USING ADVANCED ANALYTICS TO PREDICT RISK FOR GRANTS OVERSIGHT

by
Jennifer Wagner

A capstone project submitted to Johns Hopkins University in conformity with the requirements for the degree of Master of Science in Government Analytics

Baltimore, Maryland
August 2019

Abstract

While there is much discussion about applying advanced analytic methods to the auditing and oversight fields, there has been little discussion in the academic literature about using these methods for oversight. The A-133 single audit data is a unique data set that can only be maximized using advanced analytic processes due to its size and current structure. This project applies text mining and predictive modeling techniques to this data set in order to determine both the feasibility and benefits of using these methods for grants management oversight. Using these methods, I was able to identify 12 percent more findings in the audit reports than I was able to identify using established, quantitative methods. This project establishes that advanced analytics methods can be a useful for supporting grant oversight and supporting agencies' efforts to target resources on grant recipients who are highest risk for fraud, waste, and abuse.

Table of Contents

Introduction

In Federal Fiscal Year 2017, the Department of Health and Human Services (HHS) awarded over $100 Billion in grants (excluding Medicare and Medicaid Programming)[1]. HHS has increasingly used grant programs to address a variety of public health needs including the opioid epidemic, emergency preparedness, and natural disaster relief efforts. In order to do accomplish the key missions of the Department, HHS has a network of over 12,000 grant recipient organizations[2] that extend HHS's reach. However, a key risk for leveraging this network is that grant recipient organizations must be monitored to ensure that they are serving the public in good faith and acting as good stewards of federal funds. This level of monitoring is the crux of grants oversight efforts. One of the barriers for grants oversight is that much of the data about grants is in unstructured of semi-structured formats rather than structured, quantitative formats. This poses a challenge to grants management staff and oversight agencies who must collect data from multiple sources in order to identify grant recipient organizations that are not performing or may be at risk for fraud, waste, and abuse. HHS's Office of the Inspector General is very invested in understanding and applying data analytics to this problem[3].

One key data source for grant oversight is the Federal Audit Clearinghouse (FAC)

---

[1] Office of the Inspector General for Health and Human Services. "2018 Top Management & Performance Challenges Facing HHS." Washington, DC., 2018. https://oig.hhs.gov/reports-and-publications/top-challenges/2018/

[2] Office of the Assistant Secretary for Financial Resources. "TAGGS Annual Report 2017". https://taggs.hhs.gov/2017AnnualReport/pdfs/AR2017PDF.pdf

[3] "Combating Healthcare Fraud, Waste and Abuse". Department of Commerce Case Study Series. 2017. NOTE: The author is part of the federal team at HHS OIG supporting this partnership. The Partnership is also supported by NTIS, Excella Consulting, and Elder Research.

which houses the audit data related to OMB Circular A-133, the Single Audit Act of 1984, P.L. 98-502, and the Single Audit Act Amendments of 1996, P.L. 104-156. The laws and regulations require grant recipients who receive more than $750,000 in federal funds to obtain the services of an independent auditor to review the financial documents of the organization and determine if their internal controls are compliant with generally accepted government auditing standards (GAGAS) and are compliance with requirements for federal programs. The product of each A-133 single audit is a written report and a structured data form (form SF-FAC) that summarizes the findings identified in the audit report.

Grant fraud is a broad category of fraud that can impact any type of federal grant program including research grants[4],[5],[6], childcare[7], opioids[8], disaster relief[9], and others[10],[11]. The A-133 single audit reports contain important information for oversight of the grant recipients implement these programs. Because the SF-FAC form is structured and contains a large percentage of the information about risks associated

[4] Kintisch, E. "Scientific Misconduct. Researcher Faces Prison for Fraud in NIH Grant Applications and Papers." Science (New York, N.Y.) 307, no. 5717 (2005): 1851.
[5] Clark, Charles S. "Duke University Pays $112 Million to Settle Federal Grant Fraud Case." Government Executive (2019): N.PAG.
[6] 1. Samuel Reich, E. "Biologist Spared Jail for Grant Fraud." Nature 474, no. 7353 (2011): 552.
[7] Kutz, Gregory D. and Accountability Office US Government. "Head Start: Undercover Testing Finds Fraud and Abuse at Selected Head Start Centers. Testimony before the Committee on Education and Labor, House of Representatives. GAO-10-733T." US Government Accountability Office (2010).
[8] Tovino, Stacey A. "Fraud, Abuse, and Opioids." Kansas Law Review 67, (2019): 901.
[9] Pareja, Veronica. "Weathering the Second Storm: How Bureaucracy and Fraud Curtailed Homeowners' Efforts to Rebuild After Superstorm Sandy." Hofstra Law Review 47, no. 3 (2019): 925.
[10] Burnes, David, Charles R. Henderson Jr, Christine Sheppard, Rebecca Zhao, Karl Pillemer, and Mark S. Lachs. "Prevalence of Financial Fraud and Scams among Older Adults in the United States: A Systematic Review and Meta-Analysis." American Journal of Public Health 107, no. 8 (2017): e13.
[11] Kelly, Christopher and Frans Deklepper. "On the Hunt for Payroll Fraud." Internal Auditor 73, no. 2 (2016):

with compliance with federal program requirements, many oversight bodies have used the SF-FAC to help identify fraud, waste and abuse. However, the instructions for the SF-FAC indicate that auditors should only report on audit findings related to compliance with federal programs but does not provide a way to report on the findings related to GAGAS. The financial statement findings provide important information about whether a grant recipient organization can be a good steward of federal funds. This is an important insight into a dimension of risk that is currently being minimized. Without this information, any risk indicators developed from this data could be considered incomplete or biased.

Previous Efforts in Advanced Analytics for Grants Oversight

The American Recovery and Reinvestment Act (ARRA) of 2009 was enacted to provide federal funds to help stimulate the economy that was going through one of the most significant recessions in modern American history[12]. More than $200 Billion dollars in grant funds were expended during this effort[13]. There was significant concern that putting out large numbers of grants in a short time period would result in fraud, waste and misuse[14], so the act also established an unprecedented monitoring effort to ensure the fidelity and transparency of the funds was maintained.

While individual Departments' Inspectors general were, and continue to be, a key law enforcement agency involved in reducing fraud during the recovery[15], the

[12] American Recovery and Reinvestment Act, 2009. (Pub.L. 111–5) https://www.gpo.gov/fdsys/pkg/BILLS-111hr1enr/pdf/BILLS-111hr1enr.pdf
[13] Czerwinski, Stanley J., Grant Implementation Experiences Offer Lessons for Accountability and Transparency, GAO Reports (2014).
[14] Reich, Eugenie Samuel. "The Specter of Fraud." Scientific American 301, no. 1 (2009): 24.
[15] McNeil, Michele. "Federal Watchdogs Hit Trail in ARRA Oversight Effort." Education Week 30, no. 20 (2011): 14.

Recovery Accountability and Transparency Board (RATB) was the central oversight hub of this initiative. The RATB provided a mechanism for cross-agency collaboration to oversee the grant funds that were distributed as part of the ARRA initiative. The Recovery Operations Center (ROC), a division of the RATB that focused on data analytics to identify funds that were at risk for fraud, waste and abuse, used advanced analytics methods, including text mining the A-133single audits[16],[17]..

On September 30, 2015, the ROC sunset, and the functions of the board were reverted to the individual federal agencies[18]., to identify fraud, waste, or abuse of funds quickly across the federal space. This effort was key to keeping fraud rates low during the recovery[19]. Not only did the ROC meet its mission, but it did so with bipartisan support[20].

The purpose of this methodological study is to continue and update the work of the RATB and the ROC by leveraging the increasing availability of open source, low cost, text mining and analytic capabilities and apply advanced analytic methods to the question of grant oversight. For this project, I will focus only on one HHS grant program in order to demonstrate the feasibility of this approach and the benefits of using text mining for grant oversight. These methods would support federal oversight staff to

[16] Calbom, Linda. "RECOVERY ACT: California's use of Funds and Efforts to Ensure Accountability." GAO Reports (2010):
[17] Kutz, Gregory D."Thousands of Recovery Act Contract and Grant Recipients Owe Hundreds of Millions in Federal Taxes." GAO Reports (2011).
[18] Bagdoyan, Seto J. "Preserving Capabilities of Recovery Operations Center could Help Sustain Oversight of Federal Expenditures." GAO Reports (2015).
[19] Czerwinski, Stanley J. "Federal data Transparency: Opportunities Remain to Incorporate Lessons Learned as Availability of Spending Data Increases." GAO Reports (2013).
[20] Clark, Charles S. "Historic Effort to Track Stimulus Spending Wraps Up." Government Executive (2015).

quickly identify a larger number of grant recipient organizations that might be at risk than using the aggregated SF-FAC data alone.

Literature Review & Theoretical Framework

While government reports and newspapers document failures in grants oversight, providing recommendations to federal and state agencies and reporting on alleged and confirmed fraud, the academic literature determining best practices for oversight is sparse. Primarily the academic literature focuses on discussions by practitioners on the value of applying new methods to traditional oversight roles, such as auditing. The academic literature also reviews the importance of some aspects of grant oversight including internal controls, which is currently included in only one of two A-133 single audit deliverables.

For the last decade, auditors and analysts have been debating the benefits of using analytic methods to maximize the impact of the FAC and the "gold mine" of data it houses[21] However there have been some key challenges with moving forward in this area. Mr. Kull points many of them out including the lack of standards associated with the formatting of the A-133 single audit data and the there is no one agency that is tasked to make investments in maximizing the utility of this data source for risk management.

Use of Advanced Analytics in Auditing and Law Enforcement

While some sectors of the auditing field have been quick to adapt to new technologies, the auditing field is actively debating the role of analytics in conducting audits and providing oversight. Most audit and accounting practices focus on structured

[21] Kull, Joseph L. "Leveraging Technology: Creating an Interactive Single Audit Database." Journal of Government Financial Management 59, no. 2 (2010): 50.

data and using analytic tools to manipulate data structured data, but few have begun working in more advanced analytics, visual, and text analytics[22][23]. Aldhizer and other argue that auditors and accountants should consider expanding the conversation around text analytics.

Auditors have studied the application of advanced analytic methods, such as natural language processing and predictive modeling, and the impact these methods may have in the implementation of audit work[24],[25],[26],[27],[28],[29]. Researchers have found that NLP and artificial intelligence can provide insights into audit work. NLP can be used to extract information from reports and use that information to better understand how internal controls are documented, better understand sentiment associated with positive and negative trends in findings and recommendations, and to predict if documents are potentially fraudulent based on how documents are written. Some researchers believe

[22] "Aldhizer III, George R., "Visual and Text Analytics: The Next Step in Forensic Auditing and Accounting." CPA Journal 87, no. 6 (2017):30.
[23] Zhang, Chanyuan, Jun Dai, and Miklos A. Vasarhelyi. "The Impact of Disruptive Technologies on Accounting and Auditing Education." CPA Journal 88, no. 9 (2018): 20.
[24] Fisher, Ingrid E. "A Perspective on Textual Analysis in Accounting." Journal of Emerging Technologies in Accounting 15, no. 2 (2018): 11.
[25] Fisher, Ingrid E., Margaret R. Garnsey, and Mark E. Hughes. "Natural Language Processing in Accounting, Auditing and Finance: A Synthesis of the Literature with a Roadmap for Future Research." Intelligent Systems in Accounting, Finance & Management 23, no. 3 (2016): 157.
[26] Goel, Sunita and Ozlem Uzuner. "Do Sentiments Matter in Fraud Detection? Estimating Semantic Orientation of Annual Reports." Intelligent Systems in Accounting, Finance & Management 23, no. 3 (2016): 215.
[27] Loughran, Tim and Bill McDonald. "Textual Analysis in Accounting and Finance: A Survey." Journal of Accounting Research (John Wiley & Sons, Inc.) 54, no. 4 (2016): 1187.
[28] El-Haj, Mahmoud, Paul Rayson, Martin Walker, Steven Young, and Vasiliki Simaki. "In Search of Meaning: Lessons, Resources and Next Steps for Computational Analysis of Financial Discourse." Journal of Business Finance & Accounting 46, no. 3 (2019): 265.
[29] Fisher, Ingrid E., Margaret R. Garnsey, Sunita Goel, and Kinsun Tam. "The Role of Text Analytics and Information Retrieval in the Accounting Domain." Journal of Emerging Technologies in Accounting *7, (2010).*

that artificial intelligence may help auditors to complete basic audit tasks[30]. While researchers and practitioners are studying the use of analytics, there has been little discussion about best practices and standards for implementing these methods. Some auditors focus on analytics as a new space for auditors and encourage training in this space[31],[32],[33].

Law enforcement is also increasingly discussing the importance of data analytics for identifying fraud, waste, and abuse. In 2013, the Government Accountability Office (GAO) , the Council of the Inspectors General for Integrity and Efficiency (CIGIE), and the Recovery and Transparency (RAT) Board held a Data Analytics for Oversight and Law Enforcement forum to discuss the use of analytics for the prevention and detection of fraud, waste and abuse[34]. During the forum, participants discussed methods of analytics that had been helpful for oversight, however, they also indicated that there were some limitations to the use of data and analytics including the lack of information about data resources and the lack of technology to facilitate analytics.

Inspectors General are not the only law enforcement agencies investing in data analytics for oversight and enforcement. Other law enforcement agencies are using text

[30] Raschke, Robyn L., Aaron Saiewitz, Pushkin Kachroo, and Jacob B. Lennard. "AI-Enhanced Audit Inquiry: A Research Note." Journal of Emerging Technologies in Accounting 15, no. 2 (2018): 111.
[31] Pan, Gary and Poh-Sun Seow. "Preparing Accounting Graduates for Digital Revolution: A Critical Review of Information Technology Competencies and Skills Development." Journal of Education for Business 91, no. 3 (2016): 166.
[32] Bauer, Andrew M. "Data Analytics: A High-Level Introduction for Accounting Practitioners." Tax Adviser 48, no. 5 (2017):16.
[33] Zhang, Jian. "Incorporating Data Analytics in Accounting." Business Education Forum 73, no. 3 (2019): 14.
[34] Lord, Steve M.,"Data Analytics For Oversight and Law Enforcement." GAO Reports (2013a).

analytics for keyword and key concept searches related to cases[35]. Researchers have shown that data analytics can help to disrupt threats, however, law enforcement agencies must be aware of sensitivity around data they have access to[36]. Moses et al point out that law enforcement agencies using data and analytics should consider promoting transparency regarding that data in order to promote public debate about the use of data and individual rights regarding data privacy.

Advanced analytics are a disruptive new front in oversight and auditing. The literature indicates that some in the community are embracing these new, cutting edge methods while others are more hesitant. It appears there is momentum in the field to increase training in analytics for auditors, but dissemination of best practices is a challenge.

Importance of Internal Controls in Assessing Organizational Risk

Another question for the literature is how important are the GAGAS findings that are currently being left off the SF-FAC form. Federal agencies issue grants in order to implement the missions of those agencies. However, agency missions are not served when grant recipient organizations misuse the funds that are provided to them or when their employees divert those funds. Amiram and colleagues conducted a review of

[35] Chaflawee, Diamond. "The Increasing Importance of Analytics in Law Enforcement." Law Enforcement Technology 42, no. 2 (2015): 36.
[36] Moses, Lyria B. and Louis De Koker. "Open Secrets: Balancing Operational Secrecy and Transparency in the Collection and use of Data by National Security and Law Enforcement Agencies." Melbourne University Law Review41, no. 2 (2017): 530.

accounting literature that underscores the importance of internal controls in understanding financial reporting fraud[37].

Government reports also confirm the importance of monitoring internal controls in cases of grant fraud. In 2017, the GAO reviewed agencies use of the A-133 single audit reports for oversight and indicated that several of the large agencies needed to improve the way they oversee grant recipients and use the information provided in the A-133 single audits to implement that oversight function[38]. The details of fraudulent behavior manifest differently in each grant programs; however, research shows that the lack of internal controls is the overarching theme that bridges federal grant programs.

Internal controls are the policies that organizations put in place to ensure the integrity of the financial and accounting data and to prevent fraud. The lack of internal controls such as a lack of segregation of duties (ensuring checks and balances in approvals and accounting processes), challenges in cash management, and weak accounting practices may be indicators of risk for fraud, waste and abuse[39]. Weak internal controls can also negatively impact nonprofits by impacting the donor support.

GAO has conducted studies examining programs where weak internal controls have been identified and where fraud, waste, and abuse have been confirmed[40]. For

[37] Amiram, Dan, Zahn Bozanic, James D. Cox, Quentin Dupont, Johnathan M. Karpoff, and Richard Sloan. "Financial Reporting Fraud and other Forms of Misconduct: A Multidisciplinary Review of the Literature." Review of Accounting Studies 23, no.2 (2018):732.
[38] Davis, Beryl H. 2017. "Single Audits Improvements Needed in Selected Agencies' Oversight of Federal Awards." GAO Reports (2017).
[39] Petrovits, Christine, Catherine Shakespeare, and Aimee Shih. "The Causes and Consequences of Internal Control Problems in Nonprofit Organizations." Accounting Review 86, no. 1 (2011): 325.
[40] Gootnik, David. "American Samoa: Accountability for Key Federal Grants Needs Improvement: GAO-05-41." GAO Reports (2004).

example, in American Samoa, the A-133 single audit reported risk indicators such as a lack of internal controls, late reporting, and challenges with cash management that were sustained over time led to opportunity for grant fraud that resulted in local prosecution. In addition, the report cites federal agencies slow reaction to these risk factors was also concerning. While fraud was confirmed in American Samoa, there was another study reviewing risk factors in the other freely associated jurisdictions that also had risk factors associated with weak internal controls over compact funds[41].

Research grants at colleges and universities are implemented very differently than the services grant in American Samoa; however, case studies also show that internal controls contribute to fraud in these types of grant programs. Some of these cases are based on false data in applications and publications that were reported by whistle blowers at the University who reported the fraud.

In 1996, Stephen Gordon tried to increase awareness about different types of grant fraud violations that colleges and universities needed to protect against, including false claims, conspiracy, and administrative violations[42]. These standards have not changed significantly over time, though fraud schemes tend to evolve just as quickly as law enforcement methods to prevent and detect fraud[43].

In summary, the literature indicates that the GAGAS related findings contain key information about internal controls that will provide insight into organizational risk.

---

[41]Franzel, Jeanette M., "Compact of Free Association: Single Audits Demonstrate Accountability Problems Over Compact Funds: GAO-04-7." GAO Reports (2003).

[42] Gordon, Steven D. "The Liability of Colleges and Universities for Fraud, Waste, and Abuse in Federally Funded Grants and Projects." New Directions for Higher Education no. 95 (1996): 43.

[43] Luther, Megan., "From Detection to Prevention." IRE Journal 36, no. 1 (2013b): 21.

Reviewing this information would be key to assessing the risk of any grant recipient in order to determine their ability to protect federal funds and the potential risk for fraud, waste and abuse.

Data and Methods

Overview of the A-133 Single Audit Data

The hypothesis at the foundation of this project is that the text in the findings of the A-133 audit reports can be used to predict if an organization has audit findings. The second part of the hypothesis is that extracting the text and building the predictive models will identify more findings than using the SF-FAC data, which will provide more information for grants management and oversight.

For this study, the key unit of analysis is the A-133 audit. There are two measures of the dependent variable, which in this case is the number of findings identified as a result of an A-133 single audit. One of these measures is quantitative and is derived from the SF-FAC, and the other is qualitative and is derived from the A-133 single audit reports themselves. The independent variables are the characteristics of the A-133 single audit that are used as predictors for the predictive models that are applied to the text data. Both data sets were downloaded from the Federal Audit Clearinghouse website.

In FFY 2017, over 36,000 audit reports were submitted to the FAC. There are slight differences in the number of organizations represented in each data set, as all 36,000 data forms are included in the FAC aggregated database, but organizations may choose not to share the reports in the FAC, though they are still public documents. As

indicated earlier, I focused on one HHS Program, the Substance Abuse Prevention and Treatment Block Grant (SABG) due to limitations in computing power, and I also focused on the subrecipient network because the direct recipients are already monitored directly by federal staff. I analyzed each data set separately and compared the two at the end of the project.

Data Quality

One significant concern working with the A-133 single audit data is the quality of the data. The quality of the A-133 single audits, just like other audits, varies because of the guidance provided by OMB for conducting the audit and the methods auditors use to complete the audit[44]. The A-133 single audit data is an interesting data set to test for this bias in it is reviewed regularly for quality and that the Office of Management and Budget takes active steps to improve the quality of the data.

Each year, circular A-133 is augmented by a compliance statement that provides increased guidance for auditors on which audit tests should be performed in relationship with specific federal programs, which provides guidance and is intended to increase the quality of the audit reports[45].

A review of the quality of the A-133 single audits is required every six years in order to help address concerns about data quality[46] [47]. During the last review, GAO examined the A-133 single audit process and determined that federal oversight of the A-

[44] Garven, Sarah A., Amanda W. Beck, and Linda M. Parsons. "Are Audit-Related Factors Associated with Financial Reporting Quality in Nonprofit Organizations?" Auditing: A Journal of Practice & Theory 37, no. 1 (2018): 49.
[45] Brown, Clifford D. and K. Raghunandan. "Audit Quality in Audits of Federal Programs by Non-Federal Auditors." Accounting Horizons 9, (1995).
[46] Ashenfarb, David C. “Identifying Deficiencies in Single Audits.” The CPA Journal. (2018).
[47] Gannon, David J. "The Increasing Importance and Scrutiny of Single Audits." New Jersey CPA (2009): 18.

133 single audit process was not sufficient to ensure the effectiveness of the audit process. There were also concerns that audits take too long to complete and submit to be effective for correcting weaknesses in internal controls, and that grant recipients, especially the smaller ones, are concerned about the expenses and effectiveness of the single audits[48],[49]. Despite the time and expense, auditees can find the process helpful for internal oversight[50].

Independent auditors that conduct the single audits are required to consider fraud and risk factors for fraud during their audit[51]. Auditors report on fraud prevention measures and plans, but also on the risk factors or internal controls, segregation of duties, and reporting mechanisms that create opportunities for fraud to occur. Auditors must also be considerate of their own biases and how those impact audit quality[52][53].

Description of Quantitative Data

Auditors enter data about an audit in data collection form SF-FAC. The FAC aggregates those forms and makes them available as a public download from the FAC website. I downloaded three key tables from the FAC – the general table, the CFDA table, and the findings table. The general table contains summary data from all 36,081 audits submitted in 2017. The CFDA table includes information from the federal awards

[48] Franzel, Jeanette M. "Single Audit: Opportunities Exist to Improve the Single Audit Process and Oversight." GAO Reports. (2009)
[49] Franzel, Jeanette M. "Improvements Needed in Oversight and Accountability Processes." GAO Reports (2011)
[50] Manning, Troy Y. "The Recovery Act: An Auditee's Perspective." Journal of Accountancy 209, no. 6 (2010): 48.
[51] Thomas, C. W. and Juan Alejandro. "Fraud-Related Audit Issues." CPA Journal 71, no. 8 (2001): 62.
[52] Hentati-Klila, Ikhlas, Saida Dammak-Barkallah, and Habib Affes. "Do Auditors' Perceptions Actually Help Fight Against Fraudulent Practices? Evidence from Tunisia." Journal of Management & Governance 21, no. 3 (2017): 715
[53] Pennington, Robin, Jennifer K. Schafer, and Robert Pinsker. "Do Auditor Advocacy Attitudes Impede Audit Objectivity?" Journal of Accounting, Auditing & Finance 32, no. 1 (2017): 136.

table in the SF-FAC and lists all the programs each audit recipient receives. Then the findings table includes information from the findings table on the SF-FAC.

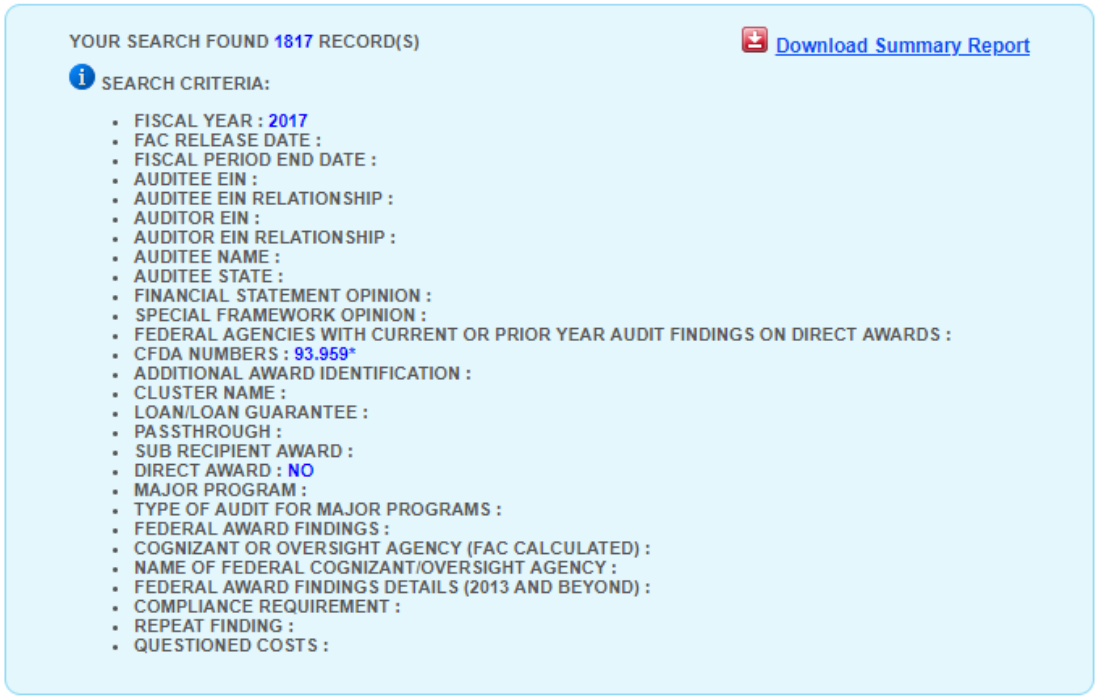
Using SAS Enterprise Guide 9.4, I first filtered the CFDA table to focus on only the audit reports with CFDA 93.959, which represents funding for the Substance Abuse Prevention and Treatment Block Grant (SABG). I then joined those filtered results to the general table using the variable DBKey, which is the unique identifier in the FAC database, to get the list of audit recipients that expended SABG dollars. I also filtered out the direct recipients in order to remove the state-level reports. This resulted in 1,805 distinct audit reports. I then joined the findings table using the variable ELECTAUDITID. This increased the number of reports to 1,824, as some EINs had multiple audit reports associated with them. From these tables, I created an analytic file which contained DBKey, EIN, State, the number of current year findings flag from the general table, the sum of the findings count from the general table, the distinct count of finding identification numbers, and the distinct count of finding reference numbers. Using this method, I identified 406 reports with current year findings from the general table, 91 reports that had audit finding identification and reference numbers from the findings table.

Description of Qualitative Data

Using the federal audit clearinghouse online search tool, I set the filters to for the Prevention and Treatment Substance Abuse Block Grant (CFDA = 93.959) and indicated that we wanted all mentions of that program. I also removed direct recipients (in this case states) because those reports are monitored directly. The search results

indicated that there were 1,817 single audit records related to the Substance Abuse Prevention and Treatment Block Grant (CFDA = 93.959, Direct Recipient = NO, removing states). Of those, 1780 were available for download.

Figure 1: FAC Download parameters for A-133 audit reports



I downloaded all 1,780 PDFS and used those as the data source for the text analytics effort.

Text Extraction

I used the Pdftools, quanteda, and TM packages in R to convert the 1,780 available PDF files to machine readable text. The text from 24 documents was not able to be extracted due to properties of the individual PDF files, including reports that were scanned rather than submitted in machine readable format, irregular fonts, and heavy graphics. This represented 1.3 percent of the total population of documents selected

for download. This first effort turned PDF files into a database that could be labeled, manipulated, and analyzed with 1,756 records.

The method I used for extracting text from the PDFs, cleaning the file, developing the analytic file is described in Appendix A.

Developing the Predictive models

Once it was determined that it was feasible to convert the PDF files into workable text, I was able to create an analytic file that could be used for modeling. After cleaning the data so that only relevant text in the findings section of the report was represented in the analytic file, I labeled each record to determine if the findings section had findings (label = FOUND), did not have findings (label = NONE), or was not readable because of the issues related to the extraction process (label = Unreadable).

Figure 2: Results from Hand Labeling A-133 records

Row Labels	Count of EINS	% of EINS
FOUND	603	34.4%
NONE	1123	63.9%
Unreadable	30	1.7%
Grand Total	**1756**	**100%**

During the labeling process, we noticed that there were several distinct differences between the records that had findings and the records that did not have findings. Key descriptive findings from the text shows that audit reports with findings have a significantly higher average word count (1519.88 words) than audit reports without findings (512.23 words). Audit report with findings also have a broader range of words than audit reports without findings (Figure 3). In addition, audit reports with

findings had findings sections with higher average character count (5,764 characters) than reports without findings (1,426 characters).

Figure3: Average Number of Frequently Used Words in Reports with Findings and Without Findings

Row Labels	Average # of Most Frequent Terms
With Findings	66.4
Without Findings	23.7
Overall Total	**38.6**

Word count, character count, and the presence or absence of frequently used terms became the predictors used to train the predictive models that would be used to detect findings in the larger data set. I removed stop words from the analytic file so that they were not included in the most frequently used words. Figure 4 contains an example from the modeling file and the full list of frequently found words is in Appendix B.

Figure 4 Example of the Modeling File

Filter

ounts	specific	implemented	determine	help	identify	yearend	word_cnt	char_cnt	target
1	0	0	0	0	1	1	568	1765	NONE
0	1	0	0	0	1	0	363	1810	FOUND
1	1	0	0	0	1	1	643	1815	NONE
0	0	0	0	0	0	0	272	639	NONE
1	0	0	0	0	1	1	631	1693	NONE
1	1	0	0	0	0	0	344	854	NONE
1	1	0	0	1	1	1	738	2198	NONE
0	0	0	0	0	0	0	185	384	NONE
1	0	0	0	1	1	1	1032	2179	NONE
0	0	0	0	0	0	0	197	470	NONE
0	0	0	0	0	0	0	77	371	NONE

Once I created the analytic file, I was able to train and test four predictive models to determine which one would be the most effective at finding findings in the text.

Results

Because hand labeling over 36,000 A-133 reports a year is not a practical oversight strategy, I trained four predictive models based on the hand labeled targets, word count, character count, and the presence or absence of frequently found terms. The target of the predictive models is to determine which audit reports have findings based on the predictor characteristics.

Predictive Models

I trained four applicable predictive models to determine which type of model would be the most appropriate for predicting the presence of findings based on the hypothesized predictors, word count, character count, and key words.

Support Vector Machine Model

The Support Vector Machine Model (SVM) uses key data points in the training set to draw a hyperplane between two key classifications – in this case, did the audit report have findings or not. While this type of model is highly respectful of my limited computing power, this model had low accuracy in the test data. The accuracy did not improve as the model was applied to the full data set. Figure 5 provides the confusion matrix for this model and figure six shows key statistics

Figure 5: Support Vector Machine Confusion Matrix

	Test Data		Full Data Set	
Prediction	Found	None	Found	None
Found	104	88	526	480
None	16	136	77	643

Figure 6: Support Vector Machine Key Statistics

	Test Data	Full Data Set
Accuracy	0.6977	0.6773
95% Confidence Interval	(0.6461, 0.7458)	(0.6547, 0.6993)
No Information Rate	0.6512	0.6506
P-Value [Acc > NIR]	0.03865	0.01048
Kappa	0.4159	0.3853
Mcnemar's Test P-value	3.352e-12	< 2e-16
Sensitivity	0.8667	0.8723
Specificity	0.6071	0.5726
Pos Pred Value	0.5417	0.5229
Neg Pred Value	0.8947	0.8931
Prevalence	0.3488	0.3494
Detection Rate	0.3023	0.3048
Detection Prevalence	0.5581	0.5829
Balanced Accuracy	0.7369	0.7224

Predicted Class: FOUND

The Kappa value, the accuracy rate, and the p-value all indicate that this model is fair, but that we might be able to do better.

Random Forest Model

The Random Forest model is a decision-tree based classifier. This model produces multiple decision trees and finds the mean predictors across the various trees. Because of the large number of variables, the random forest model can train using different parts of the dataset and bring that information together into the final model. Figure 7 shows the confusion matrix for the random forest model. Even in the small test set, there were very few misclassified reports. The percentage of misclassified reports in the full data set is also very small.

Figure 7: Random Forest Confusion Matrix

	Test Data		Full Data Set	
Prediction	Found	None	Found	None
Found	120	4	601	4

	Test Data		Full Data Set	
None	0	220	2	1119

Figure 8: Random Forest Key Statistics

	Test Data	Full Data Set
Accuracy	0.9884	0.9965
95% Confidence Interval	(0.9705, 0.9968)	(0.9924, 0.9987)
No Information Rate	0.6512	0.6506
P-Value [Acc > NIR]	<2e-16	<2e-16
Kappa	0.9746	0.9924
Mcnemar's Test P-value	0.1336	0.6831
Sensitivity	1.0000	0.9967
Specificity	0.9821	0.9964
Pos Pred Value	0.9677	0.9934
Neg Pred Value	1.0000	0.9982
Prevalence	0.3488	0.3494
Detection Rate	0.3488	0.3482
Detection Prevalence	0.3605	0.3505
Balanced Accuracy	0.9911	0.9966

Predicted Class: FOUND

From the key statistics, we can see that Kappa Value indicates that this is a very strong model. The accuracy rate is almost 100 percent and improved slightly from the

Neural Network Model

Neural network models consist of interconnected nodes that process information that calculate the impact of predictors on targets. While neural networks can be difficult to explain to non-technical audiences, they can be very accurate at predicting targets.

Figure 9: Neural Network Confusion Matrix

	Test Data		Full Data Set	
Prediction	Found	None	Found	None
Found	118	3	595	3
None	2	221	8	1120

Figure 10: Neural Network Key Statistics

	Test Data	Full Data Set
Accuracy	0.9855	0.9936
95% Confidence Interval	(0.9664, 0.9953)	(0.9886, 0.9968)
No Information Rate	0.6512	0.6506
P-Value [Acc > NIR]	<2e-16	<2e-16
Kappa	0.9681	0.986
Mcnemar's Test P-value	1	0.2278
Sensitivity	0.9833	0.9867
Specificity	0.9866	0.9973
Pos Pred Value	0.9752	0.9950
Neg Pred Value	0.9910	0.9929
Prevalence	0.3488	0.3494
Detection Rate	0.3430	0.3447
Detection Prevalence	0.3517	0.3465
Balanced Accuracy	0.9850	0.9920

Predicted Class: FOUND

From the key statistics we can see that the neural net model is also a strong model. It has similarly strong kappa statistics, accuracy rates, and p-values.

Naive Bayes Model

Naïve Bayes classifiers are often used to categorize text using word frequencies as features. The A-133 modeling file does not use word frequencies, but instead uses other features which likely impacts the quality of Naïve Bayes as a predictive model for this data set, but the analytic file could have been used to create a document term matrix that might be more suited to a Naïve Bayes model.

Figure 11: Naïve Bayes Confusion Matrix

	Test Data		Full Data Set	
Prediction	Found	None	Found	None
Found	30	0	152	1
None	90	224	451	1122

Figure 12: Naïve Bayes Key Statistics

	Test Data	Full Data Set
Accuracy	0.7384	0.7381
95% Confidence Interval	(0.6885, 0.784)	(0.7167, 0.7587)
No Information Rate	0.6512	0.6506
P-Value [Acc > NIR]	0.0003277	3.82e-15
Kappa	0.3027	0.3036
Mcnemar's Test P-value	< 2.2e-16	< 2.2e-16
Sensitivity	0.25000	0.25207
Specificity	1.00000	0.99911
Pos Pred Value	1.00000	0.99346
Neg Pred Value	0.71338	0.71329
Prevalence	0.34884	0.34936
Detection Rate	0.08721	0.08806
Detection Prevalence	0.08721	0.08864
Balanced Accuracy	0.62500	0.62559

Predicted Class: FOUND

Comparison of Hand Labels, Predictive Models, and Quantitative Methods

The true test of the hypothesis comes with the comparison of the hand labeled data, the quantitative methods, and the various predictive models. Figure 13 shows the differences between the different analytic methods. As stated in the data section, using the data from the SF-FAC form, I was able to leverage two different variables that could be used to identify audit findings, 1) the current year findings variable on the general table and 2) the findings reference numbers on the findings table. These two variables describe very different numbers of audit reports with findings. Using the current year findings flag, I was able to determine that about 22 percent of audit reports have findings, while the findings reference number variable only identifies about 5 percent of audit reports as having findings. Despite the slightly smaller population, I was able to identify 603 reports, of 34.3 percent of the reports had findings, which is about 12 percent more findings than the quantitative methods. This is primarily due to the ability to identify the financial findings that are not entered in the structural data.

Figure 13: Comparison of the Predictive Models

	Structured Data (Current Year Findings)	Structured Data (Finding Reference Numbers)	Hand-Labeled Data	Random Forest Model	Neural Net Model	Support Vector Machine (SVM) Model	Naive Bayes Model
No. of Records Tested	1824	1824	1756	1726	1726	1726	1726
No. of Correctly Identified Findings	406	91	603	601	595	526	152
Detection Rate	22%	4.9%	34.3%	34.82%	34.47%	30.48%	8.81%
Accuracy Rate	N/A	N/A	N/A	99.65%	99.36%	67.73%	73.81%
Kappa	N/A	N/A	N/A	0.9924	0.986	0.3853	0.3036

Once I was able to establish the baseline number of findings, I was able to determine the value of the predictive models. Since hand labeling all 36,000 audit reports each year isn't feasible, the predictive models are important to apply this study to the rest of the A-133 data set. The Naïve Bayes model underperformed significantly compared to the other models. This is likely because the analytic file was not developed in an idea format for this type of model, though all the information needed is available. The SVM model also underperformed in comparison to the neural net model and the random forest models. This is likely because the relationship between the predictors may not be linear and I was not able to identify the appropriate kernel for the SVM model.

Both the Random Forest Model and the Neural Network model had more than 99% accuracy rates and strong kappa statistics. While both models are performant, I would prefer to use use the Random Forest model for future work as it predicted the most positive findings, and for this work a false positive is more valuable than a false

negative. The Random Forest model also benefits form being easier to explain than an Neural Network model.

Conclusion

The purpose of this project was twofold, the first objective was to determine if it was feasible to turn the A-133 single audit reports into usable data, and the second was to determine if mining the reports for information could provide more information for oversight than may have been available using traditional methods.

The first part of the study's hypothesis was to determine if it was possible and feasible to turn the audit reports into useful data. With minimal code and enough computing power, transforming the data from PDFs into a database is not only feasible, but could, on its own, provide benefits for federal oversight. Rather than having to read more than 36,000 reports each year, having a data that can be queried would reduce the burden on federal staff responsible for grants oversight. The database could be queried for key words and concepts in order to help federal staff target grant recipient organizations that may be at higher risk or may have findings related to key programs or highly visible activities. Currently this is primarily a time consuming, manual process, but this study shows that it doesn't have to be. This process also helps to identify the GAGAS findings related to organizational internal controls. A review of the literature indicates that these types of findings are important to organizational risk assessments and should not be left out. This process is the only way to access those findings currently.

Because there are over 36,000 A-133 single audits submitted each year, the volume of reports, even with key word searches, is too large to review by hand each year. The second part of the study was aimed at determining if advanced analytics such as predictive models could help identify grant recipient organizations that have risk indicators, in this case findings, in their reports. Identifying these indicators will allow federal staff to target the grant recipient organizations have risk indicators and therefore may require additional monitoring and oversight. Using the text of the A-133 single audit reports, I was able to create two predictive models that identified almost 100 percent of the risk indicators in a population of grant recipients. These models identified about 12 percent more findings than could have been identified thought the SF-FAC data collection forms. This method also reduced the number of complex data table joins that are required for using the quantitative database provided by the FAC.

This study confirms the early efforts of the ROC that text mining the A-133 single audits is both feasible and meaningful but takes the work a further by showing that the text mining effort can be completed using open source tools. The work makes a second contribution with the addition of predictive models to help target grant recipient organizations that have risk indicators, in this case findings, within the large data set. By targeting grant recipients with risk indicators, federal staff will be able to focus resources on potentially at-risk grant recipients. Program staff will benefit form this method by directing monitoring and technical assistance efforts towards organization that are most at risk. Oversight staff will be able to use this method to develop audit, evaluation, and inspection questions.

The study was limited by available computing power. Increases in compute power would have allowed me to convert the entire PDF database to text and apply the predictive model to all 36,000 reports. In addition, increased computing power would have allowed me to skip some of the data preparation steps. In addition, I could have reformatted the analytic file to work better with SVM and Naïve Bayes models in order to optimize those two model types.

Now that it is established that the A-133 single audit report text can be converted into useful data; additional research could use that text to determine additional risk indicators that can be mined from the A-133 reports. In addition, this work could inform the policy surrounding A-133 single audit reports and the method used to collect GAGAS related findings. The literature suggests that these findings should not be excluded from risk assessments, and while the only way to access those findings currently is to use text mining and other advanced analytic methods, OMB could issue changes to include the GAGAS related findings in the SF-FAC data collection form. This would provide a more structured, accessible way to access this information for grants management and oversight and allow oversight agencies to better comply with requirements on pre-award and post-award risk assessments.

References

1. American Recovery and Reinvestment Act, 2009. (Pub.L. 111–5)
2. “Combating Healthcare Fraud, Waste and Abuse”. NTIS Case Study Series. 2017. https://www.ntis.gov/downloads/pdf/NTISCaseStudyHHSOIG.pdf
3. "Fraud in Research Grants Brings a Federal Crackdown." *U.S.News & World Report* 84, (1978)
4. "Public Safety Officers' Benefits Program: Performance Measurement would Strengthen Accountability and Enhance Awareness among Potential Claimants." *GAO Reports* (2009b):
5. Kutz, Gregory D. "Thousands of Recovery Act Contract and Grant Recipients Owe Hundreds of Millions in Federal Taxes." *GAO Reports* (2011):
6. Aldhizer III, George R. "Visual and Text Analytics: The Next Step in Forensic Auditing and Accounting." *CPA Journal* 87, no. 6 (2017): 30.
7. Amiram, Dan, Zahn Bozanic, James D. Cox, Quentin Dupont, Jonathan M. Karpoff, and Richard Sloan. "Financial Reporting Fraud and Other Forms of Misconduct: A Multidisciplinary Review of the Literature." *Review of Accounting Studies* 23, no. 2 (2018): 732.
8. Bagdoyan, Seto J. "Preserving Capabilities of Recovery Operations Center could Help Sustain Oversight of Federal Expenditures." *GAO Reports* (2015):
9. Bauer, Andrew M. "Data Analytics: A High-Level Introduction for Accounting Practitioners." *Tax Adviser* 48, no. 5 (2017):16.
10. Bloodgood, Brianna. "Particularity Discovery in Qui Tam Actions: A Middle Ground Approach to Pleading Fraud in the Health Care Sector." *University of Pennsylvania Law Review* 165, no. 6 (2017): 1435.
11. Brown, Clifford D. and K. Raghunandan. "Audit Quality in Audits of Federal Programs by Non-Federal Auditors." *Accounting Horizons* 9, (1995):
12. Burnes, David, Charles R. Henderson Jr, Christine Sheppard, Rebecca Zhao, Karl Pillemer, and Mark S. Lachs. "Prevalence of Financial Fraud and Scams among Older Adults in the United States: A Systematic Review and Meta-Analysis." *American Journal of Public Health* 107, no. 8 (2017): e13.
13. Calbom, Linda. "Recovery Act: California's use of Funds and Efforts to Ensure Accountability." *GAO Reports* (2010):
14. Chaflawee, Diamond. "The Increasing Importance of Analytics in Law Enforcement." *Law Enforcement Technology* 42, no. 2 (2015):
15. Clark, Charles S. "Duke University Pays $112 Million to Settle Federal Grant Fraud Case." *Government Executive* (2019): N.PAG.
16. Clark, Charles S. "Historic Effort to Track Stimulus Spending Wraps Up." *Government Executive* (2015):
17. Czerwinski, Stanley J. "Federal Data Transparency: Opportunities Remain to Incorporate Lessons Learned as Availability of Spending Data Increases." *GAO Reports* (2013):
18. Czerwinski, Stanley J., "Opportunities Remain to Incorporate Recovery Act Lessons Learned." *GAO Reports* (2013c)

19. Czerwinski, Stanley J. "Grant Implementation Experiences Offer Lessons for Accountability and Transparency." *GAO Reports* (2014):
20. Davis, Beryl H. "Single Audits Improvements Needed in Selected Agencies' Oversight of Federal Awards." *GAO Reports* (2017):
21. El-Haj, Mahmoud, Paul Rayson, Martin Walker, Steven Young, and Vasiliki Simaki. "In Search of Meaning: Lessons, Resources and Next Steps for Computational Analysis of Financial Discourse." *Journal of Business Finance & Accounting* 46, no. 3 (2019): 265.
22. Fisher, Ingrid E. "A Perspective on Textual Analysis in Accounting." *Journal of Emerging Technologies in Accounting* 15, no. 2 (2018): 11.
23. Fisher, Ingrid E., Margaret R. Garnsey, Sunita Goel, and Kinsun Tam. "The Role of Text Analytics and Information Retrieval in the Accounting Domain." *Journal of Emerging Technologies in Accounting* 7, (2010): 1.
24. Fisher, Ingrid E., Margaret R. Garnsey, and Mark E. Hughes. "Natural Language Processing in Accounting, Auditing and Finance: A Synthesis of the Literature with a Roadmap for Future Research." *Intelligent Systems in Accounting, Finance & Management* 23, no. 3 (2016): 157.
25. Franzel, Jeanette M. "Improvements Needed in Oversight and Accountability Processes." *GAO Reports* (2011):
26. Franzel, Jeanette M. "Single Audit: Opportunities Exist to Improve the Single Audit Process and Oversight." *GAO Reports* (2009a):
27. Frenzel, Jeanette M.,"Compact of Free Association: Single Audits Demonstrate Accountability Problems Over Compact Funds: GAO-04-7." *GAO Reports* (2003.)
28. Gannon, David J. "The Increasing Importance and Scrutiny of Single Audits." *New Jersey CPA* (2009):
29. Garven, Sarah A., Amanda W. Beck, and Linda M. Parsons. "Are Audit-Related Factors Associated with Financial Reporting Quality in Nonprofit Organizations?" *Auditing: A Journal of Practice & Theory* 37, no. 1 (2018): 49.
30. Goel, Sunita and Ozlem Uzuner. "Do Sentiments Matter in Fraud Detection? Estimating Semantic Orientation of Annual Reports." *Intelligent Systems in Accounting, Finance & Management* 23, no. 3 (2016): 215.
31. Gordon, Steven D. "The Liability of Colleges and Universities for Fraud, Waste, and Abuse in Federally Funded Grants and Projects." *New Directions for Higher Education.* no. 95 (1996): 43.
32. Gootnik, Davd.,"American Samoa: Accountability for Key Federal Grants Needs Improvement: GAO-05-41." *GAO Reports* (2004).
33. Hentati-Klila, Ikhlas, Saida Dammak-Barkallah, and Habib Affes. "Do Auditors' Perceptions Actually Help Fight Against Fraudulent Practices? Evidence from Tunisia." *Journal of Management & Governance* 21, no. 3 (2017): 715.
34. Kelly, Christopher and Frans Deklepper. "On the Hunt for Payroll Fraud." *Internal Auditor* 73, no. 2 (2016):
35. Kintisch, E. "Scientific Misconduct. Researcher Faces Prison for Fraud in NIH Grant Applications and Papers." *Science (New York, N.Y.)* 307, no. 5717 (2005): 1851.

36. Kull, Joseph L. "Leveraging Technology: Creating an Interactive Single Audit Database." *Journal of Government Financial Management* 59, no. 2 (2010): 50.
37. Kutz, Gregory D. and Accountability Office US Government. "Head Start: Undercover Testing Finds Fraud and Abuse at Selected Head Start Centers. Testimony before the Committee on Education and Labor, House of Representatives. GAO-10-733T." *US Government Accountability Office* (2010).
38. Lord, Steve M., "Data Analytics." *GAO Reports* (2013a):
39. Loughran, Tim and Bill McDonald. "Textual Analysis in Accounting and Finance: A Survey." *Journal of Accounting Research (John Wiley & Sons, Inc.)* 54, no. 4 (2016): 1187. doi:10.1111/1475-679X.12123.
40. Luther, Megan., "From Detection to Prevention." *IRE Journal* 36, no. 1 (2013b).
41. Manning, Troy Y. "The Recovery Act: An Auditee's Perspective." *Journal of Accountancy* 209, no. 6 (2010):
42. McNeil, Michele. "Federal Watchdogs Hit Trail in ARRA Oversight Effort." *Education Week* 30, no. 20 (2011): 14.
43. Moses, Lyria Bennett and Louis De Koker. "Open Secrets: Balancing Operational Secrecy and Transparency in the Collection and use of Data by National Security and Law Enforcement Agencies." *Melbourne University Law Review* 41, no. 2 (2017): 530.
44. Office of the Assistant Secretary for Financial Resources. "TAGGS Annual Report 2017". Published online at https://taggs.hhs.gov/2017AnnualReport/pdfs/AR2017PDF.pdf
45. Office of the Inspector General for Health and Human Services. "2018 Top Management & Performance Challenges Facing HHS." Washington, DC., 2018.
46. Pan, Gary and Poh-Sun Seow. "Preparing Accounting Graduates for Digital Revolution: A Critical Review of Information Technology Competencies and Skills Development." *Journal of Education for Business* 91, no. 3 (2016): 166.
47. Pareja, Veronica. "Weathering the Second Storm: How Bureaucracy and Fraud Curtailed Homeowners' Efforts to Rebuild After Superstorm Sandy." *Hofstra Law Review*47, no. 3 (2019): 925.
48. Pennington, Robin, Jennifer K. Schafer, and Robert Pinsker. "Do Auditor Advocacy Attitudes Impede Audit Objectivity?" *Journal of Accounting, Auditing & Finance* 32, no. 1 (2017): 136.
49. Petrovits, Christine, Catherine Shakespeare, and Aimee Shih. "The Causes and Consequences of Internal Control Problems in Nonprofit Organizations." *Accounting Review* 86, no. 1 (2011): 325.
50. Raschke, Robyn L., Aaron Saiewitz, Pushkin Kachroo, and Jacob B. Lennard. "AI-Enhanced Audit Inquiry: A Research Note." *Journal of Emerging Technologies in Accounting* 15, no. 2 (2018): 111.
51. Reich, Eugenie Samuel. "The Specter of Fraud." *Scientific American* 301, no. 1 (2009): 24. doi:10.1038/scientificamerican0709-24.
52. Samuel Reich, E. "Biologist Spared Jail for Grant Fraud." *Nature* 474, no. 7353 (2011): 552.

53. Thomas, C. W. and Juan Alejandro. "Fraud-Related Audit Issues." *CPA Journal* 71, no. 8 (2001): 62
54. Tovino, Stacey A. "Fraud, Abuse, and Opioids." *Kansas Law Review* 67, (2019): 901.
55. Zhang, Chanyuan, Jun Dai, and Miklos A. Vasarhelyi. "The Impact of Disruptive Technologies on Accounting and Auditing Education." *CPA Journal* 88, no. 9 (2018): 20.
56. Zhang, Jian. "Incorporating Data Analytics in Accounting." *Business Education Forum* 73, no. 3 (2019): 14.

Appendices

Appendix A — Data Processing

I used R packages pdftools, tm, and quanteda to extract the text of the PDF and turn it into plain text. First, I turned the filenames into a vector, and then I read the files into R and develop a corpus. Once the corpus has been created, I turn the corpus into a data frame. The data frame can then be written as data frame.

Figure A1: Code Snip for Text Extraction

```
library(quanteda)
library(tm)
library(pdftools)
files <- list.files(pattern = "pdf$")
auditscorp <- Corpus(URISource(files),
        readerControl = list(reader = readPDF))
mycorpus <- corpus(auditscorp)
auditframe<-data.frame(text=unlist(sapply(auditscorp, `[`, "content")),
        stringsAsFactors=T)
write.csv(auditframe, file = "MyData.csv")
```

I broke the PDFs into batches of 100 in order to stay within the limits of my computing power. As seen in Figure A2, each page in the PDF has become a row in a data set.

Figure A2: The first 20 pages of an extracted single audit report

		text
2	16602720171.pdf.content1	CROOK COUNTY, OREGON FINANCIAL REPORTFOR THE YEAR ENDED JUNE 30, 2017 12700 SW 72nd Ave. Tigard, OR 97223
3	16602720171.pdf.content2	CROOK COUNTY, OREGONCOMPREHENSIVE ANNUAL FINANCIAL REPORT For the Year Ended June 30, 2017
4	16602720171.pdf.content3	CROOK COUNTY, OREGON TABLE OF CONTENTS PAGE
5	16602720171.pdf.content4	CROOK COUNTY, OREGON TABLE OF CONTENTS (CONTINUED) PAGE
6	16602720171.pdf.content5	CROOK COUNTY, OREGON TABLE OF CONTENTS (CONTINUED) PAGE
7	16602720171.pdf.content6	CROOK COUNTY, OREGONINTRODUCTORY SECTION
8	16602720171.pdf.content7	CROOK COUNTY, OREGON Board of CommissionersName and Address Term ExpiresSeth
9	16602720171.pdf.content8	CROOK COUNTY, OREGON FINANCIAL SECTION
10	16602720171.pdf.content9	PAULY, ROGERS, AND CO., P.C. 12700 SW 72nd Ave. Tigard, OR 97223 (503) 620-2632
11	16602720171.pdf.content10	OpinionsIn our opinion, the financial statements referred to above present fairly, in all material respects, the respective financialposition of the
12	16602720171.pdf.content11	Other InformationThe Council members, as listed in the table of contents have not been subjected to the auditing proceduresapplied in the audi
13	16602720171.pdf.content12	CROOK COUNTY, OREGON MANAGEMENT'S DISCUSSION AND ANALYSIS JUNE 30, 2017_
14	16602720171.pdf.content13	CROOK COUNTY, OREGON MANAGEMENT'S DISCUSSION AND ANALYSIS JUNE 30, 2017__
15	16602720171.pdf.content14	CROOK COUNTY, OREGON MANAGEMENT'S DISCUSSION AND ANALYSIS JUNE 30, 2017__
16	16602720171.pdf.content15	CROOK COUNTY, OREGON MANAGEMENT'S DISCUSSION AND ANALYSIS JUNE 30, 20
17	16602720171.pdf.content16	CROOK COUNTY, OREGON MANAGEMENT'S DISCUSSION AND ANALYSIS JU
18	16602720171.pdf.content17	CROOK COUNTY, OREGON MANAGEMENT'S DISCUSSION AND ANALYSIS JUNE 30, 2017____
19	16602720171.pdf.content18	CROOK COUNTY, OREGON MANAGEMENT'S DISCUSSION AND ANALYSIS JUNE 30, 2017_
20	16602720171.pdf.content19	CROOK COUNTY, OREGONNOTES TO BASIC FINANCIAL STATEMENTS
21	16602720171.pdf.content20	CROOK COUNTY, OREGON STATEMENT OF NET POSITION

Unfortunately, due to computing power limitations, I found I needed to cut down the dataset to focus on the findings pages in order to complete my analysis. I cleaned the data by tagging and then removing non-finding pages from the dataset and aggregating rows that contained findings for each audit report. I created a flag variable that indicated whether or not the organization had findings (Found = the report did have findings, None = the report did not have findings, unreadable = the report could not be read so the presence of findings could not be determined). At the end of this process, each organization had one row of findings information and a label. Figure 3A displays the first 11 rows of the final text mined data set, which includes the findings text (variable = text). The current model includes 1756 labeled records (see Figure 3A).

Figure A3: Examples of Labeled Findings Pages

	text	Flag
1	UNITED COMMUNITY ACTION PROGRAM, INC. AND SUBSIDIARY Pawnee, Oklahoma SCH...	NONE
2	12 &12, Inc. and Affiliate Schedule of Findings and Questioned Costs - Continued June...	FOUND
3	LATINO COMMUNITY DEVELOPMENT AGENCY, INC. SCHEDULE OF FINDINGS AND QUESTI...	NONE
4	The Oklahoma Mental Health Council d/b/a Red Rock Behavioral Health Services Sched...	NONE
5	Family Service Association of San Antonio, Inc. Federal Awards – Schedule of Findings ...	NONE
6	Career and Recovery Resources, Inc. Schedule of Findings and Questioned Costs (Conti...	NONE
7	The Council on Recovery (Parent-Only) Schedule of Findings and Questioned Costs for ...	NONE
8	SAN ANTONIO COUNCIL ON ALCOHOL AND DRUG ABUSE Schedule of Findings and Qu...	NONE
9	FAMILY SERVICES OF SOUTHEAST TEXAS SCHEDULE OF FINDINGS AND QUESTIONED CO...	NONE
10	SAN ANTONIO LIFETIME RECOVERY, INC. dba LIFETIME RECOVERY SCHEDULE OF FINDIN...	NONE
11	INTEGRAL CARE SCHEDULE OF FINDINGS AND QUESTIONED COSTS - CONTINUED For th...	NONE

Through this process, I was able to support the first half of the hypothesis, establishing that it is possible to convert the A-133 single audit reports to usable text data for analysis. The next steps were to create the analytic file

Appendix B — Frequently Used Terms Included in Predictive Modeling

Term	Number of Reports containing Key Term
Requires	1358
Certain	1171
Errors	1158
Provided	1123
Implemented	1101
Repeat	1087
Whether	1055
Projects	1011
Policy	997
Subrecipient	967
Approved	959
Maintain	924
Journal	921
However	899
Reconciliation	882
Numbers	877
county's	797
Determine	779
Auditor	738
Contract	735
Views	732
Service	709
Amount	699
Personnel	697
Government	694
Reporting	691
Revenue	673
Development	638
Continue	631
School	622
Result	592
Testing	553
Additional	546
Block	544
Generally	536

Term	Number of Reports containing Key Term
Special	497
Public	458
Complete	452
Account	451
Ensure	443
Monthly	441
Lack	438
Prevention	427
Bank	424
Based	417
Human	411
Balance	399
Employee	398
Must	396
Received	385
Schedule	382
Activity	381
Part	378
Balances	372
Major	366
auditor's	363
Opinion	363
Regulations	363
Segregation	362
Implement	357
Preparation	350
Page	347
Accurate	346
Employees	339
Continued	338
Process	330
Expenses	324
Agency	311
Awards	306
Transactions	303

Term	Number of Reports containing Key Term
Date	301
Unmodified	296
Criteria	294
Monitoring	292
Assistance	290
Receivable	290
Costs	289
Resources	289
Reconciliations	288
Support	286
Year	282
Relating	281
Access	279
Findings	271
Threshold	269
Considered	267
Accounting	261
Duties	258
Deficiencies	256
June	251
September	250
Statement	248
Officials	248
Auditing	247
Following	247
Funds	245
Name	243
Care	241
Proper	239
Uniform	236
Information	236
Grants	233
Substance	233
Effective	233
Included	229
Cost	222
auditor's	220

Term	Number of Reports containing Key Term
Payroll	220
Federal	219
Board	219
Eligibility	216
District	215
Student	214
Type	212
Office	210
Recorded	210
Addition	209
Cfda	206
Management	205
Basis	204
Education	203
Auditee	200
Deficiency	200
Number	198
Annual	198
Guidance	197
Abuse	195
auditors'	194
Made	194
Prepare	194
Procedures	191
Report	190
Cause	189
Corrective	188
Provide	180
Fund	179
Planned	177
Audit	174
Used	174
Lowrisk	174
Section	173
Documentation	173
Matters	172
Significant	167

Term	Number of Reports containing Key Term
Financial	166
Programs	166
Award	166
Community	166
Recommendation	165
Distinguish	165
Finding	164
Also	164
Entity	164
Expenditures	163
Issued	162
Summary	162
Organization	160
Review	159
Applicable	159
None	156
Weaknesses	154
Noncompliance	150
Statements	148
Deficiencies	148
Recommend	148
Activities	147
Cash	143
State	142
Will	142
States	141
Weakness	140

Term	Number of Reports containing Key Term
Requirement	140
Amounts	136
Action	134
Health	132
Policies	128
Prior	127
requirements	126
december	125
Timely	124
Gaap	124
required	123
Services	118
Results	117
instances	116
accordance	109
Grant	106
questioned	103
Noted	102
Control	98
Place	96
principles	96
properly	95
Tested	95
including	95
Current	93
Effect	92
Entries	87

Appendix C — Bio Sketch

Jennifer Wagner currently serves as the Senior Program Analyst for Grants at the Department of Health and Human Services, Office of the Inspector General, Division of Data Analytics. Prior to joining the OIG, Jennifer served as a Senior Public Health Advisor and National Synar Coordinator at the Center for Substance Abuse Prevention, leading a statistically focused tobacco control and prevention program that aimed to reduce the youth tobacco access.

During her time at HHS OIG, Ms. Wagner has served as a federal lead and product owner for a joint venture project between the OIG, the Department of Commerce National Technical Information Service, Excella Consulting, and Elder Research to apply advanced analytics to grants data, focusing on text analytics, data fusion, and anomaly detection.

Ms. Wagner was born in the Washington, D.C. area (December 31, 1980) and her academic life has focused on schools in the area. Ms. Wagner currently holds a Bachelor's Degree in Psychology from the University of Maryland (2002) and a Graduate Certificate in Sampling and Data Analysis from the George Washington University. Ms. Wagner is currently pursuing a Master's Degree in Government Analytics from Johns Hopkins University. Her elective classwork has focused on risk management, text analytics, and public policy.