Description of document: Department of Labor (DOL) Inspector General (OIG) statement of work, reports and presentations produced for the DOL OIG by Elder Research Inc., under contracts DOLOIG14AI0006/DOLOIG14U00012, GS35F032OT/DOLF12XG21355, and GS35F0320T/DOLOIG14A0006, 2012-2017

Requested date: 30-November-2014

Released date: 10-April-2017

Posted date: 07-May-2018

Source of document:

Disclosure Officer
Office of Inspector General
U.S. Department of Labor
200 Constitution Ave., N.W., Room S-5506
Washington, DC 20210
Fax: (202) 693-7020
Email: foiarequest@dol.gov

U.S. Department of Labor

Office of Inspector General
Washington, D.C. 20210

April 10, 2017

This is in final response to your November 30, 2014 Freedom of Information Act request (215010) addressed to this office for a copy of the statement of work, reports or presentations produced for the DOL OIG by Elder Research Inc., under contracts DOLOIG14AI0006/DOLOIG14U00012, GS35F032OT/DOLF12XG21355, and GS35F0320T/DOLOIG14A0006.

The policy of the Inspector General is to make, to the extent possible, full disclosure of our identifiable records in accordance with the provisions of the Freedom of Information Act. Accordingly, I am enclosing a copy of all materials responsive to your request. However, certain information, which includes fields of information used for audit/investigative techniques, and individual's names and personal information have been redacted from the enclosed documents. The withheld information is subject to various FOIA exemptions, as discussed below.

Exemption (b)(6) authorizes the withholding of names and details of personal information in personnel, medical and similar case files, which, if disclosed to the public, would constitute an unwarranted invasion of personal privacy.

Exemption (b) (7)(e) protects law enforcement information that would disclose techniques or procedures for audits and law enforcement investigations. In this case, specific details regarding data mining techniques the OIG uses for audit/investigative purposes has been redacted on portions of several pages.

You have the right to appeal this response within 90 days from the date of this letter. Should you decide to do this, your appeal must state, in writing, the grounds for appeal, together with any statement or arguments. Such an appeal should be addressed and directed to the Solicitor of Labor, citing OIG/FOIA No.215010, Room N-2428, 200 Constitution Avenue, N.W., Washington, D.C. 20210. Please refer to the Department of Labor regulations at 29 CFR 70.22 for further details on your appeal rights.

Should you need to discuss your request, feel free to contact this office at 202-693-5116 and select Disclosure Officer, or the DOL FOIA Public Liaison, Thomas Hicks at 202-693-5427. Additionally, you may contact the Office of Government Information Services (OGIS) at the National Archives and Records Administration to inquire about the FOIA mediation services they offer. The contact information for OGIS is as follows: Office of Government Information Services, National Archives and Records Administration, 8601 Adelphi Road-OGIS, College Park, Maryland 20740-6001; e-mail at ogis@nara.gov; telephone 202-741-5770; toll free at 1-877-684-6448; or facsimile at 202-741-5769.

Congress excluded three discrete categories of law enforcement and national security records from the requirements of the FOIA. *See* 5 U.S.C. 552(c). This response is limited to those records that are subject to the requirements of the FOIA. This is a standard notification that is given to all our requesters and should not be taken as an indication that excluded records do, or do not, exist.

Finally, fees were not charged for this request. If you have any concerns regarding this letter, feel free to contact me at this office at 202-693-5116 and refer to FOIA case number 215010 on future inquiries.

Sincerely,

Kim Pacheco

Kim Pacheco
Disclosure Officer

Enclosures

SOLICITATION/CONTRACT/ORDER FOR COMMERCIAL ITEMS

OFFEROR TO COMPLETE BLOCKS 12, 17, 23, 24, & 30

1. REQUISITION NUMBER

PAGE 1 OF 4

2. CONTRACT NO. DOL-OIG-14-A-0006

3. AWARD/ EFFECTIVE DATE

4. ORDER NUMBER DOL-OIG-14-U-00012

5. SOLICITATION NUMBER

6. SOLICITATION ISSUE DATE

7. FOR SOLICITATION INFORMATION CALL: ▶ a. NAME Paula Miller-Sheelor

b. TELEPHONE NUMBER *(No collect calls)*

8. OFFER DUE DATE/LOCAL TIME

9. ISSUED BY CODE OIG

US DEPARTMENT OF LABOR
200 CONSTITUTION AVE NW S-
5506
WASHINGTON DC 20210

10. THIS ACQUISITION IS ☒ UNRESTRICTED OR □ SET ASIDE: % FOR:

□ SMALL BUSINESS
□ HUBZONE SMALL BUSINESS
□ SERVICE-DISABLED VETERAN-OWNED SMALL BUSINESS
□ WOMEN-OWNED SMALL BUSINESS (WOSB) ELIGIBLE UNDER THE WOMEN-OWNED SMALL BUSINESS PROGRAM
□ EDWOSB
□ 8(A)

NAICS: 541519

SIZE STANDARD: $25.0

11. DELIVERY FOR FOB DESTINATION UNLESS BLOCK IS MARKED □ SEE SCHEDULE

12. DISCOUNT TERMS

□ 13a. THIS CONTRACT IS A RATED ORDER UNDER DPAS (15 CFR 700)

13b. RATING

14. METHOD OF SOLICITATION □ RFQ □ IFB □ RFP

15. DELIVER TO CODE

16. ADMINISTERED BY CODE OIG

US DEPARTMENT OF LABOR
200 CONSTITUTION AVE NW S-
5506
WASHINGTON DC 20210

17a. CONTRACTOR/ OFFEROR CODE 028211527 FACILITY CODE

ELDER RESEARCH INCORPORATED
300 W. MAIN STREET, SUITE 301
CHARLOTTESVILLE VIRGINIA 22903
ATTN: GERHARD PILCHER

TELEPHONE NO. 434-973-7673

□ 17b. CHECK IF REMITTANCE IS DIFFERENT AND PUT SUCH ADDRESS IN OFFER

18a. PAYMENT WILL BE MADE BY CODE DOL

US DEPARTMENT OF LABOR
OFFICE OF THE CHIEF
FINANCIAL OFFICER
200 CONSTITUTION AVE NW
WASHINGTON DC 20210

18b. SUBMIT INVOICES TO ADDRESS SHOWN IN BLOCK 18a UNLESS BLOCK BELOW IS CHECKED □ SEE ADDENDUM

19. ITEM NO.	20. SCHEDULE OF SUPPLIES/SERVICES	21. QUANTITY	22. UNIT	23. UNIT PRICE	24. AMOUNT
	The contractor shall provide a Business Intelligence Solution (BI-SOLUTION) Covering Vast and Complex All services shall be provided in accordance with the attached Statement of Work (SOW). The Period of Performance is July 23, 2014 through July 22, 2015 *(Use Reverse and/or Attach Additional Sheets as Necessary)*				

25. ACCOUNTING AND APPROPRIATION DATA

26. TOTAL AWARD AMOUNT *(For Govt. Use Only)* $290,250.00

□ 27a. SOLICITATION INCORPORATES BY REFERENCE FAR 52.212-1, 52.212-4. FAR 52.212-3 AND 52.212-5 ARE ATTACHED. ADDENDA □ ARE □ ARE NOT ATTACHED.

□ 27b. CONTRACT/PURCHASE ORDER INCORPORATES BY REFERENCE FAR 52.212-4. FAR 52.212-5 IS ATTACHED. ADDENDA □ ARE □ ARE NOT ATTACHED.

□ 28. CONTRACTOR IS REQUIRED TO SIGN THIS DOCUMENT AND RETURN ______ COPIES TO ISSUING OFFICE. CONTRACTOR AGREES TO FURNISH AND DELIVER ALL ITEMS SET FORTH OR OTHERWISE IDENTIFIED ABOVE AND ON ANY ADDITIONAL SHEETS SUBJECT TO THE TERMS AND CONDITIONS SPECIFIED.

□ 29. AWARD OF CONTRACT: ______ OFFER DATED ______. YOUR OFFER ON SOLICITATION (BLOCK 5), INCLUDING ANY ADDITIONS OR CHANGES WHICH ARE SET FORTH HEREIN, IS ACCEPTED AS TO ITEMS:

30a. SIGNATURE OF OFFEROR/CONTRACTOR

30b. NAME AND TITLE OF SIGNER *(Type or print)* W. Gerhard Pilcher, Vice President

30c. DATE SIGNED 7/22/2014

31a. UNITED STATES OF AMERICA *(SIGNATURE OF CONTRACTING OFFICER)* Paula Miller-Sheelor

31b. NAME OF CONTRACTING OFFICER *(Type or print)* Paula Miller-Sheelor

31c. DATE SIGNED 07/23/2014

19. ITEM NO.	20. SCHEDULE OF SUPPLIES/SERVICES	21. QUANTITY	22. UNIT	23. UNIT PRICE	24. AMOUNT
1.	Provide a Business Intelligence Solution (BI-SOLUTION) Covering Vast and Complex, for the amount of $290,250.000 to be divided between OWCP and OIG. **Accounting Info:** 151521830XBR2014FSAD090414PFSADMP0000PWCP00PFSANO P90018-251037 Funding Stream: 151521830XBR2014FSAD090414PFSADMP0000PWCP00PFSANO Cost Center: P90018 Object Class: 251037 Funded: $145,125.00	1	Job	$145,125.00	$145,125.00
2.	Provide a Business Intelligence Solution (BI-SOLUTION) Covering Vast and Complex, for OWCP and OIG, for the amount of $290,250.00 to be split in half between the two agencies. **Accounting Info:** Accounting Info: 2801061414AD20140106000114G00000G0000GOIG00GAUDIT G90201-251037 Funding Stream: 2801061414AD20140106000114G00000G0000GOIG00GAUDIT Cost Center: G90201 Object Class: 251037 Funded: $145,125.00	1	Job		$145,125.00

32a. QUANTITY IN COLUMN 21 HAS BEEN

□ RECEIVED □ INSPECTED □ ACCEPTED, AND CONFORMS TO THE CONTRACT, EXCEPT AS NOTED:

32b. SIGNATURE OF AUTHORIZED GOVERNMENT REPRESENTATIVE

32c. DATE

32d. PRINTED NAME AND TITLE OF AUTHORIZED GOVERNMENT REPRESENTATIVE

32e. MAILING ADDRESS OF AUTHORIZED GOVERNMENT REPRESENTATIVE

32f. TELEPHONE NUMBER OF AUTHORIZED GOVERNMENT REPRESENTATIVE

32g. E-MAIL OF AUTHORIZED GOVERNMENT REPRESENTATIVE

33. SHIP NUMBER

□ PARTIAL □ FINAL

34. VOUCHER NUMBER

35. AMOUNT VERIFIED CORRECT FOR

36. PAYMENT

□ COMPLETE □ PARTIAL □ FINAL

37. CHECK NUMBER

38. S/R ACCOUNT NUMBER

39. S/R VOUCHER NUMBER

40. PAID BY

41a. I CERTIFY THIS ACCOUNT IS CORRECT AND PROPER FOR PAYMENT

41b. SIGNATURE AND TITLE OF CERTIFYING OFFICER

41c. DATE

42a. RECEIVED BY *(Print)*

42b. RECEIVED AT *(Location)*

42c. DATE REC'D *(YY/MM/DD)*

42d. TOTAL CONTAINERS

BASE					
Type of License	**Period (months/years)**	**Cost**		OIG	**OWCP**
RADR Perpetual License	**Perpetual**	$112,500.00		$56,250.00	**$56,250.00**
Insight Perpetual License	**Perpetual**	$125,000.00		$62,500.00	**$62,500.00**
RADR Annual Maintenance	12 Months	$21,500.00		$10,750.00	**$10,750.00**
Insight Annual Maintenance	12 Months	$31,250.00		$15,625.00	**$15,625.00**
		Total		$145,125.00	**$145,125.00**

BUSINESS INTELLIGENCE SOLUTION (BI-SOLUTION)
STATEMENT OF WORK

Background

The Office of Inspector General – Office of Audit (OIG/OA) performs information and related information technology audits of the U.S. Department of Labor (DOL) cyber networks, systems, and related database information.

The OIG has an IT audit office located in the Frances Perkins Building, Washington, DC. This office is responsible for ensuring the Inspector General meets mandatory obligations to perform annual evaluations of more than 60 DOL major information systems' security controls, and provide targeted extracts of database information, including at times establishing the reliability of the information at a summary or transaction level. The latter'activity is generally referred to as data mining or also called knowledge discovery. OIG has been using Elder Research Inc.'s RADR and Insight proprietary software for performing data mining and analytics of DOL's Office of Workers Compensation Federal Employees' Compensation Act chargeback, compensation, bill pay, case management and DOL OIG's investigative case file data.

Objective and Scope

The objective is for OIG to contract with Elder Research Inc. to provide an automated BI-Solution for performing knowledge discovery across heterogeneous DOL computing environments and related external systems that are comprised of large databases, including many within high Terabyte ranges.

Requirements

The contractor shall establish an automated BI-Solution for OIG and additional DOL program agencies using the proprietary RADR and Insight software for knowledge discovery activities that can span multiple program agencies' disparate systems, networks, and databases in order to assess and analyze the data. Information generated from these discovery activities is to further OIG audit targeting involving issues such as potential waste, fraud, and abuse and covering multiple topics such as revenue, expenses, payments, performance measurement, enforcement actions, employee safety and retirement, cost savings, entity and metadata identification, and/or combinations of any or all these elements. The use of the BI-Solution using RADR and Insight will be seen as an important tool in OIG's analysis of information from selected databases. It will allow users to analyze information from unforeseen data relationships, find correlations or patterns among dozens of data fields residing in a variety of types of databases such as flat-file, relational, and/or hierarchical databases. Databases may also be ranged across multiple operating systems platforms such as Windows, Unix, Oracle, and Open Source.

Deliverables

- RADR
  - Perpetual Multiple DOL Progam Agency License
  - Perpetual Multiple DOL Progam Agency Maintenance License

- Insight
    - Perpetual Multiple DOL Progam Agency License
    - Perpetual Multiple DOL Progam Agency Maintenance License

- The mathematical models for current and future projects

- The Insight cubes (output via OLAP) and the underlying data for current and future projects.

- IT services, including training, that is in support of the use of RADR and Insight for current and future projects.

- Specific documentation, such as process and procedures related to the access, use and reporting of RADR and Insight and the Insight cubes.

SOLICITATION/CONTRACT/ORDER FOR COMMERCIAL ITEMS

OFFEROR TO COMPLETE BLOCKS 12, 17, 23, 24, & 30

1. REQUISITION NUMBER: 14-OOIG-OIG-NAT-0024

PAGE 1 OF 19

2. CONTRACT NO. DOL-OIG-14-A-0006

3. AWARD/ EFFECTIVE DATE

4. ORDER NUMBER

5. SOLICITATION NUMBER

6. SOLICITATION ISSUE DATE

7. FOR SOLICITATION INFORMATION CALL: ▶ a. NAME Paula Miller-Sheelor

b. TELEPHONE NUMBER *(No collect calls)*

8. OFFER DUE DATE/LOCAL TIME

9. ISSUED BY CODE OIG

US Department of Labor
200 Constitution Ave, NW S-5506
Washington DC 20210

10 THIS ACQUISITION IS ☒ UNRESTRICTED OR □ SET ASIDE: % FOR:

□ SMALL BUSINESS
□ HUBZONE SMALL BUSINESS
□ SERVICE-DISABLED VETERAN-OWNED SMALL BUSINESS
□ WOMEN-OWNED SMALL BUSINESS (WOSB) ELIGIBLE UNDER THE WOMEN-OWNED SMALL BUSINESS PROGRAM
□ EDWOSB
□ 8(A)

NAICS: 541519

SIZE STANDARD: $25.0

11. DELIVERY FOR FOB DESTINATION UNLESS BLOCK IS MARKED □ SEE SCHEDULE

12. DISCOUNT TERMS As Indicated On Each Call

□ 13a. THIS CONTRACT IS A RATED ORDER UNDER DPAS (15 CFR 700)

13b. RATING

14. METHOD OF SOLICITATION □ RFQ □ IFB □ RFP

15. DELIVER TO CODE

As Indicated On Each Call

16. ADMINISTERED BY CODE OIG

US Department of Labor
200 Constitution Ave, NW S-5506
Washington DC 20210

17a. CONTRACTOR/ OFFEROR CODE 028211527 FACILITY CODE

ELDER RESEARCH INCORPORATED
635 BERKMAR CIR
CHARLOTTESVILLE VIRGINIA 229011464

TELEPHONE NO.

□ 17b. CHECK IF REMITTANCE IS DIFFERENT AND PUT SUCH ADDRESS IN OFFER

18a. PAYMENT WILL BE MADE BY CODE

As Indicated On Each Call

18b. SUBMIT INVOICES TO ADDRESS SHOWN IN BLOCK 18a UNLESS BLOCK BELOW IS CHECKED □ SEE ADDENDUM

19. ITEM NO.	20. SCHEDULE OF SUPPLIES/SERVICES	21. QUANTITY	22. UNIT	23. UNIT PRICE	24. AMOUNT
	The contractor shall provide a Business Intelligence Solution (BI-SOLUTION) Covering Vast and Complex. All services shall be provided in accordance with the attached Statement of Work (SOW) and all terms and conditions included herein and the attached items in the proposal dated May 20, 2014. The Period of Performance is Date of Award through June 30, 2019.				

25. ACCOUNTING AND APPROPRIATION DATA
As Indicated On Each Call

26. TOTAL AWARD AMOUNT *(For Govt. Use Only)* $0.00

□ 27a. SOLICITATION INCORPORATES BY REFERENCE FAR 52.212-1, 52.212-4. FAR 52.212-3 AND 52.212-5 ARE ATTACHED. ADDENDA □ ARE □ ARE NOT ATTACHED.

□ 27b. CONTRACT/PURCHASE ORDER INCORPORATES BY REFERENCE FAR 52.212-4. FAR 52.212-5 IS ATTACHED. ADDENDA □ ARE □ ARE NOT ATTACHED.

☒ 28. CONTRACTOR IS REQUIRED TO SIGN THIS DOCUMENT AND RETURN 1 COPIES TO ISSUING OFFICE. CONTRACTOR AGREES TO FURNISH AND DELIVER ALL ITEMS SET FORTH OR OTHERWISE IDENTIFIED ABOVE AND ON ANY ADDITIONAL SHEETS SUBJECT TO THE TERMS AND CONDITIONS SPECIFIED.

□ 29. AWARD OF CONTRACT: DOL-14-Q-00004 OFFER DATED May 20, 2014. YOUR OFFER ON SOLICITATION (BLOCK 5), INCLUDING ANY ADDITIONS OR CHANGES WHICH ARE SET FORTH HEREIN IS ACCEPTED AS TO ITEMS:

30a. SIGNATURE OF OFFEROR/CONTRACTOR

WGPilch

30b. NAME AND TITLE OF SIGNER *(Type or print)* W. Gerhard Pilcher Vice President

30c. DATE SIGNED 7/17/2014

31a. UNITED STATES OF AMERICA *(SIGNATURE OF CONTRACTING OFFICER)*

Paula Miller-Sheelor

31b. NAME OF CONTRACTING OFFICER *(Type or print)* William Aumand

31c. DATE SIGNED 07/21/2014

AUTHORIZED FOR LOCAL REPRODUCTION
PREVIOUS EDITION IS NOT USABLE

SOLICITATION/CONTRACT/ORDER FOR COMMERCIAL ITEMS

OFFEROR TO COMPLETE BLOCKS 12, 17, 23, 24, & 30

1. REQUISITION NUMBER 14-OOIG-OIG-NAT-0024	PAGE 1 OF 19		
2. CONTRACT NO. DOL-OIG-14-A-0006	3. AWARD/ EFFECTIVE DATE	4. ORDER NUMBER	5. SOLICITATION NUMBER
6. SOLICITATION ISSUE DATE			
7. FOR SOLICITATION INFORMATION CALL:	a. NAME Paula Miller-Sheelor	b. TELEPHONE NUMBER (No collect calls)	8. OFFER DUE DATE/LOCAL TIME

9. ISSUED BY CODE OIG

US Department of Labor
200 Constitution Ave, NW S-5506
Washington DC 20210

10. THIS ACQUISITION IS ☒ UNRESTRICTED OR □ SET ASIDE: % FOR:

□ SMALL BUSINESS
□ HUBZONE SMALL BUSINESS
□ SERVICE-DISABLED VETERAN-OWNED SMALL BUSINESS
□ WOMEN-OWNED SMALL BUSINESS (WOSB) ELIGIBLE UNDER THE WOMEN-OWNED SMALL BUSINESS PROGRAM
□ EDWOSB
□ 8(A)

NAICS: 541519

SIZE STANDARD: $25.0

11. DELIVERY FOR FOB DESTINATION UNLESS BLOCK IS MARKED □ SEE SCHEDULE

12. DISCOUNT TERMS
As Indicated On Each Call

□ 13a. THIS CONTRACT IS A RATED ORDER UNDER DPAS (15 CFR 700)

13b. RATING

14. METHOD OF SOLICITATION □ RFQ □ IFB □ RFP

15. DELIVER TO CODE

As Indicated On Each Call

16. ADMINISTERED BY CODE OIG

US Department of Labor
200 Constitution Ave, NW S-5506
Washington DC 20210

17a. CONTRACTOR/ OFFEROR CODE 028211527 FACILITY CODE

ELDER RESEARCH INCORPORATED
635 BERKMAR CIR
CHARLOTTESVILLE VIRGINIA 229011464

TELEPHONE NO.

□ 17b. CHECK IF REMITTANCE IS DIFFERENT AND PUT SUCH ADDRESS IN OFFER

18a. PAYMENT WILL BE MADE BY CODE

As Indicated On Each Call

18b. SUBMIT INVOICES TO ADDRESS SHOWN IN BLOCK 18a UNLESS BLOCK BELOW IS CHECKED □ SEE ADDENDUM

19. ITEM NO.	20. SCHEDULE OF SUPPLIES/SERVICES	21. QUANTITY	22. UNIT	23. UNIT PRICE	24. AMOUNT
	The contractor shall provide a Business Intelligence Solution (BI-SOLUTION) Covering Vast and Complex. All services shall be provided in accordance with the attached Statement of Work (SOW) and all terms and conditions included herein and the attached items in the proposal dated May 20, 2014. The Period of Performance is Date of Award through June 30, 2019.				

25. ACCOUNTING AND APPROPRIATION DATA
As Indicated On Each Call

26. TOTAL AWARD AMOUNT *(For Govt. Use Only)*
$0.00

□ 27a. SOLICITATION INCORPORATES BY REFERENCE FAR 52.212-1, 52.212-4. FAR 52.212-3 AND 52.212-5 ARE ATTACHED. ADDENDA □ ARE □ ARE NOT ATTACHED.

□ 27b. CONTRACT/PURCHASE ORDER INCORPORATES BY REFERENCE FAR 52.212-4. FAR 52.212-5 IS ATTACHED. ADDENDA □ ARE □ ARE NOT ATTACHED.

☒ 28. CONTRACTOR IS REQUIRED TO SIGN THIS DOCUMENT AND RETURN 1 COPIES TO ISSUING OFFICE. CONTRACTOR AGREES TO FURNISH AND DELIVER ALL ITEMS SET FORTH OR OTHERWISE IDENTIFIED ABOVE AND ON ANY ADDITIONAL SHEETS SUBJECT TO THE TERMS AND CONDITIONS SPECIFIED.

□ 29. AWARD OF CONTRACT: DOL-14-O-00004 OFFER DATED May 20, 2014. YOUR OFFER ON SOLICITATION (BLOCK 5), INCLUDING ANY ADDITIONS OR CHANGES WHICH ARE SET FORTH HEREIN, IS ACCEPTED AS TO ITEMS:

30a. SIGNATURE OF OFFEROR/CONTRACTOR

31a. UNITED STATES OF AMERICA *(SIGNATURE OF CONTRACTING OFFICER)*

30b. NAME AND TITLE OF SIGNER *(Type or print)*	30c. DATE SIGNED	31b. NAME OF CONTRACTING OFFICER *(Type or print)* William Aumand	31c. DATE SIGNED

AUTHORIZED FOR LOCAL REPRODUCTION
PREVIOUS EDITION IS NOT USABLE

Pursuant to BPA contract number(s) **DOL-OIG-14-A-0006**, a Blanket Purchase Agreement (BPA) is hereby established between **Elder Research, Inc.** and the U.S.DEPARTMENT OF LABOR (DOL) under the terms and conditions of the above stated contract(s) and the following terms and conditions are incorporated in this BPA:

I. AUTHORITY

This BPA is entered into pursuant to the Federal Acquisition Regulation Part 13.303-2.

II. DESCRIPTION OF AGREEMENT

This Blanket Purchase Agreement allows for ordering of services for **Business Intelligence Solution (BI-SOLUTION) Covering Vast and Complex**, by the U.S. Department of Labor, Office of Inspector General. Please be advised that the services to be provided for BPA **DOL-OIG-14-A-0006** is described below in Section III entitled "Scope of Work" The period of performance for this BPA encompasses five years.

Statement of Work

Business Intelligence solution (BI-SOLUTION)

Background

The Office of Inspector General - Office of Audit (OIG/OA) performs information and related information technology audits of the U.S. Department of Labor (DOL) cyber networks, systems, and related database information.

The OIG has an IT audit office located in the Frances Perkins Building, Washington, DC. This office is responsible for ensuring the Inspector General meets mandatory obligations to perform annual evaluations of more than 60 DOL major information systems' security controls, and provide targeted extracts of database information, including at times establishing the reliability of the information at a summary or transaction level. The latter activity is generally referred to as data mining or also called knowledge discovery. OIG has been using Elder Research Inc.'s RADR and Insight proprietary software for performing data mining and analytics of DOL's Office of Workers Compensation Federal Employees' Compensation Act chargeback, compensation, bill pay, case management and DOL OIG's investigative case file data.

Objective and Scope

The objective is for OIG to contract with Elder Research Inc. to provide an automated BI-Solution for performing knowledge discovery across heterogeneous DOL computing environments and related external systems that are comprised of large databases, including many within high Terabyte ranges.

Requirements

The contractor shall establish an automated BI-Solution for OIG and additional DOL program agencies using the proprietary RADR and Insight software for knowledge discovery activities that can span multiple program agencies' disparate systems, networks, and databases in order to assess and analyze the data. Information generated from these discovery activities is to further OIG audit targeting involving issues such as potential waste, fraud, and abuse and covering multiple topics such as revenue, expenses, payments, performance measurement, enforcement actions, employee safety and retirement, cost savings, entity and metadata identification, and/or combinations of any or all these elements. The use of the BI-Solution using RADR and Insight will be seen as an important tool in OIG's analysis of information from selected databases. It will allow users to analyze information from unforeseen data relationships, find correlations or patterns among dozens of data fields residing in a variety of types of databases such as flat-file, relational, and/or hierarchical databases. Databases may also be ranged across multiple operating systems platforms such as Windows, Unix, Oracle, and Open Source.

Deliverables

- RADR
  - o Perpetual Multiple DOL Progam Agency License
  - o Perpetual Multiple DOL Progam Agency Maintenance License

- Insight
  - o Perpetual Multiple DOL Progam Agency License
  - o Perpetual Multiple DOL Progam Agency Maintenance License

- The mathematical models for current and future projects

- The Insight cubes (output via OLAP) and the underlying data for current and future projects.

- IT services, including training, that is in support of the use of RADR and Insight for current and future projects.

- Specific documentation, such as process and procedures related to the access, use and reporting of RADR and Insight and the Insight cubes.

The Government anticipates awarding various fixed price Time & Material (T&M labor hour) Task Orders from this BPA. Please provide a list of potential labor categories and associated prices.

ADMINISTRATIVE DATA

Primary Point of Contact: **W. Gerhard Pilcher**

Vice President & Senior Scientist

855-973-7673 ext. 707 (phone)

434-973-7875 (fax)

Gerhard@dataminingl ab.com

Alternate Point of Contact: **Jeff Deal**

Vice President, Operations

434-227-5851 ext. 851 (phone)

434-973-7875 (fax)

deal@dataminingl ab.com

Are you a Small Business under NAIC Code 334118 (FAR PART 19.102)?

YES X NO _______

Are you a Small Business Administration (SBA) certified Small Disadvantaged Business (SDB)? YES _________ NO X

Are you a Woman-Owned Business? YES _______ NO X

CAGE CODE: 1GMY7

DUNS NUMBER: 028211527

TIN: 30-0000656

GSA SCHEDULE #: GS-35F0320T

SIN Number: **132-51**

Contract Expiration Date: **March 31, 2017**

Cognizant DCAA Office (Include complete address):
(Other auditing activity may be listed)

DCAA
Branch Office
514 Butler Farm Road
Suite 290
Hampton, VA 23666
Telephone: (757) 865-5520

V. PERIOD OF PERFORMANCE

This BPA shall be in effect from the date of award through five years.

VI. DESCRIPTION OF AGREEMENT

Under this agreement, the BPA holder shall provide lifecycle replacement and enhancement of the existing **Business Intelligence Solution (BI-SOLUTION) Covering Vast and Complex** that support the users across the country for the Department of Labor's OIG. The above described supplies and/or services shall be provided when ordered by an authorized Contracting Officer during the specified period stated in Item V, entitled "PERIOD OF PERFORMANCE".

VII. REPRESENTATIVE OF THE CONTRACTING OFFICER

The Department of Labor's Office of Inspector General Contracting Officer's Technical Representative (COTR) is **Keith Galayda.**

The COTR has the following authority:

To direct work but does not have authority to direct the contractor to perform work outside the scope, price, terms and conditions of the BPA's performance work statement and issued task orders, the GSA Federal Supply Schedule Pricing, terms and conditions or in excess of funding which has been obligated by the Contracting Officer for performance of work; inspection and acceptance of supplies/services; and Invoices.

SOLICITATION/CONTRACT/ORDER FOR COMMERCIAL ITEMS
OFFEROR TO COMPLETE BLOCKS 12, 17, 23, 24, & 30

1. REQUISITION NO. 4-129G- 3869

PAGE 1 OF 13

2. CONTRACT NO. GS-35F-0320T

3. AWARD/EFFECTIVE DATE SEE BLOCK 31C

4. ORDER NO. DOLF12XG21355

5. SOLICITATION NUMBER DOL12GRQ20069

6. SOLICITATION ISSUE DATE 09/10/2012

7. FOR SOLICITATION INFORMATION CALL:

a. NAME Paula Miller-Sheelor

b. TELEPHONE NO. (No Collect Calls) (202) 693-7050

8. OFFER DUE DATE/LOCAL TIME 09/04/2012 2:00 pm ES

9. ISSUED BY CODE 1604

OIG Office of Procurement Services
U. S. Department of Labor
RM S5506
200 Constitution Ave, NW
Washington DC 20210

10. THIS ACQUISITION IS ☐ UNRESTRICTED OR ☐ SET ASIDE: % FOR:
☐ SMALL BUSINESS ☐ EMERGING SMALL BUSINESS
☐ HUBZONE SMALL BUSINESS
☐ SERVICE-DISABLED VETERAN-OWNED SMALL BUSINESS ☐ 8(A)

NAICS: 541519
SIZE STANDARD:

11. DELIVERY FOR FOB DESTINATION UNLESS BLOCK IS MARKED
☒ SEE SCHEDULE

12. DISCOUNT TERMS NET 30

☐ 13a. THIS CONTRACT IS A RATED ORDER UNDER DPAS (15 CFR 700)

13b. RATING N/A

14. METHOD OF SOLICITATION ☒ RFQ ☐ IFB ☐ RFP

15. DELIVER TO CODE 1604

U. S. Department of Labor
Office of Inspector General

200 Constitution Avenue, NW
Washington DC 20210

16. ADMINISTERED BY CODE 1604

OIG - Office of Procurement Services
U. S. Department of Labor
RM S5506
200 Constitution Ave, NW
Washington DC 20210

17a. CONTRACTOR/OFFEROR CODE FACILITY CODE 028211527

Elder Research Inc
300 W Main Ste 301
Charlottesville, VA 22903

TELEPHONE NO. 540-560-3183

☐ 17b. CHECK IF REMITTANCE IS DIFFERENT AND PUT SUCH ADDRESS IN OFFER

18a. PAYMENT WILL BE MADE BY CODE

Department of labor
OASAM BRANCH OF INVOICE PAYMENT RM: S5526
US DEPARTMENT OF LABOR
200 CONSTITUTION AVENUE, NW
WASHINGTON DC 20210

18b. SUBMIT INVOICES TO ADDRESS SHOWN IN BLOCK 18a UNLESS BLOCK BELOW IS CHECKED
☒ SEE ADDENDUM

19. ITEM NO.	20. SCHEDULE OF SUPPLIES/SERVICES	21. QUANTITY	22. UNIT	23. UNIT PRICE	24. AMOUNT
	The contractor shall provide a Business Intelligence Solution (BI-SOLUTION) Covering Vast and Complex. All services shall be provided in accordance with the Attached Statement of Work (SOW) and all terms and Conditions included herein and the attached items Highlighted in the proposal. The Period of Performance is September 28, 2012 through September 27, 2013.				

(Use Reverse and/or Attach Additional Sheets as Necessary)

25. ACCOUNTING AND APPROPRIATION DATA 28-01061212AD-2012-0108000112-G00000-G0000-GOIG00-GAUDIT

26. TOTAL AWARD AMOUNT (For Govt. Use Only) $320,703.20

☐ 27a. SOLICITATION INCORPORATES BY REFERENCE FAR 52.212-1, 52.212-4. FAR 52.212-3 AND 52.212-5 ARE ATTACHED. ADDENDA ☐ ARE ☐ ARE NOT ATTACHED.

☐ 27b. CONTRACT/PURCHASE ORDER INCORPORATES BY REFERENCE FAR 52.212-4. FAR 52.212-5 IS ATTACHED. ADDENDA ☐ ARE ☐ ARE NOT ATTACHED

☐ 28. CONTRACTOR IS REQUIRED TO SIGN THIS DOCUMENT AND RETURN ______ COPIES TO ISSUING OFFICE. CONTRACTOR AGREES TO FURNISH AND DELIVER ALL ITEMS SET FORTH OR OTHERWISE IDENTIFIED ABOVE AND ON ANY ADDITIONAL SHEETS SUBJECT TO THE TERMS AND CONDITIONS SPECIFIED

☐ 29. AWARD OF CONTRACT: REF. ______ OFFER DATED ______. YOUR OFFER ON SOLICITATION (BLOCK 5), INCLUDING ANY ADDITIONS OR CHANGES WHICH ARE SET FORTH HEREIN IS ACCEPTED AS TO ITEMS:

30a. SIGNATURE OF OFFEROR/CONTRACTOR WG Pilcher

31a. UNITED STATES OF AMERICA (SIGNATURE OF CONTRACTING OFFICER) Paula Miller Sheelor

30b. NAME AND TITLE OF SIGNER (TYPE OR PRINT) W.G. PILCHER, VP

30c. DATE SIGNED 9/27/2012

31b. NAME OF CONTRACTING OFFICER (TYPE OR PRINT) PAULA Miller-Sheelor

31c. DATE SIGNED 9/27/2012

AUTHORIZED FOR LOCAL REPRODUCTION
PREVIOUS EDITION IS NOT USABLE

STANDARD FORM 1449 (REV. 3/2005)
Prescribed by GSA - FAR (48 CFR) 53.212

DOLF12XG21355

U.S. DEPARTMENT OF LABOR
OFFICE OF INSPECTOR GENERAL
OFFICE OF AUDIT

PROVIDE A BUSINESS INTELLIGENCE SOLUTION (BI-SOLUTION) COVERING VAST AND COMPLEX DATA TO EXTRAPOLATE MEANINGFUL RESULTS

STATEMENT OF WORK

Background

The Office of Inspector General – Office of Audit (OIG/OA) performs information and related information technology audits of the U.S. Department of Labor (DOL) cyber networks, systems, and related database information.

The OIG has an IT audit office located in the Frances Perkins Building, Washington, DC. This office is responsible for ensuring the Inspector General meets mandatory obligations to perform annual evaluations of the DOL's 72 major information systems' security controls, and provide targeted extracts of database information, including at times establishing the reliability of the information at a summary or transaction level. The latter activity is generally referred to as data mining or also called data or knowledge discovery (OIG's term of choice will be knowledge discovery). To achieve a reliable knowledge discovery capability, the OIG is seeking a BI Solution. The BI-Solution will be seen as a major tool in OIG's analysis of information from electronic databases to bring about knowledge that is not obvious from performing traditional audits or investigations. For example, the BI Solution results produced are expected to find meaning in data that can uncover hidden patterns, trends, anomalies and relationships that can transform the information into action items. BI results produced would be of the type that would empower data-driven decision making and place knowledge discovery in the forefront of audit planning. In addition, this knowledge is critical to the OIG in identifying high-risk areas, activities, and transactions as a way for the OIG to reduce waste, fraud and abuse in Federal government programs, contracts, and grants and may include predictive behavior modeling.

Objective and Scope

The objective is for OIG to obtain BI-Solution for performing knowledge discovery across heterogeneous DOL computing environments and related external systems that are comprised of large databases, including many within high Terabyte ranges.

Requirements

The contractor shall propose a BI-Solution for OIG to use in knowledge discovery activities that can span multiple, disparate operating systems, information networks, and databases in order to assess and analyze data from different organizational sources. The results generated should isolate issues involving potential waste, fraud, and abuse covering multiple topics such as revenue, expenses, performance measurement, enforcement actions, cost savings, entity and metadata identification, and/or combinations of any or all these elements. The BI Solution proposed should be capable of analyzing data from many different dimensions or angles, find correlations, trends or patterns among dozens of data sets and fields stored in multiple data warehouses and relational databases across multiple internal and external systems. The BI Solution's knowledge discovery results are to be displayed in a readily useable format for audit/investigation considerations and decision-making.

Proposal Contents

The contractor's oral proposal shall address the BI-Solution, and shall include the following:

1. Describe the contractor's BI Solution covering how robust the Solution is to meet the SOW Requirements, including the flexibility of the BI Solution adapting to future changes in these requirements.
2. Describe the BI-Solution in use with other organizations.
3. Describe the contractor's reputation based on key personnel's demonstrated relevant experience, especially description of Federal government experience. Include resumes of Project Manager and Technical BI Solution Expert (if not Project Manager).
4. Describe and demonstrate* how the BI-Solution covers the following areas:

   - degree of human intervention throughout process of readying/loading database information through to the completion of the BI-Solution results reporting.
   - technical functions to analyze trends, patterns, correlations, and other valued analytics
   - capability to "drill down" into summary information to view detail transactional data
   - support of forensic related functions such as analyzing associated meta data
   - wait time from initiation of query to receipt of result report(s) based on proposed BI Solution using contractor selected data sets/databases.
   - results reporting are displayed easily and quickly understood without detailed interpretations, e.g., graphs, tables, decision tree schematics, flowcharts, etc.

   * (1) OIG will arrange with the contractor a demonstration of the proposed BI Solution after receipt of the written contractor proposal.

   (2)OIG requests the demonstration be performed by the Project Manager and Technical BI Solution Expert, as appropriate.

   (3) OIG expects the contractor to perform analytics using data sets pertaining to the Department of Labor and other data sets that may be relevant to Labor's programs at the Federal Government's web site for Data.gov (http://www/data.gov) and may also include other related public accessible data bases/sets.
5. Describe how the BI Solution would be acquired and managed as:

a. a purchased product and continuous support component,
b. a contractor provided service and continuous support component, and
c. Including, an ongoing training component for the OIG for both the a. and b. proposed approaches above.

Basis of Award

The BI-Solution selected will be awarded based on the submitted proposal, technical capabilities of the BI Solution, demonstration of implementation and results, and best value to the government.

DOL12GRQ20069
Attachment 1

U.S. DEPARTMENT OF LABOR
OFFICE OF INSPECTOR GENERAL
OFFICE OF AUDIT

PROVIDE A BUSINESS INTELLIGENCE SOLUTION (BI-SOLUTION) COVERING VAST AND COMPLEX DATA TO EXTRAPOLATE MEANINGFUL RESULTS

STATEMENT OF WORK

Background

The Office of Inspector General – Office of Audit (OIG/OA) performs information and related information technology audits of the U.S. Department of Labor (DOL) cyber networks, systems, and related database information.

The OIG has an IT audit office located in the Frances Perkins Building, Washington, DC. This office is responsible for ensuring the Inspector General meets mandatory obligations to perform annual evaluations of the DOL's 72 major information systems' security controls, and provide targeted extracts of database information, including at times establishing the reliability of the information at a summary or transaction level. The latter activity is generally referred to as data mining or also called data or knowledge discovery (OIG's term of choice will be knowledge discovery). To achieve a reliable knowledge discovery capability, the OIG is seeking a BI Solution. The BI-Solution will be seen as a major tool in OIG's analysis of information from electronic databases to bring about knowledge that is not obvious from performing traditional audits or investigations. For example, the BI Solution results produced are expected to find meaning in data that can uncover hidden patterns, trends, anomalies and relationships that can transform the information into action items. BI results produced would be of the type that would empower data-driven decision making and place knowledge discovery in the forefront of audit planning. In addition, this knowledge is critical to the OIG in identifying high-risk areas, activities, and transactions as a way for the OIG to reduce waste, fraud and abuse in Federal government programs, contracts, and grants and may include predictive behavior modeling.

Objective and Scope

The objective is for OIG to obtain BI-Solution for performing knowledge discovery across heterogeneous DOL computing environments and related external systems that are comprised of large databases, including many within high Terabyte ranges.

Requirements

The contractor shall propose a BI-Solution for OIG to use in knowledge discovery activities that can span multiple, disparate operating systems, information networks, and databases in order to assess and analyze data from different organizational sources. The results generated should isolate issues involving potential waste, fraud, and abuse covering multiple topics such as revenue, expenses, performance measurement, enforcement actions, cost savings, entity and metadata identification, and/or combinations of any or all these elements. The BI Solution proposed should be capable of analyzing data from many different dimensions or angles, find correlations, trends or patterns among dozens of data sets and fields stored in multiple data warehouses and relational databases across multiple internal and external systems. The BI Solution's knowledge discovery results are to be displayed in a readily useable format for audit/investigation considerations and decision-making.

Proposal Contents

The contractor's oral proposal shall address the BI-Solution, and shall include the following:

1. Describe the contractor's BI Solution covering how robust the Solution is to meet the SOW Requirements, including the flexibility of the BI Solution adapting to future changes in these requirements.
2. Describe the BI-Solution in use with other organizations.
3. Describe the contractor's reputation based on key personnel's demonstrated relevant experience, especially description of Federal government experience. Include resumes of Project Manager and Technical BI Solution Expert (if not Project Manager).
4. Describe and demonstrate* how the BI-Solution covers the following areas:

    -- degree of human intervention throughout process of readying/loading database information through to the completion of the BI-Solution results reporting.
    -- technical functions to analyze trends, patterns, correlations, and other valued analytics
    -- capability to "drill down" into summary information to view detail transactional data
    -- support of forensic related functions such as analyzing associated meta data
    -- wait time from initiation of query to receipt of result report(s) based on proposed BI Solution using contractor selected data sets/databases.
    -- results reporting are displayed easily and quickly understood without detailed interpretations, e.g., graphs, tables, decision tree schematics, flowcharts, etc.

    * (1) OIG will arrange with the contractor a demonstration of the proposed BI Solution after receipt of the written contractor proposal.

    (2)OIG requests the demonstration be performed by the Project Manager and Technical BI Solution Expert, as appropriate.

    (3) OIG expects the contractor to perform analytics using data sets pertaining to the Department of Labor and other data sets that may be relevant to Labor's programs at the Federal Government's web site for Data.gov (http://www/data.gov) and may also include other related public accessible data bases/sets.
5. Describe how the BI Solution would be acquired and managed as:

a. a purchased product and continuous support component,
b. a contractor provided service and continuous support component, and
c. including, an ongoing training component for the OIG for both the a. and b. proposed approaches above.

Basis of Award

The BI-Solution selected will be awarded based on the submitted proposal, technical capabilities of the BI Solution, demonstration of implementation and results, and best value to the government.

DOL14A10006

Pursuant to BPA contract number(s) **DOL-OIG-14-A-0006**, a Blanket Purchase Agreement (BPA) is hereby established between **Elder Research, Inc.** and the U.S.DEPARTMENT OF LABOR (DOL) under the terms and conditions of the above stated contract(s) and the following terms and conditions are incorporated in this BPA:

I. AUTHORITY

This BPA is entered into pursuant to the Federal Acquisition Regulation Part 13.303-2.

II. DESCRIPTION OF AGREEMENT

This Blanket Purchase Agreement allows for ordering of services for **Business Intelligence Solution (BI-SOLUTION) Covering Vast and Complex**, by the U.S. Department of Labor, Office of Inspector General. Please be advised that the services to be provided for BPA **DOL-OIG-14-A-0006** is described below in Section III entitled "Scope of Work" The period of performance for this BPA encompasses five years.

Statement of Work

Business Intelligence solution (BI-SOLUTION)

Background

The Office of Inspector General - Office of Audit (OIG/OA) performs information and related information technology audits of the U.S. Department of Labor (DOL) cyber networks, systems, and related database information.

The OIG has an IT audit office located in the Frances Perkins Building, Washington, DC. This office is responsible for ensuring the Inspector General meets mandatory obligations to perform annual evaluations of more than 60 DOL major information systems' security controls, and provide targeted extracts of database information, including at times establishing the reliability of the information at a summary or transaction level. The latter activity is generally referred to as data mining or also called knowledge discovery. OIG has been using Elder Research Inc.'s RADR and Insight proprietary software for performing data mining and analytics of DOL's Office of Workers Compensation Federal Employees' Compensation Act chargeback, compensation, bill pay, case management and DOL OIG's investigative case file data.

Objective and Scope

The objective is for OIG to contract with Elder Research Inc. to provide an automated BI-Solution for performing knowledge discovery across heterogeneous DOL computing environments and related external systems that are comprised of large databases, including many within high Terabyte ranges.

Requirements

The contractor shall establish an automated BI-Solution for OIG and additional DOL program agencies using the proprietary RADR and Insight software for knowledge discovery activities that can span multiple program agencies' disparate systems, networks, and databases in order to assess and analyze the data. Information generated from these discovery activities is to further OIG audit targeting involving issues such as potential waste, fraud, and abuse and covering multiple topics such as revenue, expenses, payments, performance measurement, enforcement actions, mployee safety and retirement, cost savings, entity and metadata identification, and/or ombinations of any or all these elements. The use of the BI-Solution using RADR and Insight will e seen as an important tool in OIG's analysis of information from selected databases. It will llow users to analyze information from unforeseen data relationships, find correlations or atterns among dozens of data fields residing in a variety of types of databases such as flatile, relational, and/or hierarchical databases. Databases may also be ranged across multiple perating systems platforms such as Windows, Unix, Oracle, and Open Source.

Deliverables

- RADR
  - Perpetual Multiple DOL Progam Agency License
  - Perpetual Multiple DOL Progam Agency Maintenance License

- Insight
  - Perpetual Multiple DOL Progam Agency License
  - Perpetual Multiple DOL Progam Agency Maintenance License

- The mathematical models for current and future projects

- The Insight cubes (output via OLAP) and the underlying data for current and future projects.

- IT services, including training, that is in support of the use of RADR and Insight for current and future projects.

- Specific documentation, such as process and procedures related to the access, use and reporting of RADR and Insight and the Insight cubes.

The Government anticipates awarding various fixed price Time & Material (T&M labor hour) Task Orders from this BPA. Please provide a list of potential labor categories and associated prices.

ADMINISTRATIVE DATA

Primary Point of Contact: **W. Gerhard Pilcher**

Vice President & Senior Scientist

855-973-7673 ext. 707 (phone)

434-973-7875 (fax)

Gerhard@dataminginglab.com

Alternate Point of Contact: **Jeff Deal**

Vice President, Operations

434-227-5851 ext. 851 (phone)

434-973-7875 (fax)

deal@dataminglab.com

DOL-OIG Claimant Fraud Model

Prepared by Elder Research, Inc.

Sarah Will and Kris Hoover

Contents

Business Problem

Department of Labor Office of the Inspector General (DOL-OIG) contracted with Elder Research, Inc. (ERI) in FY 2013 to create a model to detect fraud in the Office of Workers' Compensation Programs (OWCP), specifically in the Federal Employee Claims Act (FECA). The process of modeling the data should highlight abnormalities in the data that can be used to form the basis of future audits.

Defining Fraud

DOL-OIG provided a data set, called the (b) (7)(E) [redacted] This file contains both claimant and provider cases. These cases will be used to define fraud for the modeling process.

Limitations

The OI file has both provider and claimant cases. Provider cases will not be used in this data analysis. Provider cases will be removed by virtue of joining the OI file with the Case Management file; the SSN of the provider will not be found with an associated case in the Case Management file.

The cases in the OI file represent a many-to-many situation. A claimant can have more than one criminal, administrative, or civil action taken against them. A claimant can also have multiple medical claims in FECA. Unfortunately, there is no way to know which particular case for a claimant led to the criminal/administrative/civil action; only an SSN is available from the DOL-OIG investigation results, not a claim number. Thus, one result might match to many cases in the Case Management file. We will discuss ways to limit this problem later on.

Criminal cases in the DOL-OIG FECA file have a number of different outcomes. They are

- DECL – Declined
    - The prosecution declined to prosecute the case. This could be for any number of reasons, including a lack of evidence, or an excessive workload at the district court.
- CV – Convicted
    - The claimant was convicted. The exact charge is unknown.
- DS – Dismissed
    - The case was dismissed because there wasn't enough evidence to make a decision.
- PT – Pre-trial
- AC – Acquitted
    - The claimant was found not guilty.

All cases will be considered fraud, regardless of the criminal outcome. After discussions with the team that runs the case tracking system, it was decided that all cases should be considered fraud, because the investigator believed it to be fraud. They want to be able to find cases similar to these in the future. The legal outcome was more likely a procedural side effect rather than a reflection of the merit of the case.

Administrative cases in the DOL-OIG FECA file have a number of different referral outcomes. They are:

- Benefit/Payments
- Cost Efficiencies
- Counseling
- Debarment/Suspension
- Declared Overpayment
- Forfeitures Crim/Civ
- None
- Recoveries
- Resignation
- Restitution Crim/Civ
- Revocation/Denial
- Termination
- Voluntary Restitution

As with Criminal outcome codes, all Administrative outcomes will be considered fraud for the purpose of modeling.

Civil cases are all against providers. They will not be considered in this analysis.

Data Sets

Chargeback

Data owner: DOL. The amount being charged to an agency for a case that belongs to one of their employees. Includes information from the Case Management file and summary information from the Bill Pay and Compensation files.

Compensation

Data owner: DOL. Cases that receive payment for a FECA claim. Includes repeated payments and one-time payments. Only includes active payments in a given time period.

Bill Pay

Data owner: DOL. Bill paid on behalf of DOL for a FECA case. Includes information about payment, whether it be to a pharmacy, hospital, or physician.

Case Management

Data owner: DOL. FECA cases that have been reported to DOL. Includes all open and closed cases.

OI Case File

Data owner: DOL-OIG. FECA cases that have been investigated by DOL.

Data Audit

A data audit was performed on Chargeback, Compensation, Bill Pay, and Case Management data. This included learning about the data, transforming the raw text files into denormalized tables, examining properties of the data, and much else. Here are a number of the things that we learned about each of the raw data sets:

Chargeback

Chargeback is provided in a fixed width format. The data has two record types, with different fixed widths, in the same file: Summary Records and Detail Records. There is a set of metadata rows associated with both record types. The record type is differentiated by the field "Record Type."

Chargeback data contains non-numeric characters as the last character in many of the numeric columns. This is an artifact from the way the data used to be stored in (b) (7)(E) These characters are a 1-to-1 replacement of a number. The conversion table is:

(b) (7)(E)

Chargeback contains numeric currency fields that are not identified as such. This means that decimal places for dollar values with cents are missing. Decimal places were added to the following fields:

(b) (7)(E)

Compensation

Compensation is provided in a fixed width format. The data has three record types, each with different fixed widths, in the same file. The major redefine is between Death Records, Temporary Disability Records, and Scheduled Awards Records. The record type is determined by the variable "Pay Type." 1

indicates Temporary Disability Records, 7 indicates Death Records, and 9 indicates Scheduled Award Records.

Payee addresses are broken into either physical mailing address columns or (b) information. The phrase (b) (7)(E) is found from characters from 114 to 128 if the records are [redacted] specific. Otherwise, the columns are broken down into address information.

Since the data is fixed width, the column boundaries are outlined by a data dictionary. After data inspection, it was discovered that there was one more byte of data than there should have been, as indicated by the data dictionary. A conversation with the data owner confirmed this, and they are in the process of correcting the data dictionary. The additional byte of data belongs to the final column in the data, Cash Receipt.

It was discovered during the import process that some of the Compensation files were shorter in width than other files. We determined that expanded columns were inadvertently left off of the data extract. We requested those additional columns from a subset of the months. We replaced the old extract files with the new extract files.

Compensation contains numeric currency fields that are not identified as such. This means that decimal places for dollar values with cents are missing. Decimal places were added to the following fields:

(b) (7)(E)

(b) (7)(E)

Bill Pay

Bill Pay is provided in a fixed width format. The data has three record types on the same file: Provider records, Hospital records, and Pharmacy records. The field breakdown is determined by the column "Provider Type:" F indicates a Pharmacy record, H indicates a Hospital record, and P indicates a Physician record. There are also redefines in the data depending on whether the record is from before or after 09/04/2003. Since all of our data is more recent than 2003, our record structure is set up to assume the columns for post-2003 data.

Bill Pay contains numeric currency fields that are not identified as such. This means that decimal places for dollar values with cents are missing. Decimal places were added to the following fields (the number of decimal fields is indicated in parentheses):

(b) (7)(E)

Case Management

Case Management is provided in a pipe delimited file. There are no redefines in the data. No issues with the raw data were discovered during the audit process.

Through this process, we learned about the concept of Retired Cases. These are the equivalent of historical cases. They were once active FECA cases, but have not received a payment(b) (7)(E)

When that time passes, they are moved off of the Case Management file and into the Retired database. While we could like to consider closed cases, we were told by the data owners that retired records were structured very differently and would be very difficult to retrieve. We decided to move forward with only cases that had been active recently and thus appear in the Case Management file.

Audit Files

Data audits were performed for all files mentioned above. Reports from the audits are attached in the appendix of this document.

Data Analysis

A big part of getting to know the data is to start asking questions. As an outsider without intimate knowledge of the data or the business processes, we ask the data many questions. Some of these questions are easily answered through discussions with the subject matter experts. Other questions might lead to discoveries in the data that were previously unknown. These can present potential vulnerabilities in the business process.

Case Number Overlap

To begin with, we found the common variable (case number) on all of the tables, and compared how often a specific case was found on each of the four files.

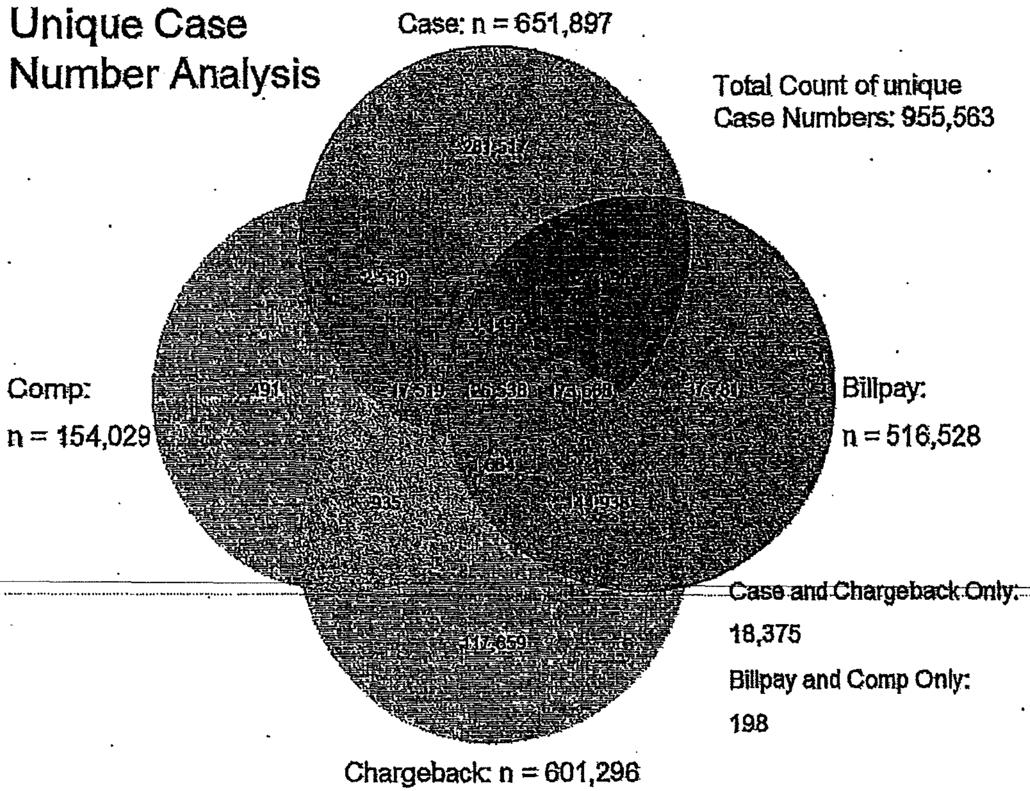


Figure 1. Unique Case Number Analysis

As you can see in Figure 1, there were a number of case numbers not represented on the Case Management table. This was a concern that we brought to the attention of the data owners who performed the extract, OWCP. They indicated that a filter removing all short-form closure cases was inadvertently left in place when the data was extracted for this project. They gave us a new extract of the data. We performed the same analysis with the new Case Management data.

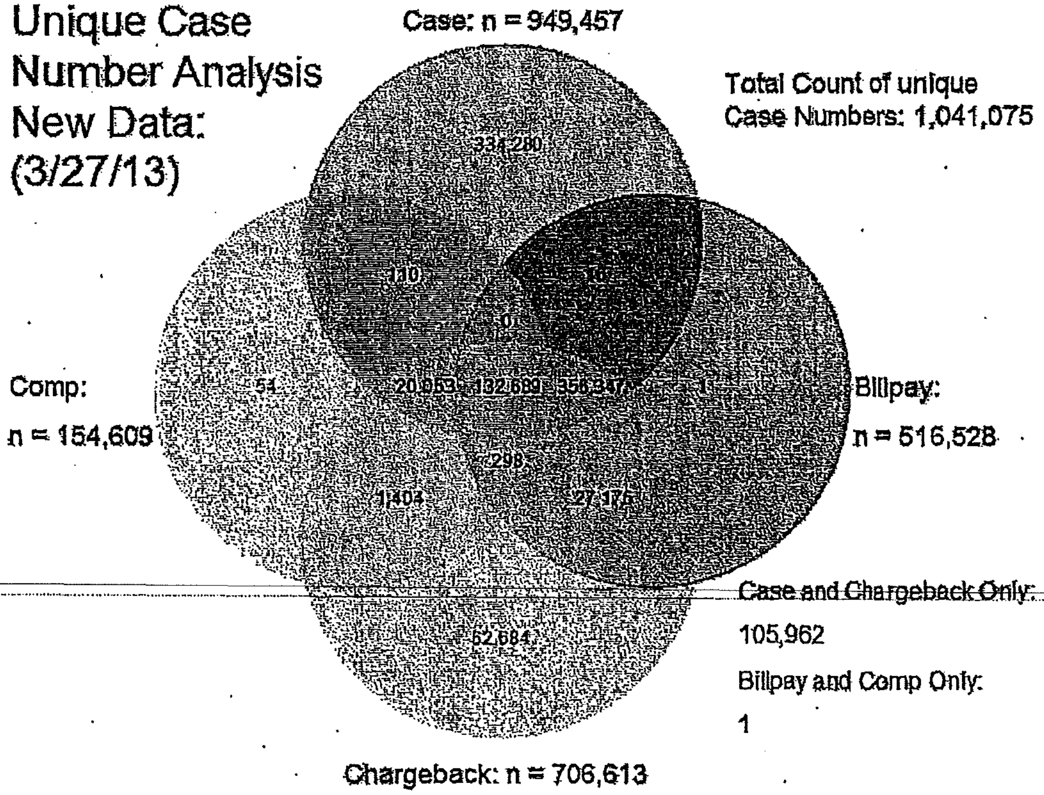


Figure 2. Update Unique Case Number Analysis

The data owners agreed that these new numbers made more sense. The lack of total overlap was of some concern. For example, that there were cases on Compensation, Chargeback, and Billpay that did not appear on the Case Management file. This was explained by Retired cases. Cases are "retired" after a certain timeframe of inactivity associated with the case. Since the Case Management file was snapshotted after the Compensation, Billpay, and Chargeback files were extracted, it is possible those discrepancies occurred because of the slight time difference.

Potential Audit Issues

Throughout the course of our data analysis, we noted what to us, as outsiders, seemed strange. Each of these could be a potential audit. Examples are presented below. These issues were presented in detail during status update meetings, and are summarized here.

Case Management

(b) (7)(E)

(b) (7)(E)

Compensation

(b) (7)(E)

Pre-Modeling

Winnowing Down the Modeling Dataset

Cases in the Case Management file were opened as long ago as 1940. This creates an issue of stale data. We would like to include only recent cases, since the nature of OWCP has surely changed in the last eight decades, but need a methodical way to do so. Luckily, this probably has been somewhat taken care by the practice of removing Retired cases.

Additional considerations include payment. Since we are modeling fraud with the goal of a monetary return to DOL and the employing agency, it would be important to only consider cases (b) (7)(E) (b) (7)(E) This was accomplished by inner joining the (b) (7) data and (b) (7)(E) data to the Case Management data on *case number*. Since we only have [redacted] data and [redacted] data since FY 2009, this will remove a number of cases from modeling consideration on the final analysis dataset.

Another consideration is undue weight given to a single claimant. This could happen if one claimant has many FECA claims. A way to avoid this is to only consider (b) (7)(E) The most recent case is determined by the (b) (7)(E) If a claimant has more than one case with the same date of injury, then the following tie breakers will be used: (b) (7)(E) (b) (7)(E)

Another consideration is to (b) (7)(E) At this point, whether or not a claim continues is out of their hands. They cannot actively be perpetrating fraud. So they will also be removed from the analysis dataset.

After all of these modifications to the Case Management dataset, which originally had 949,457 records, the new analysis dataset had 344,660 records.

Non-Fraud

As discussed earlier, some of the claims in the Case Management file are known to be fraudulent. We need a contrasting label for all other claims. In an ideal world, they would be known not fraud, meaning they had been investigated or audited and had been determined to be not fraudulent. Unfortunately, we do not have marked non-fraud cases. Instead, we classified all cases outside of the known fraud set as "unknown." From there, we winnowed down the cases in consideration, as outlined in the Creating a Modeling Dataset section. Additionally, data mining techniques were performed to attempt to identify those cases most likely to be non-fraud. That process is outlined in the Selecting Non-Fraud section.

Creating Variables

Many of the variables that were used for this data mining effort (such as injury type, injury nature, rehabilitation indicator) are categorical variables with a large number of categories. With our sparseness of known fraud cases and the large number of categories, we needed to narrow down the categories to those that are most important. We started by looking at the categorical responses that were most common in each of the variables for the universe of cases. We then performed chi-square tests to compare the frequency of these categories for fraud and non-fraud cases. This let us know the categories where fraud cases have a statistically higher-than or lower-than-expected ratio of occurrence.

(b) (7)(E) - Two variables available about each claimant is (b) (7)(E) and (b) (7) (b) (7)(E) From this, we can determine (b) (7)(E) This was put forth as a possible model variable.

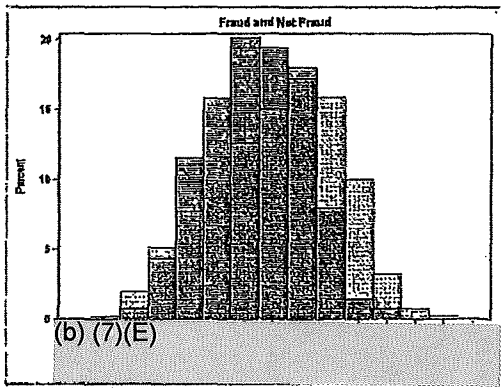


Figure 3. (b) (7)(E)

As can be seen in Figure 3, fraud cases tend to have (b) (7)(E)

Injury Variables

There are five different injury variables (b) (7)(E) Many of these variables have over 50 different possible categories. To include each category of each variable would water down the model significantly. Instead, we did analysis into the more prevalent categories from each variable. We considered each category that had at least 1% of the data. From there, we did a chi-square test for each variable, checking whether the category occurred statistically more often in fraud cases than non-fraud cases. This led to the creation of four binary variables:

(b) (7)(E)

Rehab Indicator

One available variable is whether or not a claimant is or was in rehabilitation for a specific injury. This variable was condensed down – any valid value was considered an indication of rehab, while a missing value was considered an indication of no rehab.

Number of Cases

This derived variable calculates the total number of claims within the Case Management file by (b) (7)(E)

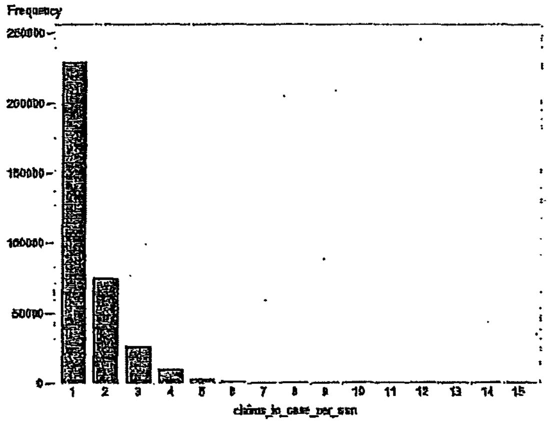


Figure 4. Claims in Case Management by (b) (7)(E)

ICD-9

There is a variable in the Case Management file called (b) (7)(E) The description from the data dictionary is (b) (7)(E) and the valid values are "Narrative, or up to 6 ICD-9 codes." Typically, we like to work with the Primary Accepted Condition. We followed up with the data owners to find out if the codes were listed in any specific order. We received the following response: "No, the codes aren't listed in any particular order and we don't consider any one code 'primary'. All codes are the injuries that are accepted in the claim, so all have equal primacy."

- 72% of cases have one or more ICD-9s
- 31% of cases have two or more ICD-9s
- 14% of cases have three or more ICD-9s
- 7% of cases have four or more ICD-9s
- 3% of cases have five or more ICD-9s
- 1% of cases have six ICD-9s

Unfortunately, the model and visualization needs a primary ICD-9, and without an established hierarchy, one had to be devised that was at least consistent among all cases. We picked the first ICD-9 and classified it as the Primary Accepted Condition.

Bill Reimbursement Percentage

Bills can either be paid directly to a provider, hospital, or pharmacy, or they can be reimbursed back to the claimant. Claimants can perpetrate fraud by having bills reimbursed to themselves that they never paid. One way to approach this is to look at the percent of bills that a claimant had reimbursed. We created this derived variable by dividing the number of reimbursed bills by the total number of bills paid. Note that this formula is blind to the monetary value of the bills.

Master and Subsidiary Cases

The Case Management file comes with two types of cases: Independent cases and "dependent cases," which are further broken down into two types: Master and Subsidiary. The following information was found on a DOL website (http://www.dol.gov/owcp/dfec/regs/compliance/DFECfolio/FECA-PT2/group5.htm) about when cases are classified as either dependent or independent:

> c. When to Double Cases. Cases should be doubled when correct adjudication of the issues depends on frequent cross-reference between files. Cases meeting one of the following tests must be doubled:
>
> > (1) A new injury case is reported for an employee who previously filed an injury claim for a similar condition or the same part of the body. For instance, a claimant with an existing case for a back strain submits a new claim for a herniated lumbar disc.
> >
> > (2) Two or more separate injuries (not recurrences) have occurred on the same date.
> >
> > (3) Adjudication or other processing will require frequent reference to a case which does not involve a similar condition or the same part of the body. For instance, an employee with an existing claim for carpal tunnel syndrome files a new claim for a mental condition which has overlapping periods of disability.

Based on these descriptions, we believe the subsidiary cases are different enough from the master case that they should be treated, and thus modeled, separately. Thus, each individual record in the Case Management file will be scored, regardless of the DOL perception of association between one or more cases.

Date of Injury

Based on a chi-square analysis, we determined that injuries were reported to have occurred more often on (b) (7)(E) and in (b) (7)(E) A binary flag for both of these concepts will be added to the analysis dataset. The analysis is shown below. In Figure 5 and Figure 6, the Day of the Week comparison is broken down.

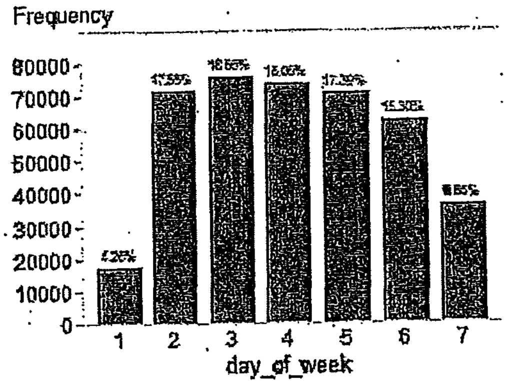


Figure 5. Number of Claims, by Day of the Week, for Non-fraud Cases

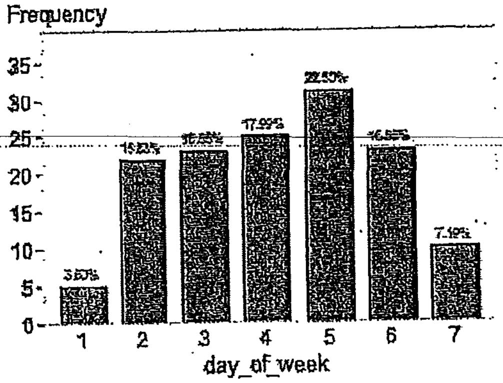


Figure 6. Number of Claims, by Day of the Week, for Fraud Cases

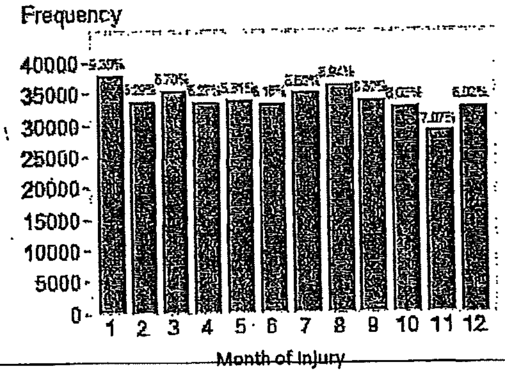


Figure 7. Number of Claims, by Month of Injury, for Non-Fraud Cases

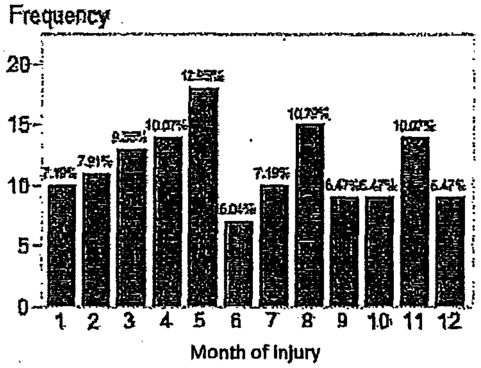


Figure 8. Number of Claims, by Month of Injury, for Fraud Cases

District

DOL has a number of different districts, based on geographic location of the claimant, which could be ~~responsible for handling a claim. We determined that District~~ (b) (7) ~~which covers~~ (b) (7)(E) ~~and~~ (b) (7)(E) has a higher than expected incidence level of fraud.

Transportation Expenses

In Bill Pay data, there are expenses related to transportation. The Office of Audit recently conducted an audit that focused on these payments and found that they were often unauthorized or extreme and did not have the necessary level of oversight. We created a metric that (b) (7)(E) (b) (7)(E) This variable became a new input into the model.

Modeling Methodology

Overview of the Modeling Process

Modeling is a three-step process: training, validation, and testing. First, we build many models using samples of the training data. This includes many types of models, such as logistic regressions and random forests, and many variations of those models. We continue to tune parameters for the model based on the insights learned from the results of previous models. We use validation data to evaluate how well a specific model with a certain set of tuned parameters works. Once we have a specific model for each methodology, such as a logistic regression and random forest, we compare those model types to each other using the testing data. Analyzing how well those models perform against each other using the testing data allows us to determine the final model that will be put into production.

Selecting Non-Fraud

The modeling approach used to identify potential fraudulent cases within the DOL data was more complex than a normal prediction model. A very low percentage of known fraud within the data prevented the usage of common methodology for identifying other fraud cases. This was caused by only having 139 observations of known fraud within the data, compared to 405,228 cases that may or may not be fraud. This means that 0.00034% of our analysis data set is truly fraud. It is extremely unlikely that 0.00034% is the true percentage of fraud cases within the data; therefore, some of the 405,228 cases whose fraud status is unknown are likely to be fraud. It is critical to identify these cases that are not classified as either fraud or not fraud but are probably fraud. Recognizing these cases is crucial because any models created from data where there are fraud cases classified as not fraud, would be biased and provide inaccurate predictions. To overcome this obstacle, a two-step modeling approach was constructed to reduce the probability of classifying fraud cases as not fraud.

Before any of the modeling was started, 30% (41 observations) of the fraud cases were removed from the list of fraud data and put into a testing data set. These 41 observations would be used at the very end of the modeling process to identify how well the model works on data it has never seen before. This process of splitting data into different subsets is known as data partitioning. Three subsets of the whole dataset are created throughout the modeling process. These three subsets are training data, validation data, and testing data. The training data is used to create many models using many combinations of variables and techniques. Once the numerous models have been created, they are then validated against the validation data set. A particular model could work very well against training data; however, the algorithm might be extremely over fit to training data. That model would perform very poorly against validation data, whereas models that do not suffer from over fitting will perform well against validation data. From this validation of the models, a final model should be selected to use in production. Once that model has been decided upon, the testing data is used to understand how effectively the production model will perform on unseen data. This allows final calculations to be made on the predictive power and accuracy of the production model. Reference to these three data partitions will be made in the rest of the Modeling Methodology section as the three partitions are created.

Once the 41 observations were set aside for model testing at the end of the modeling process, the two-step modeling approach could be started. The first step in the two-step modeling approach is designed to more accurately classify the cases that currently are unknown according to whether they are fraud or not fraud. These particular cases are then ranked based upon their likelihood of fraud. This ranking will not be exact, but will provide a more accurate representation of whether or not each data point is fraudulent. To create a ranked list of each unknown fraud observation, an initial logistic regression was created. This model was constructed using all of the variables that were designated to be used in the modeling process. Some of these variables are from the raw data, others are transformations of raw data, and several are binary indicators. The list of variables is:

(b) (7)(E)

(b) (7)(E)

All of these variables were used in a stepwise logistic regression model that had an entry alpha level of 0.18 and allowed for second order interactions to be tested. For this initial model, all of the 98 (139 initial cases, minus the 41 held out for testing) remaining fraud cases were modeled against the 405,229 cases (one case could not be used due to a missing gender) whose fraud status was unknown. The stepwise model ran for 20 iterations and identified the following variables as being important predictors:

(b) (7)(E)

lysis of Effects

DF	Wald Chi-Square	Pr > ChiSq
1	41.6643	<.0001
1	54.7440	<.0001
1	26.7040	<.0001
1	13.0133	0.0003
1	10.1836	0.0014
1	8.2502	0.0041
1	1.1536	0.2828
1	0.5615	0.4537
1	2.9614	0.0842
1	0.3950	0.5297
1	0.0491	0.8246
1	10.9505	0.0009
1	8.8371	0.0030
1	4.4409	0.0351
1	3.9471	0.0470
1	2.5385	0.1111
1	2.9470	0.0860
1	2.5967	0.1071

From this output, there are 18 parameters, aside from the intercept, that were used in the initial model. The importance of each parameter can be noticed by the Wald Chi-Square value. Also, parameters that have an '*' in their name are interaction parameters. This means that there is a relationship between these two variables that, when used together, enhances the predictive power of the model.

After the logistic regression model was created, all of the observations were scored using the model and a probability of fraud was generated for each of the cases with an unknown fraud status. These scores were then turned into an ordered list, which was used to determine which cases can be classified as not fraud. The determination of not fraud was made by the percentile of an observation's score. With an assumed response of not fraud, accurate models could then be created to predict fraud in DOL data.

Creating Modeling Datasets

Before the second step of the two-step modeling approach could commence, a sample needed to be taken from the known fraud cases and the list of unknown fraud cases to fully create the training, validation, and testing data sets. From the entire population of known fraud cases, 41 had already been set aside from the very beginning for use in the testing data set and had no impact on the first step of this methodology. The remaining 70% (98 observations), which were used in the first step of the modeling process, needed to be split into training and validation data for the second step of the modeling process. From these 98 data points, 71 were randomly sampled to be used in the training data set and the remaining 27 were used for the validation data set. This resulted with a breakdown of 50%, 20%, and 30% of known fraud cases into the training, validation, and testing data set, respectively.

The ordered list of unknown fraud observations then had to be sampled in order to complement the known fraud cases that had already been partitioned. First, observations to be put into the testing partition were selected by randomly sampling from the ordered list of unknown fraud cases below the 5th percentile mark. This means that, for the testing data set, the 5% of cases most likely to be fraud were withheld and could not be randomly selected to fall into the group selected as not fraud to complement the fraud in the testing partition. The 5th percentile cutoff is designed to mimic, in theory, the amount of fraud historically found in datasets. Based upon past research, it has been shown that 5% of any data contains fraud. Therefore, if the initial stepwise logistic regression served its purpose, the top 5% of the data in the ranked list should be the fraudulent cases. Since there should be no in fraud in the subset we are randomly selecting to complement the known fraud in the testing partition, the 5th percentile barrier was used to theoretically prevent any fraud from being classified as not fraud. This will not be completely accurate at preventing true fraud from being classified as not fraud, but this is desirable because the purpose of the testing partition is to mimic reality, and in real data, there will be actual fraud that is being considered not fraud. 778 observations were randomly selected and considered not fraud. The count of 778 was derived so that when added to the 41 observations of known fraud, the known fraud would be approximately 5% of the testing data set. Again, the 5% is the theoretical percentage of fraud in a data set.

After a sample had been taken from the ordered list of unknown fraud in order to complete the testing data set, the list had to be randomly sampled again to finish the creation of the training and validation data sets. When creating the testing data set, a 5th percentile cutoff was used; for the training and validation partition, a 15th percentile cutoff was used. The reasoning behind this is that it is important to be extremely accurate in the identification of not fraud cases for the training and validation components of the model creation process. If there are fraud cases that have been diagnosed as not fraud in either the training or validation partitions, there will be bias built into all of the created models and also in the model that is put into production based upon the validation data. 512 observations were selected for the validation partition, which results in a 5% rate of fraud when combined with the 27 known fraud cases in the validation data set. The rest of the observations below the 15th percentile were designated for the training data set. Different types of predictive models can handle vast differences in the size of the training data, and therefore, all remaining 343,231 observations were assigned to the training partition.

Partition	Total Size	Fraud Count	Not Fraud Count
Training	343,302	71	343,231

Validation	539	27	512
Testing	778	41	819

Table 2. Total Cases used in each of the modeling datasets

Models

After the data was properly partitioned, the second stage of the two-step approach was able to begin. For this part of the modeling methodology, 3 different algorithms were used to attempt to predict fraud. The 3 different algorithms used were logistic regression, random forest, and neural network. The same variables that were used in the initial model were also used for the logistic regression, random forest, and neural network.

Logistic Regression

The logistic regression was created in SAS. First, we fed all variables mentioned above into the logistic regression. We ran this 500 times. We kept track of the variables selected in each iteration of the logistic regression. From that list, we selected the variables that were selected in at least 50% of the models and always had scores of the same sign (i.e. always positive or always negative). We used these variables to run the final logistic regression. The following variables were selected:

(b) (7)(E)

This model performed very well in the out-of-sample data, as is demonstrated in the ROC Curve figures seen below. ROC (short for Receiver Operator Characteristics Curves) Curves are a common tool used to evaluate models:

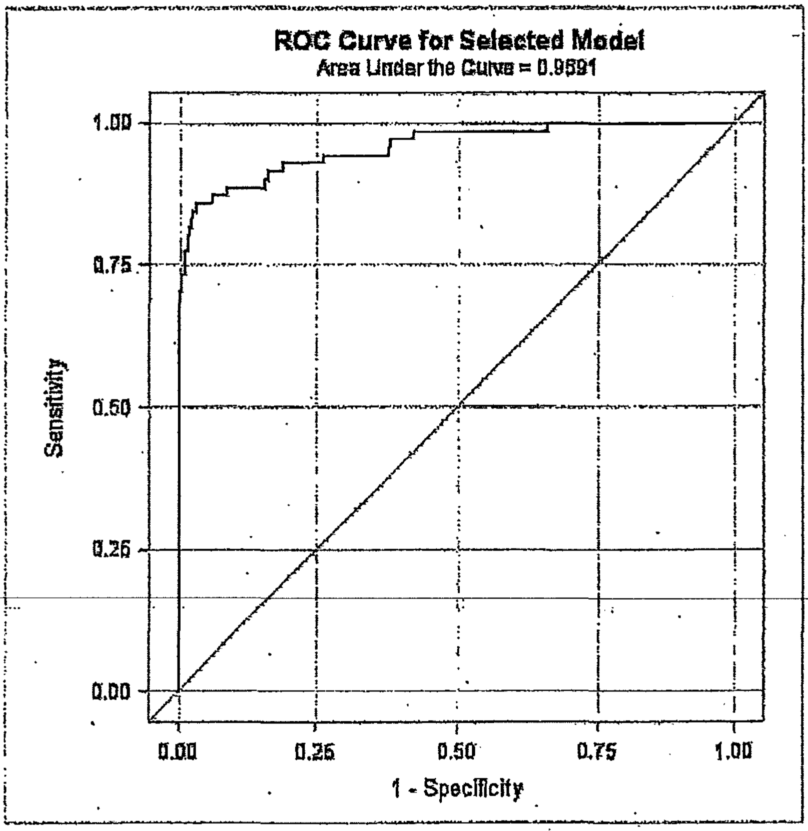


Figure 9. ROC Curve for Logistic Regression Model

Random Forest

A random forest model was also created to see how well it would stack up against the competition. A random forest model is a collection of decision trees that are ensembled together. Each individual tree uses different variables and different observations. These variables and observations are both randomly selected from the training data. In total, there were 1,000 trees generated for the random forest, each of them being built from 50 fraud cases and 950 not fraud cases.

Neural Network

Lastly, a neural network was created using the training data. For the neural network, there needed to be a defined set of not fraud observations. Because of this, 1,349 observations were taken from the not fraud training set to be paired with the 71 known fraud observations in the training set. 71 was selected to maximize the number of included fraud cases in the model. This caused the percentage of fraud to be 5% within the training data utilized for the neural network. The final neural network that was developed had a decay weight of 0.25 and contained 10 hidden layers.

Scoring

All three of these models were then scored against the validation data to understand how effective they are in predicting fraud.

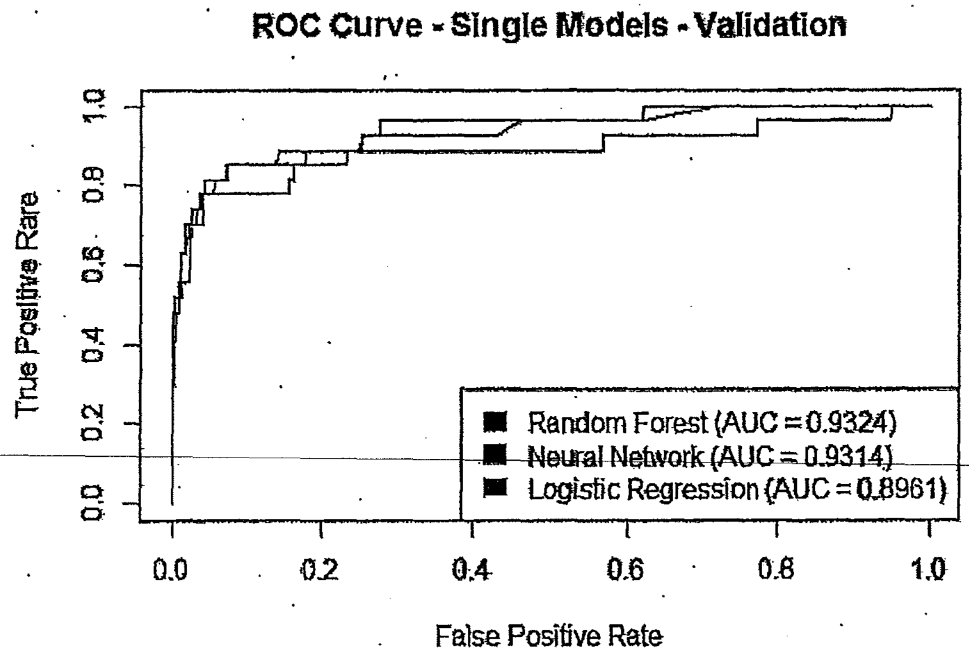


Figure 10. Comparison of Three Models' Performance on Validation Data

The measurement used to determine the effectiveness of the models is the area under the curve (AUC). This value is calculated by figuring out the area below the ROC curve. The ROC curve is a way to compare true positive rate to false positive rate. In order to compare the models together, the key idea to understand is that the model with the highest AUC value is the more desirable model. A baseline all AUC values can be compared to is an AUC of 0.5, which is pure random assignment of fraud and not fraud. Based upon the above graph, the random forest model performed the best, followed very closely by the neural network. The logistic regression performed still performed well, but not as well as the other two models. A few more plots to understand the effectiveness of each of these models are below.

The vertical line in these three plots is the cutoff between fraud and not fraud. Observations to the left of the vertical line are known fraud cases; therefore a higher predicted value is desired. Observations to the right of the vertical line are the not fraud cases and therefore would ideally have predicted values that are close to 0. This is another method to visually see how effective each of the three techniques is at predicting fraud versus not fraud.

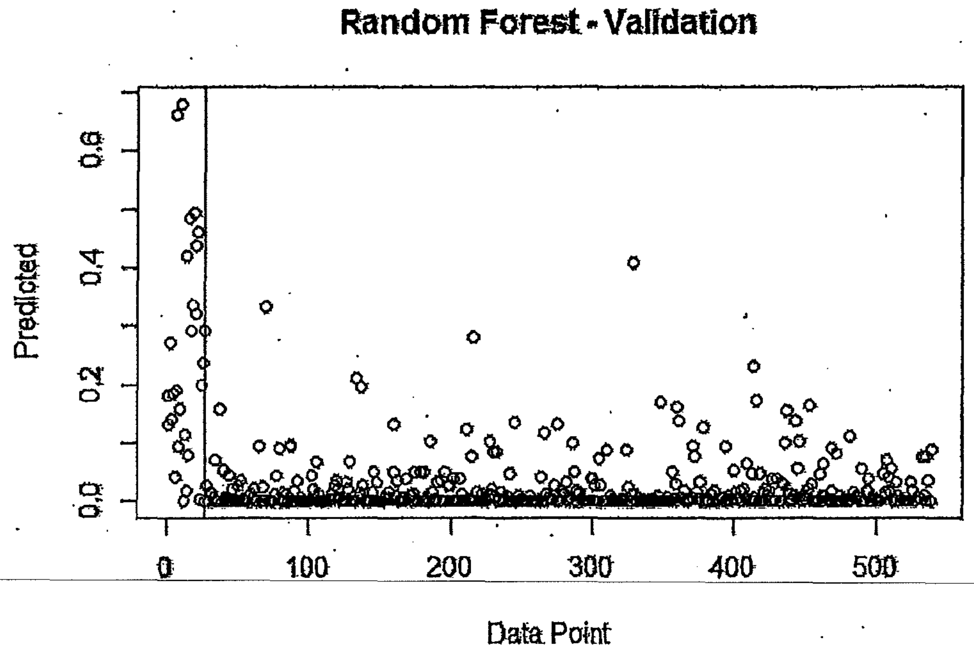


Figure 11. Random Forest Predicted Score on Validation Data

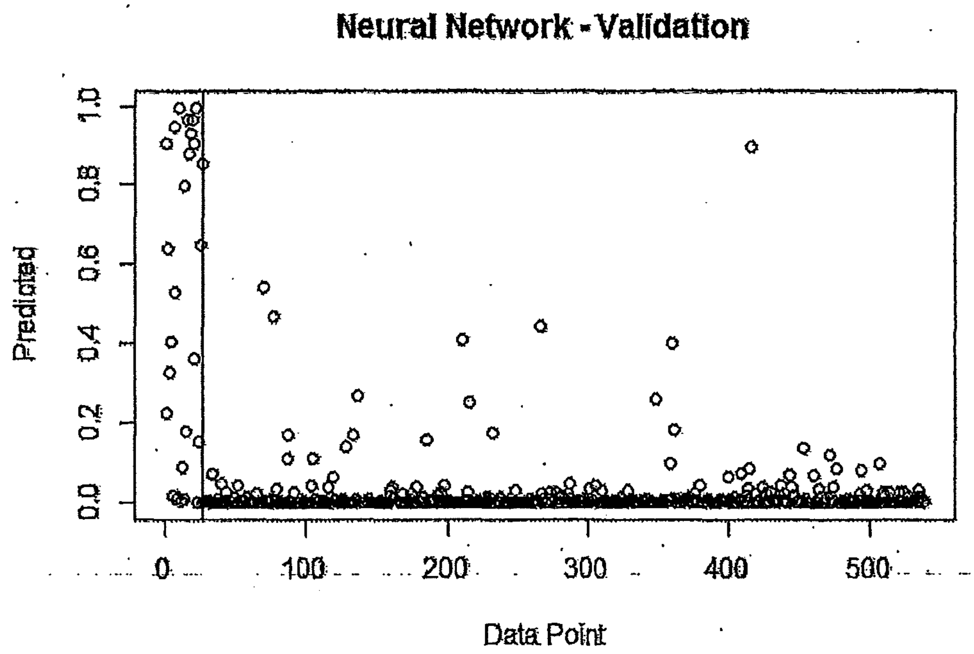


Figure 12. Neural Network Predicted Score on Validation Data

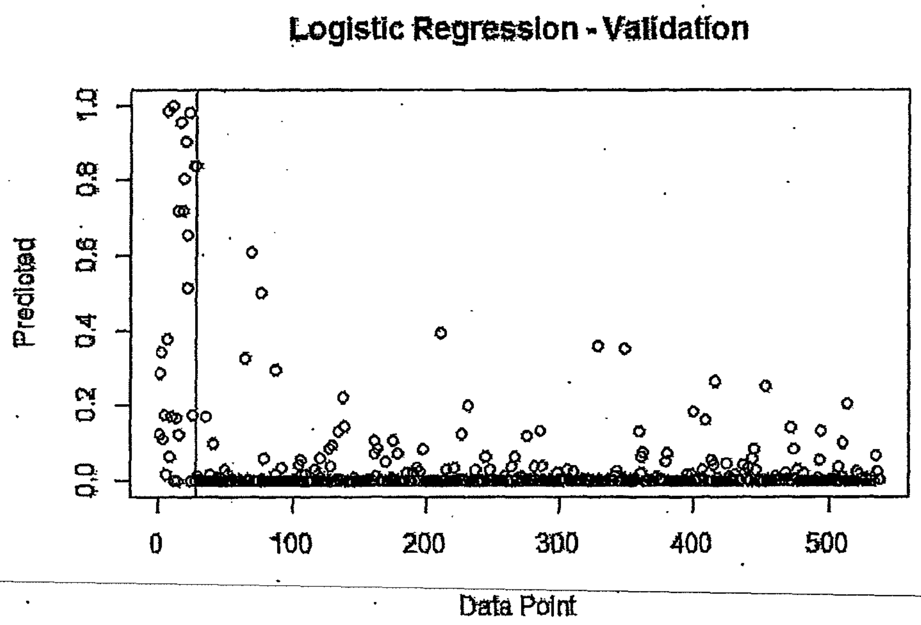


Figure 13. Logistic Regression Predicted Score on Validation Data

Ensembling

A way to improve model performance and to prevent overfitting is ensembling different modeling techniques together. Since a logistic regression, a random forest, and a neural network were all effective at predicting fraud based upon assessment statistics, the random forest was be ensembled with the neural network into a new model, and also all three were ensembled together. The ensembles were created by averaging the probabilities (predicted values) of the ensembled models together. The accuracy of the two new ensemble models can be seen in the following plots.

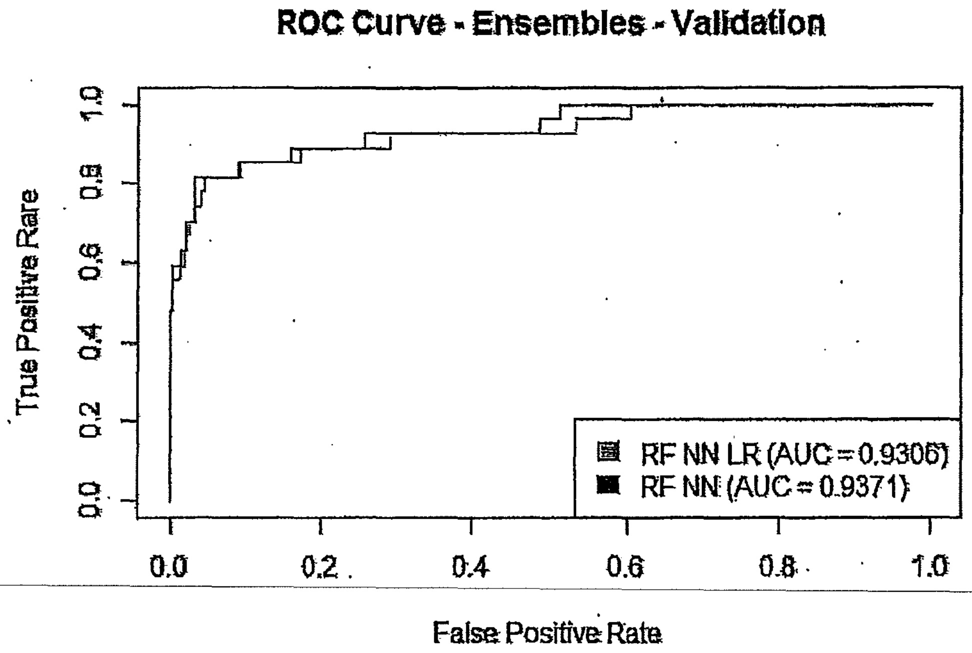


Figure 14. ROC Curve for Ensembled Models on Validation Data

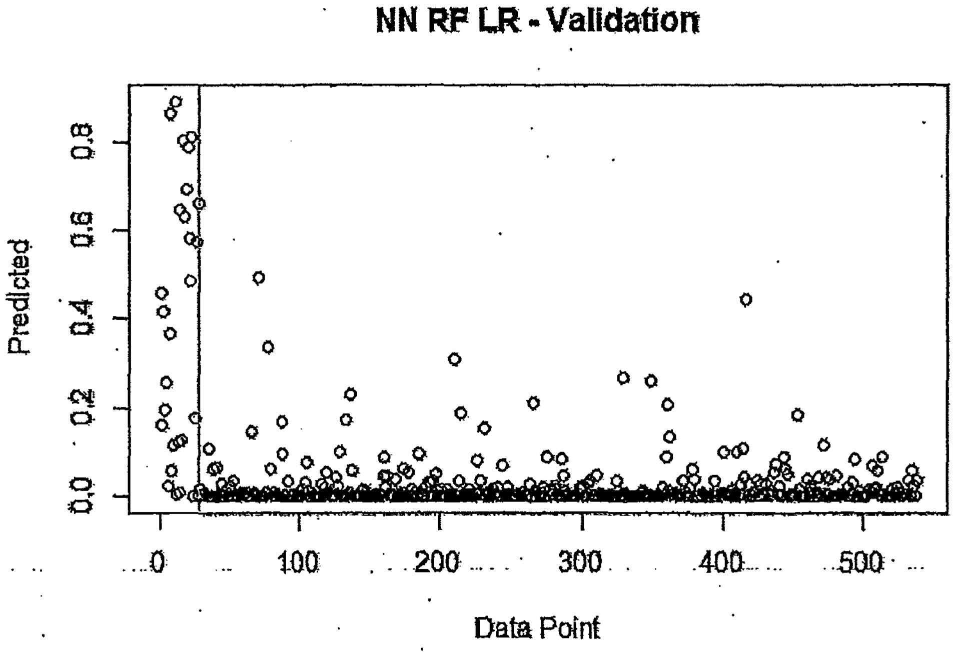


Figure 15. Ensembled Model Predicted Score on Validation Data

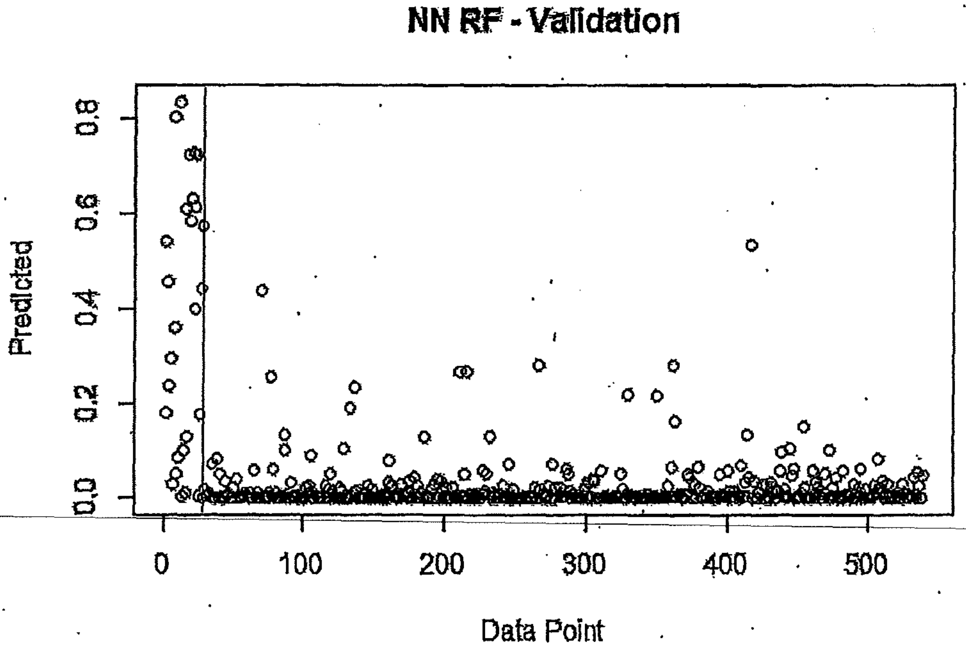


Figure 16. Predicted Scores for Ensembled Models on Validation Data

Based upon the AUC score for the new ensemble models, the random forest and neural network ensemble performed better out of the two ensembles and also better than any of the individual models.

Model	AUC (Validation Data)
Random Forest	0.9324
Neural Network	0.9314
Logistic Regression	0.8961
Logistic Regression + Random Forest + Neural Net	0.9306
Random Forest + Neural Net	0.9371

Table 3. Model Performance

Based on the performance of the random forest and neural network ensemble with the validation data, this model was selected to be the model used in production to determine which cases are fraudulent.

Model Evaluation

After the final model was chosen, the model was evaluated against the testing dataset to identify the model's effectiveness against unseen 'real world' data. For comparison, all of the other potential models were also assessed using the testing data; however, the ensemble of the random forest and neural network will be the model used in production, not any of the other models.

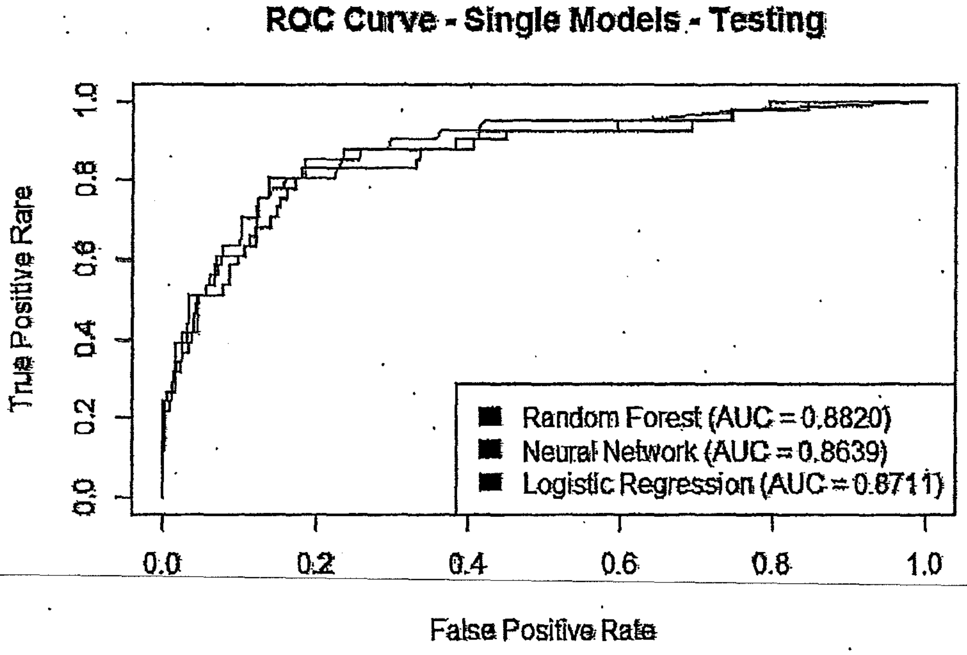


Figure 17. ROC Curve for Individual Models on Testing Data

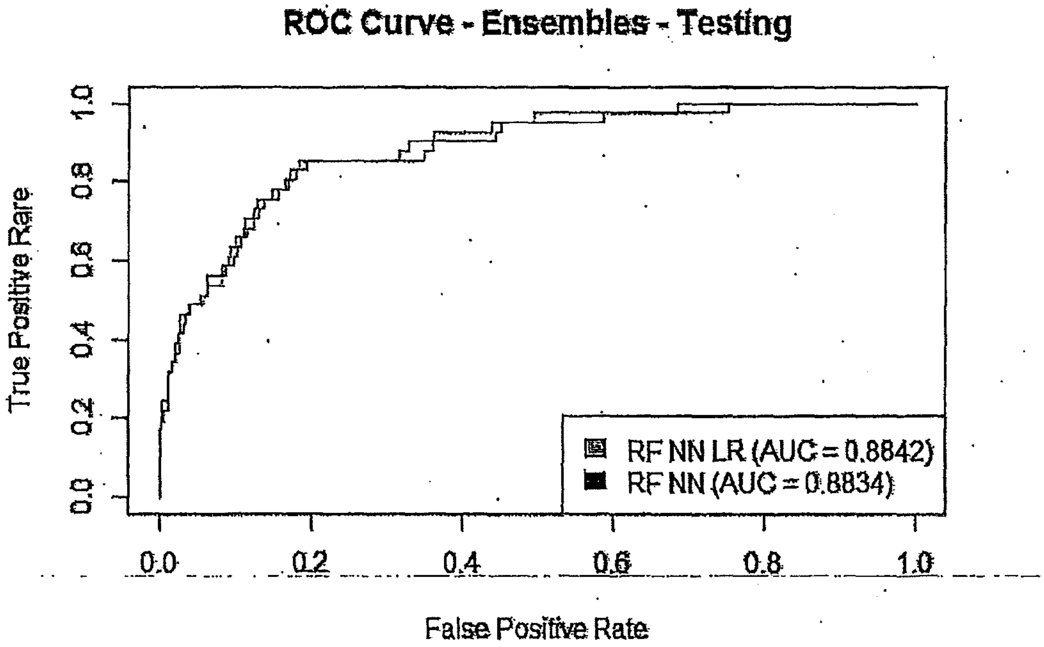


Figure 18. ROC Curve for Ensembled Models on Testing Data

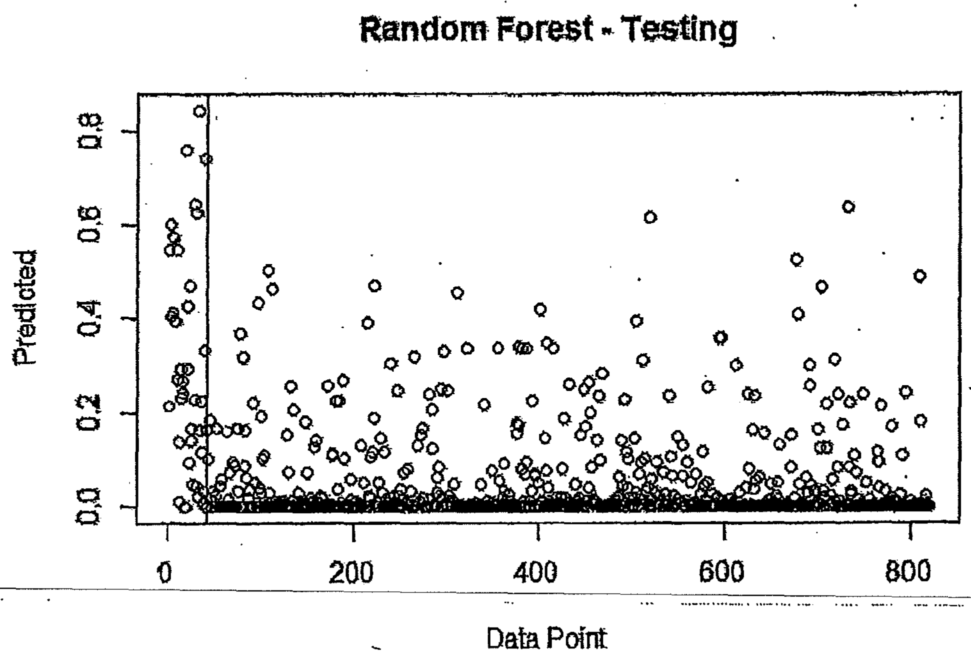


Figure 19. Predicted Values for Random Forest on Testing Data

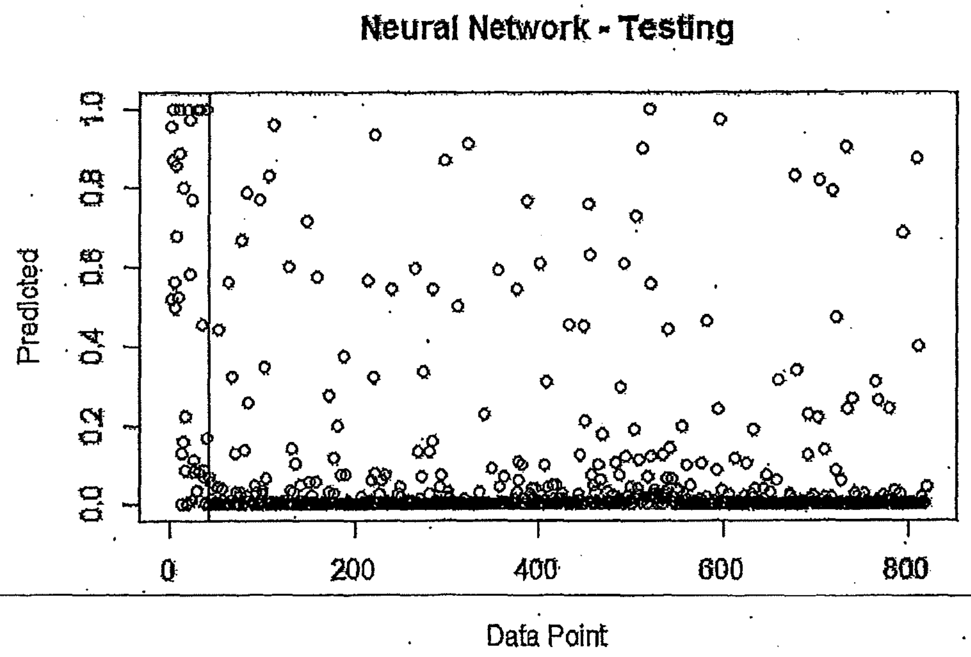


Figure 20. Predicted Values for Neural Network on Testing Data

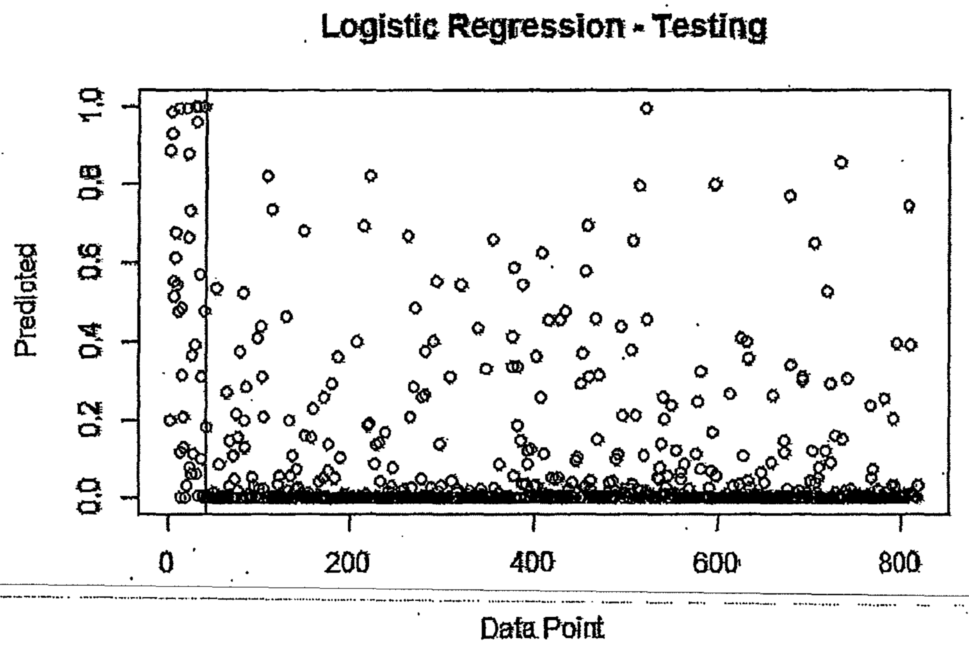


Figure 21. Predicted Values for Logistic Regression on Testing Data

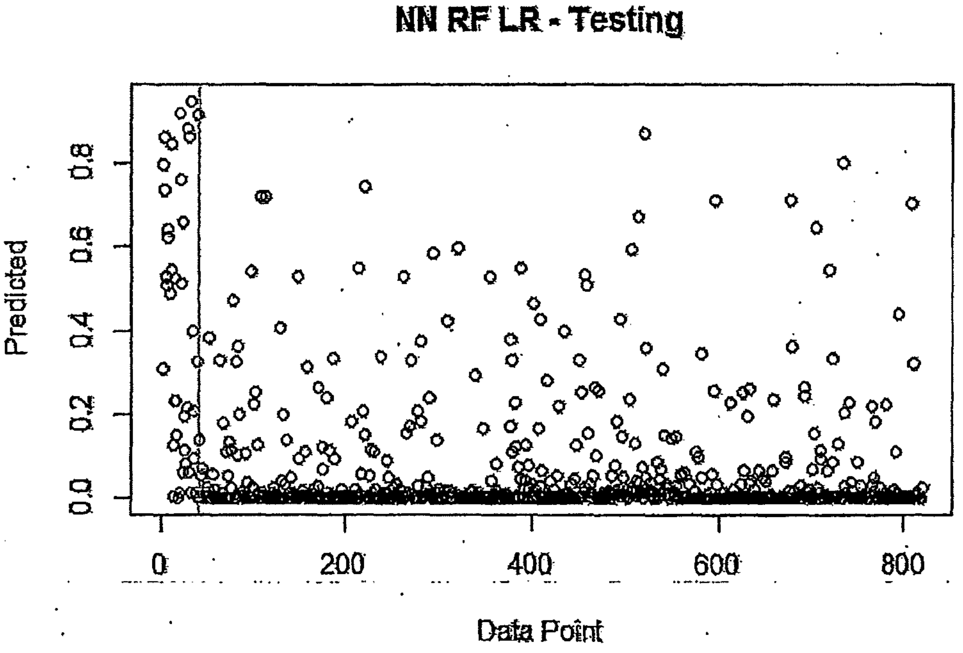


Figure 22. Predicted Values for Ensembled Model on Testing Data

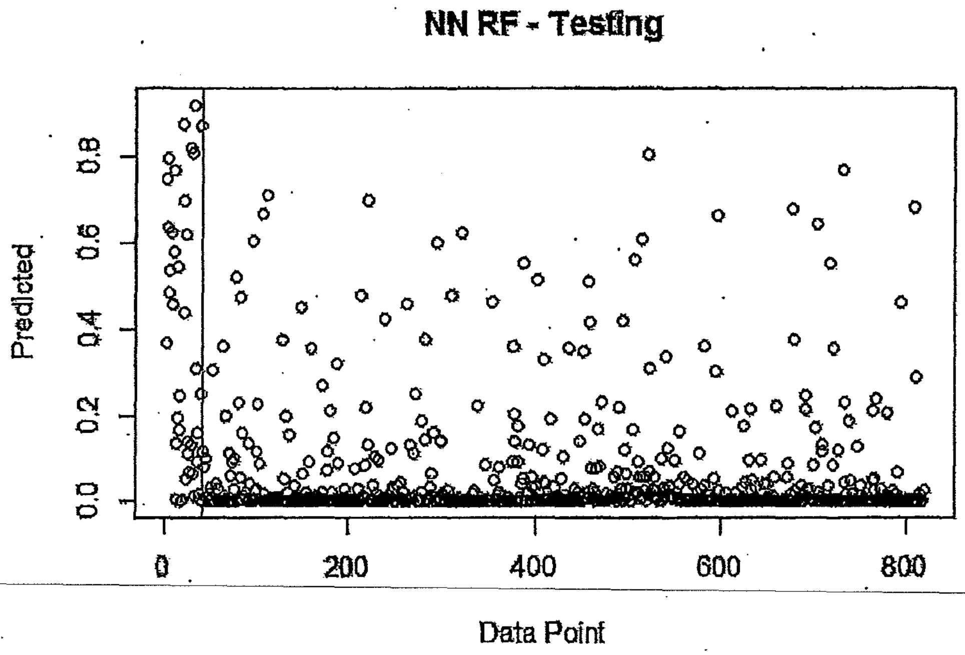


Figure 23. Predicted Values for Ensembled Models on Testing Data

From the evaluation of the models against testing data, we see the AUC scores are slightly down for each of the models. This is due to using more realistic out of sample data. The training and validation data sets were constructed below the 15th percentile whereas the testing data set was constructed below the 5th percentile. This idea can easily be visualized in the 5 plots contrasting fraud and not fraud points to the left and right of the vertical line, respectively. In the validation plots, the not fraud side of the plot had very few points that the models identified as potentially fraud, but now with the testing plots, more points are receiving a higher percent because the testing data has more points that are similar to fraud. This goes back to the notion of the 15th percentile and 5th percentile cutoffs. This fact is also the cause of the drop in AUC. All of this was completely expected based upon the differing natures of the two data sets.

Model	AUC (Testing Data)
Random Forest	0.8820
Neural Network	0.8639
Logistic Regression	0.8711
Logistic Regression + Random Forest + Neural Net	0.8842
Random Forest + Neural Net	0.8834

Table 4. Model Performance on Testing Data

Another interesting outcome of the running all of the models against the testing dataset was that the ensemble of the logistic regression, random forest, and neural network outperformed the ensemble of the random forest and neural network. The AUC difference between the 2 models is only 0.0008 and likely due to random chance, therefore there is no concern with the ensemble of the random forest and neural network being put into production.

Once the ensemble model of the random forest and neural network was put into production, every single case in the 949,457 observation data set was scored. This resulting score is the risk score that is currently being visualized in RADR. Since the cases that are known fraud cases were also scored through the model and put into RADR, it was possible to identify how much of the data would need to be analyzed in order to identify any percentage of the known fraud cases. The following chart shows how much of the data must be analyzed in order to identify the known fraud cases in 5% increments.

Percentage of fraud captured	Percentage of data investigated
5%	0.014%
10%	0.10%
15%	0.22%
20%	0.36%
25%	0.53%
30%	0.73%
35%	1.17%
40%	1.68%
45%	2.20%
50%	2.56%
55%	3.52%
60%	4.44%
65%	5.38%
70%	6.29%
75%	7.74%
80%	9.47%
85%	13.24%
90%	21.29%
95%	70.03%
100%	96.51%

Table 5. Percentage of Fraud Captured in relation to Percentage of Data Covered

DOL-OIG Data Mining

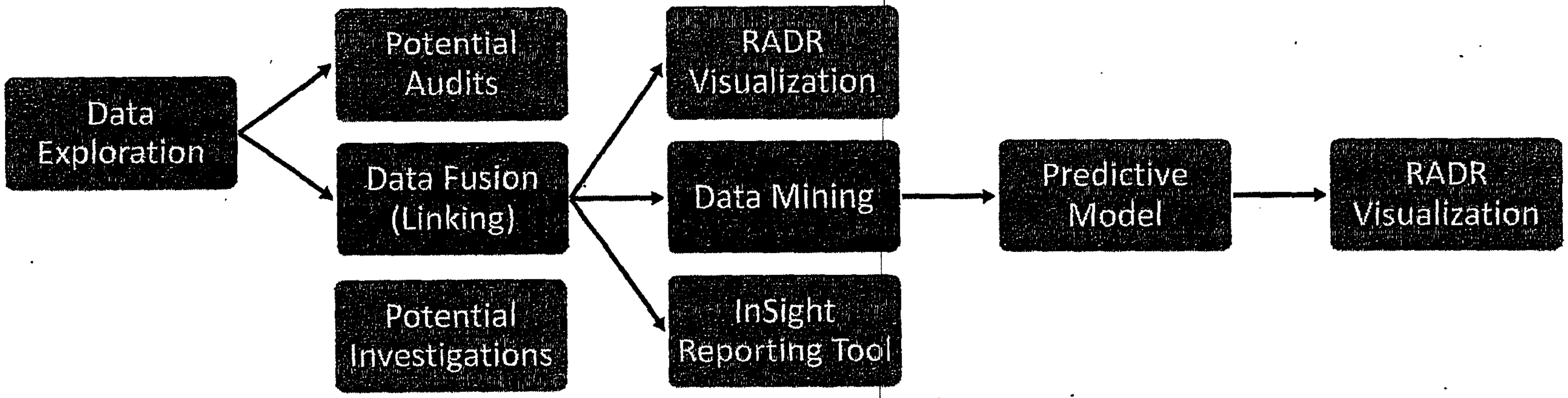
Results, Evaluations, and Next Steps

ELDER RESEARCH INC.
DATA MINING & PREDICTIVE ANALYTICS

Data Mining Project

- Healthcare Claimant (FECA) Data Mining
- Linking of four disparate data sources for analysis of 900,000+ FECA Cases
- Five Deliverables
  - Healthcare Claimant Risk Model
  - RADR
  - InSight
  - Risk Model Documentation
  - Program Analysis

The Process


Data Exploration
Potential Audits
Data Fusion (Linking)
Potential Investigations
RADR Visualization
Data Mining
InSight Reporting Tool
Predictive Model
RADR Visualization

Data Analysis

- Uncovers previously unseen patterns or insights from the data
- $4.5M in potential improper payments found from 1,750 cases labeled "medical only" still receiving compensation
- Identifies data anomalies that highlight potentially weak, non-existent controls

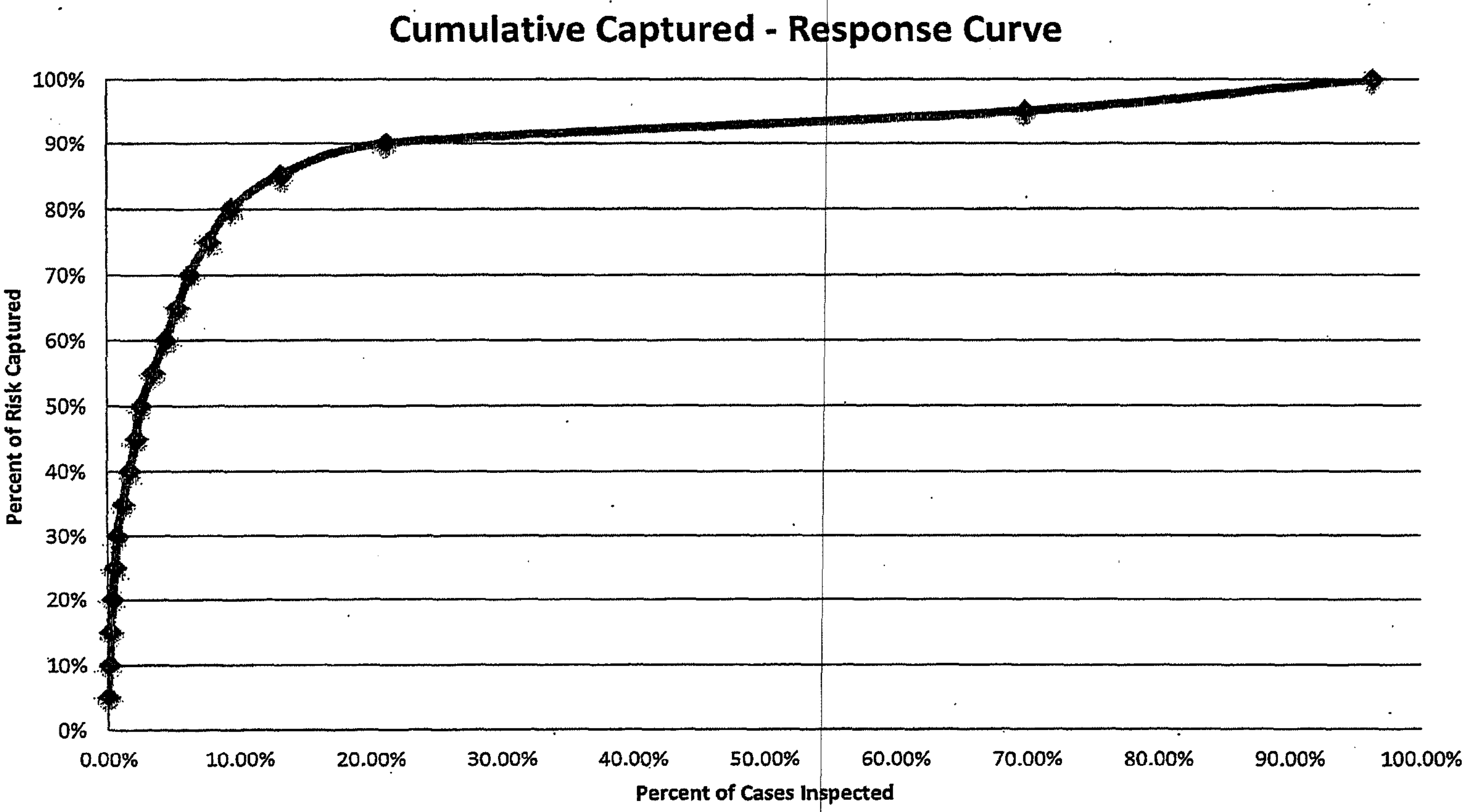
Claimant Risk Model – Data Mining

- Analysis of past data patterns and behaviors to predict future outcomes
- Produces a risk score that indicates risk magnitude and relativity
- Examines many types of compliance concerns, not just focused on a particular aspect of the claim

Claimant Risk Model – Data Mining

- Modeled on 29 variables and 400,000 case subset
- Tested three types of models
  - Logistic Regression
  - Neural Network
  - Random Forest
- Best performance came from an Ensembled Neural Network and Random Forest Model

Claimant Risk Model – Data Mining

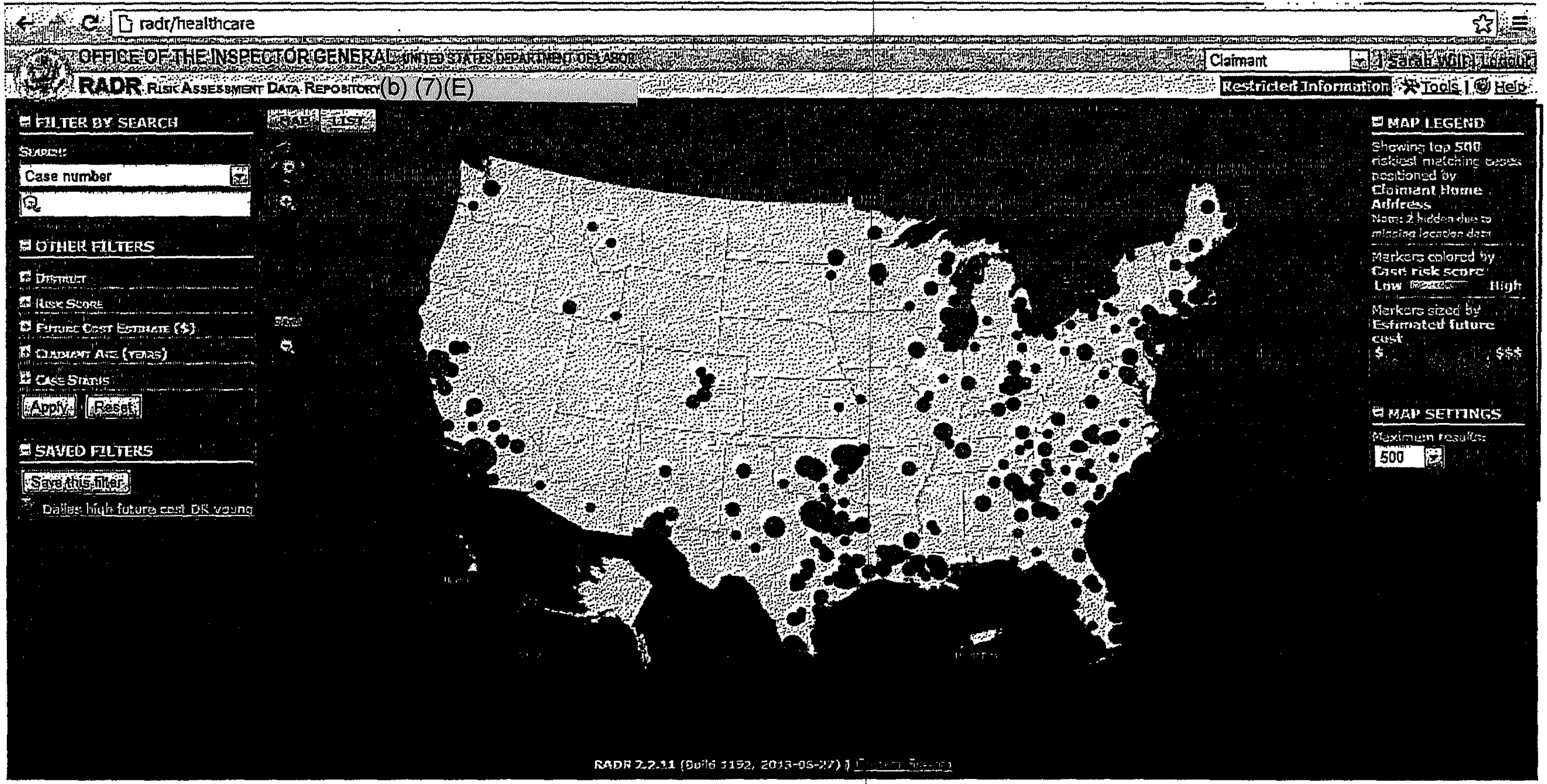


Introduction • Deliverables • Program Strategy • Program Growth • Benefits

RADR

- Claim-level data visualization
- Risk indicator and profile
- Customizable
- Expandable beyond Healthcare Claimant Risk

RADR - Map View



radr/healthcare

OFFICE OF THE INSPECTOR GENERAL UNITED STATES DEPARTMENT OF LABOR

Claimant | Sarah Wolf | Logout

RADR RISK ASSESSMENT DATA REPOSITORY (b) (7)(E)

Restricted Information | Tools | Help

FILTER BY SEARCH

Search:

Case number

OTHER FILTERS

- District
- Risk Score
- Future Cost Estimate ($)
- Claimant Age (years)
- Case Status

Apply | Reset

SAVED FILTERS

Save this filter

Dallas high future cost DR young

MAP LEGEND

Showing top 500 riskiest matching cases positioned by **Claimant Home Address**

Note: 2 hidden due to missing location data

Markers colored by **Case risk score**

Low High

Markers sized by **Estimated future cost**

$ $$$

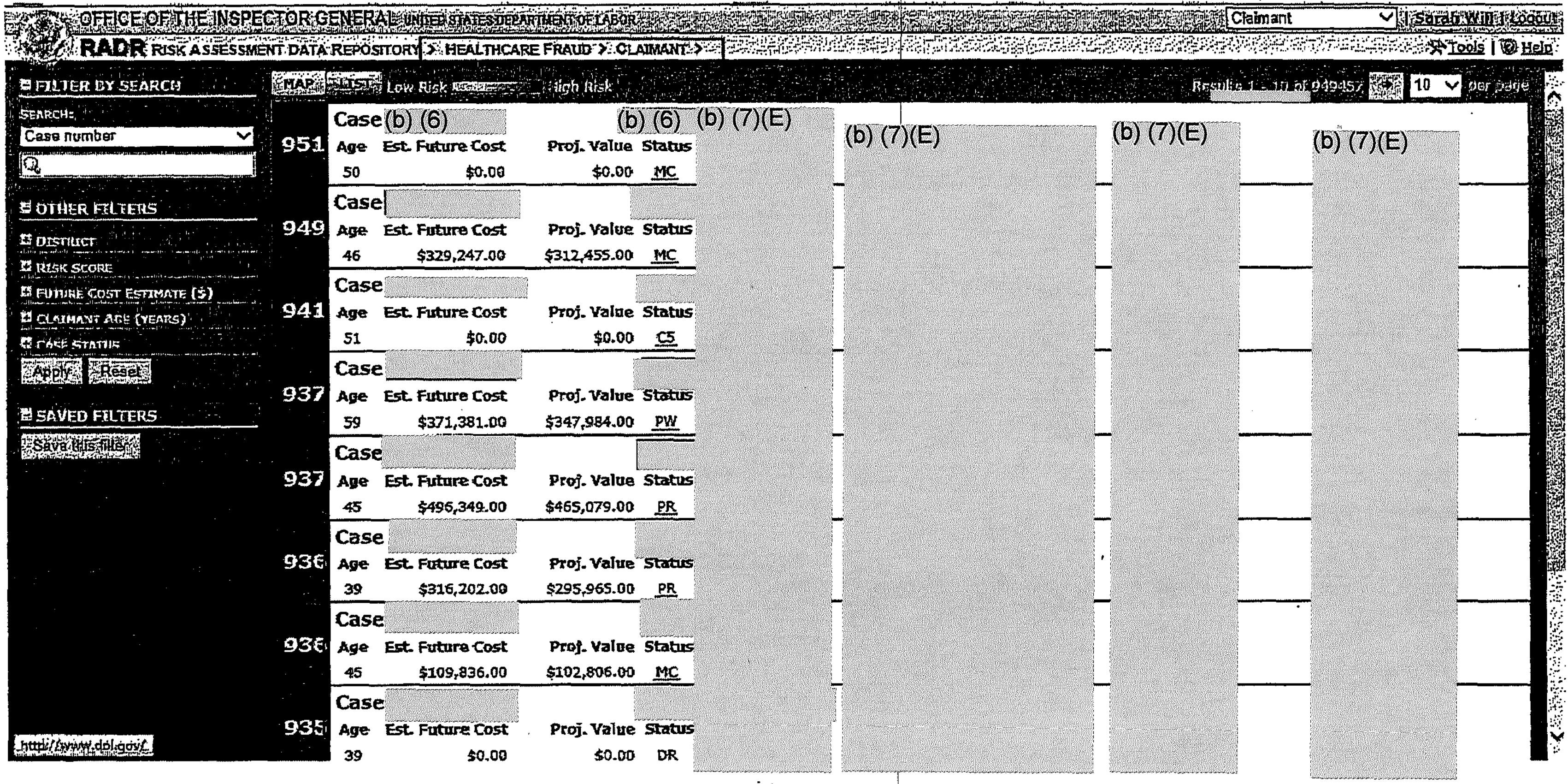
MAP SETTINGS

Maximum results:

500

RADR 2.2.11 (Build 1152, 2013-05-27) | [illegible]

RADR - List View



OFFICE OF THE INSPECTOR GENERAL UNITED STATES DEPARTMENT OF LABOR

RADR RISK ASSESSMENT DATA REPOSITORY > HEALTHCARE FRAUD > CLAIMANT >

Claimant | Sarah Will | Logout

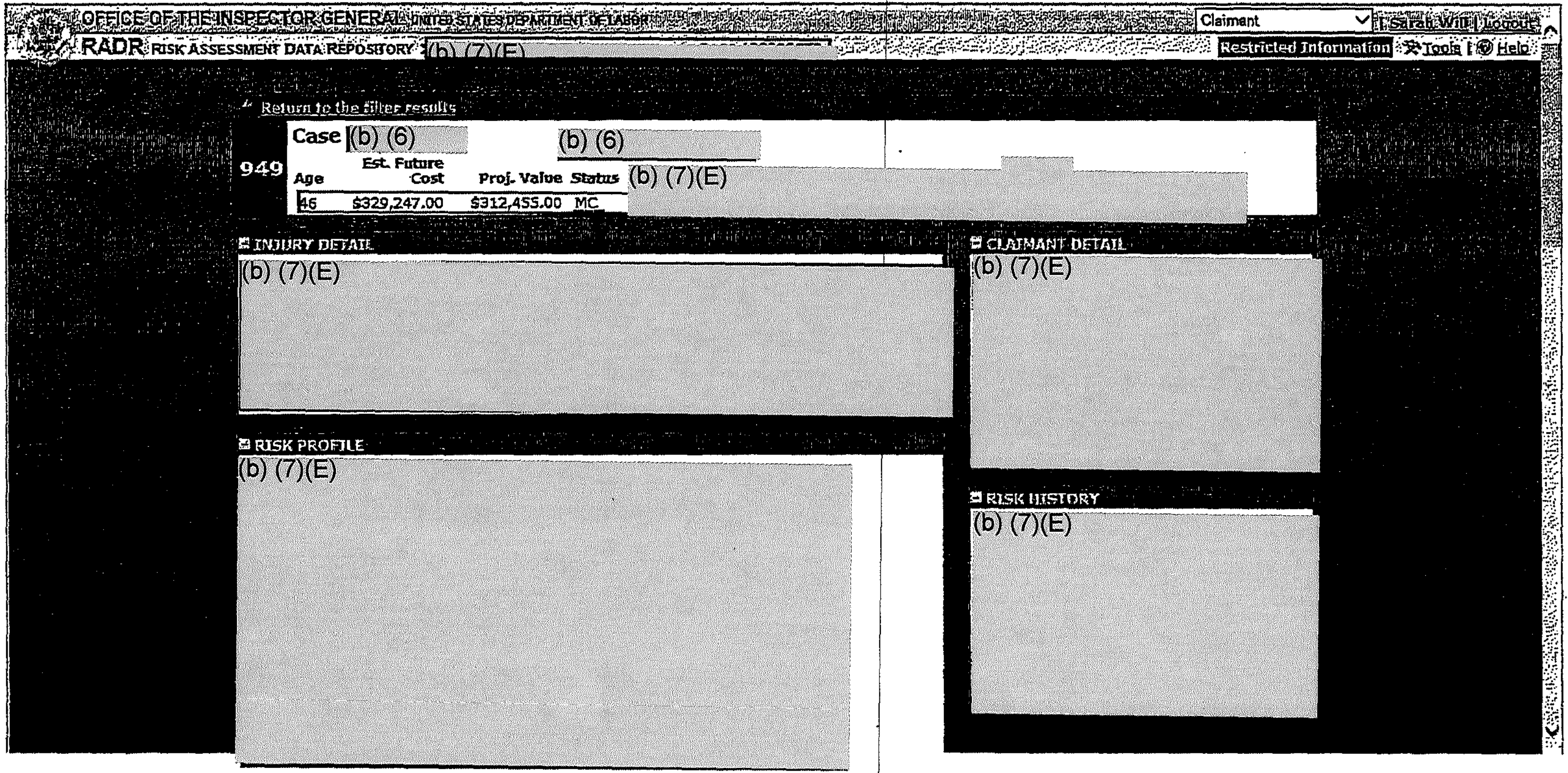
Tools | Help

Results 1 - 10 of 949457 10 per page

(b) (6) (b) (7)(E) (b) (7)(E) (b) (7)(E) (b) (7)(E)

	Case	Age	Est. Future Cost	Proj. Value	Status
951	(b) (6)	50	$0.00	$0.00	MC
949	[redacted]	46	$329,247.00	$312,455.00	MC
941	[redacted]	51	$0.00	$0.00	CS
937	[redacted]	59	$371,381.00	$347,984.00	PW
937	[redacted]	45	$496,349.00	$465,079.00	PR
936	[redacted]	39	$316,202.00	$295,965.00	PR
936	[redacted]	45	$109,836.00	$102,806.00	MC
935	[redacted]	39	$0.00	$0.00	DR

RADR - Detail View



OFFICE OF THE INSPECTOR GENERAL UNITED STATES DEPARTMENT OF LABOR
Claimant
| Sarah Will | Logout
RADR RISK ASSESSMENT DATA REPOSITORY
Restricted Information
Tools | Help
Return to the filter results

(b) (7)(E)

Case (b) (6) (b) (6)

949

Age	Est. Future Cost	Proj. Value	Status
46	$329,247.00	$312,455.00	MC

(b) (7)(E)

INJURY DETAIL

(b) (7)(E)

CLAIMANT DETAIL

(b) (7)(E)

RISK PROFILE

(b) (7)(E)

RISK HISTORY

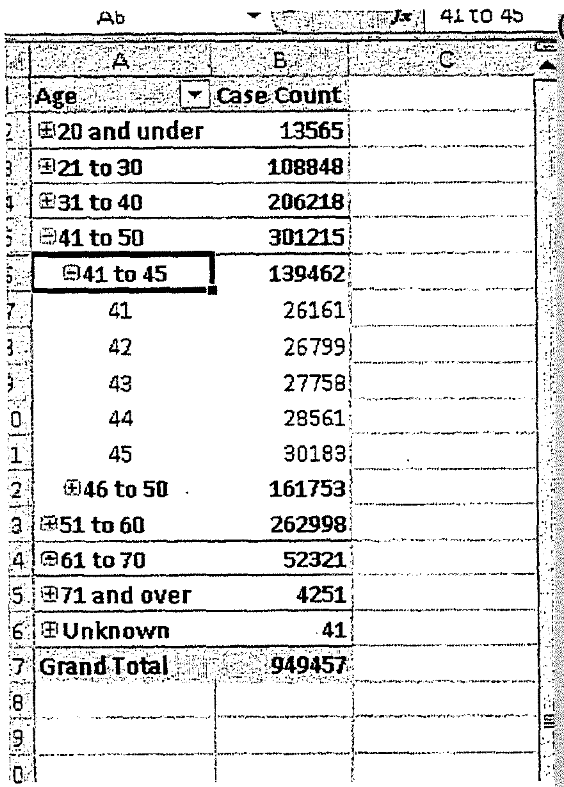
(b) (7)(E)

Introduction • Deliverables • Program Strategy • Program Growth • Benefits

InSight

- Data fusion
- Data querying and drilldown
  - Comparison of any field to any other field
  - Compute totals for subsets of data
- Audit trail

InSight

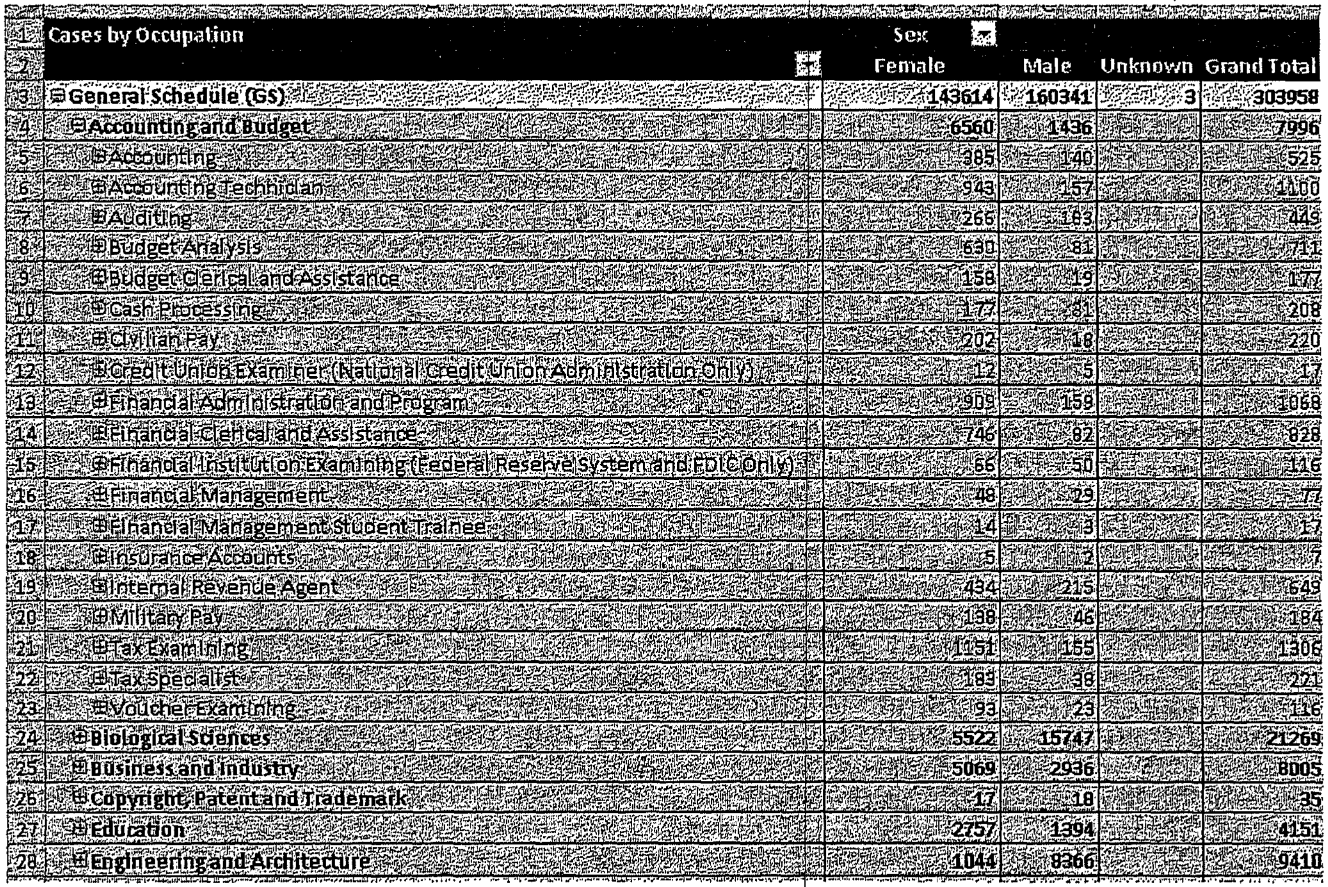


Age	Case Count
⊞20 and under	13565
⊞21 to 30	108848
⊞31 to 40	206218
⊟41 to 50	301215
⊟41 to 45	139462
41	26161
42	26799
43	27758
44	28561
45	30183
⊞46 to 50	161753
⊞51 to 60	262998
⊞61 to 70	52321
⊞71 and over	4251
⊞Unknown	41
Grand Total	949457

(b) (7)(E)

(b) (7)(E)

InSight



1	Cases by Occupation	Sex			
2		Female	Male	Unknown	Grand Total
3	⊟ **General Schedule (GS)**	143614	160341	3	303958
4	⊟ **Accounting and Budget**	6560	1436		7996
5	⊞ Accounting	385	140		525
6	⊞ Accounting Technician	943	157		1100
7	⊞ Auditing	266	183		449
8	⊞ Budget Analysis	630	81		711
9	⊞ Budget Clerical and Assistance	158	19		177
10	⊞ Cash Processing	177	31		208
11	⊞ Civilian Pay	202	18		220
12	⊞ Credit Union Examiner (National Credit Union Administration Only)	12	5		17
13	⊞ Financial Administration and Program	909	159		1068
14	⊞ Financial Clerical and Assistance	746	82		828
15	⊞ Financial Institution Examining (Federal Reserve System and FDIC Only)	66	50		116
16	⊞ Financial Management	48	29		77
17	⊞ Financial Management Student Trainee	14	3		17
18	⊞ Insurance Accounts	5	2		7
19	⊞ Internal Revenue Agent	434	215		649
20	⊞ Military Pay	138	46		184
21	⊞ Tax Examining	1151	155		1306
22	⊞ Tax Specialist	183	38		221
23	⊞ Voucher Examining	93	23		116
24	⊞ **Biological Sciences**	5522	15747		21269
25	⊞ **Business and Industry**	5069	2936		8005
26	⊞ **Copyright, Patent and Trademark**	17	18		35
27	⊞ **Education**	2757	1394		4151
28	⊞ **Engineering and Architecture**	1044	8366		9410

Introduction • Deliverables • Program Strategy • Program Growth • Benefits

9 Levels of Analytics

Descriptive Techniques:

1 – Standard Reporting

2 – Custom Reporting or "Slicing and Dicing" the Data (Excel)

3 – Queries/drilldowns (SQL, OLAP)

4 – Dashboards/alerts (Business Intelligence)

5 – Statistical Analysis

6 – Clustering (Unsupervised Learning)

Predictive Techniques:

7 – Predictive Modeling

8 – Optimization & Simulation

9 – Next Generation Analytics – Text Mining & Link Analysis

Data Analysis vs. Data Mining

Data Analysis	Data Mining
Having a human formulate questions and using the data to help answer them	Using the computer and data to figure out what questions should be asked, and helping you answer them

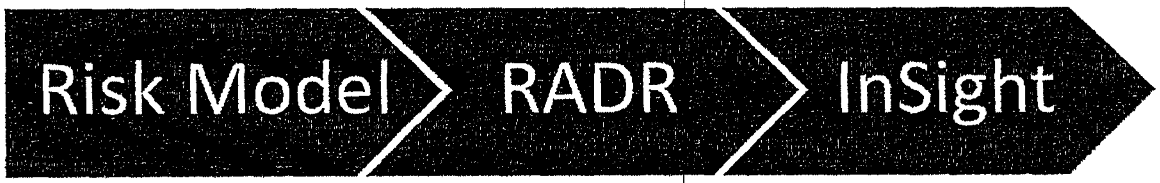
- Data Mining is a methodical combination of a multitude of data analyses
- The computer takes away the trial and error that a human would have to go through
- Allows for a multi-faceted approach to the underlying data instead of individual analyses

Skillset for Data Mining

- Unique quantitative skills from Mathematics, Statistics, Engineering, and Computer Science
- New advanced degree programs emerged to supplement on-the-job training
- Understanding advanced data mining algorithms and complex data mining software

Putting The Suite to Work

- Claimant Risk Model - gain knowledge to direct the efforts towards finding program issues and identifying specific instances of risk
- RADR - view attributes of risky claimants
- InSight - follow up with and zoom in on program issues that have been identified



Risk Model

RADR

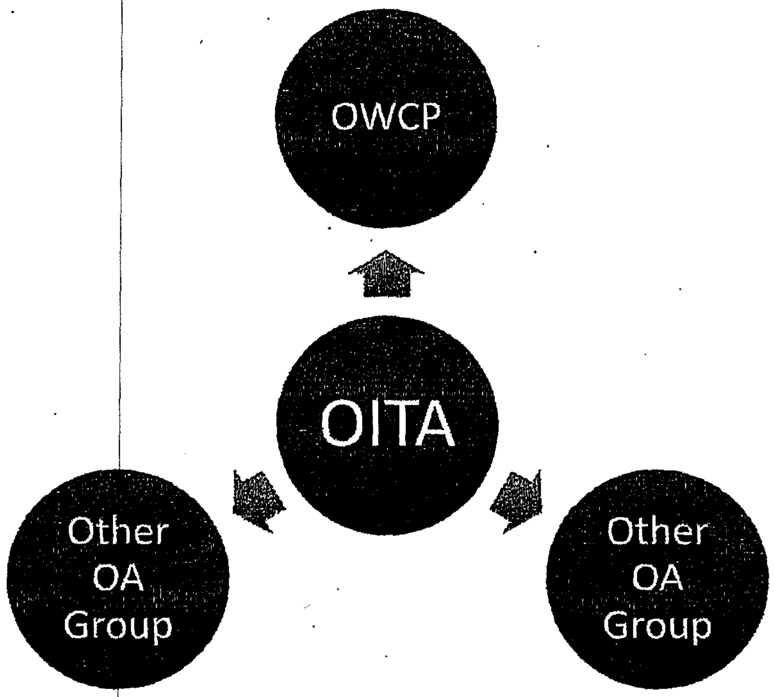
InSight

OI as an Audit Customer

- Share tools, knowledge, and leads
- RADR and InSight can make communication between multiple teams easier
- Decreases investigator reliance on tips and increases overall agency ROI

Program Strategy

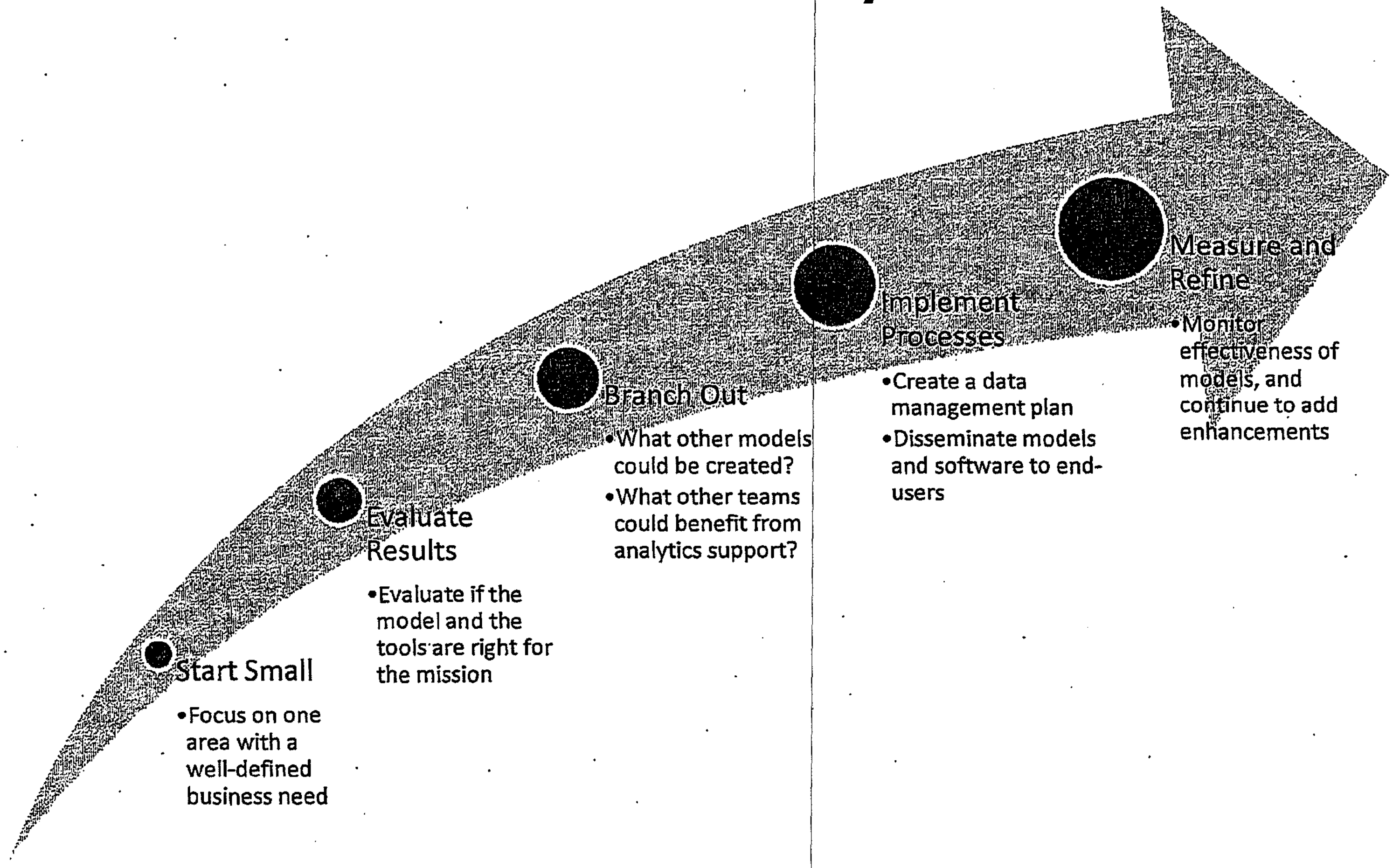
- IT Audit team oversaw the Data Mining project for the benefit of the OWCP Audit team
- Group has positioned itself to help additional teams and program areas increase their impact by adopting data mining techniques



Lessons Learned

- The process of data analysis in a new project almost always uncovers new insights
- SME review and participation
- Maximum benefit comes from experimenting with tools throughout auditor workflow

Growth of an Analytics Team



Start Small

- Focus on one area with a well-defined business need

Evaluate Results

- Evaluate if the model and the tools are right for the mission

Branch Out

- What other models could be created?
- What other teams could benefit from analytics support?

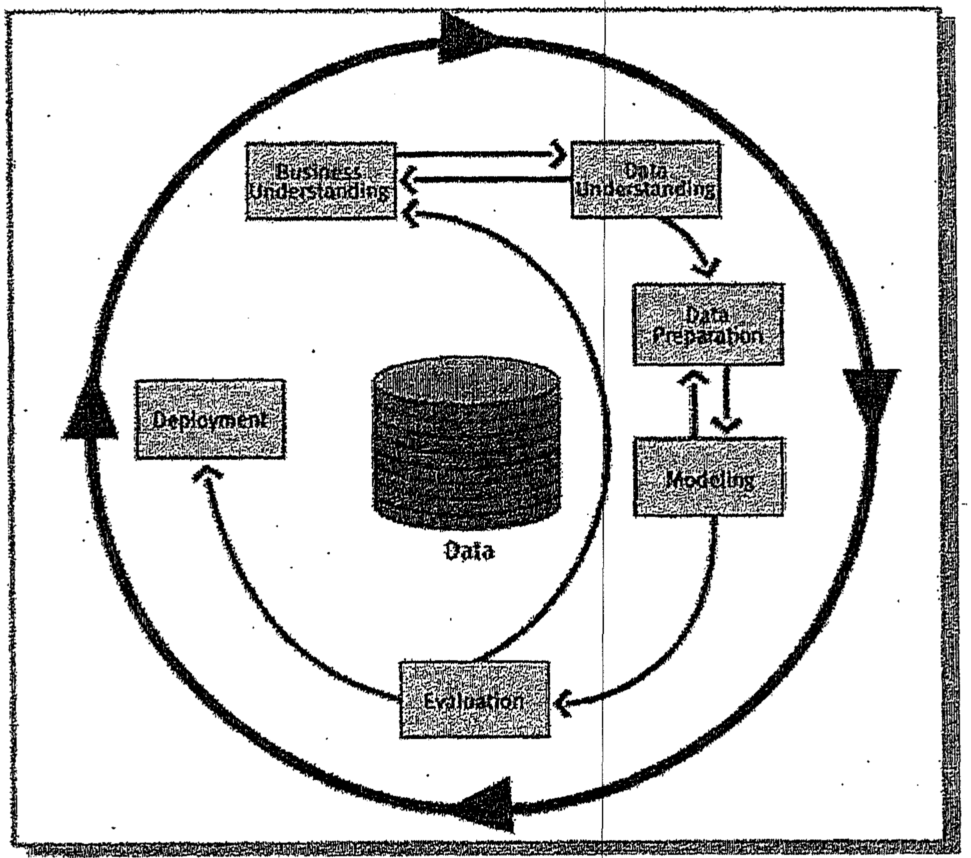
Implement Processes

- Create a data management plan
- Disseminate models and software to end-users

Measure and Refine

- Monitor effectiveness of models, and continue to add enhancements

CRISP-DM



Introduction • Deliverables • Program Strategy • Program Growth • Benefits

Potential Growth Areas

(b) (7)(E)

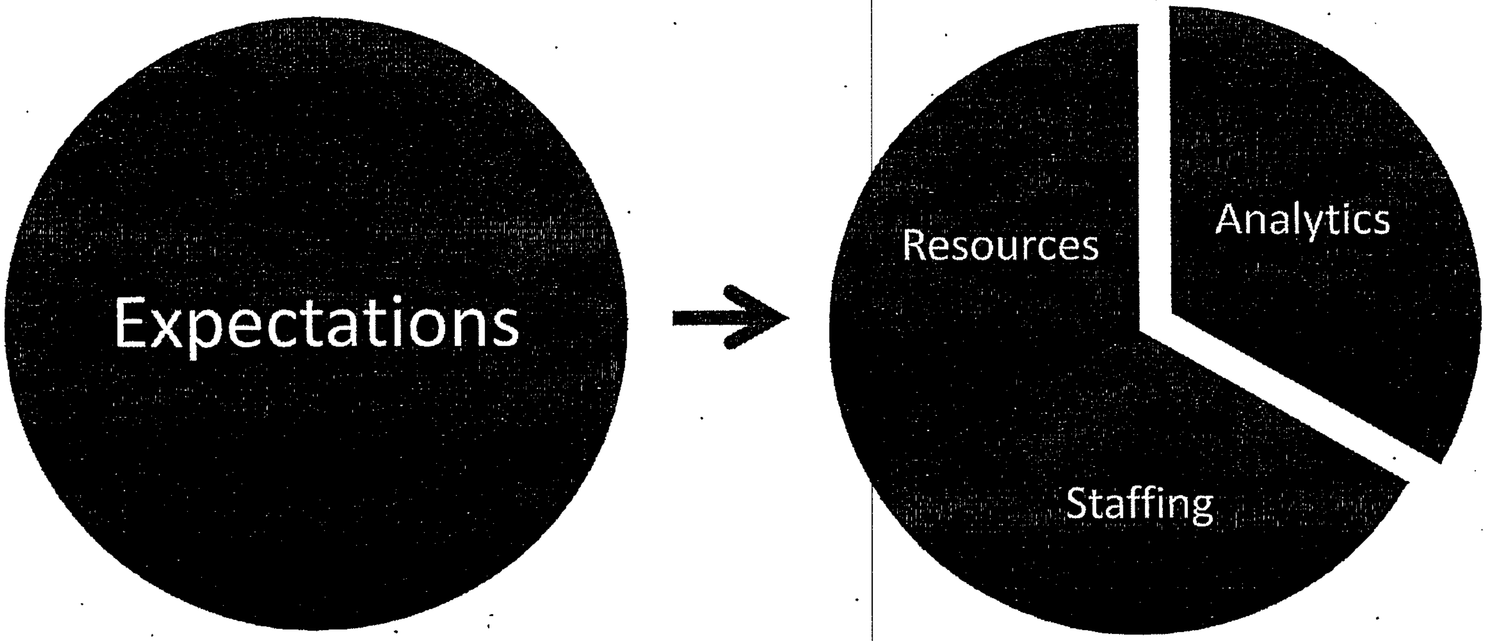
Benefits

- Knowledge Sharing
- Visibility over Controls
- Increased Performance

Agency

- DOL has 72 major information systems (and even more databases)
- Analytics critical to help the 140 auditors tackle these information systems

How Do You Fill The Gap?


Expectations
Resources
Analytics
Staffing