KeAi
CHINESE ROOTS
GLOBAL IMPACT

Available online at www.sciencedirect.com

ScienceDirect

The Journal of Finance and Data Science 8 (2022) 69–85
http://www.keaipublishing.com/en/journals/jfds/

Big data, accounting information, and valuation

Doron Nissim *

Columbia Business School, USA

Received 2 April 2022; accepted 13 April 2022
Available online 20 April 2022

Abstract

This paper reviews research that uses big data and/or machine learning methods to provide insight relevant for equity valuation. Given the huge volume of research in this area, the review focuses on studies that either use or inform on accounting variables. The article concludes by providing recommendations for future research and practice.

JEL classification: C53; G11; G12; G14; G17; G32; M41

Keywords: Big data; Machine learning; Valuation; Financial misstatements; Earnings forecasts; Stock return predictability; XBRL; Survey

1. Introduction

The term Big Data is not well defined. It is commonly used to describe a range of different concepts: from the collection and aggregation of vast amounts of data, to a plethora of advanced analytical techniques designed to reveal relevant patterns (Favaretto et al[1]). For example, the European Commission[2] defines big data as:

> *… large amounts of different types of data produced from various types of sources, such as people, machines or sensors. This data could be climate information, satellite imagery, digital pictures and videos, transition records or GPS signals. Big Data may involve personal data: that is, any information relating to an individual, and can be anything from a name, a photo, an email address, bank details, posts on social networking websites, medical information, or a computer IP address.*

Oracle defines big data as[a]

> *… larger, more complex data sets, especially from new data sources. These data sets are so voluminous that traditional data processing software just can't manage them … data that contains greater variety, arriving in increasing volumes, and with more velocity … the three Vs … variety refers to the many types of data that are*

E-mail address: dn75@columbia.edu.
Peer review under responsibility of China Science Publishing & Media Ltd.
* Helpful comments were provided by Matthias Breuer, Kai Du, Jingzhi Huang, and two anonymous reviewers.
[a] https://www.oracle.com/big-data/what-is-big-data/.

https://doi.org/10.1016/j.jfds.2022.04.003

> *available. Traditional data types were structured and fit neatly in a relational database. With the rise of big data, data comes in new unstructured data types. Unstructured and semistructured data types, such as text, audio, and video, require additional preprocessing to derive meaning and support metadata. … With big data, you'll have to process high volumes of low-density, unstructured data. This can be data of unknown value, such as Twitter data feeds, clickstreams on a web page or a mobile app, or sensor-enabled equipment. For some organizations, this might be tens of terabytes of data. For others, it may be hundreds of petabytes. … Velocity is the fast rate at which data is received and (perhaps) acted on.*

In addition, references to big data often imply or are implied by the use of artificial intelligence methods, especially machine learning algorithms.

While there is no univocal definition of big data, the use of data and analytics consistent with at least some aspects of this term (e.g., high volume, unstructured data, machine learning) has proliferated in finance and accounting research and practice over the last two decades. This paper reviews research that uses big data and/or machine learning (ML) methods to provide insight related to the use of accounting information in equity valuation. It focuses on studies providing insights about the quality and forecasting of earnings as well as on the use of accounting information in evaluating risk and predicting stock returns. Earnings quality is included in the review because it is relevant for valuation (e.g., in determining or adjusting price multiples or in forecasting the profit margin used to derive free cash flow in DCF; see Nissim[3]).[b] After reviewing insights from extant studies, the paper provides recommendations for future research and practice.

The ability to conduct granular accounting-based analyses across large sets of companies has increased in recent years due to improvements in data availability and modelling techniques. In particular, the Securities and Exchange Commission's (SEC) mandated structured disclosures in eXtensible Business Reporting Language (XBRL) format now provide comprehensive, machine-readable "as-filed" financial statement data for essentially all U.S. public companies. These data enable more accurate and consistent measurement of accounting factors, which may yield more informative insights. In addition, developments in ML methods enable the extraction of insights from large sets of financial variables as well as from unstructured data, which can be used to predict financial misstatements, earnings, and stock returns. They also allow for more efficient estimation of relevant parameters and latent variables (e.g., Breuer and Schutt's[5] model of nondiscretionary accruals).

As reviewed in this paper, most of the accounting/big data-related research to date focuses on using big data (primarily textual) and/or ML methods to extract information from or about accounting variables while paying relatively little attention to variable measurement and accounting relationships. Several recent studies employ XBRL data to facilitate more accurate measurement of accounting predictors of earnings or stock returns (e.g., Du et al[6]) and a few incorporate structured financial analysis (e.g., Binz et al[7]) This paper calls for a greater emphasis on the selection, measurement, and contextualization of accounting variables in implementing ML algorithms. It also highlights the importance of conducting big data/ML research on the measurement and valuation of intangible assets.

The paper proceeds as follows. Section 2 describes studies that use big data or ML methods to predict financial misstatements or to inform on the quality of accounting estimates. Section 3 discusses studies that provide evidence on the informativeness of big data or ML methods about future revenue or earnings. Section 4 describes research that uses big data or ML methods to inform on risk dimensions. Section 5 reviews studies that provide evidence on the usefulness of textual firm disclosures and/or ML methods applied to accounting features in predicting stock returns. Section 6 provides recommendations concerning the use of big data and ML methods in accounting research and practice.

2. Financial misstatements

Many studies demonstrate that using big data or ML methods helps in predicting financial reporting fraud or other financial misstatements. Relatedly, a few studies provide evidence that ML methods can help improve the accuracy of accounting estimates or the quality of audit (which in turn affects the reliability of accounting estimates).

[b] In addition, earnings quality is determined in part by earnings management activities, which affect equity valuation. For example, when a firm's financial misstatement is detected, its stock price usually drops significantly (e.g., Karpoff et al[4]).

2.1. Models based on quantitative information

Extending prior research that uses logistic regressions to predict financial misstatements using quantitative variables (e.g., Dechow et al[8]), many recent studies use ML methods to achieve the same goal. The following are several examples.

Cecchini et al[9] provide a methodology for detecting management fraud using basic financial data. They develop a kernel that allows for an implicit and generally nonlinear mapping of points, usually into a higher dimensional feature space. This financial kernel constructs features shown in prior research to be helpful in detecting management fraud. Support vector machines using the financial kernel correctly labeled 80% of the fraudulent cases and 90.6% of the nonfraudulent cases on a holdout set.

Amiram et al[10] create a firm-year measure that assesses the extent to which features of the distribution of a firm's financial statement numbers diverge from a theoretical distribution posited by Benford's Law. They show that the measure is correlated with proxies for accruals-based earnings management and earnings manipulation, and that it predicts material misstatements as identified by SEC Accounting and Auditing Enforcement Releases (AAER).

Kim et al[11] develop three three-class financial misstatement detection models (intentional, unintentional, and non-misstating): multinomial logistic regression, support vector machine, and Bayesian networks. To deal with class imbalance and asymmetric misclassification costs, the authors apply cost-sensitive learning using MetaCost. They use 49 variables identified by previous studies as predictors of financial misstatements, including financial statement ratios (e.g., asset turnover), off-balance sheet variables (e.g., operating leases), nonfinancial measures (e.g., abnormal change in employees), market variables (e.g., market-adjusted stock returns), and governance measures (e.g., CEO power). They derive additional features by measuring the variables relative to industry benchmarks and past values in addition to their levels, resulting in a total of 1086 features. Variables related to accruals quality, such as changes in inventory, along with industry-adjusted and change measures, show discriminatory power. Firm efficiency (measured using the relationship between revenue and resources such as fixed assets and R&D), and market variables such as the short interest ratio, are also found to be useful in detecting misstatements and deliberate fraud.

Developing models to detect financial statement fraud involves challenges related to (1) the rarity of fraud observations, (2) the relative abundance of explanatory variables identified in the prior literature, and (3) the broad underlying definition of fraud. Perols et al[12] introduce and evaluate three data analytics preprocessing methods to address these challenges. The first method addresses the imbalance between the low number of fraud observations relative to the number of non-fraud observations by creating multiple subsets of the original dataset that each contains all fraud observations and different random subsamples of non-fraud observations. The second method addresses the imbalance between the low number of fraud observations relative to the number of explanatory variables identified in the fraud prediction literature by creating multiple subsets of randomly selected explanatory variables. The third method uses a priori knowledge to partition the variables into subsets based on their relation to specific types of fraud (e.g., revenue versus expense fraud). Results from evaluating actual cases of financial statement fraud suggest that the first and third methods improve fraud prediction performance by approximately 10 percent relative to the best current techniques.

Dutta et al[13] develop predictive models for both intentional (fraudulent) and unintentional (erroneous) financial restatements using a comprehensive dataset that includes 3513 restatement cases over the period 2001 to 2014. They employ several ML techniques, including Decision Tree, Artificial Neural Network, Naïve Bayes, Support Vector Machine, and Bayesian Belief Network, and find that ANN outperforms other data mining algorithms in terms of accuracy and area under the ROC curve.[c]

Hajek and Henriques[14] examine whether an improved financial fraud detection system can be developed by combining specific features derived from financial information and managerial comments in corporate annual reports. Using a wide range of ML methods, they find that ensemble methods outperform the remaining methods in terms of true

[c] AUC—the area under the receiver operating characteristic (ROC) curve—measures the probability that the model ranks a random positive example more highly than a random negative example. AUC ranges in value from 0 to 1. A model whose predictions are 100% wrong has an AUC of 0.0; one whose predictions are 100% correct has an AUC of 1.0.

positive rate (fraudulent firms correctly classified as fraudulent).[d] In contrast, Bayesian belief networks (BBN) performed best on nonfraudulent firms (true negative rate).

Bao et al[15] develop a fraud prediction model using an ML approach. Unlike prior research, the authors use raw accounting numbers rather than financial ratios, and they employ an ML ensemble method rather than logistic regression. They show that their fraud prediction model outperforms the Dechow et al[8] logistic regression model based on financial ratios, and the Cecchini et al[9] support-vector-machine model with a financial kernel that maps raw accounting numbers into a broader set of ratios.

Bertomeu et al[16] use a total of 102 accounting, governance, audit, market, and business variables to predict accounting misstatements. They find that accounting variables do not detect misstatements well on their own, but they become important with suitable interactions with audit and market variables. The top seven variables are % of soft assets, bid-ask spread, non-audit fees divided by total fees, a dummy for qualified opinion over internal controls, changes in operating lease activity, short interest, and stock return volatility.

Hunt et al[17] utilize ML techniques to estimate the likelihood that a company switches auditors based on accounting and other variables (e.g., total assets, discretionary accruals, ROA, material M&A dummy, audit opinion, auditor tenure, etc.). They then examine whether an increased likelihood of switching is associated with audit quality. The authors find lower audit quality among companies that are more likely to switch auditors but remain with their incumbent auditor. These companies have a higher likelihood of misstatement and larger abnormal accruals, consistent with auditors sacrificing audit quality to retain clients that might otherwise switch.

Machine learning methods may be especially effective in detecting fraud when used with internal bookkeeping data. Liang et al[18] apply unsupervised machine learning methods to journal entry data from four different companies. They show that the techniques are effective in recognizing patterns in the data and thus in spotting anomalies, as evidenced by successful case studies and recalling injected anomalies including those created by audit practitioners.

2.2. Models based on text, vocal, visual, or other unstructured data

Psychological and linguistic studies provide insights that can be used to gauge deception (e.g., Vrij.[19]) The following are examples of studies that build on this research to evaluate management credibility and the likelihood of financial fraud.

Loughran and McDonald[20] develop a negative word list (e.g., loss, claim, impairment, against, adverse, restructuring, litigation …) and use it to measure the tone of 10-Ks. They show that the proportion of negative words in the 10-K predicts fraud, material weakness, and unexpected earnings.

Larcker and Zakolyukina[21] estimate linguistic-based classification models of deceptive discussions during quarterly earnings conference calls. Using data on subsequent financial restatements and a set of criteria to identify the severity of accounting problems, the authors label each call as "truthful" or "deceptive." Prediction models are then developed with the word categories that have been shown by previous psychological and linguistic research to be related to deception. They find that the out-of-sample performance of models based on CEO and/or CFO narratives is significantly better than a random guess by 6–16% and is at least equivalent to models based on financial and accounting variables. The language of deceptive executives exhibits more references to general knowledge (you know, you would agree, everybody knows, etc.), fewer non-extreme positive emotions, and fewer references to shareholder value (value for shareholders, stockholder value, investor value, etc.). In addition, deceptive CEOs use significantly more extreme positive emotions (e.g., amazing, brilliant, awesome, etc.) and fewer anxiety words (e.g., worried, fearful, nervous, etc.). Finally, a portfolio formed from firms with the highest deception scores from CFO narratives produces an annualized alpha of between −4% and −11%.

Hobson et al[22] examine whether nonverbal vocal cues (e.g., tone of voice) that are commonly associated with cognitive dissonance (psychological discomfort felt when one's actions and beliefs are discrepant) are useful for detecting financial misreporting. Using automated vocal emotion analysis software, they find that vocal dissonance markers in CEO speech during earnings conference calls are positively associated with the likelihood of irregularity restatements. The diagnostic accuracy levels are 11% better than chance and of similar magnitude to models based solely on financial accounting information. The association between vocal dissonance markers and irregularity restatements holds even after controlling for financial accounting and linguistic-based predictors.

[d] Ensemble methods combine several predictive models to get higher quality predictions than each of the models could provide on its own. Examples include random forest algorithms that combine many decision trees, ensemble of logistic regression classifiers, ensemble of support vector machine models, etc.

Brown et al[23] use a machine learning technique to assess whether the thematic content of financial statement disclosures is incrementally informative in predicting intentional misreporting. Using a Bayesian topic modeling algorithm, they determine and empirically quantify the topic content of a large collection of 10-K narratives. They find that the algorithm produces a valid set of semantically meaningful topics (e.g., "digital technology and services," "legal proceedings") that predict financial misreporting, based on samples of SEC enforcement actions (AAER) and irregularities identified from financial restatements and 10-K filing amendments. The topic algorithm significantly improves the detection of financial misreporting when added to models based on commonly used financial and textual style variables. Furthermore, models that incorporate topics significantly outperform traditional models when detecting serious revenue recognition and core expense errors.

Ryans[24] uses Naïve Bayesian classification to identify SEC comment letters associated with future restatements and write-downs. The naïve Bayesian classifier creates a statistical model based on the differences between words used in the restatement/write-down training documents, compared to the non-restatement training documents. He finds that the classifier performs well in predicting restatements and write-offs. Abnormal investor downloads of the comment letters from the SEC's EDGAR website, significantly negative abnormal returns at comment letter disclosure, revenue recognition comments, and the number of letters in a conversation also help predict these outcomes.

2.3. Accounting estimates

ML methods can also be used by managers and auditors to help improve the accuracy of accounting estimates or to inform on their quality. The following are a few examples.

Ding et al[25] show that ML can substantially improve managerial estimates used in preparing financial statements. Specifically, using insurance companies' data on loss reserves (future customer claims) estimates and realizations, they document that the loss estimates generated by ML were superior to actual managerial estimates reported in financial statements in four out of five insurance lines examined.

Commerford et al[26] conduct an experiment to examine how "algorithm aversion"—the tendency to discount computer-based advice more heavily than human advice, although the advice is identical otherwise—manifests in auditor judgments. They find that auditors receiving contradictory evidence from their firm's artificial intelligence (AI) system (instead of a human specialist) propose smaller adjustments to management's complex estimates, particularly when management develops their estimates using relatively objective (vs. subjective) inputs.

3. Revenue and earnings forecasts

This section reviews studies that provide evidence on the use of big data and/or machine learning methods in forecasting revenue or earnings. It also covers studies that inform on earnings properties that have implications for future earnings, such as earnings persistence and abnormal accruals.

3.1. Firm quantitative disclosures (excluding XBRL)

Traditional research on forecasting annual earnings shows that the simple random walk model (i.e., earnings are predicted to equal their previous year's value) often provides a reasonable starting point (e.g., Bradshaw et al[27]). In some cases, significant improvement may be obtained by allowing for a drift or by accounting for mean-reversion following extreme changes or abnormal profitability (e.g., Freeman et al[28]). Over the last decade, studies that generate earnings forecasts have primarily done so using linear combinations of firm characteristics, with coefficients obtained from panel data regressions of future earnings on current and past firm characteristics. The earnings predictors used by these studies include current earnings as well as other characteristics such as size, dividends, accruals, etc. Essentially all studies allow the intercept and the coefficient on current earnings to depend on whether the company reported a profit or loss. See Monahan[29](chapter 6) for a review and discussion of this research. More recently, several studies adopt ML algorithms in generating earnings forecasts using firm characteristics.

Anand et al[30] employ random forest classifiers to generate out-of-sample predictions of directional changes (increases or decreases) in return on equity (ROE), return on assets (ROA), return on net operating assets (RNOA),

cash flow from operations (CFO), and free cash flow (FCF).[e] With a minimum set of independent variables (e.g., past value of the same variable, and a firm dummy), the method achieves classification accuracies ranging from 57 to 64% for the profitability measures, compared to 50% for the random walk. The cash flow measures perform better than the earnings-based profitability measures. Accruals show strong incremental ability beyond cash flows in predicting future cash flows. The method is insensitive to outliers, and data used were not winsorized or standardized.

Hunt et al[31] use a set of 60 financial ratios to predict the sign of next year's earnings change. They find that the random forest ML technique significantly improves out-of-sample forecast accuracy relative to stepwise logit regressions and elastic net.[f] They further show that these forecasts are useful for generating abnormal returns.

van Binsbergen et al[32] use random forest regressions to generate earnings forecasts based on firm fundamentals and macroeconomic variables.[g] They show that the difference between analysts' expectations and the ML-based forecast is negatively associated with future stock returns, suggesting that the ML estimates provide incremental information relative to analysts' forecasts.

Cao and You[33] show that ML models, especially those accommodating nonlinearities, generate significantly more accurate and informative forecasts of corporate earnings than a host of state-of-the-art earnings prediction models in the extant literature. Further analysis suggests that ML models uncover economically sensible relationships between historical financial information and future earnings, and the new information uncovered by ML models is of considerable economic significance. The new information component is significantly associated with both future stock returns and analyst forecast errors, with stocks in the quintiles with the most favorable new information outperforming those in the least favorable quintiles by approximately 70 bps per month. The overall results suggest that limiting to linear relationships and aggregated accounting numbers substantially understates the decision usefulness of financial statement information to investors.

Easton et al[34] use a simple k-nearest neighbors (k-NN) model to forecast a subject firm's annual earnings by matching its recent earnings history to earnings histories of comparable firms, and then extrapolating the forecast from the comparable firms' lead earnings (median subsequent earnings of the matched k-neighbors). The rationale for this approach is that earnings reflect economic performance, and firms with similar past performance are more likely to perform similarly in the future. To implement the analysis, the authors first "tune" two key parameters: (1) the length of the earnings history, M, used to find matches, and (2) the number of nearest neighbors, k, that are matched to each subject firm-year. They find that the optimal value of M is two and that the optimal value of k is 90. Easton et al[34] show that their approach performs better than other commonly used models (e.g., random walk, regression-based forecasts) in forecasting out-of-sample, and that adding more features to the matching (e.g., accruals, total assets, dividends) does not lead to better forecasts. The model also generates a novel ex ante indicator of forecast inaccuracy. This indicator, which equals the interquartile range of the comparable firms' lead earnings, predicts forecast accuracy and identifies situations when the forecasts are strong (weak) predictors of future stock returns.

Hendriock[35] provides evidence of improved accuracy and bias of earnings forecasts derived using conditional probability density function (pdf) that is estimated using past earnings, accruals, and other financial measures, instead of cross-sectional OLS regressions of earnings on the same variables. The improvement is large when conditional pdfs are obtained via quantile regressions, and it is even larger when substituting artificial neural networks for quantile regressions.

Binz et al[7] use ML to estimate Nissim and Penman's[36] (NP) structural framework that decomposes profitability into four levels of increasingly disaggregated profitability drivers. They find that out-of-sample profitability forecasts obtained by applying machine learning to NP's framework are more accurate than those from benchmark models, and that investing strategies based on intrinsic values generated from the profitability forecasts yield risk-adjusted returns. Focusing on operating activities, core items and five-year-horizon forecasts improves performance while using a long time series of past information impairs performance.

[e] A random forest classifier consists of a large number of individual decision trees that operate as an ensemble. Each individual tree in a random forest classifier spits out a class prediction and the class with the most votes (from the different trees) is the model's prediction.

[f] Elastic net adds to the regression model two sets of shrinkage constrains, ridge and LASSO.

[g] A random forest regression is a non-linear and non-parametric ensemble method that averages multiple forecasts from (potentially) weak decision tree regressions and is asymptotically unbiased and can approximate any function. (A decision tree regression is similar to a decision tree classifier except that each leaf's prediction is the average value of the target variable for the corresponding subset instead of a class.)

3.2. XBRL

Since 2012, all U.S. public companies must tag quantitative amounts in financial statements and footnotes of their 10-K and 10-Q reports using XBRL. XBRL files provide many more additional variables compared to Compustat, and they allow for a much finer and more consistent construction of financial predictors. Several recent studies take advantage of these data. Their findings show some potential, but they also suggest that XBRL data have their own issues (e.g., the common use of report-specific extension elements, filling or extraction errors). Additional research and structure are needed to obtain the full benefit from these data. For example, most studies directly use XBRL reported numbers rather than use the data to construct variables (e.g., use XBRL data to carefully measure core or recurring profit margin). In addition, the quality of these data likely improves over time, as companies gain experience in preparing XBRL files.

Baranes and Palas[37] use Support Vector Machines (SVM) and comprehensive financial data extracted from SEC-mandated XBRL to predict the directional detrended change in firms' EPS. Their model classifies the companies correctly about 63.4% of the time, less than the traditional method of Stepwise Multivariate Logistic Regression (SMLR), which has an average prediction rate of 68.1%. However, SMLR's performance is similar across the industries examined, while SVM results in substantially better performance for some industries.

Using the ratio of extension elements to total elements in XBRL 10-K filings as a measure of XBRL complexity, Huang et al[38] find that firms' XBRL filings are more complex when the firms are performing poorly, an effect that is more pronounced when firms are more complex. Furthermore, complex XBRL filings are associated with less (more) persistent positive (negative) earnings.

Chen et al[39] use ML methods (random forests and stochastic gradient boosting) and high-dimensional detailed financial data (XBRL) to predict the direction of one-year-ahead earnings changes. Their models show significant out-of-sample predictive power: the area under the Receiver Operating Characteristics curve (AUC) ranges from 67.52 to 68.66 percent, significantly higher than the 50 percent of a random guess. The annual size-adjusted returns to hedge portfolios formed based on the predictions range from 5.02 to 9.74 percent. The models outperform two conventional models that use logistic regressions and small sets of accounting variables, and professional analysts' forecasts. The outperformance relative to the conventional models stems from both nonlinear predictor interactions missed by regressions and the use of more detailed financial data by machine learning.

3.3. Firms' textual disclosures

Financial reports and other firm disclosures (e.g., press releases, conference call transcripts, investor presentations, websites) contain much more information than the few hundred financial variables provided by data aggregators such as Compustat, FactSet, or Bloomberg. Over the last two decades, a significant number of studies attempt to extract insight relevant for the prediction of earnings from various textual disclosures, primarily the MD&A.

3.3.1. Financial reports

Li[40] documents a positive relationship between annual report readability and the level and persistence of the firm's earnings, where readability is measured using the length of the document (-; the number of words) and the Fog index (-; a function of the average sentence length in words and the percentage of words with more than two syllables).

Li[41] examines the information content of forward-looking statements (FLS; sentences containing words such as "will," "should," "expect," etc.) in the MD&A of 10-K and 10-Q filings. To assess the content and tone of FLS in MD&As, he relies on a Naïve Bayesian learning algorithm[h] instead of a dictionary-based approach.[i] To construct the training data, Li[41] manually classifies 30,000 randomly selected FLS into one of four tones: positive, neutral,

[h] The Naive Bayes method is a classification technique based on Bayes' Theorem with an assumption of independence among predictors (features). It uses a set of training instances (e.g., sentences, paragraphs, or even documents), which belong to known classes (e.g., because they were each examined and classified by the researcher). For each word or phrase (feature) that appears in the training set, the algorithm uses empirical frequencies to calculate the likelihood of that feature appearing in an instance belonging to each class. When presented with a new instance of unknown class, the naïve Bayesian algorithm observes the frequency of each of the features in the new instance and calculates the likelihood that the instance belongs to each class, according to Bayes Theorem. The class with the greatest likelihood is the predicted class of the new instance.

[i] The dictionary approach relies on a "mapping" algorithm (typically % of words) and assigns each word (or phrase) in a document into different categories (e.g., optimistic/pessimistic, positive/negative) based on some predefined or customized dictionaries.

negative, and uncertain. He finds that the average tone of the FLS is positively associated with future earnings even after controlling for other determinants of future performance. The tone measures based on three commonly used dictionaries (Diction, General Inquirer, and the Linguistic Inquiry and Word Count) do not positively predict future performance.

Amel-Zadeh and Faasse[42] find that changes in the text of the MD&A and footnotes and differences between the tone of the two sections predict negative stock returns and operating performance. Investors generally underreact to the information in the two narrative sections, particularly to information in the footnotes.

Peterson et al[43] use a vector space model to measure accounting consistency based on the textual (words) similarity of accounting policy footnotes disclosed in 10-K filings. They find that accounting consistency over time is positively associated with earnings persistence, predictability, accrual quality, and absolute discretionary accruals. Lower consistency relative to other firms in the industry is associated with larger absolute accrual model residuals.

Using the Fog Index to measure readability and focusing on the MD&A section of the annual report, Lo et al[44] predict and find that firms most likely to have managed earnings have MD&As that are more complex, suggesting that the Fog Index may inform on earnings management and by extension on future earnings.

Bochkay and Levine[45] combine narrative disclosures in the MD&A Section of 10-K reports with financial variables to generate one-year-ahead return on equity (ROE) forecasts. They find that models enhanced with MD&A disclosures are more accurate than models using quantitative financial variables alone. Text-enhanced models are as good as or better than analyst consensus forecasts for small firms and firms with low analyst following. Firms with large changes in future performance, negative future performance, high investor scrutiny, high distress risk, high positive accruals, and high value relevance of earnings have more informative MD&A disclosures, whereas younger firms and firms facing greater market risk and litigation exposure have less informative MD&As.

3.3.2. Press releases

Davis et al[46] show that optimistic language in earnings press releases—measured using the difference between counts of words characterized by linguistic theory as optimistic versus pessimistic—is positively associated with future return on assets (ROA).

Henry and Leone[47] evaluate alternative measures of the tone of earnings press releases. They present evidence that word-frequency tone measures based on domain-specific wordlists—compared to general wordlists—better predict the market reaction to earnings announcements, have greater statistical power in short-window event studies, and exhibit more economically consistent post-announcement drift. Further, inverse document frequency weighting, advocated in Loughran and McDonald,[20] provides little improvement to the alternative approach of equal weighting.

3.3.3. Conference calls

Li et al[48] create a culture dictionary using an ML technique applied to earnings call transcripts. For each of the core corporate cultural values of innovation, integrity, quality, respect, and teamwork, the authors identify seed words (e.g., the seven seed words for the cultural value of teamwork are collaborate, collaboration, collaborative, cooperate, cooperation, cooperative, and teamwork). They then train a neural network model to identify for each cultural value words and phrases in earnings calls that appear in close association of the seed words of that cultural value (measured using cosine similarity). Using the culture dictionary, the authors score each of the five core values based on a weighted-frequency count of the related words and phrases in the earnings call transcripts. They find that corporate culture correlates with business outcomes, including operational efficiency, risk-taking, earnings management, executive compensation design, firm value, and deal making, and that the culture-performance link is more pronounced in bad times.

3.3.4. Corporate websites

Lynch and Taylor[49] develop a measure of corporate website content (first principal component of five metrics: # of formatting tags, # of links, # of tags related to images/audio/video, and # of dynamic elements) and use it to identify large changes in corporate websites content that do not occur in close proximity to EDGAR filings and press releases ("standalone changes"). Using standard event study methods, the authors find that standalone changes in corporate websites provide significant value-relevant information to investors, reduce information asymmetry, and precede significant revisions in analyst earnings forecasts and increases in media coverage.

3.4. Analysts' reports

Analysts' reports generally contain four types of outputs:

(1) Forecasts (e.g., EPS, Revenue, EBITDA, EBIT) – typically for the current and two subsequent years, but for some firms they may extend up to five years ahead and in some cases even longer
(2) Stock recommendation – typically three levels (e.g., overweight, neutral, underweight) or five levels (e.g., strong buy, buy, hold, sell, strong sell)
(3) Target price – typically 12-month projected price
(4) Quantitative and qualitative data and analysis, including support for the other three outputs.

Substantial research in finance and accounting examines different properties of the first three outputs. However, until recently relatively little research investigated the fourth output, which is considered by buy-side analysts as more important than the other three public outputs.[j] This has changed recently, with the advent of big data and ML. The following are a few examples.

Huang et al[51] employ a Naive Bayes ML approach to extract textual opinions from a large sample of analyst reports. To construct a training dataset for the Naive Bayes ML approach, they randomly select 10,000 sentences from the sample and manually classify each sentence into one of three categories: positive, negative, and neutral. They then use this "prior" information to classify each sentence in each analyst report as belonging to one of the three categories and measure the overall opinion as the difference between the % positive and negative sentences. They find that analyst reports provide information to investors beyond that in the contemporaneously released earnings forecasts, stock recommendations, and target prices, and also assist investors in interpreting these signals. Cross-sectionally, investors react more strongly to negative than to positive text, suggesting that analysts are especially important in propagating bad news. Additional evidence indicates that analyst report text is more useful when it places more emphasis on nonfinancial topics, is written more assertively and concisely, and when the perceived validity of other information signals in the same report is low. Finally, analyst report text is shown to have predictive value for future earnings growth in the subsequent five years.

Bellstam et al[52] develop a measure of innovation using the text of analyst reports of S&P 500 firms. The text-based measure gives a useful description of innovation by firms with and without patenting and R&D. For nonpatenting firms, the measure identifies innovative firms that adopt novel technologies and innovative business practices (e.g., Walmart's cross-geography logistics). For patenting firms, the text-based measure strongly correlates with valuable patents. The text-based measure forecasts greater firm performance and growth opportunities for up to four years, and these value implications hold just as strongly for innovative nonpatenting firms.

3.5. Media

Several studies provide evidence on the usefulness of media content in forecasting earnings and stock returns. For example, Tetlock et al[53] find that: (1) a high fraction of negative words in firm-specific news stories forecasts low firm earnings; (2) firms' stock prices briefly underreact to the information embedded in negative words; and (3) the earnings and return predictability from negative words is largest for the stories that focus on fundamentals. Together these findings suggest that linguistic media content captures otherwise hard-to-quantify aspects of firms' fundamentals, which investors quickly incorporate into stock prices.

3.6. Web activity, satellite, GPS, cell phone pings, and similar data

Information about consumers' activities (e.g., online searches) or location (e.g., GPS) can be used to help forecast firms' revenue and earnings. Such data have been available for quite some time, but their volume, variety, quality, and supply are

[j] Brown et al[50] survey 344 buy-side U.S. analysts and report that these analysts—which are among the primary consumers of sell-side research—find sell-side analysts useful primarily because of their industry knowledge (average rating of 5.05 out of 6), management access (4.7), and the calls and visits that they have with them (3.91). All three ranked above the "public" output (written reports (3.76), earnings forecasts (2.67), and stock recommendations (1.76); the survey did not ask about target prices).

constantly increasing (e.g., Teoh.[54]) Research using these data has followed a similar trend. The following are a few examples.

Froot et al[55] generate real-time estimates of retail corporate sales using data obtained from MKT Mediastats, LLC pertaining to consumer activity at large U.S. retailers (e.g., web searches and downloads, primarily from mobile devices). They find that these measures explain quarterly sales growth, revenue surprises, and earnings surprises, generating average excess announcement returns of 3.4%.

Katona et al[56] find that satellite data allow sophisticated investors to formulate profitable strategies, especially by targeting the quarterly reports of retailers with bad news. Kang et al[57] show that the number of cars in the parking lots of stores, measured using satellite data, is a timely measure of store-level performance and that some institutional investors trade on and profit from this information.

Using GPS location data from customers' mobile devices, Jin et al[58] develop measures of customer loyalty and find that revenues and earnings are more persistent when customers are more loyal. Specifically, revenues and earnings are more persistent when customers (a) have more regular shopping patterns, (b) are repeat rather than one-time customers, (c) shop during the week rather than on weekends, and (d) spend more time in the store.

Using GPS data, Noh et al[59] show that foot traffic to the commerce locations of firms that sell durable goods decreases after reporting financial results that suggest an increase in solvency risk (as measured using Altman Z-score). They interpret this result as suggesting that consumers respond to information about firms' longevity conveyed by their earnings.

Li and Venkatachalam[60] use a data set of cell phone "pings" (i.e., geolocation signals from mobile devices) to track production disruptions (outages)—material events for U.S. oil refineries. They show that (1) refining firms do not voluntarily disclose refinery outages identified by cell phone pings; (2) traditional media cover only a small portion of ping-based outages; (3) the stock market finds ping-based outages to be value relevant but incorporates the information with delay. Further analysis suggests that given the incomplete media coverage and lack of firm disclosure, investors appear to learn the financial impact of such outages through subsequent earnings announcements.

3.7. Transaction-level data

While transaction-level data have been used by institutional investors for quite some time, studies examining their informativeness have just started to emerge. The following are two examples.

Dichev and Qian[61] explore whether granular consumer purchases data, obtained from the Nielsen Retail Measurement Services (RMS) scanner data, contain incremental value-relevant information about the corresponding manufacturers. Using weekly consumer purchases data generated by point-of-sale systems, the authors construct a measure of aggregated consumer purchases at the firm-quarter level and find that it strongly predicts manufacturer revenues as well as analysts' revenue forecast errors and stock returns.

Blankespoor et al[62] use transaction-level credit and debit card sales for a sample of retail firms to construct a weekly measure of abnormal revenue for each firm. They show that this measure is correlated with abnormal returns, unexpected revenue realizations, and management revenue forecast news.

3.8. Crowdsourcing

Social media has become a popular venue for individuals to share their opinions about stocks, employers, products, and other firm-specific information. The following are examples of studies that use opinions from retail investors, employees, or consumers to help predict firms' earnings and stock returns.

3.8.1. Investors

Chen et al[63] investigate the extent to which investor opinions transmitted through social media predict future stock returns and earnings surprises. The authors conduct textual analysis of articles published on Seeking Alpha, one of the most popular social media platforms for investors in the United States, and they also consider the readers' perspective as inferred via commentaries written in response to these articles. They find that the views expressed in both articles and commentaries predict future stock returns and earnings surprises.

Jame et al[64] find that earnings forecasts provided by Estimize, an open platform that solicits and reports forecasts, are incrementally useful in forecasting earnings. Results are stronger when the number of Estimize contributors is larger, consistent with the benefits of crowdsourcing increasing with the size of the crowd.

Bartov et al[65] test whether opinions of individuals tweeted just prior to a firm's earnings announcement predict its earnings and announcement returns. They find that the aggregate opinion from individual tweets successfully predicts a firm's forthcoming quarterly earnings and announcement returns. These results hold for tweets that convey original information, as well as tweets that disseminate existing information, and are stronger for tweets providing information directly related to firm fundamentals and stock trading.

Drake et al[66] examine whether investors' actions to acquire accounting information are predictive of future firm performance (unexpected earnings and other measures) because these actions partially reveal investors' private expectations of this performance. Using a database of EDGAR downloads, they find evidence that information acquisition of accounting reports by EDGAR users—especially by more sophisticated institutional users (e.g., hedge funds, investment banks)—is predictive of future firm performance.

3.8.2. Consumers

Using a data set of product reviews on Amazon.com, Huang[67] provides evidence that consumer opinions predict revenues, earnings, and stock returns.

3.8.3. Employees

Green et al[68] find that changes in crowdsourced employer ratings are associated with growth in sales and profitability and help forecast one-quarter-ahead earnings announcement surprises and stock returns. The return effect is concentrated among reviews from current employees, stronger among early firm reviews, and also stronger when the employee works in the headquarters state. Decomposing employer ratings, the authors find that the return effect is related to changing employee assessments of career opportunities and views of senior management. It is unrelated to work-life balance.

Huang et al[69] find that employee predictions of their companies' six-month business outlook (obtained from Glassdoor.com) is incrementally informative in predicting future operating performance. Its information content is greater when the disclosures are aggregated from a larger, more diverse, more knowledgeable employee base, consistent with the wisdom of crowds phenomenon. Average outlook predicts bad news events more strongly than good news events, suggesting that employee social media disclosures are relatively more important as a source of bad news.

4. Firm risk

Several studies use textual analysis and/or apply ML methods to accounting information to evaluate different risk dimensions. I provide a few examples below.

Rogers et al[70] examine the relation between disclosure tone (measured using text dictionaries) and shareholder litigation to determine whether managers' use of optimistic language increases litigation risk. They find that plaintiffs target more optimistic statements in their lawsuits and that sued firms' earnings announcements are unusually optimistic relative to other firms experiencing similar economic circumstances.

Donovan et al[71] use machine learning methods to create a comprehensive measure of credit risk based on qualitative information disclosed in conference calls and in the MD&A section of the 10-K. In out-of-sample tests, they find that their measure improves the ability to predict credit events (bankruptcies, interest spreads, and credit rating downgrades), relative to credit risk measures developed by prior research (e.g., z-score).

Using textual analysis and comparing cybersecurity-risk disclosures of firms that were hacked to others that were not (measured using similarity of vectors of frequency of 3210 cybersecurity-related words extracted from the disclosures), Florackis et al[72] construct a firm-level measure of cybersecurity risk. The authors extract the discussion on cybersecurity risk from the "Item 1A. Risk Factors" section of firms' 10-K, which contains information about the most significant risk factors for each firm. They find that cybersecurity risk is priced in the cross-section of stock returns (i.e., it is positively associated with subsequent stock returns), and that high-exposure firms perform poorly in periods of high cybersecurity risk. The measure is higher in information-technology industries, correlates with characteristics linked to firms hit by cyberattacks, and predicts future cyberattacks.

5. Stock return prediction using accounting information

Many of the studies discussed in Section 3, which provide evidence on the informativeness of big data and ML methods about future revenue and earnings, also show the stock return predictability of this information. This section reviews studies that predict stock returns using textual analysis of firms' disclosures and/or by applying ML methods to accounting information. Given that stock return predictability is the focus of much of finance research (and practice), and that accounting information is used in measuring key stock return factors, the volume of studies in this area is huge. I discuss a few examples, focusing on studies that emphasize accounting information.

5.1. Quantitative factors

Studies providing evidence on the usefulness of ML methods in processing quantitative accounting information to predict stock returns go back at least thirty years. For example, Kryzanowski et al[73] use an artificial neural network (ANA) to learn the relationships between a company's stock return one year forward and the most recent four years of financial data for the company and its industry, as well as data for seven macroeconomic variables. They show that the ANN algorithm correctly classifies 72% of the positive/negative returns. The remainder of this subsection describes a few recent studies.

Yan and Zheng[74] examine over 18,000 fundamental signals from financial statements and use a bootstrap approach to evaluate the impact of data mining on fundamental-based anomalies. They find that many fundamental signals are significant predictors of cross-sectional stock returns even after accounting for data mining. This predictive ability is more pronounced following high-sentiment periods and among stocks with greater limits to arbitrage, consistent with mispricing.

Amel-Zadeh et al[75] compare the ability of a range of ML models to predict abnormal stock returns around earnings announcements using financial statements data. They find that Random Forests produce the most accurate forecasts and the highest abnormal returns. Neural network-based models perform relatively better for predictions of extreme market reactions, while linear methods are relatively better in predicting moderate market reactions. Long-short portfolios based on model predictions generate sizable abnormal returns, which seem to decay over time. Random Forests models perform well because they select the most important predictors of free cash flows and firm characteristics that are known predictors of stock returns.

Geertsema and Lu[76] use ML to identify comparable firms and conduct relative valuation. The ML algorithm learns optimal decision rules to predict valuation multiples as weighted averages of peer firm multiples based on their comparability to the subject firm. Machine valuations behave like fundamental value; over-valued stocks decrease in price and under-valued stocks increase in price in the following month. The ML approach identifies valuation drivers that are consistent with theory—profitability ratios, growth measures and efficiency ratios are the most important value drivers.

Chen et al[77] use deep neural networks to estimate an asset pricing model for individual stock returns using many firm fundamentals and macroeconomic variables. The key innovations are to use the fundamental no-arbitrage condition as criterion function, to construct the most informative test assets with an adversarial approach and to extract the states of the economy from many macroeconomic time series. The asset pricing model outperforms out-of-sample all benchmark approaches in terms of Sharpe ratio, explained variation and pricing errors.

Du et al[6] compare the informativeness of SEC-mandated, machine-readable XBRL structured filings, or "as-filed data," with that of Compustat. They find that discrepancies between as-filed and Compustat data, potentially a result of Compustat's standardizations, affect inferences about the existence and magnitude of the accruals anomaly: accruals calculated from as-filed data do predict returns and accruals calculated from Compustat data do not. Trades of hedge funds that download structured filings correlate with the as-filed accruals signal and, especially, the discrepancy between as-filed and Compustat accruals signals. Inferences about four other accounting-based anomalies are similarly affected by discrepancies between data sources.

5.2. Text-based measures

Several studies use textual analysis of firms' financial disclosures to predict stock returns. The following are a few examples.

Feldman et al[78] use a classification scheme of positive and negative words to measure the tone change in the MD&A section of Forms 10-Q and 10-K relative to prior periodic SEC filings. They find that management's tone change adds significantly to portfolio drift returns in the window of 2 days after the SEC filing date through 1 day after the subsequent quarter's preliminary earnings announcement, beyond financial information conveyed by accruals and earnings surprises. The drift returns are affected by the ability of the tone change signals to help predict the subsequent quarter's earnings surprise but cannot be completely attributed to this ability.

Using the complete history of regular quarterly and annual filings by U.S. corporations, Cohen et al[79] show that changes to the language and construction of financial reports (full text) have strong implications for firms' future returns and operations. A portfolio that shorts "changers" and buys "nonchangers" earns up to 188 basis points per month in alpha (over 22% per year) in the future. Moreover, changes to 10-Ks predict future earnings, profitability, future news announcements, and even future firm-level bankruptcies. Unlike typical underreaction patterns, there is no announcement effect, suggesting that investors are inattentive to these simple changes across the universe of public firms.

Meursault et al[80] develop a measure of earnings call text surprise, SUE.txt. They compute it using a regularized logistic text regression that links the text to the market reaction around the call. SUE.txt generates a text-based post-earnings-announcement drift (PEAD.txt) larger than the classic PEAD. The magnitude of PEAD.txt is considerable even in recent years when the classic PEAD is close to zero. The calls' news content is shown to be about details behind the earnings number and the fundamentals of the firm.

Cao et al[81] build an AI analyst that digests corporate financial information, qualitative disclosure and macroeconomic indicators, and show that it is able to beat the majority of human analysts in stock price forecasts and in generating excess returns. In the contest of "man vs machine," the relative advantage of the AI Analyst is stronger when the firm is complex, and when information is high-dimensional, transparent, and voluminous. Human analysts remain competitive when critical information requires institutional knowledge (such as the nature of intangible assets). The edge of the AI over human analysts declines over time when analysts gain access to alternative data and to in-house AI resources. Combining AI's computational power and the human art of understanding soft information produces the highest potential in generating accurate forecasts.[k]

6. Discussion and summary

While the above review provides many examples of studies that employ big data and/or ML algorithms to inform on or extract insight from accounting variables, there are several important areas that have received relatively little attention to date. I elaborate on two such topics, which in my view are especially important: (1) the use of big data and/or ML methods to measure the amortized cost and value of intangible assets; and (2) the incorporation of economic, financial, and accounting structure in implementing ML algorithms.

Relatively few studies use big data or ML methods to provide evidence on the amortized cost or value of intangible assets. This is an extremely important area, and big data/ML methods have the potential to generate substantial insight on these assets. The amortized cost of intangibles assets is difficult to measure because (1) investments in intangible assets are often mingled with periodic costs (e.g., employees that develop and maintain software); (2) some investments in intangibles benefit the current period in addition to future periods (e.g., advertising); and (3) the pattern of the benefits is highly uncertain, so any amortization scheme is likely to produce poor matching.

The value of intangible assets is difficult to measure because the cash flows associated with investments in intangibles are less certain than those from fixed assets (e.g., Kothari et al[83]). This is due to several factors, including (1) greater uncertainty (compared to investments in fixed assets) regarding whether the investment will be successful; (2) if the investment is successful, the benefits may be much higher than those from successful investments in fixed assets due to scalability, network effects, optionality (e.g., an invention in one context can lead to other applications), and

[k] A similar study—although one that focuses on analysts' reports instead of firms' disclosures—is Bonini et al.[82] They complement traditional data sources (market, fundamentals, and analyst recommendations) by processing a large corpus of sell-side analyst reports and use machine learning (ML) to emulate a sophisticated agent who utilizes all data sources to form forward-looking portfolios. They find that analyst textual content on top of conventional data sources has a higher-order complex effect, which feeds into the portfolio selection process in a nonlinear form. Unlike a simple linear technology, a nonlinear machine learning model yields positive and strongly significant alphas. The ML long-short portfolio strategy exhibits a higher correlation with common characteristics related to R&D, leverage, and cash holdings, stressing the intertwined and complex nature of analyst content.

low variable costs; and (3) unlike most fixed assets, intangible assets have little or no exit value if the project is unsuccessful.

Measuring the amortized cost and value of intangibles has become both more important and more feasible over time. The increasing trend over recent decades in intangible intensity and, relatedly, in economic volatility, adjustment speed, and scalability (e.g., McKinsey & Company[84]) have made reported earnings and book value poor proxies for future earnings. For example, Arnott et al[85] find that the underperformance of value factors over the last decade is due in large part to accounting measures such as earnings and book value failing to capture increasingly important intangible assets. The dramatic increase in the availability of data and improvements in computing and data analysis methods enable relatively precise measurement of different types of intangible assets.[1]

My second observation can be summarized as follows: big data and ML methods should complement, not substitute for traditional data and analysis. While big data, alternative data, and modern analytical tools are increasingly important, they should not substitute for a robust framework that incorporates economic, financial, and accounting structure and uses the wealth of financial and non-financial information that is already available. Out-of-sample validity of insights from pure data mining is often questionable. Extracting information from data involves extrapolating from past relationships. In some cases, the statement "this time it's different" actually holds, making it difficult to extrapolate from past data. Relatedly, judgement often has to be incorporated, which is difficult to do when forecasts and estimates are derived from "black boxes." Simple, traditional models that incorporate economic intuition and provide more visibility into the estimation process are more suitable for incorporating judgement.

Some interpret developments in ML and big data as implying less need for structure—"let the data speak for themselves." In my view, the opposite is true. These developments suggest an increasing role for structure and domain expertise. The increasing trends in intangible intensity, economic volatility, adjustment speed, and scalability mentioned above suggest that flexibility to account for changing circumstances is particularly important these days, and the richness of available data and advancement in ML methods can facilitate such adjustments. The first wave of artificial intelligence models used expert systems, while the second shifted to machine learning. Combining the two or more generally moving from "man vs. machine to man plus machine" approach, may yield superior results (e.g., Cao et al[81]).

The structure and operations of many quant funds demonstrate this issue. They often pay relatively little attention to accounting, financial and economic considerations, or they incorporate them separately. A hedge fund may have two separate teams, one working on constructing the indicators/features and the other focusing on training models using the pool of indicators/features to predict stock returns. However, the contextual nature of the features and the interactions among them suggest that greater integration and a more robust process with respect to non-statistical considerations may yield better results. For instance, Dyer et al[86] examine quantitative investors' ability to navigate the impact of new accounting standards on financial data. They find that relative to funds that rely heavily on human discretion to make investment decisions, the returns of quantitative mutual funds temporarily decrease following the implementation of standards that change the definition of key accounting variables. This effect is particularly strong for funds that rely heavily on accounting data and invest in many stocks.

A few recent studies demonstrate the potential benefits of incorporating financial and accounting structure into the analysis. For example, Binz et al[7] use structured profitability analysis and find that it leads to more accurate forecasts, and Du et al[6] show that more careful measurement of accounting variables—facilitated using XBRL data—leads to improved stock return predictability. In my view, this direction is particularly promising. Nissim[3,87,88] dives deep into the development, measurement, and contextualization of accounting factors, thereby providing guidance for selecting, measuring, and interacting accounting-based features when implementing ML algorithms.

As noted, without structure and transparency (as is often the case with ML methods), the out-of-sample validity of pure data mining is questionable and the ability to adjust the model and estimates for changing circumstances is limited. If the training period does not include events or circumstances similar to those underlying out-of-sample predictions—or more generally when past co-movements or patterns are not a good guide for the future—the estimates are likely to be highly imprecise. In such cases, incorporating economic, financial, and accounting considerations is particularly important. This can be done through: (1) careful selection, definition, and measurement of the variables (features); (2) imposing structure on the specifications (e.g.,

[1] Kai Wu (https://www.sparklinecapital.com/research) provides examples of using big data and ML methods to estimate the value of various intangible assets.

incorporating no-arbitrage conditions, allowing for select interactions, grouping variables into clusters, etc.); (3) adjusting the variables and specifications to the context; and (4) modelling causal or quasi-causal effects when such identification is feasible.

Recent research increasingly recognizes the importance of incorporating expert knowledge. For instance, Chen et al[77] implement deep neural networks to estimate an asset pricing model for individual stocks using the no-arbitrage condition as a criterion function. They conclude: "a successful use of machine learning methods in finance requires both subject specific domain knowledge and a state-of-the-art technical implementation." I believe that this conclusion extends to accounting.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Favaretto M, De Clercq E, Schneble CO, Elger BS. What is your definition of Big Data? Researchers' understanding of the phenomenon of the decade. *PLoS One*. 2020;15(2):e0228987.
2. European Commission. *Directorate-General for Justice and Consumers, the EU Data Protection Reform and Big Data*. Publications Office; 2018. https://data.europa.eu/doi/10.2838/190200.
3. Nissim D. *Earnings Quality*. Columbia Business School; 2022a. Available at: https://ssrn.com/abstract=3794378.
4. Karpoff JM, Lee DS, Martin GS. The cost to firms of cooking the books. *J Financ Quant Anal*. 2008;43(3):581–611.
5. Breuer M, Schütt HH. Accounting for uncertainty: an application of Bayesian methods to accruals models. *Rev Account Stud*. 2021:1–43.
6. Du K, Huddart SJ, Jiang D. *Lost in Standardization: Revisiting Accounting-Based Return Anomalies Using As-Filed Financial Statement Data*. 2022. *Available at*: *SSRN 3781979*.
7. Binz O, Schipper K, Standridge K. What can analysts learn from artificial intelligence about fundamental analysis?. https://ssrn.com/abstract=3745078; 2022.
8. Dechow PM, Ge W, Larson CR, Sloan RG. Predicting material accounting misstatements. *Contemp Account Res*. 2011;28(1):17–82.
9. Cecchini M, Aytug H, Koehler GJ, Pathak P. Detecting management fraud in public companies. *Manag Sci*. 2010;56(7):1146–1160.
10. Amiram D, Bozanic Z, Rouen E. Financial statement errors: evidence from the distributional properties of financial statement numbers. *Rev Account Stud*. 2015;20(4):1540–1593.
11. Kim YJ, Baik B, Cho S. Detecting financial misstatements with fraud intention using multi-class cost-sensitive learning. *Expert Syst Appl*. 2016;62:32–43.
12. Perols JL, Bowen RM, Zimmermann C, Samba B. Finding needles in a haystack: using data analytics to improve fraud prediction. *Account Rev*. 2017;92(2):221–245.
13. Dutta I, Dutta S, Raahemi B. Detecting financial restatements using data mining techniques. *Expert Syst Appl*. 2017;90:374–393.
14. Hajek P, Henriques R. Mining corporate annual reports for intelligent detection of financial statement fraud–A comparative study of machine learning methods. *Knowl Base Syst*. 2017;128:139–152.
15. Bao Y, Ke B, Li B, Yu YJ, Zhang J. Detecting accounting fraud in publicly traded US firms using a machine learning approach. *J Account Res*. 2020;58(1):199–235.
16. Bertomeu J, Cheynel E, Floyd E, Pan W. Using machine learning to detect misstatements. *Rev Account Stud*. 2020:1–52.
17. Hunt JO, Rosser DM, Rowe SP. Using machine learning to predict auditor switches: how the likelihood of switching affects audit quality among non-switching clients. *J Account Publ Pol*. 2021;40(5):106785.
18. Liang PJ, Wang A, Akoglu L, Faloutsos C. *Pattern Recognition and Anomaly Detection in Bookkeeping Data*. Carnegie Mellon University; 2021. Working Paper.
19. Vrij A. *Detecting Lies and Deceit: Pitfalls and Opportunities*. 2nd ed. Chichester, UK: John Wiley & Sons; 2008.
20. Loughran T, McDonald B. When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *J Finance*. 2011;66(1):35–65.
21. Larcker DF, Zakolyukina AA. Detecting deceptive discussions in conference calls. *J Account Res*. 2012;50(2):495–540.
22. Hobson JL, Mayew WJ, Venkatachalam M. Analyzing speech to detect financial misreporting. *J Account Res*. 2012;50(2):349–392.
23. Brown NC, Crowley RM, Elliott WB. What are you saying? Using topic to detect financial misreporting. *J Account Res*. 2020;58(1):237–291.
24. Ryans JP. Textual classification of SEC comment letters. *Rev Account Stud*. 2021;26(1):37–80.
25. Ding K, Lev B, Peng X, Sun T, Vasarhelyi MA. Machine learning improves accounting estimates: evidence from insurance payments. *Rev Account Stud*. 2020;25(3):1098–1134.
26. Commerford BP, Dennis SA, Joe JR, Ulla JW. Man versus machine: complex estimates and auditor reliance on artificial intelligence. *J Account Res*. 2022;60(1):171–201.
27. Bradshaw M, Drake M, Myers J, Myers L. A re-examination of analysts' superiority over time-series forecasts of annual earnings. *Rev Account Stud*. 2012;17(4):944–968.
28. Freeman RN, Ohlson JA, Penman SH. Book rate-of-return and prediction of earnings changes: an empirical investigation. *J Account Res*. 1982:639–653.

29. Monahan SJ. Financial statement analysis and earnings forecasting. *Foundations and Trends® in Accounting*. 2018;12(2):105–215.
30. Anand V, Brunner R, Ikegwu K, Sougiannis T. *Predicting Profitability Using Machine Learning*. 2019. *Available at*: *SSRN 3466478*.
31. Hunt J, Myers J, Myers L. Improving earnings predictions with machine learning. *Unpubl. Work. Pap.* 2019.
32. van Binsbergen JH, Han X, Lopez-Lira A. *Man Vs. Machine Learning: The Term Structure of Earnings Expectations and Conditional Biases (No. W27843)*. National Bureau of Economic Research; 2020.
33. Cao K, You H. *Fundamental Analysis via Machine Learning*. 2020. Working Paper.
34. Easton PD, Kapons M, Monahan SJ, Schütt HH, Weisbrod EH. *Forecasting Earnings Using K-Nearest Neighbor Matching*. 2021. Available at: SSRN 3752238.
35. Hendriock M. *Forecasting Earnings with Predicted, Conditional Probability Density Functions*. 2021. Available at: SSRN 3901386.
36. Nissim D, Penman SH. Ratio analysis and equity valuation: from research to practice. *Rev Account Stud*. 2001;6:109–154.
37. Baranes A, Palas R. Earning movement prediction using machine learning-support vector machines (SVM). *Journal of Management Information and Decision Sciences*. 2019;22(2):36–53.
38. Huang F, No WG, Vasarhelyi MA. Do managers use extension elements strategically in the SEC's tagged data for financial statements? Evidence from XBRL complexity. *J Inf Syst*. 2019;33(3):61–74.
39. Chen X, Cho YH, Dou Y, Lev B. *Predicting Future Earnings Changes Using Machine Learning and Detailed Financial Data*. Journal of Accounting Research; 2022. forthcoming.
40. Li F. Annual report readability, current earnings, and earnings persistence. *J Account Econ*. 2008;45(2–3):221–247.
41. Li F. The information content of forward-looking statements in corporate filings—a naive Bayesian machine learning approach. *J Account Res*. 2010;48(5):1049–1102.
42. Amel-Zadeh A, Faasse J. *The Information Content of 10-K Narratives: Comparing MD&A and Footnotes Disclosures*. 2016. *Available at*: *SSRN 2807546*.
43. Peterson K, Schmardebeck R, Wilks TJ. The earnings quality and information processing effects of accounting consistency. *Account Rev*. 2015;90(6):2483–2514.
44. Lo K, Ramos F, Rogo R. Earnings management and annual report readability. *J Account Econ*. 2017;63(1):1–25.
45. Bochkay K, Levine CB. Using MD&A to improve earnings forecasts. *J Account Audit Finance*. 2019;34(3):458–482.
46. Davis A, Piger J, Sedor L. Beyond the numbers: measuring the information content of earnings press release language. *Contemp Account Res*. 2012;29(3):845–868.
47. Henry E, Leone AJ. Measuring qualitative information in capital markets research: comparison of alternative methodologies to measure disclosure tone. *Account Rev*. 2016;91(1):153–178.
48. Li K, Mai F, Shen R, Yan X. Measuring corporate culture using machine learning. *Rev Financ Stud*. 2021;34(7):3265–3315.
49. Lynch B, Taylor DJ. *The Information Content of Corporate Websites*. 2021. *Available at*: *SSRN 3791474*.
50. Brown LD, Call AC, Clement MB, Sharp NY. The activities of buy-side analysts and the determinants of their stock recommendations. *J Account Econ*. 2016;62(1):139–156.
51. Huang A, Zang A, Zheng R. Evidence on the information content of text in analyst reports. *Account Rev*. 2014;89(6):2151–2180.
52. Bellstam G, Bhagat S, Cookson JA. A text-based analysis of corporate innovation. *Manag Sci*. 2021;67(7):4004–4031.
53. Tetlock PC, Saar-Tsechansky M, Macskassy S. More than words: quantifying language to measure firms' fundamentals. *J Finance*. 2008;63(3):1437–1467.
54. Teoh SH. The promise and challenges of new datasets for accounting research. *Account Org Soc*. 2018;68:109–117.
55. Froot K, Kang N, Ozik G, Sadka R. What do measures of real-time corporate sales say about earnings surprises and post-announcement returns? *J Financ Econ*. 2017;125(1):143–162.
56. Katona Z, Painter M, Patatoukas PN, Zeng J. On the capital market consequences of alternative data: evidence from outer space. In: *9th Miami Behavioral Finance Conference*. 2018, July.
57. Kang JK, Stice-Lawrence L, Wong YTF. The firm next door: using satellite images to study local information advantage. *J Account Res*. 2021;59(2):713–750.
58. Jin H, Stubben S, Ton K. *Customer Loyalty and the Persistence of Revenues and Earnings*. 2021. *Available at*: *SSRN 3744417*.
59. Noh S, So EC, Zhu C. *Financial Reporting and Consumer Behavior*. 2021. *Available at*: *SSRN*.
60. Li B, Venkatachalam M. Leveraging big data to study information dissemination of material firm events. *J Account Res*. 2021.
61. Dichev ID, Qian J. *The Benefits of Transaction-Level Data: The Case of Nielsen Scanner Data*. 2021. *Available at*: *SSRN 3740052*.
62. Blankespoor E, Hendricks BE, Piotroski JD, Synn C. *Real-time Revenue and Firm Disclosure*. 2022. Available at: SSRN.
63. Chen H, De P, Hu YJ, Hwang BH. Wisdom of crowds: the value of stock opinions transmitted through social media. *Rev Financ Stud*. 2014;27(5):1367–1403.
64. Jame R, Johnston R, Markov S, Wolfe MC. The value of crowdsourced earnings forecasts. *J Account Res*. 2016;54(4):1077–1110.
65. Bartov E, Faurel L, Mohanram PS. Can Twitter help predict firm-level earnings and stock returns? *Account Rev*. 2018;93(3):25–57.
66. Drake MS, Johnson BA, Roulstone DT, Thornock JR. Is there information content in information acquisition? *Account Rev*. 2020;95(2):113–139.
67. Huang J. The customer knows best: the investment value of consumer opinions. *J Financ Econ*. 2018;128(1):164–182.
68. Green TC, Huang R, Wen Q, Zhou D. Crowdsourced employer reviews and stock returns. *J Financ Econ*. 2019;134(1):236–251.
69. Huang K, Li M, Markov S. What do employees know? Evidence from a social media platform. *Account Rev*. 2020;95(2):199–226.
70. Rogers JL, Van Buskirk A, Zechman SLC. Disclosure tone and shareholder litigation. *Account Rev*. 2011;86(6):2155–2183.
71. Donovan J, Jennings J, Koharki K, Lee J. Measuring credit risk using qualitative disclosure. *Rev Account Stud*. 2021;26(2):815–863.
72. Florakis C, Louca C, Michaely R, Weber M. *Cybersecurity Risk (No. W28196)*. National Bureau of Economic Research; 2020.
73. Kryzanowski L, Galler M, Wright DW. Using artificial neural networks to pick stocks. *Financ Anal J*. 1993;49(4):21–27.

74. Yan X, Zheng L. Fundamental analysis and the cross-section of stock returns: a data-mining approach. *Rev Financ Stud*. 2017;30(4):1382–1423.
75. Amel-Zadeh A, Calliess JP, Kaiser D, Roberts S. *Machine Learning-Based Financial Statement Analysis*. 2020. *Available at*: *SSRN 3520684*.
76. Geertsema P, Lu H. *Relative Valuation with Machine Learning*. 2021. *Available at*: *SSRN 3740270*.
77. Chen L, Pelger M, Zhu J. Deep learning in asset pricing. Working Paper. Available at: https://arxiv.org/pdf/1904.00745.pdf; 2021.
78. Feldman R, Govindaraj S, Livnat J, Segal B. Management's tone change, post earnings announcement drift and accruals. *Rev Account Stud*. 2010;15(4):915–953.
79. Cohen L, Malloy C, Nguyen Q. Lazy prices. *J Finance*. 2020;75(3):1371–1415.
80. Meursault V, Liang PJ, Routledge B, Scanlon M. *PEAD. Txt: Post-Earnings-Announcement Drift Using Text*. 2021. *Available at*: *SSRN 3778798*.
81. Cao S, Jiang W, Wang JL, Yang B. *From Man vs. Machine to Man+ Machine: The Art and Ai of Stock Analyses (No. W28800)*. National Bureau of Economic Research; 2021.
82. Bonini S, Gultekin M, Shohfi T, Simaan M. *Analysts, Fundamentals, and Portfolio Selection: A Machine Learning Approach*. 2021. *Available at*: *SSRN 3399746*.
83. Kothari SP, Laguerre TE, Leone AJ. Capitalization versus expensing: evidence on the uncertainty of future earnings from capital expenditures versus R&D outlays. *Rev Account Stud*. 2002;7(4):355–382.
84. McKinsey, Company. Getting tangible about intangibles: the future of growth and productivity?. https://www.mckinsey.com/business-functions/marketing-and-sales/our-insights/getting-tangible-about-intangibles-the-future-of-growth-and-productivity; 2021.
85. Arnott RD, Harvey CR, Kalesnik V, Linnainmaa JT. Reports of value's death may be greatly exaggerated. *Financ Anal J*. 2021;77(1):44–67.
86. Dyer T, Guest NM, Yu E. *New Accounting Standards and the Performance of Quantitative Investors*. 2021. *Available at*: *SSRN 3969442*.
87. Nissim D. *Profitability Analysis*. Columbia Business School; 2022b. Available at: https://ssrn.com/abstract=4064824.
88. Nissim D. *Reformulated Financial Statements*. Columbia Business School; 2022c. Available at: https://ssrn.com/abstract=4064722.