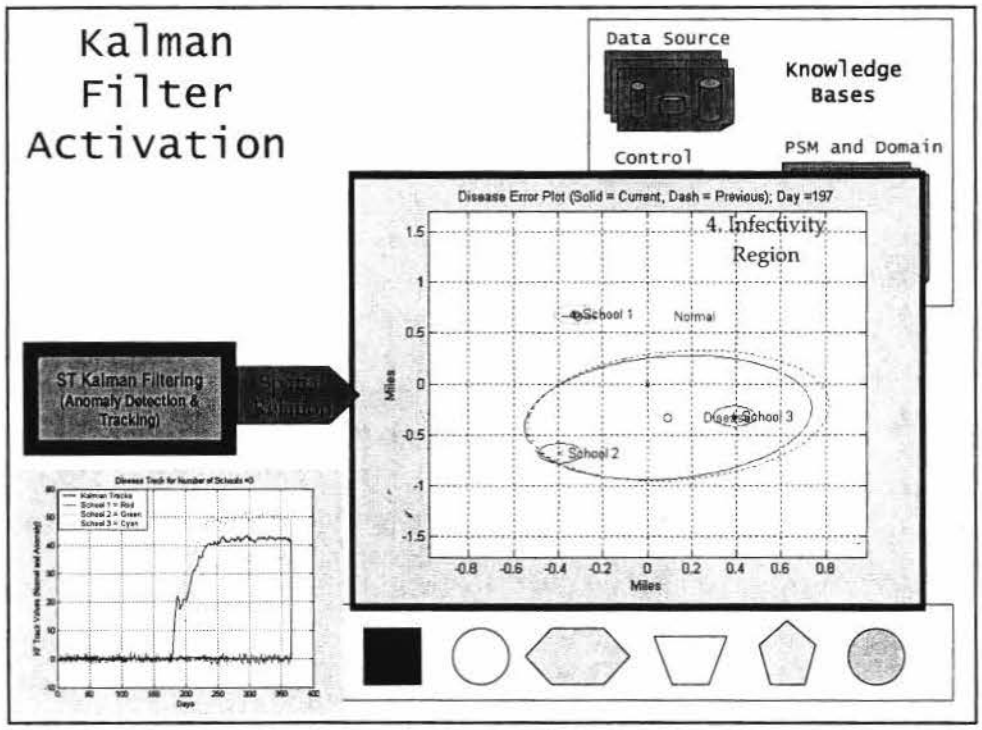
Data Fusion

Veridian-Stanford Team

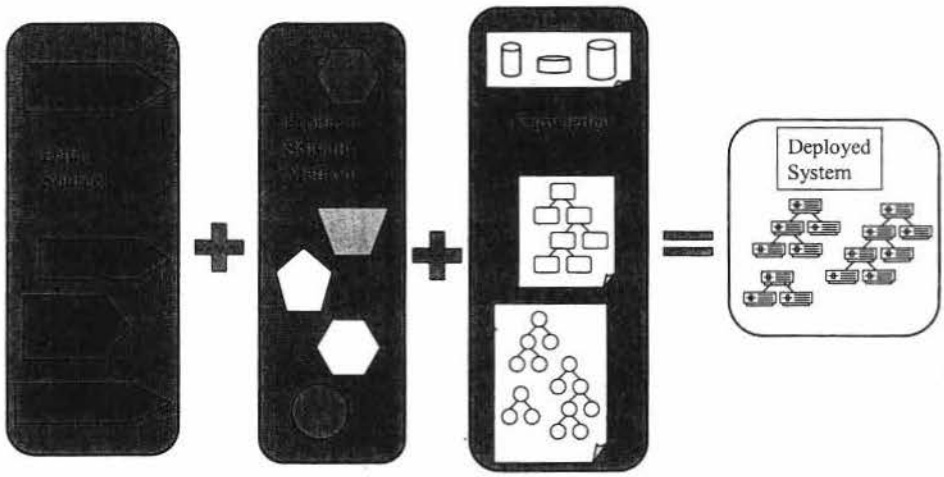
Coupling of Fusion Techniques with Knowledge Based Temporal Processing

- Statistical techniques normalize the data
- Spatio-Temporal Kalman filters (fusion of similar sources) provides:
  - Early **anomaly detections**
  - Early **spatial temporal processing**
- Both treated simply as Problem Solving Methods (PSM) within the Knowledge-based infrastructure
- Knowledge-based process provides situation specific models that characterize the results of the fused detections
- Provides foundation for higher level fusion
  - Information fusion
  - Multiple hypotheses

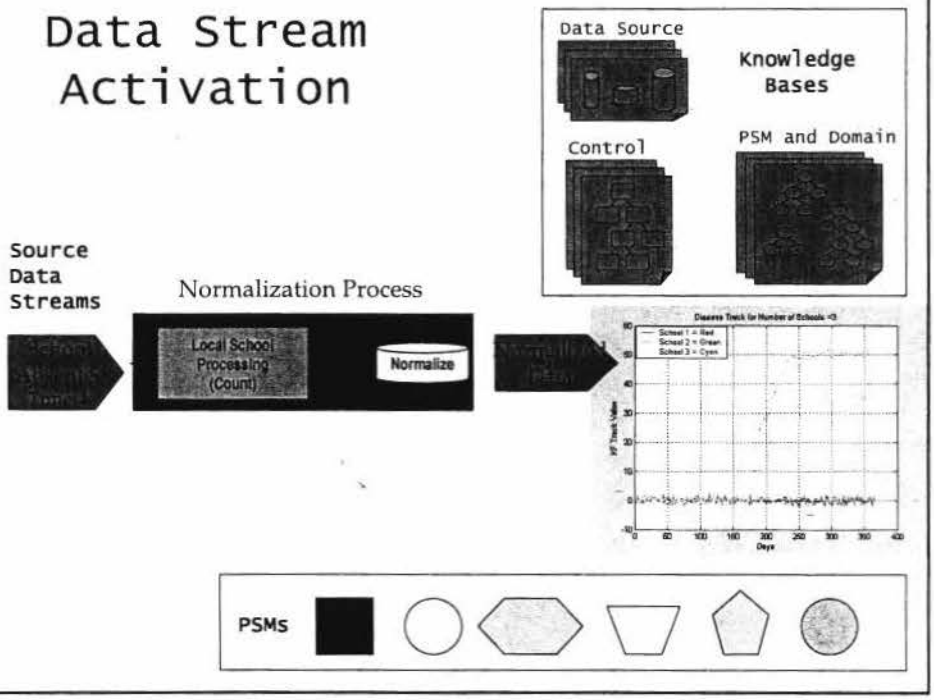

Kalman Filter Activation
Data Source
Knowledge Bases
Control
PSM and Domain
Anomaly Models
ST Kalman Filtering (Anomaly Detection & Tracking)
Normalcy Model
PSMs


Kalman Filter Activation
Data Source
Knowledge Bases
Control
PSM and Domain
ST Kalman Filtering (Anomaly Detection & Tracking)
Disease Error Plot (Solid = Current, Dash = Previous); Day =197
4. Infectivity Region
School 1
Normal
School 3
School 2
Miles
Miles

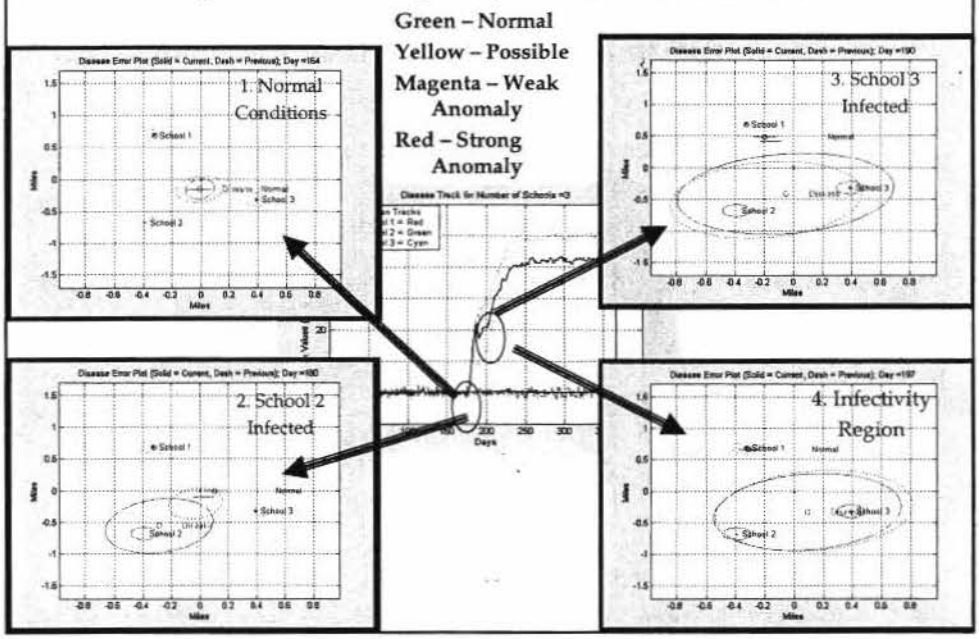
The Problem: Integrating Data Source, and Problem Solving Methods within a Domain Knowledge Framework



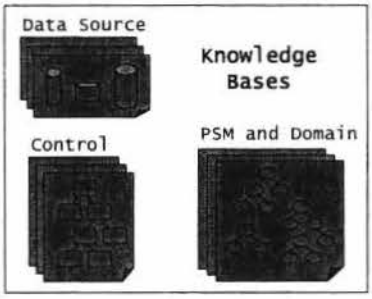
Data Stream Activation



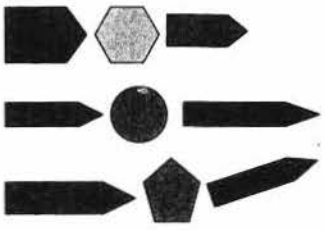
Kalman Filter Effects (Schools) on Spatial Clustering: Days 164, 180, 190, 197



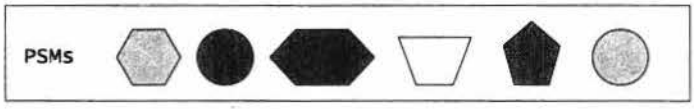
Knowledge-Based Abstraction



Source Processing Streams



- Knowledge-Based Temporal Abstraction
- Information Fusion
- Multiple Hypotheses



Traditional Kalman Filtering vs. Spatial-Temporal Kalman Filtering

- **Traditional Kalman Filtering**
  - **State function** jointly evaluates stochastic model (state model) with the statistical model (observation model)
  - **Gain function** adjusts dependency between system and observation models
  - Multiple models create means to monitor different conditions simultaneously
  - Adaptive estimation of background makes specification of background models less critical (Normalcy).
  - Shorter normalcy baseline needed than for classical regression or time-series techniques (Steady-state).
  - Naturally handles gaps in data.
- **Spatial-Temporal Kalman Filter**
  - Retains all advantages of traditional Kalman filter
  - Allows multiple data sources to contribute to simultaneous adjustment of state model (Anomaly)
  - Evaluates each source contribution at each time update (Adaptive weights)
  - Jointly calculates source effects on both normalcy and anomaly models
    - Calculates degree of anomaly for each source
  - Adds spatial correlation as a natural by-product
  - Spatially maps progress of anomaly over time

Veridian Approach Employs the Following Technologies

- Spatial-Temporal Kalman Filter
  - General Methods
    - Statistical
    - Stochastic
    - Matrices
    - Spatial–Temporal Correlators
    - Spatial–Temporal Clustering
  - Fusion Techniques
    - Fusion of Similar Data
    - Anomaly Detectors
    - Syndromic Classification Detectors
  - Optimization Techniques
    - Minimum Mean Square Error
  - Processing Techniques
    - Batch or Sequential Processing
  - Measures of Performance
    - Covariance uncertainty errors
    - Multiple Models allow for comparison against baseline normalcy models
- Information Filter
  - Fusion of Dissimilar Sources
  - Centralized or Distributed
  - Accounts for Double Counting
- Knowledge-Based Temporal Abstraction Process
  - Context Interpretation
- Multiple Hypotheses
  - Fusion of Dissimilar Data
  - Fusion of Detection Methods with Classification Methods

Fusion Techniques & Those Used by Veridian-Stanford Team (Part 1)

Fusion Method	Advantages	Disadvantages	Any Data type?	Any Venue?	Any Infection Path?	Teams tried it at all	Teams but little used	Teams and used	Under development
General Methods			Types: Behavioral – B, Non Traditional Medical – NTM, Traditional Medical – TM						
Statistical	Makes full use of data	Ignores underlying model	B, NTM, TM	Yes	Yes			x	
Auto-Regressive Moving Average (ARMA)	Characterizes model	Complex, requires large datasets	B, NTM, TM	Yes	Yes	No			
Stochastic Processing	Characterizes model	Ignores dataset	B, NTM, TM	Yes	Yes			x	
Matrices	Multi-parameters	Difficult to handle	B, NTM, TM	Yes	Yes			x	
Temporal Correlators	Correlates dynamic relationships	Requires normalcy models	B, NTM, TM	Yes	Yes				x
Spatial Correlators (Association)	Correlates spatial activity	Requires some link between datasets	B, NTM, TM	Yes	Yes			x	
Temporal Clustering Methods	Localizes activity over time	Well-behaved data sets within window	B, NTM, TM	Yes	Yes				x
Spatial Clustering Methods	Localizes activity for cluster development	Dimensionality	B, NTM, TM	Yes	Yes				x
Knowledge Based Methods									
Blackboards	Links features or attributes	Well-defined task structure	B, NTM, TM	Yes	Yes	No			
Expert Systems	Defines abstractions to interpret events	Highly structured & Need for expert decomposition of events	B, NTM, TM	Yes	Yes			x	
Evidential Reasoning	Extends probability theory to include "don't know" conditions	Possibility theory is generally ad-hoc	B, NTM, TM	Yes	Yes	No			
Bayesian Networks	Determines impact of features on classification and identification	Calculating relationships	B, NTM, TM	Yes	Yes			x	
Classification Clustering	Cluster dataset activity to class structure (e.g. syndromic classification)	Test data needs to be developed	B, NTM, TM	Yes	Yes				x

Fusion Techniques & Those Used by Veridian-Stanford Team (Part 2)

Fusion Method	Advantages	Disadvantages	Any Data type?	Any Venue?	Any Infection Path?	Teams tried it at all	Teams but little used	Teams and used	Under development
Fusion Techniques									
Kalman Filters	Evaluates Statistical and Stochastic methods jointly	Models must be representative of conditions	B, NTM	Yes	Yes			x	
Fusion of Similar Data	Weighs each source contribution (e.g. KF) (Good	Limited to data source quality	B, NTM	Yes	Yes			x	
Fusion of Dissimilar Data	Capitalizes on feature diversity (e.g. MHT)	Requires structured system	B, NTM, TM	Yes	Yes				x
Centralized Fusion	Makes full use of processing	Single point of failure (Double counting if layered)	B, NTM, TM	Yes	Yes			x	
Distributed Fusion	Allows multiple processing centers to reach conclusions	Increased Processing Complexity (Double	B, NTM, TM	Yes	Yes				x
Anomaly Detectors	Compares datasets against test data (normal)	Requires sufficient "clean" data to develop "normal" metrics	B, NTM	Yes	Yes			x	
Syndromic Classification Detectors	Compares datasets against syndromically conditioned test data (normal)	Mainly nonexistent	B, NTM, TM	Yes	Yes				x
Learning Techniques									
Neural Networks	Data matching ability	Training	B, NTM, TM	Yes	Yes	x			
Hidden Markov Models	Develops underlying Markov structure	Training	B, NTM	Yes	Yes	No Hidden Markov development			Markov models under development
Optimization Techniques									
Maximum A-Posteriori	Bayesian techniques well understood	Requires priors	B, NTM	Yes	Yes				x
Maximum Likelihood	Non-Bayesian approach	May not be optimal	B, NTM, TM	Yes	Yes				x
Least Squares	Effective with linear systems; Deterministic	Requires complexity to handle non-linear systems	B, NTM, TM	Yes	Yes			x	
Minimum Mean Square Error	Random	Requires density	B, NTM, TM	Yes	Yes				x
Weighted Average	Simplest process	Non-optimal	B, NTM, TM	Yes	Yes			x	
Processing Techniques									
Batch Processing	Global evaluation	Data storage	B, NTM, TM	Yes	Yes			x	
Sequential Processing	Temporal characterization	Local evaluation	B, NTM, TM	Yes	Yes				x
Measures of Performance	Quantifies performance	Requires baseline or truth data	B, NTM, TM	Yes	Yes				x

Veridian Data Fusion Concepts

- Detection Fusion
  - Fusion of multiple sources for early detectability
- Spatial-Temporal Kalman Filter
  - Combine multiple source contributions of similar type
  - Two Kalman filter models run simultaneously
    - Normal model and Anomaly model
  - Adaptively weighs each source contribution
  - Operate normal and anomalous models jointly
  - Cluster & assess those sources that give rise to anomalous behavior
  - Spatial representation of afflicted sources
  - Temporally combines uncertainty errors for joint evaluation of data-dependent statistical models and stochastic models
- Information Fusion
  - Combines results over dissimilar sources by fusing multiple ST Kalman filters to assess anomaly conditions
  - Fusion performed by combining uncertainties from each source type & removing redundant information (compensates for double counting)
- Classification and Identification Fusion
  - Fusion of multiple sources
  - Combines selected attributes from NT behavioral sources with NT medical sources
  - Three approaches considered
- Knowledge-Based Temporal Abstraction Process
  - Provides an interpretive quality to the data – creating abstractions at multiple levels
    - Based on data types
    - Context defines interpretation
  - Links NT behavioral & NT medical conditions
- Multiple Hypotheses
  - Means to fuse detection results with syndrome classification results
  - Hypotheses compensate for time-varying uncertainty
- Bio-STORM
  - Evaluates traditional medical results

TITLE PAGE

Briefing Materials concerning TIA BIOALERT 1

FOLDER 2.b TOC 6

DATA SOURCES

- 2.b.1 Bio-surveillance Data Space (consolidated Team response)
    - Pitt-CMU
    - IBM T.J.Watson Research Center
    - Johns Hopkins U Applied Physics Lab
    - Verdian-Stanford U
- 2.b.1.a Templates listing traditional/non-traditional Sources
- 2.b.1.b Charts – Status of sources input
    - Pitt-CMU

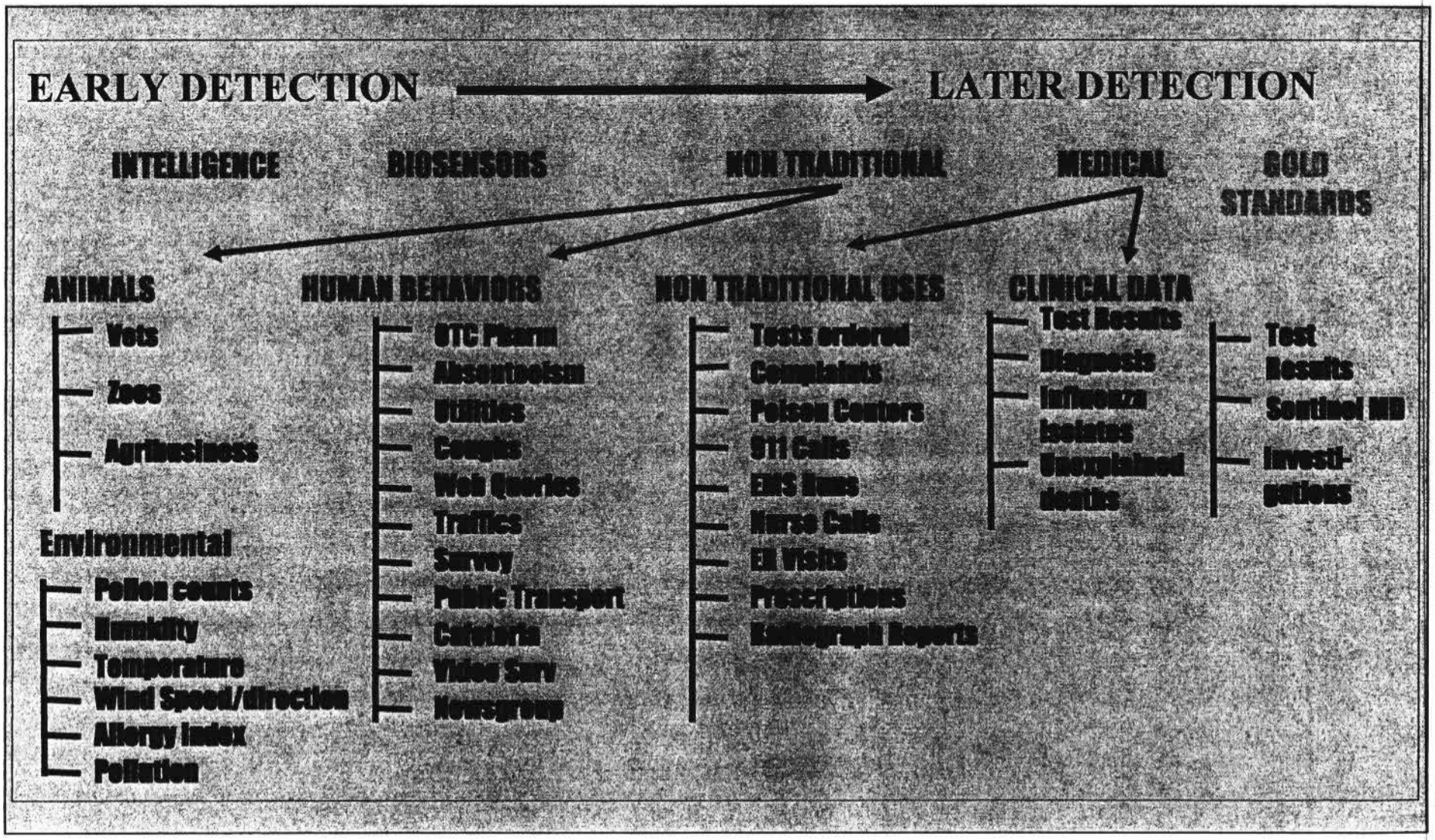
FOLDER 2.b TOC 6 –Cont'd – Page 2

- 2.b.2 Scalable Bio-surveillance Systems: Focus Areas
  - Pitt-CMU
    - Biomed Security Institute (MPC Corp)
    - Pittsburgh Poison Center
- 2.b.3 Status of JHU/APL Data Sources
  - Johns Hopkins U Applied Physics Lab
- 2.b.4 Data Sources
  - IBM TJ Watson Research Center
- 2.b.5 Data Sources
  - Veridian-Stanford Team -2nd PI Meeting
- 2.b.5.a Data Sources (Hampton Roads)
  - RPHD
- 2.b.5.b In-hand or Imminent Data Sources
- 2.b.5.c Norfolk Utility Data – Water Usage
- 2.b.5.d Surveillance Data, Inc. (SDI)
  - SDI

- 2.b.6 Traditional/non-Traditional Data Sources
  - IBM TJ Watson Research Center
  - Pitt-CMU
  - Johns Hopkins U Applied Physics Lab
  - Verdian
- 2.b.7 Bio-surveillance Data Space (2)
- 2.b.7.a Chart – RPHD Data Sources
  - Veridian-Stanford
- 2.b.7.b Data Sources: Pharmacy Data
  - Walter Reed
- 2.b.7.c. Data Sources: Tricare Nurse Hotline Data
  - Walter Reed
- 2.b.7.d. Data Elements Nurse Calls
- 2.b.7.e. Data Sources: Veterans Administration
  - VA

Biosurveillance Data Space

(Consolidated from Team's Responses)



(b)(4)

Traditional Data Sources

Data Source	Organization	Issues	Utility	Frequency of Collection	Latency	Temporal Resolution	Geographic localization / granularity	Electronic Data	Status	Next Step	Period (over what dates)	IRB?	Personal ID
[redacted]	Patient admission records			Real-time	Seconds	Timestamp	Address	Yes	Receiving data feed		5 years	Approved	Yes
[redacted]	Patient admission records			Real-time	Seconds	Timestamp	Address	Yes	Receiving data feed		5 months	Approved	Yes

2.B.1.c.A.

Non-Traditional (Clinical) Data Sources

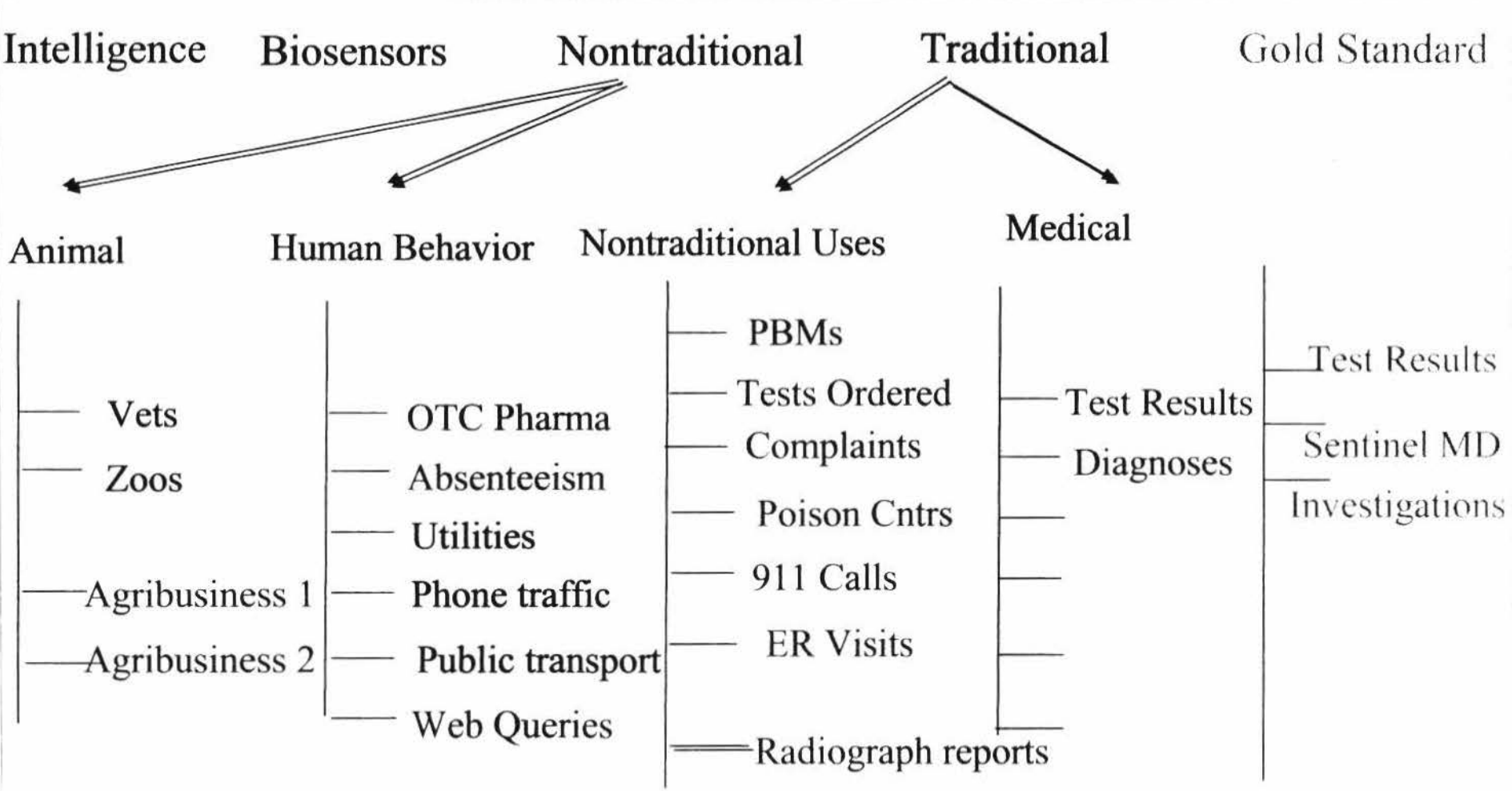
Data Source	Organization	Issues	Utility	Frequency of Collection	Latency	Temporal Resolution	Geographic localization / granularity	Electronic Data	Status	Next Step	Period (over what dates)	IRB?	Personal ID
Health website queries	Individual queries	May be difficult to get location information							Under investigation				
(b)(4)									Under discussion				
		Declined to purchase data							Not pursuing		>3 years		
									Will be approached				
	ndividual laims	No repsonse to inquiries							Not pursuing		5 years		

Non-Traditional Data Sources

Data Source	Organization	Issues	Utility	Frequency of Collection	Latency	Temporal Resolution	Geographic localization / granularity	Electronic Data	Status	Next Step	Period (over what dates)	IRB?	Personal ID
Animal													
A Beef Producers	Individual cows			One time	N/A			Yes	Awaiting delivery	Preliminary analyses	5 years	Approved	No
A Pork Lot	Individual hogs			One time	N/A			Yes	Awaiting delivery	Preliminary analyses	5 years	Approved	No
(b)(4)		Declined to provide data							Not pursuing				
		Declined to provide data							Not pursuing				
Animal Control Office									Uninvestigated idea	Investigate feasibility			
Call Center													
(b)(4)	Individual calls			One time	N/A			Yes	Data delivered	De-identify data	2 years	Approved	Yes
	Individual calls			One time	N/A	Timestamp	Address	Yes	Data delivered	De-identify data	5 years	Approved	Yes
	Individual calls			One time	N/A			Yes	Received 2 years of data	Preliminary analyses	5 years	Approved	No
									Awaiting approval from county				
		Their CAD vendor must assemble data							Awaiting approval from county				
									Under discussion				
									Under discussion				
									Contacted				
									Under discussion				
Purchasing													
Anonymous Supermarket									Awaiting delivery	Preliminary analyses	1.5 to 2 years		
Anonymous Drugstore									Under discussion				
IRI data	Classes of OTC remedies	Level of temporal/spatial aggregation too coarse				Weekly	County		Not pursuing		5 years		

AC Nielson	Generic chemical classes/names	Level of temporal/spatial aggregation too coarse				(Not daily)	Store		Not pursuing		3 years			
(b)(4) [redacted]	Inidvidual Rx					Timestamp	Home address		Under discussion		2 years			
(b)(4) [redacted]	Classes of Rx drugs	Level of temporal/spatial aggregation too coarse							Not pursuing		2 years			
(b)(4) [redacted]	Classes of Rx drugs	Level of temporal/spatial aggregation too coarse					Census district		Not pursuing		54 months			
(b)(4) [redacted]	MSA	Level of temporal/spatial aggregation too coarse							Not pursuing					
Grocery data									Uninvestigated idea	Investigate feasibility				
Indirect Evidence of Absenteeism														
A phone company		Legal department refuses to participate							Continuing to pursue using other avenues					
Video store									Uninvestigated idea	Investigate feasibility				
Highway cameras									Uninvestigated idea	Investigate feasibility				
ISP		Try Verizon, AT&T, WorldNet, Earthlink							Uninvestigated idea	Investigate feasibility				
(b)(4) [redacted]		Does not monitor whether cable is being viewed							Not pursuing					
A light company									Under discussion					
(b)(4) [redacted]		Data latency too high							Not pursuing					
A transit company									Contacted					

Pitt-CM



Legend: Red: we have; Blue: under contract; Black: we have tried to obtain or are trying at present

2.8.1.5

2.8.3

Status of JHU/APL Data Sources

	Military	Civilian	Coverage	Value	Relationship with Provider	Privacy
Health Care						
(b)(4) ER Visits	Current	Current	[redacted]	Medium	OK	High
Electronic Claims	Current	Current		High	Good	High
Requests for Lab Tests		Current		Med-High	OK	High
EMS Runs	Future	Future		Low		High
Medical Examiner	Future	Future		Low	Good	Med
(b)(4) Nursing Homes		Past	[redacted]	Low		Med
Human Behaviors						
(b)(4) Absenteeism Schools		Current	[redacted]	High	OK	High
OTC Medications		Current		High	OK	Medium
Dial-A-Nurse	Pending	Future		High	Military Good, Civilian Poor	Med-High
911 Calls	N/A	Future		Med-High	Only approached the city so far	?
(b)(4) Poison Center Calls		Past	[redacted]	Low	Good	High
Animal						
(b)(4) Zoos & Parks		Current	[redacted]	Medium	Good	Low
Vet. Hospitals	Pending	Current		Medium	OK	Medium
Research		Current		Medium	OK	Medium
Vet. Assistance Centers		Future		Medium		Medium
Animal Refuge		Current		Low	OK	Low
Animal Control		Pending		Medium	Good	Low
DNR		Future		Low		Low
Sensors						
(b)(4) Environmental	Pending		[redacted]	High	Good	Secure
Pathogen Specific Tickets	Pending			High	Good	Secure
Gene Chips	Future	Future		High		High

2.b.8
JHUAPL
alpldata

ESSENCE II Indicators Traditional

Data Source	Organization	Period	Value	Freq. of Collection	
Diagnostic Lab Results	State Labs	11/99 to present	Low	daily	
Diagnostic Lab Results	Private	not yet	Moderate to High	real-time	
Nursing Homes	Private	11/99 to 4/2000	Low	daily	same day

Non- Traditional (clinical)

	Data Source	Organization	Period	Value	Freq. of Collection	Latency
	Emergency Room	Individual Hospitals	11/99 to present	High	daily / real-time	real-time
	Mil. Ambulatory Data	Tricare	99 to present	High	8hrs / real-time	1-3 days
	Civ. Ambulatory Data	HMOs	11/99 to present	High	daily / real-time	1 day to a year
	Mil. Prescription Meds.	Web MD	7/01 to present	High	8 hrs / real-time	seconds
	Mil. Nurse Hotline	Tricare	in progress	High	daily / real-time	real-time
	Civ. Nurse Hotline	HMOs	in progress	High	daily / real-time	real-time
(b)(4)	Poison Center	[redacted]	99'	Low	daily / real-time	real-time
	Electronic Vital Signs		in progress	High	daily / real-time	real-time
	Civ. Inpatient Data	Insurers (3 states)	99-02	High	daily	real-time*
	Civ. Prescriptions	HMOs, Insurers	99-02	Hgh	daily	real-time*

Non-Traditional

(b)(4)	Work Absenteeism	[redacted]	99-00	unknown	could get daily	end of day
	School Absenteeism	some jurisdictions	11/99 to present	Mod	daily	same day
	Over the counter meds	major chains	11/99 to present	Mod to High	daily / real-time	real-time
	Animal health	several sources	9/01 to present	High for certain diseases	daily	daily
	TV viewing patterns	TV ratings	in progress	unknown	daily	1 day
(b)(4)	Metro rider ship	[redacted]	00-01	unknown	daily	?
	Airline rider ship		98-present	unknown	daily	same day
	Health care web sites	HMOs, Insurers	99-02	High, population specific	real-time	
(b)(4)	Video surveillance metro	[redacted]	02	unknown	daily	1 day
	Death data, notices	national (3 states)	96-present	unknown	daily	1-2 days
	Grocery (not incl over counter)	City	94-98	Mod	daily	1 day
	Search engines (Web)	Search engine	02	unknown	real-time*	real-time*

Temporal Res.	Spatial Res.	Electronic	Automated	Personal ID	Status	IRB	Issues
# of positives	Zip Code	yes	Can be	no	active	yes	minimal coverage
# of positives	Zip Code	yes	Can be	no	inactive	no	coverage resolved
# of cases	exact location	some	no	no	inactive	no	

Temporal Res.	Spatial Res.	Electronic	Automated	Personal ID	Status	IRB	Issues
chief complaint	Zip Code	yes	yes	can be obtained	active	yes	number participating
ICD-9	Zip Code	yes	yes	yes - on Denver	active	yes - for certain projects	speeding up access
ICD-9	Zip Code	yes	yes	can be obtained	active	no	per capita coverage
prescriptions	Zip Code	yes	yes	yes - on Denver	active	yes - for certain projects	
chief complaint	unknown	yes	?	can be obtained	in progress	not yet	legal difficulties in obtaining
chief complaint	Zip Code	yes	yes	can be obtained	in progress	no	redirection of FY02 funds
chief complaint	Phone Number	yes	yes	no	inactive	no	not as useful as first thought
vital signs and ICD9	unknown	yes	yes	?	in progress	?	
ICD-9	ZIP res & work	yes	yes	yes	historical	no, privacy exempt	
prescriptions	ZIP res & work	yes	yes	yes	historical	no, privacy exempt	

# absent	employee	yes	yes	no	inactive	no	need other sources also
# absent by school	school location	yes	some	no	active	no	need to understand all factors
# of products	store location	yes	yes	no	active	no	need to understand all factors
syndromes/species	facility location	some	no	no	active	no	doesn't cover all possibilities
3 of viewers	jurisdiction	yes	yes	no	in progress	no	needs detailed investigation
rider destination	station	yes	?	smart pass only	investigating	no	needs detailed investigation
rider ship by city	airport	yes	yes	no	?	no	disease model use
ICD-9 categories	ZIP res & work	yes	yes	yes	active	no, privacy exempt	best: families with children
persons	point location	no/convert	yes	no	historical	no, privacy exempt	under development
persons	point location	yes	yes	yes	active	no, privacy exempt	
household	ZIP, point loc	yes	yes	yes (household)	historical	no, privacy exempt	
current interest	IP address->city	yes	yes	no, IP address	historical	no, privacy exempt	

Data Sources

Murray Campbell
IBM TJ Watson Research Center

Overview

- Site/Regional data sources
  - Survey of self-assessed health
  - Site attendance
  - Web queries (outgoing)
  - Phone calls (outgoing)
  - Other (utilities, cafeteria, cough)
- Insurance claims data (SDI)
- Environmental
- Other SDI data sources

Site-Specific Surveillance

- Applicable to military bases, work sites, campuses, etc.
  - Geographically constrained (single building or small set of buildings)
  - Relatively fixed population of reasonable size (i.e., at least hundreds)
- Evaluate nontraditional data sources that focus on specific sites
- Some site-specific data sources have regional analogs

Survey of Self-Assessed Health

- Hypothesis: A disease outbreak causes a detectable decrease in a population's overall self-assessed health level
- Poll a population, asking individuals to self-assess their current health
  - If ill, can ask about symptoms
- Advantages: Can detect early symptoms, real-time, can focus on specific areas or sub-populations, accuracy controllable by sample size, various channels (email, phone, web) available, symptom information provides specificity
- Concerns: Privacy, reliability, missing responses, legal liability, time to establish baseline
- Applicable regionally

Watson Health Survey

- 400 participants (2200 total employees)
- Polled weekly via email
- Active from 1/25/2002-5/31/2002
- Privacy protection
- 92% of responses show up same day (73% by noon)
- Minimal dropouts (< 2%)
- Sick individuals always report at least one symptom

Site Attendance

- Hypothesis: A disease outbreak causes a detectable reduction in site attendance
- Record site attendance electronically using badge-in data
- Advantages: Results early in day are predictive of daily total, real-time, data already being collected for other purposes
- Concerns: Only applicable to restricted access sites, exit data may not be recorded, non-specific, not sensitive to mild illness, affected by external factors (e.g., weather, school holidays, day of week), privacy

Watson Site Attendance

- About 2100 unique badges seen each weekday
- Ongoing collection since 12/2001
- Privacy issues: anonymized badge ID
- Data cleansing issues (e.g., transient badges from other sites)
- Final attendance for day predictable from early entries (linear regression using 10am data gives R-squared of .99)
- Strong effects from snow days, school closings/holidays

Web Queries

- Hypothesis: A disease outbreak at a site causes a detectable increase in web activity on medically-related topics
- Outgoing web accesses may be recorded on a proxy server
  - Source IP, destination IP, URL (optional)
- Advantages: Real-time, some specificity if URL information recorded, data already being collected for other purposes
- Concerns: Not all sites use proxy server, privacy, volume of data, identifying medically-related queries
- May be applicable regionally (ISPs)

Watson Web Queries

- Medically-related sites manually identified
- Proxy logs from 1/1/2002 – 4/30/2002
- Anonymized source IP, actual destination IP available. No URL recorded at Watson due to large storage requirements
- 20-60 unique individuals access medically-related sites on a typical weekday

Phone Calls

- Hypothesis: A disease outbreak at a site causes a detectable increase in phone calls/callers to medically-related numbers
- Outgoing calls may be recorded at a site for billing purposes
  - Source number, called number, time, length
- Advantages: Real-time, data being collected for other purposes
- Concerns: Privacy, identification of medically-related numbers
- Applicable regionally (from phone company)

Watson Phone Calls

- About 5000 medically-related phone numbers identified in Westchester County and vicinity
  - Scan yellow pages, web directories
- Data available since 10/2001, broken down by site (Yorktown, Hawthorne)
- On a typical weekday
  - 5000 calls, 1500 callers
  - 2% of calls are to medically-related numbers
  - 5% of callers make medically-related calls

Other Site-Specific Sources

- Utility-based (electric, water, sewage)
  - Large daily variations precluded use at Watson
- Cafeteria
  - Being evaluated
- Cough counting
  - Technology exists, need method to estimate number of coughers
- Site departure
  - Not recorded electronically at Watson, car counting tool may be of use here

Insurance Claims Dataset

- Supplied by Surveillance Data Inc.
- August 2001 through May 2002
- ICD9 codes 001-139 (Infectious diseases) and 460-519 (Respiratory)
- NY, Pittsburgh, DC areas
- Includes anonymized patient identifier, service date, process date, up to 4 ICD9 codes, age, gender, zipcode
- 1.2 million claims
- SDI data covers nationwide back to 1998

Environmental Dataset

- Archive of environmental data (temperature, precipitation, wind speed, wind direction, etc.)
- Currently being recorded for a set of states in Northeast
- Not a detector, but may affect population health level
- Near-real-time data, ~1 km resolution if used with satellite data to interpolate

Other Datasets Available Through SDI

- OTC medications
  - National coverage at store level, 1 day lag
- Lab tests ordered
  - National coverage by zipcode, 1-3 day lag
- Prescription medications
  - National coverage by zipcode, near-real-time possible
- ER admissions, discharges
  - Limited coverage currently available, near-real-time

Backup

Polling Pilot Study Details

- Polling conducted via email
- Each participant contacted once weekly, 1/5 of population on each weekday, randomly partitioned each week
- Self-assessed health is requested for the day poll is **sent**, not received
  - I.E., Retrospective self-assessment
- Demographic information supplied by participants

Demographic Questions

- Age?
- Children in household?
- Do you smoke?
- Have you had a flu shot this fall?
- Reside in Westchester county?
- Primary work location?
- Do you use mass transit regularly?

Polling Questions (1)

- Please assess how you feel today:
  - As usual
  - Worse than usual (sick)
  - Much worse than usual (very sick)
- Note that the wording is dynamic, and changes if the message is viewed on a day after the poll was sent, I.E., "Please assess how you felt on <date>"

Polling Questions (2)

- If sick, the following additional questions are asked:
  - Which symptoms do you have? (Fever, headache, muscle ache, fatigue, cough, stuffy/runny nose, sore throat, nausea, vomiting, diarrhea, other)
  - How many days have you been unwell prior to today?
  - Was anyone else in your household sick in the prior week?
  - Do you feel like you are getting better?
  - Have you been in a large crowd in the past two weeks?

2.B.5

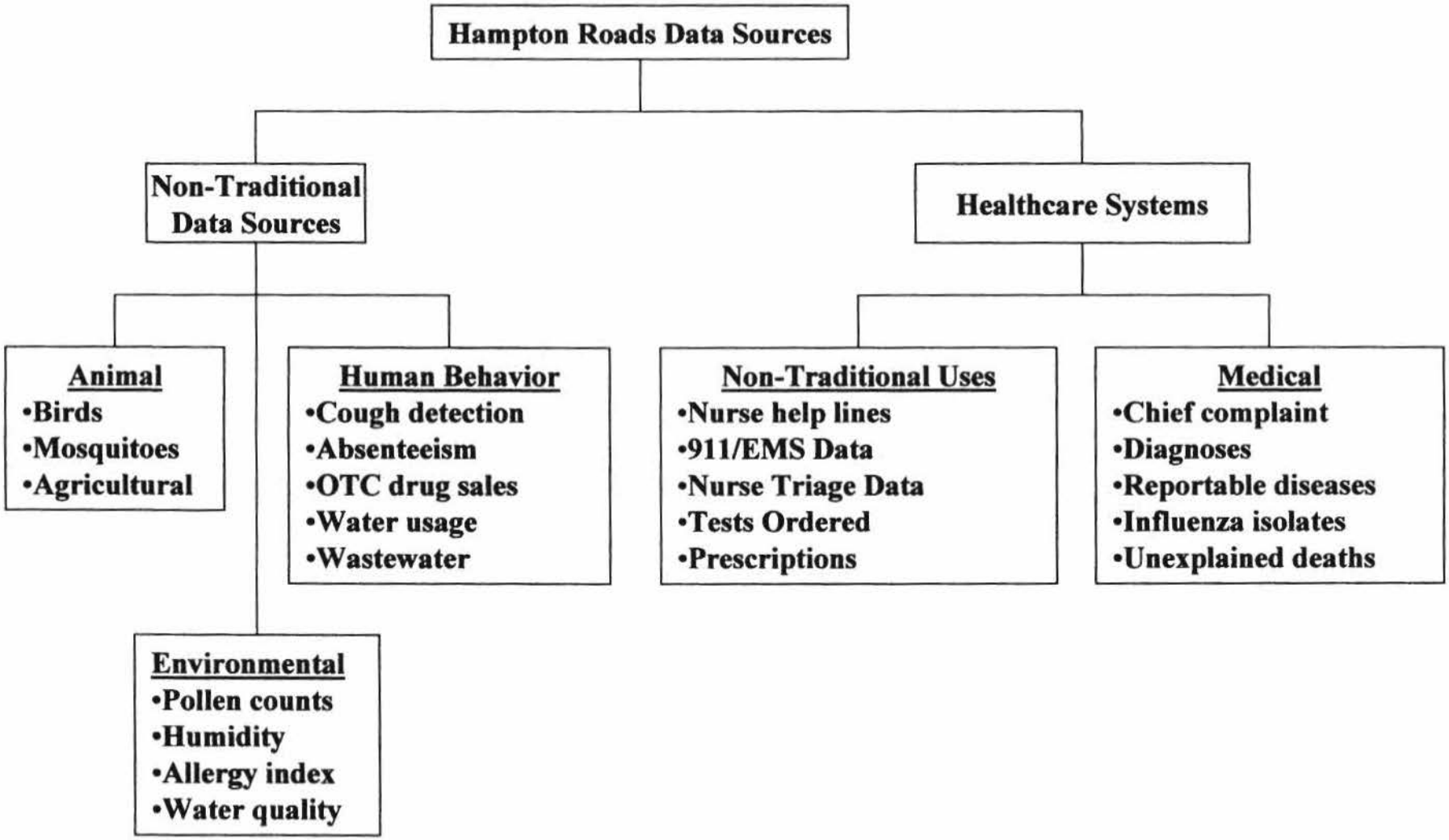
DATA SOURCES

Veridian-Stanford Team

2nd PI Meeting

4-6 June 2002

RPHD Data Sources

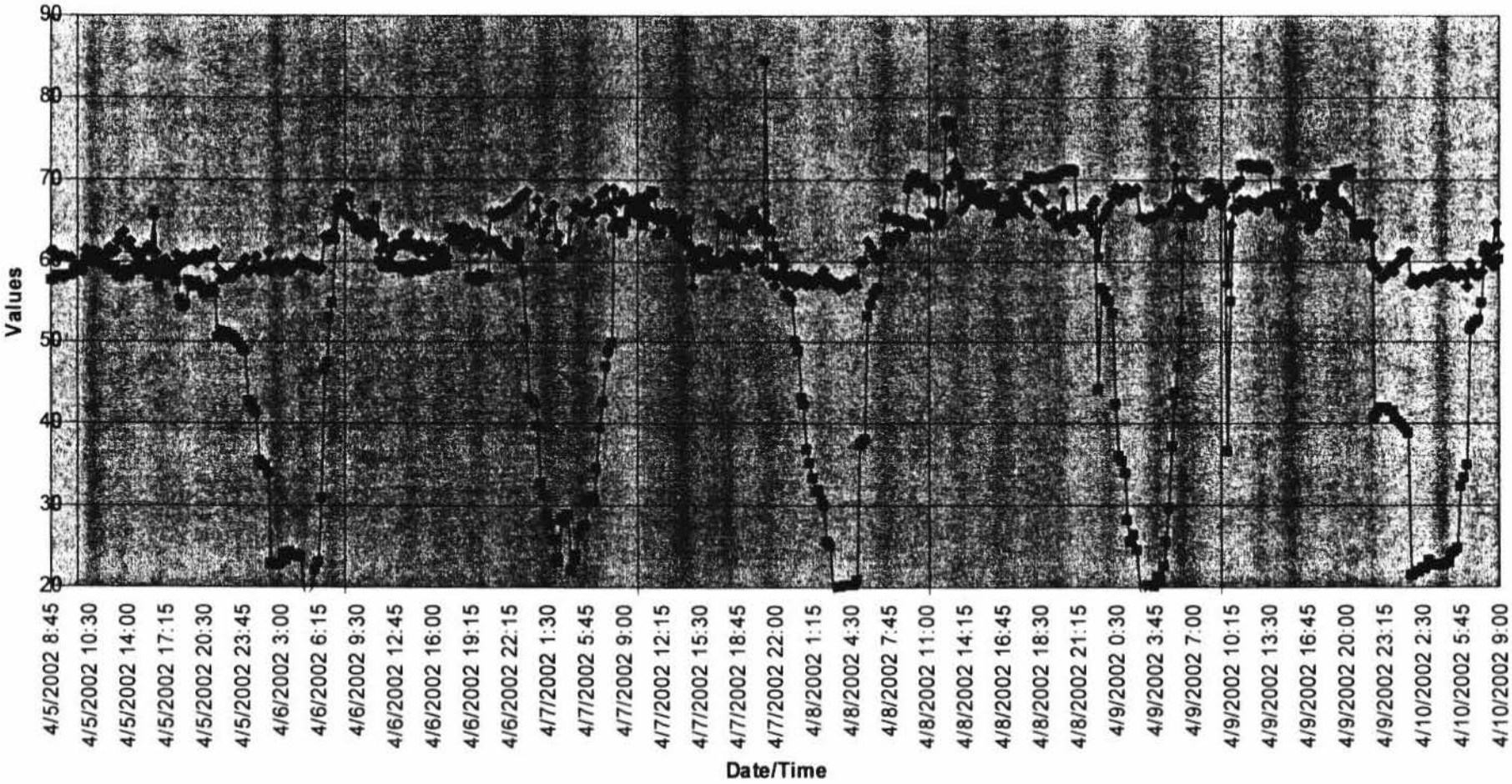


2.b.5.B

In-hand or Imminent Data Sources

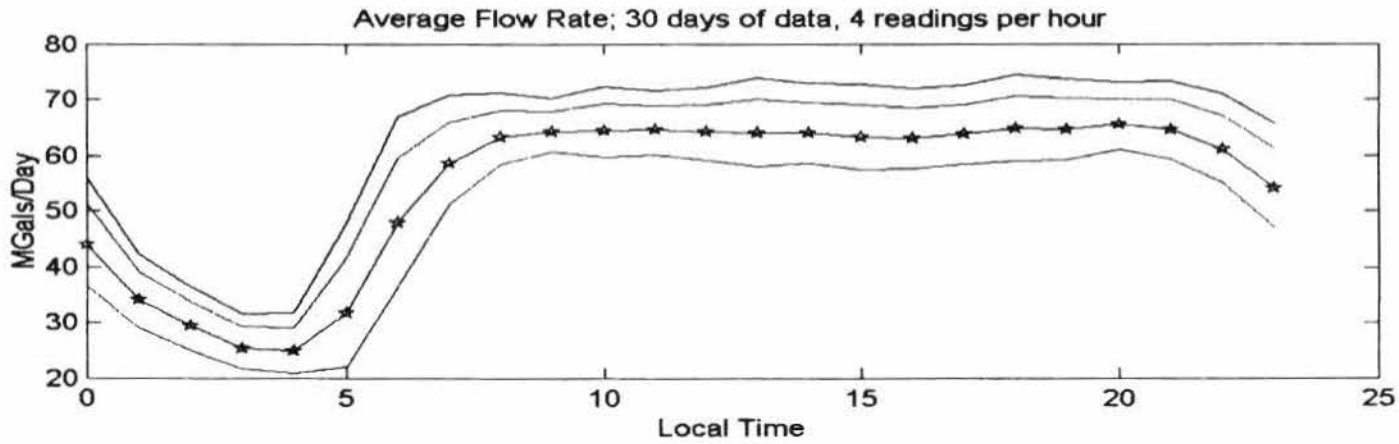
- Naval Medical Center, Portsmouth CHCS, includes are military acute care clinics
- Syndromic Surveillance Data collected by Regional Public Health
- Norfolk Water and Wastewater
- Norfolk Fire and Rescue 911/EMS
- Commercially Available Data (Surveillance Data, Inc.)
  - Large sample of doctor's offices (75 in Norfolk)
  - Flu isolates

Norfolk Utility Data - - Water Usage



What is Detectable in Water Usage?

30 Day Avg Flow Rate and Pressure (4 Readings Per Hour) With 1 std interval (~+/-6 MG/Day) and Upper 95% prediction limit



Population of Norfolk/Portsmouth is over 330,000.

Rough Calculation

100,000 people ill with 5 flushes (2 gal ea.) extra over 8 hours is about one standard deviation.

2.B.3D

Surveillance Data, Inc. (SDI)

About SDI

Surveillance Data Inc., based in Plymouth Meeting, Pennsylvania

- Has been in business since 1981
- The company provides healthcare data and market research to pharmaceutical, consumer products, and healthcare companies.
- SDI's mission is to provide healthcare data products with high business relevance and the means and know-how to put this data into action.
- SDI is the leading provider of real-time localized illness tracking and modeling data to the pharmaceutical and healthcare industries.
- SDI is providing data for IBM

Veridian status with SDI

Have obtained pricing and are negotiating modified Agreement

Possible Data from SDI

Historical Data

- **3 years of historical Diagnoses data (Same ICD-9 Codes as above)**
- **3 seasons of historical Mosquito Indices (April-September)**
- **3 seasons of historical Laboratory Confirmed Influenza isolates**
- **3 seasons of historical Rapid Test Diagnoses of Influenza**
- **3 years historical Allergy Alert Index (February – October)**
- **Historical Pollen Counts (when available)**

Weekly or Daily Reported Data

- **Raw diagnoses of ICD-9 codes attached (52 weeks)**
- **Mosquito Index (April-September)**
- **Laboratory Confirmed Influenza Isolates (October-March)**
- **Rapid Test Diagnoses of Influenza (October-March)**
- **Pollen Count data (Total, Grass, Tree, Weed) when available**
- **Allergy Alert Index (February - October)**

2.B.6

Non-Traditional Data Sources

Data Source	Organization	Issues	Utility	Frequency of Collection	Latency	Temporal Resolution	Geographic localization / granularity	Electronic Data	Status	Next Step	Period (over what dates)	IRB?	Personal ID
	IBM												
Health survey: Regularly poll a population, asking individuals to self-assess their health	IBM Watson Research Lab	Survey must be organized on a site/regional basis. Modeling missing responses. Time to establish baseline.	High, particularly in cases of heightened alert. Provides some specificity if symptoms are queried when ill. Sensitivity controllable by sample size.	Real-time	Symptoms showing on a number of persons in survey area	Second	Site/Regional. Can be adjusted as needed	Yes	Initial survey complete. 400 individuals out of 2200 participated, were polled weekly via email	Rerun at Watson, 2002-2003 with higher sample size. Test at other sites.	1/25/2002-5/31/2002	Yes at Watson, approved Jan 2002	Yes, but we anonymized
Site attendance: Electronically recorded from badge entry	IBM Watson Research Lab	Not all sites record attendance electronically. Affected by external factors.	Moderate if applicable	Currently daily, potentially real-time	Symptoms showing on a number of persons at site	Second	Site	Yes	Data collection ongoing	Model seasonality as more data become available. Test at other sites.	12/2001 - present	Not required, but Watson HR, Legal approval needed	Yes, but we anonymized
Proxy server logs: Outgoing accesses to medically-related websites	IBM Watson Research Lab	Not all sites use a proxy server. Identifying medically related accesses.	Moderate if applicable and sufficient volume. May provide some specificity if URL is tracked	Currently monthly, potentially real-time	Symptoms showing on a number of persons at site	Second	Site	Yes	Four month sample collected	Model seasonality as more data become available. Test at other sites.	1/1/2002-4/31/2002	Not required, but Watson HR, Legal approval needed	Yes, but we anonymized
Phone calls: Outgoing calls to medically-related numbers	IBM Watson Research Lab	Identifying and maintaining list of medically-related phone numbers	Moderate if applicable and sufficient volume	Currently daily, potentially real-time	Symptoms showing on a number of persons at site	Second	Site	Yes	Data collection ongoing	Model seasonality as more data become available. Test at other sites.	10/2001 - present	Not required, but Watson HR, Legal approval needed	Yes, but we anonymized

Data Source	Organization	Issues	Utility	Frequency of Collection	Latency	Temporal Resolution	Geographic localization / granularity	Electronic Data	Status	Next Step	Period (over what dates)	IRB?	Personal ID
Utility usage: Water, electric, sewage	IBM Watson Research Lab	Normal daily variations can swamp any signal	Low at Watson. May be a proxy for site attendance.	Currently monthly, potentially real-time	Symptoms showing on a number of persons at site	Second	Site	Yes	Large daily variations precluded use at Watson	Test at sites with more typical usage patterns.	1/2002-5/2002	Not required	No
Cafeteria transactions	IBM Watson Research Lab	Normal daily variations can swamp any signal	Low at Watson. May also be a proxy for site attendance.	Currently monthly, potentially real-time	Symptoms showing on a number of persons at site	Second	Site	Yes	Three month sample collected	None planned	1/2002-2/2002	Not required	No
Cough counting: Recorded in public areas	IBM Watson Research Lab	May be difficult to distinguish single cougher from multiple coughers. Specific to respiratory	Low	Potentially real-time	Symptoms showing on a number of persons at site	Second	Site	Potentially	Cough incidence too low for use at Watson	None planned	N/A	Unknown, but likely Watson HR, Legal approval needed	No
Site departure	IBM Watson Research Lab	Not all sites electronically record departures	Moderate if applicable	Potentially real-time	Symptoms showing on a number of persons at site	Second	Site	Potentially	Not recorded at Watson. Could be tracked with car-counting technology	Investigate use of automatic car-counting to detect site departure.	Not yet available	Not required	No
Newgroups: Health-related articles	N/A	Little geographic specificity, small article counts	Low	Potentially real-time	Symptoms showing on a number of persons in geographical area	Second	Poor geographic specificity	Yes	Not useful due to lack of geographic specificity, small sample sizes	None planned	Google maintains archive going back ~20 years.	Not required	May have personal identifiers
Environmental (temperature, moisture, pollution, etc.)	NOAA	Not a detector, but may be useful for predicting normal health patterns	Moderate	Real-time	N/A	Hour	1 km, using station readings with satellite data for interpolation	Yes	Currently archived for CT, NY, NJ, MD, PA, DC	Investigate effects of environmental conditions on health.	11/2001 - present	N/A	N/A
OTC medication	Surveillance Data Inc.	TBD	TBD	Daily	Symptoms showing on a number of person	Day	Store (national coverage)	Yes	Under discussion with SDI	Obtain, analyze data	TBD	Not required	TBD

Data Source	Organization	Issues	Utility	Frequency of Collection	Latency	Temporal Resolution	Geographic localization / granularity	Electronic Data	Status	Next Step	Period (over what dates)	IRB?	Personal ID
(b)(4) [redacted]													
Animal													
Agribusiness 1				One time	N/A			Yes	Awaiting delivery	Preliminary analyses	5 years	Approved	No
Agribusiness 2				One time	N/A			Yes	Awaiting delivery	Preliminary analyses	5 years	Approved	No
(b)(4) [redacted]		Declined to provide data							Not pursuing				
		Declined to provide data							Not pursuing				
Animal Control Office									Uninvestigated idea	Investigate feasibility			
Call Center													
(b)(4) [redacted]	Individual calls			One time	N/A			Yes	Data delivered	De-identify data	2 years	Approved	Yes
	Individual calls			One time	N/A	Timestamp	Address	Yes	Data delivered	De-identify data	5 years	Approved	Yes
	Individual calls			One time	N/A			Yes	Received 2 years of data	Preliminary analyses	5 years	Approved	No
									Awaiting approval from county				
		Their CAD vendor must assemble data							Awaiting approval from county				
									Under discussion				
									Under discussion				
									Contacted				
									Under discussion				
Purchasing													
Anonymous Supermarket									Awaiting delivery	Preliminary analyses	1.5 to 2 years		
Anonymous Drugstore									Under discussion				
(b)(4) [redacted]	Classes of OTC remedies	Level of temporal/spatial aggregation too coarse				Weekly	County		Not pursuing		5 years		

Data Source	Organization	Issues	Utility	Frequency of Collection	Latency	Temporal Resolution	Geographic localization / granularity	Electronic Data	Status	Next Step	Period (over what dates)	IRB?	Personal ID
(b)(4)	Generic chemical classes/names	Level of temporal/spatial aggregation too coarse				(Not daily)	Store		Not pursuing		3 years		
	Inidvidual Rx					Timestamp	Home address		Under discussion		2 years		
	Classes of Rx drugs	Level of temporal/spatial aggregation too coarse							Not pursuing		2 years		
	Classes of Rx drugs	Level of temporal/spatial aggregation too coarse					Census district		Not pursuing		54 months		
	MSA	Level of temporal/spatial aggregation too coarse							Not pursuing				
Grocery data									Uninvestigated idea	Investigate feasibility			
Indirect Evidence of Absenteeism													
A phone company		Legal department refuses to participate							Continuing to pursue using other avenues				
Video store									Uninvestigated idea	Investigate feasibility			
Highway cameras									Uninvestigated idea	Investigate feasibility			
(b)(4) ISP									Uninvestigated idea	Investigate feasibility			
(b)(4)									Not pursuing				
A light company									Under discussion				

Data Source	Organization	Issues	Utility	Frequency of Collection	Latency	Temporal Resolution	Geographic localization / granularity	Electronic Data	Status	Next Step	Period (over what dates)	IRB?	Personal ID
(b)(4) [redacted]		Data latency too high							Not pursuing				
A transit company									Contacted				

JHU/APL

(b)(4) Work Absenteeism	[redacted]		unknown	could get daily	end of day	# absent	employee	yes	need other sources also		99-00	no	smart pass only
School Absenteeism	some jurisdictions		Mod	daily	same day	# absent by school	school location	yes	need to understand all factors		11/99 to present	no	no
Over the counter meds	major chains		Mod to High	daily / real-time	real-time	# of products	store location	yes	need to understand all factors		11/99 to present	no	no
Animal health	several sources		High for certain diseases	daily	daily	syndromes/species	facility location	some	doesn't cover all possibilities		9/01 to present	no	no
TV viewing patterns	TV ratings		unknown	daily	1 day	3 of viewers	jurisdiction	yes	needs detailed investigation		in progress	no	no
(b)(4) Metro rider ship	[redacted]		unknown	daily	?	rider destination	station	yes	needs detailed investigation		00-01	no	smart pass only
(b)(4) Airline rider ship	[redacted]		unknown	daily	same day	rider ship by city	airport	yes	disease model use		98-present	no	no
Health care web sites	HMOs, Insurers		High, population specific	real-time		ICD-9 categories	ZIP res & work	yes	best: families with children		99-02	no, privacy exempt	yes
(b)(1),(b)(4) Video surveillance metro	[redacted]		unknown	daily	1 day	persons	point location	no/convert	under development		02	no, privacy exempt	no
Death data, notices	national (3 states)		unknown	daily	1-2 days	persons	point location	yes			96-present	no, privacy exempt	yes
Grocery (not incl over counter)	City		Mod	daily	1 day	household	ZIP, point loc	yes			94-98	no, privacy exempt	yes (household)
Search engines (Web)	Search engine		unknown	real-time*	real-time*	current interest	IP address->city	yes			02	no, privacy exempt	no, IP address

Verdian

Cough Detection [redacted]	Military common areas	Determine possible locations	H	Continuous	Minute	Minute	Address	Yes	Under development	Demonstrate prototype	Current	No	No

Data Source	Organization	Issues	Utility	Frequency of Collection	Latency	Temporal Resolution	Geographic localization / granularity	Electronic Data	Status	Next Step	Period (over what dates)	IRB?	Personal ID
Pharmaceuticals – OTC	Norfolk area Base/Post Exchanges		H	Continuous	Day	Hour	Area	Yes	POCs contacted	Brief Base Commandant	?	?	Yes
Environmental Data - Mosquito count	SDI	$$, agreement. Data collected April-Sept.	L	Daily, maybe weekly reporting	Day - Week	Day	Area	Yes	SDI Contract in approval process	Finalize agreement	3 years + Current	No	No
Environmental Data - Humidity and Pollen	Pollen.com	Pollen count combines all types of pollen	L	Every 15 minutes	Minute	Minute	Area	Yes	Automatically being collected at Veridian	Analyze data	3 years + Current	No	No
School Absenteeism	Norfolk Public Schools	Automation, forewarned that data accuracy is a problem. They have 3 year historical data.	H	Daily	Daily	Day	By school district	Yes	Have bypassed normal research proposal process, CIO for Public Schools is setting up Oracle queries for us (time consuming)	Complete MOU with Norfolk. (already signed) Help integrate data	3 years + Current	No	No
Utility Usage	Norfolk Public Works	Probable low signal to noise	H	Every 15 minutes	Minute	Minute	Eastern and Western half of Norfolk	Yes	Currently Obtaining active data. Obtained 5-year historical data. Analyzing data.	Analyzing data and correlating against known Public health outbreaks	5 years + Current	No	No
Utility Usage	Stanford University Campus	Probable low signal to noise. 5 year historical data available.	H	Every 15 minutes	Minute	Minute	Area	Yes	Initial data transfer.	Follow-up	2 years + current	No	No
Sanitation data	Hampton Roads Sanitation Data	Probable low signal to noise	H	Daily	Day	Day	Area per plant	Yes	Made formal request from HRSD. Consulting with Hydrologist.	Awaiting reply -- Follow-up with HRSD.	unknown	No	No

Data Source	Organization	Issues	Utility	Frequency of Collection	Latency	Temporal Resolution	Geographic localization / granularity	Electronic Data	Status	Next Step	Period (over what dates)	IRB?	Personal ID
Potable Water		Probable low signal to noise	H				Area per plant		Requested contact. Waiting for reply.	Follow-up	unknown	No	No
Veterinary Data - Animal Death Reports	CDC monitoring pet deaths at Petting Zoos	May need state DOH approval	L	Daily, but collated at other frequency	Day	Day	Area	Yes	We discovered that Dr. Love does not collect animal data. Obtaining leads from state Vet – he does not collect this kind of data.	Pursue State Vet leads.	unknown	No	Yes
Veterinary Data - Agricultural	California Health	Negotiations in process	L	Intermittent	Unknown	Day	Unknown	Unknown	Data promised	Negotiate access rights	Current	Yes	Yes
Veterinary Data –Urban	Santa Clara Public Health		L	Daily	Day	Day	Address	Yes	Initial discussion	Continue discussion and identify access issues	Current	No	Maybe
Veterinary Data - Zoo	San Francisco Public Health		L	Intermittent	Month	Day	By zoo	No	Data under negotiation	Negotiate access rights	Current	No	No
Allergy Index	SDI	$$, Agreement,	L	Daily, maybe weekly reporting	Day - Week	Day	Area	Yes	SDI Contract in approval process	Finalize agreement	3 years + Current	No	No

Non-Traditional (Clinical) Data Sources

Data Source	Organization	Issues	Utility	Frequency of Collection	Latency	Temporal Resolution	Geographic localization / granularity	Electronic Data	Status	Next Step	Period (over what dates)	IRB?	Personal ID
IBM													
Lab tests ordered	Surveillance Data Inc.	TBD	TBD	TBD	A number of patients have been seen by physicians	Day	Originating zipcode	Yes	Under discussion with SDI	Obtain, analyze data	TBD	No	Yes, anonymization needed
Prescription	Surveillance Data Inc.	TBD	TBD	Near real-time	A number of patients have been seen by physicians	Day	Originating zipcode (in some cases only first 3 or 4 digits)	Yes	Under discussion with SDI	Obtain, analyze data	1-3 years available	No	Yes, anonymization needed
		(b)(4)											
Health website queries	Individual queries	May be difficult to get location information							Under investigation				
(b)(4)									Under discussion				
		Declined to purchase data							Not pursuing		>3 years		
									Will be approached				
	Individual claims	No repsonse to inquiries							Not pursuing		5 years		
JHU/APL													
Emergency Room	Individual Hospitals	number participating	High	daily / real-time	real-time	chief complaint	Zip Code	yes	active		11/99 to present	yes	can be obtained
Mil. Ambulatory Data	Tricare	speeding up access	High	8hrs / real-time	1-3 days	ICD-9	Zip Code	yes	active		99 to present	yes - for certain projects	yes - on Denver
Civ. Ambulatory Data	HMOs	per capita coverage	High	daily / real-time	1 day to a year	ICD-9	Zip Code	yes	active		11/99 to present	no	can be obtained
Mil. Prescription Meds.	Web MD		High	8 hrs / real-time	seconds	prescriptions	Zip Code	yes	active		7/01 to present	yes - for certain projects	yes - on Denver
Mil. Nurse Hotline	Tricare	legal difficulties in obtaining	High	daily / real-time	real-time	chief complaint	unknown	yes	in progress		in progress	not yet	can be obtained

Civ. Nurse Hotline	HMOs	redirection of FY02 funds	High	daily / real-time	real-time	chief complaint	Zip Code	yes	in progress		in progress	no	can be obtained
Poison Center	MD	not as useful as first thought	Low	daily / real-time	real-time	chief complaint	Phone Number	yes	inactive		99'	no	no
Electronic Vital Signs	VA Hospitals		High	daily / real-time	real-time	vital signs and ICD9	unknown	yes	in progress		in progress	?	?
Civ. Inpatient Data	Insurers (3 states)		High	daily	real-time*	ICD-9	ZIP res & work	yes	historical		99-02	no, privacy exempt	yes
Civ. Prescriptions	HMOs, Insurers		Hgh	daily	real-time*	prescriptions	ZIP res & work	yes	historical		99-02	no, privacy exempt	yes

Verdian

911 Call Codes	Norfolk Fire Rescue	Need state DOH approval	M	Continuous	Day	Minute	Address	Yes	MOU negotiated, submitted by Veridian 5/21/02 – follow-on arrangements in progress	Meeting with Fire Official Dave Palmer	Unknown	No	Yes
911 Call Codes	San Francisco Public Health	Negotiating level of spatial resolution	M	Continuous	1-4 weeks	Minute	Address or Census Block	Yes	Final approval pending	Resubmit to IRB	2 years + Current	Yes	Yes
Nurse Triage Data	Portsmouth Naval Medical Center ER		H	Continuous	1-2 days	Minute	Address	Yes	Agreement near	Brief Admiral Adams, arrange MOU between AFRL Rome and NMCP	1 years + Current	?	Yes
Military Sick Call	Military Acute Care Clinics		H	Continuous	1-2 days	Minute	Address	Yes	Agreement in principle	MOU	1 years + Current	?	Yes
Physician test orders	Portsmouth Naval Medical Center		H	Continuous	Daily, but not always closed out properly at end of day	Minute	Address	Yes	Agreement in principle	MOU	1 years + Current	?	Yes
Pharmaceuticals – Prescription	Portsmouth Naval Medical Center		H	Continuous	Day	Day	Address	Yes	Agreement in principle	MOU	1 years + Current	?	Yes
Pharmaceuticals – OTC	Palo Alto V.A. Med Center		H	Continuous	Day	Minute	City	Yes	Completed obtaining pilot data	Refine data extraction routines based on pilot data	Current	Yes	Yes

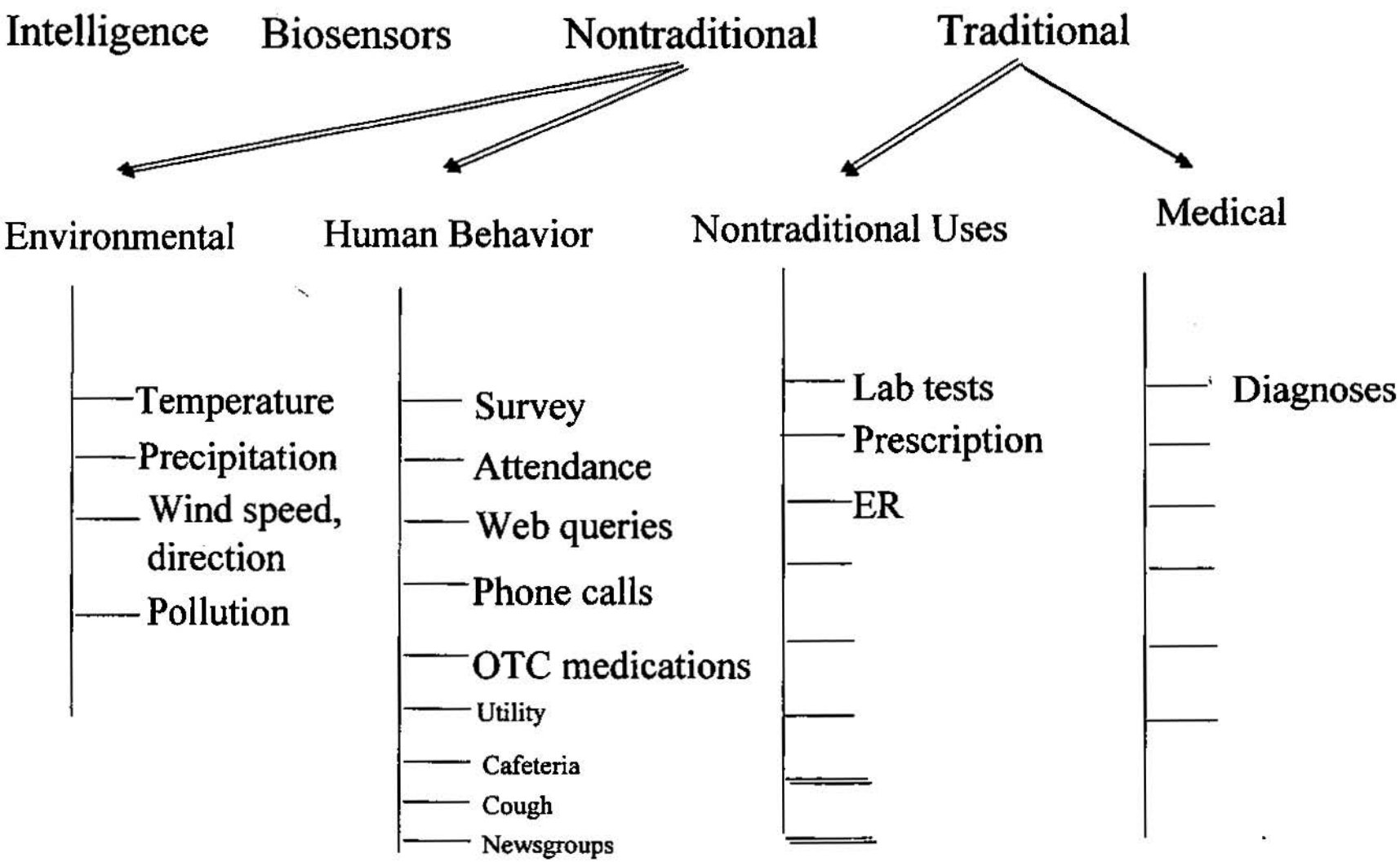
Traditional Data Sources

Data Source	Organization	Issues	Utility	Frequency of Collection	Latency	Temporal Resolution	Geographic localization / granularity	Electronic Data	Status	Next Step	Period (over what dates)	IRB?	Personal ID
IBM													
Insurance claims: ICD9 codes with associated service date, gender, age, zip code	Surveillance Data Inc.	Claims may come in slowly after service (1-4 days may be needed to detect an anomaly)	High for modeling background health, Moderate to low for detection	Real-time	A number of patients have been seen by physicians	Daily	Zipcode	Yes	Sample obtained for NY, Pgh, DC areas, 8/01 - 5/02, selected ICD9 codes	Obtain larger samples to model seasonality, variability between regions	8/01-5/02. Data goes back to 1998	No	Yes, but was anonymized. Linkable attributes were only examined in aggregate.
(b)(4)													
(b)(4)	Patient admission records			Real-time	Seconds	Timestamp	Address	Yes	Receiving data feed		5 years	Approved	Yes
	Patient admission records			Real-time	Seconds	Timestamp	Address	Yes	Receiving data feed		5 months	Approved	Yes
JHU/APL ESSENCE II Indicators													
Diagnostic Lab Results	State Labs	minimal coverage	Low	daily	Up to a month	# of positives	Zip Code	yes	active		11/99 to pres	yes	no
Diagnostic Lab Results	Private	coverage resolved	Moderate to	real-time	days	# of positives	Zip Code	yes	inactive		not yet	no	no
Nursing Homes	Private		Low	daily	same day	# of cases	exact location	some	inactive		11/99 to 4/20	no	no
Verdian													
Patient findings	Portsmouth Naval Medical Center		H	Continuous	Daily, but not always closed out properly at end of day	Minute	Address	Yes	Agreement near	MOU	1 year historical + current	?	Yes
Patient Syndromes	Santa Clara Public Health	Need to negotiate	H	Continuous	Day	Day	Address	Yes	Initial discussion	Continue discussion and identify access issues	Unknown	Yes	Yes
Pharmaceuticals – OTC	Palo Alto V.A. Med Center		H	Continuous	Day	Minute	City	Yes	Completed obtaining pilot data	Refine data extraction routines based on pilot data	Current	Yes	Yes
Pharmaceuticals – Prescription	Palo Alto V.A. Med Center		H	Continuous	Day	Minute	City	Yes	IRB approval granted	Work out data extraction logistics	Current	Yes	Yes
Patient findings	Palo Alto V.A. Med Center		H	Continuous	Day	Minute	City	Yes	IRB approval granted	Work out data extraction logistics	Current	Yes	Yes
Findings, Physician test orders, Chief Complaints, Nurse Triage Data													
Physician test orders	Palo Alto V.A. Med Center		H	Continuous	Day	Day	Address	Yes	IRB approval granted	Work out data extraction logistics	2-Jun		

Traditional Data Sources

Data Source	Organization	Issues	Utility	Frequency of Collection	Latency	Temporal Resolution	Geographic localization / granularity	Electronic Data	Status	Next Step	Period (over what dates)	IRB?	Personal ID
IBM													
Insurance claims: ICD9 codes with associated service date, gender, age, zip code	Surveillance Data Inc.	Claims may come in slowly after service (1-4 days may be needed to detect an anomaly)	High for modeling background health, Moderate to low for detection	Real-time	A number of patients have been seen by physicians	Daily	Zipcode	Yes	Sample obtained for NY, Pgh, DC areas, 8/01 - 5/02, selected ICD9 codes	Obtain larger samples to model seasonality, variability between regions	8/01-5/02. Data goes back to 1998	No	Yes, but was anonymized. Linkable attributes were only examined in aggregate.
(b)(4)													
(b)(4)	Patient admission records			Real-time	Seconds	Timestamp	Address	Yes	Receiving data feed		5 years	Approved	Yes
	Patient admission records			Real-time	Seconds	Timestamp	Address	Yes	Receiving data feed		5 months	Approved	Yes
JHU/APL ESSENCE II Indicators													
Diagnostic Lab Results	State Labs	minimal coverage	Low	daily	Up to a month	# of positives	Zip Code	yes	active		11/99 to pres	yes	no
Diagnostic Lab Results	Private	coverage resolved	Moderate to	real-time	days	# of positives	Zip Code	yes	inactive		not yet	no	no
Nursing Homes	Private		Low	daily	same day	# of cases	exact location	some	inactive		11/99 to 4/20	no	no
Verdian													
Patient findings	Portsmouth Naval Medical Center		H	Continuous	Daily, but not always closed out properly at end of day	Minute	Address	Yes	Agreement near	MOU	1 year historical + current	?	Yes
Patient Syndromes	Santa Clara Public Health	Need to negotiate	H	Continuous	Day	Day	Address	Yes	Initial discussion	Continue discussion and identify access issues	Unknown	Yes	Yes
Pharmaceuticals – OTC	Palo Alto V.A. Med Center		H	Continuous	Day	Minute	City	Yes	Completed obtaining pilot data	Refine data extraction routines based on pilot data	Current	Yes	Yes
Pharmaceuticals – Prescription	Palo Alto V.A. Med Center		H	Continuous	Day	Minute	City	Yes	IRB approval granted	Work out data extraction logistics	Current	Yes	Yes
Patient findings	Palo Alto V.A. Med Center		H	Continuous	Day	Minute	City	Yes	IRB approval granted	Work out data extraction logistics	Current	Yes	Yes
Findings, Physician test orders, Chief Complaints, Nurse Triage Data													
Physician test orders	Palo Alto V.A. Med Center		H	Continuous	Day	Day	Address	Yes	IRB approval granted	Work out data extraction logistics	2-Jun		

Biosurveillance Data Space



TITLE PAGE

Briefing Materials concerning TIA BIOALERT 1

FOLDER 2.c TOC 7

DETECTION

- 2.c.1 Detection Method Template
- 2.c.2 Bio-Surveillance Panel on Detection Strategies
    - IBM T.J. Watson Research Center, JHU
- 2.c.3 Summary on Detection Algorithms
    - Johns Hopkins U Applied Physics Lab
    - GEIS
- 2.c.4 Bio-surveillance Detection Progress: June 2002
    - Pitt-CMU
- 2.c.5 Detection Approaches and Algorithms
    - Veridian-Stanford Team 2nd PI Meeting

2.C.1

Detection Method	Advantages	Disadvantages	Any Venue?	Infection Path?	Which Teams tried it at all?	Tried but little used	Tried and used	Under development
Time-weighted averaging	Quick and easy	Univariate			Pitt/Mellon	Pitt/Mellon		
Serfling	Quick and easy	Univariate			Pitt/Mellon		Pitt/Mellon	
ARIMA	Quick and easy	Univariate			Pitt/Mellon	Pitt/Mellon		
SARIMA + External Factors	Quick and easy	Univariate			Pitt/Mellon		Pitt/Mellon	
Univariate HMM	Quick and easy	Univariate			Pitt/Mellon	Pitt/Mellon		
Univariate Kalman Filter	Quick and easy	Univariate			Pitt/Mellon	Pitt/Mellon		
Recursive Least Squares	Quick and easy	Univariate						
Support Vector Machine	Nonlinear	Depends how it's used			Pitt/Mellon	Pitt/Mellon		
Neural Nets	Nonlinear	Depends how it's used			Pitt/Mellon	Pitt/Mellon		
CuSUM	Easy	Mainly for financial			Pitt/Mellon	Pitt/Mellon		
Randomization	Important	Expensive			Pitt/Mellon		Pitt/Mellon	
Spatial Scan Statistics	Sensitive				Pitt/Mellon		Pitt/Mellon	Pitt/Mellon
Bayesian Networks	Good probabilistic representation	Where to get structure from			Pitt/Mellon		Pitt/Mellon	
Contingency Tables	Classic	Needs discretization			Pitt/Mellon		Pitt/Mellon	
Scalar Outlier (SQC)	Classic, very easy	Misses multivariate stuff. Can't handle cyclic or drifting data			Pitt/Mellon		Pitt/Mellon	
Multivariate Anomalies	Well-founded	Only looks at individual cases			Pitt/Mellon		Pitt/Mellon	
Change-point statistics	Very sensitive	univariate			Pitt/Mellon		Pitt/Mellon	
False Discovery Rate Tests	Useful if multiple testing				Pitt/Mellon		Pitt/Mellon	
WSARE (Recent patterns)	Generic	Tuned for general anomalies…not specific threats			Pitt/Mellon		Pitt/Mellon	Pitt/Mellon
PANDA (Causal Model)	Probabilistic	Currently only univariate			Pitt/Mellon		Pitt/Mellon	Pitt/Mellon
FLUMOD (space/Time HMM)	May help model flu background	Not yet tested			Pitt/Mellon		Pitt/Mellon	Pitt/Mellon
Multivariate Kalman					Not Pitt/Mellon			

Bio-Surveillance
Panel on Detection Strategies

IBM Thomas J. Watson Research Center
Johns Hopkins University

Detection Techniques

Probabilistic	**AR Time series** Geo-differential **Kalman filter** **Graphical Models (Factor Graphs)** Demographic discriminator Unsupervised Clustering
Deterministic	Regression **Kernel-methods (SVM)** Holt-winter forecasting

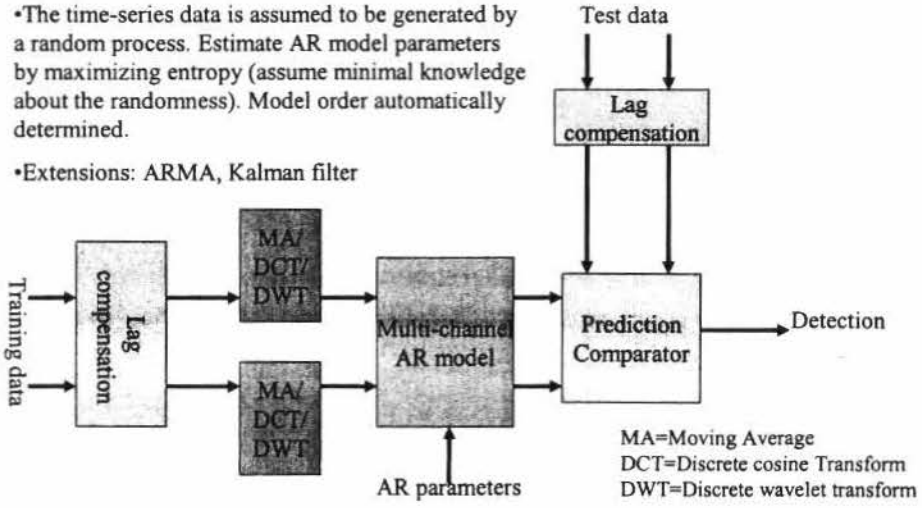
Unsupervised Incremental Clustering

- A statistical algorithm (variant of *incremental k-means*).
  - Start with N clusters one of which is seeded with high values of source features.
  - Update clusters based on data source values observed till date.
  - Choose that cluster as abnormal which has largest norm (L1).
- Characteristics.
  - No use of prior or historical data of normal behavior or other outbreak behavior.
  - No online supervision for detecting abnormalities either.
- Extensions.
  - Markov chain based smoothing.
  - Historical normal data to develop robust "normal clusters."

Auto-regressive Prediction-Based Model

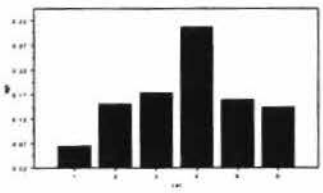
•The time-series data is assumed to be generated by a random process. Estimate AR model parameters by maximizing entropy (assume minimal knowledge about the randomness). Model order automatically determined.

•Extensions: ARMA, Kalman filter

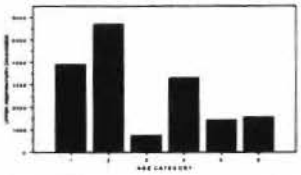


Demographic Based Detection

CASE STUDY: age distribution



COUNTY DEMOGRAPHIC PROFILE →



COUNTY UPPER RESPIRATORY ICD9 PROFILE

ASSUMPTIONS

- Cases will be reflective of exposed and/or susceptible population
- **Reporting biases will be altered in a catastrophic event**
- Early diagnoses will fall into upper respiratory ICD9s

ADVANTAGES

- Age distribution of common illnesses consistent over time
- Sliding window controls for seasonal effects

LIMITATIONS

- Success depends on timeliness of data feed; reporting delay distribution needs to be modeled
- **Age-biases for physician's visits pre- and post-attack are not understood**

Geo-Differential Detection

- Detect local attack in a target region by comparing with unaffected reference region.
  - Model and quantify normal variation in tracking.
  - Detect significant deviations in the tracking.
- Used phone call monitoring at two IBM research sites at Yorktown, Hawthorne which are 10 miles apart.
- Simplifying Poisson process assumption: Calls/callers on successive days are not always independent.
- Define two time windows.
  - ←History window→←test window→.
- Model ratio of counts in a window for target and reference regions.
- Use chi-squared statistic to detect deviations.
- Pros & cons.
  - Does not model underlying phenomena.
  - Requires good tracking and moderate volumes.

Adaptive Forecasting

- Method
  - Forecast one-day ahead based on a modified **Holt-Winters forecaster**
  - Compute threshold based on differences between observed and forecast data
  - Trigger alerts if an observed data point is greater than the threshold
  - Replace an outlier with an extrapolation for future forecast
- Pros & Cons
  - Handles base, trend, day of week seasonality, special days, missing data
  - No preprocessing is needed
  - Single data channel detector
    - –Data channel must exhibit good correlation with attacks
    - –Normal level activities must be satisfactorily forecasted

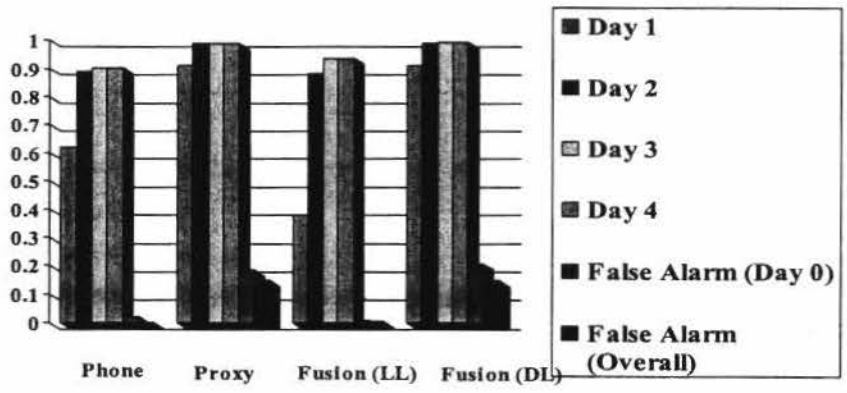
Data Sources and Attack Simulation

- Data Sources.
  - Number of Phone calls to medical centers.
  - Web proxy logs from Watson to specific medical Web sites.
  - SDI Data: Insurance Claims in doctors' offices and hospitals.
- Attack Simulations.
  - Not more than one attack in a scenario.
  - Multi-channel coordinated attack (Jan 02 2002-March 30 2002).
  - 1000 runs of attack scenarios with varying.
    - date of attack.
    - Relative strength of attack (0-30 %).
    - Number infected grows according to anthrax epi-curve over a 5 day period.
- Average detection and false alarm rate over 1000 attack scenarios were computed.
- Average Lag: How many days after the first date of attack is the attack detected.

Results: Summary for Simulated Attacks

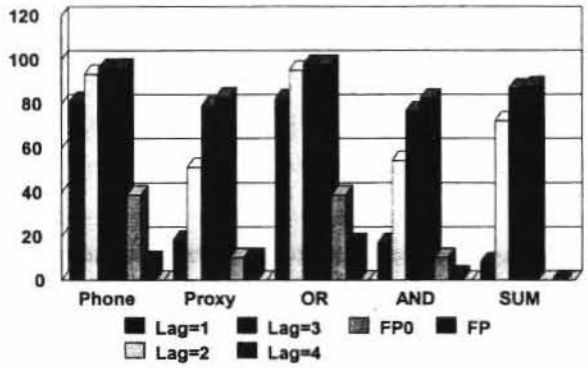
	Supervision	Detection (Day 3)	False Alarm	Lag for exceeding 90 % detection	# channels used
Clustering	No	98.77	14.7	Day 2	2
Auto-regressive	Yes	95.2	17.2	Day 3	2
Adaptive Forecasting	No	77.2	7.1	Day 4	1
Demographic	No	94	1.5	Day 2	1

Results: Unsupervised Clustering



Results AR

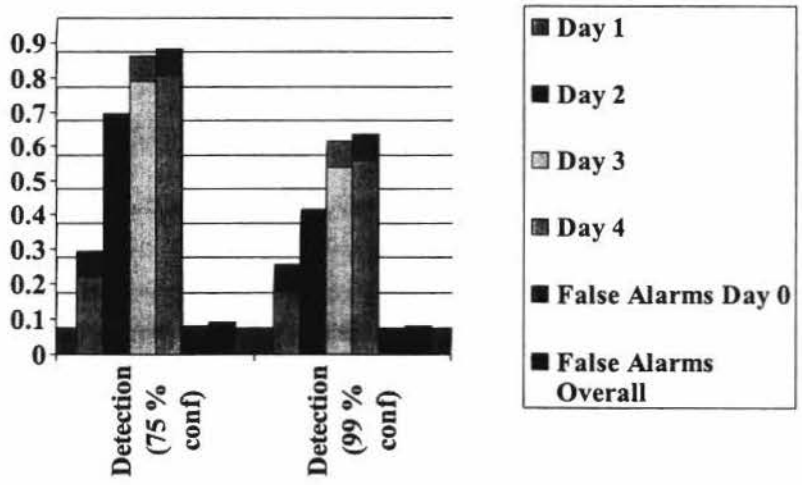
Two-channel (phone + proxy) channel results



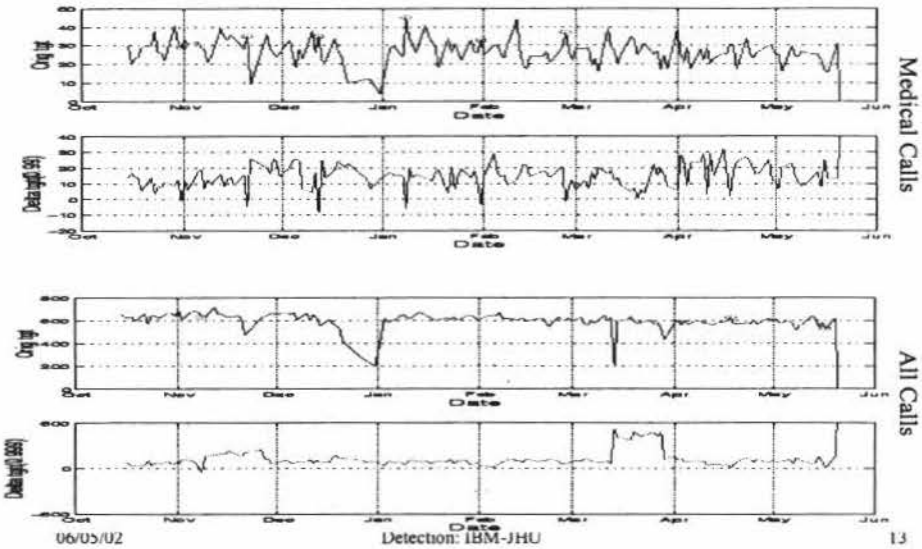
Confidence Level = 90%

rank 1: 100%

Demographic Based Detection

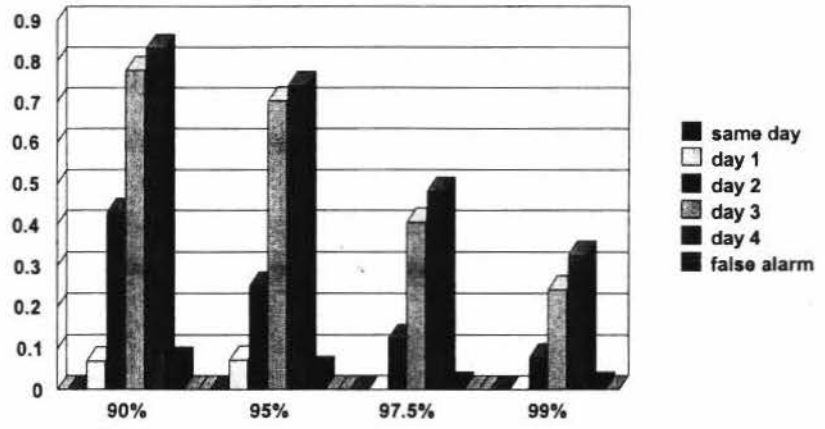


Differential Detection



Adaptive Forecasting Results

Detection of Simulated Attacks on Proxy Log



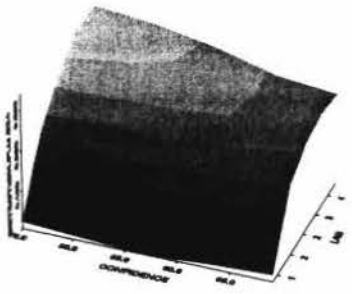
Cough Detection

- 10-30 ms audio features
- HMM-based detection technique
- Detection in the 95 % + range with false alarms in less than 10 % range
- Detection suffers if the background is a cafeteria as against a meeting room
- Multimodality can help make this usable
  - Visual assistance in assigning cougher ids..
  - Techniques borrowed from audiovisual detection of events such as "explosions", "gunshots" for video databases which is a different application domain.

BACKUP

RESULTS: INFLUENZA ATTACK SIZE 524

MAX 0.34



CONFIDENCE

LAG

DETECTION RATES

CONFIDENCE / LAG	75%	90%	95%	99%
1	.02	.02	.02	.02
2	.22	.21	.21	.21
3	.32	.27	.24	.23
4	.34	.28	.24	.23

CONFIDENCE	FALSE ALARM RATE
75%	.005
90%	.005
95%	.005
99%	negligible
% of false alarms on lag 0	0.015
on lags 5-7	0.004

Challenges

- Handling variations under normal conditions.
  - Base, trend, day of week seasonality, missing data, holidays and special-day events are automatically handled.
- Amount of normal data required to bootstrap.
- Incremental Update for algorithms.

JHU – GEIS Summary on Detection Algorithms

Howard Burkom & Steve Magruder

JHU/APL

Yevgeniy Elbert

GEIS

Algorithms Considered

- Temporal (Useful for Diffuse Outbreaks)
  - CUSUM, Based on EWMA Smoothing
  - Periodic Moving Average
  - AR*
- Space-Time (Useful for Localized Outbreaks)
  - Kulldorff Scan Statistic*, Conventional and Stratified
  - Odds Ratio
  - (Adaptive) Matched Filter
  - D^2

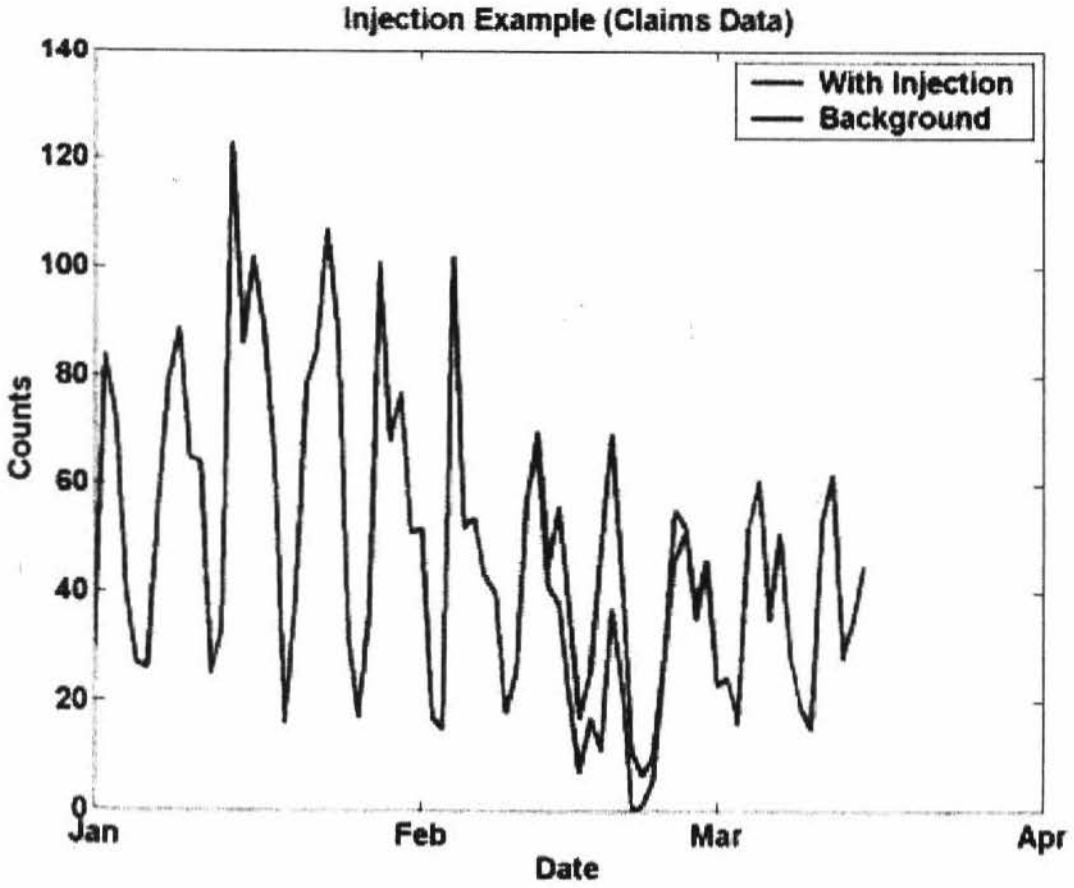
Discussed in Data Fusion Session

* Current Focus of Development

Temporal Algorithm Challenges

- DOW Effect (& Holidays)
- Background Disease / Short-Term Trends
- Unpredicted Aberrant Events

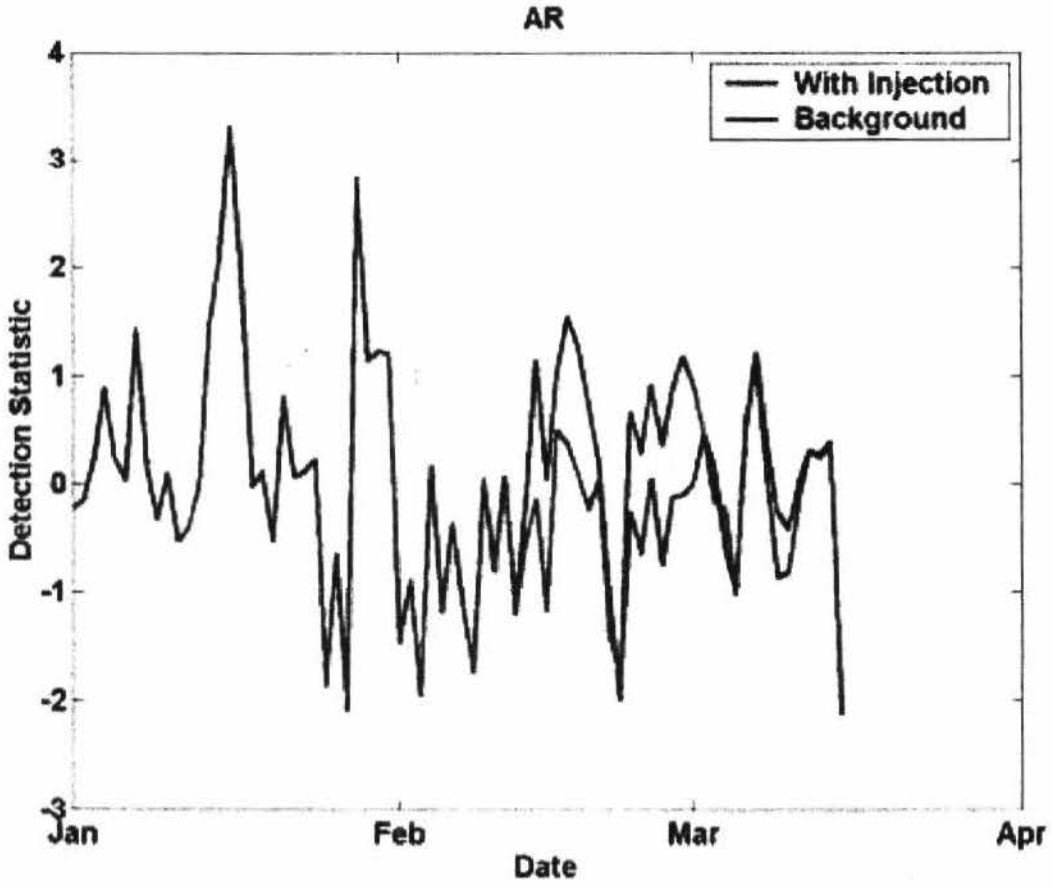
Signal-Injection Example



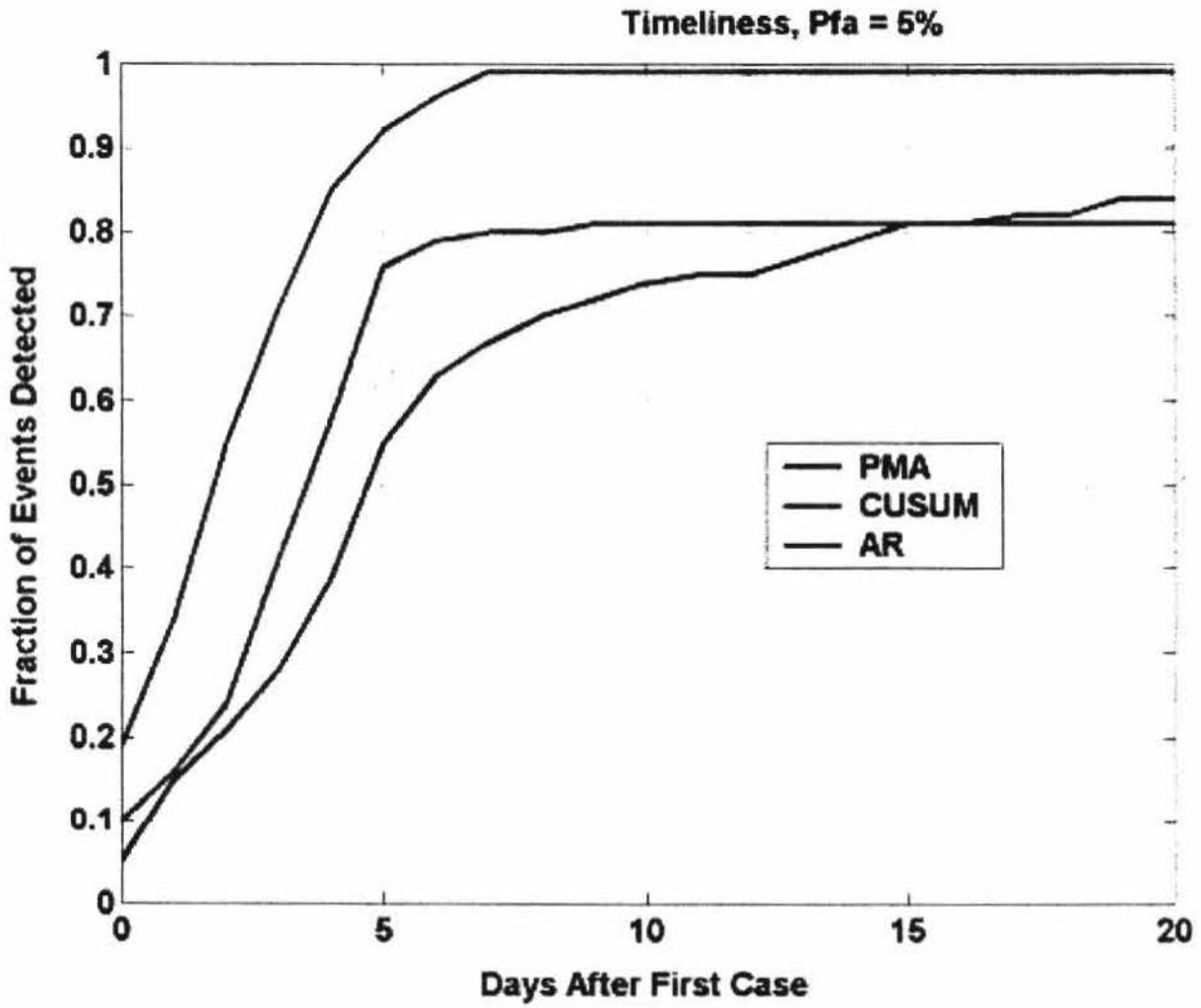
Injection:
200 Smallpox Cases

Data Set:
Winter '02,
Military Claims,
XX County,
29 "ili" codes

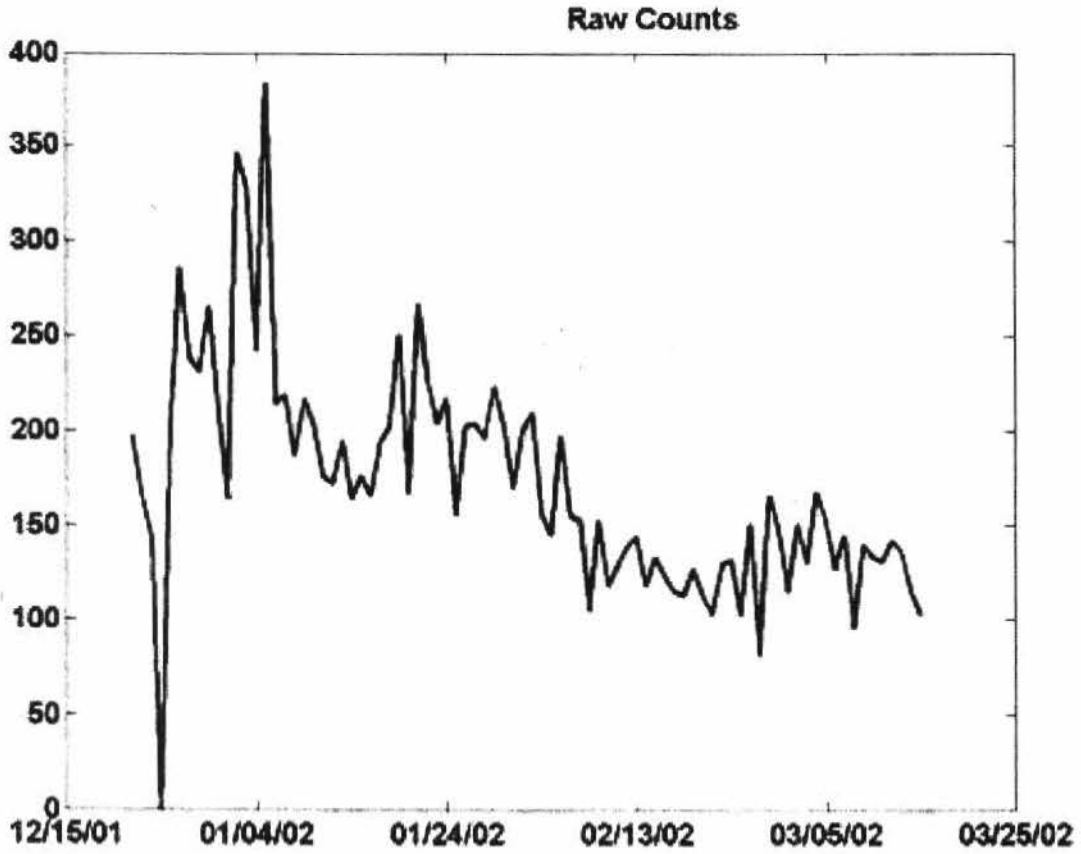
AR Output


AR
With Injection
Background
Detection Statistic
4
3
2
1
0
-1
-2
-3
Jan
Feb
Mar
Apr
Date

Summary After 100 Injections



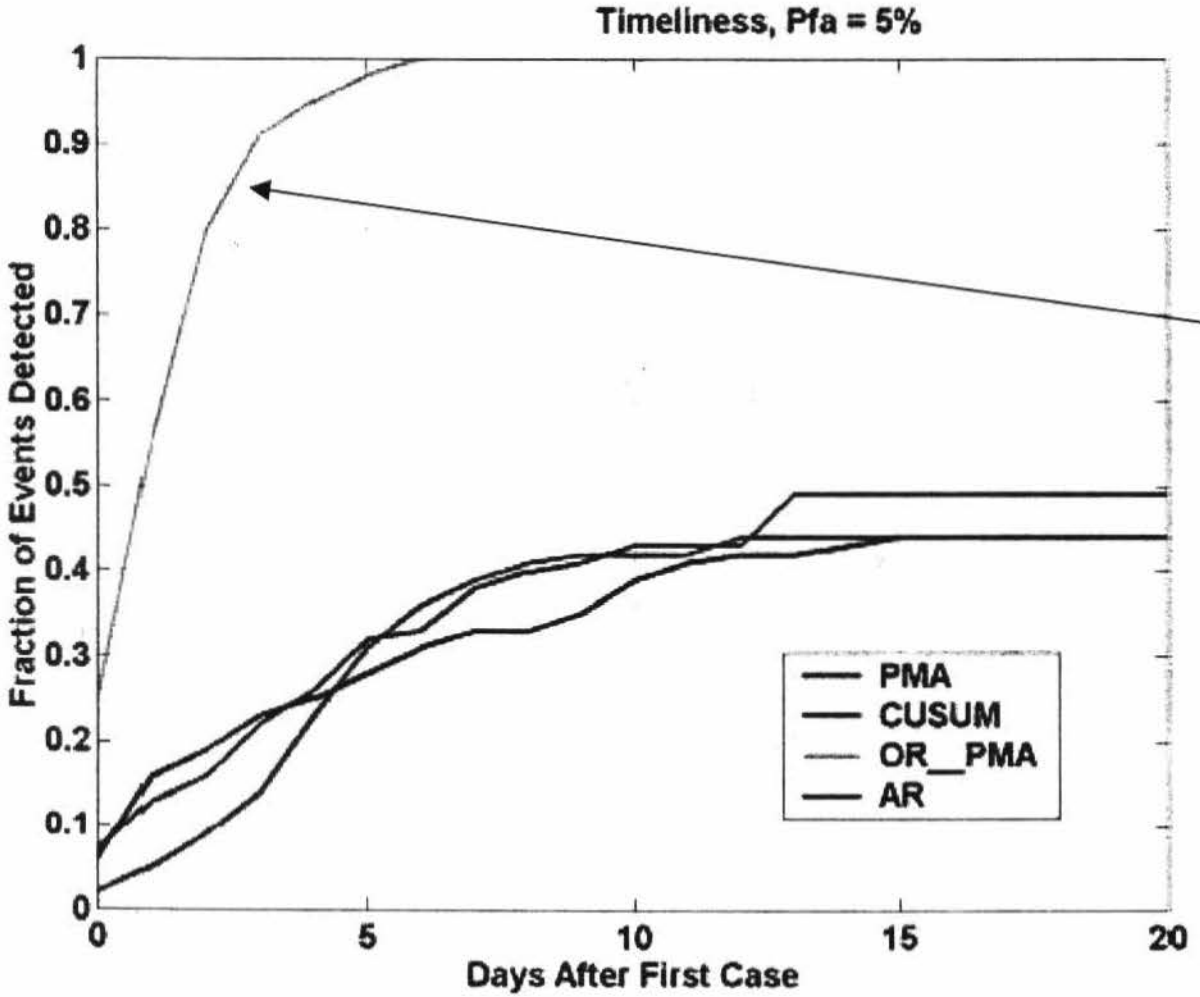
Some Shorter-Time Trends in OTC



Injection: 200
Smallpox Cases

Data Set:
Winter '02,
OTC anti-FLU
XX County

Detector Performance on OTC Data

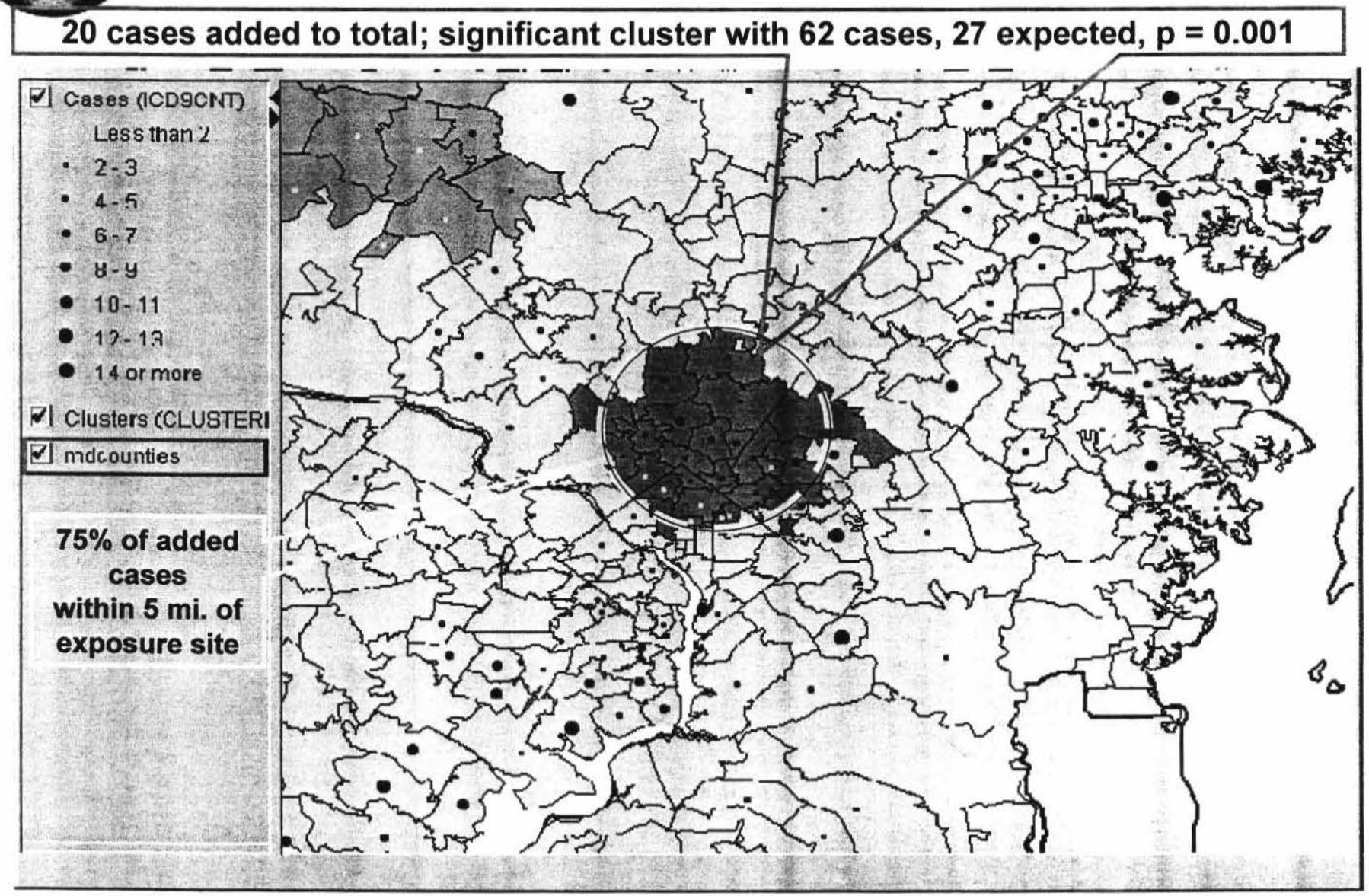


"OR" Uses Spatial Info!

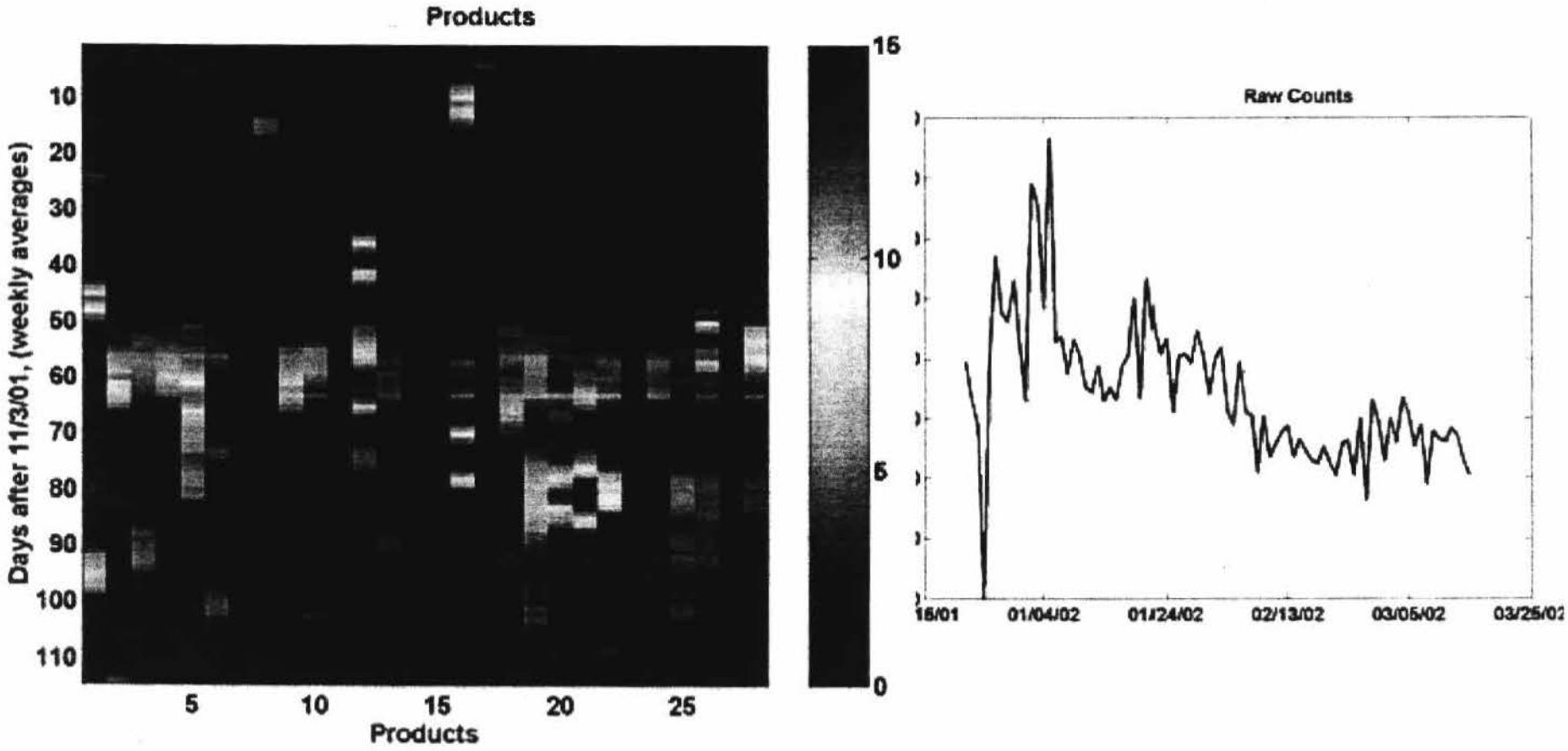
Spatial Detection / Clustering

- First Order Issue Is Matching of Spatial Filter Structure to Outbreak Spatial Structure
- Scan Statistic Approach Address This Problem
  - "Kulldorff" Statistic Seems to do Best
- Generalized Scan Statistic Approach:
  - Several Statistics Used to Combine Multiple Data Sets (See Later Session)

Simulation Day 9: Abnormal Cluster Indicated Fri. Dec. 21 (739 cases)



Another Dimension of Structure



Siegrist's Questions

- **Advantages, disadvantages and applicability of each approach they use in context (eg, data sources, attack scenarios).**
  - AR can adapt quickly to trends, DOW, but still responds on shorter time scales to a new outbreak. Doesn't make many assumptions.
  - Spatial techniques exploit outbreak spatial structure. Kulldorff scan statistic is robust to outbreak scale & location, and supports data fusion. Also makes few assumptions.
  - Additional dimensionality provides additional opportunities
- **Overcoming the false alarm problem.**
  - Tune detectors to reject time scales not of interest
  - Remaining FA rejection depends on accumulation of information (multiple data sources, or multiple descriptors within a data source)
  - Some FAs will require human investigation

Siegrist's Questions (Cont'd)

- **Are there important advantages in complementary detection approaches, rather than only one best way?**
  - Simultaneous temporal and space-time is one example.
- **What experiments could be performed to determine the answer?**
  - A working surveillance system, stimulated with simulated outbreaks is the best we can do.
- **What metrics are best to measure effectiveness?**
  - Minimum outbreak level that can be detected and classified, for a given false alert level & time late, for various outbreak scenarios.
- **Others? (Q: How can we improve sensitivity further?)**
  - Exploit structure on additional dimensions

Pitt/CMU Biosurveillance Detection Progress: June 2002

Center for Biomedical Informatics
University of Pittsburgh and Carnegie Mellon University

Greg Cooper (U Pitt, CBMI)
Jeremy Espino (U Pitt, CBMI)
Andrew Moore (CMU, Comp Sci)
Bob Olszewski (U Pitt, CBMI)
Jeff Schneider (CMU, Robotics)
Rich Tsui (U Pitt, CBMI)
Mike Wagner (U Pitt, CBMI)
Weng-Keen Wong (CMU Comp Sci)

Contact:
Andrew Moore
awm@cs.cmu.edu
412-268-7599
www.cs.cmu.edu/~awm

Pitt/CMU Status

Method	Has Pitt/CMU tried it?	Tried but little used	Tried and used	Under development	Multivariate signal tracking?	Spatial?
Time-weighted averaging	Yes	Yes				
Serfling	Yes		Yes			
ARIMA	Yes	Yes				
SARIMA + External Factors	Yes		Yes			
Univariate HMM	Yes		Yes			
Kalman Filter	Yes	Yes				
Recursive Least Squares	Yes		Yes			
Support Vector Machine	Yes	Yes				
Neural Nets	Yes	Yes				
CuSUM	Yes	Yes				
Randomization	Yes		Yes	Yes		
Spatial Scan Statistics	Yes			(w/ Howard Burkom)	Yes	Yes
Bayesian Networks	Yes			Yes	Yes	
Contingency Tables	Yes		Yes			
Scalar Outlier (SQC)	Yes	Yes				
Multivariate Anomalies	Yes		Yes		Yes	
Change-point statistics	Yes			Yes		
FDR Tests	Yes		Yes		Yes	
WSARE (Recent patterns)	Yes		Yes	Yes	Yes	Yes
PANDA (Causal Model)	Yes			Yes	Yes	Yes
FLUMOD (space/Time HMM)				Yes	Yes	Yes

Details of these methods and bibliography available from "Summary of Biosurveillance-relevant statistical and data mining technologies" by Moore, Cooper, Tsui and Wagner. Downloadable (PDF format) from www.cs.cmu.edu/~awm/biosurv-methods.pdf

WSARE v2.0

- What's Strange About Recent Events?
- Designed to be easily applicable to any date/time-indexed biosurveillance-relevant data stream.

3

WSARE v2.0

- Inputs:
  - 1. Date/time-indexed biosurveillance-relevant data stream
  - 2. Time Window Length
  - 3. Which attributes to use?

4

WSARE v2.0

- Inputs:
  - 1. Date/time-indexed biosurveillance-relevant data stream
  - 2. Time Window Length
  - 3. Which attributes to use?

"last 24 hours"

"ignore key and weather"

Example

Primary Key	Date	Time	Hospital	ICD9	Prodrome	Gender	Age	Home			Work			Recent Flu Levels	Recent Weather	(Many more...)
								Large Scale	Medium Scale	Fine Scale	Large Scale	Medium Scale	Fine Scale			
h6r32	6/2/2	14:12		781	Fever	M	20s	NE	15217	A5	NW	15213	B8	2%	70R	...
t3q15	6/2/2	14:15		717	Respiratory	M	60s	NE	15222	J3	NE	15222	J3	2%	70R	...
t5hh5	6/2/2	14:15		322	Respiratory	F	80s	SE	15210	K9	SE	15210	K9	2%	70R	...
:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:

5

(b)(4)

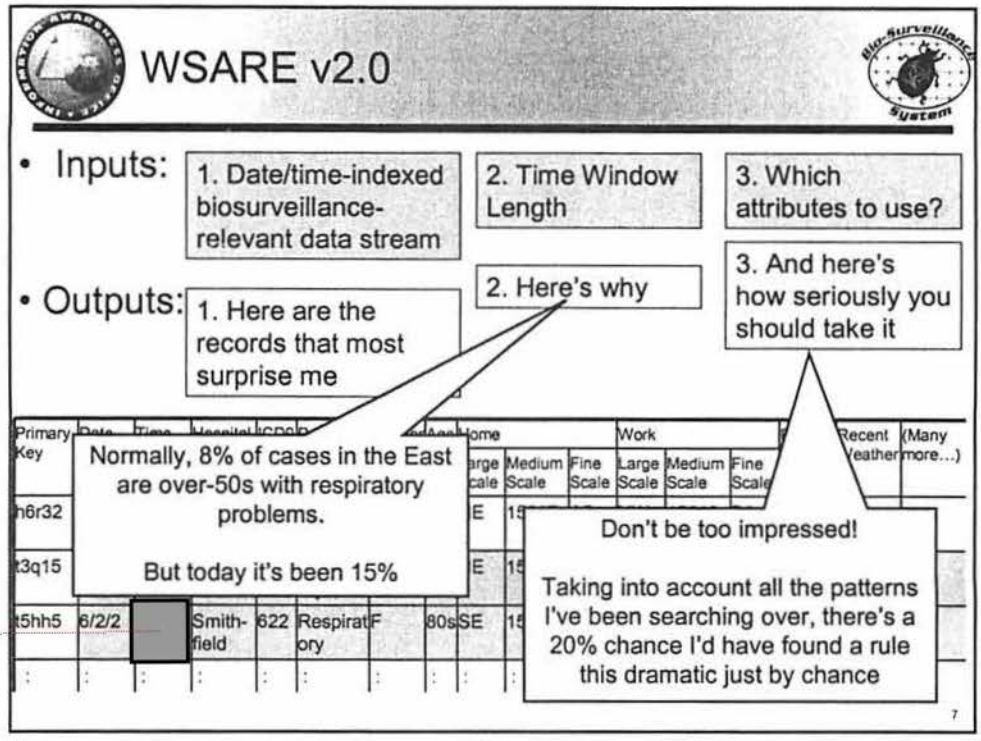
WSARE v2.0

- Inputs:
  - 1. Date/time-indexed biosurveillance-relevant data stream
  - 2. Time Window Length
  - 3. Which attributes to use?
- Outputs:
  - 1. Here are the records that most surprise me
  - 2. Here's why
  - 3. And here's how seriously you should take it

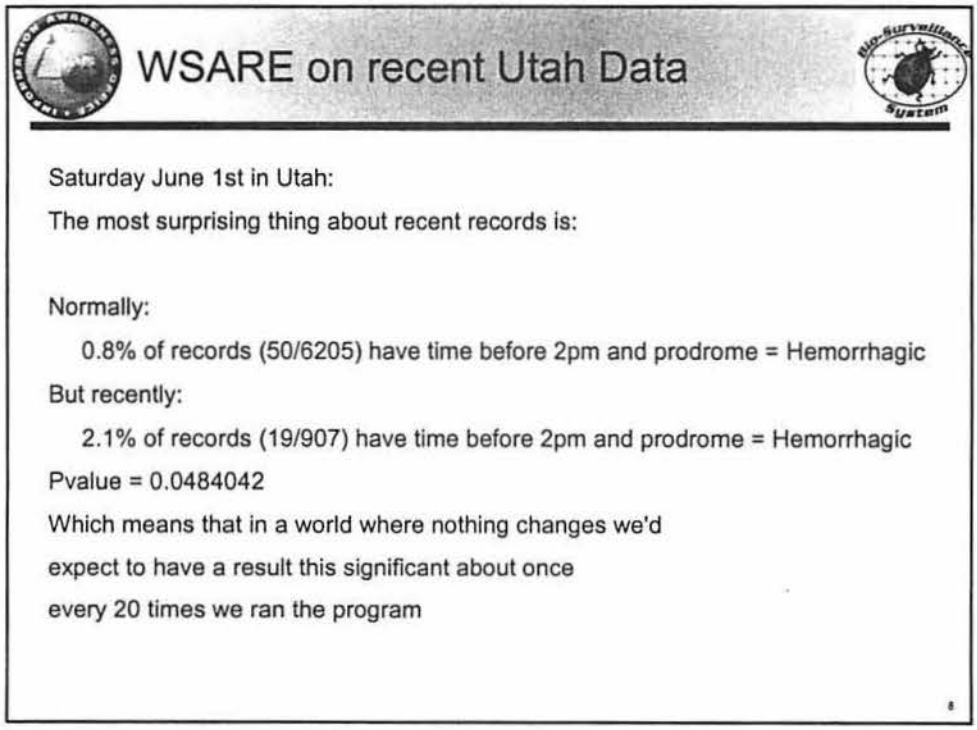
Primary Key	Date	Time	Hospital	ICD9	Prodrome	Gender	Age	Home			Work			Recent Flu Levels	Recent Weather	(Many more...)
								Large Scale	Medium Scale	Fine Scale	Large Scale	Medium Scale	Fine Scale			
h6r32	6/2/2	14:12		781	Fever	M	20s	NE	15217	A5	NW	15213	B8	2%	70R	...
t3q15	6/2/2	14:15		717	Respiratory	M	60s	NE	15222	J3	NE	15222	J3	2%	70R	...
t5hh5	6/2/2	14:15		322	Respiratory	F	80s	SE	15210	K9	SE	15210	K9	2%	70R	...
:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:

6

(b)(4)


WSARE v2.0
• Inputs:
1. Date/time-indexed biosurveillance-relevant data stream
2. Time Window Length
3. Which attributes to use?
• Outputs:
1. Here are the records that most surprise me
2. Here's why
3. And here's how seriously you should take it
Normally, 8% of cases in the East are over-50s with respiratory problems.
But today it's been 15%
Don't be too impressed!
Taking into account all the patterns I've been searching over, there's a 20% chance I'd have found a rule this dramatic just by chance
Primary Key
Home
Work
Recent Weather
(Many more...)
Large Scale
Medium Scale
Fine Scale
Large Scale
Medium Scale
Fine Scale
h6r32
t3q15
t5hh5
6/2/2
Smith-field
622
Respiratory
F
80s
SE
7

(b)(4)

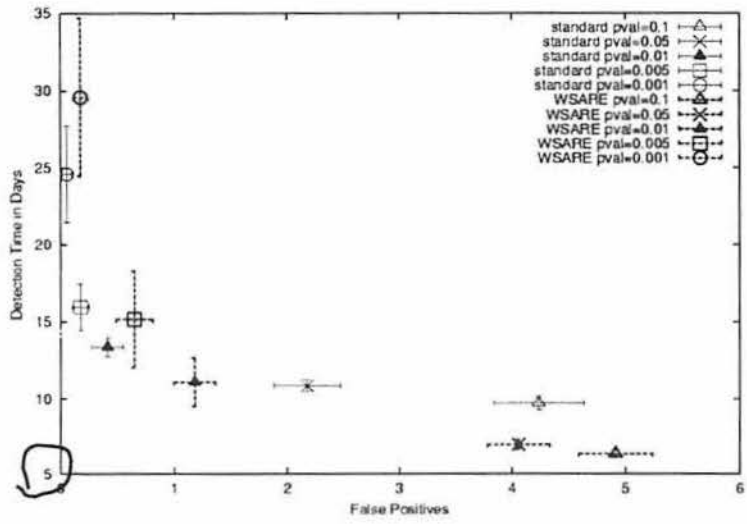

WSARE on recent Utah Data
Saturday June 1st in Utah:
The most surprising thing about recent records is:
Normally:
0.8% of records (50/6205) have time before 2pm and prodrome = Hemorrhagic
But recently:
2.1% of records (19/907) have time before 2pm and prodrome = Hemorrhagic
Pvalue = 0.0484042
Which means that in a world where nothing changes we'd
expect to have a result this significant about once
every 20 times we ran the program
8

GRIDSIM-based validation

- We need to understand sensitivity and specificity.
- Have developed a very fast simulator in which people go to work, go out for food, live in families.
- Background of mild diseases: some food-borne, some occupational, some contagious etc.
- Warning: Simulation designed for speed and complex dynamics: not reality (will use Carley's and Fridsma's work for that)
- A nasty disease is introduced at some point.
  - Can WSARE detect it?
  - How does WSARE compare to a standard sensible epidemiological alternative?

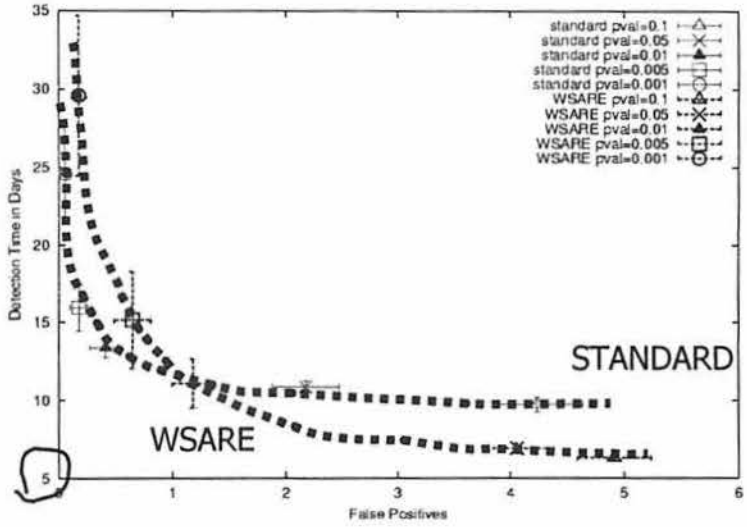
9

WSARE Simulation Sensitivity vs Specificity on GRIDSIM



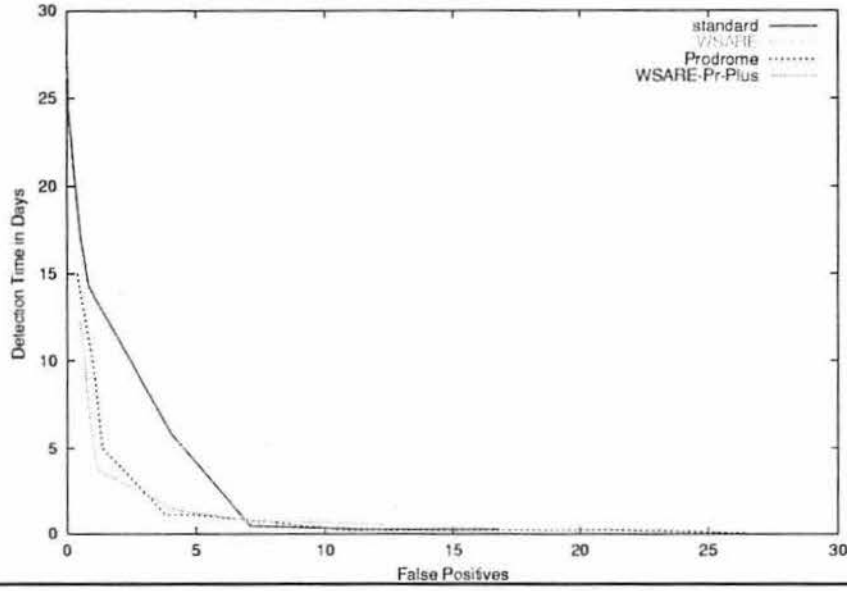
10

WSARE Simulation Sensitivity vs Specificity on GRIDSIM



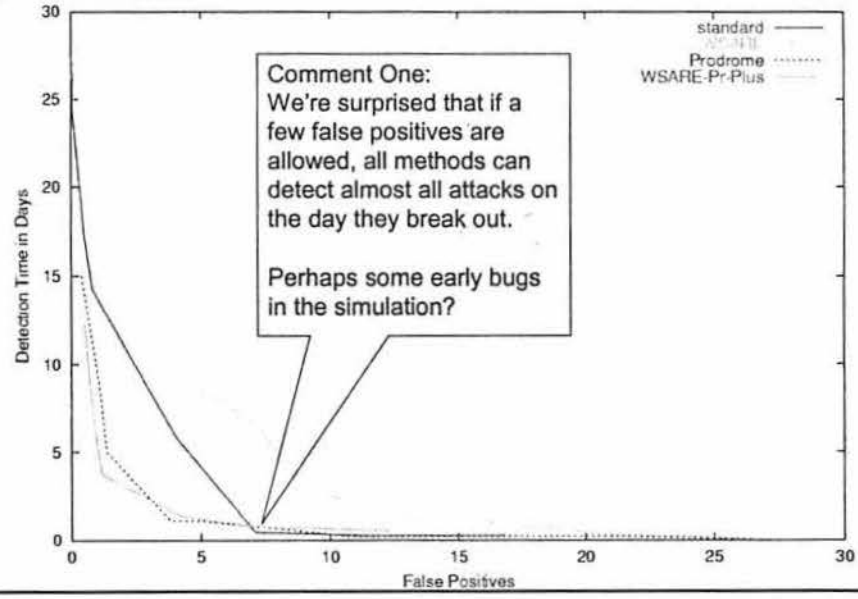
11

Validating WSARE on "Biowar"



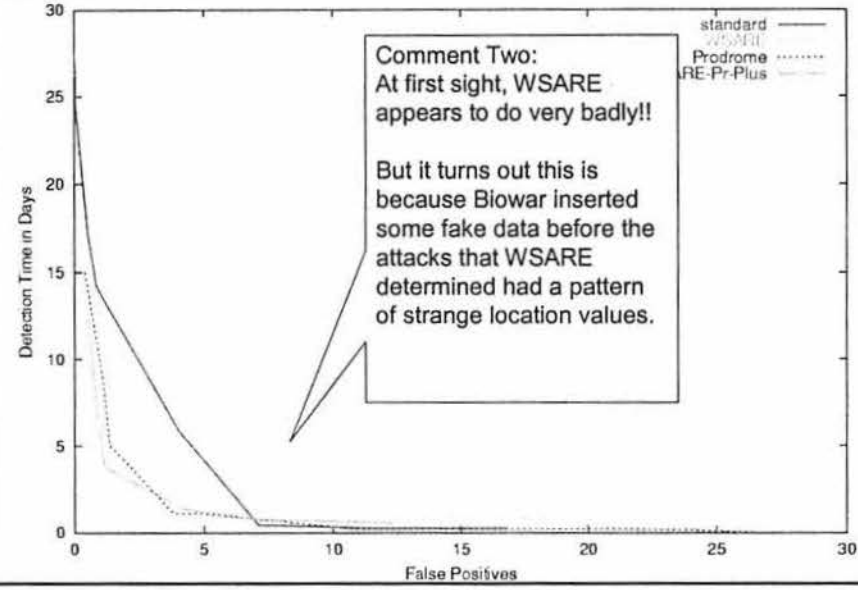
12

Validating WSARE on "Biowar"



13

Validating WSARE on "Biowar"



14

WSARE 3.0 [Prototype scheduled for September 2002]

- ANOVA-style WSARE

- Suppose it is very very hot weather.
- WSARE v2 would report "I'm very surprised to see so many people coming in with heat symptoms"
- WSARE v3 will report "Heat symptoms are high but explained by the weather"

15

WSARE 3.0 [Prototype scheduled for September 2002]

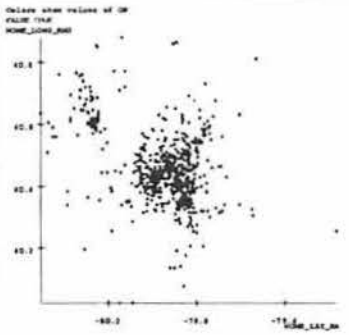
- "Taking into account recent flu levels..."
- "Taking into account that today is a public holday..."
- "Taking into account that this is Spring..."
- "Taking into account recent heatwave..."
- "Taking into account that there's a known natural Food-borne outbreak in progress..."

- Accomplished by
- Not comparing against mediumly recent events
- But instead against the predicted Joint Distribution of attributes predicted by a Bayes Net that has been trained on all historical data

Bonus: More efficient use of previous data

16

QSCAN (Quick SCAN Statistics) [Collaboration with Howard Burkom]



Standard scan statistic question:
Given the geographical locations of occurrences of a phenomenon, is there a region with an unusually high (low) rate of these occurrences?

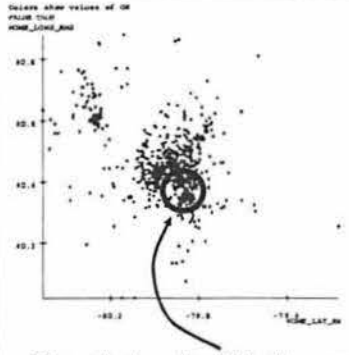
Standard approach:

1. Compute the likelihood of the data given the hypothesis that the rate of occurrence is uniform everywhere, L_0
2. For some geographical region, W, compute the likelihood that the rate of occurrence is uniform at one level inside the region and uniform at another level outside the region, L(W).
3. Compute the likelihood ratio, $L(W)/L_0$
4. Repeat for all regions, and find the largest likelihood ratio. This is the scan statistic, S^*_W
5. Report the region, W, which yielded the max, S^*_W

See [Glaz and Balakrishnan, 99] for details

17

QSCAN: significance testing



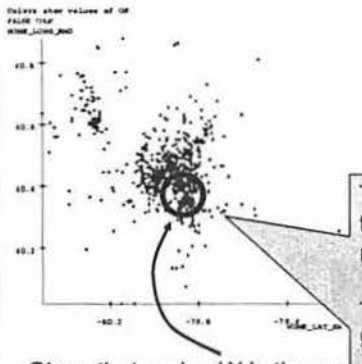
Given that region W is the most likely to be abnormal, is it significantly abnormal?

Standard approach:

1. Generate many randomized versions of the data set by shuffling the labels (positive instance of the phenomenon or not).
2. Compute S^*_W for each randomized data set. This forms a baseline distribution for S^*_W if the null hypothesis holds.
3. Compare the observed value of S^*_W against the baseline distribution to determine a p-value.

18

QSCAN: significance testing



Given that region W is the mo
likely to be abnormal, is it
significantly abnormal?

Standard approach:

1. Generate many randomized versions of the data set by shuffling the labels (positive instance of the phenomenon or not)

The log likelihood for the data assuming this region has a difference occurrence rate is -15406.

That value is higher than that found in 99% of the randomized cases.

Therefore, we conclude that there is a different occurrence rate in this region with a p value of 0.01.

Sound the ALARM!!!

ɔmized
line
l

e of S^*_W
tion to

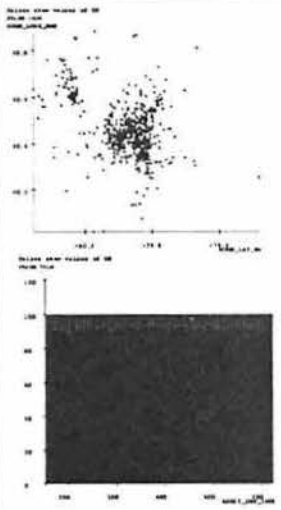
19

QSCAN: computational efficiency

- Store the data points in a kd-tree
  - result: order of magnitude speedup on the 70K record data set.
- Store the regions in a kd-tree (using center location and radius)
  - result: a similar order of magnitude speedup
- Store both the regions and the data in a kd-tree
  - result: surprisingly, no further speedup
- Store and process all randomized labels together in the kd-trees
  - result: computation is sub-linear in the number of randomized trials

20

QSCAN: compensating for inherent spatio-temporal fluctuations



Problem: there are natural fluctuations in occurrence rates both temporally and spatially.

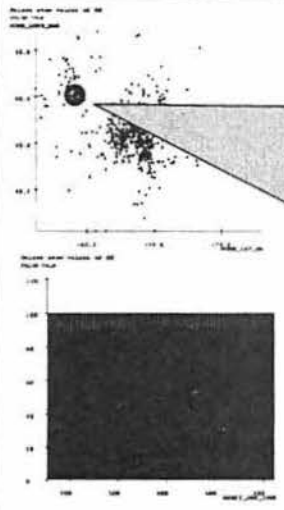
Solution: assume a linear contribution of fluctuations:
expected rate = global_average_rate +
$\text{deviation}_{\text{region}}$ + $\text{deviation}_{\text{time}}$

Problem: the usual method of computing p values with randomization no longer applies directly

Solution: assume a prior history of data collected during "normal" situations. Compute a baseline from that.

21

QSCAN: compensating for inherent spatio-temporal fluctuations



The global average rate is 9.16%. During the recent 10 days that average is 0.25% higher.

This region, however, is normally 0.35% lower than the global average.

Its 17.24% occurrence rate and corresponding 12% chance that it could score this strangely by chance makes it the most worrisome region today.

Checking against historical data, 51% of previous time periods had a region this strange.

No Need to Sound an ALARM

currence

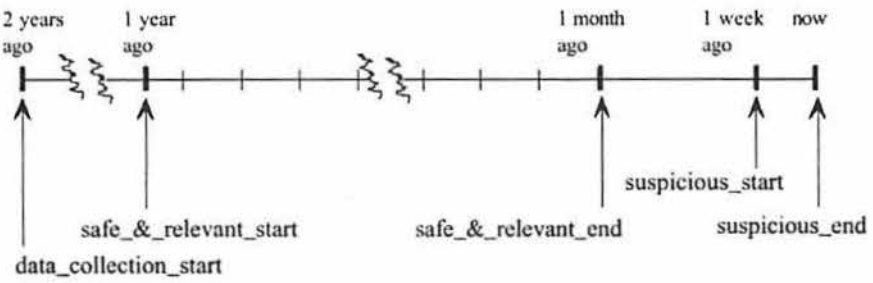
ctuations:
ıte +
me

values
es directly

ected
te a

22

PANDA (Patient-based ANomaly Detection Algorithm)

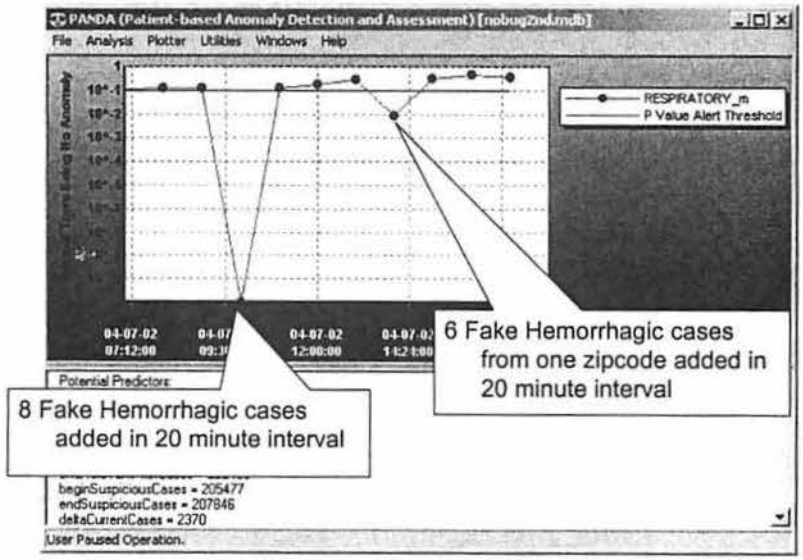


Puts a prior distribution on Respiratory problems using data from a known safe period

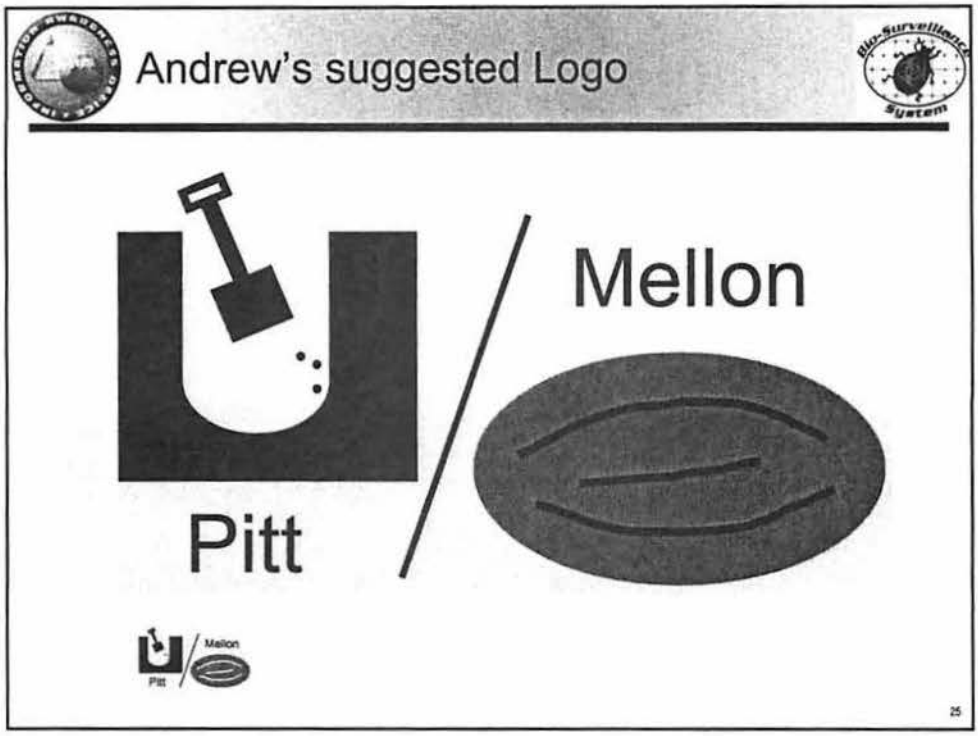
Estimates same parameters over varying recent time window and does a Bayesian test of whether a different process is in operation

23

PANDA Example



24


Andrew's suggested Logo
Mellon
Pitt
25

DETECTION APPROACHES AND ALGORITHMS

Veridian-Stanford Team

2nd PI Meeting

4 - 6 June 2002

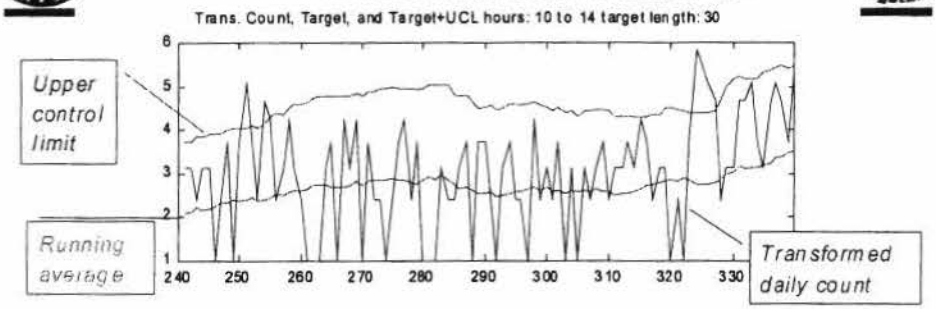
Technical Approach

- What Advanced Analytics are being Used/Developed for Detection/Characterization?
  - BASIS automated time series analysis
  - Automated spatial change detection
  - Scan Statistic
  - Spatio-temporal abstraction – Stanford presentation later today

Technical Approach

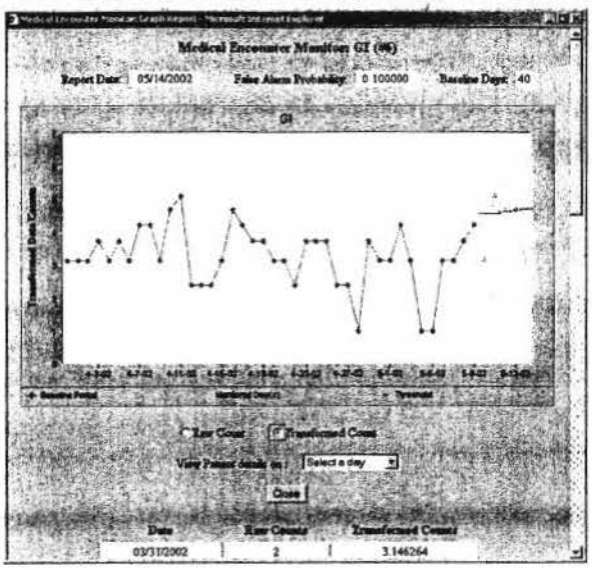
- First utilize methods developed during DARPA support for CDC's Enhanced Surveillance Project
  - Fully automated alerting on time series data
  - Generalize to spatial distribution
- Then develop/implement new methods

Control Chart Method



- Process Control Chart to detect clustering at the most recent end of a time series
- Simple, well-known statistical method
- Diurnal, seasonal trends, and other localized, dynamic variability are accounted for
- Abrupt, abnormal deviations detected against baseline
- Most useful for relatively large, stable, population samples

Set Upper Control Limit to Tolerable False Alarm Rate - - May Be Adaptive

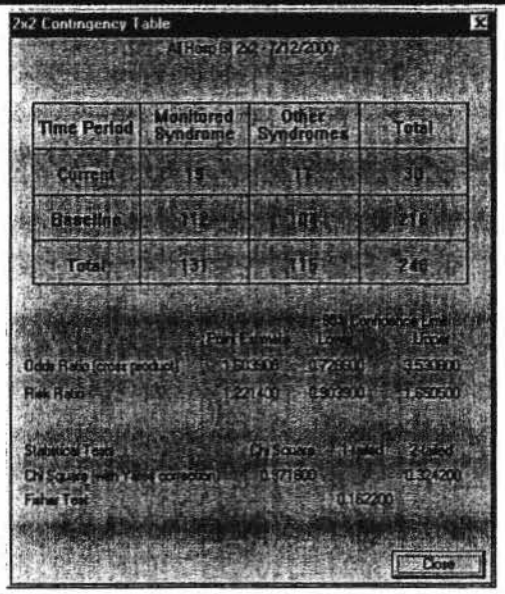


2 X 2 Contingency Table Method for Time Series Data

	Sentinel Filter		Reference Filter
	Sentinel Case count	Non-sentinel Count	**Total Case Count**
Present	*a*	*b*	*R1=a+b*
Past ("baseline")	*c*	*d*	*R2=c+d*
Totals	*C1=a+c*	*C2=b+d*	*N=a+b+c+d*

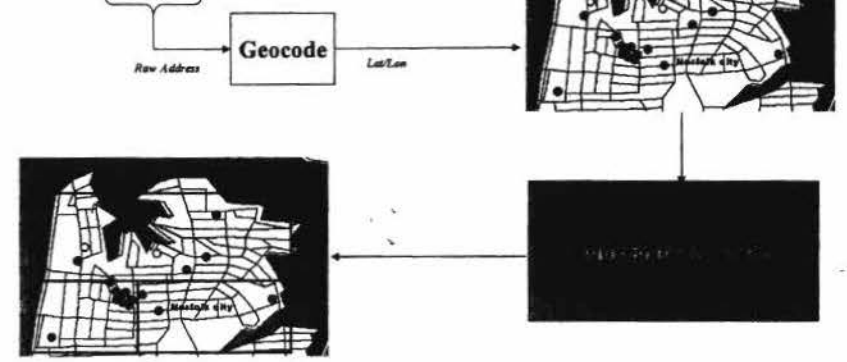
- Simple, well-known statistical method
  - Routinely used in epidemiology
  - Determines whether there is correlation between two ways of categorizing data
  - Useful for small samples (with Fischer exact test) as well as large
- Monitors relative disease rates in our application
- Tolerates imperfect baseline data
- Provides useful testing of accumulating data prior to end of observational period

As Implemented in BASIS Automated Surveillance Software



Developing Similar Test for Spatial Distribution of Reported Events

Name	Address	Syndromes	Date
John Doe	1065 Main St.	Upper Respiratory,Fever	06/04/02
Jane Smith	1253 East Elm St.	Gastro-Intestinal	06/04/02
Jenifer Jones	2198 Green St.	Fever	06/04/02



Test for Change in Spatial Distribution

- Comparison of two epochs of proportional allotment of event counts among several (N) locations
  - "Baseline period" (distribution from past)
  - "Current period" (present distribution)
- Use 2 x N contingency table with chi-squared test to look for deviation from "null" hypothesis" of no change in spatial distributional

2xN Contingency Table

- Divide surveillance area into *N* tracts with expected number of cases per epoch not too small, e.g. at least ~5-10

	Tract 1	*Tract 2*	...	*Tract N*
Current	n_{11}	n_{12}		n_{1N}
Baseline	n_{21}	n_{22}	...	n_{2N}

- Does not require explicit knowledge of spatial distribution of subject population
- Tolerates imperfect baseline and accumulating data
- Test does not localize the deviation