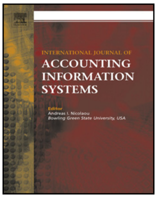
International Journal of Accounting Information Systems xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

International Journal of Accounting Information Systems



A taxonomy to guide research on the application of data mining to fraud detection in financial statement audits☆

Glen L. Gray [a,*], Roger S. Debreceny [b]

[a] David Nazarian College of Business and Economics, California State University at Northridge, United States
[b] School of Accountancy, Shidler College of Business, University of Hawai'i at Mānoa, United States

ARTICLE INFO

Article history:
Received 1 June 2013
Received in revised form 26 December 2013
Accepted 15 May 2014
Available online xxxx

Keywords:
Auditing
Fraud
Data mining

ABSTRACT

This paper explores the application of data mining techniques to fraud detection in the audit of financial statements and proposes a taxonomy to support and guide future research. Currently, the application of data mining to auditing is at an early stage of development and researchers take a scatter-shot approach, investigating patterns in financial statement disclosures, text in annual reports and MD&As, and the nature of journal entries without appropriate guidance being drawn from lessons in known fraud patterns. To develop structure to research in data mining, we create a taxonomy that combines research on patterns of observed fraud schemes with an appreciation of areas that benefit from productive application of data mining. We encapsulate traditional views of data mining that operates primarily on quantitative data, such as financial statement and journal entry data. In addition, we draw on other forms of data mining, notably text and email mining.

1. Introduction

This study explores the targeted application of data mining techniques to fraud detection as a core component of financial statement audits.[1] Data mining refers to the extraction of knowledge from large

☆ The authors thank Sev Grabski and Eckhardt Kriel for their valuable comments and suggestions, as discussants of an earlier version of this paper presented at the 2013 University of Waterloo Research Symposium on Information Integrity & Information Systems Assurance.

* Corresponding author.

E-mail addresses: glen.gray@csun.edu (G.L. Gray), roger@debreceny.com (R.S. Debreceny).

[1] While the paper focuses on the detection of fraud within the financial statement audit conducted by external auditors, most of the key messages are also relevant for internal auditors.

http://dx.doi.org/10.1016/j.accinf.2014.05.006
1467-0895/© 2014 Elsevier Inc. All rights reserved.

volumes of data (Han and Kamber, 2006. 5). Data mining involves acquisition, loading and integration of data; application of specialist data mining tools and, finally, human interpretation of the discovered meaning.[2] The decision to incorporate data mining into financial audits is both a firm-level decision for accounting firms and an engagement-level decision. Firm-level decisions preclude engagement-level decisions in that if firm management does not see a beneficial reason to invest resources in software, infrastructure, training, and staffing then data mining will likely not be a cost-effective option for engagement teams. Larger accounting firms and some specialist providers offer a variety of data mining services. Currently, data mining is used in specialized audits (e.g., fraud audits or forensic audits) by expert staff in the professional services firms; however, data mining is seldom used in financial statement audits. When used, it is for identified high-risk clients by the firm's data mining specialists. The majority of this paper is focused on engagement-level data mining activities; however, we revisit firm-level issues and research opportunities in the concluding parts of the paper.

Applying data mining to fraud detection as part of a routine financial audit can be challenging and, as we will explain, data mining should be used when the potential payoff is high. In general, when it comes to fraud detection for a given audit client, the audit team would make three major decisions: (1) What specific types of fraud (e.g., revenue recognition, understated liabilities, etc.) should be included in the audit plan for a particular client? (2) What sources of data (e.g., journal entries, emails, etc.) would provide evidence of each type of fraud? (3) Which data mining technique(s) (e.g., directed or undirected techniques) would be the most effective for finding potential evidence of fraud in the selected data? Developing answers for each of these questions is significant individually, but, in combination, answering these questions is challenging. These challenges may encourage the audit team to continue to use traditional – but less diagnostic – analytical and substantive procedures. However, as we will discuss in this paper, each of the populations for each of these three questions can be intelligently reduced so that the application of data mining to fraud detection becomes more manageable and will have a higher potential for a successful payoff. We also recognize that data mining techniques and associated software can have a steep learning curve. Further, if used improperly, data mining can produce many false positives and spurious patterns that will require auditors to expend time to subsequently investigate. The primary contribution of this paper is in identifying specific fraud and evidence combinations where data mining would be the most effective in traditional financial audits as well as those combinations where data mining would be least effective. Identifying the more effective use of data mining could encourage auditors to include data mining as a regular element of their audit programs. Future researchers can build on our exploratory findings to further refine the application of data mining in financial statement audits.

Specifically, the paper proposes a taxonomy that includes three components, namely, account schemes and evidence schemes (components of fraud schemes as defined by Gao and Srivastava (2011)),[3] and data mining functionality to identify the most effective combinations of those three components. Data mining has the potential to enhance the efficiency and effectiveness of the audit. Productive data mining toolsets are now more widely available and auditors have access to a cornucopia of audit-relevant data both internal and external to the client organization. Internal data can include financial data, non-financial data, and email archives. Externally, a vast array of quantitative and qualitative information on organizations is now available on the Internet and in commercial financial and textual databases. These include news reports, blog postings, Facebook postings, and Twitter feeds. For public companies, regulatory filings such as the filings made on the U.S. Securities and Exchange Commission's (SEC's) EDGAR database in XBRL format are available.

There has been an increased interest in data mining for fraud detection in the regulatory and professional domain. For example, the SEC has developed an "Accounting Quality Model," (commonly referred to as the Robocop) designed to identify anomalous financial statement filings to the Commission. The tool mines the XBRL data repository along with other datasets (Lewis, 2012; Rohman and Berg, 2013). A recent report by the Financial Executives International (FEI) also points to a variety of tools to data mine the XBRL filings to the SEC (FERF, 2013). The Advisory Committee on the Auditing Profession (ACAP) to the U. S. Treasury recommended

[2] CRISP-DM sets out a methodology that sets out a process model that sees data mining flowing from the development of a business context for the mining; understanding data sources; data preparation; modeling; evaluation and deployment (Chapman et al., 2000).

[3] Gao and Srivastava (2011) divided fraud schemes into two components: account schemes reflecting the accounts impacted by the fraud (e.g., fictitious revenue) and evidence schemes reflecting how the fraudster implemented the fraud (e.g., fake documents).

that the Public Company Accounting Oversight Board (PCAOB) establish a fraud center that would, in part, facilitate "sharing of practices, and data and innovation in fraud prevention and detection methodologies and technologies" (ACAP, 2008, VII:1). To assist auditors with data availability, the American Institute of Certified Public Accountants (AICPA) is promoting its Audit Data Standards (ADS) (Zhang et al., 2012; Titera, 2013).[4] The objective of ADS is to facilitate ready extraction of entity-level transactional data for audit interrogation. This standardization of data elements that can be requested from different audit clients will increase the economy of scale of using data mining software, which in turn can help justify the cost of learning data mining software and developing a portfolio of common data mining routines that can be applied to multiple clients.

In the academic arena, there has been a small but growing literature that applies data mining techniques to auditing in general and fraud detection in particular (Jans et al., 2010; Perols, 2011; Ravisankar et al., 2011; Alden et al., 2012). This comes after a hiatus of more than a decade and results in part from the increased availability of datasets, such as the EDGAR database, and from improved data mining tools. Despite this recent spurt in research, much of the existing research does not achieve the greatest possible return on research investment. Researchers take a scatter-shot approach, investigating patterns in financial statement disclosures, text in annual reports and MD&As, and the nature of journal entries without appropriate guidance drawn from lessons in known fraud patterns. The taxonomy we present in this paper combines research on where patterns of fraud are most frequently observed with an appreciation of those areas where data mining can be productively applied in the audit process.

The organization of the remainder of the paper is as follows. Section 2 explores the nature of data mining and recent research on the application of data mining to the financial statement audit. Section 3 discusses the application of data mining to each phase of the audit to illustrate that data mining can have wide applicability to an audit. Section 4 describes the development of our taxonomy of fraud schemes (account schemes and evidence schemes) and data mining techniques. Section 5 provides concluding comments and sets out future research directions.

2. Literature review

In this section, we introduce and compare the key techniques of data analysis, data extraction, and data mining that auditors can apply in an audit. We provide some examples from the audit domain. We then introduce recent research that applies a variety of data mining techniques to the audit and fraud context.

2.1. Data analysis and extraction versus data mining

Practicing auditors are rather imprecise when considering the nature of data mining, confounding data extraction, data analysis, and data mining. Fig. 1 illustrates the conceptual relationship between data extraction, data analysis, and data mining. As the auditor moves from data extraction to data analysis to data mining, the software becomes more sophisticated in terms of functionality and provides a greater amount of diagnostic and predictive power. The triangle in Fig. 1 indicates the relative frequency of the use of these different techniques by auditors based on our review of professional literature. The following paragraphs provide an overview of the different categories of data examination tools included in Fig. 1.

2.1.1. Data extraction and query

For financial statement audits, Excel, ACL, and CaseWare IDEA are illustrative of the most frequently used tools to examine client data. These tools are essentially used as data query and extraction tools that perform data analysis with a variety of descriptive statistics and a limited range of statistical techniques. ACL and CaseWare IDEA also include tools to prepare samples of populations for further audit procedures. These and similar tools traditionally have been called *computer assisted audit tools and techniques* (CAATTs) (Coderre, 2009). The auditor subsequently performs any actual *analysis*, in the investigative sense of the word, of the data. With these tools, the auditor has already made the decision as to what data he/she wants to extract. For example, if a company has an internal control that requires the controller to sign checks to vendors greater than $5000, the auditor will use one of these tools to extract check numbers and related data

[4] The Audit Data Standards are available at: http://www.aicpa.org/InterestAreas/FRC/AssuranceAdvisoryServices/Pages/AuditDataStandardWorkingGroup.aspx.

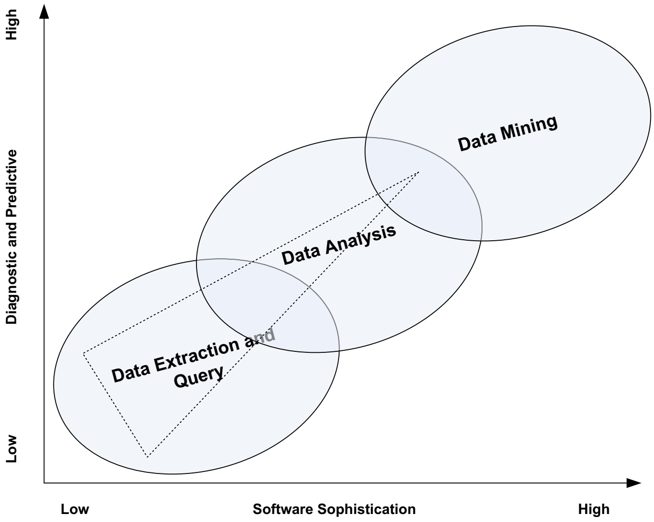


Fig. 1. Relationship of various data examination tools.

for all checks between, say, $4900 and $4999. This extraction is used because the auditor believes that there is a risk that a fraudster will write fraudulent checks just below the $5000 threshold to avoid review by the controller. The auditor will subsequently locate and investigate those checks and supporting documentation to determine if any checks violate the control or are fraudulent.

We do not discount data extraction tools for enhancing audit efficiency. Indeed, data extraction software provides powerful audit tools. Queries that fully exploit the functions built into ACL and CaseWare IDEA can be effective in identifying suspicious data patterns in client datasets. These tools, however, are essentially in use as data query and extraction tools to reduce the client's full population of data to a smaller, more-manageable population of data. After extraction, the auditor performs an analysis or investigation (frequently manually) of the extracted sample from the population. It is up to the auditor's training and experience to identify any suspicious extracted items that require additional investigation.

2.1.2. Data analysis

Data analysis toolsets available to auditors provide a range of analytical techniques from simple to relatively sophisticated. For example, at the low end, data analysis includes basic descriptive statistics such as counts, minimums, maximums, dispersion, and means. Low-end analysis also may include ratio analysis. At the higher end, data analysis can include inferential statistics such as univariate and multivariate regression, canonical correlation analysis and other outputs of statistical software. Like data extraction, the principal limiting factor on the use of data analysis tools comes from the collective knowledge of the auditors. For large organizations, it is easy for auditors to be overwhelmed by significant volumes of data (Eppler and Mengis, 2004).

2.1.3. Data mining

As illustrated in Fig. 1, data analysis and data mining definitions and examples overlap. A term frequently associated with data mining is *knowledge discovery* (Delen and Al-Hawamdeh, 2009). Data mining is about discovering patterns, rules, or models based on one or more populations of data. The results (patterns, rules, or models) are used to predict future outcomes. Actual outcomes that fall outside the predicted ranges or patterns are red flags that the auditor should investigate.

Data mining techniques can be divided into two broad categories—*directed* (or top-down approach) and *undirected* (or bottom-up approach). With directed data mining, there is identification of a specific targeted variable of interest. Data mining technology finds relationships between that variable and a selected population of other variables. With undirected data mining, there is no specific targeted variable (dependent variable); instead the objective is to find any relationship between any variables in a population of data.

Another way to characterize these two broad categories is to say that the top-down approaches test specific hypotheses and the bottom-up approach generates new hypotheses.

Examples of directed data mining include:

Classification*:* Developing a set of rules based on existing data (training sets) to, in turn, correctly classify a new object into relatively discrete classifications. For example, classification techniques may categorize an audit client as low, medium, or high risk based on a broad set of independent variables such as financial statement variables, movements in share prices and volumes. The quality of classification data mining techniques can be measured by accuracy of classification, robustness and ability to interpret the classification outcomes (Han and Kamber, 2006). Classification techniques include decision tree induction (Siciliano and Conversano, 2006), Bayesian belief networks (Bashir et al., 2006), neural networks (Zhang, 2000; Smith, 2006), support vector machines (Steinwart and Christmann, 2008) and genetic algorithms (Cox, 2005).

Estimation*:* Whereas classification is in use for relatively discrete outcomes, estimation techniques use continuous variables. For example, rather than categorizing risk as low/medium/high, estimating techniques would generate a risk score (e.g., from 0 to 10). These estimations can subsequently classify an object, through a technique similar to logistic regression. For example, a high-risk client has a risk score over 7.5. Estimations could be used to rank order a population of objects and then to develop a threshold score. For example, all audit clients above the threshold score of 8.0 must have a forensic auditor assigned to the audit.

Prediction: Classification and estimation are data mining techniques that are in use to reveal characteristics of previously identified variables in a dataset. Data mining authors often differentiate "prediction" techniques from classification and estimation (Han and Kamber, 2006). When employing classification and estimation techniques we are not seeking to test the correctness of the outcome. Conversely, the objective of prediction data mining is to test systematically hypotheses as well as to find outliers (red flags) where a factor falls outside a predicted range. Like classification techniques, these prediction techniques use training sets for initial model building as well holdout samples or bootstrapping methods on final datasets.

Examples of undirected data mining include:

Affinity grouping: The objective of affinity grouping is to find relationships between variables in the dataset. There is no specific dependent and independent variable. The data analyst is looking for previously unknown connections between variables. In auditing, data mining techniques could be in use to find relationships in the population of journal entries provided to the auditor. An example of such an affinity grouping could be journal entries flowing between property, plant and equipment and plant maintenance. The auditor would investigate those journal entries that fall outside the expected affinity grouping. Relations can also build on data that follow a given sequence such as a temporal pattern. Sequential pattern mining looks for relationships in terms of the order in which related events occur. This class of data mining has clear implications for identifying patterns of, for example, a series of inappropriate, but relatively small, transfers between accounts.

Clustering*:* The objective is to cluster variables into subgroups based on relationships in the dataset. The objectives of directed classification, discussed above, and undirected clustering are similar. The implementation and outcomes are quite different, however. With directed classification, the subgroups have predefined classes based on the pre-selected variable(s). Conversely, with undirected clustering the subgroups are not pre-selected. The auditor must interpret the underlying meaning of the sub-groupings in clustering using their knowledge of the accounting model and the business models employed by the client. In a set of journal entries, clustering will identify expected clusters of journal entries. An example of such a cluster would be entries involving sales revenue, receivables, product costs, and inventory. Equally, clusters of journal entries may be identified that do not fall within an established or known pattern. Clustering techniques include hierarchical methods (Berkhin, 2006), self-organizing maps based on neural networks (Kohonen, 2000) and density-based techniques (Berkhin, 2006). An important subset of clustering for fraud detection is outlier mining. Techniques to identify outliers include distance- and density-based algorithms (Han and Kamber, 2006).

Description and visualization*:* The general objective of the data mining is to discover "interesting" descriptive statistics in the database. Like affinity grouping and clustering, it will be up to the auditor to interpret the results. However, this is a good starting point in that the results of this data mining will motivate the auditor to conduct more focused data mining – even directed data mining – to prove or disprove his/her initial interpretations of the descriptive results.

While, data mining has been more traditionally applied to highly structured data, text mining is a growing form of data mining that understands statistical and linguistic patterns in bodies of text (Witten, 2005; Srivastava and Sahami, 2009; Berry and Kogan, 2010; Weiss, 2010). With enhanced availability of large volumes of text from the Internet and elsewhere, text mining has become an increasingly important and widely used family of techniques. When text is communicated within a community of interest, for example by emails, social network analysis can be applied to the corpus (Debreceny and Gray, 2011; Scott, 2013; Worrell et al., 2013). As will be discussed later, process data mining and user roles (for segregation of duties analysis) are even newer data mining domains that have a great deal of potential in fraud detection.

2.2. Application of data mining to auditing and fraud

In recent years, there has been an increase in the application of data mining techniques to financial statement audits. In this subsection, we review research that uses data mining to understand potential fraudulent patterns in financial statement disclosures, business processes within the corporation, journal entries and in text generated within the corporation for internal and external use.

The most long standing application of data mining has been in mining financial statement disclosures (Green and Choi, 1997; Fanning and Cogger, 1998; Feroz et al., 2000). After a hiatus of a decade, there again has been an increasing attention to mining of financial statement data for identification of potentially fraudulent corporate reports. For example, Ravisankar et al. (2011) data mine core financial statement data points (e.g. Net and Gross Profit) and ratios (e.g. Long term debt/Total capital and reserves) to identify fraudulent reports. Their dataset is 101 Chinese corporations with known frauds, matched with non-fraud corporations. Ravisankar et al. (2011) employ a range of data mining techniques including support vector machines, neural networks, and genetic programming. One of the neural network techniques, Probabilistic Neural Network, was the most productive, with a surprising greater than 90% ability to identify correctly the corporations in the dataset in directed data mining tasks. Alden et al. (2012) employ Evolutionary Algorithms (EAs) to identify fraud in financial statements, with relatively high discriminant capability.

Perols (2011) addresses issues that arise with the identification of fraud in financial statements. These include low levels of observed fraud in a given population of audited entities, variations in cost between, for example, false positives and false negatives (with the latter being dramatically more expensive to the auditor and capital markets than the former), and the messy nature of financial statement data. After accounting for these characteristics in the design of his dataset, Perols finds that logistic regression and the widely used support vector machines (SVM) classification technique both perform well in identifying fraudulent companies and outperform other techniques such as different forms of neural networks. As Perols notes, logistic regression and SVM are both well understood and relatively efficient to deploy in production environments.

Another growing strand in the application of data mining to auditing is process mining (Van der Aalst, 2011; Jans et al., 2013). This technique extracts business process knowledge from event logs generated by corporate information systems, typically ERP systems. Process mining research has been carried out on corporate decision making processes (van der Aalst et al., 2011), compliance and risk management (Caron et al., 2013), and control structures (Elsas, 2008). While process mining could involve a variety of data sources, such as workflow and role analysis in ERP systems, most recent research concentrates on mining the process knowledge inherent in activity logs. Alles et al. (2006) and Jans et al. (2010) both develop systems to capture and then mine, processes from large scale log files. In the area of employee fraud, Jans et al. (2010) employ univariate and multivariate clustering techniques to identify potentially fraudulent transactions in a large dataset of purchase requisitions within one corporation. As is typical in this class of data mining, the analysis cannot categorically identify actual frauds—the analysis identifies outliers and other transactions that do not fit the predicted trends or patterns. In addition, the number of questionable purchase orders identified can be too large for human investigation. As such, a sample must be taken and subject to further inspection.

In establishing the opportunity to commit fraud, understanding the roles that users can play within enterprise systems is an important input to the audit of internal controls and particular transactions. Understanding of roles closely aligns to the analysis of segregation of duties. In one of the first papers on role mining, Colantonio et al. (2011) establish a methodology for role modeling and apply this methodology to a large dataset from a major corporation.

Arguably, the area where auditors currently conduct the most analysis of large-scale datasets is the analysis of client journal entries required under SAS No. 99. Debreceny and Gray (2010) take the first step in data mining journal entries. Journal entries have unusual characteristics that derive from the underlying business processes that give rise to the entries. Debreceny and Gray (2010) mine a corpus of 29 sets of journal entries from practice. They show that some techniques, such as Benford's Law, do not apply to this corpus. Unfortunately, the authors did not have access to known fraudulent journal entries and as a result do not employ more sophisticated data mining techniques. Argyrou (2012) employs Self Organizing Map (SOM) to identify suspicious transactions in a set of journal entries for a single corporation. Argyrou seeds errors in the set. The SOM technique identifies the errors with a high degree of accuracy and analyzes the cost of misclassification of entries between Type I and II errors.

An increasing area of interest is the application of text data mining techniques to the audit. One of the most important sources of internal corporate communications is email, which is an all-pervasive method for communication within entities and between staff in those entities and the ecosystem of suppliers, customers, advisors, and consultants. Text is at the heart of the email conversation. The semi-structured nature of emails with date, recipient, and subject fields add important temporal and social network dimensions to the data mining of emails. Further, given the importance of maintaining emails for discovery in future litigation and regulatory requirements has meant that organizations increasingly keep well-structured email archives. Debreceny and Gray (2011) introduce and analyze the text mining and social network analysis that apply in the data mining of emails from a fraud and audit perspective. They provide a practical example of social network analysis of emails, based on an audit-driven analysis of the publicly available Enron archive of emails. They provide a structured evaluation of potential research in the application of data mining of emails to audit and fraud detection.

An active thread in the application of text mining to fraud is deception analysis. This approach to research leverages well known cues of intentional misrepresentation by content deception in language made by senders intending to deceive information recipients (Debreceny and Gray, 2011). Known patterns of deceptive language include higher levels of active language and negative emotion in the writing. Foundational theories include Information Manipulation Theory (IMT) (McCornack, 1992) and Interpersonal Detection Theory (Buller and Burgoon, 1996). Humpherys et al. (2011) employ linguistic cues in Management's Discussion and Analysis (MD&A) to identify possible deception that may be indicative of fraudulent misrepresentation. Humpherys et al. (2011) take 101 actions by the SEC in the Commission's Accounting and Auditing Enforcement Releases (AAERs) as being examples of financial statement fraud. They matched 101 fraudulent examples with a similar number of non-fraudulent examples. Using a model that identified characteristics, such as active language, affect, word length and sentence complexity, they were able to develop a model that correctly identified 67% of fraud and non-fraud examples. Glancy and Yadav (2011) analyze deception in both the MD&A and notes to the financial statements. They propose a new model for detecting deception, which they term the Computational Fraud Detection Model (CFDM), with generation of Singular Value Decomposition vectors (SVD) that is typical of data mining, underpinning their model. They use a commercial mining product (SAS's Enterprise Miner) to undertake the mining. As with Humpherys et al. (2011), Glancy and Yadav (2011) use SEC AAERs to identify fraud. Glancy and Yadav are able to identify potentially fraudulent corporations with reasonable accuracy, using textual disclosures that predate the AAER. In other words, these very preliminary results are able to proactively identify potentially fraudulent corporations.

While this early flowering and rejuvenation of research on data mining are encouraging, there is little direction. Most of the papers seem opportunistic in nature. With the important exception of Perols (2011) and to some extent of (Humpherys et al., 2011), the research is largely devoid of understanding the audit environment, known patterns in fraud or of the implementation of data mining in professional service firms. We address some of these issues in the next sections.

3. Data mining at each phase of the audit

This section introduces the objectives of the audit and its various phases from a fraud detection perspective. The section explores expanding the auditor's data domain to improve auditor performance in fraud detection. While auditors may feel overwhelmed by the data they currently examine in the audit, such data comprise only a small part of the client's potentially fraud-relevant data inside a company. In addition, there is a wide range of data external to the client that is potentially relevant in assessing possible material misstatements arising from fraud. SAS No. 1 mandates that the auditor "has a responsibility to plan and perform the [financial statement] audit to obtain reasonable assurance about whether the financial statements are free of material misstatement, whether caused by error or fraud" (PCAOB, 2003). The focus of this study is on the detection of fraud. SAS No. 99 decomposes fraud into fraudulent financial reporting and misstatements arising from the misappropriation of assets. While the latter class of fraud is important, it is not, as we will discuss in more detail later, as significant as fraudulent financial reporting (PCAOB, 2002). SAS No. 99 (para. 06) defines fraudulent financial reporting as "intentional misstatements or omissions of amounts or disclosures in financial statements *designed to deceive* financial statement users where the effect causes the financial statements not to be presented, in all material respects, in conformity with generally accepted accounting principles (GAAP)" (emphasis added) (PCAOB, 2002). SAS No. 99 and the other auditing standards require that the auditor undertake a variety of analytical and planning tasks and substantive audit procedures to support the detection of errors arising from fraudulent financial reporting. A stylized view of the various phases of the audit is shown in Fig. 2.

3.1. Phases of the audit

We now address the nature of each of these phases in turn; pointing to the potential role that data mining may play in each phase.

3.1.1. Understanding the client

The assessment of risk of material misstatement whether from error or fraud drives the audit. Developing an understanding of the client is a key aspect of the high-level audit planning. As part of the early stage of the planning process, the auditor must understand a variety of client-related risk factors. These include ownership and organizational structures; the nature of value-adding processes; business partner and related party relationships, and the regulatory environment in which the client operates. Given the complexity of the typical audit client in terms of business processes, organizational structures, and business partner relationships, there is potential to exploit data mining tools and techniques to

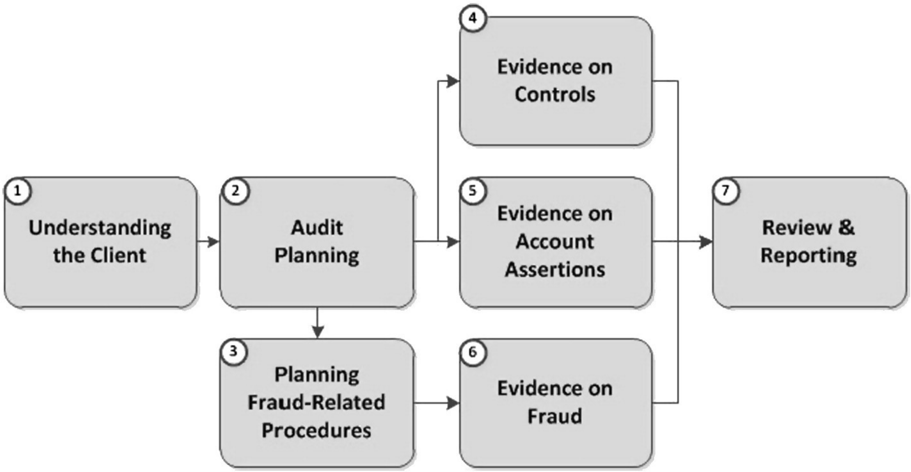


Fig. 2. Phases of the audit (adapted from Debreceny and Gray (2011)).

increase the auditor's analytical power over client performance data, external relationships, and networks. Typically, much of this pre-audit data mining will focus on readily available data external to the client such as published financial statements, press releases and analyses, analyst reports, share prices, regulatory filings, etc. Higher quality performance data within the client are also open to data mining. Because every client has data structures that have some unique aspects to them, the cost of acquiring and interpreting this latter class of data is relatively high, however, and careful attention is necessary to ensure that appropriate data mining techniques are employed for high value outcomes.

3.1.2. Audit planning

Detailed risk-based audit planning follows this high-level risk assessment of the client. As part of the planning process, the auditor normally will undertake a variety of analytical procedures to develop expectations of realized account balances such as levels of debt, cash generation, and levels of accruals. Often auditors will undertake relatively simplistic ratio analysis of core financial statement data points to generate these expectations. Data mining that combines analysis of external data of national and industry trends with client-level data may make this part of the audit more effective and efficient. National and industry levels of output, revenue and profitability are averages, and clients will necessarily perform at a different level than these averages, nonetheless, data mining of this combined dataset is likely to deliver superior results in the development of audit plans.

3.1.3. Planning fraud-related procedures

SAS No. 99 requires that the auditor systematically address the risks of potential error arising from fraud as part of the planning process. Auditors are required to discuss fraud risks in the audit planning process in a methodical fashion, including undertaking active brainstorming by the audit team. Completing a variety of analytical procedures is also an important component of this phase of the audit. SAS No. 99 appropriately notes that traditional analytical procedures may be at too high a level of aggregation. For example, SAS No. 99 notes in respect of potentially fraudulent reporting of revenue:

> *In planning the audit, the auditor also should perform analytical procedures relating to revenue with the objective of identifying unusual or unexpected relationships involving revenue accounts that may indicate a material misstatement due to fraudulent financial reporting. An example of such an analytical procedure that addresses this objective is a comparison of sales volume, as determined from recorded revenue amounts, with production capacity. An excess of sales volume over production capacity may be indicative of recording fictitious sales. As another example, a trend analysis of revenues by month and sales returns by month during and shortly after the reporting period may indicate the existence of undisclosed side agreements with customers to return goods that would preclude revenue recognition. (para. .29)*

While these examples are relatively naïve trend ratios analyzed temporally, they provide a flavor of the audit risk assessment procedures that can be undertaken with data mining tools. For example, rather than a straightforward ratio analysis of known metrics (e.g. inventory turnover and product line profitability), data mining of several metrics could lead to the development of a more sophisticated model of the client's value adding processes and risk profile. Another vital part of the planning stage is to consider the risks arising from weaknesses in internal controls. SAS No. 99 (paras. .44 and .45) requires auditors to assess the design and implementation of controls intended to prevent fraud. As part of their commitment to monitor the effectiveness of internal controls, clients often prepare extensive matrices of account assertions and the design and existence of controls. These matrices can be subject to relatively straightforward data analysis for assessment of risks and key controls for subsequent testing.

3.1.4. Evidence on controls

Following the planning phase, the collection of audit evidence commences. A key aspect of this process is to assess the operating effectiveness of internal controls. Process mining on the operation of internal controls throughout the operating cycle is a potential application of data mining. Process mining typically operates on the system logs of, for example, ERP systems that underpin the general ledger. Conducting process mining on control processes, and attempted and actual control overrides is an example of evidence collection on the operation of internal controls.

3.1.5. Evidence on potential fraud considerations

When undertaking audit fieldwork, the auditor conducts substantive procedures, some of which relate specifically to the detection of fraud. For example, SAS No. 99 notes that substantive analytical procedures in respect of revenue may exploit "disaggregated data, for example, comparing revenue reported by month and by product line or business segment." Data mining on corporate performance data and revenue databases and other client databases will be potentially important as examples of fraud-related analytical procedures. Apart from substantive analytical procedures, SAS No. 99 notes that changes in the nature, timing, and extent of substantive audit procedures are required as the auditor responds to fraud-specific risks assessments.

3.1.6. Concluding the audit, review and reporting

SAS No. 99 requires the auditor to evaluate the likelihood of material misstatement due to fraud "at or near the completion of fieldwork" (para. .74) and respond appropriately. Such responses may include re-evaluating the use of data mining procedures rejected as cost-ineffective earlier in the planning process or opening up new data mining procedures because of red flags observed during the fieldwork. For example, substantive analytical procedures or other tests on revenue may indicate need for mining of, for example, key executive emails as we discuss in more detail later in this paper. Following this process, final review of audit evidence is undertaken, leading to decisions on the nature of the final audit report.

3.2. Planning the use of data mining

The discussion in the previous paragraphs shows how data mining might be employed at each phase of the audit. When considering the use of data mining the auditor must undertake a data examination process, as illustrated in Fig. 3.

First, the audit team must decide what variables (and data sources that include those variables) to examine to meet their audit objectives (which includes in part the fraud types that came out of the brainstorming session)—and to define what constitutes red flags. These client-related internal and external variables and data sources would be selected based on the audit firm's policies and procedures; audit templates that are developed by the firm for specific audit objectives; the fraud-related brainstorming session of the audit team, and the variables that might be suggested by the individual auditors based on their own experiences or something they have seen at the client during the current audit or past audits.

Second, those internal and external variables in appropriate data sources are mapped to the variables in the data analysis and mining software that the auditor employs. For example, if the auditor wanted to run a query that included vendor ID, check ID, check amount, and check date of the client's database, the auditor would have to determine the appropriate field names and corresponding tables in the client's database. Third, the auditor would have to determine which procedures and tools they would use to examine the collected data. If the auditor was considering using data mining tools at this point, the auditor would have to determine which data mining techniques are most appropriate considering the characteristics of the data. Fourth, the auditor must then investigate the resulting red flags and determine what actions to take next based on the investigation results.

In terms of the actual investigation, if the auditor uses simple data queries that have little or no built in diagnostic value, the auditor is essentially conducting the complete investigation. With data analysis tools, the tools can help pinpoint transactions that are more suspicious and, thereby, increase the auditor's productivity. Finally, data mining tools would have the highest level of diagnostic ability and, as such, would reduce the amount of subsequent investigation that the audit team must undertake.

4. Fraud patterns and data mining taxonomy

The application of data mining techniques to fraud detection requires: 1) analysis of fraud risks; 2) identification of potential methods undertaken to commit the potential fraud; 3) determination of the availability of indicators of the fraud and associated methods, and 4) selection of appropriate data mining techniques most likely to discover these indicators. The previous sections pointed to possible use of data mining techniques and sources of data in the context of the financial statement audit. It would be a major undertaking if auditors attempted to evaluate *all* data mining techniques and *all* data elements in *all* data sources for *all* types of fraud. The objective of this section is to provide guidance as to the most effective

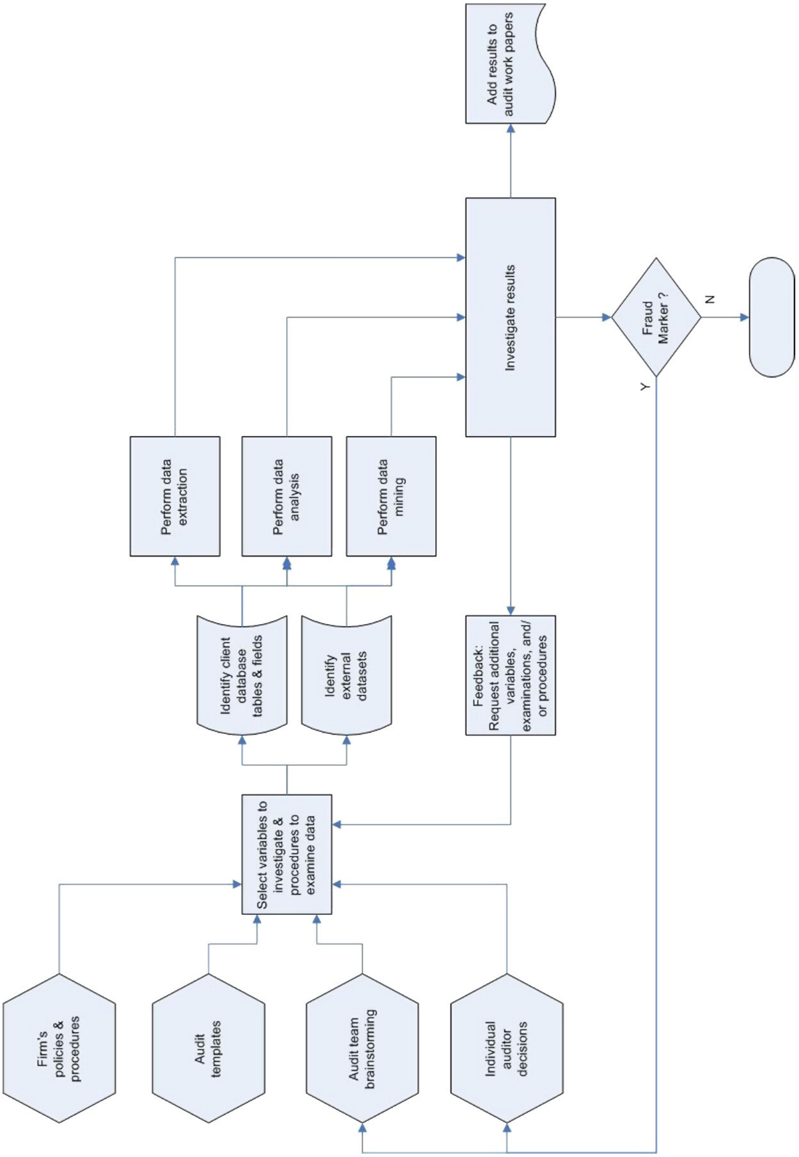

Firm's policies & procedures
Audit templates
Audit team brainstorming
Individual auditor decisions
Select variables to investigate & procedures to examine data
Identify client database tables & fields
Identify external datasets
Perform data extraction
Perform data analysis
Perform data mining
Investigate results
Feedback: Request additional variables, examinations, and/ or procedures
Add results to audit work papers
Fraud Marker ?
Y
N

Fig. 3. Data examination processes in the audit.

Table 1
Occurrence of account schemes (fraud types) (n = 148).
Source: Gao and Srivastava (2011).

Account Schemes	Occurrence
Premature Revenue Recognition	34%
Fictitious Revenues	25%
Overvalued Assets and Understated Expenses	11%
Omitted or Understated Expenses/Liabilities	8%
Other Methods to Overstate Revenues	5%
Omitted or Improper Disclosure	5%
Fictitious Assets	5%
Overvalued Assets/Equity	3%
“Wrong Way” Frauds	1%
Miscellaneous	1%

and least effective applications of data mining for fraud detection. We develop a taxonomy with three components. The first two components are the two components of fraud schemes, namely, account schemes (accounts impacted by the fraud) and associated evidence schemes that fraudsters have adopted to undertake the fraud drawing upon the work of Gao and Srivastava (2011). The third component is the applicability of data mining techniques to each combination of the account scheme and evidence scheme based primarily on the characteristics of the evidence scheme. The most fundamental question is: Does the evidence scheme include something that can be data mined? Some of the evidence schemes identified by Gao and Srivastava can cost effectively be subjected to direct examination by data mining techniques. Other evidence schemes require data mining and assessment of indirect signals and still other evidence schemes do not provide data mining opportunities. We draw this discussion together in a set of recommendations for areas of research and data mining projects that the researchers and the auditing profession might productively invest resources.

4.1. Classes of account schemes and associated evidence schemes

While theft or misappropriation of assets represents a significant area of concern, earnings management has represented the most significant frauds in recent decades. Earnings management is placed on a continuum from genuine and justifiable choices on accruals to deliberate actions that are entirely without merit and are fraudulent (POB, 2000). Gao and Srivastava (2011) develop a fraud taxonomy that considers both the type of fraud and the method employed to implement and hide the fraud. Gao and Srivastava describe this latter aspect as the *evidence scheme*. They studied SEC Enforcement Actions from 1997 until 2002 as the foundation for the instantiation of their fraud taxonomy. They found 100 enforcement actions in that period for which information existed on evidence schemes, representing 148 individual fraudulent events. Table 1 lists the account schemes (classes of fraud), in descending order of occurrence. Unsurprisingly, given the traditional focus on fraudulent revenues and revenue recognition, Gao and Srivastava identify significantly more than half of the fraudulent actions as involving revenues.

Gao and Srivastava also studied the evidence schemes involved in each of these frauds. Fraudsters use evidence schemes for dissimulation (“hiding the real”) or simulation (“showing the false”).[5] The taxonomy of evidence schemes include fake documents (e.g. “fictional shipping documents” or “fabricated stock certificates”); hidden documents (e.g. “keeping secret collection memoranda to track collections of contingent transactions”), and collusion with third parties (e.g. “requesting customers to provide false audit confirmations”). Table 2 lists the evidence schemes, again in descending order of occurrence.

Gao and Srivastava provide data that correlate occurrence of frauds and evidence schemes. Fig. 4 provides a visual representation of this correlation. The three black cells represent the highest combinations of account scheme and evidence scheme frequencies. These cells represent areas where data mining could potentially

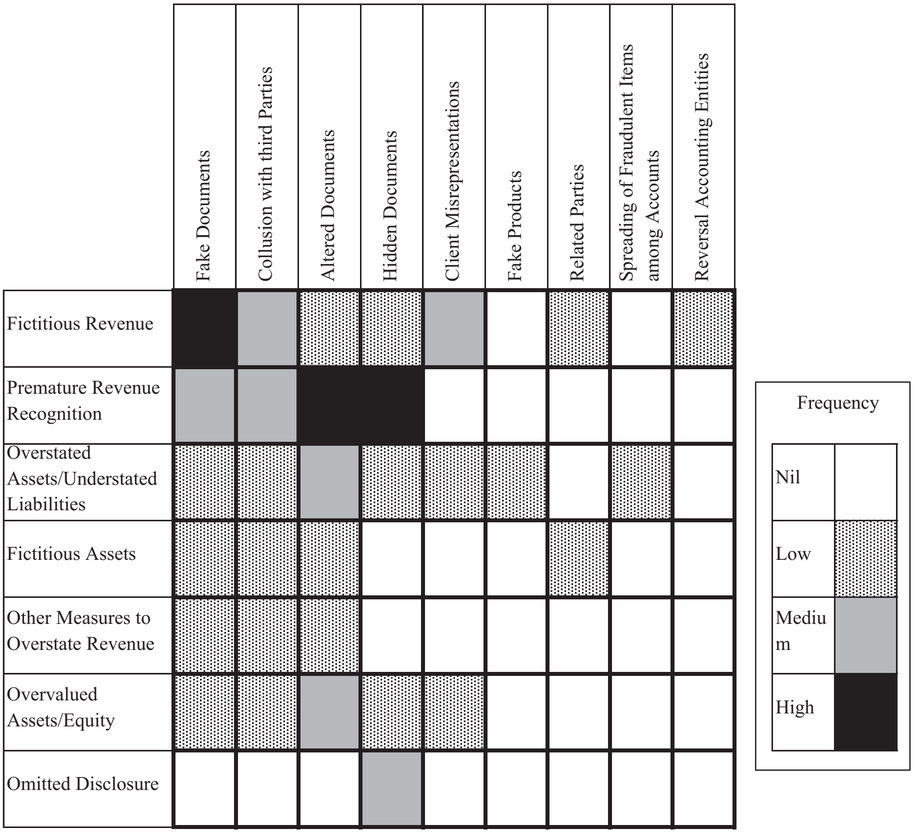
[5] Gao and Srivastava draw this formulation from Bowyer (1982).

Table 2
Evidence schemes employed (n = 185[a]).
Source: Gao and Srivastava (2011).

Evidence Schemes	Occurrence
Hidden Documents/Information	25%
Altered Documents	22%
Fake Documents	19%
Collusion with Third Parties	15%
Client Misrepresentations	8%
Improper Related Party Transactions	3%
Shifts and/or Spreading of Fraudulent Items among Accounts	3%
Fake Products/Equipment	2%
Reversal Accounting Entries	2%
Miscellaneous	1%

[a] There are multiple evidence schemes employed in some frauds.

provide significant returns. The many white cells represent no intersection between the particular account scheme and the particular evidence scheme. There would be little reason to invest significant research or other resources, at least in the short term, in building data mining techniques or skills for these null cells.



	Fake Documents	Collusion with third Parties	Altered Documents	Hidden Documents	Client Misrepresentations	Fake Products	Related Parties	Spreading of Fraudulent Items among Accounts	Reversal Accounting Entities
Fictitious Revenue	█	▒	░	░	▒		░		░
Premature Revenue Recognition	▒	▒	█	█					
Overstated Assets/Understated Liabilities	░	░	▒	░	░	░		░	
Fictitious Assets	░	░	░				░		
Other Measures to Overstate Revenue	░	░	░						
Overvalued Assets/Equity	░	░	▒	░	░				
Omitted Disclosure				▒					

Frequency	
Nil	
Low	░
Medium	▒
High	█

Fig. 4. Occurrence of fraud schemes: account scheme and evidence scheme combinations.

Please cite this article as: Gray GL, Debreceny RS, A taxonomy to guide research on the application of data mining to fraud detectio..., Int J Account Inf Syst (2014), http://dx.doi.org/10.1016/j.accinf.2014.05.006

4.2. Fraud scheme indicators

The relative frequencies of combinations of account schemes and evidence schemes shown in Fig. 4 provide guidance for auditors in both the planning and fieldwork phases of the audit. The key question is how can data mining identify fraud indicators for these combinations? As discussed below, some of the evidence schemes are subject to direct and cost-effective identification by data mining techniques. Other evidence schemes will require identification of secondary evidence for indicators that may lead the auditor to the direct evidence within that scheme.

Evidence schemes that directly influence information within the accounting information system, broadly defined, are tractable with data mining approaches. For example, frauds involving the “spreading of fraudulent items among accounts” and “reversal accounting entries” are likely to be highly tractable with mining of journal entries or transactions that underlie the journal entries. Fraud indicators for evidence schemes that do not directly affect the accounting information system are more difficult to access and require analysis of secondary evidence. For example, it is difficult, but not impossible, to mine application transaction databases for “hidden” or “altered” documents. In the case of revenue frauds, for example, there must be journal entries in the sales systems or general ledger to book revenue for fabricated sales transactions. This class of transaction might well be highlighted as an outlier in a database of journal entries because of multiple indicators (e.g. some unusual combination of the date of the transaction(s), atypical pattern of transactions for a single client, journal entry description etc.).

Collusion with third parties and fraudulent related party transactions are particularly difficult to detect directly. However, it is possible that prima facie evidence of collusion will come from the accounting information system. For example, in a fraudulent revenue scenario, there may be indicators of unusual relationships such as dramatically increased sales to a client or higher profit margins in a product group or client, perhaps coupled with higher days outstanding in receivables. Supplementary mining of emails with the name(s) of the product group, customers and managers associated with approval of sales may bring evidence of collusion to light. Of course, if executives are determined to hide collusion, they will ensure that not all interactions with the relevant third parties or with other executives within the corporation use the corporate email system. The reality, however, is that these systems are so much part of the corporate DNA that executives fall back on them even for communications that are subsequently clear evidence of wrongdoing.

SAS No. 99 provides some guidance on these types of secondary evidential matter. The standard suggests that auditors analyze key performance metrics longitudinally and cross-sectionally in comparison to peers for evidence of unusual and potentially fraudulent financial reporting patterns. At a much more detailed level, as we discussed earlier, the standard suggests that evidence of potentially fraudulent revenue may rest in a variety of detailed analytical procedures. These procedures may require a drill down into the nature, type, and timing of sales transactions. The auditor should also assess revenue-associated events affecting accounts such as inventory, accounts receivable, and cash. Related indicators may also include analysis of sales levels with production volume. The procedures suggested by SAS No. 99 imply that not only does the auditor have to maintain a detailed understanding of the value-adding processes of the client, but also makes an in-depth analysis of key performance metrics and financial statement relationships. These metrics and relationships will need to be assessed for the client as a whole as well as core subsets such as geographic and product categories.

Paragraph .68 of SAS No. 99 provides guidance to the auditor on the assessment of fraud risks. Some of the risks identified in Paragraph .68 are identified during the conduct of the audit as a by-product of other procedures. For example, SAS No. 99 alerts auditors to “inconsistent, vague, or implausible responses from management or employees arising from inquiries or analytical procedures,” “complaints by management about the conduct of the audit or management intimidation of audit team members, particularly in connection with the auditor's critical assessment of audit evidence or in the resolution of potential disagreements with management,” and “undue time pressures imposed by management to resolve complex or contentious issues.” These risk factors provide color in the conduct of the audit, but are not subject to testing directly by substantive procedures involving data mining or otherwise.

Conversely, guidance in Paragraph .68 to auditors to assess the following risks relates in some way to potential data mining:

- Last-minute adjustments that significantly affect financial results
- Unsupported or unauthorized balances or transactions

- Evidence of employees' access to systems and records inconsistent with that necessary to perform their authorized duties
- Significant unexplained items on reconciliations

The accounting DNA is represented in the journal entries in the client's general ledger. Sections .58 through .61 of SAS No. 99[6] mandate a set of procedures to test the accounting information system and the journal entries therein for potential misstatements arising from fraud. As SAS No. 99 notes, a number of financial statement frauds have involved "inappropriate or unauthorized" journal entries as well as adjustments made outside the formal accounting system, such in the consolidation process leading to a set of consolidated financial statements or even by manual entries to the draft of financial statements. These latter adjustments are beyond the scope of data mining. Evidence of such frauds is seen directly within, for example, specialized computerized consolidation systems or by a process of reconciliation from the final adjusted trial balance to the final reported financial statements.[7]

4.3. Overlaying data mining onto the fraud scheme taxonomy

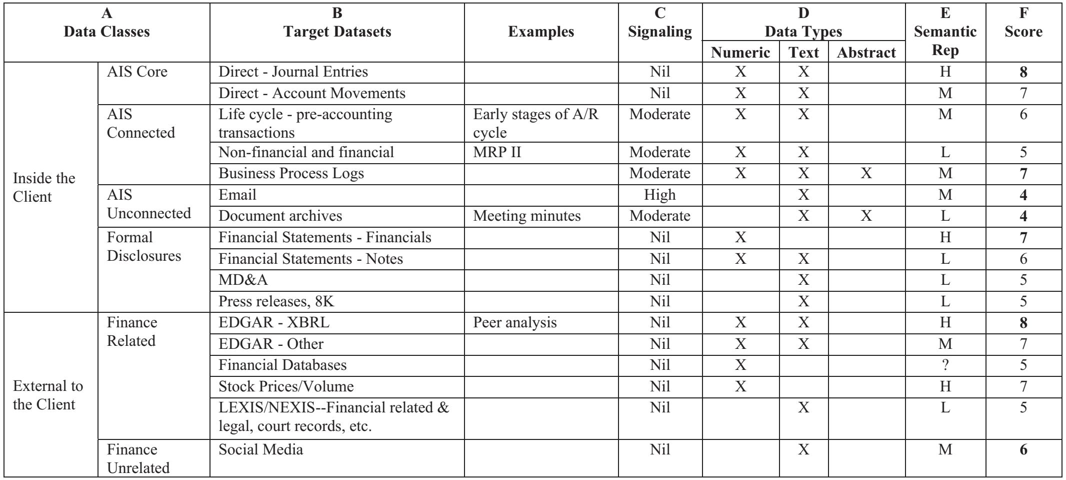
Earlier in this section, we set out a taxonomy of fraud schemes including account schemes and evidence schemes. In the following subsections, we first set up a categorization of data mining target datasets that is fine-tuned for the external auditing domain, illustrated in Fig. 5. The critical concept underlying Fig. 5 is that it provides an analysis of *where* data mining can be productively applied within the context of the external audit. The likelihood that data mining can be applied successfully in the audit setting is scored. A variety of data mining techniques can be applied to each of the datasets; however, not all datasets are equally productive. Factors include the connection of the information source to the key concerns of the financial statement audit, the availability of data to the audit, and the quality of the semantic representation of the information source. As a result, we rank the productivity of each of the rows in Fig. 5, and return to the method for ranking shortly. First, we explain each of the columns in Fig. 5.

The first dimension represents the sources of information that exist within the client's information system and those that are external to the client (column A). We then further sub-divide the information sources within each of these two principal classes (column B). The first two classes of datasets within the client's information system are those that connect in some ways to the client's accounting information system. The "AIS Core" represents the heart of the accounting information system, represented primarily by the General Ledger or equivalent system or module in enterprise systems. There are two datasets for data mining within the AIS core — journal entries that represent the totality of transactions linking all accounts within a designated time period. The second dataset within the AIS core is analysis of transactions over time within a particular account or class of accounts.

"AIS Connected" is a transaction system that feeds journal entries to the AIS core. The first component of AIS connected is the set of applications that exist within the enterprise system that are tightly coupled with the General Ledger. Examples include customer relationship management, inventory management, sales and payable applications. The focus of this dataset is on aspects of the early part of the lifecycle of the transaction. The second dataset in this category is a variety of less tightly coupled application systems that pass summary journal entries to the General Ledger but, more importantly, contain huge volume of non-accounting and transaction data. Examples include maintenance and production systems, in the manufacturing environment. The third dataset is the business process logs that are maintained by enterprise and other transaction systems. For both "AIS core" and "AIS connected," we recognize that there will be considerable variation between clients depending on factors such as the client's adoption of enterprise systems, geographical distribution, and number of divisions. The next category of datasets is the class of client information systems that do not connect to the General Ledger, but which have audit-relevant information. This category has two classes of datasets: corporate email and other document and information repositories.

[6] SAS No. 99 followed the report of the then Public Oversight Board's *Panel on Audit Effectiveness* (POB, 2000). The Panel's report provides further elucidation of the factors involved in financial statement fraud involving journal entries and otherwise.

[7] A closely related class of financial statement fraud is inappropriate disclosure in footnotes or misclassification within the financial statements.



A **Data Classes**		**B** **Target Datasets**	**Examples**	**C** **Signaling**	**D** **Data Types**			**E** **Semantic Rep**	**F** **Score**
					Numeric	**Text**	**Abstract**		
Inside the Client	AIS Core	Direct - Journal Entries		Nil	X	X		H	**8**
		Direct - Account Movements		Nil	X	X		M	7
	AIS Connected	Life cycle - pre-accounting transactions	Early stages of A/R cycle	Moderate	X	X		M	6
		Non-financial and financial	MRP II	Moderate	X	X		L	5
		Business Process Logs		Moderate	X	X	X	M	**7**
	AIS Unconnected	Email		High		X		M	**4**
		Document archives	Meeting minutes	Moderate		X	X	L	**4**
	Formal Disclosures	Financial Statements - Financials		Nil	X			H	**7**
		Financial Statements - Notes		Nil	X	X		L	6
		MD&A		Nil		X		L	5
		Press releases, 8K		Nil		X		L	5
External to the Client	Finance Related	EDGAR - XBRL	Peer analysis	Nil	X	X		H	**8**
		EDGAR - Other		Nil	X	X		M	7
		Financial Databases		Nil	X			?	5
		Stock Prices/Volume		Nil	X			H	7
		LEXIS/NEXIS--Financial related & legal, court records, etc.		Nil		X		L	5
	Finance Unrelated	Social Media		Nil		X		M	**6**

Fig. 5. Data mining targets in the audit environment.

The next category of datasets is the formal disclosures which the auditor is directly responsible for auditing or which are closely associated with the audit objective. These datasets include the financial statements and notes and the MD&A as well as press releases and regulatory filings, such as 8-K filings with the SEC, closely associated with financial reporting process. The financial statements and MD&A will include both prior quarterly and annual statements, as appropriate, as well as the draft of the current reports. The press releases and regulatory filings will occur throughout the fiscal year.

The second principal source of information for data mining exists external to the client, divided between those that are related to finance and broader sources. Finance-related datasets include the XBRL filings within the SEC's EDGAR system and other EDGAR filings. Collectively, these filings provide a large set of highly disaggregated, entity-specific disclosures. We differentiate the XBRL filings to the SEC from analysis of the audit client's financial statements that we discuss above. The former allows peer analysis that utilizes all the filings on EDGAR. There is also a variety of financial databases (e.g. S&P Capital IQ) that can be mined for fraud detection. In addition, stock prices and volumes can provide pointers to potential fraud. Finally, in this group of datasets is LEXIS/NEXIS, which can provide analysis of news items, court filings, and a variety of other information sources. Information sources with potential for data mining beyond finance include social media, such as Twitter and Facebook.

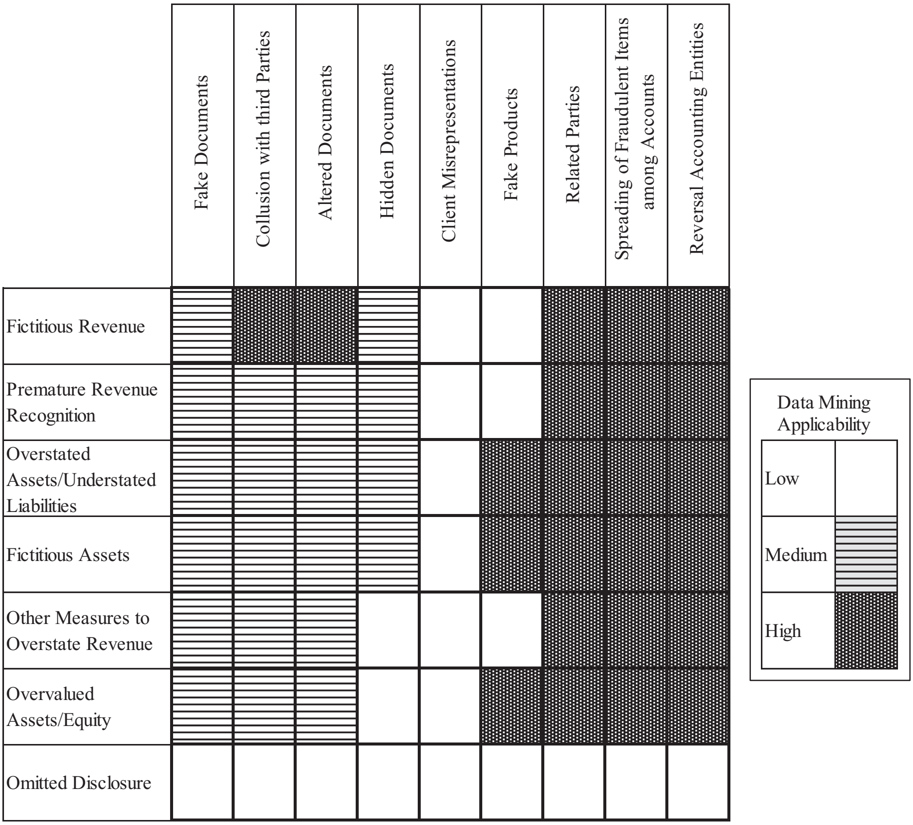
We now introduce our scoring system for rating each of the target datasets. The first factor we title "Signaling" (column C). In the context of the financial statement audit, some requests by the auditor for additional information will be seen by the client as a relatively normal part of the audit. Other requests would be seen by the client as highly unusual and may trigger a significant level of concern from the client as to the direction that the audit is taking. As a result, some types of data requests would only be undertaken when the benefit from mining this data was expected to be especially rewarding. An example of a request for additional data that might raise only moderate concern from the client would be data from maintenance systems (relevant to questions of asset valuation) or sales systems (relevant to revenue recognition). These data are closely aligned to established flows of accounting transactions in the General Ledger. An example of requests for additional data that would clearly raise concerns from the client would be a request for the content or headers (metadata) of senior management emails. At the same time, the likely fraud detection payoff from mining these emails is likely to be high. A secondary factor is the cost of acquiring and understanding data sources. It has taken the major audit firms some years to set up business processes to acquire and understand client journal entries, required by SAS No. 99. Signaling is scored as: 1) "nil" (no signaling) because the auditor receives the data from the client in the normal course of the audit (e.g. journal entries or financial statements) (scored as three), 2) "moderate" signaling where a clear connection to standard audit practices can be established (e.g. maintenance data) (scored as two) and 3) "high" signaling (e.g. emails) (scored as one).

The next attribute of these target datasets is the data types that could be in the datasets (column D). The data types of interest include numeric, textual and abstract representations. Different data mining techniques apply to each of these data types. For example, when considering journal entries, the account numbers and the associated economic values in the journal entries are the most obvious subject of data mining. At the same time, however, the text in comment fields and elsewhere can be mined. Each data type contributes a score of one (maximum of three). The next dimension of data mining attributes is the level of semantic representation of the information source (column E). Higher semantic representation will assist the efficacy of data mining. Again, using journal entries as an example, data mining can leverage the internal logic of the double entry bookkeeping system and the chart of accounts. We score this as: ?low = one, medium = two and high = three. The total score for each dataset is shown in column F. The maximum score possible is nine. In our analysis, we take the highest score for each class (column A) of datasets. For example, as shown in Fig. 5, we identify five finance-related sources of information sources external to the client. The score for the first of these datasets, the SEC's EDGAR XBRL repository, has a score of eight. We apply that score to this class of target datasets.

4.4. Applying scoring system to fraud and evidence schemes

Having developed this scoring scheme, we then apply it to each cell in the account scheme and evidence scheme matrix (Fig. 4). We score the likelihood that a particular class of datasets will apply to a particular combination of fraud and evidence scheme. Where we see a high degree of application of the particular category of datasets to the given combination of fraud and evidence scheme, the score shown in Column F of Fig. 5 is applied in full. If the datasets are believed to be only of moderate utility, a weight of 50% is set. Finally, if datasets are considered to be only of low utility, the weight is set at zero. The scores for all classes of datasets are summed for each combination of fraud and evidence scheme. The scores are then ranked and trifurcated into low, medium and high levels of application of data mining to each combination of fraud and evidence scheme. Fig. 6 shows the heatmap of the fraud and evidence schemes and the possible contribution of data mining (low, medium or high contribution) in identifying that fraud scheme (a specific account scheme and evidence scheme combination). The figure sets out a matrix of fraud schemes (rows) and evidence schemes (columns). Examples of combinations of fraud and evidence schemes that we see as having high application of data mining are a variety of fraud schemes associated with revenue (e.g. fictitious revenue) coupled with the "spreading of fraudulent items among accounts" evidence scheme. This evidence scheme typically requires journal entries to implement. This means that the "AIS core" group of target datasets (journal entries and account movements) are likely to be of significant value this specific account/evident fraud scheme combination and, as a result, are highly rated. At the same time, there are likely to be significant benefits from the "AIS connected" target datasets and, to a lesser extent, from the use of external datasets such as the XBRL filings.

For the "Overvalued Assets/Understated Assets" and "Fictitious Assets" schemes, data mining would operate primarily on indirect evidence rather than directly on the client misrepresentation. For example, we indicate that frauds involving overvalued assets and understated expenses (accounts) that employ posting of reversal entries (evidence) would be tractable with data mining. Not only is most or all of the information needed for data mining within a database of journal entries, but there is also triangulation between two



	Fake Documents	Collusion with third Parties	Altered Documents	Hidden Documents	Client Misrepresentations	Fake Products	Related Parties	Spreading of Fraudulent Items among Accounts	Reversal Accounting Entities
Fictitious Revenue	Medium	High	High	Medium	Low	Low	High	High	High
Premature Revenue Recognition	Medium	Medium	Medium	Medium	Low	Low	High	High	High
Overstated Assets/Understated Liabilities	Medium	Medium	Medium	Medium	Low	High	High	High	High
Fictitious Assets	Medium	Medium	Medium	Medium	Low	High	High	High	High
Other Measures to Overstate Revenue	Medium	Medium	Medium	Low	Low	Low	High	High	High
Overvalued Assets/Equity	Medium	Medium	Medium	Low	Low	High	High	High	High
Omitted Disclosure	Low	Low	Low	Low	Low	Low	Low	Low	Low

Fig. 6. Application of data mining.

classes of accounts and across time (prior to period-end and the reversal in the next period). Direct mining of the evidence is feasible and cost-effective. The assessment of the potential for data mining in other cells draws on the same two factors.

Evidence schemes that involve fake, altered, or hidden documents form a key part of revenue frauds as well as frauds in areas such as fictitious assets. We show each fraud class for these evidence schemes as having either moderate or high application of data mining. Examples of documents are hidden side agreements with customers or altered or fraudulent sales contracts. It is less likely that direct data mining over sales systems will lead to high quality evidence of fraud markers. Indirect approaches may be more cost-effective and there are two areas of text data mining that are of particular interest that are relevant to frauds that use this class of evidence scheme. First, mining of comment fields and other textual inputs that support relevant sales and other journal entries for unusual patterns would identify potentially fraudulent transactions. The power of such analyses increases if triangulated with other data mining approaches. Textual analysis may be one way to reduce the number of transactions identified for subsequent examination. Textual analysis can use both known stop words as well as changes in text patterns. For example, in frauds involving revenue, when clients make sales regularly to an established customer, the descriptive text is unlikely to change dramatically. Second, text mining of emails for key executives involved in questionable sales transactions would seem to be an area of vital importance for many cells in the heatmap in Fig. 6.

Where frauds involve data triangulation with data sources that are more readily available to the auditor, data mining is of increased interest. The SEC database has a high proportion of frauds associated with revenue, including premature recognition and fictitious revenue. This class of fraud provides an interesting case study of where either directed or undirected data mining techniques for fraud markers are feasible. Typically, underlying entries for these sales are in supporting sales systems ("AIS connected") and in some cases evidence appearing in subsequent journal entries in the General Ledger ("AIS core"). Direct data mining of sales systems seems to be an area of particular interest. The objective would be to identify unusual transactions for subsequent audit enquiries. Automated analysis of patterns of sales and receivables databases should discover outliers (rapid increase in sales volume, high margins, high levels of outstanding receivables etc.). Final determination of the evidence of fraud will require understanding of the value-adding processes of the client and making enquiries of the client and inspection of underlying documentation.

There are a number of combinations of fraud and evidence schemes where we see relatively low application of data mining. For example, the "omitted disclosures" fraud scheme is categorized as having low applicability of data mining for all evidence schemes. The rationale for this categorization rests on the nature of the fraud, which involves either failure to disclose material information in the financial statements or misrepresentation of the real state of affairs. The underlying information is available in the accounting information system. The fraud arises from the way the financial statements reports on this information. Then consider the evidence scheme of "client misrepresentation." This evidence scheme involves the client's responses to auditor enquiry being, at worst, a direct falsehood or, at best, dissembling. There is, then, limited direct transactional evidence upon which data mining can operate that we might classify as client misrepresentation. As a result, we show low application of data mining for this evidence scheme for all account schemes.

4.5. Combined heatmap

Fig. 7 overlays the two heat maps of fraud schemes (account schemes and evidence schemes) and possible contribution of data mining in identifying these fraud schemes. For each combination of fraud (rows) and evidence scheme (columns), the upper block reflects the observed frequency of frauds (from Fig. 4). The lower block shows the likely benefits to be gained from data mining (from Fig. 6). For example, frauds that involved fictitious revenue that employed fake documents (upper left cell) were one of the more common types of account/evidence combinations (so it is colored in black). The likely contribution of data mining is moderate in the identification of this fraud scheme (so it is colored gray with horizontal lines).

In the remainder of this sub-section, we analyze the ratings of the application of data mining to the particular fraud schemes. As we note above, the role played by revenue in frauds is well known and

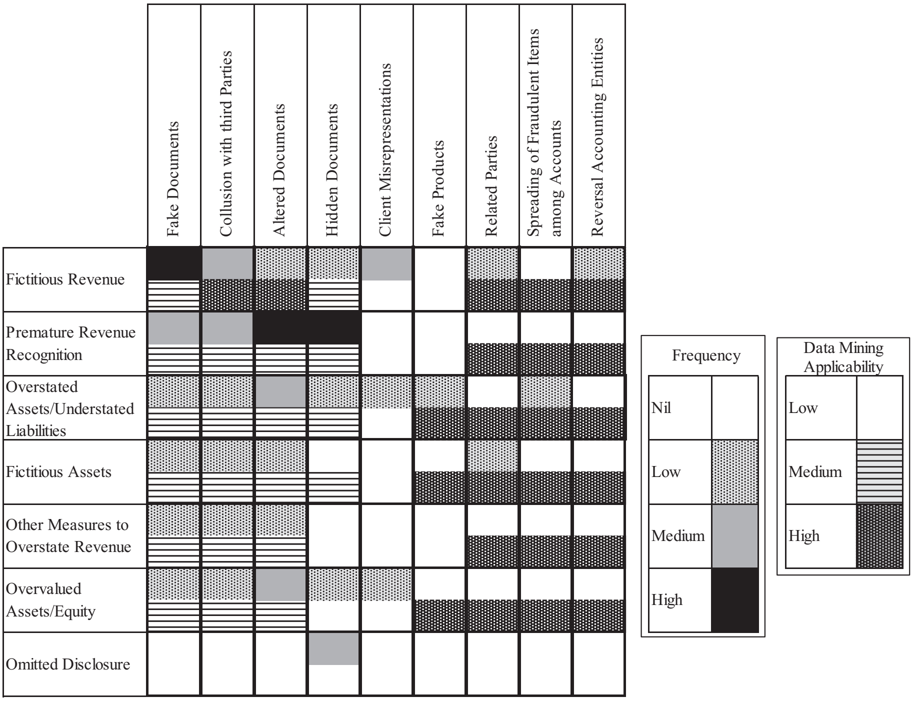

Fake Documents
Collusion with third Parties
Altered Documents
Hidden Documents
Client Misrepresentations
Fake Products
Related Parties
Spreading of Fraudulent Items among Accounts
Reversal Accounting Entities
Fictitious Revenue
Premature Revenue Recognition
Overstated Assets/Understated Liabilities
Fictitious Assets
Other Measures to Overstate Revenue
Overvalued Assets/Equity
Omitted Disclosure
Frequency
Nil
Low
Medium
High
Data Mining Applicability
Low
Medium
High

Fig. 7. Fraud and evidence schemes and application of data mining.

explicitly addressed in SAS No. 99. Interestingly, the three fraud schemes involving revenue ("fictitious revenue," "premature revenue recognition" and "other measures to overstate revenue") each have either moderate or high application of data mining for most evidence schemes. Similarly, frauds involving fraudulent measurement of assets and/or liabilities ("overstated assets/understated liabilities," "fictitious assets" and "overvalued assets/equity") each have moderate or high applicability of data mining for several evidence schemes.

There are, however, some clear mismatches between the frequency of account/evidence combinations and data mining applicability. As can be seen from Fig. 7, where there is high applicability of data mining, we do not see moderate or high frequency of frauds (right side of Fig. 7). As noted above, frauds that involve spreading of fraudulent items over several accounts or reversal of accounting entries can be productively addressed with data mining. Unfortunately, there are few frauds that involve these techniques.

Further, the likely contribution of data mining is not independent of the incidence of frauds. When frauds are very rare, it is more difficult for data mining to be successful in identifying fraud indicators (red flags). Further, merely because data mining is likely to be productive does not necessarily mean that there will be a payoff in areas of high productivity. The cost-benefit relationship of data mining techniques to evidence schemes falls along a continuum with implications for choice of mining target datasets.

5. Conclusions and suggestions for future research

The increasing value of data mining as a financial statement auditing tool is due to the convergence of several factors: (1) increasing emphasis on fraud detection in audits by regulators and standard setters, which provides motivation to identify and use tools to increase auditor productivity; (2) growing use of data mining tools as a forensic tool within accounting firms, which means there is a growing population of people within the firms with data mining experience as well as a general data mining awareness; and (3) the evolution of more robust and easier to use data mining tools. In addition, the expanding use of data mining as a *de rigueur* part of e-discovery in law suits have provided many examples of how data mining can be used for forensic investigations and, because of competition in the marketplace, e-discovery has accelerated the development of improved data mining tools. The growing general firm-level awareness of data mining and success of data mining in the legal profession may promote the firm-level decision to use data mining in more financial audits.

Fraud detection is a vital component of the modern financial statement audit. There are many forms of data mining that are available to auditors as part of their fraud detection activities. Equally, there are numerous potential data mining datasets. Some of these datasets are within the client organization (e.g. journal entries and email archives). Other datasets are external sources (e.g. EDGAR filings, social media). Just as there are many potential datasets that can be searched for indications of fraud, there are a variety of methods by which audit clients typically commit fraud. It is difficult, à priori, to decide which data mining tools can be cost effectively applied to which potential type of fraud. In this paper, we have addressed these decisions by developing a taxonomy that integrates observations of historical patterns of fraud schemes (combination of “account schemes” and “evidence schemes”) with an analysis of the application of data mining to these fraud schemes. The taxonomy shown in Fig. 7 helps identify where data mining could be the most effective (as well as the least effective). The taxonomy builds on the real world in that it reflects the frequency of different types of fraud schemes (combinations of account schemes and evidence schemes). The taxonomy also indicates that data mining will not be an effective tool to discover every type of fraud scheme and therefore data mining should be employed judiciously. There are several limitations to this research, several of which we address in the next subsection on future research. First, the scheme for scoring the application of data mining has not yet been fully tested with auditors, data mining specialists, or software and services providers. Second, while the study addresses the costs and benefits of data mining in the audit environment, these costs and benefits have not been directly measured and assessed. Third, there are a variety of issues and challenges that must be addressed in the implementation of data mining in the audit setting. These issues are beyond the scope of the paper, but do present limitations to the study.

5.1. Potential research questions

Because the current use of data mining is relatively ad hoc and new in the financial statement auditing domain, they are many potential research questions that should be addressed in future research. An important theme in the paper is the potential application of data mining publicly available information on audit clients. What benefits for auditing can arise from employing advanced data mining tools on this external information? A range of data mining tools could be investigated on a variety of data sources to assess broad and targeted (e.g. revenue) fraud indicators. In Section 2, we said that data mining tools fall into two broad categories: directed and undirected. Within each of these categories there are several specific data mining approaches. Determining which specific approach is most effective with specific evidence schemes would be very valuable information for the audit community.

An important recent development within the realm of external information is the availability of fully tagged financial statement data in the XBRL format. XBRL data are now available for SEC filers for several years. Additionally, the market for XBRL tools to mine these data is growing more robust. What will be the impact on audit planning and analytical procedures from the detailed disclosures made to EDGAR in XBRL? From this research, we could more fully understand the likely impact of widespread adoption of XBRL on audit planning and analytical procedures.

What is the value of textual data as a fraud-detection audit tool? What textual data mining tools seem to work best in the audit environment? Emails and other textual materials have been incriminating evidence is several highly-visible legal proceedings and congressional hearings. Textual data mining software have become more sophisticated. Further, there are products specifically designed for email data mining. Two or three text data mining tools could apply to two or three bodies of textual materials (e.g., email). Comparison of these tools can be on various dimensions such as content and readability of output reports and ease of use. Of particular importance are granularity and the number of false positives.

The relationship between data mining and continuous monitoring and auditing is an important area of future research. In general, data mining is batch processing of very large datasets, at infrequent time intervals. Conversely, continuous monitoring and auditing involve short-term (even real time) monitoring of transaction flows to spot anomalies and outliers when they occur. So, on the surface, it may seem that there is a fundamental mismatch between data mining and continuous monitoring. On the other hand, one of the key aspects of data mining is model building. After data mining software has identified a model, that model could then be the basis to continuously monitor subsequent transactions against the model to identify anomalies. Besides testing each transaction against the model, each new transaction would incorporate into the model so that the model becomes a dynamic rolling model to reflect the changes in the business' environment.

In an environment of information overload, auditors who attempt to incorporate data mining techniques throughout the audit life cycle must select from a range of audit techniques (How to select the best technique?) and potential data sources (How to select the best data sources? How to test the quality of non-financial data?) and prevent false-positives and spurious patterns.

As we noted above, we recognize that there is a variety of issues associated with accounting firms expanding their data mining activities that should be researched before data mining can achieve its potential. An important issue with data mining is the client's potential unwillingness to give auditors full access to applications and databases so that the auditors can easily expand the sources of data that they mine. Currently, clients do not typically give auditors direct access to their databases. Instead, the clients normally prepare copies of their data for the auditors that include only the records and fields that the auditors explicitly request. Subsequently requesting additional records and/or fields may cause delays in the audit. There are a number of prospective solutions to this issue that are open to future research. These include creation of an electronic “sandbox” in which the auditor can conduct data mining activities separate from the client's live data. Current and future versions of the Audit Data Standard may go some way to overcoming these concerns.

A theme in the paper is that auditors can include non-accounting (non-financial) data in their data mining domain. There are, however, no external standards that apply to non-financial data. As such, auditors will need to develop their own set of processes to determine the quality of non-financial data and the consistency of that data with the financial statements. This is a significant area for future research because of the growing activities of the Global Reporting Initiative (GRI), the Integrated Reporting Council (IIRC), and the World Intellectual Capital Initiative (WICI) to develop global non-financial and financial reporting standards.[8]

There are a variety of future research areas that address the potential issues that are likely to arise when data mining is more widely used in accounting firms. For example, while there is enhanced awareness of data mining and provision of data mining services in non-audit settings within the firms, the vast majority of audit engagement staff and partners do not have a technology background. As such, they are likely to have little understanding or awareness of data mining technology or what data mining tools can do to improve the effectiveness and efficiency of fraud detection. Selection of the appropriate data elements from the client's datasets and interpretation of data mining results are critical skills. Research on the development of these and other skills at all levels of the firms will be necessary. Research is also needed on the technological aspects of extensive use of data mining. We know little on the nature of learning curve costs with respect to understanding the client's various datasets.

An issue internal to services firms is: How does the data mining impact the risk profile of the accounting firm? When taking traditional small samples, if the “smoking gun” related to a fraud is not in their sample; the firm's defense is that their audit sampling followed industry practices. However, data mining can be considered the equivalent to taking a 100% sample. If the smoking gun is in that sample, but the auditors missed it, then the auditors no longer have their traditional industry-practice defense.

[8] For example, see http://www.wici-global.com/.

References

ACAP. Final report of the Advisory Committee on the Auditing Profession to the U.S. Department of the Treasury, VIIWashington, DC: Advisory Committee on the Auditing Profession to the U.S. Department of the Treasury; 2008. p. 1.

Alden M, Bryan D, Lessley B, Tripathy A. Detection of financial statement fraud using evolutionary algorithms. J Emerg Technol Account 2012;9.

Alles M, Brennan G, Kogan A, Vasarhelyi MA. Continuous monitoring of business process controls: A pilot implementation of a continuous auditing system at Siemens. Int. J. Account. Inf. Syst. 2006;7(2):137–61.

Argyrou A. Auditing journal entries using Self-Organizing Map. 18th Americas Conference on Information Systems. Seattle: Association for Information Systems; 2012.

Bashir A, Khan L, Awad M. Bayesian networks. In: Wang J, editor. Encyclopedia of data warehousing and mining: I–Z. Hershey, PA: Idea Group Inc.; 2006. p. 89–92.

Berkhin P. A survey of clustering data mining techniques. In: Kogan J, Nicholas C, Teboulle M, editors. Grouping multidimensional data: Recent advances in clustering. Heidelberg: Springer; 2006. p. 25–71.

Berry MW, Kogan J. Text mining: Applications and theory. Chichester, U.K: Wiley; 2010.

Bowyer JB. Cheating. New York, NY: St. Martin Press; 1982.

Buller DB, Burgoon JK. Interpersonal detection theory. Commun Theory 1996;6(3):203–42.

Caron F, Vanthienen J, Baesens B. Comprehensive rule-based compliance checking and risk management with process mining. Decis. Support. Syst. 2013;54(3):1357–69.

Chapman P, Clinton J, Kerber R, Khabaza T, Reinartz T, Shearer C, et al. CRISP-DM 1.0. Copenhagen: CRISP Consortium; 2000.

Coderre DG. Computer-aided fraud prevention and detection: A step-by-step guide. Hoboken, NJ: John Wiley & Sons; 2009.

Colantonio A, Di Pietro R, Ocello A, Verde NV. A new role mining framework to elicit business roles and to mitigate enterprise risk. Decis. Support. Syst. 2011;50(4):715–31.

Cox E. Fuzzy modeling and genetic algorithms for data mining and exploration. San Francisco: Morgan Kaufmann; 2005.

Debreceny RS, Gray GL. Data mining journal entries for fraud detection: An exploratory study. Int. J. Account. Inf. Syst. 2010;11(3):157–81.

Debreceny RS, Gray GL. Data mining of electronic mail and auditing: A research agenda. J. Inf. Syst. 2011;25(2):195–226.

Delen D, Al-Hawamdeh S. A holistic framework for knowledge discovery and management. Commun. ACM 2009;52(6):141–5.

Elsas P. X-raying segregation of duties: Support to illuminate an enterprise's immunity to solo-fraud. Int. J. Account. Inf. Syst. 2008;9(2):82–93.

Eppler MJ, Mengis J. The concept of information overload: A review of literature from organization science, accounting, marketing, MIS, and related disciplines. Inf. Soc. 2004;20(5):325–44.

Fanning K, Cogger K. Neural network detection of management fraud using published financial data. Int J Intell Syst Account Finance Manag 1998;7(1):21–41.

FERF. Data mining with XBRL. Morristown: NJ Financial Executives Research Foundation; 2013.

Feroz E, Kwon T, Pastena V, Park K. The efficacy of red flags in predicting the SEC's targets: An artificial neural networks approach. Int J Intell Syst Account Finance Manag 2000;9(3):145–57.

Gao L, Srivastava RP. The decomposition of management fraud schemes: Analyses and implications. Indian Account Rev 2011;15(1):1–23.

Glancy FH, Yadav SB. A computational model for financial reporting fraud detection. Decis. Support. Syst. 2011;50(3):595–601.

Green BP, Choi JH. Assessing the risk of management fraud through neural network technology. Audit. J. Pract. Theory 1997;16(1):14–28.

Han J, Kamber M. Data mining: Concepts and techniques. 2nd ed. San Francisco: Morgan Kaufmann; 2006.

Humpherys SL, Moffitt KC, Burns MB, Burgoon JK, Felix WF. Identification of fraudulent financial statements using linguistic credibility analysis. Decis. Support. Syst. 2011;50(3):585–94.

Jans M, Lybaert N, Vanhoof K. Internal fraud risk reduction: Results of a data mining case study. Int. J. Account. Inf. Syst. 2010;11(1):17–41.

Jans M, Alles M, Vasarhelyi M. The case for process mining in auditing: Sources of value added and areas of application. Int. J. Account. Inf. Syst. 2013;14(1):1–20.

Kohonen T. Self-organizing maps. 3rd ed. Heidelberg: Springer; 2000.

Lewis CM. Risk modeling at the SEC: The accounting quality model. Securities and Exchange Commission; 2012 [13 December 2012 [cited 24 April 2013]. Available from http://www.sec.gov/news/speech/2012/spch121312cml.htm].

McCornack SA. Information manipulation theory. Commun. Monogr. 1992;59:1–16.

PCAOB. AU section 316—Consideration of fraud in a financial statement audit. Public Company Accounting Oversight Board; 2002 [URL, cited 10 October 2013. Available from http://pcaobus.org/standards/auditing/pages/au316.aspx].

PCAOB. Consideration of fraud in a financial statement audit—Interim statements on auditing standards (SAS) no. 99. Washington, DC: Public Company Accounting Oversight Board; 2003.

Perols J. Financial statement fraud detection: An analysis of statistical and machine learning algorithms. Auditing 2011;30(2):19–50.

POB. Report and recommendations of the panel on audit effectiveness. New York: American Institute of Certified Public Accountants, Public Oversight Board; 2000.

Ravisankar P, Ravi V, Raghava Rao G, Bose I. Detection of financial statement fraud and feature selection using data mining techniques. Decis. Support. Syst. 2011;50(2):491–500.

Rohman L, Berg J. Q&A with an expert: The SEC is developing tools that use XBRL data to discover accounting anomalies and improve financial disclosures. Merrill Corporation; 2013 [16 April 2013 [cited 24 April 2013]. Available from http://goo.gl/gS2Ya].

Scott J. Social network analysis. Thousand Oaks, CA: Sage Publications; 2013.

Siciliano R, Conversano C. Decision tree induction. In: Wang J, editor. Encyclopedia of data warehousing and mining: I–Z. Hershey, PA: Idea Group Inc.; 2006. p. 353–8.

Smith KA. Neural networks for prediction and classification. In: Wang J, editor. Encyclopedia of data warehousing and mining: I–Z. Hershey, PA: Idea Group Inc.; 2006. p. 865–9.

Srivastava A, Sahami M. Text mining: Classification, clustering, and applications. Boca Raton, FL: CRC Press; 2009.

Please cite this article as: Gray GL, Debreceny RS, A taxonomy to guide research on the application of data mining to fraud detectio..., Int J Account Inf Syst (2014), http://dx.doi.org/10.1016/j.accinf.2014.05.006

Steinwart I, Christmann A. Support vector machines. New York, NY: Springer Science; 2008.
Titera WR. Updating audit standards — Enabling audit data analysis. J. Inf. Syst. 2013;27(1).
Van der Aalst WMP. Process mining: Discovery, conformance and enhancement of business processes. 1st. ed. New York: Springer; 2011.
van der Aalst W, van Hee K, van der Werf JM, Kumar A, Verdonk M. Conceptual model for online auditing. Decis. Support. Syst. 2011; 50(3):636–47.
Weiss SM. Fundamentals of predictive text mining. New York: Springer; 2010.
Witten IH. Text mining. In: Singh MP, editor. Practical handbook of internet computing. Boca Raton, FL: Chapman & Hall/CRC Press; 2005. p. 14-11–22.
Worrell J, Wasko M, Johnston A. Social network analysis in accounting information systems research. Int. J. Account. Inf. Syst. 2013; 14(2):127–37.
Zhang GP. Neural networks for classification: A survey. IEEE Trans Syst Man Cybern 2000;30(4):451–62.
Zhang L, Pawlicki AR, McQuilken D, Titera WR. The AICPA assurance services executive committee emerging assurance technologies task force: The Audit Data Standards (ADS) initiative. J. Inf. Syst. 2012;26(1):199–205.