Big data techniques in auditing research and practice: Current trends and future opportunities

Gepp, Adrian; Linnenluecke, Martina K; O'Neill, Terence J

Published in:
Journal of Accounting Literature

DOI:
10.1016/j.acclit.2017.05.003

Accepted for publication in Journal of Accounting Literature.

BIG DATA TECHNIQUES IN AUDITING RESEARCH AND PRACTICE: CURRENT TRENDS AND FUTURE OPPORTUNITIES

Abstract

This paper analyzes the use of big data techniques in auditing, and finds that the practice is not as widespread as it is in other related fields. We first introduce contemporary big data techniques to promote understanding of their potential application. Next, we review existing research on big data in accounting and finance. In addition to auditing, our analysis shows that existing research extends across three other genealogies: financial distress modelling, financial fraud modelling, and stock market prediction and quantitative modelling. Auditing is lagging behind the other research streams in the use of valuable big data techniques. A possible explanation is that auditors are reluctant to use techniques that are far ahead of those adopted by their clients, but we refute this argument. We call for more research and a greater alignment to practice. We also outline future opportunities for auditing in the context of real-time information and in collaborative platforms and peer-to-peer marketplaces.

Keywords

Auditing; Big Data; Data Analytics; Statistical Techniques

[Graphical Abstract provided in separate file]

[Bullet-Point Highlights provided in separate file]

Accepted for publication in Journal of Accounting Literature.

1. INTRODUCTION

This paper analyzes the use of big data techniques in auditing, and finds that the practice is not as widespread as it is in other related fields. We first introduce contemporary big data techniques and their origins in the multivariate statistical literature to help unfamiliar auditors understand the techniques. We then review existing research on big data in accounting and finance to ascertain the state of the field. Our analysis shows that – in addition to auditing – existing research on big data in accounting and finance extends across three other genealogies: (1) financial distress modelling, (2) financial fraud modelling, and (3) stock market prediction and quantitative modelling. Compared to the other three research streams, auditing is lagging behind in the use of valuable big data techniques. Anecdotal evidence from audit partners indicates that some leading firms have started to adopt big data techniques in practice; nevertheless, our literature review reveals a general consensus that big data is underutilized in auditing. A possible explanation for this trend is that auditors are reluctant to use techniques and technology that are far ahead of those adopted by their client firms (Alles, 2015). Nonetheless, the lack of progress in implementing big data techniques into auditing practice remains surprising, given that early use of random sampling auditing techniques put auditors well ahead of the practices of their client firms.

This paper contributes to bridging the gap between audit research and practice in the area of big data. We make the important point that big data techniques can be a valuable addition to the audit profession, in particular when rigorous analytical procedures are combined with audit techniques and expert judgement. Other papers have looked at the implications of clients' growing use of big data (Appelbaum, Kogan, & Vasarhelyi, in press) and the sources of useful big data for auditing (e.g., Vasarhelyi, Kogan, and Tuttle (2015);

Zhang et al. (2015)); our work focuses more on valuable opportunities to use contemporary big data techniques in auditing. We contribute to three research questions regarding the use of big data in auditing, raised by Appelbaum et al. (in press) and Vasarhelyi et al. (2015): "What models can be used?", "Which of these methods are the most promising?" and "What will be the algorithms of prioritization?" We provide key information about the main big data techniques to assist researchers and practitioners understand when to apply them. We also call for more research to further align theory and practice in this area; for instance, to better understand the application of big data techniques in auditing and to investigate the actual usage of big data techniques across the auditing profession as a whole.

This paper also integrates research in big data across the fields of accounting and finance. We reveal future opportunities to use big data in auditing by analyzing research conducted in related fields that have been more willing to embrace big data techniques. We offer general suggestions about combining multiple big data models with expert judgement, and we specifically recommend that the audit profession make greater use of contemporary big data models to predict financial distress and detect financial fraud.

The paper proceeds as follows. Section 2 introduces big data techniques, including their origin in the multivariate statistical literature and relates it to the modern mathematical statistics literature. Section 3 offers a systematic literature review of existing research on big data in accounting and finance. This section highlights how auditing substantially differs from the other major research streams. Section 4 identifies novel future research directions for using big data in auditing. Finally, Section 5 concludes the paper with important recommendations for the use of big data in auditing in the 21st century and a call for further research.

2. AN INTRODUCTION TO BIG DATA TECHNIQUES

This section presents an overview of big data and big data techniques to promote a greater understanding of their potential application. Auditors that use more advanced techniques need to understand them (Appelbaum et al., in press). An introduction to big data provides the necessary background to present the main big data techniques available and the key information needed to determine which are appropriate in a given circumstance. Appendix A describes the main big data techniques, summarizes their key features and provides suggested references for readers who want more information.

Big data refers to structured or unstructured data sets that are commonly described according to the four Vs: Volume, Variety, Velocity, and Veracity. Volume refers to data sets that are so large that traditional tools are inadequate. Variety reflects different data formats, such as quantitative, text-based, and mixed forms, as well as images, video, and other formats. Velocity measures the frequency at which new data becomes available, which is increasingly often at a very rapid rate. Finally, the quality and relevance of the data can change dramatically over time, which is described as its veracity. The auditing profession has a large and growing volume of data available to it, of increasing variety and veracity. Textual information obtained online is one new type of data, and we discuss this phenomenon later in the paper. Auditors also face an increasing velocity of data, particularly in the context of real-time information, and this is described in Section 4.

Big data comes in a variety of flavors – “small p, large n”, “large p, small n”, and “large p, large n”, where n refers to the number of responses and p the number of variables measured at each response. These categorizations are important because they can influence which technique is the most suitable. The big data techniques described in Appendix A are

suited to different categorizations; for instance, Random Forests[1] is particularly useful for "large p, small n" problems. High-frequency trading generates massive data sets of both high volume and high velocity, creating major challenges for data analysis. Nevertheless, such "small p, large n" problems are perhaps the easiest of the three scenarios and the analytic tools used are, in the main, adaptations of existing statistical techniques. The "large p, small n" scenario is best exemplified by genomics. A single human genome contains about 100 gigabytes of data. Essentially the data is a very long narrow matrix with each column corresponding to an individual and each row corresponding to a gene. The cost of sequencing a genome has now fallen to a point where it is possible for individuals to purchase their own genome. As a consequence, genomics is rapidly transitioning to the "large p, large n" scenario. Climate change research is another example of science at the forefront of the big data "large p, large n" scenario, with multivariate time-series collected from a world-wide grid of sites over very long time frames.

Big data also refers to the techniques and technology used to draw inferences from the variety of flavors of data. These techniques often seek to infer non-linear relationships and causal effects from data which is often very sparse in information. Given the nature of the data, these techniques often have no or very limited distributional assumptions. Computer scientists approach big data from the point of view of uncovering patterns in the complete record – this is often called the algorithmic approach. The patterns are regarded as approximations of the complexity of the data set. By comparison, statisticians are more inclined to treat the data as observations of an underlying process and to extract information and make inferences about the underlying process.

[1] Random Forests for regression-type problems uses bootstrap samples to develop multiple decision trees (usually thousands) and then aggregates them together by averaging. See Appendix A for more information.

The statistical techniques used in big data necessitate more flexible models, since highly structured traditional regression models are very unlikely to fit big data well. Furthermore, the volume (as well as variety and velocity) of big data is such that it is not feasible to uncover the appropriate structure for models in many cases. The popularity of more flexible approaches dates back to Efron's (1979) introduction of the bootstrap at a time when increasing computer power made such new techniques feasible. The bootstrap is a widely applicable statistical tool that is often used to provide accuracy estimates, such as standard errors that can be used to produce confidence intervals. Regularization is another widely used technique which imposes a complexity penalty that shrinks estimated parameters towards zero to prevent over-fitting or to solve ill-posed problems. Ridge regression, which uses a L2 penalty[2], was initially proposed by Hoerl and Kennard (1970) in the 1970s; however, it has only become popular in recent decades with the advent of increased computing power. More recently, regularization techniques have become popular alternatives, such as LARS (least angle regression and shrinkage) proposed by Bradley Efron, Hastie, Johnstone, and Tibshirani (2004) and Tibshirani's (1996) Lasso (least absolute shrinkage and selection operator) which uses an L1 penalty[3]. The use of an L1 penalty is important because it is very effective in variable reduction and so results in sparse models that are easier to interpret. These simpler models are often easier to communicate to clients. Penalties that are a mixture of L1 and L2 are also available (Friedman, Hastie, & Tibshirani, 2010); indeed, contemporary statistics scholars continue to investigate new penalties for regularization.

Supervised learning develops explanatory or predictive models from data with known outcomes to apply to data with unknown outcomes. Some popular ways to conduct

[2] A L2 penalty penalises a model for complexity based on the sum of all the squared coefficients.
[3] An L1 penalty uses the absolute value of coefficients rather than the squared coefficients used in L2 penalties.

supervised learning include artificial neural networks, classification and regression trees (decision trees), Random Forests, Naïve Bayes, regularized regression[4] (as mentioned above), support vector machines, and multivariate adaptive regression splines (MARS). In contrast, unsupervised learning seeks to uncover patterns in unlabeled data. Popular methods are unsupervised neural networks, latent variable models, association rules, and cluster analysis. Machine learning is an overarching term that encompasses both supervised and unsupervised learning. The techniques mentioned in this paragraph are briefly described in Appendix A.

3. THE USE OF BIG DATA IN ACCOUNTING AND FINANCE RESEARCH

This paper offers a systematic literature review of the use of big data techniques in auditing research and practice and follows methodical steps for collecting data to arrive at a comprehensive data set of articles to include in the review. First, we searched the Social Sciences Citation Index for 'big data' papers, searching for articles that contained the key words "big data" or "analytics" or "data mining" in the title, abstract, or keywords. To ensure that the search was not too broad, we limited the search to articles that also contained the keywords "accounting" or "financ*" in the title, abstract, or keywords. Our search identified a total of 286 records as of November 2016. Next, we screened the resulting articles to only retain those of interest to the current research. This reduced the original article base to 45 records. Excluded articles discussed other big data and quantitative applications in the context of business decision-making (e.g., improving customer retention in financial services, see Benoit and Van den Poel (2012)). Next, we conducted further searches via cited references and Google Scholar to manually add another 47 articles into the data set. The articles were then assessed by the author team and categorized according to their main

[4] Regularization is a general concept that can be applied to regression, but also commonly to the other models mentioned to help prevent over-fitting.

research focus. The analysis revealed four main genealogies, which we review below: (1) financial distress modelling, (2) financial fraud modelling, (3) stock market prediction and quantitative modelling, and (4) auditing. We find that there has been much progress in the first three fields, but that auditors have been slow to implement research findings into practice. We then proceed to address the lack of uptake of big data measures.

3.1 Financial Distress Modelling

Papers in the financial distress modelling stream use data mining techniques to detect and forecast the financial distress (or financial failure) of companies and these techniques are also of interest to auditors to assist with their going concern evaluations.

Multiple studies have used decision tree based models. Sun and Li (2008) apply data mining techniques based on decision trees in order to predict financial distress. Starting with 35 financial ratios and 135 listed company-pairs, the researchers design and test a prediction model to show theoretical feasibility and practical effectiveness. Koyuncugil and Ozgulbas (2012b) use data mining methods to design a financial distress early warning system for small- to medium-sized enterprises. They test the model on over 7,000 small- to medium-sized enterprises and develop a number of risk profiles, risk indicators, early warning systems, and financial road maps that can be used for mitigating financial risk. Similar work has also been undertaken by Koyuncugil and Ozgulbas (2012a) and Kim and Upneja (2014). Li, Sun, and Wu (2010) use classification and regression tree methods to estimate financial distress and failure for a sample of Chinese listed companies, while Gepp, Kumar, and Bhattacharya (2010) use US listed companies.

Chen and Du (2009) propose a different approach and apply data mining techniques in the form of neural networks to build and test financial distress prediction models. Using 37 ratios across 68 listed companies, they demonstrate the feasibility and validity of their

modelling. Additional research supports their approach and suggests that neural networks perform better for financial distress modelling than decision trees and alternative approaches such as support vector machines (Geng, Bose, & Chen, 2015).

Zhou, Lu, and Fujita (2015) compare the performance of financial distress prediction models based on big data analytics versus prediction models based on predetermined models from domain professionals in accounting and finance. They find that there is no significant difference in the predictions. However, a combination of both approaches performs significantly better than each on its own (Zhou et al., 2015). Lin and McClean (2001) also find that a hybrid approach of professional judgement and data mining produces more accurate predictions. Kim and Han (2003) go one step further and argue that analyses should incorporate qualitative data mining approaches to elicit and represent expert knowledge about bankruptcy predictions from data sets such as loan management databases.

The literature recognises that financial distress might not be limited to a company, but may also extend to corporate stakeholders. Khandani, Kim, and Lo (2010) use machine learning techniques to construct models of consumer credit risk at the level of the individual and the customer, rather than the corporation. They combine customer transactions and credit bureau data and are able to use machine learning to significantly improve classification rates on credit card default and delinquencies. Singh, Bozkaya, and Pentland (2015) were inspired by animal ecology studies to analyse the transactions of thousands of people; they found that individual financial outcomes are associated with spatio-temporal traits (e.g., exploration and exploitation) and that these traits are over 30% better at predicting future financial difficulties than comparable demographic models.

Auditors could harness big data techniques and methods for forecasting financial distress and, combined with their professional judgement, be better able to judge the future

financial viability of a firm. This would improve the going concern evaluations required in audits by the Statement on Auditing Standards, No. 59 for public companies (AICPA, 1988). Incorporating big data models should help avoid the costly error of issuing an unmodified opinion prior to bankruptcy. Read and Yezegel (2016) found that this problem is particularly apparent in non-Big 4 firms within the first five years of an audit engagement. The authors do not offer an underlying reason, but it may be that smaller audit firms are reluctant to issue modified going concern opinions early in an engagement for fear of losing clients. If this is the case, then smaller audit firms may be better able to justify modified opinions to their clients by presenting them with objective results from big data models, and thereby increasing the independence of the going concern evaluations. The use of these models also represents an opportunity to increase the efficiency of the going concern evaluation part of the audit, notwithstanding the initial overhead cost of becoming familiar with big data models and techniques.

Although it is likely that the focus will be on one-year predictions that relate to going concern opinions, financial distress models could also be used for longer forecasts. These longer forecasts could be used by internal auditors who tend to have longer time-horizons than external auditors. Financial distress models that are supplemented by the opinion of the internal audit team as to the veracity of the forecasts could provide valuable information for senior management and the Board of Directors. Longer range forecasts and opinions give management more time to make strategic changes to minimize the likelihood that predicted financial distress will occur.

3.2 Financial Fraud Modelling

A second major research stream centers on modelling financial fraud, which can help auditors assess the risk of fraud (Bell & Carcello, 2000) when conducting fraud risk

assessments. Section 200 of the Statement on Auditing Standards No. 122/123 requires that external auditors "obtain reasonable assurance about whether the financial statements as a whole are free from material misstatement, whether due to fraud or error" (AICPA, 2011). By adopting contemporary big data models, auditors could provide this assurance, notwithstanding the current debate as to the exact meaning of "reasonable assurance" (Hogan, Rezaee, Riley, & Velury, 2008).

Financial fraud is a substantial concern for organizations and economies around the world.[5] The Association of Certified Fraud Examiners (2016) estimates that the typical organization loses 5% of revenue each year to fraud. Applying this to the Gross World Product for 2014, global fraud loss amounts to nearly 4 trillion US dollars. These numbers have prompted researchers to consider the application of big data techniques to fraud detection, prediction, and prevention. For instance, R. Chang et al. (2008) suggest using visual data analytics to interactively examine millions of bank wire transactions—they argue that this approach is both feasible and effective. In contrast, Abbasi, Albrecht, Vance, and Hansen (2012) model financial fraud using meta-leaning, which is a specialized form of machine learning that combines the outputs of multiple machine learning techniques in a self-adaptive way to improve accuracy. They find the method to be more effective than other single approaches.

Other approaches use supervised neural networks (Green & Choi, 1997; Krambia-Kapardis, Christodoulou, & Agathocleous, 2010) or unsupervised neural networks based on a growing hierarchical self-organizing map (e.g., Huang, Tsaih, and Lin (2014); Huang, Tsaih, and Yu (2014)) to build a financial fraud detection model. The approach proposed by Huang,

[5] An excellent review of financial fraud modelling using big data techniques is also provided by Ngai, Hu, Wong, Chen, and Sun (2011) who offer a classification framework for the existing literature. West and Bhattacharya (2016) also review computational intelligence-based approaches, such as neural networks and support vector machines.

Tsaih, and Lin (2014) involves three stages: first, selecting statistically significant variables; second, clustering into small sub-groups based on the significant variables; and third, using principal component analysis to reveal the key features of each sub-group. Huang, Tsaih, and Yu (2014) apply this model to 144 listed firms and find that the approach can effectively detect fraudulent activity. Ravisankar, Ravi, Rao, and Bose (2011) use neural networks, support vector machines, and genetic programming to identify firms engaging in financial fraud. They find that probabilistic neural networks and genetic programming outperform other methods and are similarly accurate. Building on the work of Busta and Weinberg (1998), Bhattacharya, Xu, and Kumar (2011) proposed a genetic algorithm to optimize a neural network based on Benford's Law. They used simulated data to conclude that their algorithm showed promise for detecting fraud in financial statements. Meanwhile, Kirkos, Spathis, and Manolopoulos (2007) found a Bayesian network that outperformed an artificial neural network, as well as a decision tree. A support vector machine developed using the output from principal components analysis has also been studied (Sadasivam, Subrahmanyam, Himachalam, Pinnamaneni, & Lakshme, 2016).

The best approach to financial fraud modelling is heavily debated. C. C. Lin, Chiu, Huang, and Yen (2015) compare the differences between data mining approaches and the judgement of experts, and find that neural networks and decision trees achieve a correct classification rate of over 90% on a holdout sample. The judgement of experts is shown to be more consistent with the decision tree approach. Perols (2011) reviews the performance of popular statistical and machine learning techniques and finds that logistic regression and support vector machines perform well relative to competing models such as neural networks and decision trees. Given that these papers come to opposing conclusions, there is clearly uncertainty in the field. Chen (2016) constructs a financial statement fraud model using a

two-stage process which appears to offer advantages over the one-step approach used in Ravisankar et al. (2011) and Perols (2011). The first stage involves selecting the major variables using two decision tree algorithms: classification and regression tree (CART) and Chi-squared automatic interaction detector (CHAID). The second stage constructs the financial fraud model using the variables from stage one. The second stage uses a number of approaches including the two approaches from stage one, as well as Bayesian belief network, support vector machines, and neural networks. Chen (2016) finds that the combination of CHAID in stage one and CART in stage two proves to be the most accurate methodology for detecting financial statement fraud. Zhou and Kapoor (2011) concur that a combination of professional judgement and big data techniques provides a more effective and efficient approach.

There has also been research into the process of analyzing financial statement text for the purposes of detecting fraud, which is well summarized by Gray and Debreceny (2014). More recently, Purda and Skillicorn (2015) developed a language-based tool that relies on data to identify important indicators of fraud (see also Van Den Bogaerd and Aerts (2011)). The language-based tool has an initial training period which uses a decision tree approach to analyze reports of known fraud firms and obtain a rank order list of words best able to distinguish fraud versus non-fraud. The second stage uses vector order machines to predict the fraud status of financial reports and assign a truth probability. The approach is able to generate correct classification rates of over 80%.

The above review shows that studies have used big data techniques to model the occurrence of financial fraud as a binary dependent variable, which implicitly treats all fraud as equal. Even though the cost of financial fraud varies greatly between cases and has obvious economic implications, very few studies have modelled the cost of financial fraud.

There is also an opportunity for fraud models to take advantage of the fact that collusion between multiple offenders often occurs in fraud cases. Free and Murphy (2015) conclude that that the social nature of fraud may assist in identifying distinctive features. These features could be incorporated into fraud models to improve their accuracy.

External auditors can improve their fraud risk assessments by using big data financial fraud models that advance standard regression models, such as the well-known F-score fraud model based on logistic regression (Dechow, Ge, Larson, & Sloan, 2011). These big data financial fraud models are developed using data from past frauds. They offer valuable information to auditors because past research has revealed that auditors often have little real experience of fraud (Humpherys, Moffitt, Burns, Burgoon, & Felix, 2011). Nevertheless, auditors tend to be reluctant to rely on decision aids to detect fraud (Eining, Jones, & Loebbecke, 1997), so there is an opportunity for future research to investigate how to best use big data fraud models in conjunction with auditor expertise. This research topic also encompasses how to best present the analysis and output from big data models to auditors. Hogan et al. (2008) also called for future research into incorporating more sophisticated fraud models into audits. This is particularly relevant because big data models offer different information than the more familiar and traditional regression models (such as the F-score model).

Internal auditors could also use these models to draw attention to situations that require investigation. Forensic accountants and forensic auditors could also use these models to determine the probability of fraud having occurred, in order to provide initial corroboration.

3.3 Stock Market Prediction and Quantitative Modelling

In addition to the two research streams outlined above, a third stream is focused on stock market predictions and other quantitative modelling. This stream of research is particularly interested in predictive analysis and providing investment advice to managers and investors. Although this stream is not directly relevant to auditing, relevant lessons will be uncovered from the ways in which big data techniques are applied in this area.

Chun and Kim (2004) use neural networks and case-based reasoning, and a choice of two markets and a choice of passive or active trading strategy, to generate financial predictions substantially in excess of buy-and-hold returns. Lam (2004) employs neural networks and predicts market returns using financial ratios and macroeconomic variables. Chun and Park (2006) later find that a hybrid model further outperforms a pure case-based reasoning approach in predicting a stock market index, although the result is not statistically significant. Equity portfolios that outperform a benchmark index portfolio have also been constructed using popularity in Google searches (Kristoufek, 2013) and changes in Google search queries (Preis, Moat, & Stanley, 2013). Guerard, Rachev, and Shao (2013) also study equity portfolio construction and Pachamanova and Fabozzi (2014) review other studies on the topic. In addition, Zhang et al. (2015) use a genetic algorithm-based model to generate stock trading rules (quantitative investment), which outperforms both a decision tree and a Bayesian network.

Curme, Preis, Stanley, and Moat (2014) find that an increase in Google and Wikipedia searches on politics or business are related to subsequent stock market falls. Li, Ma, Wang, and Zhang (2015) use the Google search volume index as a measure of investor attention and find a significant association between the search index and trader positions and future crude oil prices. Adopting a different approach, Sun, Shen, and Cheng (2014) use individual stock

transaction data to create a trading network to characterize the trading behaviour of stocks investors. They show that trading networks can be used to predict individual stock returns. Shapira, Berman, and Ben-Jacob (2014) model the stock market as a network of many investors, while Gui, Li, Cao, and Li (2014) model it as a network of communities of stocks.

Many studies have analysed news articles in order to make stock market predictions. Tetlock (2007) uses daily content from a popular *Wall Street Journal* column and finds that when media pessimism is high stock prices decline but then return to fundamentals. Additionally, unusually high or low media pessimism helps predict high trading volume. Alanyali, Moat, and Preis (2013) find the daily number of mentions of a stock in the *Financial Times* is positively correlated with daily volume, both before and on the day of the news release. Piskorec et al. (2014) construct a news cohesiveness index based on online financial news and show that this is correlated with and driven by volatility in financial markets. Research has also examined the sentiment of news articles (Smales, 2014a, 2014b, 2015). Jensen, Ahire, and Malhotra (2013) find a significant association between firm-specific news sentiment and intraday volatility persistence, especially for bad news. Nardo, Petracco-Giudici, and Naltsidis (2016) review the literature and conclude that while there is merit in using online news to predict changes in financial markets, the gains from implementing such an approach are usually less than 5%. However, Ranco et al. (2016) find substantial benefit in coupling news sentiment with web browsing data. Some studies (Dhar, 2014; Kao, Shyu, & Huang, 2015; Zheludev, Smith, & Aste, 2014) have also incorporated non-traditional online sources of information such as social media, blogs, and forums, and proposed many questions for future research.

Other examples of quantitative modelling include: service architecture for capital market systems management (Rabhi & Benatallah, 2002); managing metadata in financial

analytics software (Flood, 2009); identifying successful initial public offerings (Martens et al., 2011); high-frequency financial data mining (Sun & Meinl, 2012); identifying drivers of firm value (Kuzey, Uyar, & Delen, 2014); sentiment analysis for predicting economic variables (Levenberg, Pulman, Moilanen, Simpson, & Roberts, 2014); volatility of returns (Sun, Chen, & Yu, 2015); option pricing (Thulasiram, Thulasiraman, Prasain, & Jha, 2016; Xiao, Ma, Li, & Mukhopadhyay, 2016); and market basket analysis (Videla-Cavieres & Rios, 2014), which is the identification of sets of products or services that are sold together.

Quantitative modelling and stock market prediction, particularly that which uses online textual information and sentiment analysis, is an active area of research that is leveraging the usefulness of big data techniques. This has been especially true in recent years; most of the articles mentioned above were published in or after 2013.

Big data sentiment analysis has potential applications in auditing. Negative sentiment appearing in online news, social media, and other online sources may influence a risk-based audit. For example, consistent negative sentiment about certain products might steer auditors to examine allowances for product returns or warranty claims. Online sentiment about a client might also influence an auditing firm's decision to accept or continue an engagement.

Conducting a sentiment analysis of company emails might help an auditor understand the company under review and reveal areas at higher risk of fraud. For instance, inconsistent email sentiment within a business unit could indicate internal disharmony and signal that internal controls have been breached or that fraud has occurred. When email sentiment at the senior management level of an organization is positive, but turns to negative at lower levels, this may signal that employees are aware of and unhappy that management has committed control breaches (or fraud). Similarly, an auditor might be encouraged to look more closely at a business unit that presents a profile of email sentiment that is inconsistent with that of the

rest of the company. Sentiment analysis focused on the co-occurrence of words and social networks could also be used to search for collusive parties in internal or forensic audit investigations. These are a few examples of how auditors could benefit from sentiment analysis, and could be the subject of a thorough cost-benefit analysis in future research.

Other potential uses for sentiment analysis in auditing might be discovered by studying its application in other domains. Ravi and Ravi (2015) review a study that analysed Enron emails (Mohammad, 2012) to reveal marked differences by gender in the use of emotional words, particularly those about trust. Would knowledge of the pattern of use, and any outliers, help an audit team understand its client and the risks it faces when planning an audit? Additionally, would the outliers in email usage assist internal auditors to identify risks such as compliance or control breaches and unauthorised activities?

Sentiment analysis is also an opportunity to add value to the audit service (external or internal) with novel and valuable information, such as providing clients with a list of their business units, ranked by employee sentiment.

3.4 Auditing

Given the well-developed literature on financial distress, financial fraud modelling, and stock market prediction, it is surprising that the auditing profession has been slow to adopt big data techniques. Anecdotal evidence from partners at some leading audit firms indicates they have begun to use big data, but the true extent of its use in practice is unknown and would be the subject of valuable future research. Many scholars have lamented the lack of big data in auditing (e.g., Acito and Khatri (2014); Alles (2015); Brown-Liburd, Issa, and Lombardi (2015); Cao, Chychyla, and Stewart (2015); Earley (2015); Griffin and Wright (2015); Krahel and Titera (2015); Werner and Gehrke (2015); Zhang, Yang, and Appelbaum (2015)). Earley (2015) acknowledges that big data could be a game-changer in auditing, and

Schneider, Dai, Janvrin, Ajayi, and Raschke (2015) predict that data analytics will significantly change the way auditors work. Cao et al. (2015) contend that big data can improve financial statement audits. Furthermore, Griffin and Wright (2015) refer to the slow uptake of big data as possibly the greatest risk in the field and call for it to be more widely used in practice, education, and research.

Alles (2015) argues that, to maintain credibility, auditors need to be aligned with the practices of their clients. However, the argument for auditors to only use big data once their clients embrace it is not on a sound footing; indeed, auditors' early use of random sampling techniques has already put them ahead of client firms. Furthermore, as data-driven approaches become more prevalent, audit clients are likely to view the use of big data techniques as commonplace. In fact, it is already happening in some places; the International Auditing and Assurance Standards Board has stated that clients in some regions are enquiring more about the use of data analytics, and in some cases are already expecting to see it used in audits (IAASB, 2016). Appelbaum et al. (in press) identify a growing use of big data by audit clients, which they link to an urgency for auditors to follow suit.

Krahel and Titera (2015) argue that accounting and auditing standards have not kept up with technological change and still emphasize presentation, aggregation, and sampling. On the other hand, big data enables auditors to analyze the processes that generate data, including full population testing, which adds value to the auditing and accounting profession and to the clients for whom they work. The call for a change in standards is also taken up by Moffitt and Vasarhelyi (2013), Vasarhelyi et al. (2015) and Appelbaum et al. (in press), who point out that practitioners, academics, and students would all benefit from knowing more about big data.

Brown-Liburd et al. (2015) examine the behavioral effects of big data on auditor judgement, and discuss issues such as information overload, information relevance, pattern recognition, and ambiguity. They conclude that adding big data techniques to the set of tools used in the audit process would add value. They also note that it is important to use the technique and data set most appropriate to each circumstance, which points to the need for more research in this area. Yoon, Hoogduin, and Zhang (2015) also argue that big data offers a complementary source of evidence for the audit function, and that its use should be evaluated according to the audit evidence criteria frameworks of sufficiency, reliability, and relevance. Moffitt and Vasarhelyi (2013) also support the use of big data in new forms of audit evidence.

In addition to financial distress modelling and financial fraud modelling, big data offers many other advantages to the audit profession. Process mining, which analyses the event logs of business systems (Jans, Alles, & Vasarhelyi, 2014), has been shown to improve audit results when tested on real world data sets (Werner & Gehrke, 2015). Big data video, audio, and textual information processing can also improve accounting and auditing functions (Crawley & Wahlen, 2014; Warren, Moffitt, & Byrnes, 2015). For instance, in addition to verifying transactions against invoices and receipts, auditors could also use non-traditional information such as photos, videos, and GPS location (Moffitt & Vasarhelyi, 2013).

Overall, Hagel (2013) and Smith (2015) make a case for accountants and auditors to ‘own’ big data, not just because it provides better information, but because doing so will help move the profession up the value chain to become a true business partner, rather than a transactional service provider. Examples of how auditors could use financial distress models and sentiment analysis to contribute to this aim have been provided in previous sections.

4. DISCUSSION AND NOVEL RESEARCH DIRECTIONS

R. M. Chang, Kauffman, and Kwon (2014) argue that there has been a paradigm shift in the research questions that can be asked and the research methods that can be used. They argue that social networks, blogs, political discourse, company announcements, digital journalism, mobile phones, home entertainment, online gaming, online financial services, online shopping, social advertising, and social commerce are just some of the new contexts in which research questions can be examined. This context, and big data analytic tools, provide researchers with opportunities to do frequent, controlled, and meaningful research on real world issues. S. H. Kim (2000) also sees a paradigm shift, with big data offering the opportunity to harvest an ocean of online data, filter information, and generate new knowledge. D. S. Zhang and Zhou (2004) see big data as the way to find the 'golden nugget'. Amoore (2011, p. 24) poetically describes the paradigm shift as 'the analytic of the data derivative – a visualized risk flag or score drawn from an amalgam of disaggregated fragments of data, inferred across the gaps between data and projected onto an array of uncertain futures'.

It is clear that big data techniques represent a valuable opportunity for the auditing profession. However, this opportunity has not yet been capitalized on to the degree it has in related areas. As previously mentioned, auditing would benefit from adopting modern big data models to predict financial distress and detect financial fraud. Updated standards may help overcome the auditing profession's apparent reluctance to engage with big data techniques. There is no doubt that having access to frequently updated big data sets that incorporate non-traditional information would be of great value to the audit function. As stated in Section 2, traditional tools are not adequate for analyzing big data, because it is so large, arrives so rapidly, and its variability or relevance changes dramatically over time. It is

also known that auditors can have difficulty integrating multiple pieces of evidence in some circumstances (Moeckel, 1991), while big data techniques excel at integrating diverse pieces of information into decision aids. Hence, the big data techniques listed in Appendix A would be a valuable addition to the auditing profession and to audit research.

Big data techniques can also be applied to traditional, smaller data sets to gain additional insights. For example, Read and Yezegel (2016) use logistic regression to analyze the relationship between audit tenure and audit reporting. The authors use squared terms in the model to control for a potential nonlinear relationship, but this still imposes the constraint of a quadratic relationship. The use of a non-parametric big data technique, such as a decision tree or MARS (see Appendix A), could reveal the presence of any non-quadratic relationships. Furthermore, models produced using either of these techniques can be easily visualized, communicated and explained. Lennox and Kausar (2017) also use squared terms to consider potential nonlinearities in a supplementary analysis, but they also had to handle skewness in their data. However, decision tree models are unaffected by skewness and so this would not have been a concern for such models. A further example is Xu and Zhang (2009), who use a stepwise method to remove variables from their bankruptcy regression models because of highly correlated independent variables. An alternative would have been a Lasso regularized regression (see Appendix A), which has more flexibility to handle correlated independent variables. In addition to being able to exclude variables as done by stepwise methods, a Lasso regularized regression has the ability to shrink coefficients towards zero without removing them all together.

The non-auditing research streams reviewed above are more developed in their use of big data techniques and offer some important findings relevant to auditing.

1) Combining multiple techniques has been shown to outperform the use of a single technique (e.g. Abbasi et al. (2012); Chen (2016)).
2) Big data techniques are best used to complement, not replace, human experts (e.g. Zhou et al. (2015)). This could be an important argument for overcoming reluctance to use big data techniques.
3) Non-traditional sources of information such as text offer additional value (e.g., using online news to predict stock market movements). For instance, future research in auditing could benefit from advances in natural language processing (NLP), which is used to process and interpret natural language in context. A potential application is analyzing unstructured contracts in audits. Using the context of the text, NLP can be applied to automatically extract constructs such as company or person names, or key terms and conditions, which could then be analyzed using other big data techniques. For instance, a network of extracted names could be used to identify those that appear in multiple contracts. Each name could also be matched against email correspondence and then sentiment scores computed based on associated emails and online information. Models could then risk-sort contracts either purely based on anomalies in the data mentioned or by also incorporating expectations based on the auditor's knowledge of the particular engagement. NLP could also be used to advance fraud detection models that analyze text, from either emails (see Gray and Debreceny (2014)) or the Management Discussion & Analysis section of financial reports (Purda & Skillicorn, 2015). The NLP Group at Stanford University has made their CoreNLP software freely available[6]. This software can be applied to many different languages and can be tailored by training it on documents containing, for example, financial or legal language. This is important,

[6] See http://nlp.stanford.edu/software/.

because finance-specific language solutions have been shown to perform substantially better than general solutions when used in a finance context (Loughran & McDonald, 2011).

Other examples of future research directions include real-time accounting and financial information, and collaborative platforms and peer-to-peer marketplaces.

4.1 Real-Time Accounting and Financial Information

How would audits adapt in the face of a real-time information paradigm? People have become used to seeing their bank account information in real-time. The same sort of information could be provided by firms, superannuation funds, and governments. Big data techniques could allow financial information to be made available in real-time, instead of via traditional quarterly or annual reports. Real-time information also poses an important question about how to provide auditing and assurance services in such a setting. How do auditing and governance practices handle a system where new information is available well before a traditional audit can take place? Real-time auditing processes are required. The existing literature on continuous auditing (Chiu, Liu, & Vasarhelyi, 2014) refers to a continuous cycle of auditing; this work could be enhanced by big data techniques that are well-suited to quickly analyzing and adapting to new data. As mentioned in Appendix A, there are big data techniques that can automatically and computationally efficiently handle new data sets with characteristics such as missing values, or irrelevant or highly-correlated data. These are important features for real-time systems in which such data issues cannot be manually addressed.

Much has been written on the ‘user-unfriendliness’ of company financial reports, government budgets, and superannuation fund reports. Using big data tools, information that is collected in real-time could be displayed using state-of-the-art visualizations and

customized dashboards in a way that is more user-friendly than traditional financial reports. Furthermore, the tools could be set to display changes over time, not just a snapshot, and this may influence market participants to be less focused on the short-term. The issue would then become how these new visualizations and dashboards would be audited for the assertions of *existence, completeness, classification, and understandability*, and *accuracy and valuation.* Changing the way such information is presented will likely require substantial shifts in audit procedures, although practices relating to the *existence* assertion might remain similar.

Overall, real-time financial reporting to the public would necessitate a fundamental change for auditors, from providing assurances about numbers to assurances about real-time systems (that subsequently produce numbers). However, financial reporting to the public is a long way from being a reality. Corporations in many parts of the world still report less frequently than quarterly, including in Australia, New Zealand, and the United Kingdom. A sensible first step would be real-time financial reporting to senior management, who then might be more likely to support real-time reporting to the public. Robust research on the impacts of such a change would also help provide confidence during what would be a paradigm shift.

Real-time reporting to management still raises important questions for the financial statement audit. The information included on management's real-time dashboard (or other visualization) could be used by the auditor to better understand the company and its environment, how it is managed, and its potential risks. For example, an energy company's dashboard which includes substantial information about the derivatives market might indicate a high risk if the auditor discovers it is not predominantly for hedging purposes. In fact, that might have been the case for Enron, if real-time dash-boards had been available at that time. These visualizations could also improve the efficiency of the audit process. For example, a

dashboard that listed the age of each piece of inventory would help auditors substantiate inventory value. However, what tests would auditors need to conduct in order to be confident in the reliability of the dashboard? This question represents a shift towards providing assurances of systems, which, as mentioned above, would be needed for real-time reporting to the public. Thus, real-time reporting to management would likely also help auditors prepare for a potential move to real-time reporting to the public.

The concept of real-time information is not limited to auditing. For example, fraud modelling should take advantage of additional information by using big data techniques set up to automatically update as new data becomes available. There are already examples of data sources moving to real-time information. The Federal Reserve Bank of Chicago provides financial statement data for holding companies on a daily basis in a simple downloadable format, although no summary statistics or visualizations are available[7]. Daily updates incorporate any revisions or new information that become available between the traditional quarterly reports. Does this daily stream of information provide useful information for fraud detection models? Research should take advantage of this and other more frequently updated data.

4.2 Collaborative Platforms and Peer-to-Peer Marketplaces

Peer-to-peer marketplaces are changing the way business is done. Traditionally, firms made profits by standing in-between businesses and individuals wanting to sell and buy products and services, such as banking, insurance, employment, accommodation, and transport. The advent of big data means that buyers and sellers can be brought together via collaborative platforms. This eliminates the need for the middle broker. Insurance companies, banks, and other brokers who provide matching services represent some of the most

[7] See https://www.chicagofed.org/banking/financial-institution-reports/bhc-data.

profitable and successful business models; thus, the advent of peer-to-peer marketplaces has the potential to dramatically reshape the goods and services business landscape. Additionally, peer-to-peer marketplaces are not constrained to traditional (geographic) borders, which poses another line of research as well as different future data sources. For example, one of the most popular new ways to source accommodation is via Airbnb, which is a peer-to-peer marketplace that does not own any accommodation assets.

As is often the case with new technologies, including those facilitated by big data, peer-to-peer marketplaces also present challenges about how we think about audit and verification to ensure confidence in the marketplace. What information do market participants use to assess the reliability of their counterparts and their financials? How can this information be verified and what role can audits play in providing meaningful assurances to market participants? Testing controls could be very important, because participants likely expect that they are implemented by the software in a standardized manner. However, does the vast number of market participants mean that going concern evaluations and fraud risk assessments primarily become outputs from big data models for predicting financial distress and detecting fraud, respectively? There are many important questions such as these. Answering them will involve analyzing platforms and marketplaces which hold huge amounts of various types of data, much of which is changing in real time and does not involve primary documentation. Once again, big data techniques are well-suited to this analysis. One approach is to cross-reference information from multiple secondary sources to obtain a reasonable probability (assurance) of correctness, as is done in the Airbnb platform.

5. CONCLUSION AND FUTURE OPPORTUNITIES

This paper reviews research in accounting and finance concerning data analytics and big data in order to better understand the use of big data techniques in auditing. We first point out the origins of big data techniques in the multivariate statistical literature and then categorize big data accounting and finance research under several research groupings. Our analysis shows that, in addition to auditing, there are influential papers across financial distress modelling, financial fraud modelling, and stock market prediction and quantitative modelling. We review each of these streams of research to ascertain their main contributions and to outline knowledge gaps. Unlike financial distress and financial fraud modelling, auditing has been slow to make use of big data techniques. Auditing would greatly benefit from embracing the use of big data techniques, regardless of whether client firms are using them or not. Findings from accounting and finance research suggest combining multiple big data models instead of applying an individual model, and using big data models to complement human experts.

There are many opportunities to use big data techniques in auditing, particularly when rigorous analytical procedures are combined with traditional audit techniques and expert judgement. Audits could benefit from harnessing the improvements in recent big data financial distress and financial fraud models. Sentiment analysis and natural language processing are other promising auditing tools that require more research. There are also novel research directions for auditing which are well-suited to big data techniques, such real-time information settings, and collaborative platforms and peer-to-peer marketplaces.

The rapid growth of big data across all fields means that academic publications have been leapfrogged by discourse in popular outlets Gandomi and Haider (2015). Going forward, there is a challenge to conduct robust research that better informs audit practice in a

timely manner. This includes the future research suggested above that evaluates the effectiveness of different big data techniques in an auditing context, as well as associated cost-benefit analyses and studies that consider the best ways to combine big data modelling with expert judgement.

Research has an important role to play in bringing theory and practice into closer alignment. Academic literature has lamented the slow integration of big data into auditing. However, anecdotal evidence from partners at some leading audit firms indicates they have begun to use big data. Indeed, the websites of some audit firms promote data analytics as part of their innovation in auditing. For example, KPMG describes their audit as "powered by Data & Analytics (D&A) innovations" (KPMG, 2016) and Deloitte's Chief Innovation Officer mentions the use of natural language processing and other big data techniques in auditing (Raphael, 2015)[8]. On the other hand, while the academic literature had referred to big data as potentially a "game-changer" that represents a "paradigm shift", one KPMG partner has stated that "From the perspective of an auditor, the rise of D&A does not represent a fundamental shift in what we do" (O'Donnell, 2016). This statement might not be representative, but it flags that practitioners do not yet realize the potential of big data. Overall, the prevalence of big data techniques in audit practice remains largely unknown.

To help align research and practice, it is important to understand the prevalence and nature of big data techniques in audit practice. A qualitative, interview-based study is needed to fill this knowledge gap. It should cover as broad a range of firms as possible, from Big 4 through to small audit firms, because usage probably varies widely. Findings from such research could be used to direct future research towards scientifically (in-)validating the effectiveness of current uses, as well as providing clear guidance on the effectiveness of

[8] The author does not use the term "big data", but nevertheless discusses some big-data techniques.

techniques not yet used. This might encourage research findings to be more quickly implemented in practice.

ACNOWLEDGEMENTS

[Acknowledgements have been removed so that authors are not identified]

REFERENCES

Abbasi, A., Albrecht, C., Vance, A., & Hansen, J. (2012). Metafraud: A meta-learning framework for detecting financial fraud. *Mis Quarterly, 36*(4), 1293-1327.

Acito, F., & Khatri, V. (2014). Business analytics: Why now and what next? *Business Horizons, 57*(5), 565-570.

AICPA. (1988). *Statement on Auditing Standards (SAS) No. 59: The Auditor's Consideration of an Entity's Ability to Continue as a Going Concern*. New York, USA: American Institute of Certified Public Accountants.

AICPA. (2011). *Statement on Auditing Standards (SAS) Nos. 122–124: No. 122, Statements on Auditing Standards: Clarification and Recodification; No. 123, Omnibus Statement on Auditing Standards; No. 124, Financial Statements Prepared in Accordance With a Financial Reporting Framework Generally Accepted in Another Country*. New York, USA: American Institute of Certified Public Accountants.

Alanyali, M., Moat, H. S., & Preis, T. (2013). Quantifying the relationship between financial news and the stock market. *Scientific Reports, 3*, 6.

Alles, M. G. (2015). Drivers of the use and facilitators and obstacles of the evolution of big data by the audit profession. *Accounting Horizons, 29*(2), 439-449.

Alston, C. L., Mengersen, K. L., & Pettitt, A. N. (2012). *Case studies in bayesian statistical modelling and analysis*. United Kingdom: Wiley.

Amoore, L. (2011). Data derivatives on the emergence of a security risk calculus for our times. *Theory Culture & Society, 28*(6), 24-43.

Appelbaum, D., Kogan, A., & Vasarhelyi, M. A. (in press). Big data and analytics in the modern audit engagement: Research needs. *Auditing: A Journal of Practice & Theory*.

Association of Certified Fraud Examiners. (2016). Report to the nations on occupational fraud and abuse. http://www.acfe.com/rttn2016.aspx.

Bell, T. B., & Carcello, J. V. (2000). A decision aid for assessing the likelihood of fraudulent financial reporting. *Auditing: A Journal of Practice & Theory, 19*(1), 169-184.

Benoit, D. F., & Van den Poel, D. (2012). Improving customer retention in financial services using kinship network information. *Expert Systems with Applications, 39*(13), 11435-11442.

Bhattacharya, S., Xu, D., & Kumar, K. (2011). An ANN-based auditor decision support system using Benford's law. *Decision Support Systems, 50*(3), 576-584.

Brown-Liburd, H., Issa, H., & Lombardi, D. (2015). Behavioral implications of big data's impact on audit judgment and decision making and future research directions. *Accounting Horizons, 29*(2), 451-468.

Busta, B., & Weinberg, R. (1998). Using Benford's law and neural networks as a review procedure. *Managerial Auditing Journal, 13*(6), 356-366.

Cao, M., Chychyla, R., & Stewart, T. (2015). Big data analytics in financial statement audits. *Accounting Horizons, 29*(2), 423-429.

Chang, R., Lee, A., Ghoniem, M., Kosara, R., Ribarsky, W., Yang, J., . . . Sudjianto, A. (2008). Scalable and interactive visual analysis of financial wire transactions for fraud detection. *Information Visualization, 7*(1), 63-76.

Chang, R. M., Kauffman, R. J., & Kwon, Y. (2014). Understanding the paradigm shift to computational social science in the presence of big data. *Decision Support Systems, 63*, 67-80.

Chen, S. D. (2016). Detection of fraudulent financial statements using the hybrid data mining approach. *Springerplus, 5*, 16.

Chen, W. S., & Du, Y. K. (2009). Using neural networks and data mining techniques for the financial distress prediction model. *Expert Systems with Applications, 36*(2), 4075-4086.

Chiu, V., Liu, Q., & Vasarhelyi, M. A. (2014). The development and intellectual structure of continuous auditing research. *Journal of Accounting Literature, 33*(1–2), 37-57.

Chun, S. H., & Kim, S. H. (2004). Data mining for financial prediction and trading: application to single and multiple markets. *Expert Systems with Applications, 26*(2), 131-139.

Chun, S. H., & Park, Y. J. (2006). A new hybrid data mining technique using a regression case based reasoning: Application to financial forecasting matter. *Expert Systems with Applications, 31*(2), 329-336.

Crawley, M., & Wahlen, J. (2014). Analytics in empirical/archival financial accounting research. *Business Horizons, 57*(5), 583-593.

Curme, C., Preis, T., Stanley, H. E., & Moat, H. S. (2014). Quantifying the semantics of search behavior before stock market moves. *Proceedings of the National Academy of Sciences of the United States of America, 111*(32), 11600-11605.

Dechow, P. M., Ge, W., Larson, C. R., & Sloan, R. G. (2011). Predicting material accounting misstatements. *Contemporary Accounting Research, 28*(1), 17-82.

Dhar, V. (2014). Can big data machines analyze stock market sentiment? *Big Data, 2*(4), 177-181.

Earley, C. E. (2015). Data analytics in auditing: Opportunities and challenges. *Business Horizons, 58*(5), 493-500.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics, 7*(1), 1-26.

Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics, 32*(2), 407-499.

Eining, M. M., Jones, D. R., & Loebbecke, J. K. (1997). Reliance on decision aids: an examination of auditors' assessment of management fraud. *Auditing: A Journal of Practice & Theory, 16*(2), 1-19.

Finch, W. H., & French, B. F. (2015). *Latent variable modeling with R*. New York: Routledge, Taylor & Francis Group.

Flood, M. D. (2009). Embracing change: financial informatics and risk analytics. *Quantitative Finance, 9*(3), 243-256.

Free, C., & Murphy, P. R. (2015). The ties that bind: The decision to co-offend in fraud. *Contemporary Accounting Research, 32*(1), 18-54.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software, 33*(1), 1-22.

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management, 35*(2), 137-144.

Geng, R. B., Bose, I., & Chen, X. (2015). Prediction of financial distress: An empirical study of listed Chinese companies using data mining. *European Journal of Operational Research, 241*(1), 236-247.

Gepp, A., Kumar, K., & Bhattacharya, S. (2010). Business failure prediction using decision trees. *Journal of Forecasting, 29*(6), 536-555.

Gray, G. L., & Debreceny, R. S. (2014). A taxonomy to guide research on the application of data mining to fraud detection in financial statement audits. *International Journal of Accounting Information Systems, 15*(4), 357-380.

Green, B. P., & Choi, J. H. (1997). Assessing the risk of management fraud through neural network technology. *Auditing: A Journal of Practice & Theory, 16*(1), 14-28.

Griffin, P. A., & Wright, A. M. (2015). Introduction: Commentaries on big data's importance for accounting and auditing. *Accounting Horizons, 29*(2), 377-379.

Guerard, J. B., Rachev, S. T., & Shao, B. P. (2013). Efficient global portfolios: Big data and investment universes. *Ibm Journal of Research and Development, 57*(5), 11.

Gui, X. Q., Li, L., Cao, J., & Li, L. (2014). Dynamic communities in stock market. *Abstract and Applied Analysis*.

Hagel, J. (2013). Why accountants should own big data. *Journal of Accountancy, 216*(5), 20.

Hastie, T., Friedman, J., & Tibshirani, R. (2009). *The elements of statistical learning: data mining, inference and prediction* (2nd ed. ed.). New York: Springer.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics, 12*(1), 55-67.

Hogan, C. E., Rezaee, Z., Riley, R. A., & Velury, U. K. (2008). Financial statement fraud: Insights from the academic literature. *Auditing: A Journal of Practice & Theory, 27*(2), 231-252.

Huang, S. Y., Tsaih, R. H., & Lin, W. Y. (2014). Feature extraction of fraudulent financial reporting through unsupervised neural networks. *Neural Network World, 24*(5), 539-560.

Huang, S. Y., Tsaih, R. H., & Yu, F. (2014). Topological pattern discovery and feature extraction for fraudulent financial reporting. *Expert Systems with Applications, 41*(9), 4360-4372.

Humpherys, S. L., Moffitt, K. C., Burns, M. B., Burgoon, J. K., & Felix, W. F. (2011). Identification of fraudulent financial statements using linguistic credibility analysis. *Decision Support Systems, 50*(3), 585-594.

IAASB. (2016). *Exploring the growing use of technology in the audit, with a focus on data analytics*. Available at https://www.ifac.org/publications-resources/exploring-growing-use-technology-audit-focus-data-analytics: International Auditing and Assurance Standards Board (IAASB).

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning : with applications in R*. New York: Springer.

Jans, M., Alles, M. G., & Vasarhelyi, M. A. (2014). A field study on the use of process mining of event logs as an analytical procedure in auditing. *The Accounting Review, 89*(5), 1751-1773.

Jensen, J. B., Ahire, S. L., & Malhotra, M. K. (2013). Trane/Ingersoll Rand combines lean and operations research tools to redesign feeder manufacturing operations. *Interfaces, 43*(4), 325-340.

Kao, Y.-C., Shyu, J., & Huang, J.-Y. (2015). eWOM for stock market by big data methods. *Journal of Accounting, Finance & Management Strategy, 10*(2), 93.

Khandani, A. E., Kim, A. J., & Lo, A. W. (2010). Consumer credit-risk models via machine-learning algorithms. *Journal of Banking & Finance, 34*(11), 2767-2787.

Kim, M.-J., & Han, I. (2003). The discovery of experts' decision rules from qualitative bankruptcy data using genetic algorithms. *Expert Systems With Applications, 25*(4), 637-646.

Kim, S. H. (2000). An architecture for advanced services in cyberspace through data mining: A framework with case studies in finance and engineering. *Journal of Organizational Computing and Electronic Commerce, 10*(4), 257-270.

Kim, S. Y., & Upneja, A. (2014). Predicting restaurant financial distress using decision tree and AdaBoosted decision tree models. *Economic Modelling, 36*, 354-362.

Kirkos, E., Spathis, C., & Manolopoulos, Y. (2007). Data mining techniques for the detection of fraudulent financial statements. *Expert Systems with Applications, 32*(4), 995-1003.

Koyuncugil, A., & Ozgulbas, N. (2012a). Early warning system for financially distressed hospitals via data mining application. *Journal of Medical Systems, 36*(4), 2271-2287.

Koyuncugil, A., & Ozgulbas, N. (2012b). Financial early warning system model and data mining application for risk detection. *Expert Systems with Applications, 39*(6), 6238-6253.

KPMG. (2016). *KPMG audit, powered by data + analytics*. Available from https://assets.kpmg.com/content/dam/kpmg/pdf/2016/03/data-and-analytics-tools.pdf.

Krahel, J. P., & Titera, W. R. (2015). Consequences of big data and formalization on accounting and auditing standards. *Accounting Horizons, 29*(2), 409-422.

Krambia-Kapardis, M., Christodoulou, C., & Agathocleous, M. (2010). Neural networks: The panacea in fraud detection? *Managerial Auditing Journal, 25*(7), 659-678.

Kristoufek, L. (2013). Can Google Trends search queries contribute to risk diversification? *Scientific Reports, 3*.

Kuzey, C., Uyar, A., & Delen, D. (2014). The impact of multinationality on firm value: A comparative analysis of machine learning techniques. *Decision Support Systems, 59*, 127-142.

Lam, M. (2004). Neural network techniques for financial performance prediction: integrating fundamental and technical analysis. *Decision Support Systems, 37*(4), 567-581.

Lennox, C. S., & Kausar, A. (2017). Estimation risk and auditor conservatism. *Review of Accounting Studies, 22*(1), 185-216.

Levenberg, A., Pulman, S., Moilanen, K., Simpson, E., & Roberts, S. (2014). Predicting economic indicators from web text using sentiment composition. *International Journal of Computer and Communication Engineering, 3*(2), 109.

Li, H., Sun, J., & Wu, J. (2010). Predicting business failure using classification and regression tree: An empirical comparison with popular classical statistical methods and top classification mining methods. *Expert Systems With Applications, 37*(8), 5895-5904.

Li, X., Ma, J., Wang, S. Y., & Zhang, X. (2015). How does Google search affect trader positions and crude oil prices? *Economic Modelling, 49*, 162-171.

Lin, C. C., Chiu, A. A., Huang, S. Y., & Yen, D. C. (2015). Detecting the financial statement fraud: The analysis of the differences between data mining techniques and experts' judgments. *Knowledge-Based Systems, 89*, 459-470.

Lin, F. Y., & McClean, S. (2001). A data mining approach to the prediction of corporate failure. *Knowledge-Based Systems, 14*(3), 189-195.

Loughran, T. I. M., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of Finance, 66*(1), 35-65.

Martens, D., Vanhoutte, C., De Winne, S., Baesens, B., Sels, L., & Mues, C. (2011). Identifying financially successful start-up profiles with data mining. *Expert Systems with Applications, 38*(5), 5794-5800.

Moeckel, C. (1991). Two factors affecting an auditor's ability to integrate audit evidence*. *Contemporary Accounting Research, 8*(1), 270-292.

Moffitt, K. C., & Vasarhelyi, M. A. (2013). AIS in an age of big data. *Journal of Information Systems, 27*(2), 1-19.

Mohammad, S. M. (2012). From once upon a time to happily ever after: Tracking emotions in mail and books. *Decision Support Systems, 53*(4), 730-741.

Nardo, M., Petracco-Giudici, M., & Naltsidis, M. (2016). Walking down wall street with a tablet: A survey of stock market predictions using the web. *Journal of Economic Surveys, 30*(2), 356-369.

Negnevitsky, M. (2011). *Evolutionary computation*. Harlow, England: Addison Wesley/Pearson.

Ngai, E. W. T., Hu, Y., Wong, Y. H., Chen, Y. J., & Sun, X. (2011). The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. *Decision Support Systems, 50*(3), 559-569.

O'Donnell, R. (2016). *Data, analytics and your audit*. Avaialble from https://home.kpmg.com/us/en/home/insights/2016/02/data-analytics-audit-article.html: KPMG.

Pachamanova, D. A., & Fabozzi, F. J. (2014). Recent trends in equity portfolio construction analytics. *Journal of Portfolio Management, 40*(3), 137-+.

Perols, J. (2011). Financial statement fraud detection: An analysis of statistical and machine learning algorithms. *Auditing: A Journal of Practice & Theory, 30*(2), 19-50.

Piskorec, M., Antulov-Fantulin, N., Novak, P. K., Mozetic, I., Grcar, M., Vodenska, I., & Smuc, T. (2014). Cohesiveness in financial news and its relation to market volatility. *Scientific Reports, 4*, 8.

Preis, T., Moat, H. S., & Stanley, H. E. (2013). Quantifying trading behavior in financial markets using Google Trends. *Scientific Reports, 3*.

Provost, F., & Fawcett, T. (2013). *Data science for business*. Sebastopol, UNITED STATES: O'Reilly Media.

Purda, L., & Skillicorn, D. (2015). Accounting variables, deception, and a bag of words: assessing the tools of fraud detection. *Contemporary Accounting Research, 32*(3), 1193-1223.

Rabhi, F. A., & Benatallah, B. (2002). An integrated service architecture for managing capital market systems. *Ieee Network, 16*(1), 15-19.

Ranco, G., Bordino, I., Bormetti, G., Caldarelli, G., Lillo, F., & Treccani, M. (2016). Coupling news sentiment with web browsing data improves prediction of intra-day price dynamics. *Plos One, 11*(1), 14.

Raphael, J. (2015). How artificial intelligence can boost audit quality. *CFO Magazine (CFO.com), June 15*.

Ravi, K., & Ravi, V. (2015). A survey on opinion mining and sentiment analysis: Tasks, approaches and applications. *Knowledge-Based Systems, 89*, 14-46.

Ravisankar, P., Ravi, V., Rao, G. R., & Bose, I. (2011). Detection of financial statement fraud and feature selection using data mining techniques. *Decision Support Systems, 50*(2), 491-500.

Read, W. J., & Yezegel, A. (2016). Auditor tenure and going concern opinions for bankrupt clients: Additional evidence. *Auditing: A Journal of Practice & Theory, 35*(1), 163-179.

Rokach, L., & Maimon, O. (2014). *Data mining with decision trees: theory and applications* (2nd edition.. ed.). Hackensack, USA: World Scientific.

Sadasivam, G. S., Subrahmanyam, M., Himachalam, D., Pinnamaneni, B. P., & Lakshme, S. M. (2016). Corporate governance fraud detection from annual reports using big data analytics. *International Journal of Big Data Intelligence, 3*(1), 51-60.

Schneider, G. P., Dai, J., Janvrin, D. J., Ajayi, K., & Raschke, R. L. (2015). Infer, predict, and assure: Accounting opportunities in data analytics. *Accounting Horizons, 29*(3), 719-742.

Shapira, Y., Berman, Y., & Ben-Jacob, E. (2014). Modelling the short term herding behaviour of stock markets. *New Journal of Physics, 16*, 16.

Singh, V. K., Bozkaya, B., & Pentland, A. (2015). Money walks: Implicit mobility behavior and financial well-being. *Plos One, 10*(8).

Smales, L. A. (2014a). News sentiment in the gold futures market. *Journal of Banking & Finance, 49*, 275-286.

Smales, L. A. (2014b). Reaction to nonscheduled news during financial crisis: Australian evidence. *Applied Economics Letters, 21*(17), 1214-1220.

Smales, L. A. (2015). Time-variation in the impact of news sentiment. *International Review of Financial Analysis, 37*, 40-50.

Smith, S. S. (2015). Strategy, analytics, and the role of accountancy. *Strategic Management Review, 9*(1), 87-93.

Sun, E. W., Chen, Y. T., & Yu, M. T. (2015). Generalized optimal wavelet decomposing algorithm for big financial data. *International Journal of Production Economics, 165*, 194-214.

Sun, E. W., & Meinl, T. (2012). A new wavelet-based denoising algorithm for high-frequency financial data mining. *European Journal of Operational Research, 217*(3), 589-599.

Sun, J., & Li, H. (2008). Data mining method for listed companies' financial distress prediction. *Knowledge-Based Systems, 21*(1), 1-5.

Sun, X. Q., Shen, H. W., & Cheng, X. Q. (2014). Trading network predicts stock price. *Scientific Reports, 4*, 6.

Sutton, C. D. (2005). Classification and regression trees, bagging, and boosting. In E. J. W. C.R. Rao & J. L. Solka (Eds.), *Handbook of Statistics* (Vol. Volume 24, pp. 303-329): Elsevier.

Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *Journal of Finance, 62*(3), 1139-1168.

Thulasiram, R. K., Thulasiraman, P., Prasain, H., & Jha, G. K. (2016). Nature-inspired soft computing for financial option pricing using high-performance analytics. *Concurrency and Computation-Practice & Experience, 28*(3), 707-728.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B, 58*(1), 267-288.

Van Den Bogaerd, M., & Aerts, W. (2011). Applying machine learning in accounting research. *Expert Systems With Applications, 38*(10), 13414-13424.

Vasarhelyi, M. A., Kogan, A., & Tuttle, B. M. (2015). Big data in accounting: An overview. *Accounting Horizons, 29*(2), 381-396.

Videla-Cavieres, I. F., & Rios, S. A. (2014). Extending market basket analysis with graph mining techniques: A real case. *Expert Systems with Applications, 41*(4), 1928-1936.

Warren, J. D., Moffitt, K. C., & Byrnes, P. (2015). How big data will change accounting. *Accounting Horizons, 29*(2), 397-407.

Werner, M., & Gehrke, N. (2015). Multilevel process mining for financial audits. *Ieee Transactions on Services Computing, 8*(6), 820-832.

West, J., & Bhattacharya, M. (2016). Intelligent financial fraud detection: A comprehensive review. *Computers & Security, 57*, 47-66.

Xiao, S., Ma, S. H., Li, G., & Mukhopadhyay, S. K. (2016). European option pricing with a fast fourier transform algorithm for big data analysis. *Ieee Transactions on Industrial Informatics, 12*(3), 1219-1231.

Xu, M., & Zhang, C. (2009). Bankruptcy prediction: the case of Japanese listed companies. *Review of Accounting Studies, 14*(4), 534-558.

Yoon, K., Hoogduin, L., & Zhang, L. (2015). Big data as complementary audit evidence. *Accounting Horizons, 29*(2), 431-438.

Zhang, D. S., & Zhou, L. (2004). Discovering golden nuggets: Data mining in financial application. *Ieee Transactions on Systems Man and Cybernetics Part C-Applications and Reviews, 34*(4), 513-522.

Zhang, J., Yang, X. S., & Appelbaum, D. (2015). Toward effective big data analysis in continuous auditing. *Accounting Horizons, 29*(2), 469-476.

Zhang, S., & Wu, X. (2011). Fundamentals of association rules in data mining and knowledge discovery. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 1*(2), 97-116.

Zhang, X. Z., Hu, Y., Xie, K., Zhang, W. G., Su, L. J., & Liu, M. (2015). An evolutionary trend reversion model for stock trading rule discovery. *Knowledge-Based Systems, 79*, 27-35.

Zheludev, I., Smith, R., & Aste, T. (2014). When can social media lead financial markets? *Scientific Reports, 4*, 12.

Zhou, L. G., Lu, D., & Fujita, H. (2015). The performance of corporate financial distress prediction models with features selection guided by domain knowledge and data mining approaches. *Knowledge-Based Systems, 85*, 52-61.

Zhou, W., & Kapoor, G. (2011). Detecting evolutionary financial statement fraud. *Decision Support Systems, 50*(3), 570-575.

APPENDIX A: BRIEF DESCRIPTIONS OF COMMON BIG DATA TECHNIQUES

Technique	Brief Description
Regularized Regression (also known as shrinkage)	Aims to prevent over-fitting by shrinking variable coefficients towards zero. This shrinkage reduces the variance of the coefficient estimates that can adversely affect prediction accuracy, particularly with "highly correlated, large p" problems. It can also be used solve ill-formed problems. *Further reading*: James, Witten, Hastie, and Tibshirani (2013, pp. 214-228) provide further detail in Chapter 6, particularly Section 6.2.
- Ridge Regression	Uses an L2 penalty based on the sum of squared coefficients, which performs well when all variables are likely to be important in relatively similar magnitudes.
- Lasso/LARS	Uses an L1 penalty based on the sum of the absolute value of coefficients. The important advantage of this penalty is that it is effective at variable selection and so results in simpler models that are often desirable for their improved interpretability.
- Elastic-Net	Uses a weighted average of the L1 and L2 penalty. The weighting can be automatically chosen based on the data using a process called cross-validation. This weighted average can result in substantially improved model accuracy in some cases.
Tree-based Methods	Comprise single tree model or an ensemble of them. Tree models are non-parametric models that are built in a recursive process of splitting the data into homogenous groups (usually two). *Further reading*: Rokach and Maimon (2014) cover single trees in detail, while Sutton (2005) also cover ensembles.

Technique	Brief Description
- Single Classification and Regression Trees (CART) also known as decision trees	The advantages of a single tree are that they: are resistant to outliers and irrelevant variables, automatically model interactions between variables, and do not require any variable transformations. Relatively small models are also easy to interpret and display visually. However, single trees are very sensitive to changes in the data (as are some neural networks) and so have high variance.
- Ensembles of decision trees including Random Forests (enhanced bagging) and Multiple Additive Regression Trees (MART or gradient boosting)	Ensembles of decision trees that are combined through an averaging process (Random Forests) or iterative improvement process (MART). This reduces the high variance of individual trees and usually results in increased accuracy. Random Forests are particularly good at "large p, small n" problems. Ensemble models are inherently more difficult to interpret, but there are procedures to extract information in interpretable ways.
Splines - Multivariate Adaptive Regression Splines (MARS)	Splines involve dividing the range of independent variables into sections and fitting separate polynomials to each section. This is particularly useful when there are known breakpoints that separate different distributions. For example, the distribution for retail sales is different during holiday periods. Alternatively, MARS is one particular spline technique that automatically chooses the number of sections and where to place the breakpoints (and then fits linear models to each section). Other types of splines include **natural regression splines** and **smoothing splines**. **Local regression** is a popular alternative to splines. *Further reading*: James et al. (2013, pp. 271-282) cover splines and local regression in Chapter 7, particularly Sections 7.4–7.6. MARS is more complex and only covered in a more technical book by Hastie, Friedman, and Tibshirani (2009, pp. 321-329) in Section 9.4.

Technique	Brief Description
Support Vector Machines (SVMs)	SVMs are popular for classification problems, but are not applicable to regression problems. SVMs place hyperplanes in the data to attempt to separate it into the desired groups. Kernel SVMs offer non-linear extensions. Major drawbacks include no variable selection and no easy way to calculate the associated probabilities of classification. Logistic regression with an L1 or L2 penalty is an alternative to binary classification that overcomes these drawbacks. *Further reading*: Provost and Fawcett (2013, pp. 89-94) briefly introduce SVMs, starting with a comparison to standard logistic regression.
Naïve Bayes and Bayesian (Belief) Networks	A simple model that assumes the variables (or features) are (conditionally) independent. This assumption is almost always violated, but it can still perform well in some circumstances, because of the low variance associated with the simple assumption. It also easily handles "large p" problems. Bayesian belief networks are generalisations of Naïve Bayes that relax some of the independence assumptions by defining a network of conditional dependencies between variables. *Further reading*: Provost and Fawcett (2013, pp. 233-244) introduce the basic concepts of Naïve Bayes and Alston, Mengersen, and Pettitt (2012, pp. 348-360) cover Bayesian Networks in Chapter 20.

Technique	Brief Description
Genetic Algorithms (GAs) including Genetic Programming (GP)	Types of evolutionary algorithms that are heavily based on Darwin's survival of the fittest principle to evolve better solutions to a problem. They are non-parametric, and able to handle missing values and model interactions, but there are a large number of model parameters to set based on user expertise. GAs can be used for both supervised learning and unsupervised learning, and often to optimise the parameters of other models. *Further reading*: Negnevitsky (2011, pp. 219-257) cover evolutionary algorithms in Chapter 7.
Artificial Neural Networks (ANNs), sometimes called Neural Networks or Neural Nets	ANNs are non-parametric models designed on the inner processes of the human brain, primarily with respect to pattern learning. There are many different types of ANNs and they can be trained using **supervised** or **unsupervised** methods (such as **self-organising maps**). Procedures (such as genetic algorithms) are available to automate the numerous model parameters. Advantages include their ability to model non-linear relationships and handle highly correlated variables and outliers. However, the black-box nature makes interpretation difficult, although techniques are available to extract some information. *Further reading*: Negnevitsky (2011, pp. 165-217) provide more detail in Chapter 6.
Association Rules	Unsupervised learning approaches that attempt to find simple rules to describe frequently occurring patterns. For example analysing a department store database might reveal that customers who buy jeans also often buy music. *Further reading*: S. Zhang and Wu (2011) provide an overview of association rules.

Technique	Brief Description
Clustering or **Data Segmentation**	A large collection of unsupervised learning techniques designed to find sub-groups within the data, such that the data is more homogenous within each sub-group. *Further reading*: Provost and Fawcett (2013, pp. 163-183) provide more detail in Chapter 6, particularly in the section titled "Clustering".
Latent Variable Models	A class of models that assumes there are one or more influential quantities that are hidden and unobservable. Popular examples include principal components analysis, principal curves, item response theory and multidimensional scaling, which attempt to model the complete set of data with a smaller set of latent variables. Such methods can also be used as a first step that feeds into a second supervised learning step. *Further reading*: Finch and French (2015) provide information on a variety of latent variable models.
Ensembles	Many of contemporary techniques, including those listed above, combine the results of multiple underlying models. Other techniques to combine multiple models include averaging outputs, a majority vote decision, a hierarchical approach, and more sophisticated processes such as stacking, bagging, and boosting. Ensemble models often outperform individual models in terms of accuracy, but they are inherently more complex to interpret. *Further reading*: Sutton (2005) introduces bagging and boosting (in Sections 1.2, 5 and 6), two popular methods to create ensemble models.