

Analytics for Finance and Accounting: Data Structures and Applied AI



Sean Cao

University of Maryland

Wei Jiang

Emory University

Lijun Lei




















*University of North Carolina
at Greensboro*


Copyright © 2025


















All Rights Reserved

No part of this publication may be reproduced, distributed, or transmitted in any form or by any means, including photocopying, recording, or other electronic or mechanical methods, without the author's prior written permission, except in the case of brief quotations embodied in critical reviews and certain other non-commercial uses permitted by copyright law. For permission requests, please get in touch with the author.

Contents

Preface.....	1
Forewords	2
Chapter 1 Data Analytics in Finance and Accounting.....	4
1.1. How to leverage data science for corporate stakeholders	4
 The rising use of big data for decision-making.....	4
 How to leverage data science for corporate stakeholders: the importance of domain knowledge.....	5
 Extra contents on summarizing how industries use AI.....	5
1.2. Overview of structured and unstructured data	12
 An overview of available business data.....	12
 An overview of unstructured and structured data analytics	122
1.3. Theory-driven and machine-learning approach of data analytics	17
 Theory-driven and machine-learning approach of data analytics	17
1.4. The advantages of applying machine-learning approaches.....	19
 The advantages of applying machine-learning approaches.....	19
References	22
Chapter 2 Analyzing Annual Reports	23
2.1. Data structure in annual reports and 10-K filings	23
 Data structure of the 10-K filing.....	23
 Item 1 Business Description.....	23
 Item 1A risk disclosure	23
 Item 7: Management's Discussion and Analysis	23
2.2. Conventional textual analysis approach.....	32
 Textual analysis: Keyword search and LDA	32
2.3. Empirical examples: Analyzing corporate filings for making business decisions	38
 An empirical example of analyzing 10-K filings.....	38
Appendix 2A: Project 1a How to crawl annual reports	43
Appendix 2B: Project 1b How to parse unstructured data	43
Appendix 2C: Solution	44
 How to crawl annual reports.....	44
 How to parse unstructured data	49
References	52
Chapter 3 Emerging AI Technology in Textual Analysis.....	53
3.1. Procedures for applying machine learning models	53
 Procedures for applying machine learning models	53
3.2. Fundamental concepts of pre-training in machine learning	61
 Basic concept and foundation of ML.....	60
 Application of supervised learning, ensemble learning, and model selection in Fintech	60
3.3. Pre-trained phrase-level word embedding	62
3.4. Pre-trained sentence level-word embedding	64
 Textual analysis: Word representation and sentence level analysis using Google Bert	64

3.5. Prompt engineering for large language models	70
 Prompt engineering for large language models	70
3.6. Reinforcement learning	75
3.7. Man and machine	77
Appendix 3A: Evaluating machine learning models	82
Appendix 3B: Prompting engineering	82
References	83
Chapter 4 Analyzing Earnings Conference Calls.....	84
4.1. Data structure in earnings conference calls	84
 Data structure of conference calls	84
4.2. Standard dependence parser	86
 Textual analysis: sentence level analysis using NLP parser	86
4.3. Empirical example: Measuring corporate culture using machine learning	90
 Empirical examples: business application of using conference call transcripts	89
4.4. Empirical example: From words to syntax: Identifying context-specific information in textual analysis	91
Appendix 4: Applying GPT to analyze conference call transcripts using both API and web interface	94
 Applying GPT to analyze conference call transcripts	94
Chapter 5 Analyzing Material Company News.....	96
5.1 Data structure in 8-K filings.....	96
 Data structure of the 8-K filing	96
5.2. Empirical example: Technological peer pressure and product disclosure	106
 Empirical example: Technological peer pressure and product disclosure.....	106
5.3. Empirical example: A game of disclosing “other events”	107
 Empirical example: A game of disclosing “other events”	107
References	109
Chapter 6 Analyzing Data from Social Media.....	110
6.1. What is social media?.....	110
 What is social media?.....	110
6.2. Data from social media	112
 Data from social media platforms	112
6.3. Empirical Example: Negative Peer Disclosure	120
 Empirical example: Negative peer disclosure.....	118
References	122
Chapter 7 Data Analytics in Environmental, Social, and Governance	124
7.1. Corporate Governance	124
 Corporate governance: data and technology.....	124
7.2. Textual data for corporate governance	126
 Textual data for corporate governance	126
 Environmental, social, and governance (ESG) disclosures.....	126
 Analytics of Regulators’ comments and IPO.....	126
7.3. Emerging technologies as governance mechanisms	132
 Emerging technologies as governance mechanisms	132

7.4. Empirical example: Auditing and blockchain	136
 Empirical example: Auditing and blockchain.....	136
References	139
Chapter 8 Analyzing Unstructured Data from Fund Managers.....	141
8.1. Mutual fund disclosure in Form N-CSR	141
 Mutual fund disclosure in Form N-CSR.....	141
8.2. Empirical example: Extracting fund managers' private information and risk assessment from mutual fund shareholder reports.....	146
 Empirical example 1: Extracting fund managers' private information and risk assessment.....	146
 Empirical example 2: Extracting fund managers' private information and risk assessment.....	146
References	151
Chapter 9 Analyzing Image Data	152
9.1. Images in corporate executive presentations.....	152
 Images in corporate executive presentations.....	152
9.2. Empirical example: Visual information in the age of AI	154
 Empirical example: Visual information in the age of AI	154
References	158
Chapter 10 Analyzing the Balance Sheet	159
10.1. Data structure in balance sheets.....	159
 Data structure of the balance sheet	159
 Debate about fair value accounting	161
10.2. Empirical example: Analyzing data in the balance sheet.....	162
 Analyzing the balance sheet	162
 An empirical example of analyzing the balance sheet	162
10.3. Machine learning applications for balance sheet data.....	170
Appendix 10. Regression Methods	174
 An overview of regressions	174
 Three examples of regressions	174
 Fama-Macbeth regression	175
 Two-way sorting	175
 Risk-adjusted return sorting	175
References.....	177
Chapter 11 Analyzing the Income Statement	178
11.1. Data structure in income statements	178
 Data structure of the income statement	178
11.2. Earnings and stock prices.....	180
11.3. Empirical example: Post-earnings announcement drift (PEAD) anomaly.....	183
 An empirical example of analyzing the income statement.....	183
11.4. Machine learning application on income statement data	189
Appendix 11. Key variable explanations	191
References	194

Preface



[Learn about the book](#)

Accounting and finance students with a keen interest in applying artificial intelligence (AI)-based tools in research often encounter a challenge: the need to enroll in separate programming courses that operate independently from their core curriculum. This creates an educational void, compelling our students to juggle the amalgamation of these two disciplines. This book addresses this gap by equipping students with the necessary skills to integrate applied AI with domain-specific data and knowledge in the fields of accounting and finance.

Unlike conventional approaches that commence with programming training, this begins by acquainting students with domain data, textual features, and use cases associated with these data. Relevant emerging technologies are then introduced along with use cases. Upon completing the book, students could be expected to apply AI and machine learning tools to generate and analyze unstructured financial data, such as conference call transcripts, press releases, annual reports, ESG, or other social media disclosures, product/operational images, and fund managers' disclosures. To enhance the learning experience, the book is complemented by a video library, featuring educational videos corresponding to each chapter, including tutorials on how to use the GPT API.

The book offers two flexible learning approaches. Firstly, it can serve as foundational material for the equivalent of a business school graduate-level course, emphasizing the data-driven nature of these disciplines. Secondly, for instructors teaching traditional courses such as financial accounting or corporate finance, relevant chapters (e.g., the chapter on annual report) can function as supplementary material. Prerequisite coding requirements are minimized for both instructors and students throughout the book. Even in instances where coding is necessary, tutorial videos and exercises facilitate a nuanced understanding of coding workflows. Upon finishing this book, students should have a clear picture of diverse use cases of AI tools in the fields of accounting and finance. The materials in this book could also serve as preparation for students to pursue more technical courses.

Forewords

Gareth M. James

We have all seen the power of AI to impact every facet of our lives from picking TV shows on streaming feeds, to self-driving taxis, to generative AI helping with your term paper. But how can an individual without extensive coding skills take advantage of this technology? "Integrating Artificial Intelligence in Accounting and Finance" addresses this challenge for students seeking proficiency in AI applications to finance areas. This book immerses learners in domain-specific data and knowledge, gradually introducing AI and machine learning tools. Minimizing coding prerequisites, it ensures a seamless journey into the intersection of AI and finance, preparing students for diverse technical challenges. By delving into unstructured financial data, from conference call transcripts to social media disclosures, students develop practical skills. Enhanced by a comprehensive video library and accommodating various learning approaches, this book serves as both a foundational guide for modern data-oriented graduate-level courses and supplementary material for more traditional classes.

Kai Li

One of the first endeavors to devote an entire book on applied AI for the finance and accounting audience by researchers who are earlier adoptors of those techniques. A must read to scholars and doctoral students for an introduction to the increasingly important research tools of big data and machine learning.

Andrew Karolyi

The world of finance and accounting is experiencing yet another revolution. This one is about big data analytics. And it is transformative.

Capital markets data is now everywhere. Whether in the form of data feeds from trading platforms and exchanges, mandatory filings by corporates to regulators, or social media or news feed coverage of those companies and markets. It is structured and unstructured, textual and audiovisual. And there are no barriers to managing these data for competitive advantage because AI and big data analytics are advancing as quickly as the various forms of data are being generated.

Portfolio managers, analysts, auditors, corporate financial officers and capital market regulators know they need upskilling in tools and techniques and fast. This dizzying number of new approaches to big data analytics need organizing principles to create logic and order. The new book by Cao, Jiang, and Lei could not have arrived soon enough. It is well-written, accessible to general readers, and the chapters are effectively supplemented with useful videos to showcase methods and techniques. The core of the book focuses on distilling textual data, but my favorite is the back third of the book that showcases how traditional data on balance sheets, income statements and stock returns can be integrated using machine learning approaches.

Acknowledgements

The authors appreciate Jackie Cardello, Will Cong, Ilia Dichev, Jennifer Disharoon, Lawrence Gordon, Gareth M. James, Andrew Karolyi, Kai Li, Mark Liu, Vojislav Maksimovic, Lei Zhou, and seminar participants at Renmin University for helpful feedback and comments on the textbook. Sean Cao appreciates support from the Smith AI initiative from Capital Market Research at University of Maryland and financial support from GRF CPAs & Advisors. We also appreciate the support from FinTech at Cornell Initiative and Digital Economy and Financial Technology (DEFT) Lab.

Chapter 1 Data Analytics in Finance and Accounting

1.1. How to leverage data science for corporate stakeholders

The rising use of big data for decision-making



[The rising use of big data for decision-making](#)

Data analytics is a comprehensive field covering diverse activities that involve collecting, organizing, and analyzing raw data. Propelled by advancements in computing power, mass storage, and machine learning, the utilization of data analytics has surged over the past decade. It possesses the capacity to analyze various types of information, including both structured and unstructured data.

Capital markets produce a plethora of information that is crucial for efficient contracting, risk-sharing, and resource allocation. The nature of a company is shaped by the structure of its contractual arrangements with stakeholders. These contractual relations are essential to companies, forming a network of interconnected contracts involving stakeholders such as customers, suppliers, employees, investors, communities, and others who have a stake in the company. In the traditional system, information provision and consumption centered on corporate disclosure. Managers of a company decide the amount of information to supply by weighing the costs and benefits. Although managers can exert control over what information a company supplies and when, regulatory agencies consistently intervene in this process by establishing a baseline level of information that must be released with various disclosure requirements.

In today's capital markets, information can be generated by and extracted from a multitude of sources, including and beyond mandatory filings required by regulators, companies' voluntary disclosures, information produced by financial analysts, shared by competitors, uncovered by news media, etc. Accordingly, the decision-making process has evolved from a model in which

managers primarily rely on their experience to a data analytics-driven approach. However, the shift presents an inherent challenge: the necessity for large-scale data collection from different sources and in various formats. Data analytics techniques assist stakeholders in collecting relevant information, organizing both structured and unstructured data, and then conducting appropriate analyses to reveal trends and metrics that would otherwise be obscured by the abundance of information. Given the central role that information plays in capital markets, corporate stakeholders increasingly recognize the value and importance of data analytics. As revealed in Cao, Jiang, Yang, and Zhang (2023), there is a discernible uptick in the application of data analytic tools in analyzing regulatory company filings downloadable from the Securities and Exchange Commission's Electronic Data Gathering, Analysis, and Retrieval (EDGAR) database system. Specifically, the proportion of automatic machine downloads of annual and quarterly regulatory company filings (i.e., 10-Ks and 10-Qs) surged from under 40 percent in 2003 to over 80 percent after 2015 (Figure 1).

What separates us from computer science and statistics majors: The importance of domain knowledge

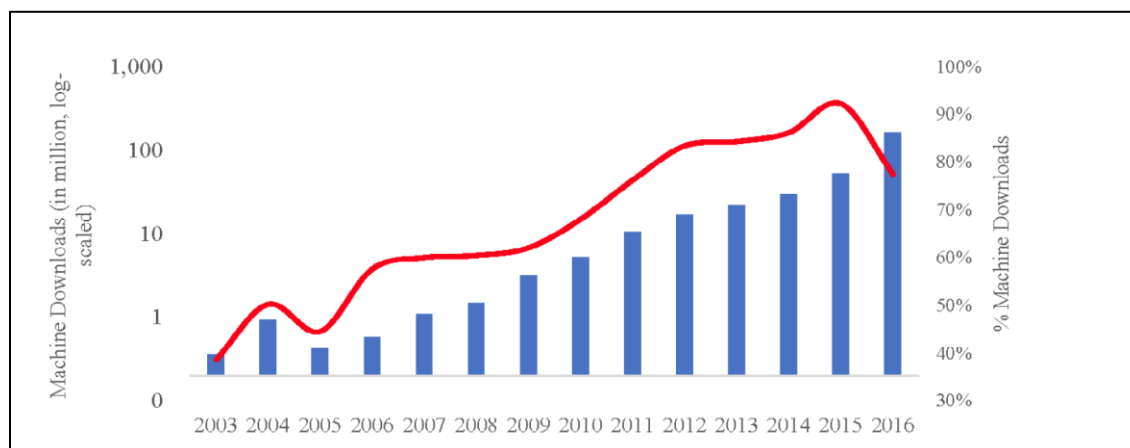


[How to leverage data science for corporate stakeholders: the importance of domain knowledge](#)



[How industries use AI](#)

If computers begin to play an increasingly important role in data analytics, an interesting question arises: can computers outperform humans? High-profile human-computer competitions began in chess. One of the most famous chess computers is Deep Blue because of the chess match between Deep Blue and World Chess Champion Garry Kasparov in 1997.

Figure 1. Machine downloads of 10-K and 10-Q filings

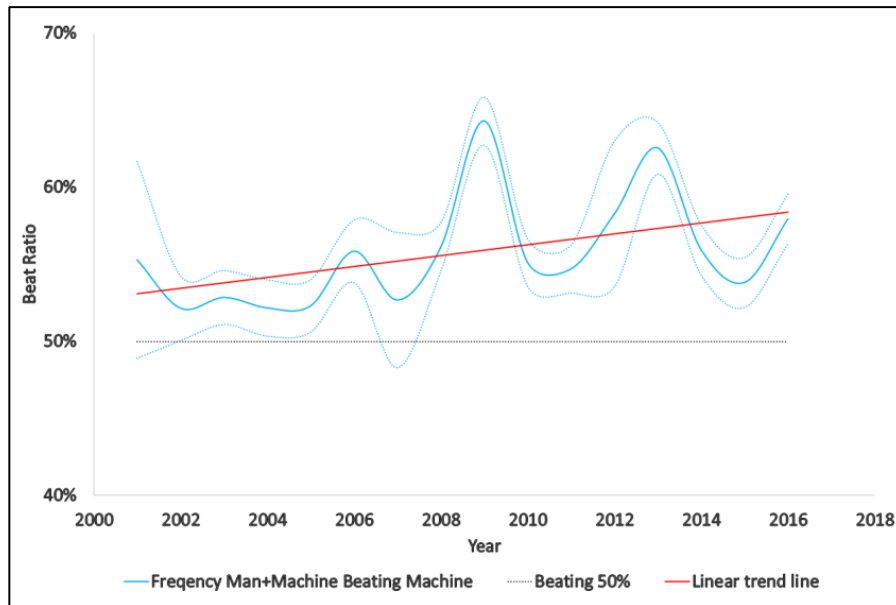
This figure plots the annual number of machine downloads (blue bars and left axis) and the annual percentage of machine downloads over total downloads (red line and right axis) across all 10-K and 10-Q filings from 2003 to 2016. Machine downloads are defined as downloads from an IP address downloading more than 50 unique firms' filings daily. The number of machine downloads and the number of total downloads for each filing are recorded as the respective downloads within seven days after the filing becomes available on EDGAR.

Source: Cao et al. (2023)

Cao, Jiang, Wang, and Yang (2021) build an AI analyst capable of processing corporate financial information, qualitative disclosure, and macroeconomic indicators. They find that an AI analyst built with the currently available technology could indeed beat a majority of human analysts in stock price forecasts (Figure 2). The relative advantage of the AI analyst is more pronounced when the firm is complex, and when information is high-dimensional, transparent and voluminous. Nevertheless, human analysts retain their competitive advantage when critical information requires institutional knowledge. More importantly, the edge of AI over human analysts diminishes over time as human analysts gain access to alternative data and in-house AI resources. Unsurprisingly, combining AI's computational power and the human art of understanding soft information proves to have the highest potential for generating accurate forecasts.

Data analytics techniques can be complicated and rapidly evolving. Yet, the first step in utilizing them is relatively simple: identifying the required information and devising a strategy to gather it. For instance, analysts must understand the objectives of the decision-making process, the pertinent information necessary to facilitate decision-making, and potential sources of useful information. This creates a pressing demand for business professionals who possess both domain knowledge in business *and* a practical understanding of data analytics techniques. Business professionals with both business expertise and data analytics skills can play a critical bridging role that decipher information needs of decision-makers, conduct preliminary analyses, and lead a team to formalize and implement quantitative models.

Figure 2. The performance of AI-assisted analyst vs human analysts



This figure plots the proportion of AI-assisted Analyst recommendations that are more accurate than the Analyst recommendations alone on an annual basis. The blue line in the middle gives the annual AI-assisted Analyst beat ratios, the blue-dotted lines above and below are the 95% confidence interval of the beat ratio, and the red line gives the best linear approximation of the trend in beat ratios.

Source: Cao et al. (2024)

Furthermore, the advancement of Artificial intelligence (AI) opens the avenue to various domain-specific applications in, for example, investing, compliance, marketing, etc. AI

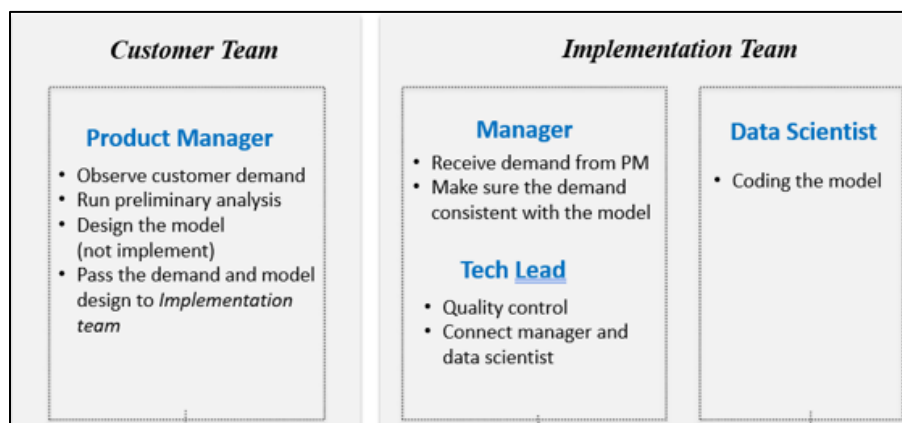
applications do not just help in improving productivity, but also in managing associated legal and security risks. On the other hand, some abilities, such as reasoning-based intelligence, are still unique to humans though could be enhanced by AI. In the new eras, the development of AI will only increase the demand for professionals with domain expertise.

Figure 3 illustrates a typical data analytics team composed of a customer team and an implementation team. The customer team includes the client-facing product manager, who must understand customer demand, perform preliminary analysis, and communicate both the demand and desirable solutions to the implementation team. An ideal product manager should possess business knowledge and be proficient in applying basic data analytics tools. The implementation team consists of a lead team serving as the bridge between the customer team and the implementation team, and a data team specializing in implementation. The lead team requires professionals with strong business knowledge and data analytic skills since they are responsible for receiving demand from the product manager, conveying data analytics solutions to data scientists, and performing quality control. The data team mostly comprises data scientists, with backgrounds in computer science or statistics and strong programming skills. Business professionals with data analytics skills can excel in roles that require integrated skills, such as the product manager and the tech lead.

The objective of this book is to introduce computational tools and AI technologies to business students and to link these tools to business domain knowledge in order to prepare them to study programming. In an era of increasingly data-driven decision-making in business, understanding the domain-specific data features, questions, and use cases of these technologies will be crucial to maintaining a competitive advantage in the market. Let's say you are assigned the task of identifying new products in company reports. Traditional approaches that do not

leverage AI, such as manually searching various websites to obtain product information, tend to be resource-intensive and less focused. Having basic knowledge of computer science would be helpful, but it would only familiarize you with available computational tools and algorithms that do not address specific data sources. To accomplish this task in the most efficient way possible, you need a business-domain-specific understanding of how these tools can be applied. This textbook not only teaches students to select appropriate data sources, such as product descriptions from Item 1 of the company's 10-K filings, but also demonstrates how to effectively extract relevant information, in this case by identifying new product sentences by comparing Item 1 of 10-Ks from two consecutive years. It also explains how to leverage computational tools to implement the method. This integrated approach bridges the gap between business domain knowledge and computer science knowledge and equips readers with the knowledge and skills to solve common real-world problems.

Figure 3. The importance of data analytics and business domain knowledge



Tailoring data science to the needs of different corporate stakeholders

Data analytics for corporate talent

Managers and **employees** are heavily invested in their company's current and future financial well-being. This creates a robust demand for information on the company's operating and financial

condition, including profitability, and future prospects, as well as comparative information on competing peers and business opportunities. Such information is also required to design compensation and incentive contracts. Information extracted using data analytics tools could assist managers in addressing all these questions, including, for example:

- What product lines, geographic areas, or other segments are performing well in comparison to our peer companies and our own benchmarks?
- Should we consider expanding or contracting our business?
- How will current profit levels impact incentive- and share-based compensation?
- What capital structure is suitable for our business?
- How to improve cash flow management?
- What is an appropriate dividend payout policy?
- How are we doing compared to competitors?

Data analytics for shareholders

Suppliers of capital to the companies, including **shareholders** and **creditors**, rely on data analytics to gain insights into the company's financial health, performance, and risks to ascertain a proper level of cost of capital for the firm. Expectations of future profitability and cash generation impact a company's stock price and its ability to borrow money on favorable terms. Investors also use company information to evaluate managerial performance. Here are some examples of questions that information extracted using data analytics could assist investors in addressing:

- What are the expected future profits, cash flows, and dividends for input into stock-price models?
- Is the company financially solvent and able to meet its financial obligations?

- How do expectations about the economy, interest rates, and the competitive environment affect the company?
- Is company management demonstrating good stewardship of the resources to which they have been entrusted?
- Do we have the information we need to critically evaluate strategic initiatives proposed by management?

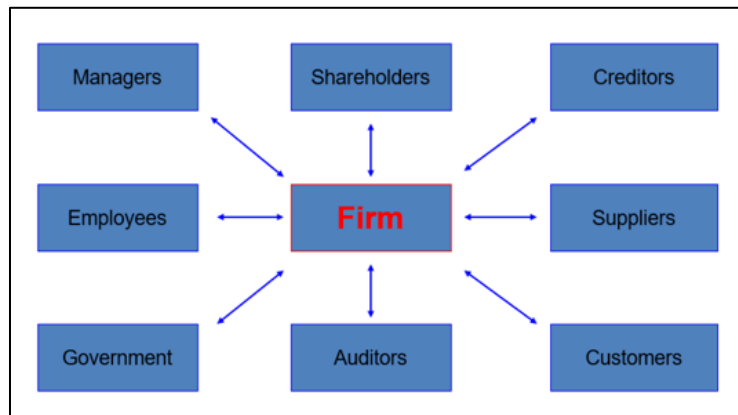
Data analytics for supply chain partners

Suppliers and **customers** need company information to make important decisions regarding their financial transactions and business relationships. For example, suppliers can use data analytics to establish credit terms and evaluate their long-term commitment to supply-chain relations. They also use company information to monitor and adjust their contracts and commitments. Customers seek company information to assess a company's ability to provide products or services, as well as its staying power and reliability. Here are some examples of questions that creditors and suppliers can address with the help of data analytics:

- Should we extend credit in the form of a loan or line of credit for inventory purchases?
- Should we procure raw materials from this supplier?

Data analytics for other stakeholders

Auditors rely on company information to detect potential financial misstatements. **Governments** demand company information to ensure compliance with laws and regulations. As illustrated in Figure 4, all stakeholders can uncover critical insights for their decision-making through the application of data analytics.

Figure 4. Firm as a nexus of contracts

1.2. Overview of structured and unstructured data



[An overview of available business data](#)



[An overview of unstructured and structured data analytics](#)

Unstructured data analytics

Advances in data analytics techniques have led to new sources of information that enable us to tackle a broader range of problems. Recently, there have also been innovations in the methods used to create new data. Information collected from these fresh sources or generated through these new mechanisms is largely **unstructured qualitative data**.

More and more information users are drawn to **texts in firm regulatory filings**. Despite containing financial statements and pages of tables and charts, annual reports and other company regulatory filings still consist mostly of text. The recent reduction in computer storage costs and an increase in computer processing capabilities have made textual analysis of these disclosures more feasible. Regulatory filings that are widely available for analysis include annual reports, current reports, proxy statements, initial public offering (IPO) prospectuses, and more.

Conference call transcripts enable analysts to capture and analyze information disclosed during corporate conference calls. These calls provide an opportunity for managers to announce and discuss the firm's financial performance, while allowing analysts and investors to pose relevant questions about the company.

Environmental, social, and responsibility (ESG) reports serve as internal and external communications detailing a company's ESG initiatives and their impact on the environment and society. While some countries mandate the annual publication of ESG reports, many companies in regions without such requirements also voluntarily release them.

Social media has become a vital communication channel for businesses. Social media platforms enable the rapid dissemination of information to millions of people within seconds. This evolution in information exchange has opened up a new range of opportunities for companies to inform and interact with stakeholders. Company information shared on social media platforms, whether by the company itself or by investors, consumers, competitors, and others, offers an additional perspective on a company's operations, performance, and risks.

Audio data pertaining to business activities can also enhance decision-making processes. For example, in addition to analyzing textual transcripts of conference calls and investment presentations, audio recordings of these events can provide valuable nuances to analysts.

Video and image data are more widely used than ever before due to the progress of video and image capture devices. Development in computer algorithms enables the processing and interpretation of static images and the derivation of objective information from videos. Product-related images provided by companies or shared by customers are examples of potentially valuable image data. Videos of investor presentations, product releases, and other company events could be valuable for managers, investors, and other decision-makers.

Structured data analytics

Traditionally, company information is financial in nature and comprises **structured quantitative data** aggregated and used to prepare financial statements for internal and external information users. Information intermediaries and other marketplace agents also produce company information for capital market participants.

Financial statements provide critical financial information in accordance with applicable accounting standards to ensure the relevancy, reliability, and comparability of firm information. Companies use four financial statements to periodically report on their business activities: the balance sheet, income statement, statement of stockholders' equity, and statement of cash flows. The balance sheet reports on a company's financial position at a specific point in time, while the income statement, statement of stockholders' equity, and statement of cash flows report on performance over a period of time. These three statements link the balance sheet from the beginning to the end of a period.

Executive compensation disclosures provide information concerning the amount and type of compensation paid to a firm's chief executive officer, chief financial officer, and the other three highest-paid executive officers. The company must also disclose the criteria used for executive compensation decisions and the relationship between the company's executive compensation practices and corporate performance. The Summary Compensation Table, included in the proxy statement, is the cornerstone of the required disclosures. The Summary Compensation Table provides a comprehensive overview of a company's executive compensation practices in a single chart. It is followed by other tables and disclosures with more specific information on the components of compensation for the last completed fiscal year, for example, information about

grants of stock options, stock appreciation rights, long-term incentive plan awards, pension plans, employment contracts and related arrangements.

Financial analyst forecasts and recommendations provide useful processed information from financial experts. Financial analysts provide short-term and long-term forecasts on various financial metrics, including earnings, sales, capital expenditures, etc. Financial analysts also regularly issue investment recommendations.

Loan agreements include details such as loan terms, amounts, interest rates, and required collateral. Loan agreements often include contractual requirements, called covenants, that restrict a company's behavior in some fashion. Violation of loan covenants can lead to early repayment or other compensation demanded by the lender. Information in loan agreements thus reflects the creditor's assessment of a company's credit risk.

Patents and citations are valuable for understanding the innovations a company has in development. Patent filings protect the most significant discoveries, providing a wealth of technological, geographical, and industry data. The relationship between an invention's economic importance and patent data is well-documented. Unsurprisingly, patent data has become increasingly used in business analytics.

In this book, unstructured textual and image data are discussed in chapter 2 to chapter 9; structured numerical data are introduced in chapter 10 and chapter 11.

Figure 5. Structured and unstructured business data

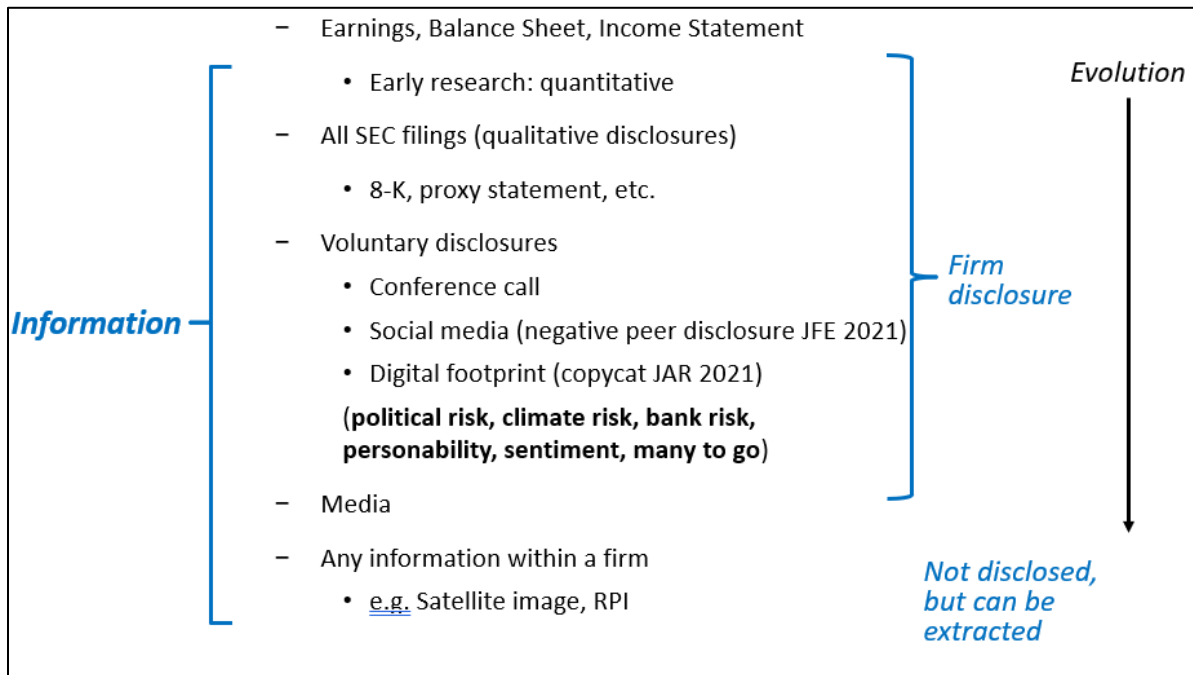
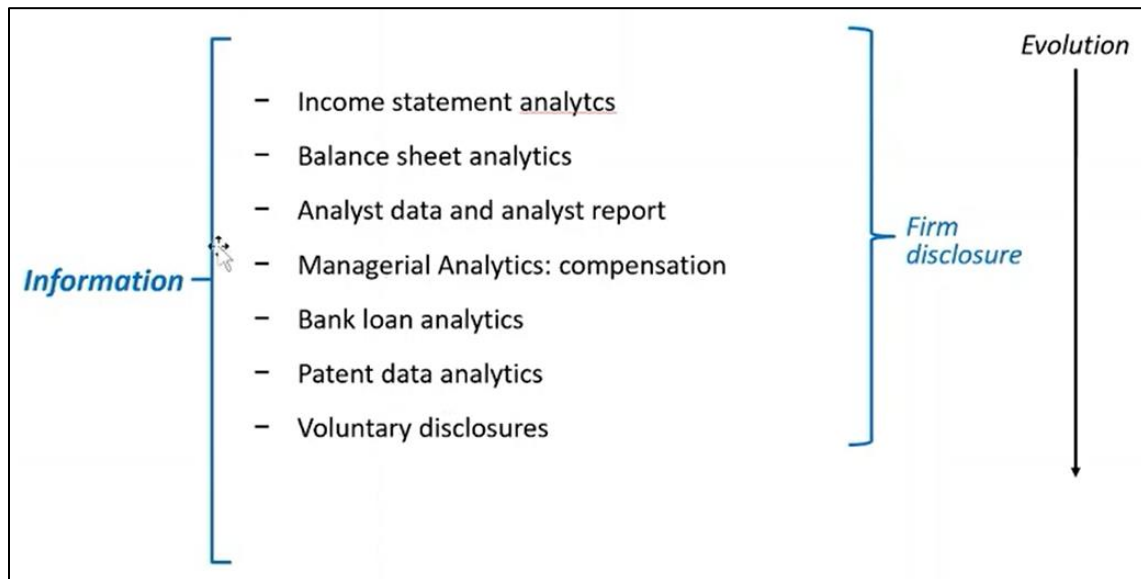


Figure 6. Data analytics based on structured and unstructured business data



1.3. Theory-driven and machine-learning approach of data analytics



[Theory-driven and machine-learning approach of data analytics](#)

Theory-driven approach vs. machine-learning approach

When analyzing either quantitative or qualitative data, two approaches can be taken in data analytics. The **theory (hypothesis)-driven approach** relies on theory-based hypotheses to guide the direction of data analytics, whereas the **machine-learning approach** starts with supplying data to a computer model to train itself to identify patterns or make predictions. The theory-driven approach resembles the human thinking process, making it intuitive and interpretable. In contrast, the machine-learning approach leverages the computational power of machines to yield strong predictive capability, but the machine learning process remains a “black box.”

As an example, the Securities and Exchange Commission (SEC) charged Luckin Coffee Inc. with material misstatement of financial statements to falsely appear to achieve rapid growth and increased profitability and to meet the company’s earnings estimates. The fraud was uncovered by Muddy Waters LLC, an investment research firm specializing in detecting financial fraud. The firm received an anonymous tip and mobilized 92 full-time and 1,418 part-time staff on the ground to run surveillance, recording store traffic for 981 store-days covering 100% of the operating hours of 620 stores. The investigation resulted in more than 11,200 hours of videotaping and led to the conclusion that the number of items per store was inflated by at least 69 percent in the third quarter of 2019 and 88 percent in the fourth quarter. This is a typical investigation following the theory (hypothesis)-driven approach, where Muddy Waters LLC first formed a hypothesis that Luckin Coffee Inc. misstated financial statements and then conducted an investigation to examine the hypothesis. In contrast, financial statement auditors are required to perform analytical procedures that aim at detecting potential anomalies in financial reporting without necessarily having a

hypothesis that a company misstates financial statements. This is an example of machine-learning-driven approach.

While availability of a sea of new data make the machine-learning approach an attractive opportunity, it is important to remember that theory is the guiding principle that helps us think through the patterns uncovered by machine-learning techniques. There is ample room for both theory (hypothesis)-driven approach and machine-learning approach in the field of accounting and finance (Goldstein et al. 2019).

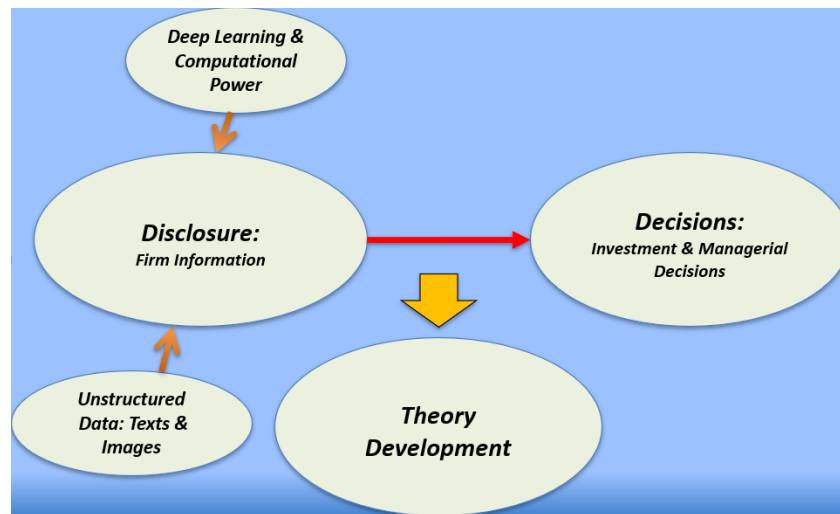
Both approaches can be employed to perform three types of data analytics: descriptive, inference, and predictive. Descriptive analytics summarize data and describe observable patterns, focusing on understanding what has happened over a period of time. Descriptive analytics techniques include descriptive statistics and cross-tabulation of data. However, conventional descriptive analytics rely mostly on structured numeric data. To offer a more complete picture, AI provides the necessary technologies to retrieve novel data from various structured and unstructured sources.

Inference analytics seek to understand what happened and why, involving more diverse data inputs and a deeper dive into the data. Inference analytic techniques include correlations, regressions, and other statistical methods. For instance, inference analytics can be applied to investigate the economic implications of AI applications in accounting and finance, such as, virtual currency, digital payment, peer-to-peer loans, crowdfunding, blockchain, and robo-investing (Goldstein, Jiang, and Karolyi 2019).

Finally, predictive analytics explores what is likely to happen or what will happen “if” something else happens. For example, when faced with new technologies, a critical decision is whether to embrace the new technology, stick with the existing technology, or invest in both.

Would tablets ultimately replace laptops? Would all-electric cars dominate hybrid cars eventually? Chandrasekaran, Tellis, and James (2022) draw on the theory of disruptive change to develop a model of the diffusion of successive technologies that helps managers estimate and predict technological leapfrogging, cannibalization, and coexistence. To predict what will happen, we first need to understand what happened, how, and why; hence, predictive analytics builds on descriptive and diagnostic analytics. Conventional predictive analytic techniques involve building models using past data and statistical techniques, including regression and a deep understanding of cause and effect. Machine-learning algorithms allow patterns to be learned from training a dataset, and predictive models to be built with limited human intervention.

Figure 7. Theory-driven and machine-learning approach of data analytics



1.4. The advantages of applying machine-learning approaches



[The advantage of applying machine-learning approaches](#)

Machine learning has several unique advantages compared with conventional data analysis techniques. First, traditional statistical methods often struggle with large amounts of data. In contrast, machine learning algorithms, such as convolutional neural networks (CNNs), can select

the best features to process information effectively. Another advantage of machine learning is its ability to handle nonlinear relationships in data. Along with the information explosion, the types of information available to analysts have grown from mostly numerical data to more complex data involving both text and images. Machine learning can identify nonlinear patterns and make predictions using various types of data, making it particularly valuable for tasks such as natural language processing and computer vision. Machine learning also offers the ability to make out-of-sample predictions, which is useful in cases where data is limited. In traditional statistical methods, repeated optimization (learning) is often required to improve the accuracy of the model. However, in machine learning, this process is streamlined and only requires one-time optimization. For tasks involving time series data, machine learning algorithms, such as long short-term memory (LSTM) networks, can be applied to identify time-series patterns. Finally, machine learning algorithms are designed to be efficient, which is important given the increasing amounts of data being generated. By using optimized algorithms, machine learning can process data quickly and effectively, making it a valuable tool in fields such as healthcare and finance, where time is of the essence.

Machine learning techniques include supervised and unsupervised learning, self-supervised learning, transfer learning, ensemble learning, and more. Supervised learning involves training a model with labeled data. On the contrary, unsupervised learning does not require labeled data and instead focuses on finding patterns and relationships within the data itself. Self-supervised learning is a variant of unsupervised learning that uses the data itself to generate labels, such as using stock returns to label news positivity. Transfer learning is a powerful technique that uses a pre-trained model to tackle new tasks with less training data. This approach is useful in cases where the cost of obtaining labeled data is high or where the amount of labeled data available is limited. These machine-learning techniques are discussed in detail in later chapters in the book.

Although machine-learning techniques are powerful tools for analyzing data, we should still keep the *Occam's razor* principle in mind: we should try the simpler models as well as more complex machine-learning techniques, and make an informed tradeoff between performance and complexity (James, Witten, Hastie, and Tibshirani 2023).

References

- Cao, S., Jiang, W., Wang, J., and Yang, B. 2024. From man vs. machine to man + machine: The art and AI of stock analysis. *Journal of Financial Economics*, 160, 103910.
- Cao, S., Jiang, W., Yang, B., and Zhang, A. 2023. How to talk when a machine is listening? Corporate disclosure in the age of AI. *Review of Financial Studies*, 36(9), 3603-3642.
- Chandrasekaran, D., Tellis, G., and James, G. 2022. Leapfrogging, cannibalization, and survival during disruptive technological change: The critical role of rate of disengagement. *Journal of Marketing*, 86(1), 149-166.
- Goldstein, I., Jiang, W., and Karolyi, G.A. 2019. To Fintech and beyond. *Review of Financial Studies*, 32(5), 1647-1661.
- James, G., Witten, D., Hastie, T., and Tibshirani, R. 2023. An introduction to statistical learning. Springer.

Chapter 2 Analyzing Annual Reports

2.1. Data structure in annual reports and 10-K filings

[Data structure of the 10-K filing](#)

Each year, U.S. public companies are required to produce a Form 10-K and file it with the U.S. Securities and Exchange Commission (SEC) within 60 days of the end of the fiscal year. The SEC's Electronic Data Gathering, Analysis, and Retrieval (EDGAR) database system allows anyone to retrieve a company's 10-K report. Some companies also post their 10-K reports on their websites. In addition, SEC rules mandate that companies send an annual report to their shareholders in advance of annual meetings. While both annual reports and 10-K filings provide an overview of the company's performance for the given fiscal year, annual reports tend to be much more visually appealing than 10-K filings. Companies put effort into designing their annual reports, using graphics and images to communicate data, while 10-K filings only report numbers and other qualitative information, devoid of design elements or additional flair.

2.1.1. Data structure in 10-K filings

[Item 1 Business description](#)

[Item 1A Risk disclosure](#)

[Item 7 Management discussion and analysis](#)

A comprehensive Form 10-K contains four parts and 15 items. Researchers are often most interested in Item 1, "Business," Item 1A, "Risk Factors," and Item 7, "Management's Discussion and Analysis of Financial Condition and Results of Operations." Therefore, we begin our discussion with these three items of Form 10-K.

Item 1, "Business," appears in Part I. It gives a detailed description of the company's business, including its main products and services, its subsidiaries, and in which markets it operates. To gain

an understanding of a company's operations and its primary products and services, Item 1 serves as an excellent starting point. Figure 1 shows an excerpt of Item 1 of Apple Inc.'s 2021 10-K. It introduces the company's background and provides information on Apple's main products.

Figure 1. Excerpt from Item 1 of Apple Inc.'s 2021 Form 10-K

<p>PART I</p> <p>Item 1. Business</p> <p>Company Background</p> <p>The Company designs, manufactures and markets smartphones, personal computers, tablets, wearables and accessories, and sells a variety of related services. The Company's fiscal year is the 52- or 53-week period that ends on the last Saturday of September. The Company is a California corporation established in 1977.</p> <p>Products</p> <p><i>iPhone</i></p> <p>iPhone® is the Company's line of smartphones based on its iOS operating system. During 2020, the Company released a new iPhone SE. In October 2020, the Company announced four new iPhone models with 5G technology: iPhone 12 and iPhone 12 Pro were available starting in October 2020, and iPhone 12 Pro Max and iPhone 12 mini are both expected to be available in November 2020.</p> <p><i>Mac</i></p> <p>Mac® is the Company's line of personal computers based on its macOS® operating system. During 2020, the Company released a new 16-inch MacBook Pro®, a fully redesigned Mac Pro®, and updated versions of its MacBook Air®, 13-inch MacBook Pro and 27-inch iMac®.</p> <p><i>iPad</i></p> <p>iPad® is the Company's line of multi-purpose tablets based on its iPadOS® operating system. During 2020, the Company released an updated iPad Pro®. In September 2020, the Company released an eighth-generation iPad and introduced an all-new iPad Air®, which was available starting in October 2020.</p> <p><i>Wearables, Home and Accessories</i></p> <p>Wearables, Home and Accessories includes AirPods®, Apple TV®, Apple Watch®, Beats® products, HomePod®, iPod touch® and other Apple-branded and third-party accessories. AirPods are the Company's wireless headphones that interact with Siri®. During 2020, the Company released AirPods Pro®. Apple Watch is the Company's line of smart watches based on its watchOS® operating system. In September 2020, the Company released Apple Watch Series 6 and a new Apple Watch SE. In October 2020, the Company announced HomePod mini™, which is expected to be available in November 2020.</p>
--

Item 1A, "Risk Factors," is also found in Part I of Form 10-K. It outlines the most significant risks faced by the company or its securities. In practice, this section focuses on the risks themselves, not on how the company addresses those risks. The outlined risks may pertain to the entire economy or market, the company's industry sector or geographic region, or be unique to the company itself. Figure 2 shows an excerpt of Item 1A of Apple Inc.'s 2021 10-K, which discusses business risks arising from the COVID-19 pandemic, such as disruptions in the supply chain and logistical services and store closures.

Item 7, "Management's Discussion and Analysis of Financial Condition and Results of Operations," presents the company's perspective on its financial performance during the prior

fiscal year. This section, commonly referred to as the MD&A, allows company management to summarize its recent business in its own words. The MD&A presents:

- The company's operations and financial results, including information about the company's liquidity and capital resources and any known trends or uncertainties that could materially affect the company's results. This section may also present the management's views on key business risks and how they are being addressed.
- Material changes in the company's results compared to the prior period, as well as off-balance-sheet arrangements and contractual obligations.
- Critical accounting judgments, such as estimates and assumptions.

Figure 3 is an excerpt from Item 7 of Target's 2021 10-K. It begins with highlights of the fiscal year, provides a summary of financial outcomes, and then continues to analyze key performance indicators, such as the gross margin.

Figure 2. Excerpt from Item 1A of Apple Inc.'s 2021 Form 10-K

<p>Item 1A. Risk Factors</p> <p>The following discussion of risk factors contains forward-looking statements. These risk factors may be important to understanding other statements in this Form 10-K. The following information should be read in conjunction with Part II, Item 7, "Management's Discussion and Analysis of Financial Condition and Results of Operations" and the consolidated financial statements and accompanying notes in Part II, Item 8, "Financial Statements and Supplementary Data" of this Form 10-K.</p> <p>The business, financial condition and operating results of the Company can be affected by a number of factors, whether currently known or unknown, including but not limited to those described below, any one or more of which could, directly or indirectly, cause the Company's actual financial condition and operating results to vary materially from past, or from anticipated future, financial condition and operating results. Any of these factors, in whole or in part, could materially and adversely affect the Company's business, financial condition, operating results and stock price.</p> <p>Because of the following factors, as well as other factors affecting the Company's financial condition and operating results, past financial performance should not be considered to be a reliable indicator of future performance, and investors should not use historical trends to anticipate results or trends in future periods.</p> <p><i>The Company's business, results of operations, financial condition and stock price have been adversely affected and could in the future be materially adversely affected by the COVID-19 pandemic.</i></p> <p>COVID-19 has spread rapidly throughout the world, prompting governments and businesses to take unprecedented measures in response. Such measures have included restrictions on travel and business operations, temporary closures of businesses, and quarantines and shelter-in-place orders. The COVID-19 pandemic has significantly curtailed global economic activity and caused significant volatility and disruption in global financial markets.</p> <p>The COVID-19 pandemic and the measures taken by many countries in response have adversely affected and could in the future materially adversely impact the Company's business, results of operations, financial condition and stock price. Following the initial outbreak of the virus, the Company experienced disruptions to its manufacturing, supply chain and logistical services provided by outsourcing partners, resulting in temporary iPhone supply shortages that affected sales worldwide. During the course of the pandemic, the Company's retail stores, as well as channel partner points of sale, have been temporarily closed at various times. In many cases, where stores and points of sale have reopened they are subject to operating restrictions to protect public health and the health and safety of employees and customers. The Company has at times required substantially all of its employees to work remotely.</p> <p>The Company is continuing to monitor the situation and take appropriate actions in accordance with the recommendations and requirements of relevant authorities. The full extent of the impact of the COVID-19 pandemic on the Company's operational and financial performance is currently uncertain and will depend on many factors outside the Company's control, including, without limitation, the timing, extent, trajectory and duration of the pandemic, the development and availability of effective treatments and vaccines, the imposition of and compliance with protective public safety measures, and the impact of the pandemic on the global economy and demand for consumer products. Additional future impacts on the Company may include, but are not limited to, material adverse effects on: demand for the Company's products and services; the Company's supply chain and sales and distribution channels; the Company's ability to execute its strategic plans; and the Company's profitability and cost structure.</p> <p>To the extent the COVID-19 pandemic adversely affects the Company's business, results of operations, financial condition and stock price, it may also have the effect of heightening many of the other risks described in this Part I, Item 1A of this Form 10-K.</p>
--

Figure 3. Excerpt from Item 7 of Target's 2021 Form 10-K

Item 7. Management's Discussion and Analysis of Financial Condition and Results of Operations

Executive Overview

We continue to make strategic investments to support our durable operating and financial model that further differentiates Target and is designed to drive sustainable sales and profit growth. During 2021, in support of our enterprise strategy described in [Item 1 on page 2](#) of this Form 10-K, we:

- Expanded our digital fulfillment capabilities, including adding permanent storage capacity in more than 200 high-volume stores, adding thousands of new items to the list available for Order Pickup and Drive Up, and doubling the number of Drive Up parking stalls compared with last year. During 2021, over 50 percent of our digital sales were fulfilled by our same-day fulfillment options: Order Pickup, Drive Up, and delivery via ShipIt.
- Continued the steady stream of newness across our assortment and continued to introduce new owned brands, including our arts and crafts owned brand, Mondo Llama™, our sweet and savory food brand, Favorite Day™, our pet food brand, Kindful™, and our first dedicated storage and home organization owned brand, Brightroom™. For the first time in history, 11 brands delivered \$1 billion or more in sales, with 4 brands delivering over \$2 billion in sales, driven by strength in Apparel, Home Furnishings & Decor and Food & Beverage.
- Launched Ulta Beauty at Target on Target.com and in about 100 Target locations, and expanded our Apple and Disney experiences.
- Remodeled 145 stores.
- Opened 32 new stores, including 28 additional small format stores in key urban markets and on college campuses.
- Invested significantly in our team, including recognition bonuses and launch of a new debt-free education assistance program.

Financial Summary

2021 included the following notable items:

- GAAP diluted earnings per share were \$14.10.
- Adjusted diluted earnings per share were \$13.55.
- Total revenue increased 13.3 percent, driven by an increase in comparable sales.
- Comparable sales increased 12.7 percent, driven by a 12.3 percent increase in traffic:
 - Comparable store originated sales grew 11.0 percent.
 - Comparable digitally originated sales increased 20.8 percent.
- Operating income of \$8.9 billion was 35.8 percent higher than the comparable prior-year period.
- We recognized a \$335 million pretax gain on the sale of Dermstore.

Sales were \$104.6 billion for 2021, an increase of \$12.2 billion, or 13.2 percent, from the prior year. Operating cash flow provided by continuing operations was \$8.6 billion for 2021, a decrease of \$(1.9) billion, or (18.1) percent, from \$10.5 billion for 2020. The drivers of the operating cash flow decrease are described on [page 27](#).

Earnings Per Share From Continuing Operations	2021	2020	2019	Percent Change	
				2021/2020	2020/2019
GAAP diluted earnings per share	\$ 14.10	\$ 8.84	\$ 8.34	63.1 %	36.3 %
Adjustments	(0.53)	0.78	0.05		
Adjusted diluted earnings per share	\$ 13.55	\$ 9.42	\$ 8.39	44.0 %	47.4 %

Note: Amounts may not foot due to rounding. Adjusted diluted earnings per share from continuing operations (Adjusted EPS), a non-GAAP metric, excludes the impact of certain items. Management believes that Adjusted EPS is useful in providing period-to-period comparisons of the results of our continuing operations. A reconciliation of non-GAAP financial measures to GAAP measures is provided on [page 24](#).

We report after-tax return on invested capital (ROIC) from continuing operations because we believe ROIC provides a meaningful measure of our capital-allocation effectiveness over time. For the trailing twelve months ended January 29, 2022, after-tax ROIC was 33.1 percent, compared with 23.5 percent for the trailing twelve months ended January 30, 2021. The calculation of ROIC is provided on [page 28](#).

Gross Margin Rate

Our gross margin rate was 28.3 percent in 2021 and 28.4 percent in 2020. This decrease reflected the net impact of:

- supply chain pressure related to increased compensation and headcount in our distribution centers, partially offset by the small net benefit of a higher percentage of digital sales fulfilled through our lower-cost same-day fulfillment options;
- higher merchandise and freight costs partially offset by historically low promotional and clearance markdown rates; and
- favorable mix in the relative growth rates of higher and lower margin categories.

Other Items in Form 10-K

Part I of Form 10-K

Part I of the report comprises two additional items in addition to Item 1 and Item 1A. Item 1B, “Unresolved Staff Comments,” requires the company to explain certain comments received from SEC staff on previously filed reports that have not been resolved after an extended period of time.

Item 2, “Properties,” describes the company’s significant physical properties, such as principal plants, mines, and other materially important physical properties. Figure 4 displays Item 2 from

Apple Inc.'s 2021 10-K. It reveals that Apple Inc. owns and leases facilities and land within the U.S. and outside the U.S.

Figure 4. Excerpt from Item 2 of Apple Inc.'s 2021 Form 10-K

Item 2. Properties

The Company's headquarters are located in Cupertino, California. As of September 26, 2020, the Company owned or leased facilities and land for corporate functions, R&D, data centers, retail and other purposes at locations throughout the U.S. and in various places outside the U.S. The Company believes its existing facilities and equipment, which are used by all reportable segments, are in good operating condition and are suitable for the conduct of its business.

Item 3, "Legal Proceedings," requires companies to disclose information about significant pending lawsuits or other legal proceedings other than ordinary litigation. Figure 5 displays Item 3 from Apple Inc.'s 2021 10-K. It is worth noting that it is not uncommon for companies to be involved in legal proceedings. Item 4 has no required information and is reserved by the SEC for future rulemaking.

Figure 5. Excerpt from Item 3 of Apple Inc.'s 2021 Form 10-K

Item 3. Legal Proceedings

The Company is subject to legal proceedings and claims that have not been fully resolved and that have arisen in the ordinary course of business. The Company's material legal proceedings are described in Part II, Item 8 of this Form 10-K in the Notes to Consolidated Financial Statements in Note 10, "Commitments and Contingencies" under the heading "Contingencies."

The outcome of litigation is inherently uncertain. If one or more legal matters were resolved against the Company in a reporting period for amounts above management's expectations, the Company's financial condition and operating results for that reporting period could be materially adversely affected. The Company settled certain matters during the fourth quarter of 2020 that did not individually or in the aggregate have a material impact on the Company's financial condition or operating results.

Part II of Form 10-K

Part II of Form 10-K comprises seven items in addition to Item 7. Item 5, "Market for Registrant's Common Equity, Related Stockholder Matters and Issuer Purchases of Equity Securities," provides information about the company's equity securities, including market information, the number of shareholders, dividends, stock repurchases by the company, and other relevant information. Figure 6 provides an example of Item 5 in Apple Inc.'s 2021 10-K.

Item 6, “Selected Financial Data,” provides a summary of certain financial information from the past five years. As shown in Figure 7, Item 6 of Apple Inc.’s 2021 Form 10-K reports selected financial information from 2016 to 2021. More detailed financial information for the past three years is included in a separate section: Item 8, “Financial Statements and Supplementary Data,” which includes the company’s balance sheet, income statement, cash flow statement, and notes to the financial statements.

Figure 6. Excerpt from Item 5 of Apple Inc.’s 2021 Form 10-K

PART II				
Item 5. Market for Registrant's Common Equity, Related Stockholder Matters and Issuer Purchases of Equity Securities				
The Company's common stock is traded on The Nasdaq Stock Market LLC under the symbol AAPL.				
Common Stock Split				
On August 28, 2020, the Company effected a four-for-one stock split to shareholders of record as of August 24, 2020. All share, restricted stock unit ("RSU") and per share or per RSU information has been retroactively adjusted to reflect the stock split.				
Holders				
As of October 16, 2020, there were 22,797 shareholders of record.				
Purchases of Equity Securities by the Issuer and Affiliated Purchasers				
Share repurchase activity during the three months ended September 26, 2020 was as follows (in millions, except number of shares, which are reflected in thousands, and per share amounts):				
Periods	Total Number of Shares Purchased	Average Price Paid Per Share	Total Number of Shares Purchased as Part of Publicly Announced Plans or Programs	Approximate Dollar Value of Shares That May Yet Be Purchased Under the Plans or Programs ⁽¹⁾
June 28, 2020 to August 1, 2020:				
Open market and privately negotiated purchases	67,990	\$ 94.68	67,990	
August 2, 2020 to August 29, 2020:				
May 2020 ASR	3,115	⁽²⁾	3,115	
Open market and privately negotiated purchases	40,004	\$ 115.99	40,004	
August 30, 2020 to September 26, 2020:				
Open market and privately negotiated purchases	60,725	\$ 114.00	60,725	
Total	171,834			\$ 56,353
(1) As of September 26, 2020, the Company was authorized to purchase up to \$225 billion of the Company's common stock under a share repurchase program announced on April 30, 2020, of which \$168.6 billion had been utilized. The remaining \$56.4 billion in the table represents the amount available to repurchase shares under the authorized repurchase program as of September 26, 2020. The Company's share repurchase program does not obligate it to acquire any specific number of shares. Under this program, shares may be repurchased in privately negotiated and/or open market transactions, including under plans complying with Rule 10b5-1 under the Exchange Act.				
(2) In May 2020, the Company entered into an accelerated share repurchase arrangement ("ASR") to purchase up to \$6.0 billion of the Company's common stock. In August 2020, the purchase period for this ASR ended and an additional 3 million shares were delivered and retired. In total, 64 million shares were delivered under this ASR at an average repurchase price of \$94.14.				

Item 7A, “Quantitative and Qualitative Disclosures about Market Risk,” mandates disclosure of the company’s exposure to market risks arising from, for example, fluctuations in interest rates, foreign currency exchanges, commodity prices, or equity prices. This section may also include

information on how the company manages these risks. Figure 8 provides an excerpt from Item 7A in Apple Inc.'s 2021 Form 10-K.

Figure 7. Excerpt from Item 6 of Apple Inc.'s 2021 Form 10-K

Item 6. Selected Financial Data					
The information set forth below for the five years ended September 26, 2020, is not necessarily indicative of results of future operations, and should be read in conjunction with Part II, Item 7, "Management's Discussion and Analysis of Financial Condition and Results of Operations" and the consolidated financial statements and accompanying notes included in Part II, Item 8 of this Form 10-K to fully understand factors that may affect the comparability of the information presented below (in millions, except number of shares, which are reflected in thousands, and per share amounts).					
	2020	2019	2018	2017	2016
Total net sales	\$ 274,515	\$ 260,174	\$ 265,595	\$ 229,234	\$ 215,639
Net income	\$ 57,411	\$ 55,256	\$ 59,531	\$ 48,351	\$ 45,687
Earnings per share:					
Basic	\$ 3.31	\$ 2.99	\$ 3.00	\$ 2.32	\$ 2.09
Diluted	\$ 3.28	\$ 2.97	\$ 2.98	\$ 2.30	\$ 2.08
Cash dividends declared per share	\$ 0.795	\$ 0.75	\$ 0.68	\$ 0.60	\$ 0.545
Shares used in computing earnings per share:					
Basic	17,352,119	18,471,336	19,821,510	20,868,968	21,883,281
Diluted	17,528,214	18,595,651	20,000,435	21,006,767	22,001,126
Total cash, cash equivalents and marketable securities	\$ 191,830	\$ 205,898	\$ 237,100	\$ 268,895	\$ 237,585
Total assets	\$ 323,888	\$ 338,516	\$ 365,725	\$ 375,319	\$ 321,686
Non-current portion of term debt	\$ 98,667	\$ 91,807	\$ 93,735	\$ 97,207	\$ 75,427
Other non-current liabilities	\$ 54,490	\$ 50,503	\$ 48,914	\$ 44,212	\$ 39,986

Item 8, "Financial Statements and Supplementary Data," mandates the company's audited financial statements, which include the company's income statement, balance sheet, statement of cash flows, and statement of stockholders' equity. The financial statements are accompanied by notes that elucidate the information presented in the financial statements. An independent accountant audits these financial statements and, for large companies, also reports on their internal controls over financial reporting.

Item 9, "Changes in and Disagreements with Accountants on Accounting and Financial Disclosure," requires companies that have changed accountants to discuss any disagreements they had with those accountants. Such disclosure is often seen as a red flag by many investors. Item 9A, "Controls and Procedures," discloses information about the company's disclosure controls and procedures, as well as its internal controls over financial reporting. Item 9B, "Other Information,"

requires companies to provide any information that should have been reported on another form during the fourth quarter of the year covered by the 10-K but was not disclosed.

Figure 8. Excerpt from Item 7A of Apple Inc.'s 2021 Form 10-K

<p>Item 7A. Quantitative and Qualitative Disclosures About Market Risk</p> <p>Interest Rate and Foreign Currency Risk Management</p> <p>The Company regularly reviews its foreign exchange forward and option positions and interest rate swaps, both on a stand-alone basis and in conjunction with its underlying foreign currency and interest rate exposures. Given the effective horizons of the Company's risk management activities and the anticipatory nature of the exposures, there can be no assurance these positions will offset more than a portion of the financial impact resulting from movements in either foreign exchange or interest rates. Further, the recognition of the gains and losses related to these instruments may not coincide with the timing of gains and losses related to the underlying economic exposures and, therefore, may adversely affect the Company's financial condition and operating results.</p> <p>Interest Rate Risk</p> <p>The Company's exposure to changes in interest rates relates primarily to the Company's investment portfolio and outstanding debt. While the Company is exposed to global interest rate fluctuations, the Company's interest income and expense are most sensitive to fluctuations in U.S. interest rates. Changes in U.S. interest rates affect the interest earned on the Company's cash, cash equivalents and marketable securities and the fair value of those securities, as well as costs associated with hedging and interest paid on the Company's debt.</p> <p>The Company's investment policy and strategy are focused on the preservation of capital and supporting the Company's liquidity requirements. The Company uses a combination of internal and external management to execute its investment strategy and achieve its investment objectives. The Company typically invests in highly rated securities, with the primary objective of minimizing the potential risk of principal loss. The Company's investment policy generally requires securities to be investment grade and limits the amount of credit exposure to any one issuer. To provide a meaningful assessment of the interest rate risk associated with the Company's investment portfolio, the Company performed a sensitivity analysis to determine the impact a change in interest rates would have on the value of the investment portfolio assuming a 100 basis point parallel shift in the yield curve. Based on investment positions as of September 26, 2020 and September 28, 2019, a hypothetical 100 basis point increase in interest rates across all maturities would result in a \$3.1 billion and \$2.8 billion incremental decline in the fair market value of the portfolio, respectively. Such losses would only be realized if the Company sold the investments prior to maturity.</p>

Part III of Form 10-K

Part III of the 10-K includes five items. Item 10, "Directors, Executive Officers and Corporate Governance," requires information about the background and experience of the company's directors and executive officers, the company's code of ethics, and certain qualifications for directors and committees of the board of directors. Item 11, "Executive Compensation," requires detailed disclosures of the company's compensation policies and programs, as well as how much compensation was paid to its top executive officers in the past fiscal year.

In Item 12, "Security Ownership of Certain Beneficial Owners and Management and Related Stockholder Matters," companies provide information about the shares owned by the company's directors, officers, and certain large shareholders. This item also includes information about shares covered by equity compensation plans.

Item 13, “Certain Relationships and Related Transactions, and Director Independence,” includes information about relationships and transactions between the company and its directors, officers, and their family members. It also includes information about whether each director of the company is independent.

Item 14, “Principal Accountant Fees and Services,” requires companies to disclose fees paid to their accounting firm for various types of services during the year. Although this disclosure is required as part of Form 10-K, most companies provide this information in a separate document called the proxy statement. Companies distribute the proxy statement among their shareholders in preparation for annual meetings. If the information was provided in a proxy statement, Item 14 will include a message from the company directing readers to the proxy statement document. The proxy statement is typically filed a month or two after the 10-K. Part III of Apple Inc.’s 2021 10-K is provided in Figure 9 as an example.

Part IV of 10-K

Part IV contains Item 15, “Exhibits, Financial Statement Schedules,” which outlines the financial statements and exhibits included as part of the 10-K filing. Many exhibits are mandatory, including documents such as the company’s bylaws, copies of its material contracts, and a roster of the company’s subsidiaries.

2.1.2. Data structure in annual reports

Similar to 10-Ks, annual reports are comprehensive reports detailing companies’ performance and activities during a fiscal year. Many companies choose to incorporate a lot of graphics and images instead of large amounts of text in their annual reports to create more visually appealing documents than 10-Ks. For example, in Figure 10, Procter and Gamble provide both numeric and graphic information regarding its financial performance in the 2021 annual report.

The structure of annual reports varies across companies, but they typically include several common sections such as: (1) letter to shareholders, (2) performance and highlights, (3) corporate strategies, (4) non-financial information such as environmental, social, and governance (ESG) information, (5) financial information, (6) leadership information, and any other pertinent information the company wishes to share.

Figure 9. Part III of Apple Inc.'s 2021 Form 10-K

<p>PART III</p> <p>Item 10. Directors, Executive Officers and Corporate Governance</p> <p>The information required by this Item is set forth under the headings "Corporate Governance," "Directors," "Executive Officers" and, if applicable, "Other Information—Security Ownership of Certain Beneficial Owners and Management" in the Company's 2021 Proxy Statement to be filed with the SEC within 120 days after September 26, 2020 in connection with the solicitation of proxies for the Company's 2021 annual meeting of shareholders, and is incorporated herein by reference.</p> <p>Item 11. Executive Compensation</p> <p>The information required by this Item is set forth under the heading "Executive Compensation," under the subheadings "Board Oversight of Risk Management" and, if applicable, "Compensation Committee Interlocks and Insider Participation" under the heading "Corporate Governance" and under the subheadings "Compensation of Directors" and "Director Compensation—2020" under the heading "Directors" in the Company's 2021 Proxy Statement to be filed with the SEC within 120 days after September 26, 2020, and is incorporated herein by reference.</p> <p>Item 12. Security Ownership of Certain Beneficial Owners and Management and Related Stockholder Matters</p> <p>The information required by this Item is set forth under the headings "Other Information—Security Ownership of Certain Beneficial Owners and Management" and "Other Information—Equity Compensation Plan Information" in the Company's 2021 Proxy Statement to be filed with the SEC within 120 days after September 26, 2020, and is incorporated herein by reference.</p> <p>Item 13. Certain Relationships and Related Transactions, and Director Independence</p> <p>The information required by this Item is set forth under the subheadings "Role of the Board of Directors," "Board Committees", "Review, Approval, or Ratification of Transactions with Related Persons" and "Transactions with Related Persons" under the heading "Corporate Governance" in the Company's 2021 Proxy Statement to be filed with the SEC within 120 days after September 26, 2020, and is incorporated herein by reference.</p> <p>Item 14. Principal Accountant Fees and Services</p> <p>The information required by this Item is set forth under the subheadings "Fees Paid to Auditors" and "Policy on Audit Committee Pre-Approval of Audit and Non-Audit Services Performed by the Independent Registered Public Accounting Firm" under the heading "Ratification of Appointment of Independent Registered Public Accounting Firm" in the Company's 2021 Proxy Statement to be filed with the SEC within 120 days after September 26, 2020, and is incorporated herein by reference.</p>

2.2. Conventional textual analysis approach

2.2.1. Conventional Approach Review (Bag of Words)

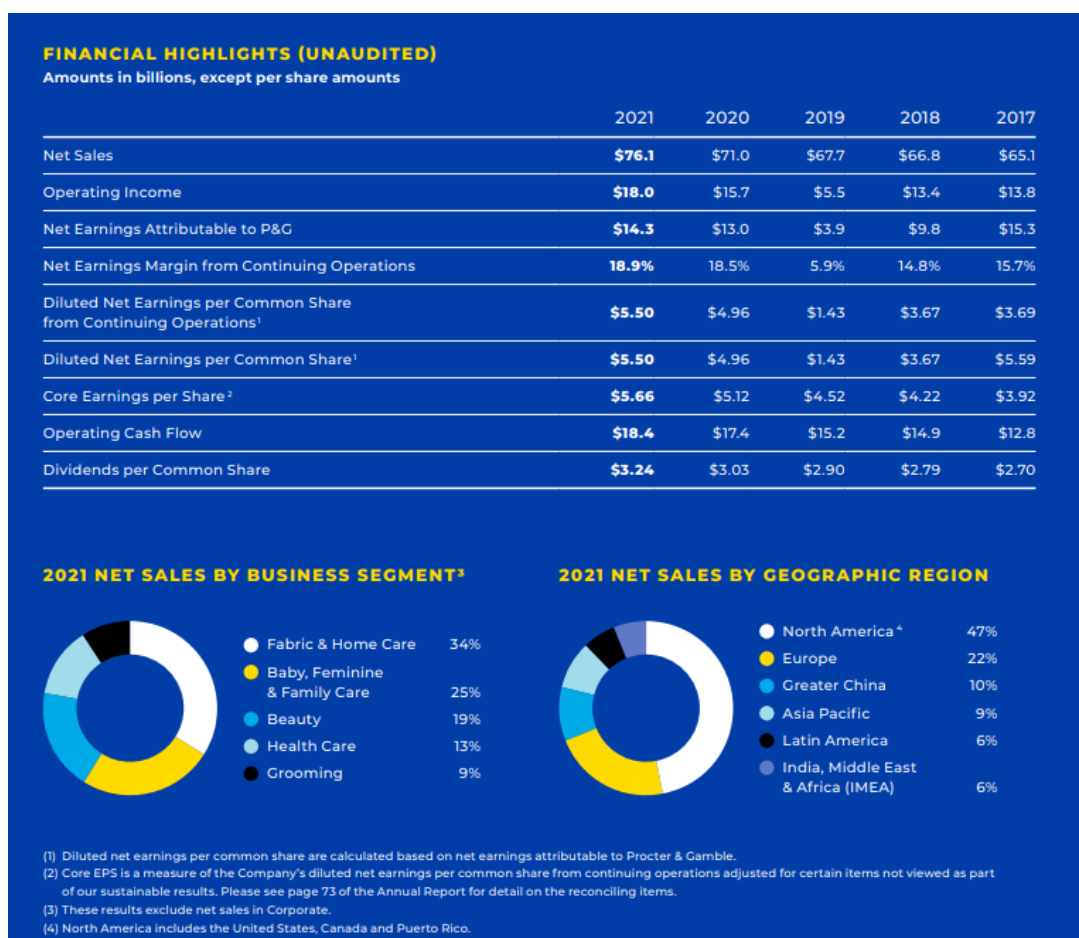


[Keyword search and LDA](#)

The “bag of words” technique is a Natural Language Processing (NLP) technique used for textual modeling. Text data can be messy and unstructured, posing challenges for machine learning algorithms in analysis. These algorithms prefer structured, well-defined, fixed-length inputs. A “bag of words” is a textual representation of the occurrence of words within a document. To create

this representation, analysts track the frequency of word occurrences in a document, disregarding grammatical details and word orders. The term “bag” is used because information about the order or structure of words in the document is discarded, and all words are collected *en masse* as if in a bag. Using this technique, variable-length texts can be converted into a fixed-length vector. The bag-of-words approach is a simple and flexible method to extract features from documents.

Figure 10. Financial highlights in P&G’s 2021 annual report



Building a Keyword Dictionary

A pre-established set of keywords is required to utilize the bag-of-words approach. Sentiment analysis, for instance, can be conducted by computing the frequency of pre-determined negative

and positive words. By comparing the number of negative words to provide words, the bag-of-words can identify the sentiment of a text as negative without the need to read the entire document.

Existing keyword lists

There is a range of well-established keyword lists available for textual analyses. In sentiment analysis, for example, the Harvard-IV-4 Dictionary is a general-purpose dictionary that provides a list of positive and negative words developed by Harvard University. The Loughran-McDonald Sentiment Word Lists are widely used in technical accounting and finance texts. Other researchers have developed similar keyword lists for non-English languages (Du, Huang, Wermers, and Wu 2022) or for purposes other than sentiment analysis, such as forward-looking statements, extreme sentiment, deception, financial constraints, uncertainty, financial performance, research and development. Technology, intangible assets, culture, big data and AI, litigation, social affiliation, supply chain, etc. (Cao, Ma, Tucker, and Wan 2018; Hassan, Hollander, Lent, and Tahoun 2019, etc.).

Self-defined keywords

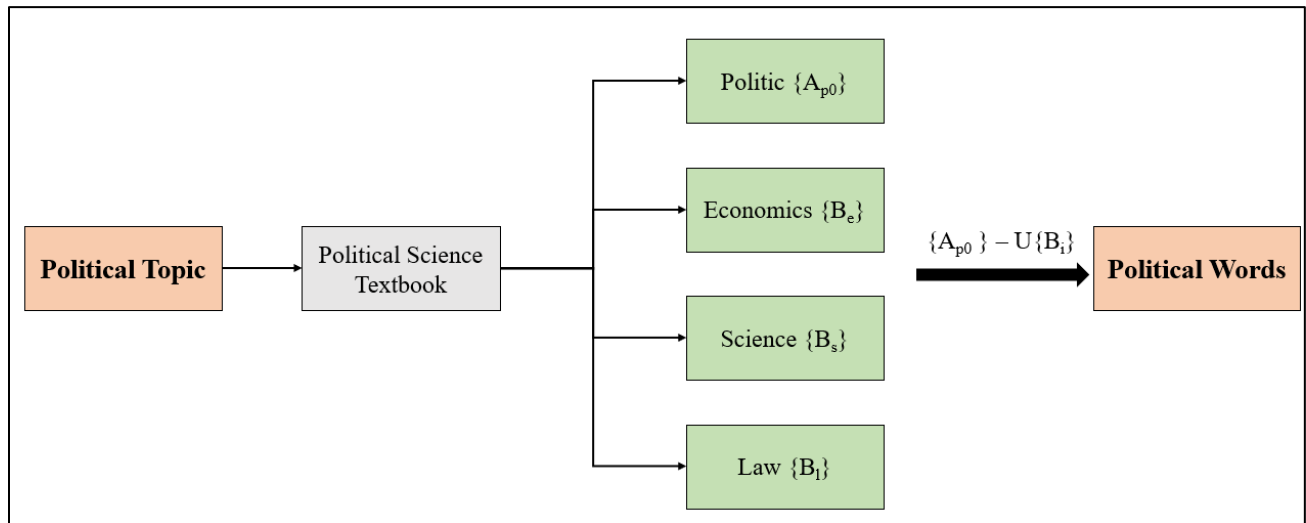
When a suitable keyword list is not available for a specific research question, we can create a customized one by reading a small sample of related texts and selecting the most relevant keywords. This approach is easy to implement, but it can also be arbitrary. Below, we discuss two structured approaches to developing self-defined keyword lists.

Corpus approach

The corpus approach begins with gathering textual contents relevant to the topic of interest, from which a set of frequently used words $\{A\}$ is extracted. This set often includes noisy keywords unrelated to the topic. To eliminate this noise, we then identify irrelevant topics and generate a list of frequently used words for each irrelevant topic $\{B_i\}$. A robust keyword list for the topic of

interest is then obtained by subtracting irrelevant topic keywords from the preliminary high-frequency word list, or $\{A\} - U\{B_i\}$. For example, to generate a list of political keywords $\{A_p\}$, one might start with political science textbooks to generate a high-frequency word list $\{A_{p0}\}$. This preliminary high-frequency word list might contain keywords relating to economics, law, science, etc. To remove these irrelevant topics, we could use a similar approach to generate lists of high-frequency words for each irrelevant topic $\{B_{\text{economics}}\}$, $\{B_{\text{law}}\}$, $\{B_{\text{science}}\}$, etc. Finally, we subtract these irrelevant keywords from the preliminary political keyword list, resulting in a clean political keyword list, or $\{A_p\} = \{A_{p0}\} - U\{B_i\}$. Figure 11 illustrates the process of using the corpus approach to develop a dictionary.

Figure 11. Using the corpus approach to develop keyword lists

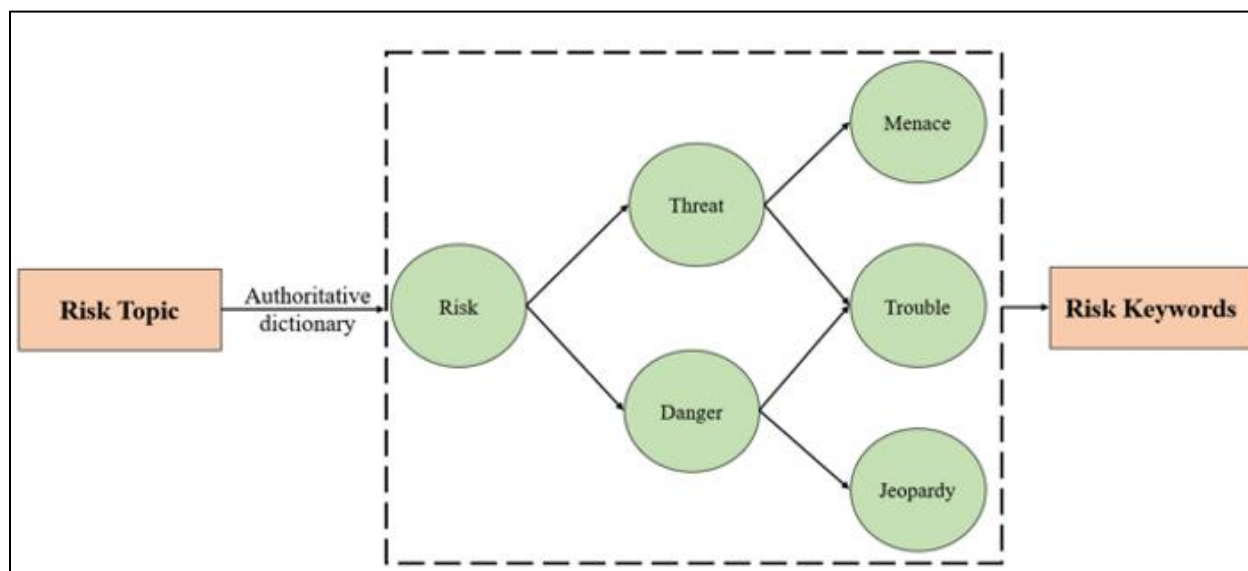


Dictionary expansion

The “dictionary expansion” approach generates an expanded keyword list by searching for synonyms of key topical words in authoritative dictionaries. For instance, to create a keyword list for “risk,” we can begin with the single word “risk” and look up all synonyms of “risk” in the Merriam-Webster Dictionary. This may yield words like “threat” and “danger.” Subsequently, we

can look up the synonyms of these synonyms, which could provide words such as “menace,” “jeopardy,” and “trouble.” The process can be continued until the additional synonyms are no longer closely related to the original concept of “risk.” (Figure 12).

Figure 12. Using the dictionary expansion approach to develop keyword lists

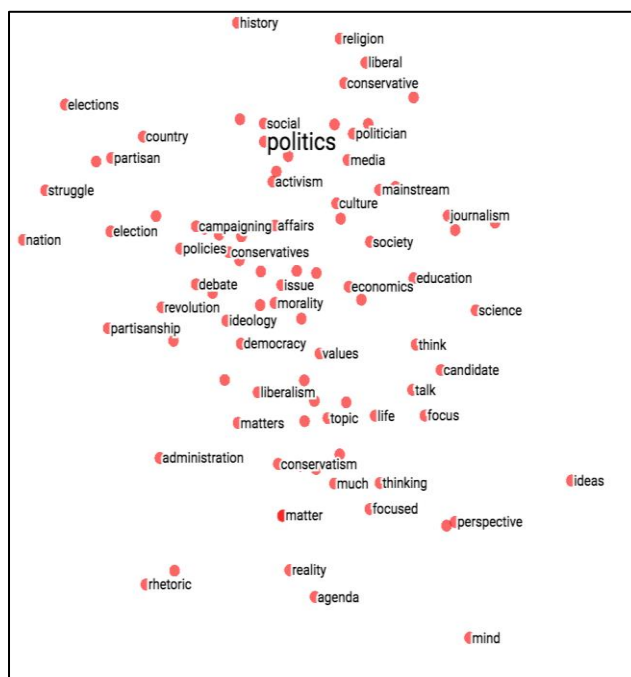
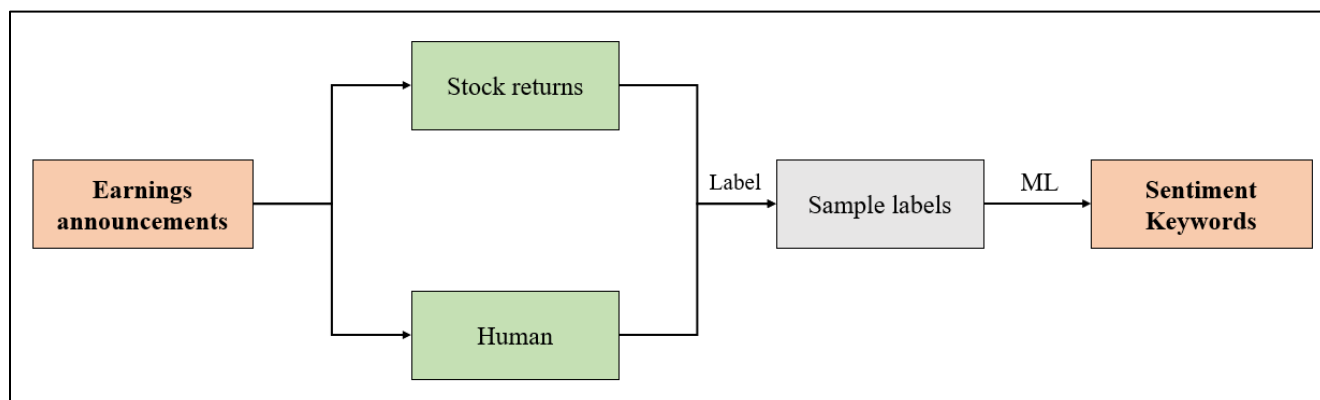


IBM Research-Almaden developed a “human-in-the-loop” approach for AI dictionary expansions (Alba, Gruhl, Ristoski, and Welch 2018). The approach not only discovers new instances from the input text corpus but also predicts new “unseen” terms not currently in the corpus. The approach runs in two phases. Continuing with the political word example, during the explore phase, the model calculates a similarity score between words in the Merriam-Webster Dictionary and the single word “politics” to identify instances in the dictionary that are similar to the word “politics,” such as “activism,” “legislature,” or “government.” In the exploit phase, the model generates new phrases based on a word’s co-occurrence score or how often words appear together. For example, “government policy” may not appear in the Merriam-Webster Dictionary, but “political policies” and “science of government” often appear together and can be used to build the more complex phrase “government policy.”

2.2.2. Latent Dirichlet Allocation (LDA)

Latent Dirichlet Allocation (LDA) is often used for dimensionality reduction. Unsupervised LDA proves useful in exploring unstructured text data by inferring relationships between words in a set of documents. A common application of unsupervised LDA is topic modeling. When given a sample of textual data and a pre-determined number of topics, K , an LDA algorithm can generate K topics that best fit the data. Determining the appropriate number of topics is somewhat arbitrary. The best practice is to review the textual sample to obtain a sense of the contents, generate a word frequency table to examine the high-frequency words, and then determine the number of topics in an informed manner. Figure 13 illustrates the keywords for the topic “politics” developed using an unsupervised LDA algorithm.

In addition to unsupervised LDA, LDA can also be supervised. Supervised LDA requires humans to read a small sample of the textual contents and label the topics for each textual input. The labeled sample is then used to train a model that predicts the topics of the remaining texts in the sample. The self-supervised approach involves using an existing label to label keywords. For instance, subsequent stock returns can be used to label positive and negative keywords when determining positive and negative keywords in earnings announcements (Figure 14).

Figure 13. Keywords for the topic “politics” generated using unsupervised LDA**Figure 14. Using supervised LDA to develop keyword lists**

2.3. Empirical examples: Analyzing corporate filings for making business decisions

 [Empirical example: Analyzing corporate filings](#)

10-Ks and 10-Qs

When companies modify 10-Ks, this often provides a signal about future operations (Brown and Tucker 2011; Cohen, Malloy, and Nguyen 2020). However, Cohen et al. (2020) document that investors tend to neglect the valuable information embedded in the changes. Constructing a

portfolio that shorts companies making significant changes to their 10-Ks or 10-Qs while buying those not making significant changes could yield returns of 30 to 50 basis points per month over the subsequent year.

Cao, Jiang, Yang, and Zhang (2023) is the first study exploring the feedback effect on corporate disclosure in response to technology. They document that firms with higher machine downloads of their SEC filings prepare 10-Ks and 10-Qs in a more machine-friendly manner that facilitates machine parsing, scripting, and synthesizing. Furthermore, firms anticipating high machine downloads avoid negative words in Loughran and McDonald (2011) after 2011, the year of publication of the LM dictionary.

Item 1 of 10-K

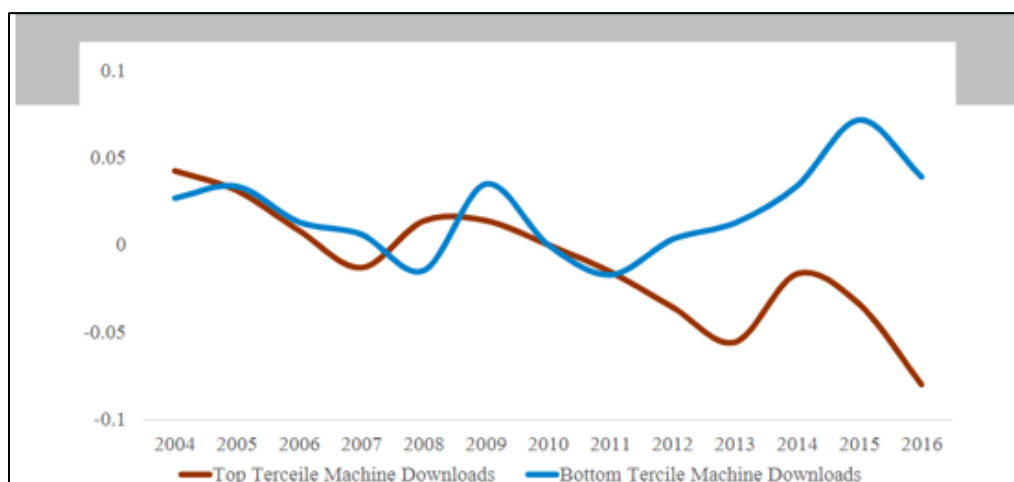
In 10-K Item 1, companies elaborate on their products and services. The textual descriptions in Item 1 can be leveraged to construct a stream of measures based on product similarity. Hoberg and Phillips (2016) extract product descriptions from 10-K Item 1 and represent word usage with a binary vector. The cosine similarity score between a pair of companies' product descriptions then captures the similarity of the products between the two companies. This product similarity measure is useful in evaluating the level of competition a company encounters. If a large number of companies provide highly similar products or services, the given company is likely to face intense competition in the product market. The measure can also refine industry classification, especially as many modern companies span multiple traditional industries. For example, Amazon Inc. operates as a retailer in the retailing industry, a streaming service provider in the entertainment industry, an electronic device maker in the manufacturing industry, and a software provider in the computer and business service industry. The product similarity measure provides an avenue to define an "industry" for Amazon that consists of companies providing a similar set of products

and services rather than arbitrarily assigning Amazon Inc. to a traditional industry. In a related vein, measuring the time-series similarity of Item 1 could assist analysts in detecting whether a company launches new products or services and implements new strategies.

Item 1A of 10-K

In Item 1A, companies delineate risk factors impacting their business. Henley and Hoberg (2019) developed an emerging risks database for banks based on risk factor disclosures in Item 1A. They employ topic modeling to obtain a 25-factor Latent Dirichlet Allocation (LDA) model, which is then used to extract 625 bigrams. Figure 16 provides an overview of the 25 emerging risk topics and the five most prevalent words in each topic. The bigrams are then converted into a set

Figure 15. Frequency of Loughran and McDonald (2011) negative words



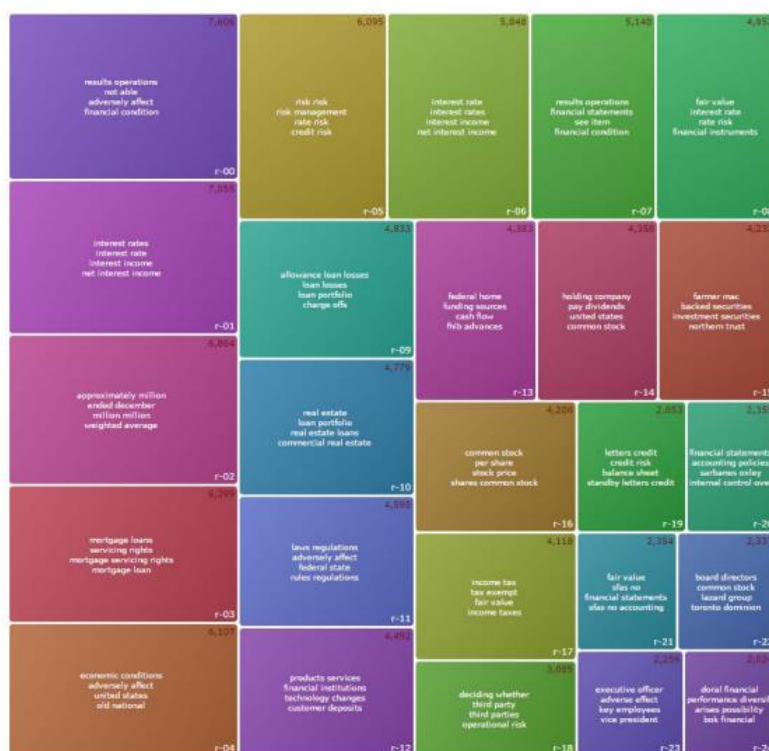
This figure plots LM – Harvard Sentiment of 10-K and 10-Q filings and compares the sentiment of the documents filed by firms with high machine downloads with that of the low group. LM – Harvard Sentiment is the difference between LM Sentiment and Harvard Sentiment. LM Sentiment is defined as the number of Loughran-McDonald (LM) finance-related negative words in a filing divided by the total number of words in the filing. Harvard Sentiment is defined as the number of Harvard General Inquirer negative words in a filing divided by the total number of words in the filing. Filings are sorted into top tercile or bottom tercile based on Machine Downloads. LM Sentiment and Harvard Sentiment sentiments are normalized to one, respectively, in 2010 within each group, one year before the publication of Loughran and McDonald (2011). The dotted lines represent the 95% confidence limits.

Source: Cao et al. (2023)

of interpretable risk factors in the form of word vectors using semantic vector analysis. The cosine similarity between the vocabulary list associated with each risk theme and the raw text of a bank's Item 1A disclosure reflects the intensity of the bank's discussion of each emerging risk. Using the risk loading, Henley and Hoberg (2019) show that risks related to real estate, prepayment, and commercial paper are elevated as early as mid-2005, prior to the 2008 financial crisis. They also find that individual bank exposure to emerging risk factors strongly predicts stock returns, bank failure, and return volatility.

Segment Information

Recruiting CEOs whose skills and attributes align with company needs is critical for company success. The inherent challenge in external CEO hiring arises from the heterogeneity of both job candidates and companies. Effective recruiting involves not only identifying competent managers but also optimizing the quality of the match between companies and CEOs. Cao, Li, and Ma (2022) find that segment information in 10-K filings helps companies find CEOs who fit their needs. For instance, when Ford hired Allan Mulally as the CEO from Boeing's Commercial Airplanes in 2006, they were seeking a leader with experience in turning around a troubled corporate giant. Allan Mulally happened to possess that experience, as revealed by the segment information disclosed in Boeing's 1999 10-K (Figure 17). The Commercial Airplanes segment of Boeing incurred a loss of 1,589 million in 1997 but achieved a profit of 2,016 million in the following two years. This experience perfectly matched what Ford valued in CEO candidates.

Figure 16. Emerging risks using LDA with 25 topics

This figure provides an overview of the 25 risk factors detected using topic modeling from 10-Ks of banks for fiscal year 2006. Each box is ranked and sized relative to its importance in the document and contains the five most prevalent words or common grams in the topic (Henley and Hoberg 2019).

Source: Henley and Hoberg (2019)

Figure 17. Segment information in Boeing's 1999 10-K

FORM 10-K For the fiscal year ended December 31, 1999						
Segment Information						
Note 24 to the consolidated financial statements						
	Net earnings (loss)			Sales and other operating revenues		
(Dollars in millions)	-----			-----		
Year ended December 31,	1999	1998	1997	1999	1998	1997
=====	=====	=====	=====	=====	=====	=====
Commercial Airplanes	\$2,016	\$ (266)	\$(1,589)	\$38,409	\$36,880	\$27,479

Appendix 2A: Project 1a How to crawl annual reports

Download ten 10-K filings and twenty 8-K filings of a company of your choice using the SEC API. Randomly read 10 files that you download and check whether the downloaded filings are correct and complete.

Appendix 2B: Project 1b How to parse unstructured data

Parse Item 1, Item 1A, and Item 7 of each of the ten 10-K filings you downloaded. Read the output and check whether the output is complete and accurate by comparing it with the original files.

Appendix 2C: Solution

How to crawl annual reports



How to crawl annual reports

```
import sys
from dataclasses import dataclass
from datetime import date, datetime
from datetime import datetime as dt
from pathlib import Path
from typing import Optional, Set, Union
import requests
from requests import Response
from collections import deque
from typing import Dict, Any, List

FULL_SUBMISSION_FILENAME = "full-submission.txt"
PRIMARY_DOC_FILENAME_STEM = "primary-document"

DATE_FORMAT_TOKENS = "%Y-%m-%d"
DEFAULT_BEFORE_DATE = date.today()
DEFAULT_AFTER_DATE = date(1994, 1, 1)

STANDARD_HEADERS = {
    "Accept-Encoding": "gzip, deflate",
}

URL_SUBMISSIONS = "https://data.sec.gov/submissions/{submission}"
URL_FILING =
("https://www.sec.gov/Archives/edgar/data/{cik}/{acc_num_no_dash}/{document}")

@dataclass
class DownloadMetadata:

    download_folder: Path
    form: str
    cik: str
    limit: int = sys.maxsize
    after: date = DEFAULT_AFTER_DATE
    before: date = DEFAULT_BEFORE_DATE
    include_amends: bool = False
    download_details: bool = False
    ticker: Optional[str] = None
    accession_numbers_to_skip: Optional[Set[str]] = None

@dataclass
class ToDownload:
    raw_filing_uri: str
    primary_doc_uri: str
    accession_number: str
    details_doc_suffix: str

DownloadPath = Union[str, Path]

Date = Union[str, date, datetime]

def secjson(uri: str, user_agent: str, host: str) -> Response:
    resp = requests.get(
        uri,
```

```

        headers={
            "Accept-Encoding": "gzip, deflate",
            "User-Agent": user_agent,
            "Host": host,
        },
    )
    resp.raise_for_status()
    return resp

def get_list_of_filings(uri: str, user_agent: str) -> Any:
    return secjson(uri, user_agent, "data.sec.gov").json()

def get_to_download(cik: str, acc_num: str, doc: str) -> ToDownload:
    cik = cik.lstrip("0")
    acc_num_clean = acc_num.replace("-", "")
    raw_filing_uri = URL_FILING.format(
        cik=cik, acc_num_no_dash=acc_num_clean, document=f"{acc_num}.txt"
    )
    primary_doc_uri = URL_FILING.format(
        cik=cik, acc_num_no_dash=acc_num_clean, document=doc.rsplit("/")[-1]
    )
    primary_doc_suffix = Path(doc).suffix.replace("htm", "html")

    return ToDownload(
        raw_filing_uri,
        primary_doc_uri,
        acc_num,
        primary_doc_suffix,
    )

def get_location(
    download_metadata: DownloadMetadata,
    accession_number: str,
    save_filename: str,
) -> Path:
    company_identifier = (
        download_metadata.ticker
        if download_metadata.ticker is not None
        else download_metadata.cik
    )
    return (
        download_metadata.download_folder
        / "sec-edgar-filings"
        / company_identifier
        / download_metadata.form
        / accession_number
        / save_filename
    )

def download_filing(uri: str, user_agent: str) -> bytes:
    return secjson(uri, user_agent, "www.sec.gov").content

def save_document(filing_contents: Any, save_path: Path) -> None:
    save_path.parent.mkdir(parents=True, exist_ok=True)
    save_path.write_bytes(filing_contents)

def validate_and_parse_date(input_date: Date) -> date:
    if isinstance(input_date, datetime):
        return input_date.date()
    elif isinstance(input_date, date):
        return input_date
    elif isinstance(input_date, str):
        try:

```

```

        return dt.strptime(input_date, DATE_FORMAT_TOKENS).date()
    except ValueError as exc:
        raise ValueError(
            "Incorrect date format. Please enter a date string of the form YYYY-
MM-DD."
        ) from exc
    else:
        raise TypeError(
            "Incorrect date input. Must be of type string, date, or datetime."
        )

def is_cik(ticker_or_cik: str) -> bool:
    try:
        int(ticker_or_cik)
        return True
    except ValueError:
        return False

def validate_and_convert_ticker_or_cik(
    ticker_or_cik: str, ticker_to_cik_mapping: Dict[str, str]
) -> str:
    ticker_or_cik = str(ticker_or_cik).strip().upper()

    if not ticker_or_cik:
        raise ValueError("Invalid ticker or CIK. Please enter a non-blank value.")

    if is_cik(ticker_or_cik):
        if len(ticker_or_cik) > 10:
            raise ValueError("Invalid CIK. CIKs must be at most 10 digits long.")
        return ticker_or_cik.zfill(10)

    cik = ticker_to_cik_mapping.get(ticker_or_cik)

    if cik is None:
        raise ValueError(
            f"Ticker {repr(ticker_or_cik)} is invalid and cannot be mapped to a CIK. "
            "Please enter a valid ticker or CIK."
        )

    return cik

def get_ticker_metadata(user_agent: str) -> Any:
    return secjson("https://www.sec.gov/files/company_tickers_exchange.json",
user_agent, "www.sec.gov").json()

def get_ticker_to_cik_mapping(user_agent: str) -> Dict[str, str]:
    ticker_metadata = get_ticker_metadata(user_agent)
    fields = ticker_metadata["fields"]
    ticker_data = ticker_metadata["data"]

    cik_idx = fields.index("cik")
    ticker_idx = fields.index("ticker")

    return {
        str(td[ticker_idx]).upper(): str(td[cik_idx]).zfill(10)
        for td in ticker_data
    }

def aggregate_filings_to_download(
    download_metadata: DownloadMetadata, user_agent: str

```

```

) -> List[ToDownload]:
  filings_to_download: List[ToDownload] = []
  fetched_count = 0
  submissions_uri = URL_SUBMISSIONS.format(
    submission="CIK{cik}.json".format(cik=download_metadata.cik)
  )
  additional_submissions = None

  while fetched_count < download_metadata.limit:
    resp_json = get_list_of_filings(submissions_uri, user_agent)
    if additional_submissions is None:
      filings_json = resp_json["filings"]["recent"]
      additional_submissions = deque(resp_json["filings"]["files"])
    else:
      filings_json = resp_json

    accession_numbers = filings_json["accessionNumber"]
    forms = filings_json["form"]
    documents = filings_json["primaryDocument"]
    filing_dates = filings_json["filingDate"]

    for acc_num, form, doc, f_date in zip( # noqa: B905
      accession_numbers, forms, documents, filing_dates
    ):
      is_amend = form.endswith("/A")
      form = form[:-2] if is_amend else form
      if (
        form != download_metadata.form
        or (not download_metadata.include_amends and is_amend)
      ):
        continue

      filings_to_download.append(
        get_to_download(download_metadata.cik, acc_num, doc)
      )

      fetched_count += 1
      if fetched_count == download_metadata.limit:
        break

    if len(additional_submissions) == 0:
      break

    next_page = additional_submissions.popleft()["name"]
    submissions_uri = URL_SUBMISSIONS.format(submission=next_page)

  return filings_to_download

def fetch_and_save_filings(download_metadata: DownloadMetadata, user_agent: str) ->
int:
  successfully_downloaded = 0
  to_download = aggregate_filings_to_download(download_metadata, user_agent)

  for td in to_download:
    try:
      location = get_location(
        download_metadata, td.accession_number, FULL_SUBMISSION_FILENAME
      )
      if not location.exists():
        raw_filing = download_filing(td.raw_filing_uri, user_agent)
        save_document(raw_filing, location)

```

```

        if download_metadata.download_details:
            primary_doc_filename = (
                f"{PRIMARY_DOC_FILENAME_STEM}{td.details_doc_suffix}"
            )
            location = get_location(
                download_metadata, td.accession_number, primary_doc_filename
            )
            if not location.exists():
                primary_doc = download_filing(td.primary_doc_uri, user_agent)
                save_document(primary_doc, location)
    except Exception as e:
        print(
            td.accession_number,
            e,
        )
        continue

    successfully_downloaded += 1

return successfully_downloaded

class Downloader:

    def __init__(
        self,
        company_name: str,
        email_address: str,
        download_folder: Optional[DownloadPath] = None,
    ) -> None:
        self.user_agent = f"{company_name} {email_address}"

        if download_folder is None:
            self.download_folder = Path.cwd()
        elif isinstance(download_folder, Path):
            self.download_folder = download_folder
        else:
            self.download_folder = Path(download_folder).expanduser().resolve()

        self.ticker_to_cik_mapping = get_ticker_to_cik_mapping(self.user_agent)

    def get(
        self,
        form: str,
        ticker_or_cik: str,
        *,
        limit: Optional[int] = None,
        after: Optional[Date] = None,
        before: Optional[Date] = None,
        include_amends: bool = False,
        download_details: bool = False,
        accession_numbers_to_skip: Optional[Set[str]] = None,
    ) -> int:

        cik = validate_and_convert_ticker_or_cik(
            ticker_or_cik, self.ticker_to_cik_mapping
        )

        if limit is None:
            limit = sys.maxsize
        else:
            limit = int(limit)
            if limit < 1:

```

```

        raise ValueError(
            "Invalid amount. Please enter a number greater than 1."
        )

    if after is None:
        after_date = DEFAULT_AFTER_DATE
    else:
        after_date = validate_and_parse_date(after)

        if after_date < DEFAULT_AFTER_DATE:
            after_date = DEFAULT_AFTER_DATE

    if before is None:
        before_date = DEFAULT_BEFORE_DATE
    else:
        before_date = validate_and_parse_date(before)

    if after_date > before_date:
        raise ValueError("After date cannot be greater than the before date.")

    num_downloaded = fetch_and_save_filings(
        DownloadMetadata(
            self.download_folder,
            form,
            cik,
            limit,
            after_date,
            before_date,
            include_amends,
            download_details,
            ticker=ticker_or_cik if not is_cik(ticker_or_cik) else None,
            accession_numbers_to_skip=accession_numbers_to_skip,
        ),
        self.user_agent,
    )

    return num_downloaded

dl = Downloader("CCC Inc.", "###@Email.com", " C:/Users/")
dl.get(
    form="10-K",
    ticker_or_cik="AAPL",
    limit=100,
    after="2021-01-01",
    before="2021-12-31",
    include_amends=False,
    download_details=True,
)

```

How to parse unstructured data



[How to parse unstructured data](#)

```

import re

import glob
from bs4 import BeautifulSoup
import csv
from pathlib import Path

```

```

files=glob.glob("C:/Users/**/*.*.html", recursive=True)

def parser(text,section):

    def extract_text(text, item_start, item_end):
        item_start = item_start
        item_end = item_end
        starts = [i.start() for i in item_start.finditer(text)]
        ends = [i.start() for i in item_end.finditer(text)]
        positions = list()
        for s in starts:
            control = 0
            for e in ends:
                if control == 0:
                    if s < e:
                        control = 1
                        positions.append([s,e])
        item_length = 0
        item_position = list()
        for p in positions:
            if (p[1]-p[0]) > item_length:
                item_length = p[1]-p[0]
                item_position = p

        item_text = text[item_position[0]:item_position[1]]

        return(item_text)

    if section == 1 or section == 0:
        try:
            item1_start = re.compile("item\s*[1][\.;\:\-\_]*\s*\\b", re.IGNORECASE)
            item1_end = re.compile("item\s*1a[\.;\:\-\_]\s*Risk|item\s*2[\.;\:\-\_]\s*Risk|item\s*2[\.;\:\-\_]\s*Prop", re.IGNORECASE)
            businessText = extract_text(text, item1_start, item1_end)
        except:
            businessText = "Something went wrong!"

    if section == 2 or section == 0:
        try:
            item1a_start = re.compile("(?!,\s)item\s*1a[\.;\:\-\_]\s*Risk",
re.IGNORECASE)
            item1a_end = re.compile("item\s*2[\.;\:\-\_]\s*Prop|item\s*[1][\.;\:\-\_]\s*\\b", re.IGNORECASE)
            riskText = extract_text(text, item1a_start, item1a_end)
        except:
            riskText = "Something went wrong!"

    if section == 3 or section == 0:
        try:
            item7_start = re.compile("item\s*[7][\.;\:\-\_]*\s*\\bM", re.IGNORECASE)
            item7_end = re.compile("item\s*7a[\.;\:\-\_]\s*Quanti|item\s*8[\.;\:\-\_]\s*", re.IGNORECASE)
            mdaText = extract_text(text, item7_start, item7_end)
        except:
            mdaText = "Something went wrong!"

    if section == 0:
        data = [businessText, riskText, mdaText]
    elif section == 1:
        data = [businessText]
    elif section == 2:

```

```
        data = [riskText]
    elif section == 3:
        data = [mdaText]
    return(data)

i=1
while i<###:
    content=open(files[i], encoding='utf-8').read()
    soup = BeautifulSoup(content, features="html.parser")
    for script in soup(["script", "style"]):
        script.extract()
    text = soup.get_text()
    output=parser(text,1)
    parent_dir="C:/Users"
    savepath=Path(parent_dir,str(i)+".txt")
    outfile = open(savepath, 'w+', newline = ' ',encoding="utf-8")
    outfile.write(output[0])
    i+=1
```


References

- Alba, A., Gruhhl, D., Ristoski, P., and Welch, S. 2018. Interactive dictionary expansion using neural language models. Second International Workshop on Augmenting Intelligence with Humans in the Loop.
- Brown, S., and Tucker, J. 2011. Large-sample evidence on firms' year-over-year MD&A modifications. *Journal of Accounting Research*, 49(2), 309-346.
- Cao, S., Li, Y., and Ma, G. 2022. Labor market benefit of disaggregated disclosure. *Contemporary Accounting Research*, 39(3), 1726-1757.
- Cao, S., Ma, G., Tucker, J., and Wan, C. 2018. Technological peer pressure and product disclosure. *The Accounting Review*, 93(6), 95-126.
- Cao, S., Jiang, W., Yang, B., Zhang, A. 2023. How to talk when a machine is listening? Corporate disclosure in the age of AI. *Review of Financial Studies*, 36(9), 3603-3642.
- Cohen, L., Malloy, C., and Nguyen, Q. 2020. Lazy prices. *Journal of Finance*, 3, 1371-1415.
- Du, Z., Huang, A., Wermers, R., and Wu, W. 2022. Language and domain specificity: A Chinese financial statement dictionary. *Review of Finance*, 26(3), 673-719.
- Hassan, T., Hollander, S., Lent, L., and Tahoun, A. 2019. Firm-level political risk: Measurement and effects. *Quarterly Journal of Economics*, 134(4), 2135-2202.
- Harris, Z. 1954. Distributional structure. *Word*, 10(23), 146-162.
- Hoberg, G., and Hanley, K.W. 2019. Dynamic interpretation of emerging risks in the financial sector. *Review of Financial Studies*, 32(12), 4543-4603.
- Hoberg, G., and Phillips, G. 2016. Text-based network industries and endogenous product differentiation. *Journal of Political Economy*, 124(5), 1423-1465.
- Loughran, T., and McDonald, B. 2011. When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *Journal of Finance*, 66(10), 35-65.

Chapter 3 Emerging AI Technology in Textual Analysis

3.1. Procedures for applying machine learning models



[Procedures for applying machine learning models](#)

3.1.1. Data cleaning, parsing, and feature selection

The initial step in building machine learning models is to preprocess raw data, or data cleaning, which is essential for improving data quality. This process involves eliminating redundant entries and boilerplate text, handling missing data and outliers, and correcting improperly formatted data. The step ensures that the model learns from consistent and relevant data, which will improve the final model's performance. For instance, it is common practice to exclude sections such as the analyst disclaimer when working with analyst reports, as they typically offer minimal or no value for machine learning tasks. Eliminating this type of "boilerplate" text allows the machine to better discern patterns by focusing its attention on the report's essential content.

The next key aspect of the model-building process is data parsing, entailing data conversion from one format to another. The objective is to transform unstructured raw data into a cohesive, structured representation that machines can easily comprehend and utilize. Consider, for instance, an HTML webpage. Parsing the HTML allows us to transform it into organized formats such as CSV or JSON, simplifying the extraction of specific details from the data. Regular expressions are commonly employed to extract specific patterns or sequences of characters from the data to further enhance the data organization and usability.

Building machine learning models also involves feature selection. Feature selection is the process of identifying input variables essential for building a high-performing model. The inclusion or exclusion of relevant features could affect the quality of the model's output. As the saying goes, "garbage in, garbage out," which underscores that the output's reliability is inherently

tied to the quality of the input. If a model is trained on a dataset that contains numerous irrelevant features, it is more likely to produce unreliable or erroneous results. Domain knowledge is thus crucial to successful feature selection. Experts with a deep understanding of the subject matter can leverage their knowledge and experience to identify key features. For instance, in their “AI Analyst” model, Cao, Jiang, Wang, and Yang (2021) specifically choose firm-level, industry-level, macroeconomic variables, and textual information from firms’ disclosures as inputs. Their selection is informed by prior studies demonstrating a robust correlation between these variables and earnings forecasts. This informed approach underscores the importance of leveraging domain expertise when choosing relevant features for machine learning models.

3.1.2. Machine learning model selection

Once the data has undergone preprocessing, the next step is building the machine learning models. This process is not a random selection; instead, it follows a systematic approach to identify the most suitable model for the given data and problem at hand.

Model selection relies heavily on understanding both the dataset’s characteristics and different models’ relative strengths and limitations. Since different models excel in different tasks, we select an initial model based on our knowledge and experience regarding each model’s capabilities, strengths, and weaknesses. For instance, random forest models are often employed for classification tasks due to their capacity to handle both numerical and categorical data and their resistance to overfitting. Long short-term memory (LSTM) models are particularly effective for time series analyses as they can capture long-term dependencies in the data. On the other hand, transfer models are commonly used in tasks where previously acquired knowledge can be utilized, such as image or natural language processing. For example, Cao, Jiang, Wang, and Yang (2021) start with two quasi-linear ML models, Elastic-Net and Support Vector Regressions that are

particularly useful for tasks with a large number of variables. They then incorporate three highly nonlinear machine learning models--Random Forest, Gradient Boosting, and Long Short-Term Memory (LSTM) Neural Networks. Random Forest and Gradient Boosting excel in capturing complex and hierarchical interactions among the input variables, while the LSTM model is designed for modeling time-series patterns in the data. This approach aligns the machine learning models with the specific characteristics of the data and expert knowledge of each model's advantages.

Once we have selected our initial set of models, we execute each model and evaluate its performance. This process aims to quantify the effectiveness of each model and validate the initial choices. Notably, ensemble models are able to amalgamate knowledge from multiple models, often resulting in improved outcomes as compared to a single model alone. Therefore, leveraging an ensemble model allows us to integrate predictions from the top-performing models. Cao, Jiang, Wang, and Yang (2021) use an ensemble consisting of the three best-performing models as their primary model. By doing so, they can harness these models' collective strengths and insights to enhance their analysis's overall predictive power and reliability. Like feature selection, model selection in machine learning also hinges on domain knowledge, especially when deciding the appropriate level of analysis, whether at the individual, industry, or market level. In this regard, researchers must draw upon their understanding of the research question at hand and the specific domain at large. Through leveraging their expertise, they can make informed decisions regarding the scope and granularity of the analysis, ensuring that the selected model aligns with the objectives and requirements of the study.

3.1.3. Hyperparameter tuning

Grid search

Grid search is a technique to identify the optimal combination of hyperparameters for a machine learning model. Its name is derived from the grid-like structure it creates, with each dimension representing a different hyperparameter. The process systematically evaluates all possible combinations by iterating through the grid.

To implement the grid search, the initial step is to identify the hyperparameters requiring tuning and define the range of values to explore for each hyperparameter. As an example, one might specify a list of learning rates like [0.001, 0.005, 0.01] and a list of batch sizes like [30, 50, 60]. Creating a grid encompassing all possible hyperparameter combinations, the model is trained and evaluated for each combination within a loop. The best model is ultimately determined by identifying the hyperparameter combination that yields the highest performance.

Cross-validation

Cross-validation is a widely used technique in machine learning for evaluating model performance. Generally, the available data is partitioned into three main subsets: the training, validation, and test sets. The models are trained on the training set, fine-tuned using the validation set, and evaluated for accuracy on the test set. This allows us to gauge how effectively the trained model generalizes to unseen data, thereby mitigating the risk of overfitting, wherein a model excels on the training data but fails to generalize to new and unseen data.

Figure 1 illustrates the process of k-fold cross-validation, the most common form of cross-validation. In k-fold cross-validation, the data is divided into k equal-sized folds. The model is trained k times, with each iteration using k-1 folds as the training set and a different fold as the validation set. Following the k training iterations, k individual evaluation scores are obtained, and

the average of these scores represents the model's overall performance. Employing k-fold cross-validation is an efficient method of maximizing the utilization of available data, particularly when the dataset size is limited.

Imagine a dataset comprising 10 million photos of pets, including some featuring dogs. The objective is to identify the dog photos. With a set of 10,000 labeled photos categorized as dog or non-dog photos, applying a five-fold cross-validation involves dividing the 10,000 labeled photos into five folds, with 2,000 images in each subset. The classification model is then constructed through five iterations, with four folds as the training set and the fifth as the validation set during each iteration. Following these five training cycles, five individual evaluation scores are obtained, collectively reflecting the model's overall performance.

Figure 1. Five-fold cross-validation



3.1.4. Model evaluation

There are several metrics for evaluating the performance of classification machine learning models, such as accuracy, precision, recall, and F1 score. Let us start by exploring the confusion matrix and related terms to enhance comprehension of these metrics.

The confusion matrix provides a comprehensive summary of a model's predictions and their alignment with the actual labels of a test dataset. It yields information on four categories of classification results: true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). A true positive (TP) refers to a scenario where the model correctly predicts a positive condition, while a true negative (TN) indicates a scenario where the model correctly predicts a negative condition. On the contrary, a false positive (FP) arises when the model incorrectly predicts a positive outcome, that is, in reality, negative, and a false negative (FN) occurs when the model incorrectly predicts a negative outcome, but the actual outcome is positive. Figure 2 depicts a 2x2 confusion matrix that represents the four possible outcomes:

Figure 2. A confusion matrix

		Actual Values	
		Positive	Negative
Predicted Values	Positive	TURE POSITIVE TP	FALSE POSITIVE FP
	Negative	FALSE NEGATIVE FN	TRUE NEGATIVE TN

Using the components outlined in the confusion matrix above, we can calculate four evaluation metrics: accuracy, precision, recall, and the F1 score.

Accuracy

Accuracy measures the overall correctness of the model's predictions. It is defined as the ratio of correct predictions (i.e., true positives and true negatives) to the total number of predictions. Higher accuracy indicates that the model's proficiency in making correct predictions.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision

Precision reflects the ratio of true positive predictions to all positive predictions made by the model. It is calculated by dividing the number of true positives by the total number of positive predictions. Higher precision values indicate a lower probability of the model falsely labeling negative instances as positive.

$$Precision = \frac{TP}{TP + FP}$$

Recall

Recall evaluates the model's ability to correctly identify true positive cases. It is defined as the number of true positives divided by the sum of true positives and false negatives.

$$Recall = \frac{TP}{TP + FN}$$

F1 Score

The F1 score combines precision and recall into a unified metric, providing insight into a model's ability to handle false positives and false negatives. It is calculated as follows:

$$F1\ Score = \frac{2 * (Precision * Recall)}{Precision + Recall}$$

When working with imbalanced data, a common challenge, the “Accuracy Paradox,” often emerges. This issue occurs when relying solely on accuracy as a metric, potentially resulting in misleading conclusions. In such scenarios, it becomes crucial to factor in precision as an important metric.

Consider an example involving an imbalanced dataset for email spam detection, wherein 98% of the emails in the dataset are not spam (negative), while only 2% are identified as spam (positive). Now, let's assume we build a classification model specifically designed to detect spam. The model yields the following results:

- True Positives (TP): 150
- False Positives (FP): 50
- True Negatives (TN): 9800
- False Negatives (FN): 50

If we calculate the accuracy of the model, we obtain:

$$\text{Accuracy} = (\text{TP} + \text{TN}) / (\text{TP} + \text{TN} + \text{FP} + \text{FN}) = (150 + 9800) / (150 + 9800 + 50 + 50) = 99.3\%.$$

At first glance, this model would appear to be impressively accurate. However, it's important to consider the context of the imbalanced dataset at hand. In this dataset, the majority of emails (approximately 98%) are non-spam. Hence, even if we were to classify all emails as non-spam, the substantial number of true negatives would still yield a high rate of accuracy. Therefore, when evaluating the model's performance, it becomes essential to incorporate additional metrics that provide a more comprehensive assessment of its effectiveness, especially in the presence of imbalanced class distributions.

To gain further insights, let's calculate the precision of the model. The precision will be:

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FP}) = 150 / (150 + 50) = 75\%.$$

While the accuracy is 99.3%, the precision is only 75%. This means that, among all the emails predicted as spam, only 75% are truly spam. This highlights the importance of computing precision for an imbalanced dataset, as the presence of false positives can result in substantial costs or consequences. Precision thus serves as a valuable evaluation metric, particularly in scenarios where the dataset exhibits class imbalance.

3.2. Fundamental concepts of pre-training in machine learning



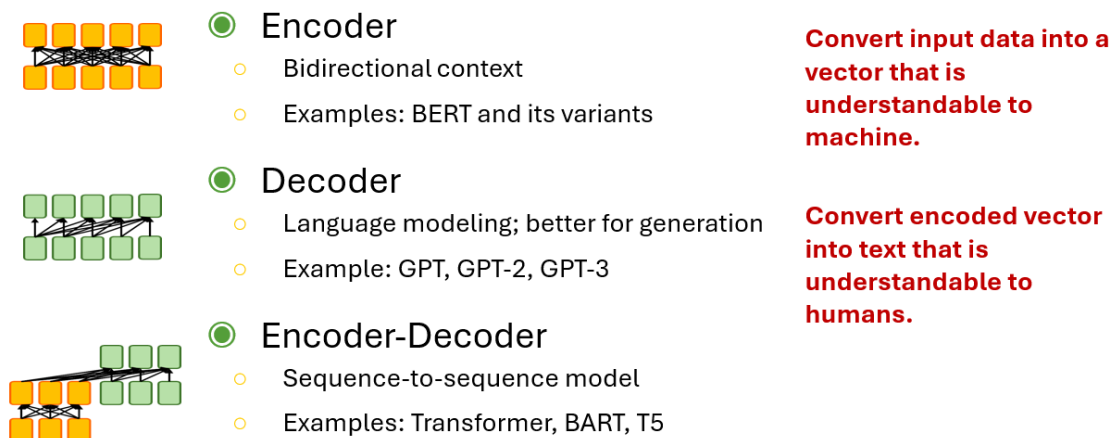
[Basic concept and foundation of machine learning](#)



[Application of supervised learning, ensemble learning, and model selection in Fintech](#)

There are three fundamental concepts in the pre-training process for machine learning algorithms: encoders, decoders, and transformers. An encoder is responsible for processing and compressing input data into a fixed-size representation that captures the essential features and context of the input. A decoder takes this compressed representation and generates the output, often translating it into a sequence of words or predictions. Finally, transformers integrate both encoders and decoders into a unified framework.

Figure 3. Three types of pre-training



3.3. Pre-trained phrase-level word embedding

Textual representation

The methods for textual analysis discussed thus far have primarily relied on word frequency. Frequency-based techniques ignore syntax and contextual meaning, potentially leading to potentially inaccurate analysis. For instance, a frequency-based algorithm might treat “many persons” and “people” as unrelated textual inputs because it lacks the ability to recognize their shared meaning. Word-embedding was developed to address this frequency-based technique limitation to incorporate meaning into textual analysis. This method represents a word with a semantic vector, essentially creating a new bag of words related to the word of interest. Word embeddings are grounded in Zellig Harris’s “distributional hypothesis,” which posits that words used in proximity to each other typically share similar meanings (Harris, 1954). To construct a semantic vector for a given phrase, such as “cash flow,” a word-embedding algorithm selects words from surrounding text inputs that can accurately predict the presence of the phrase. For example, in the text input “earnings present cash flow, which helps future investment” (Figure 3), “investment,” “earnings,” “present,” “which,” “help,” and “future” are all adjacent words of “cash flow.” The algorithm would select “investment” and “earnings” for the semantic vector representing “cash flow,” but not “which” or “help” as these words could be adjacent to numerous phrases other than “cash flow.” In other words, when “cash flow” is concealed, “earnings” and “investment” can relatively accurately predict the presence of “cash flow,” whereas “which” or “help” cannot. Hence, the word vector [Earnings, investment] represents “cash flow.”

Figure 4. Textual representation

Earnings present cash flow, which helps future investment.
Earnings present (marked), which helps future investment.

Advantages of phrase-level word-embedding

Using textual representation, algorithms can be trained to understand semantic relationships between words much the same way humans do. As previously mentioned, a frequency-based algorithm might not be able to recognize that “person” and “people” have similar meanings. However, word-based embedding creates comparable semantic vectors for “person” and “people.” For example, “person” could be represented by a word vector [human, man, woman, men, women, they], and “people” might be represented by a word vector [human, men, women, they]. This form of word embedding enables an algorithm to recognize their shared meaning through semantic computation.

Even now, researchers are working out ways to combine the computational efficiency and interpretability of frequency-based algorithms with the scalability and complex semantic and syntactic handling of word embeddings. For example, Cong, Liang, Zhang, and Zhu (2024) propose a textual-factor framework that offers a scalable, interpretable, and data-driven approach to analyze unstructured data. Their method consists of two stages. In Stage I, neural networks map each word in an unstructured text into a vector that retains the word’s semantic and syntactic characteristics. These vectors are then clustered to develop a topic model, which helps extract topic factors. In Stage II, a loading is created for each textual factor. The loading reflects how closely the unstructured text is associated with a particular textual factor. Such an approach, which aims to balance model complexity with interpretability, represents one of many promising directions for future data analytics applications (Cong, Liang, Yang, and Zhang 2021).

3.4. Pre-trained sentence level-word embedding



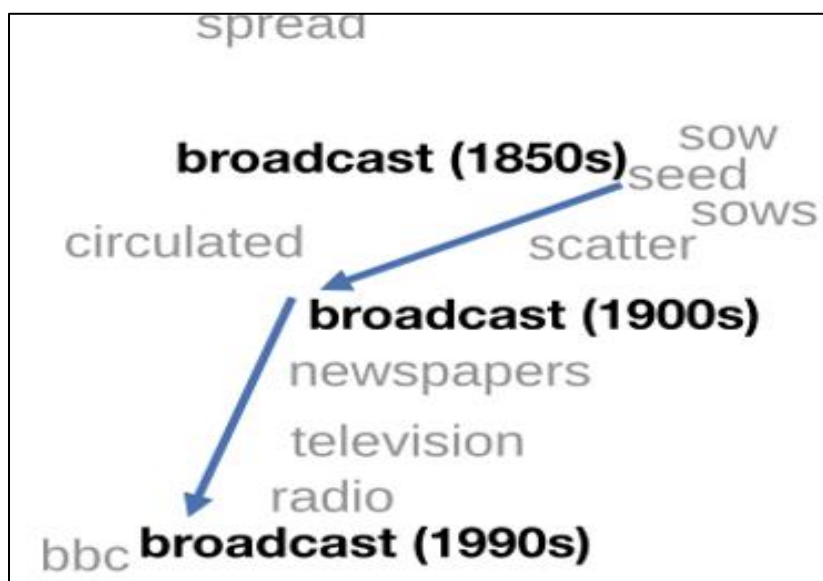
[Textual analysis: Word representation and sentence level analysis using Google Bert](#)

While phrase-level word embedding considers the meaning of words, it ignores the wealth of additional information present in sentences, potentially introducing challenges in textual analysis. To begin with, single words often carry multiple meanings that differ according to context. As an example, “liability” means a responsibility or burden in general language but represents a neutral line item representing resources contributed by creditors in financial statements. Word-embedding would struggle to differentiate between these different meanings. Consequently, the word “liability” should be represented by different semantic vectors in these two contexts in order to capture context-dependent meanings. Second, the meanings of words frequently change over time. For example, in the 1850s, “broadcast” was used to describe the way a farmer disperses seeds, whereas today it is almost always used to describe media. Therefore, a semantic vector representing “broadcast” from the 1850s would likely include words such as “sow” and “seed,” while a contemporary semantic vector for “broadcast” would feature terms like “television,” “radio,” and “newspapers” (Figure 4).

Sentence embedding addresses these problems by mapping sentences to vectors. This can be achieved by leveraging the hidden layer outputs of transformer models or by aggregating word embeddings into sentence embeddings.

3.4.1. Bidirectional Encoder Representations from Transformers

One of the most advanced tools for natural language processing (NLP) is Bidirectional Encoder Representations from Transformers (BERT), developed by Google. BERT originates from pre-training contextual representations. BERT was trained on two tasks: language modeling and next sentence prediction, using the Toronto BookCorpus and English Wikipedia.

Figure 5. The changing meaning of “broadcast” over time

In language modeling, 15% of words were selected for prediction, with the training objective being to predict the selected word given its context. The selected word is masked with a probability of 80%, replaced with a random word with a probability of 10%, and not replaced with a probability of 10%. For example, in the sentence “he is nice,” with three words, the third word, “nice,” is selected for prediction. The input text would be “he is [MASK]” with a probability of 80%, “he is kind” with a probability of 10%, and “he is nice” with a probability of 10%.

Next Sentence Prediction (NSP) training enables the model to understand how sentences are interrelated, thereby determining whether sentence B should precede or follow sentence A. As previously noted, context-free phrase-level word-embedding models create a single-word embedding representation for each word. The main advantage of BERT is its use of bi-directional learning. BERT is able to read information from right to left and from left to right; by doing so simultaneously, it can account for the context of each occurrence of a given word much more effectively than earlier tools. For example, BERT were presented with the sentences “I went to the bank to deposit a check” and “We walked along the river bank,” it would be able to use right-to-

left clues such as “debit card” and left-to-right clues such as “river.” As a result of this training process, BERT acquires contextual embeddings for words. Once pre-training concludes, the same model can be fine-tuned for a variety of downstream tasks.

BERT’s architecture is founded on the “transformer,” a deep learning model primarily used in natural language processing (NLP). A distinctive advantage of the transformer is that it relies entirely on self-attention for computing input and output representations. The “BERT base” model comprises 12 encoders, each equipped with 12 bi-directional self-attention heads, totaling 110 million parameters. “BERT large” employs 24 encoders and 16 bi-directional attention heads, encompassing 340 million parameters.

3.4.2. Generative Pre-trained Transformers

GPT (Generative Pre-trained Transformer) is a series of language models developed by OpenAI. The GPT series consists of four major versions: GPT-1.0, GPT-2.0, GPT-3.0, GPT-3.5., and GPT-4.0. OpenAI released GPT-1.0 in 2018, representing a significant breakthrough in the field of natural language processing. With 117 million parameters, it effectively understood and generated natural language. GPT-1.0 was primarily used for language translation, text completion, and question-answering tasks. The subsequent version, GPT-2.0, unveiled in 2019, boasted 1.5 billion parameters, enabling it to tackle more complex natural language processing tasks, including story generation, text summarization, and even image captioning. In 2020, GPT-3.0 was released, featuring an autoregressive language model using a transformer architecture with 175 billion parameters, making it one of the largest language models ever developed. The latest addition to the series, GPT-4.0, is the largest yet, six times larger than the GPT-3 model, with approximately one trillion parameters.

Differences between GPT and BERT

Architecture

BERT and GPT use different machine-learning models. As previously discussed, BERT is designed for bidirectional context representation, which means it processes text from both left-to-right and right-to-left directions, facilitating comprehensive context capture. This enhances BERT's understanding of sentence context and meaning. Unlike BERT models, GPT operates as an autoregressive model, generating text sequentially from left to right by predicting the next word in a sentence based on preceding words. This characteristic allows GPT to generate highly coherent and natural-sounding text.

Training data

BERT undergoes training using a masked language model on a large-scale text corpus. The original training data include all English-language Wikipedia articles and BooksCorpus, a dataset containing approximately 11,000 unpublished books, totaling about 800 million words. In contrast, GPT-3 was trained on the WebText dataset, a vast corpus containing web pages from sources like Wikipedia, books, and articles. Additionally, GPT-3 incorporated text from Common Crawl, a publicly available archive of web content.

Pre-training approach

GPT is a generative model that is trained to predict the next word in a sentence or generate an entirely new sentence. This pre-training approach equips GPT with proficiency in language generation and text completion tasks. On the other hand, BERT functions as a discriminative model, meaning that it is trained to classify whether or not a given sentence is coherent. This pre-training strategy allows BERT to excel in sentiment analysis and text classification tasks.

Usability

To use BERT, you need to download the originally published Jupyter Notebook for BERT and then set up a development environment using tools like Google Colab or TensorFlow. If you prefer to avoid the complexities of using a Jupyter Notebook or lack technical expertise, an alternative is to leverage the GPT model using ChatGPT, a straightforward process that involves simply logging into a website.

Application of GPT

GPT has found its way into academic research. For example, Li, Mai, Shen, Yang, and Zhang (2024) employ generative AI to organize financial analysts' view of corporate culture. They find that financial analysts consider business strategy as a key factor for shaping cultural values, and innovation and adaptability are the two cultures that affect almost all aspects of business operations.

The integration of GPT into business applications has given rise to various domain-specific generative AI applications. For instance, Salesforce debuted Einstein GPT, the world's first generative AI designed for customer relationship management. Bloomberg has also developed its own generative AI, BloombergGPT, because the specialized terminology of the financial domain warrants a domain-specific model. BloombergGPT is a large-scale generative AI model trained on a wide range of financial data, catering to a diverse set of NLP tasks within the financial industry. And these, of course, are only the beginning. Future applications of GPT in the field of accounting and finance might include:

Language generation and modeling

GPT can be used to build language models capable of generating new text in specific styles or genres. Its proficiency in generating natural-sounding language makes it useful for applications

such as chatbots, language translation, and content creation. Furthermore, GPT can streamline financial reporting processes by automating the generation of financial statements, analysis reports, and other financial documents, thereby enhancing efficiency and accuracy.

Text summarization and analysis

GPT is capable of condensing large text, such as news articles or research papers, into shorter summaries that encapsulate crucial information. Moreover, it can analyze the sentiment of a given text, allowing businesses to monitor customer feedback and sentiments regarding their products or services. This functionality could prove valuable for financial analysts.

Quantitative forecasting and analysis

GPT can analyze financial data, forecast future trends and outcomes, and identify potential risks associated with investments, loans, and other financial products. Such capabilities assist companies in making well-informed decisions about investments, budgeting, and other financial matters. GPT can also help companies mitigate risks, identify potential instances of fraud or financial irregularities, and ensure compliance with regulations and industry standards.

Other applications of AI in accounting and finance

In recent years, the investment management industry has experienced a rapid surge in adopting AI and machine learning technologies. These technologies have been applied in various areas within the industry, such as identifying trading patterns for generating alphas, streamlining customer support and prospect identification, and managing risk exposure. A notable example of AI and machine learning implementation is Kensho Technologies, a Massachusetts-based startup. Kensho developed an algorithm named "Warren" (in honor of Warren Buffett), which utilizes big data from capital markets and applies machine learning to discover correlations and exploit arbitrage opportunities. Another illustration comes from the hedge funds that have embraced AI

and algorithmic trading. According to a 2018 survey conducted by BarclayHedge with 2,135 hedge fund professionals, 56% of respondents reported using AI/machine learning in their investment strategies, with the primary application being idea generation and portfolio construction.

As AI continues expanding its presence in industrial applications, more attention is being paid to issues of privacy and ethics. In response to privacy concerns, federated learning techniques have emerged. Unlike traditional centralized machine learning methods, federated learning enables AI algorithms to undergo training without sharing or transmitting sensitive data. Overall, the growth of AI in the investment management industry has opened up significant opportunities and will surely continue to do so.

3.5. Prompt engineering for large language model



[Prompt engineering for large language models](#)

Prompt Engineering is a critical aspect for optimizing the performance of language models without altering their underlying parameters. In the context of AI, this means crafting prompts that effectively guide the model to produce desired outputs. Model fine-tuning, on the other hand, involves adjusting the model's parameters to improve its overall performance. Put another way, model fine-tuning is like a student who prepares for their final exam by taking numerous practice exams, whereas prompt engineering is more reminiscent of a student who does not take practice exams but instead takes meticulous notes that are specifically tailored to the content that will appear on the final.

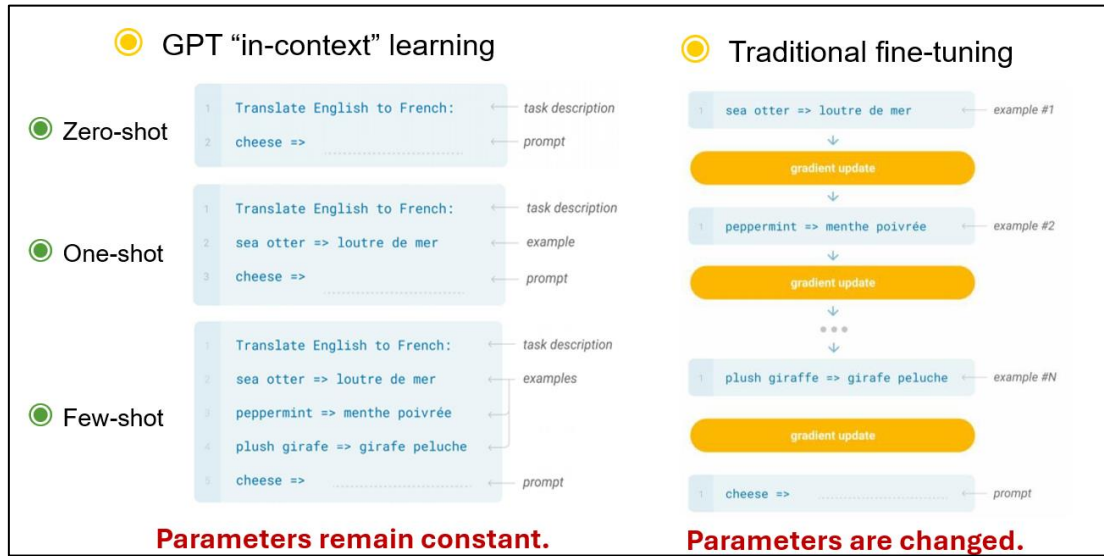
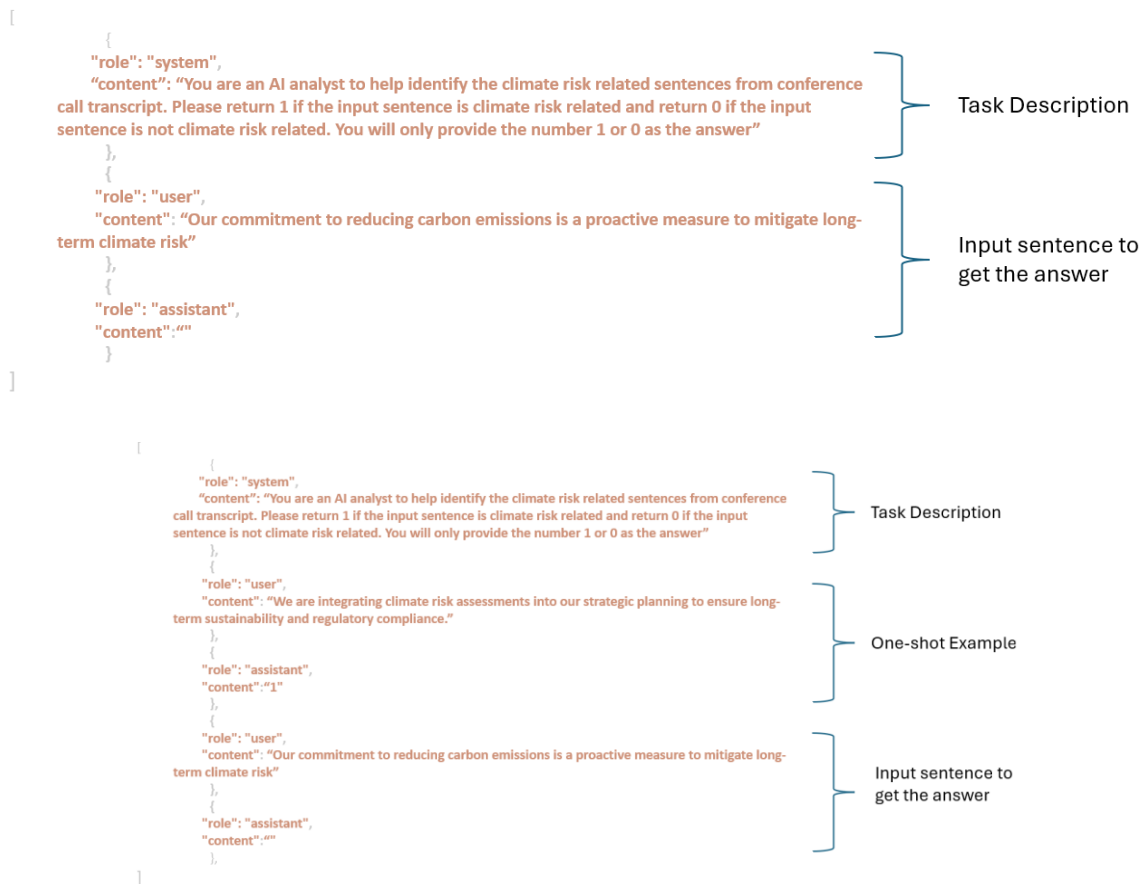
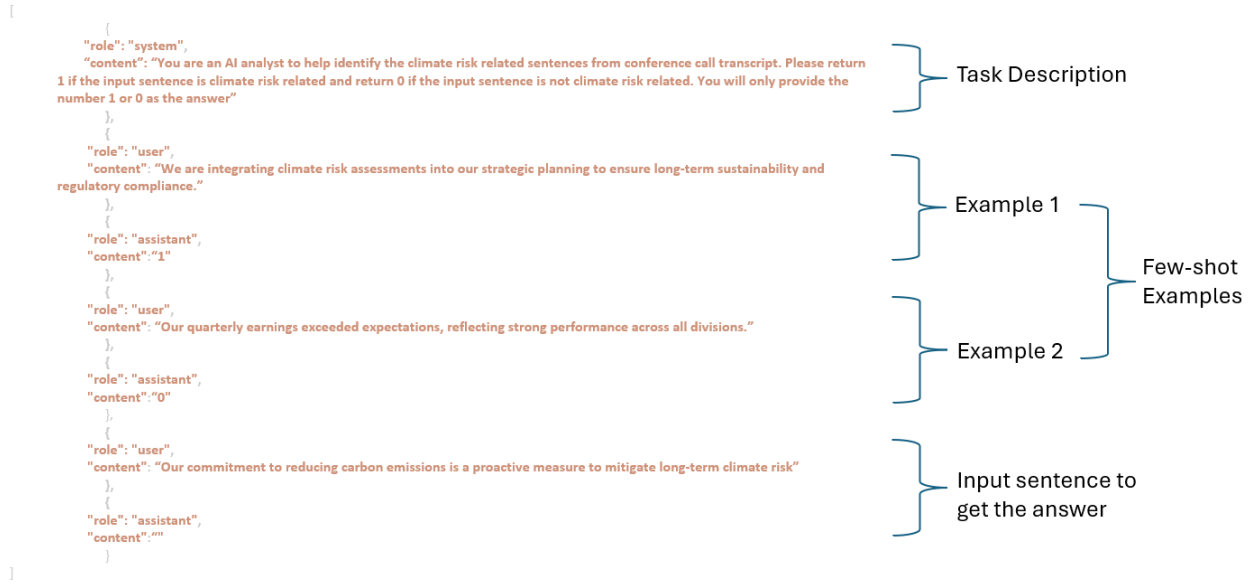
Figure 6. A comparison between prompt engineering and model fine-tuning**Figure 7. Prompt engineering**

Figure 7. Prompt engineering (continued)

There are two primary types of prompt engineering, hard prompt-tuning and soft prompt-tuning, both of which encompass various specific techniques. Hard prompt-tuning involves manually designing natural language prompts that are understandable to humans. For instance, few-shot prompt engineering is a hard prompt-tuning technique in which users provide the model with a few examples or prompts related to the task at hand. This method capitalizes on the model's ability to generalize from a small number of examples. By carefully crafting these prompts to include representative examples and instructions, users can effectively direct the model's responses and enhance its performance on tasks that it has not been explicitly trained for. Figure 7 provides a sample approach. The task is to determine if an input sentence is related to climate risk. A “zero-shot” prompt does not provide any examples while a “few-shot” prompt provides one or more examples to help the model make better decisions.

Chain-of-thought (CoT) prompting is another form of hard prompt-tuning. It requires users to craft prompts designed to guide a language model through a step-by-step reasoning process to arrive at an answer. By structuring the prompt to include a sequence of thought processes, CoT

prompting enables the model to handle more intricate tasks that go beyond simple question-and-answer formats. This method is particularly useful for addressing complex questions that require intermediate steps or logical reasoning to solve. However, creating effective CoT prompts demands a high level of domain knowledge, as the designer must understand the logical steps necessary to construct the reasoning chain that leads to the correct answer.

Figure 8 contrasts standard prompting with CoT prompting. In standard prompting, the model is given a question and an answer without any reasoning steps. As you can see, this approach can lead to errors, as the model might miscalculate or misunderstand the question. In contrast, CoT prompting breaks down the problem into smaller logical steps. In the example, the prompt explicitly calculates the total number of golf balls by first adding the balls from the boxes and then combining them with the existing balls. Then, for the orange problem, the model follows a step-by-step process to subtract the used oranges and then add the newly purchased ones, arriving at the correct answer. The step-by-step reasoning in the prompt helps the model handle more complex questions by focusing on intermediate steps, reducing the likelihood of an error.

Soft prompt-tuning takes a more nuanced approach. These approaches operate at the embedding level, where they adjust the input representations rather than the natural language itself. In the example in Figure 9, the task is to translate a sentence in English to Chinese. At first, the model fails to find a Chinese character for “Sean.” The translation is then improved by optimizing the underlying embedding of the input, after which the model is able to find a Chinese character for “Sean.”

Prefix-tuning is a soft prompt-tuning method that appends a learnable prefix to the input text. Unlike traditional prompt-tuning, where the entire prompt is adjusted, prefix-tuning focuses on optimizing a fixed-length, trainable prefix that is prepended to the input sequence. The prefix,

which consists of a sequence of tokens with associated embeddings, is learned through a process that finetunes the model on specific tasks or domains without altering the core model weights. In terms of time and space, prefix-tuning is a more efficient form of training than fine-tuning the entire model, like taking a couple of electives to expand and enhance your knowledge in a certain area. Fully fine-tuning the entire model is more like committing to a double major. It is both more costly and time-consuming but could also yield the most optimal results.

Figure 8. Chain-of-thought (CoT) prompting

Standard prompting	CoT prompting
Model input Q: James has 5 golf balls. He buys 2 more boxes of golf balls. Each box has 5 golf balls. How many golf balls does he have now? A: 15 Q: The cafeteria had 20 oranges. If they used 10 oranges to make lunch and bought 6 more, how many oranges do they have?	Model input Q: James has 5 golf balls. He buys 2 more boxes of golf balls. Each box has 5 golf balls. How many golf balls does he have now? A: James started with 5 golf balls. 2 boxes of 5 golf balls total 10 golf balls. $5+10=15$. The answer is 15. Q: The cafeteria had 20 oranges. If they used 10 oranges to make lunch and bought 6 more, how many oranges do they have?
Model output A: The answer is 36. (incorrect)	Model output A: The cafeteria had 20 oranges. They used 10 to make lunch. $20-10=10$. They bought 6 more. $10+6=16$. The answer is 16. (correct)

Figure 9. Soft prompt-tuning

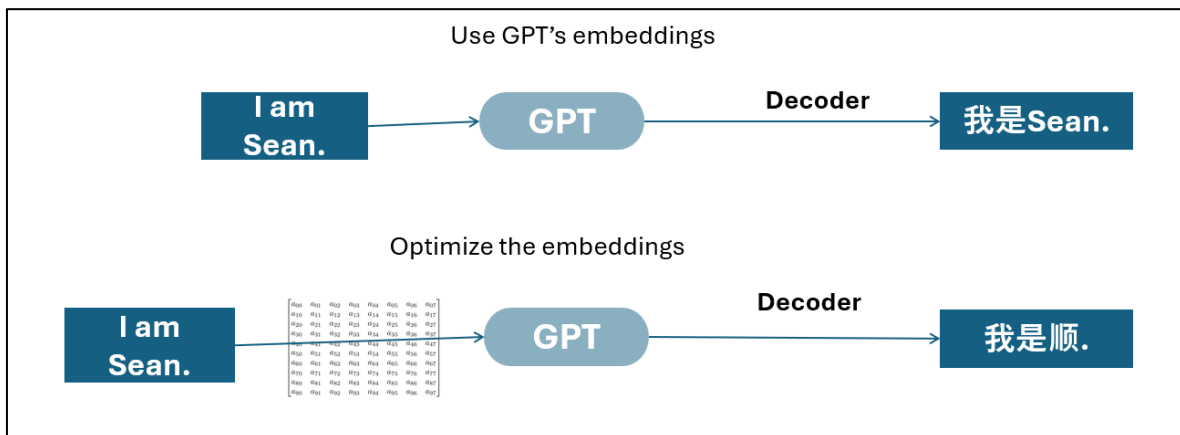
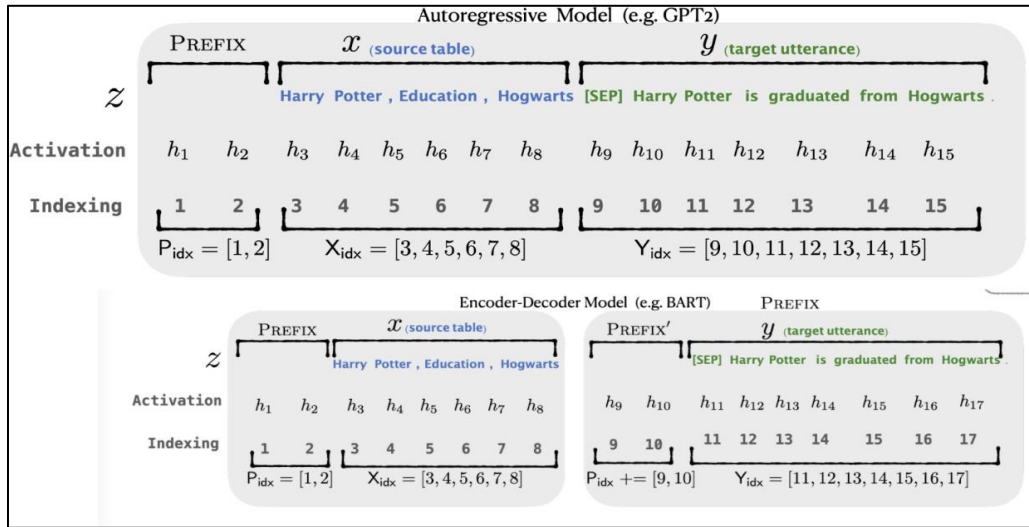


Figure 10. Prefix-tuning

When deciding between prompt engineering and model fine-tuning for adapting language models, it is essential to consider the context and requirements of your task. Prompt engineering leverages a pre-trained model's existing capabilities and involves minimal coding expertise. This method is effective when using high-intelligence language models (LLMs) and is particularly suitable for scenarios where the model's general knowledge is sufficient for the task at hand. On the other hand, model fine-tuning requires large, annotated datasets and significant computing power, because it aims to adjust the model's parameters through either partial or full tuning. Despite these drawbacks, it is particularly useful when working with sensitive data, as it allows for fine-tuning models to maintain control and privacy. Furthermore, for non-open-source LLMs, where access to internal parameters is restricted, parameter tuning is not an option.

3.6. Reinforcement learning

Reinforcement learning (RL) is an emerging field of machine learning that focuses on the actions agents should take in an environment to maximize cumulative reward. It relies on learning through interactions in which an agent explores its environment, receives feedback in the form of

rewards or penalties, and uses this feedback to learn and improve its performance over time. The agent is the learner that interacts with the environment. The environment refers to everything with which the agent interacts and which responds to the agent's actions. The environment provides feedback in the form of a reward signal, which indicates the immediate benefit or cost of the agent's actions. This may seem abstract, but in fact it mimics the everyday trial-and-error processes by which humans acquire many basic and complex skills. In the context of machine learning, however, the goal of the agent is to maximize the cumulative reward over time, often referred to as the return.

Reinforcement learning has numerous practical applications. In robotics, RL is used to teach robots to perform tasks such as grasping objects, walking, and flying. In gaming, RL has been used to develop agents that can play complex games like Go, chess, and Dota 2. In autonomous driving, RL is employed to train self-driving cars to navigate complex road conditions, avoid obstacles, and make real-time decisions. In finance, RL is applied to optimize trading strategies, manage portfolios, and make investment decisions.

Despite its wide application, there are several factors that can make RL difficult to implement. One major challenge is sample inefficiency as many RL algorithms require a large number of interactions with the environment to learn effectively. Another issue is that, due to the constantly evolving actions of the agent, complex RL algorithms can be unstable. Safety and ethics are also critical considerations in reinforcement learning, especially in applications involving autonomous systems and human interaction.

3.7. Man and machine

3.7.1. Machine philosophy

Initialization

In machine learning, initialization strategies reflect a philosophy of iterative improvement that contrasts sharply with many more perfectionist human approaches. While human logic often emphasizes thorough preparation (i.e., studying extensively before taking an exam), machine learning models typically begin with a less refined state and improve over time. Initialization in this context involves starting with a set of initial parameters or answers, which are then refined through iterative learning.

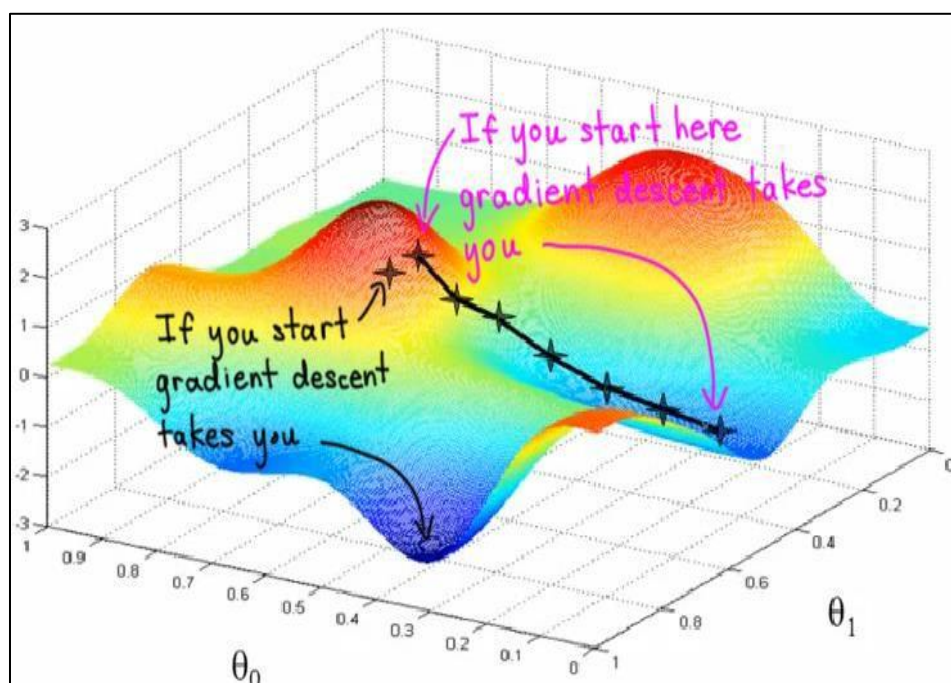
Figure 11. A provoking question

Q: Should we adopt a machine-like philosophy of iterative improvement rather than striving for perfection from the outset?

A: ?

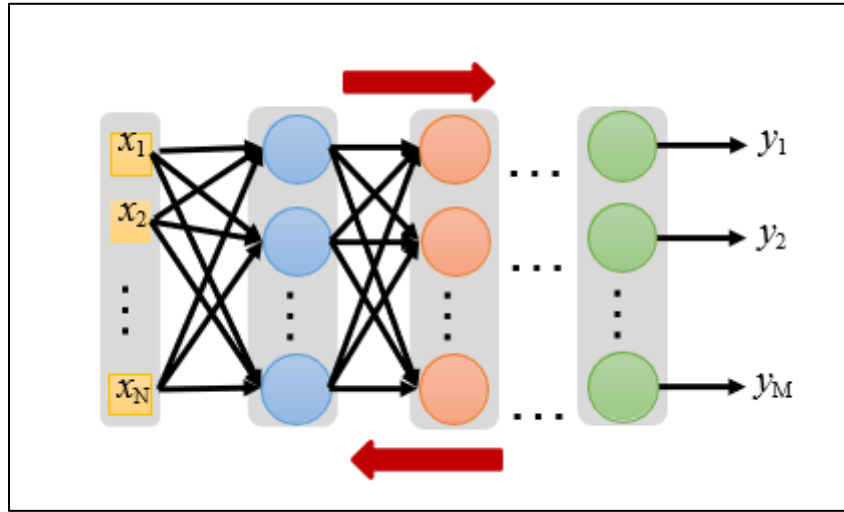


Two common initialization strategies are random initiation and zero initiation. Random initiation starts with randomly chosen values, which, while imprecise, are often closer to the eventual solution and can lead to faster learning and convergence. In contrast, zero initiation begins with uniform values (e.g., zero) and adjusts based on feedback, which can sometimes slow down the learning process.

Figure 12. The impact of initialization parameters

Forward and backpropagation

Human learning often follows a forward propagation approach. We begin by acquiring foundational knowledge—such as learning mathematics—and then apply that knowledge progressively, until, for example, we are able to run regressions with real data. This sequential learning process builds understanding step by step. Machines, however, use both forward and backward propagation to optimize their learning. Forward propagation in machine learning involves applying the model to input data and generating predictions or outputs. Then, moving on to backpropagation, the machine uses the difference between the predicted output and the target output to adjust and refine its parameters through a backward pass. In essence, after generating predictions, the machine revisits and tweaks its previous computations to minimize errors and thereby to achieve iterative improvement.

Figure 13. Forward and backpropagation

Emerging ability

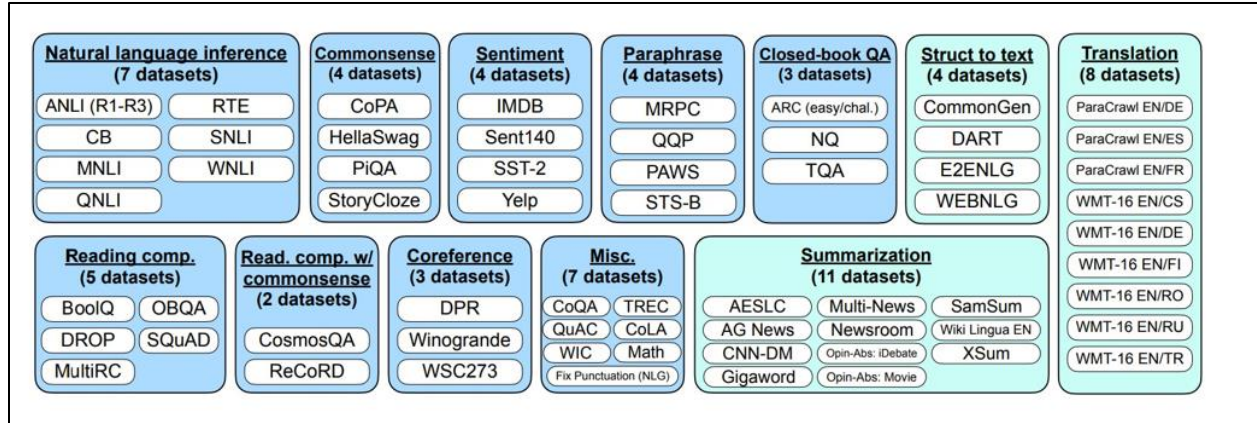
In machine learning, emerging ability refers to the capability of models to handle new, previously unseen tasks by leveraging their training on multiple tasks. Models trained on diverse and varied task prompts develop a broad set of skills and base of knowledge that can be applied to new scenarios. When faced with an unseen task, such a model can assess the similarities and overlaps between the new task and the tasks it has previously encountered. As illustrated in Figure 14, a model trained to perform such tasks as sentiment analysis, paraphrasing, reading comprehension, etc. can eventually draw on this training to perform the task of translation. Essentially, the model's prior experience with multiple tasks equips it with a flexible and adaptable skill set, enabling it to effectively tackle new challenges by recognizing and utilizing commonalities between the new task and the tasks it has already mastered.

3.7.2. Competitive advantages of man and machine

The rapid advancement of AI technologies has given new life to the age-old question of whether and when machines will be able to replace humans. Although AI technologies are increasing the capabilities of machines exponentially, humans and machines each have unique

competitive advantages. These advantages play a pivotal role in determining the tasks where AI could potentially replace humans.

Figure 14. Emerging ability of machines



Firstly, human intelligence is both reasoning-based and probability-based. Reasoning-based intelligence involves making correct decisions through logical deduction, which is inherently challenging for AI. As Cao et al. (2024) revealed, AI underperforms humans in analytical tasks with limited data where reasoning-based intelligence is crucial. On the other hand, AI outperforms humans in probability-based intelligence, which refers to the ability to make decisions based on probability. Consistent with the notion, Cao et al. (2024) find that AI models can understand and analyze large amounts of numerical and textual data and make decisions based on these data. Consequently, AI might replace humans in tasks that mostly rely on probability-based intelligence.

Secondly, spiritual pursuits are unique to humans, and machines possess no such needs or desires. While an AI can assist in constructing a church or translating a piece of scripture, it would not be able to grasp their spiritual dimensions. Similarly, AI has none of humans' capacity for emotions. This emotional distinction represents one of the most significant differences between humans and machines. While machines can design and provide objects facilitating the generation of positive emotions for humans, such as a humanoid robot offering companionship in place of a

deceased loved one, AI cannot produce those emotions in itself, and thus falls short in such tasks as artistic creation and expression.

Thirdly, machines currently lack the human capacity for curiosity. Human curiosity serves as a driving force for acquiring and accumulating knowledge, fostering the development and advancement of AI technologies. The prospect of AI becoming curious one day suggests the potential for the creation of a new generation of AI.

In conclusion, AI can potentially replace humans in certain tasks, such as those requiring probability-based intelligence. However, complete replacement is unfeasible in tasks encompassing unique human traits such as reasoning-based intelligence, spiritual pursuits, emotions, and curiosity.

Appendix 3A: Evaluating machine learning models

Dataset 1 and Dataset 2 contain the validation outcomes of two machine-learning models for detecting spam emails.

- (1) Review sample dataset 1 and calculate accuracy, precision, recall, and F1 score.
- (2) Review sample dataset 2 and calculate accuracy, precision, recall, and F1 score.
- (3) Which model do you think performs better? Explain your reasoning.

Appendix 3B: Prompting engineering

Download an earnings conference call and extract remarks relating to climate risk. Try using different prompts to achieve the best outcome.

References

- Cao, S., Jiang, W., Wang, J., and Yang, B. 2024. From man vs. machine to man+machine: The art and AI of stock analyses. *Journal of Financial Economics*, 160, 103910.
- Cong, L., Liang, T., Zhang, X., and Zhu, W. 2024. Textual factors: A scalable, interpretable, and data-driven approach to analyzing unstructured information. *Management Science*, forthcoming.
- Cong, L., Liang, T., Yang, B., and Zhang, X. 2021. Analyzing textual information at scale. In *Information for Efficient Decision Making: Big Data, Blockchain and Relevance*, 239-271.
- Li, K., Mai, F., Shen, R., Yang, C., Zhang, T. 2023. Dissecting corporate culture using generative AI-Insights from analyst reports. Working paper.

Chapter 4 Analyzing Earnings Conference Calls

4.1. Data structure in earnings conference calls



[Data structure of conference calls](#)

Many U.S. public companies host quarterly conference calls, usually within a month following the close of the fiscal quarter, to discuss their financial performance with investors. These calls typically include key company executives, investors, and financial analysts. During a conference call, company executives review financial information and address major issues impacting company performance in the preceding quarter. They also provide insights into the company's expectations for the upcoming quarters. The format often involves semi-formal presentations by company executives, followed by interactive question-and-answer sessions where investors and financial analysts can seek clarification on any aspect.

In the past, earnings conference calls were only available to professional financial analysts and institutional investors. Nowadays, nearly all public companies to stream their conference calls online or provide on-demand audio recordings of them, extending access to average investors. Furthermore, various online stock research platforms offer access to earnings conference call transcripts. The broad availability of audio recordings and transcripts presents opportunities to deploy AI and machine learning methods for prompt and thorough information processing.

For example, let us look examine the transcript of Microsoft's earnings conference call for the first quarter of fiscal year 2023. The call was held on October 25, 2022. Microsoft's Vice President of Investor Relations, Brett Iversen, hosted the call. Other Microsoft participants included CEO Satya Nadella; CFO Amy Hood; the Chief Accounting Officer, Alice Jolla,; and the Deputy General Counsel, Keith Dolliver,.

Figure 1. Introduction in Microsoft's Fiscal Year 2023 Q1 earnings conference call

BRETT IVERSEN:

Good afternoon and thank you for joining us today. On the call with me are Satya Nadella, chairman and chief executive officer, Amy Hood, chief financial officer, Alice Jolla, chief accounting officer, and Keith Dolliver, deputy general counsel.

Following Brett Iversen's introduction and overview of the structure and principles of the earnings conference call, CEO Satya Nadella stepped in to provide a high-level summary of Microsoft's strategies, the financial performance of major business units such as Microsoft Cloud, and expectations for the upcoming quarters.

Figure 2. CEO remarks in Microsoft's fiscal year 2023 Q1 earnings conference call

SATYA NADELLA:

Thank you, Brett.

To start, I want to outline the principles that are guiding us through these changing economic times: First, we will invest behind categories that will drive the long-term secular trend, where digital technology as a percentage of the world's GDP will continue to increase. Second, we will prioritize helping our customers get the most value out of their digital spend, so that they can do more with less. And, finally, we will be disciplined in managing our cost structure.

With that context, this quarter, the Microsoft Cloud again exceeded \$25 billion in quarterly revenue, up 24 percent and 31 percent in constant currency. And, based on current trends continuing, we expect our broader commercial business to grow at around 20 percent in constant currency this fiscal year, as we manage through the cyclical trends affecting our consumer business.

Following Satya Nadella's high-level summary, the CFO, Amy Hood, shared detailed financial information and provided her interpretation of the data from the company's perspective. For example, she explained that increased operating expenses primarily resulted from the growth in headcount, while a shift in sales mix and foreign exchange fluctuations adversely impacted the operating margin. Additionally, Hood presented an outlook for the second quarter of the fiscal year, at both the company and segment levels.

Figure 3. CFO remarks in Microsoft's Fiscal year 2023 Q1 earnings conference call**AMY HOOD:**

Thank you, Satya, and good afternoon everyone. Our first quarter revenue was \$50.1 billion, up 11 percent and 16 percent in constant currency. Earnings per share was \$2.35 – and increased 4 percent and 11 percent in constant currency, when adjusted for the net tax benefit from the first quarter of fiscal year 22. Driven by strong execution in a dynamic environment, we delivered a solid start to our fiscal year, in line with our expectations, even as we saw many of the macro trends from the end of the fourth quarter continue to weaken thru Q1. ...

Operating expense increased 15 percent and 18 percent in constant currency driven by investments in cloud engineering, LinkedIn, Nuance and commercial sales. At a total company level, headcount grew 22 percent year-over-year as we continued to invest in key areas just mentioned, as well as customer deployment. Headcount growth included roughly 6 points from the Nuance and Xandr acquisitions, which closed last Q3 and Q4, respectively.

Operating income increased 6 percent and 15 percent in constant currency. And operating margins decreased roughly 2 points year-over-year to 43 percent. Excluding the impact of the change in accounting estimate, operating margins declined roughly 4 points year-over-year driven by sales mix shift to cloud, unfavorable FX impact, Nuance, and the lower Azure margin noted earlier.

Finally, the floor was open to questions from other participants in the earnings conference call. During this call, eight investors and analysts asked questions that spanned both Microsoft's operating and financial decisions. For instance, an analyst from Stanley sought clarification on how Microsoft formulated the outlook guidance. Other analysts and investors inquired about future plans for various business segments, such as Microsoft Cloud, Windows, and advertising, in light of the past performance of Microsoft and its competitors.

4.2. Standard dependence parser



[Textual analysis: sentence level analysis using NLP parser](#)

Chapter 2 and Chapter 3 introduce three textual analysis methods: the Bag-of-Word approach, phrase-level word embedding, and sentence-level word embedding. Bag-of-words is a frequency-based method that summarizes textual data with numeric information but disregards the meanings and contexts of words. Phrase-level word embedding incorporates word meanings and recognizes words with similar meanings, yet it cannot handle the instances where the same word has different meanings in distinct sentences. In contrast, sentence-level word embedding allows algorithms to

recognize not only distinct words sharing similar meanings and different meanings of the same words in various sentences. Nevertheless, all these methods only consider the relationships among individual words, disregarding the grammatical relationships within a sentence.

Figure 4. Q&A session in Microsoft’s Fiscal year 2023 Q1 earnings conference call

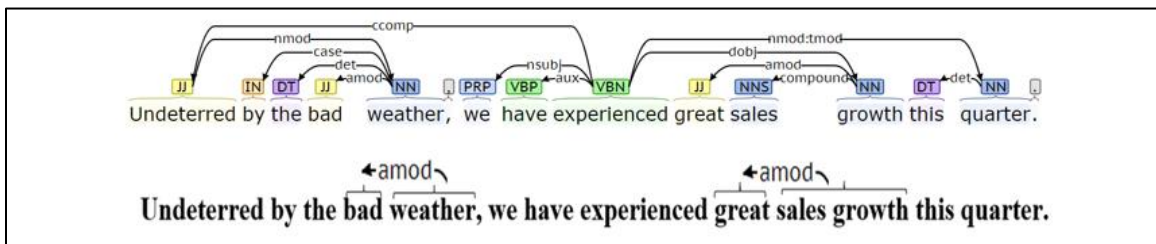
BRETT IVERSEN: Thanks, Amy. We’ll now move over to Q&A. Out of respect for others on the call, we request that participants please ask only one question. Jessi, can you please repeat your instructions?

(Operator Direction.)

KEITH WEISS, Morgan Stanley: Excellent. Thank you, guys for taking the question, and impressive results in what is obviously a very difficult environment. A question for Amy, and it pains me to ask a question about a percentage point, but I think this is what’s on a lot of investors’ minds, is this is two quarters in a row now where Azure constant currency growth came in a little bit below your guidance. And I think what investors are worrying about or sort of wondering about is there an inherent volatility in that business that’s just harder to forecast?

In linguistics, words in a sentence are classified into parts of speech (e.g., nouns, verbs, adverbs, etc.) and are interconnected to form certain dependency relationships. As an example, in the sentence, “Undeterred by the bad weather, we have experienced great sales growth this quarter,” the word “weather” functions as a noun modified by the adjective “bad,” while the noun “growth” is modified by the adjective “great.” Table 1 provides a summary of the most common dependency relationships.

Figure 5. Dependency relationship



A *dependency parser* is a data analytics tool used to analyze the grammatical structure of a sentence. It can identify the “head” word in a sentence and the words that modify it. The Nature

Language Processing (NLP) group at Stanford University has pioneered a neural network model that trains an algorithm to identify “part-of-speech” and “dependency relationships.”

Table 1. Dependency relation table

Core dependents of clausal predicates			Non-core dependents of clausal predicates			Special clausal dependents		
Nominal dep	Predicate dep		Nominal dep	Predicate dep	Modifier word	Nominal dep	Auxiliary	Other
nsubj	csubj		nmod	advcl	advmod	vocative	aux	mark
nsubjpass	csubjpass				neg	discourse	auxpass	punct
dobj	ccomp	xcomp				expl	cop	
iobj								
Noun dependents			Compounding and unanalyzed			Coordination		
Nominal dep	Predicate dep	Modifier word						
nummod	acl	amod	compound	mwe	goeswith	conj	cc	punct
appos		det	name	foreign				
nmod		neg						
Case-marking, prepositions, possessive			Loose joining relations			Other		
case			list	parataxis	remnant	Sentence head	Unspecified dependency	
			dislocated		reparandum	root	dep	

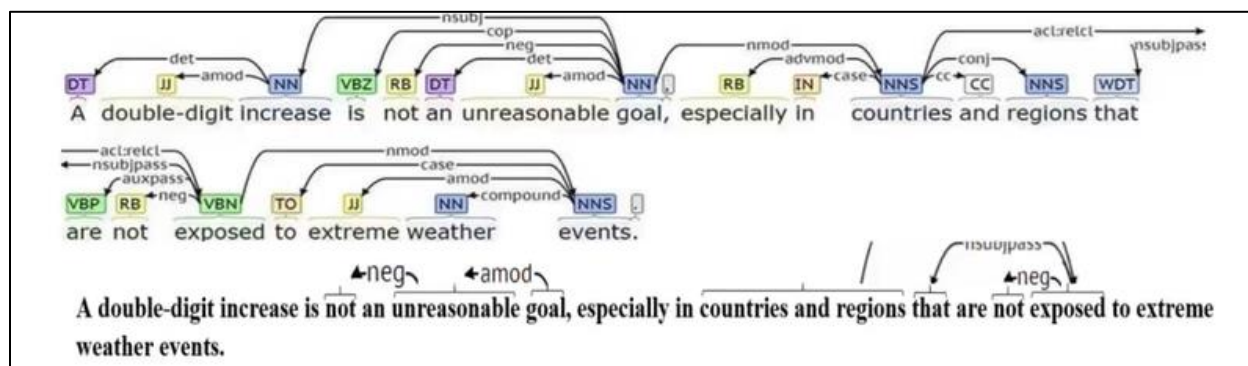
The parser builds a parse by performing a linear-time scan across the words of a sentence. At every step, it maintains a partial parse, a stack of words currently under processing, and a buffer of words yet to be processed. The parser uses a neural network classifier at every stage to determine grammatical relationships among the words. This classifier is trained using a sample of three million words from various sources such as *Wall Street Journal* articles, IBM computer manuals, nursing notes, and transcribed telephone conversations. The researchers divided the three million words into ten groups, using 90 percent of the sample for model training. They then used the trained model to predict the dependency relationships of the remaining 10 percent of the sample so as to evaluate the parser’s accuracy. To accurately recognize word dependency relationships, this iterative process spans years, refining and enhancing the model’s robustness. The reported accuracy of the current model is 92 percent.

As discussed previously, a transfer model is a machine learning model initially pre-trained on one task and then fine-tuned for a different yet related task. The concept behind transfer learning is rooted in the notion that a model trained on a large and diverse dataset can be repurposed for other tasks. In contrast, a neural model is a type of machine learning model designed to emulate the structure and functionality of the human brain. These models are composed of interconnected layers of artificial neurons and undergo training using large amounts of data to learn patterns and make predictions.

Advantages of a human-supervised standard dependency parser

The standard dependency parser has distinct advantages over other textual analysis methods. Let's return to the sentence, "Undeterred by the bad weather, we have experienced great sales growth this quarter." While a frequency-based bag-of-words approach can only identify one positive word, "great", and one negative word, "bad," the standard dependency parser can distinguish that "bad," modifying "weather," is unrelated to firm performance, whereas "great," modifying "sales," is related to it.

Consider another example (Figure 6): "A double-digit increase is not an unreasonable goal, especially in countries and regions not exposed to extreme weather events." The frequency-based bag-of-words approach would merely identify two negative words, "unreasonable" and "exposed." By contrast, the standard dependency parser can discern that "not" and "unreasonable" together form a double negative that modifies "goal," indicating a positive sentiment related to firm performance. Meanwhile, "not" and "exposed" form another double negative that modifies "countries and regions." This, too, is a positive sentiment, but unlike "goal," "countries and regions" are unlikely to be relevant to firm performance.

Figure 6. Using dependency relationships to process double negatives

4.3. Empirical example: Measuring corporate culture using machine learning



[Business application of using conference call transcripts](#)

Corporate culture serves as the invisible force that guides various aspects of a firm when decisions and actions cannot be precisely regulated. Despite its significance, the elusive and multidimensional nature of corporate culture poses challenges in quantifying it in a large sample. Li, Mai, Shen, and Yan (2023) develop a semisupervised machines learning algorithm to measure corporate culture.

They start with five core corporate values, namely innovation, integrity, quality, respect, and teamwork. For each of these values, they construct a seed word list. As an example, the seven seed words for the culture value of teamwork are collaborate, collaboration, collaborative, cooperate, cooperation, cooperative, and teamwork. Using the word embedding method, they create a “culture dictionary.” Specifically, they train a neural network model using *word2vec* that learns the meanings of words and phrases in earnings call transcripts. This is achieved by encoding words and phrases as numeric vectors. A cosine similarity is then computed between each unique word or phrase in earnings call transcripts and the seed words. Words and phrases with the closest association to the seed words are considered as culture-related words and phrases. After compiling the “culture dictionary”, each firm’s cultural value in each of the five dimension is measured as

the weighted count of the number of words associated with each value divided by the total number of words in the document.

The study documents that corporate culture is associated with operational efficiency, risk-taking, earnings management, executive compensation, firm value, and deal making. Furthermore, there is evidence that corporate culture is shaped by major corporate events such as mergers and acquisitions.

4.4. Empirical example: From words to syntax: Identifying context-specific information in textual analysis

Cao, Kim, Wang, and Xiao (2020) adapt the Stanford neural NN-based dependency tree parser to construct a performance-based tone measure for conference calls. Their approach accounts for syntactical features. They employ the Stanford NN-based parser to organize conference call narratives into a tree structure representation to achieve this. This structural transformation helps extract both lexicalized and higher-order interactions between words. This method facilitates the differentiation of sentiment words linked to performance from unrelated sentiment words, improving the calculation of performance-orientated tone.

To construct a performance-related tone measure, Cao et al. (2022) also need to create a comprehensive list of performance-related keywords. They employ word embedding to generate a corporate-context-specific keyword dictionary of performance keywords, which does not involve human labor or judgement. The following example illustrates the intuitive principles underlying word embedding. In everyday conversation, the term “liability” means “expected responsibility,” whereas, in corporate disclosures, it means “debt.” The objective is to teach the machine to understand the nuanced context of each word within the document. Word embedding achieves this by associating a word’s meaning with its contextual neighbors. For example, in the vicinity of of

“liability,” one might find words such as “debt” occurring 100 times, “borrowing” 50 times, and “cash” 10 times. Word embedding assigns a vector to represent the meaning of “liability” in this example [100 (debt), 50 (borrowing), 10 (cash)]. The machine then discerns that the definition of “liability” most pertinent to the business context is the one related to debt.

To implement word embedding, Cao et al. (2022) start with the SeedList, a foundational list of performance-related words including terms like return, margin, EPS, net, income, earning, result, gain, profit, performance, and sale. Following the following procedure, they then expand it to form an extensive compilation of performance-related words. Figure 7 illustrates the resulting word cloud representing these performance-related words.

- Calculate 100-dimensional vectors to represent the semantic meaning of each word in the vocabulary list (W) in the corpus.
- Determine the vector representation for all words in the SeedList, calculating their average vector to serve as the ultimate semantic representation for the SeedList of performance-related words.
- Compute the cosine similarity between each word in the vocabulary list (W) and the mean vector of the SeedList.
- If the angle between them is less than 45 degrees, which corresponds to a cosine value exceeding $\cos 45 \text{ degrees} = 0.707$, indicating an acute angle of less than 45 degrees, the word is considered to be similar to the seed list,.

Figure 7 Word Cloud of Performance-related Words



This figure reports the list of performance-related words. Font size of a word is proportional to the frequency of each word in conference call transcripts.

Source: Cao et al. (2022)

Appendix 4: Applying GPT to analyze conference call transcripts using both API and web interface



[Applying GPT to analyze conference call transcripts](#)

Download 5 earnings conference calls and extract the CEO's prepared remarks. Use OpenAI ChatGPT API to perform LDA topic modeling and classify the remark into an appropriate number of topics. Report the topics, their weight in the CEO's prepared remark, and the top five most frequent words in each topic. Then, use the ChatGPT web interface to perform the same task. Compare the outputs.

References

- Cao, S., Kim, Y., Wang, A., and Xiao, H. 2020. From words to syntax: Identifying context-specific information in textual analysis. Working paper.
- Li, K., Mai, F., Shen, R., and Yan, X. 2021. Measuring corporate culture using machine learning. *Review of Financial Studies*, 34(7), 3265-3315.

Chapter 5 Analyzing Material Company News

5.1 Data structure in 8-K filings



[Data structure of the 8-K filing](#)

In addition to filing annual and quarterly reports, public companies are mandated to report certain material corporate events to their shareholders on a more current basis using Form 8-K, or a “current report.” The types of information that trigger Form 8-K filings are generally considered “material.” As such, companies are obligated to disclose such information promptly rather than waiting until the end of a fiscal period as they would for Forms 10-Q and 10-K.

In March 2004, the SEC adopted sweeping changes to the Form 8-K disclosure requirements. The revised rules introduced new items and events to be disclosed in Form 8-K and require Form 8-K to be filed within four business days of the triggering event and, in some cases, even earlier. The rest of this section reviews each type of information disclosed in Form 8-K and provides examples of 8-K filings.

Section 1 Registrant’s Business and Operations

Item	Added post-2004?
Item 1.01 Entry into a Material Definitive Agreement	Yes
Item 1.02 Termination of a Material Definitive Agreement	Yes
Item 1.03 Bankruptcy or Receivership	No

Item 1.01 Entry into a Material Definitive Agreement

This item pertains to business agreements outside the ordinary course of business, including material amendments to those agreements. For instance, if a company secures a substantial loan from a bank or enters into a long-term lease that is material to the company, the agreement must be reported under Item 1.01 by filing a current report. However, if a retailer with an established

chain of stores signs a lease for one additional store, the new lease generally would be considered part of the ordinary course of business and would not need to be reported here.

The required disclosure includes the date of entering into or amending the agreement, the parties' identity, and a brief description of the terms and conditions. Figure 1 provides an example of Item 1.01 filed by Amazon Inc., where it disclosed a credit agreement with Bank of America, N.A., on September 5, 2014. The agreement provided Amazon Inc. with a credit facility that had a borrowing capacity of up to \$2 billion at an initial interest rate of the London Interbank offered rate (LIBOR) plus 0.625%.

Figure 1. Item 1.01 filed by Amazon Inc.

<p>ITEM 1.01. ENTRY INTO A MATERIAL DEFINITIVE AGREEMENT.</p> <p>On September 5, 2014, Amazon.com, Inc. (the "Company") Bank of America, N.A., as administrative agent, and the lenders party thereto entered into a credit agreement (the "Credit Agreement"). The Credit Agreement provides the Company with an unsecured revolving credit facility with a borrowing capacity of up to \$2.0 billion. The term of the Credit Agreement is two years, but it may be extended for up to three additional one-year terms if approved by the lenders.</p> <p>The initial interest rate applicable to outstanding balances under the Credit Agreement is the London interbank offered rate ("LIBOR") plus 0.625%, with a commitment fee of 0.06% on the undrawn portion of the credit facility under our current credit ratings. If the Company's credit ratings are downgraded these could increase to as much as LIBOR plus 1.00% and up to 0.10%, respectively.</p> <p>Borrowings under the Credit Agreement may be used for working capital, capital expenditures, acquisitions, and other corporate purposes. The Company currently has no borrowings outstanding under the Credit Agreement, but expects to borrow under the Credit Agreement from time-to-time in the ordinary course of business.</p> <p>The Credit Agreement contains customary representations and warranties, covenants, and events of default, but does not contain financial covenants. Upon an event of default that is not cured within applicable grace periods or waived, any unpaid amounts under the Credit Agreement may be declared immediately due and payable and the commitments may be terminated.</p> <p>The foregoing description is qualified by reference to the full text of the Credit Agreement, which is filed as Exhibit 10.1 to this Current Report on Form 8-K.</p> <p>The financial institutions party to the Credit Agreement and their respective affiliates are full service financial institutions engaged in various activities, which may include sales and trading, commercial and investment banking, advisory, investment management, investment research, principal investment, hedging, market making, brokerage, and other financial and non-financial activities and services. Certain of these financial institutions and their respective affiliates have provided, and may in the future provide, a variety of these services to the Company and to persons and entities with relationships with the Company, for which they received or will receive customary fees and expenses.</p>

Item 1.02 Termination of a Material Definitive Agreement

This item concerns the termination of material business agreements. For example, if a company procures most of its raw material through a long-term procurement agreement with one significant supplier, and that supplier terminates the agreement, the termination must be reported under this item. In contrast, if the agreement merely expires according to its terms, it need not be reported on Form 8-K. The required disclosure includes the date of the termination of the material definitive agreement, the identity of the parties to the agreement, a brief description of the terms and conditions of the agreement, a brief description of the material circumstances causing the termination, and any material early termination penalties.

Item 1.03 Bankruptcy or Receivership

In the event of a potential bankruptcy, the company must disclose this information on Form 8-K along with its reorganization plan (under Chapter 11) or liquidation (under Chapter 7). This information is important for shareholders as they need to evaluate their potential losses and consider the likelihood of the company emerging from bankruptcy.

Section 2 Financial Information

Item	Added post-2004?
Item 2.01 Completion of Acquisition or Disposition of Assets	No
Item 2.02 Results of Operations and Financial Condition	No
Item 2.03 Creation of a Direct Financial Obligation or an Obligation under an Off-Balance Sheet Arrangement of a Registrant	Yes
Item 2.04 Triggering Events That Accelerate or Increase a Direct Financial Obligation or an Obligation under an Off-Balance Sheet Arrangement	Yes
Item 2.05 Costs Associated with Exit or Disposal Activities	Yes
Item 2.06 Material Impairments	Yes

Item 2.01 Completion of Acquisition or Disposition of Assets

If a company acquires or divests a significant amount of assets, including acquiring or merging with another company or selling a business unit, the company must file an 8-K to describe the terms of the transaction.

Item 2.02 Results of Operations and Financial Condition

Many companies announce their quarterly and annual results simultaneously in an 8-K filing. If the company plans to hold an earnings conference call, this information is also disclosed in the 8-K filing. As shown in Figure 2, Amazon Inc. filed a Form 8-K along with the announcement of

its financial results from both the fourth quarter of 2020 and the year ending on December 31, 2020.

Figure 2. Item 2.02 filed by Amazon Inc.

<p>ITEM 2.02. RESULTS OF OPERATIONS AND FINANCIAL CONDITION.</p> <p>On February 2, 2021, Amazon.com, Inc. announced its fourth quarter 2020 and year ended December 31, 2020 financial results. A copy of the press release containing the announcement is included as Exhibit 99.1 and additional information regarding the inclusion of non-GAAP financial measures in certain of Amazon.com, Inc.'s public disclosures, including its fourth quarter 2020 and year ended December 31, 2020 financial results announcement, is included as Exhibit 99.2. Both of these exhibits are incorporated herein by reference.</p>
--

Item 2.03 Creation of a Direct Financial Obligation or an Obligation under an Off-Balance Sheet Arrangement of a Registrant

Companies must report the fundamental terms of material financial obligations, such as long-term debt, capital or operating leases, and short-term debt outside the ordinary course of business. The mandatory disclosure includes the date when the company incurred the direct financial obligation, a concise description of the transaction that led to the obligation, the amount of the direct financial obligation, and a brief overview of the other terms and conditions of the transaction. Illustrated in Figure 3 is an example of Item 2.03 in a Form 8-K filed by USHG Acquisition Corp. on March 29, 2022, disclosing the issuance of an unsecured non-interest-bearing promissory note in the principal amount of \$500,000 on March 29, 2022.

Figure 3. Item 2.02 filed by USHG Acquisition Corp.

<p>Item 2.03 Creation of a Direct Financial Obligation or an Obligation under an Off-Balance Sheet Arrangement of a Registrant.</p> <p>On <u>March 29, 2022</u> USHG Acquisition Corp. (the "Company") issued <u>an unsecured promissory note</u> (the "Note") in the principal amount of <u>\$500,000</u> to affiliates of USHG Investments, LLC (the "Sponsor"). The Note does not bear interest and is repayable in full upon consummation of the Company's initial business combination (a "Business Combination"). If the Company does not complete a Business Combination, the Note shall not be repaid and all amounts owed under it will be forgiven except to the extent that the Company has funds available to it outside of its trust account established in connection with its initial public offering. Upon the consummation of a Business Combination, the affiliates of the Sponsor shall have the option, but not the obligation, to convert the principal balance of the Note, in whole or in part, to warrants of the Company, at a price of \$1.50 per warrant (the "Warrants"). The terms of the Warrants will be identical to the terms of the warrants issued by the Company to the Sponsor in a private placement that took place simultaneously with the Company's initial public offering. The Note is subject to customary events of default, the occurrence of which, in certain instances, would automatically trigger the unpaid principal balance of the Note and all other sums payable with regard to the Note becoming immediately due and payable.</p> <p>The Note was issued pursuant to the exemption from registration contained in Section 4(a)(2) of the Securities Act of 1933, as amended.</p> <p>The Note is attached as Exhibit 10.1 to this Current Report on Form 8-K and is incorporated herein by reference. The disclosure set forth in this Item 2.03 is intended to be a summary only and is qualified in its entirety by reference to the Note.</p>
--

Item 2.04 Triggering Events That Accelerate or Increase a Direct Financial Obligation or an Obligation under an Off-Balance Sheet Arrangement

This item captures events that accelerate or increase financial obligations, such as loan defaults. If a company defaults on a loan, its creditors can demand immediate payment of the entire outstanding amount. In such a case, the company must disclose the date of the triggering event, a brief description of the triggering event, the amount to be repaid, the repayment terms, and any other financial obligations that might arise from the initial default.

Item 2.05 Costs Associated with Exit or Disposal Activities

This item requires companies to disclose material charges associated with restructuring plans. The required disclosure includes the date of the commitment to the exit or disposal activities, a description of the plan, and an estimate of the total expected cost.

Item 2.06 Material Impairments

A company must disclose certain material impairments under this item, including the date of the conclusion that a material charge is required, a description of the impaired assets, the facts leading to the conclusion, and an estimate of the amount of the impairment charge.

Section 3 Securities and Trading Markets

Item	Added post-2004?
Item 3.01 Notice of Delisting or Failure to Satisfy a Continued Listing Rule or Standard; Transfer of Listing	Yes
Item 3.02 Unregistered Sales of Equity Securities	No
Item 3.03 Material Modification to Rights of Security Holders	No

Item 3.01 Notice of Delisting or Failure to Satisfy a Continued Listing Rule or Standard; Transfer of Listing

This item mandates companies to disclose the delisting of their stock if they receive notification from the stock exchange that they no longer meet the requirements for continued listing. A company receiving this type of notice must disclose the date it was received, the rule or standard for continued listing that they failed to meet, and any response the company has determined to make. Delisting due to non-compliance with listing requirements often signals red flags to investors.

Item 3.02 Unregistered Sales of Equity Securities

Under this item, companies must disclose any private issuance of securities exceeding one percent of outstanding shares of that class.

Item 3.03 Material Modification to Rights of Security Holders

This item requires firms to disclose material changes to or limitations on the rights of shareholders that result from the issuance or modification of another class of securities.

Section 4 Matters Related to Accountants and Financial Statements

Item	Added post-2004?
Item 4.01 Changes in Registrant's Certifying Accountant	No
Item 4.02 Non-Reliance on Previously Issued Financial Statements or a Related Audit Report or Completed Interim Review	Yes

Item 4.01 Changes in Registrant's Certifying Accountant

When a company changes their independent auditor, it can raise concerns regarding the integrity of financial statements. As a result, companies must disclose any such changes regardless of whether the independent auditor is involuntarily dismissed, voluntarily resigns, or declines to

stand for reappointment. If they hire a new auditor, the company should disclose that in Form 8-K as well.

Item 4.02 Non-Reliance on Previously Issued Financial Statements or a Related Audit Report or Completed Interim Review

This item requires companies to inform information users when previously issued financial statements contain errors or when previously issued audit reports or interim reviews should no longer be relied upon. Companies must disclose the date when they concluded that these statements were not reliable, identify the financial statements and years or periods covered, and briefly describe the facts underlying the conclusion.

Section 5 Corporate Governance and Management

Item	Added post-2004?
Item 5.01 Changes in Control of Registrant	No
Item 5.02 Departure of Directors or Certain Officers; Election of Directors; Appointment of Certain Officers; Compensatory Arrangements of Certain Officers	No
Item 5.03 Amendments to Articles of Incorporation or Bylaws; Change in Fiscal Year	No
Item 5.04 Temporary Suspension of Trading Under Registrant's Employee Benefit Plans	No
Item 5.05 Amendments to the Registrant's Code of Ethics, or Waiver of a Provision of the Code of Ethics	No
Item 5.06 Change in Shell Company Status	Yes
Item 5.07 Submission of Matters to a Vote of Security Holders	Yes

Item 5.01 Changes in Control of Registrant

Whenever there is a change in control of the registrant, companies must disclose the persons who have acquired control and any arrangements between the old and new control groups.

Item 5.02 Departure of Directors or Certain Officers; Election of Directors; Appointment of Certain Officers; Compensatory Arrangements of Certain Officers

Companies are obligated to disclose any alterations to the board of directors or high-level executive officers, including changes to the compensation of existing high-level officers. When there is a change in the board, this disclosure must include the date of the director's resignation, refusal to stand for re-election or removal, any roles held by the director on any committee of the board of directors at the time, and a brief description of the circumstances that management believes caused the director's departure. Figure 4 shows item 5.02, filed by WeTrade Group Inc., relating to the resignation of the Chief Executive Officer on September 1, 2020.

Item 5.03 Amendments to Articles of Incorporation or Bylaws; Change in Fiscal Year

If a company amends its articles of incorporation or bylaws or changes its fiscal year, the company should disclose the changes under Item 5.03.

Figure 4. Item 5.02 filed by WeTrade Group Inc.

<p>Item 5.02 Departure of Directors or Certain Officers; Election of Directors; Appointment of Certain Officers; Compensatory Arrangements of Certain Officers.</p> <p>(b) On September 1, 2020 WeTrade Group Inc.'s Chief Executive Officer Dai Zheng resigned as Chief Executive Officer. Dai Zheng will remain as a Director of WeTrade Group Inc. As of September 1, 2020 Mr. Dai Zheng will serve as the Chairman of WeTrade Group Inc.</p> <p>Mr. Dai is a graduate of Fuzhou Finance University in PRC and majored in Finance and Economics. Mr. Dai began his career in internet and information technology industry in 1998. Between 2000 to 2004, he served as Chief Technology Officer ("CTO") for China Interaction Media Group. Between 2006 to 2012, he was a co-founder and Vice President of Qunar Cayman Islands Limited (stock code: QUNR). Since 2014, Mr. Dai has served on a number of boards that represent timeshare owners and their interests. Mr. Dai's prime duty for the Company will be to leverage his existing industry connections to assist in the implementation of the business plan.</p>

Item 5.04 Temporary Suspension of Trading Under Registrant's Employee Benefit Plans

Companies are required to file Item 5.04 when they temporarily suspend trading of the company's equity securities by participants in an individual account plan, such as a 401(k) plan, due to a blackout period.

Item 5.05 Amendments to the Registrant's Code of Ethics, or Waiver of a Provision of the Code of Ethics

If a company changes the ethics code that applies to the high-level officers, it must disclose them in Item 5.05. The company must also disclose any waivers granted to the high-level officers.

Item 5.06 Change in Shell Company Status

Companies are required to submit a Form 8-K under Item 5.06 when a transaction leads the company to cease being a shell company. This is illustrated in Figure 5, wherein WeTrade Group Inc. disclosed that the company ceased to be a shell company due to the commencement of regular revenue-generating operations and controls of the WePay System.

Item 5.07 Submission of Matters to a Vote of Security Holders

Companies must disclose the results of shareholder votes at annual or special meetings by filing Form 8-K under this item.

Figure 5. Item 5.06 filed by WeTrade Group Inc.

Item 5.06 Change in Shell Company Status
WeTrade Group Inc. has ceased to be a shell company (as defined under Rule 405 of the Securities Exchange Act of 1934, as amended) as of the result of commencement of active business operations, development of assets and generation of revenue therefrom by its wholly owned subsidiary, Yue Shang Information Technology (Beijing) Co Limited. The information contained in this 8-K constitutes the current "Form 10 information" necessary to satisfy the conditions contained in Rule 144(i)(2) under the Securities Act of 1933, as amended.
In January 2020, WeTrade Group appointed a 3rd party software company to develop an auto-billing management system ("WePay System") at the cost of RMB 400,000 (\$57,143) in order to provide online payment services for its online store customers in PRC. The resulting WePay System is an asset currently valued at \$56,191
On March 1, 2020 WeTrade Group Inc.'s wholly owned subsidiary Yue Shang Information Technology (Beijing) Co Limited entered into a Technical Entrust (Agency) Agreement with Global Joy Trip Technology (Beijing) Co Limited. WeTrade Group Inc.'s CEO and Director Dai Zheng is a majority shareholder in Global Joy Trip Technology (Beijing) Co Limited and commenced operations.
As per the terms of the agreement subsidiary Yue Shang Information Technology (Beijing) Co Limited provides social e-commerce revenue management system services, "the WePay System," to Global Joy Trip Technology (Beijing) Co Limited.
Global Joy Trip Technology (Beijing) Co Limited pays 2% of the actual Gross Merchandise Volume ("GMV") generated during the operational period to Yue Shang Information Technology (Beijing) Co Limited as the system service fee for using the WePay System. As a result of the foregoing activity, WeTrade Group Inc.'s subsidiary Yue Shang Information Technology (Beijing) Co., Ltd. has generated \$21,700 in revenue as of March 31, 2020.
WeTrade Group Inc. is no longer a shell company as defined under 17 CFR §240.12b-2 and Rule 144(i)(1)(i) as it has commenced regular revenue generating operations and controls significant assets, the WePay System.

Section 7 Item 7.01 Regulation FD

To comply with Regulation FD, companies must disclose material events under this item. Regulation FD requires companies to provide material information to the public simultaneously with its disclosure to specific individuals or entities. If a company discloses certain information to

select institutional investors and financial analysts during an investor event, it can file a Form 8-K under this item to make the same information available to the public. Please refer to Figure 6 for an example of Item 7.01 filed by CVS Health.

Figure 6. Item 7.01, filed by CVS Health

Item 7.01	Regulation FD Disclosure
A press release related to the matters described in Item 2.01 of this Current Report on Form 8-K is included in Exhibit 99.1. The information in Exhibit 99.1 of this Current Report on Form 8-K is being furnished, not filed. Accordingly, the information in Exhibit 99.1 of this Current Report will not be incorporated by reference into any registration statement filed by CVS Health under the Securities Act of 1933, as amended, unless specifically identified therein as being incorporated by reference.	

Section 8 Item 8.01 Other Events

If a company believes an event is important but does not fall into any other categories, the company can disclose this event under Item 8.01. For example, in Figure 7, Facebook Inc. disclosed that its board of directors authorized a share repurchase program.

Figure 7. Item 8.01 filed by Facebook Inc.

Item 8.01 Other Events.
Facebook's board of directors has authorized a share repurchase program of its Class A common stock, which commenced in 2017 and does not have an expiration date. As of December 31, 2020, \$8.6 billion remained available and authorized for repurchases under the program. On January 27, 2021, Facebook announced an increase of \$25 billion in the amount authorized for repurchases under the program. The timing and actual number of shares repurchased under the program depend on a variety of factors, including price, general business and market conditions, and other investment opportunities. Shares may be repurchased through open market purchases or privately negotiated transactions, including through the use of trading plans intended to qualify under Rule 10b5-1 under the Exchange Act.

Section 9 Item 9.01 Financial Statements and Exhibits

Under this item, a company must file certain financial statements and list the exhibits it has filed. For example, if a company discloses in Item 2.01 that it has acquired a business, Item 9.01 would require the company to provide the financial statements of the business. In addition, the company must present “pro forma” financial statements that demonstrate what the company’s financial results might have been if the transaction had occurred earlier. Similarly, if the company discloses in Item 1.01 that it has entered into a material agreement, that agreement may be filed as an exhibit in the 8-K under Item 9.01.

5.2. Empirical example: Technological peer pressure and product disclosure



[Empirical example: Technological peer pressure and product disclosure](#)

In theory, companies decrease their level of disclosure in response to increased competition. However, empirical testing of the relationship is challenging due to the multifaceted nature of competition and disclosure. For example, Amazon competes with Walmart for customers and distribution channels but competes with Google in information technology; Intel competes with ARM in CPU architecture design but competes with Samsung in mobile CPU sales. Companies also provide a variety of types of disclosure, such as management earnings forecasts, segment reporting, information discussed in conference calls, risk-related disclosures, compensation-related disclosures, CSR-related disclosures, etc. The impact of competition on corporate disclosure may thus depend on the specific type of disclosure and its alignment with the competitive dimension.

Cao, Ma, Tucker, and Wan (2018) construct a measure of technological competition, “technological peer pressure” (TPP), which captures the aggregate technological advances of companies that compete with a given company in the product market relative to the given company’s own technological preparedness. This type of competition is aligned with product disclosure, which reveals where the firm invests in technology for product development and improvement and how these investments have progressed. Product disclosure is quantified with the number of words in product-disclosure press releases issued by a company. Using the two measures, Cao et al. (2018) find that TPP has a significantly negative and strong economic association with product disclosure: a company that moves from the lowest decile of TPP to the highest decile reduces its product disclosure by 44.7%. In contrast, they find that TPP is not

associated with the frequency of management earnings forecasts, which is a type of disclosure supposedly not aligned with technological competition.

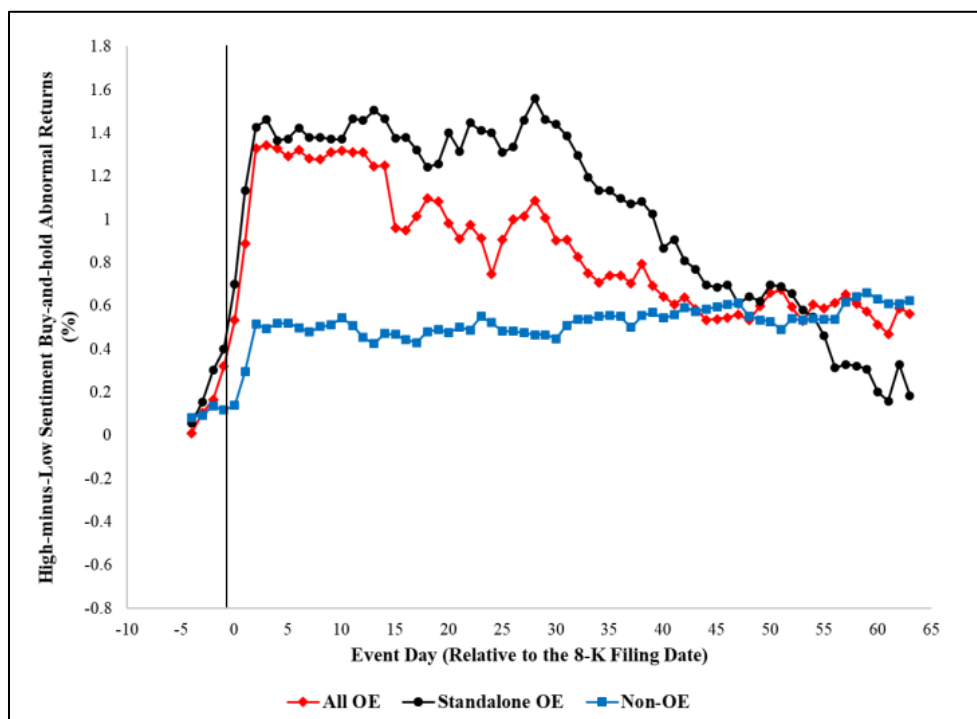
5.3. Empirical example: A game of disclosing “other events”



[Empirical example: A game of disclosing “other events”](#)

“Other Events” (OE) disclosure is a flexible and voluntary item in 8-K filings. Cao, Qin, and Shu (2023) document that firms tend to manipulate the sentiment in OE disclosures. This leads to a pronounced price response during the event window, followed by a subsequent reversal that exceeds the initial reaction. Consistent with sentiment distortion, OE sentiment negatively predicts firm performance. These patterns, interestingly, are not evident among non-OE disclosure in 8-K filings, which are subject to more stringent regulations. Figure 8 plots the variation in buy-and-hold abnormal returns (BHAR) between the high-sentiment (top 10 percent) and low-sentiment (bottom 10 percent) subgroups of OE disclosures. In both the All OE and Standalone OE samples, this differential BHAR experiences a sharp increase up to four days after disclosure, reaching approximately 1.5%, and then gradually reverts over the next three months. In contrast, the differential BHAR for non-OE sentiment displays a gradual upward drift after the initial spike without any significant reversal.

Further, the study reveals that the adverse repercussions predominantly affect retail investors rather than their sophisticated counterparts. Retail investors' actions related to information gathering and trading are responsible for the initial mispricing and also contribute to the delay in rectifying this mispricing. Notably, the manipulation of sentiment within OE disclosures seems to favor managers engaged in insider sales and option grants and companies involved in equity offerings and mergers and acquisitions.

Figure 8. Buy-and-hold abnormal returns around 8-K disclosures

This figure plots the high-minus-low sentiment buy-and-hold-abnormal returns (BHARs) starting 5 days before to 63 days after disclosures of 8-Ks. We compute industry- and past return-adjusted sentiment breakpoints by first sorting 8-Ks by its Fama and French 48 industry and then, within each industry, by the median pre-event abnormal return (PreFFAlpha). We then use sentiment's 10th and 90th percentiles as the breakpoints to identify low-sentiment and high-sentiment subsamples, respectively. The high-minus-low sentiment event BHAR is the high sentiment group's average BHAR minus the low sentiment group's average BHAR on each event day relative to the filing date. We plot average BHARs for the "All OE," "Standalone OE," and "Non-OE" samples. The sample period is from August 23, 2004, to October 23, 2020.

Source: Cao et al. (2023)

References

- Cao, S., Ma, G., Tucker, J., and Wan, C. 2018. Technological peer pressure and product disclosure. 2018. *The Accounting Review*, 93(6), 95-126.
- Cao, S., Qin, Z., and Shu, T. 2023. A game of disclosing “other events”: a message to retail investors. Working Paper

Chapter 6 Analyzing Data from Social Media

6.1. What is social media?



[What is social media?](#)

Social media has become an incredibly influential aspect of modern-day disclosure. It encompasses any digital technology that facilitates the sharing of ideas, thoughts, and information through virtual networks and communities. Anyone with an internet connection can make a social media profile and can use that profile to post nearly any content they like. Hence, personalized profiles and user-generated content are defining features of social media platforms.

Originating in the late 1970s, social media was initially designed as a platform for people to interact with friends, family, and shared interest communities. Nowadays, social media has evolved into a multifaceted tool, serving as a platform for people to discover career opportunities, make romantic connections, and share their insights and perspectives online. In addition, businesses use social media for advertising, customer communication, and increasing brand awareness. As of October 2021, more than 4.5 billion people worldwide use social media. Its ability to instantly post photographs, share viewpoints, and record events has revolutionized how people live and conduct business.

Various types of social media platforms offer a variety of services. Social networks, such as Facebook and Snapchat, allow users to share ideas, opinions, and content, and hence, most content on these platforms consists of text, images, or a combination of the two. Media networks like YouTube and TikTok facilitate the sharing of media assets, including images, videos, and other content. Review networks like Yelp are designed for evaluating products and services. Discussion networks such as Reddit and Quora provide a forum for people to discuss problems, ask questions, and debate issues. Finally, business platforms like LinkedIn, Glassdoor, and Blind cater to

professionals, fostering networking and collaboration with other professionals or potential clients.

Table 1 lists the most popular social media platforms categorized by their primary functions.

Table 1. Popular social media platforms

Type	Data	Social Media Platforms	Purpose
Social network	Textual or image	Snapchat, Twitter, Facebook, WeChat	Send messages privately or publish at-the-moment content
	Audio	Clubhouse, Spotify	Listen to live conversations on specific topics
Media network	Image	Pinterest, Instagram	Send short messages privately and publish conveniently, at-the-moment content
	Video	YouTube, TikTok, Twitch	Broadcast live video to viewers
Discussion network	Textual	Reddit, Quora	Debate and discuss, network, form communities around a subject, and share views on internet-driven topics
Business Platforms	Textual/Image	LinkedIn, Glassdoor, Blind	Collaborate with professionals or with potential clients

6.2. Data from Social Media



[Data from social media platforms](#)

Businesses have found a wide array of uses for social media platforms. Many social media users are aware that these platforms collect user data for personalized advertisements, and retailers and artisans use these platforms to market products globally. However, there are other less visible ways in which capital market participants, such as financial analysts and investors, use social media. Leveraging AI and machine learning technologies, they explore new ways to transform the vast amounts of data generated by social media users into valuable financial insights. At the same time, companies are capitalizing on social media as an alternative channel for disclosing information and communicating with the market, benefiting from comparatively lower oversight. This trend has garnered significant attention among finance scholars. In this section, we will examine how specific platforms are being used by FinTech researchers in innovative ways.

6.2.1. Twitter

Launched in 2006, Twitter (now called X) is a social media “microblogging” platform where users share and engage with messages, media, and images contained in “tweets.” Initially capped at 140 characters, the tweet length was extended to 280 in November 2017. Figure 1 shows a tweet from Microsoft on October 14, 2018. The tweet summarizes Microsoft’s first-quarter financial performance, including revenue, income, and earnings per share. The tweet has attracted considerable attention from Twitter users, evidenced by 594 “Retweets,” 116 “Quotes,” 1,760 “Likes,” and seven “Bookmarks.”

Using Twitter to predict firm performance

Twitter users have produced vast amounts of information in their tweets, leading researchers to ask whether this user-generated content has informational value in business and finance. Several studies have found that, on an aggregate level, certain types of tweets have predictive capabilities. In one such study, Bartov, Faurel, and Mohanram (2018) find that opinions expressed in individual tweets predict a firm's upcoming quarterly earnings and announcement returns. Relatedly, Tang (2018) demonstrates that aggregated third-party-generated product information on Twitter can predict firm-level sales.

Figure 1. An earnings announcement tweet by Microsoft



Using Twitter for public relations and information management

In recent years, companies have increasingly used Twitter as a tool for strategic purposes, a phenomenon that has sparked scholarly interest. For example, Blankespoor, Miller, and White (2014) examine whether firms could effectively reach more investors and thus reduce information asymmetry by sharing news through tweets. Using a sample of technology firms, they find that tweeting links to press releases helps spread earnings information among investors, which reduces information asymmetry. Jung, Naughton, Tahoun, and Wang (2018) investigate the strategic dissemination of news by companies on Twitter. They find that companies are inclined to tweet negative earnings news compared to positive earnings news.

At the same time, the autonomous nature of social media imposes challenges for companies because they have limited control over the information or opinions shared on these social media platforms. Lee, Hutton, and Shu (2015) find that a corporation's use of social media can help counterbalance negative price reactions to recall announcements. However, with the arrival of Facebook and Twitter, firms cede some control over their social media content, and the attenuation benefits of corporate social media lessen.

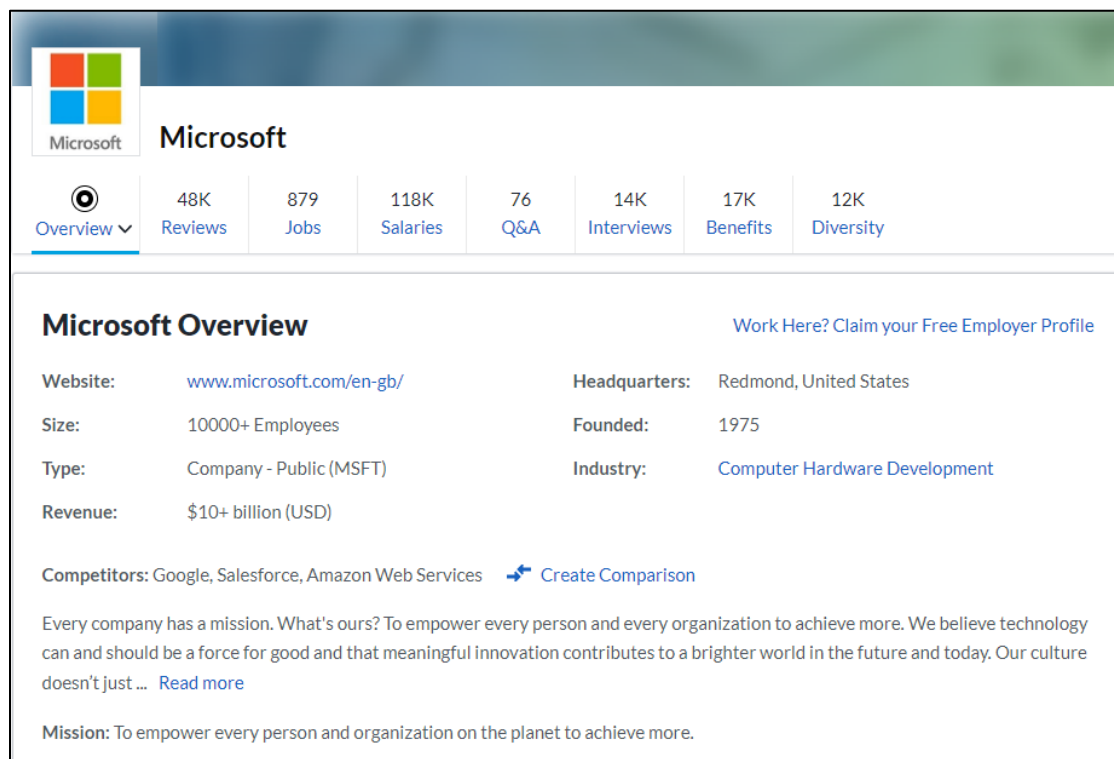
6.2.2. Glassdoor.com

Glassdoor.com is a large recruiting platform where users can explore and apply for jobs. Beyond job postings, Glassdoor also hosts a social media platform where present and former employees can anonymously review companies across various criteria, including internal CEO approval ratings, salary details, interview difficulty and questions, compensation and benefits assessments, overall company outlook, etc. Glassdoor thus provides firsthand information from employees who voluntarily express their opinions about the companies they have been associated with. Figure 2 shows the company profile page of Microsoft on Glassdoor.com.

Using employee reviews to predict firm outcomes

Since Glassdoor.com opens a “window” through which outsiders can easily access comments and opinions from a company’s employees, does that mean we can use it to glean valuable information incremental to what is disclosed by company management? Using data from Glassdoor, Hale, Moon, and Sweson (2018) find that employee opinions are useful in predicting earnings growth and management forecast news. Huang, Li, and Markov (2020) document that the average employee outlook offers incremental information in predicting future operating performance, particularly when aggregated from a larger, more diverse, and more knowledgeable employee base. Interestingly, the average outlook predicts bad news events more strongly than good news events.

Figure 2. Glassdoor company profile of Microsoft



In addition to its predictive power, Dube and Zhu (2021) show that employee opinions on Glassdoor.com prompt companies to improve their workplace practices, particularly in areas like employee relations and diversity. Such improvement is particularly notable for firms initially receiving negative reviews and those with high labor intensity. Green, Huang, Wen, and Zhou (2019) find that companies experiencing improvements in employee opinions significantly outperform firms with declines. The return effect is concentrated among reviews from current employees.

6.2.3. Stock Message Boards

Message boards have been integral to digital communication since the introduction of USENET in 1979. These platforms primarily serve as forums where users can express their thoughts, engage with others who share similar interests or tap into the expertise of individuals in specific fields. Stock message boards allow investors to connect with other investors of varying expertise levels and learn more about profitable investing strategies. Many stock message boards center around specific topics, such as options trading, precious metals, exchange-traded funds (ETFs), or commodities. Figure 3 shows the stock message board Seeking Alpha, where discussions span diverse subjects like basic materials, bonds, closed-end funds, commodities, cryptocurrency, and more. Seeking Alpha users can actively share and exchange opinions by posting or commenting on analysis articles.

Using online “crowd wisdom” to predict stock performance

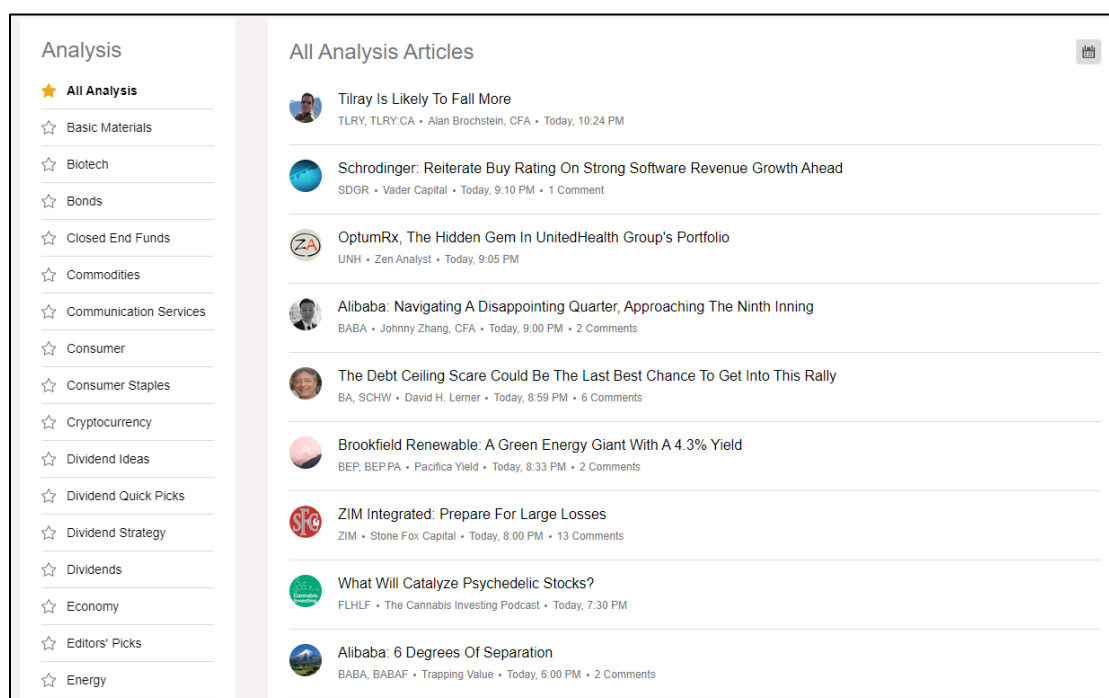
Stock message boards provide investors with a platform to exchange ideas and opinions, potentially generating “the wisdom of crowds”—the idea that the collective perspective of a large group of people is sometimes more accurate than that of a single expert. If proven true, the “crowd wisdom” derived from stock message boards could aid investors in making more informed

predictions about a company's future performance. Drake, Moon, Twedt, and Warren (2022) find that market reaction to sell-side analyst research is substantially diminished when the analyst research is preceded by reports from “social media analysts” (SMAs)—individuals posting equity research online via social media investment platforms and that this is particularly true of sell-side analysts’ earnings forecasts.

Credibility of online contributors

As in many social media contexts, concerns about accuracy and credibility on stock message boards have prompted scrutiny regarding the legitimacy of content producers. To shed light on this issue, Campbell, DeAngelis, and Moon (2019) investigate whether stock-holding positions by SMAs have a negative effect on analyst objectivity. They find no evidence that a SMA’s position reduces investor responses to the posts. In fact, they show that holding a position magnifies investor responses to the SMA’s articles. Their findings suggest that SMAs’ stock positions do not decrease the credibility and informativeness of their analyses.

Figure 3. Stock analyses on Seeking Alpha



6.2.4. YouTube

Established in 2005, YouTube is a prominent social media platform enabling users to upload, share, store, and watch videos. It is immensely popular, with a reported 2.5 billion monthly users as of June 2022, making it the second most visited website globally. The platform garners over a billion hours of video views each day, attesting to its widespread influence.

Using YouTube to extract multi-dimensional information

Given the vast content repository on YouTube, researchers are understandably intrigued by its potential insights. However, processing video content requires powerful technological tools. Hu and Ma (2022) collect startup self-introductory pitch videos from YouTube and another video-sharing website, Vimeo. Using machine learning algorithms to process these pitch videos, they are able to measure the persuasiveness of delivery in start-up pitches across visual, vocal, and verbal dimensions. They find that passionate and warm pitches increase funding probability; however, conditional on funding outcomes, those with excessively positive pitches tend to underperform.

6.2.5. LinkedIn

Connections and interactions on certain social media platforms are likely to be extensions of real-world relations. While you probably don't know everyone on WallStreetBets, there's a good chance that your Facebook friends are people you have met. Networking-based social media platforms thus allow researchers to deduce real-world connections from relations on social media platforms.

LinkedIn is a social network tailored for career and business professionals seeking connections. LinkedIn users create their professional profiles by sharing their educational backgrounds, employment histories, skills, and career interests. LinkedIn is geared toward building strategic

relationships, unlike other social networks where connections might be formed with a broad spectrum of individuals.

Using LinkedIn to uncover personal ties among financial professionals

Of all the networks on LinkedIn, researchers show a particular interest in connections between financial analysts, fund managers, and corporate executives. Jiang, Wang, and Wang (2018) use professional profiles on LinkedIn to identify revolving rating analysts with structured finance rating experience. They find that as companies issuing debt securities enlist such analysts, there is a higher likelihood that the ratings of their debt securities become inflated compared to similarly rated securities without such analyst involvement.

Bradley, Gokkaya, and Liu (2020) examine professional connections among executives and analysts formed through overlapping historical employment. They search for each analyst on LinkedIn.com to capture pre-analyst employment history. They find that analysts with professional connections to coverage firms have more accurate earnings forecasts and issue more informative buy and sell recommendations.

Using LinkedIn to extract facial information of financial professionals

In addition to employment histories, many professionals post their photos on social media networks. Li, Lin, Lu, and Venstra (2020) gather analysts' photos from their LinkedIn profiles and explore the connection between physical appearance and professional recognition. They find that, while female analysts are more likely to be voted All-Star analysts in the United States, good-looking female analysts are paradoxically less likely to be voted All-Stars. On the contrary, female analysts in China are less likely to be voted as All-Stars, but the likelihood increases with their facial attractiveness. These findings implicate a beauty penalty for female analysts in the United States and gender discrimination against female analysts in China.

6.3. Empirical Example: Negative Peer Disclosure



[Empirical example: Negative peer disclosure](#)

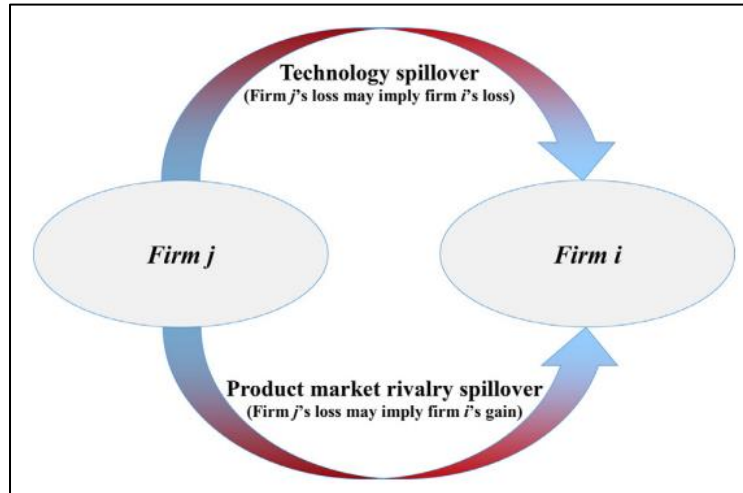
Most corporate social media posts tend to fall into common categories such as product announcements, earnings disclosures, industry awards, community engagement, etc. Cao, Fang, and Lei (2021) uncover an emerging —and entirely different— type of corporate social media posts: negative peer disclosure (NPD). NPD refers to a phenomenon that firm A discloses negative information about competitor firm B without mentioning anything about itself. Here is an example of what happened between Dropbox/Box and Globalscape, two companies that compete in the online file storage space. In 2014, news broke of a Dropbox security flaw that exposed its users' private data. Globalscape responded by retweeting a news article with this headline: “Dropbox and Box Leak Files in Security Through Obscurity Nightmare.”

When the negative news came out about Dropbox and Box, Globalscape could have been affected in two ways. On the one hand, it could have been positive in terms of product market competition since Globalscape didn't have any security breakdowns as its competitor did. At the same time, it could also have been negative from the technology spillover perspective—the market might have assumed Globalscape was subject to the same technological vulnerability (Figure 4). Hence, the NPD tweet signaled to the market that what happened to Dropbox doesn't apply to Globalscape.

To build a dataset of NPDs, Cao, Fang, and Lei (2021) collect tweets that mention a competitor from corporate Twitter accounts and employ sentiment analysis to identify negative peer disclosures. The study find that NPDs are issued by well-known and successful companies such as Nvidia, T-Mobile, Symantec, and others. The propensity of NPD increases with the degree of product market rivalry and technological proximity. The approach appears to work. Consistent

with NPDs being implicit positive self-disclosures, disclosing firms experience a two-day abnormal return of 1.6–1.7% over the market and industry. Firms using NPDs tend to outperform their non-NPD-using peers in the product markets.

Figure 4. The implications of negative peer disclosure



Source: Cao et al. (2021)

References

- Bartov, E., Faurel, L., and Mohanram, P. 2018. Can Twitter help predict firm-level earnings and stock returns? *The Accounting Review*, 93, 25-57.
- Blankespoor, E., Miller, G., and White, H. 2014. The role of dissemination in market liquidity: Evidence from firms' use of Twitter. *The Accounting Review*, 89(1), 79-112.
- Bradley, D., Gokkaya, S., Liu, X., and Xie, F. 2017. Are all analysts created equal? Industry expertise and monitoring effectiveness of financial analysts. *Journal of Accounting and Economics*, 63(2-3), 179-206.
- Campbell, J., DeAgelis, M., and Moon, J. 2019. Skin in the game: personal stock holdings and investors' response to stock analysis on social media. *Review of Accounting Studies*, 24, 731-779.
- Cao, S., Fang, V., and Lei, L. 2021. Negative peer disclosure. *Journal of Financial Economics*, 140(3), 815-837.
- Drake, M., Moon, J., Twedt, B., and Warren, J. 2023. Social media analysts and sell-side analyst research. *Review of Accounting Studies*, 28, 385-420.
- Dube, S., and Zhu, C. 2021. The disciplinary effect of social media: Evidence from firms' responses to Glassdoor Reviews. *Journal of Accounting Research*, 59(5), 1783-1825.
- Green, T., Huang, R., Wen, Q., and Zhou, D. 2019. Crowdsourced employer reviews and stock returns. *Journal of Financial Economics*, 134(1), 236-251.
- Hales, J., Moon, J., and Swenson, L. 2018. A new era of voluntary disclosure? Empirical evidence on how employee postings on social medial relate to future corporate disclosures. *Accounting, Organizations and Society*, 68-69, 88-108.
- Hu, A., and Ma, S. 2021. Persuading investors: A video-based study. Working paper.

- Huang, K., Li, M., and Markov, S. 2020. What do employees know? Evidence from a social media platform. *The Accounting Review*, 95(2), 199-226.
- Jiang, J., Wang, I., and Wang, K. 2018. Revolving rating analysts and ratings of mortgage-backed and asset-backed securities: Evidence from LinkedIn. *Management Science*, 64(12), 5461-5959.
- Jung, M., Naughton, J., Tahoun, A., and Wang C. 2018. Do firms strategically disseminate? Evidence from corporate use of social media. *The Accounting Review*, 93(4), 225-252.
- Lee, L., Hutton, A., and Shu, S. 2015. The role of social media in the capital market: Evidence from consumer product recalls. *Journal of Accounting Research*, 53(2), 367-404.
- Li, C., Lin, A., Lu, H., and Veenstra, K. 2020. Gender and beauty in the financial analyst profession: evidence from the United States and China. *Review of Accounting Studies*, 25, 1230-1262.
- Tang, V. (2018). Wisdom of Crowds: Cross-sectional variation in the informativeness of third-party-generated product information on Twitter. *Journal of Accounting Research*, 56(3), 989-1034.

Chapter 7 Data Analytics in Environmental, Social, and Governance

7.1. Corporate Governance



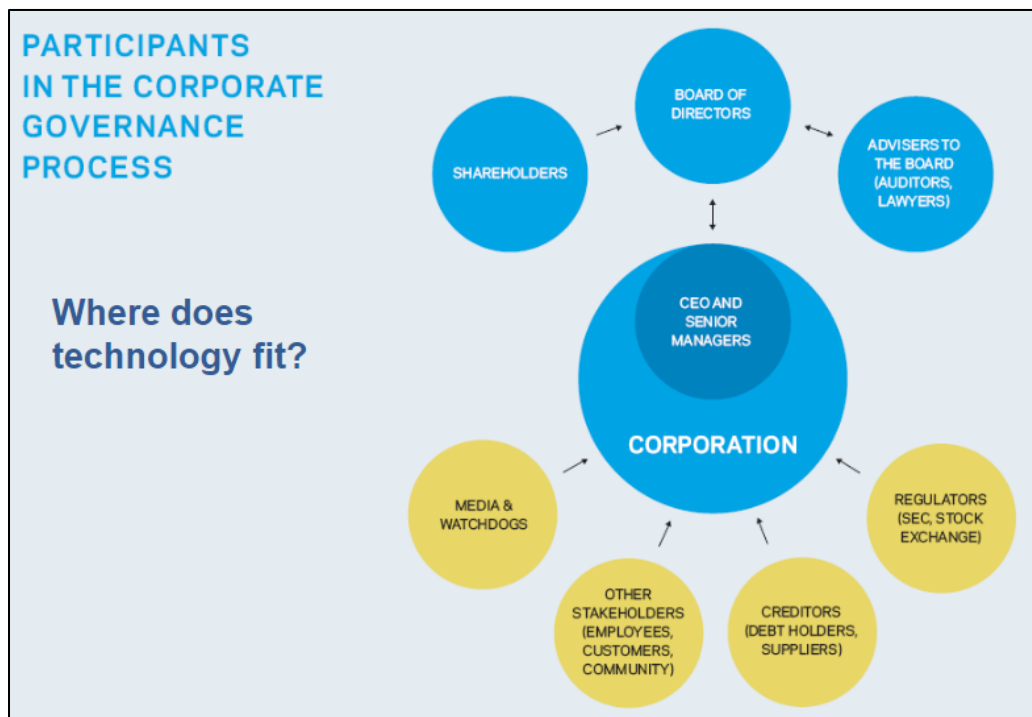
[Corporate governance: Data and technology](#)

Corporate governance refers to the set of processes, customs, policies, laws, and institutions that impact how a corporation is directed, administered, or controlled. The need for corporate governance arises from the separation of ownership and control and the presence of information asymmetry. In the era preceding the Industrial Revolution, businesses were typically owned and managed by a small group of individuals known as sole proprietors. This simple ownership structure eventually led to the more complicated corporate form, wherein individuals invest in a corporation, and managers, chosen by those investors, assume managerial responsibilities. This evolution introduces agency relationships where corporate managers serve as agents controlling corporate resources, while investors function as principals who are not directly involved in daily corporate operations.

Separation of ownership and control inevitably raises concerns about whether the agent is acting in the principal's best interests. It also allows managers to acquire private information by exercising their control. Hence, the separation of ownership and control leads to a divergence of ownership and information, granting managers an informational advantage over investors. Following the paradigm outlined by Myers and Majluf (1984), information asymmetry occurs "...when firms have information that investors do not have." To address agency and information asymmetry concerns, the traditional corporate governance system involves shareholders appointing boards of directors to oversee senior managers. These boards may seek advice from various parties, such as auditors and legal counsels. Figure 1 describes this corporate governance environment.

Traditionally, the primary focus of corporate governance has been to maximize shareholder value and safeguard shareholders' interests. However, in recent years, there has been a growing emphasis on environmental, social, and governance (ESG), broadening corporate governance's scope. Today's corporate governance system is expected to ensure the interests of all the firm's stakeholders. In addition to shareholders, corporate managers are now accountable to multiple external monitors and stakeholders, including creditors, regulators, employees, customers, the community, etc.

Figure 1. Participants in the corporate governance system



7.2. Textual data for corporate governance



[Textual data for corporate governance](#)



[Environmental, social, and governance \(ESG\) disclosures](#)



[Analytics of Regulators' comments and IPO](#)

7.2.1. Proxy statement

When seeking information about a company, individuals typically turn to financial statements, annual reports, and conference calls as primary sources. However, the proxy statement can be just as informative, if not more so, as it provides in-depth insights into corporate officers' business relationships, professional backgrounds, and compensation details.

A proxy statement is a document issued by public companies to their shareholders, guiding them on voting procedures at shareholder meetings and aiding them in making informed decisions on how to delegate their votes to a proxy. It covers various issues, such as proposals for new board members, details on directors' compensation, bonus and options plans for directors, corporate actions like proposed mergers or acquisitions, dividend payouts, and any other declarations put forth by the company's management. Below is some of the information you can glean from this important document.

- Important voting issues. Figure 2 shows the business items with board voting recommendations in Apple's Notice of 2022 Annual Meeting of Shareholders. Figure 3 lists the items of business, board voting recommendations, and voting instructions in Walmart's 2022 proxy statement. Typical items for shareholder voting at annual meetings include electing directors, ratifying the appointment of an independent registered public accounting firm, approving an advisory vote on executive compensation, approval of employee stock plans, and considering shareholder proposals.

- Details about management, their experience, and qualifications.
- Management compensation and whether their compensation structure is aligned with shareholder interests. Figure 4 presents the summary compensation table from Apple Inc.'s 2022 proxy statement. It provides a comprehensive breakdown of executive compensation for the top five executives in the past three years. Additionally, Figure 5 outlines Apple's analysis of the CEO's Restricted Stock Units (RSU) vesting. The analysis provides insights into the conditions required for RSU vesting, the corresponding performance outcomes, and the resulting vesting outcomes.
- Potential conflicts of interest, such as related-party transactions, may not benefit the company.
- Loans advanced to senior executives. These loans can deprive the company of capital, are often made on generous terms, and sometimes are forgiven, leaving shareholders to foot the bill.

Figure 2. Apple's Notice of 2022 Annual Meeting of Shareholders

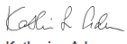
Date and Time: March 4, 2022 9:00 A.M. Pacific Time		Virtual Meeting Site: www.virtualshareholdermeeting.com/AAPL2022
		Who Can Vote: Shareholders of record at the close of business on January 3, 2022
Items of Business and Board Voting Recommendation		
1	Election of Directors: James Bell, Tim Cook, Al Gore, Alex Gorsky, Andrea Jung, Art Levinson, Monica Lozano, Ron Sugar, and Sue Wagner	FOR each of the nominees
2	Ratification of Appointment of Independent Registered Public Accounting Firm	FOR
3	Advisory Vote to Approve Executive Compensation	FOR
4	Approval of the Apple Inc. 2022 Employee Stock Plan	FOR
5-10	Shareholder Proposals if properly presented	AGAINST
And such other business as may properly come before the Annual Meeting and any postponements or adjournments thereof.		
Sincerely,		
 Katherine Adams Senior Vice President, General Counsel and Secretary Cupertino, California January 6, 2022		

Figure 3. Walmart's 2022 proxy statement

Items of Business

1

To elect as directors the 11 nominees identified in this proxy statement.

(PAGE 8) →

Vote **FOR**

2

To vote on a non-binding, advisory resolution to approve the compensation of Walmart's named executive officers.

(PAGE 40) →

Vote **FOR**

3

To ratify the appointment of Ernst & Young LLP as the company's independent accountants for the fiscal year ending January 31, 2023.

(PAGE 77) →

Vote **FOR**

4-10


To vote on the 7 shareholder proposals described in the accompanying proxy statement, if properly presented at the meeting.

(PAGE 82) →


Vote **AGAINST** each Shareholder Proposal

Shareholders may also transact any other business properly brought before the 2022 Annual Shareholders' Meeting or any adjournment or postponement thereof.


How to Cast Your Vote → (PAGE 104)




INTERNET (BEFORE THE MEETING)
www.proxyvote.com




CALL
1-800-690-6903



MOBILE DEVICE
Scan the QR code on your proxy card, notice of internet availability of proxy materials, or voting instruction form




MAIL
Mail your signed proxy card or voting instruction form



DURING THE VIRTUAL MEETING
Please see pages 103-107 for details about how to attend and vote your Shares during the virtual meeting.

April 21, 2022
By Order of the Board of Directors,



Rachel Brand
Executive Vice President, Global Governance, Chief Legal Officer, and Corporate Secretary

Figure 4. Summary Compensation Table in Apple Inc's 2022 proxy statement

Name and Principal Position	Year	Salary (\$)	Bonus (\$)	Stock Awards ⁽¹⁾ (\$)	Non-Equity Incentive Plan Compensation ⁽²⁾ (\$)	All Other Compensation (\$)	Total (\$)
Tim Cook Chief Executive Officer	2021	3,000,000	—	82,347,835	12,000,000	1,386,559 ⁽³⁾	98,734,394
	2020	3,000,000	—	0	10,731,000	1,038,259	14,769,259
	2019	3,000,000	—	0	7,671,000	884,466	11,555,466
Luca Maestri Senior Vice President, Chief Financial Officer	2021	1,000,000	—	21,959,620	4,000,000	18,883 ⁽⁴⁾	26,978,503
	2020	1,000,000	—	21,657,687	3,577,000	18,583	26,253,270
	2019	1,000,000	—	21,633,416	2,557,000	19,221	25,209,637
Kate Adams Senior Vice President, General Counsel and Secretary	2021	1,000,000	—	21,959,620	4,000,000	14,533 ⁽⁵⁾	26,974,153
	2020	1,000,000	—	21,657,687	3,577,000	14,310	26,248,995
	2019	1,000,000	—	21,633,416	2,557,000	41,384	25,231,800
Deirdre O'Brien Senior Vice President, Retail + People	2021	1,000,000	—	21,959,620	4,000,000	61,191 ⁽⁶⁾	27,020,811
	2020	1,000,000	—	21,657,687	3,577,000	37,684	26,272,371
	2019	877,500	—	16,469,527	1,795,000	17,753	19,159,780
Jeff Williams Chief Operating Officer	2021	1,000,000	—	21,959,620	4,000,000	17,437 ⁽⁷⁾	26,977,057
	2020	1,000,000	—	21,657,687	3,577,000	17,137	26,251,824
	2019	1,000,000	—	21,633,416	2,557,000	17,503	25,207,919

Shareholder voting on important corporate issues can escalate into proxy contests, commonly referred to as proxy battles. This occurs when a coalition of shareholders joins forces to challenge and oust the existing management or board of directors, essentially creating a battle for control of the company between shareholders and senior management. Proxy fights (Figure 6) are typically triggered by disgruntled shareholders who unite with others to exert pressure on management and the board of directors for necessary changes within the company. Shareholders express their dissatisfaction and assert their influence over the board by casting votes against them during the annual general meeting (AGM).

Figure 5. CEO RSU vesting analysis in Apple Inc's 2022 proxy statement

2021 RSU Results

CEO RSU Vesting

This year marked the 10-year anniversary of Mr. Cook's tenure as Apple's CEO and the vesting of the final tranche of the long-term RSU award granted upon his promotion to CEO in 2011 (the "2011 RSU Award"). Mr. Cook earned significant long-term incentives over the course of a decade, aligning his compensation with the extraordinary value created for our shareholders under his leadership.



At grant, the 2011 RSU Award had a time-based vesting schedule with 50% of the RSUs vesting on each of the 5- and 10-year anniversaries of the grant date. The 2011 RSU Award was significantly modified in 2013, at Mr. Cook's request, to put a portion of the award at risk by applying a performance condition based on Apple's Relative TSR performance over specified performance periods. The 2011 RSU Award was not modified to include dividend equivalents.

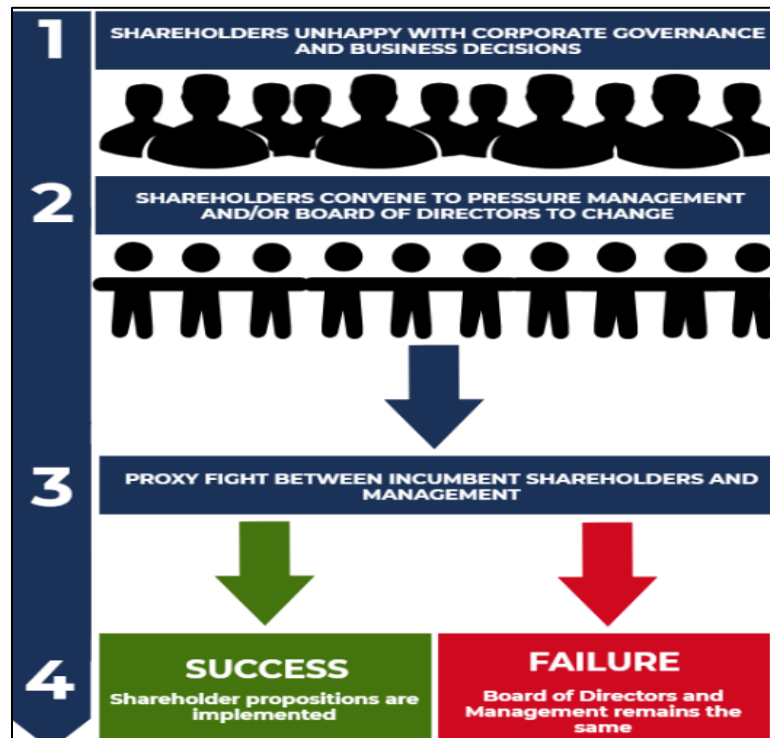
The performance condition for Mr. Cook's 2011 RSU Award required Apple to outperform two-thirds of the companies that were included in the S&P 500 for the entirety of each performance period in order for 100% of the performance-based RSUs allocated to that period to vest. The 2011 RSU Award only had downside risk to Mr. Cook. It did not contain any upside vesting opportunity above 100% of the target number of RSUs, and there was no interpolation for results between the Relative TSR levels set at the bottom, middle, and top third of companies in the S&P 500.

Relative TSR Percentile v. S&P 500 Companies	Performance-Based RSUs Vesting
Top Third	100%
Middle Third	50%
Bottom Third	0%

For the final three-year performance period under the 2011 RSU Award from August 25, 2018 through August 24, 2021, Apple's Relative TSR was at the 97th percentile of the companies that were included in the S&P 500 for the entire performance period. As a result, 100% of the target 1,120,000 performance-based RSUs for this performance period vested on August 24, 2021. Apple's total shareholder return during this performance period was 191.83%. On August 24, 2021, Mr. Cook also vested in the remaining 3,920,000 time-based RSUs under his 2011 RSU Award.

	Relative TSR Percentile Ranking for Three-Year Performance Period	TSR Results for Three-Year Performance Period
Apple	97th Percentile	191.83%
S&P 500 Companies	Top Third	≥72.26%
	Middle Third	24.94-72.25%
	Bottom Third	<24.94%

Figure 6. Proxy voting



7.2.2. Environmental, social, and governance (ESG) disclosures

Environmental, social, and governance (ESG) is an integral part of corporate governance designed to guarantee a company's operations are ethical and beneficial to all stakeholders in the broader society. Good corporate governance could be a prerequisite for making ethical and sustainable decisions on environmental and social issues (Starks, 2023). Although the interpretation of ESG is expansive and may differ across companies, the core principle is to conduct business in a sustainable manner that is economically, socially, and environmentally responsible.

Given the recent rapid development in the ESG social and regulatory landscape, ESG information is crucial in helping stakeholders assess firms' risks and opportunities. For instance, studies find that investors demand return premiums for pollution and carbon footprint due to

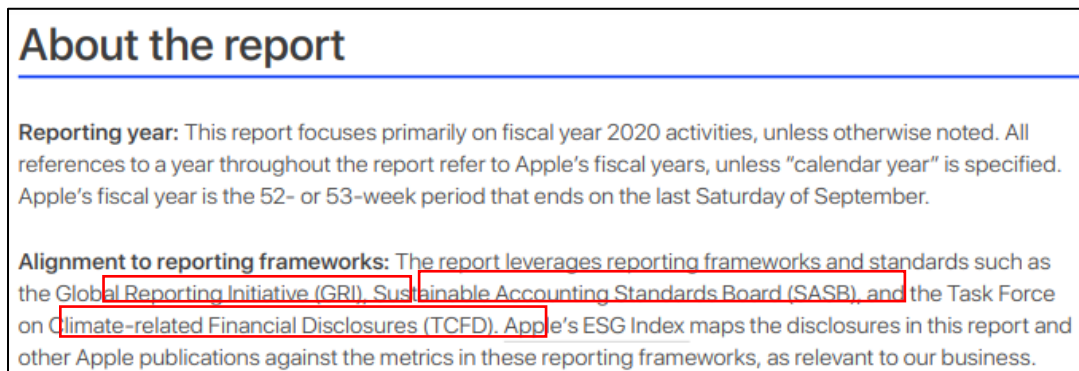
climate concerns and environmental policy uncertainty (Bolton and Kacperczyk 2021; Ilhan, Sautner, and Vilkov 2021; Pastor, Stambaugh, and Taylor 2022). ESG information could also be useful for ESG-related investment decisions such as financing of public and private investment supporting mitigation of and adaptation to climate change (Hong, Karolyi, and Scheinkman 2020) or mobilizing investment to maintain ecosystem integrity and biodiversity (Karolyi and Puente 2023). Indeed, institutional investors believe that ESG-related risks have financial implications for their investments (Krueger, Sautner, and Starks 2020).

Information regarding ESG can be obtained from both internal and external outlets. Sustainability reports serve as a primary source of ESG information furnished by companies. These reports contain information about the ESG impacts of a company's operations. There is a growing demand from investors and other stakeholders for enhanced transparency in companies' sustainability and ESG approaches. Additionally, many legislative documents either mandate or are expected to mandate the disclosure of non-financial ESG information.

The Global Reporting Initiative (GRI) is an international organization that establishes independent standards for companies to report non-financial information. These standards are designed to help businesses identify their impacts on climate change, the environment, human rights, and corporate governance. Although the GRI standards are non-mandatory and non-binding, they serve as the foundation for the proposed Corporate Sustainability Reporting Directive (CSRD). The forthcoming mandatory European Sustainability Reporting Standards (ESRS) are also based on the GRI structure. Like the International Financial Reporting Standards (IFRS), ESRS is a set of standards companies must comply with when reporting sustainability information. The issuance of sustainability reports is an effective way for companies to answer a wide variety of stakeholders' questions in a single document. In addition to the GRI standards, the

Sustainability Accounting Standards Board (“SASB”) also offers guidance on preparing informative ESG information. Other ESG reporting frameworks include those proposed by the Task Force on Climate-related Financial Disclosures (TCFD) and the United Nations Sustainable Development Goals. Figure 7 presents the reporting frameworks adopted in Apple’s 2021 ESG report.

Figure 7. Reporting Frameworks of Apple’s 2021 ESG Report



In addition to sustainability reports, companies also frequently disclose ESG information on their websites, in proxy statements, and on social media platforms. ESG information can also be collected from alternative sources external to companies, such as government agencies or non-governmental watchdog organizations.

7.3. Emerging technologies as governance mechanisms



[Emerging technologies as governance mechanisms](#)

Emerging technologies have profound impact on corporate governance (Jiang and Li 2024). On the one hand, revolutionized how information is accessed, processed, and used by shareholders and other stakeholders. The ability to access corporate information in alternative forms and sophisticated data analysis tools could create a new form of information asymmetry.

On the other hand, new technologies, such as blockchain, has the potential to improve corporate governance processes.

7.3.1. Governance with the availability of alternative data

Regulators have conventionally focused on ensuring equal access to information generated within firms. However, the advent of big data and data analytics means that a large amount of information is generated outside firms by tracking “footprints,” including satellite images, internet traffic, credit card scans, sensors, social media posts, etc. This alternative information could be ahead of or incremental to managerial information. Zhu (2019) shows that externally generated alternative data predict firm performance and mitigate insider advantages. The growing power of data, providing insights into ownership structures, leadership quality, shareholder sentiment, governance risks, and more, holds significant promise for governance. In addition, there is an increasing demand and supply of information about environmental, social, and governance issues, which would help hold managers accountable along these important dimensions.

Access to alternative information is widespread but unevenly distributed based on the skills and resources available. The rise of big data creates information asymmetry even in accessing traditional data. For example, the Securities and Exchange Commission (SEC) estimates that “as much as 85% of the documents visited are by internet bots”. The capability to process big data has become increasingly critical for establishing an informational advantage. Cao, Jiang, Wang, and Yang (2021) find that when alternative data becomes available, analysts affiliated with brokerage firms equipped with AI capacity provide more accurate forecasts. Additionally, they document that increases in machine downloads of SEC filings are associated with decreases in time to the first trade; however, such increases also widen bid-ask spreads.

On the flip side, firms are adapting to the increasing use of machines in processing corporate information. As noted earlier, Cao, Jiang, Yang, and Zhang (2023) find that the publication of

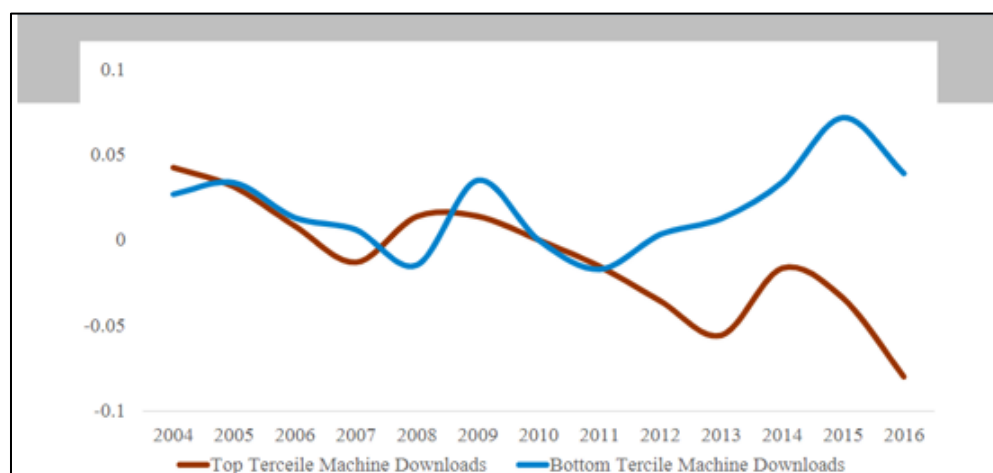
Loughran and McDonald (2011) prompts firms to reduce the use of negative words from Loughran and McDonald (2011) in corporate filings (Figure 8). Since machine processing of corporate filings is largely rule-based, the change in managerial behavior would impact the effectiveness of machine learning in corporate governance. Therefore, it's crucial to recognize that, given the influence of big data and data analytics, managers are incentivized to modify their behavior to shape and manipulate the outcomes of machine processing.

7.3.2. Governance with distributive ledger and blockchains: Shareholder voting and smart contracting

Shareholding voting empowers shareholders to directly participate in corporate governance. A proxy contest is a campaign to solicit votes (or proxies) in opposition to management at an annual or special meeting of stockholders or through action by written consent. Presently, the most prevalent forms of proxy contests involve activist stockholders vying for board representation or control, typically intending to optimize returns on the activist's investment in the short term. Proxy contests serve as a tool to instigate change. Brav, Jiang, Li, and Pennington (2021) document that approximately one percent of firms become targets of proxy contests in any given year.

The importance of shareholder voting indicates the pivotal role of shareholding records in corporate governance. In this context, blockchains offer a solution by ensuring transparent shareholding records and resolving the problem of “double voting” (Yermack 2017). Blockchains can also be implemented to add new features to the existing shareholding voting system. For example, tenure-based voting, a system that awards greater voting power to shares held for a longer duration (Edelman, Jiang, and Thomas 2019), and voting power tied to the performance of outside shares. Furthermore, the implementation of decentralized autonomous organizations can act as self-sufficient proxy advisories that empower retail investors in the decision-making process.

Figure 8. Frequency of Loughran and McDonald (2011) negative words in 10-K and 10-Q filings



This figure plots LM – Harvard Sentiment of 10-K and 10-Q filings and compares the sentiment of firms with high machine downloads with that of the low group. LM – Harvard Sentiment is the difference between LM Sentiment and Harvard Sentiment. LM Sentiment is defined as the number of Loughran-McDonald (LM) finance-related negative words in a filing divided by the total number of words in the filing. Harvard Sentiment is defined as the number of Harvard General Inquirer negative words in a filing divided by the total number of words in the filing. Filings are sorted into top tercile or bottom tercile based on Machine Downloads. LM Sentiment and Harvard Sentiment sentiments are normalized to one, respectively, in 2010 within each group, one year before the publication of Loughran and McDonald (2011). The dotted lines represent the 95% confidence limits.

Source: Cao, Jiang, Yang, and Zhang (2023)

In addition to improving voting, blockchain technology enables “smart contracts.” These digital contracts allow terms to be contingent on a decentralized consensus that is tamper-proof and typically self-enforcing through automated execution (Cong and He 2019). Smart contracts address traditional moral hazards by verifying “hidden actions,” reducing enforcement costs and deterring strategic behavior. Smart contracts have been applied in global trade finance to alleviate frictions caused by human errors and supply chain delays. In 2017, Maersk, a shipping and logistics company, established a blockchain platform for securely sharing shipping data. In the same year, in collaboration with IBM, Maersk also announced the completion of an end-to-end digitalized supply chain pilot. In this pilot, the blockchain shadowed the entire process, from raising a

purchase order to the goods being delivered. Every relevant document and approval was captured on the blockchain.

7.4. Empirical example: Auditing and blockchain



[Empirical example: Auditing and blockchain](#)

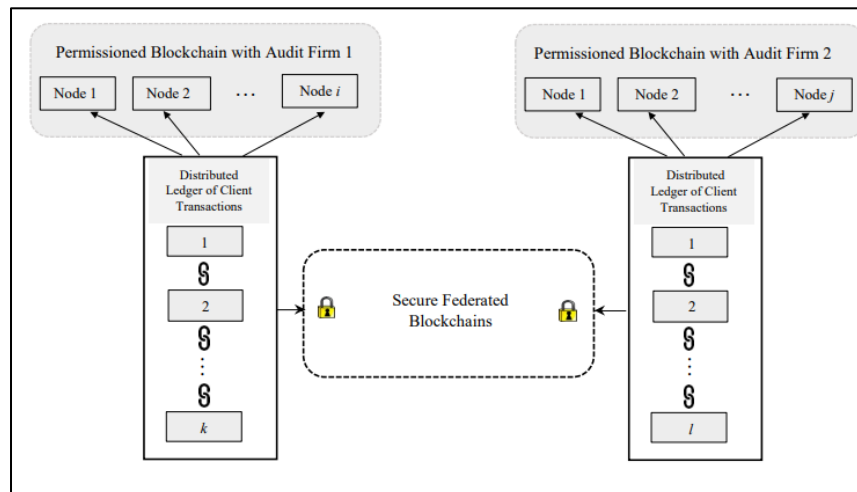
Unbiased financial reporting is crucial in financial markets; hence, regulatory agencies have constantly sought ways to improve reporting integrity. Because firms are liable to misreport, one main cost in ensuring quality reporting comes from auditors' verification of their clients' transactions. Cao, Cong, and Yang (2023) take an initial step towards understanding how blockchain technologies could transform auditors' monitoring/inspection of financial reports in the future.

If a seller claims \$1 million in accounts receivable sales, it boosts auditors' confidence in the number if the buyer can verify \$1 million in accounts payable purchases. The buyer has little incentive to collude with the seller, because if the buyer overstates the purchase for the seller's overstated sales, it implies a lower net income for the buyer. The collusion cost for buyers implies that the information that buyers provide to verify sellers' transactions is sometimes more reliable than the information that sellers themselves provide. However, such cross-party information verification is costly in the traditional auditing system. In those cases, an auditor has to contact the transaction counter-party directly to request records and manually verify the information, or outsource this labor-intensive cross-party verification to a third party, such as confirmation.com, at significant expense.

Figure 9 demonstrates how a federated blockchain with an encryption protocol has the potential to facilitate collaborative auditing and cross-party verification. In a federated blockchain, each auditor operates a permissioned blockchain for clients or has access to the blockchain ecosystem

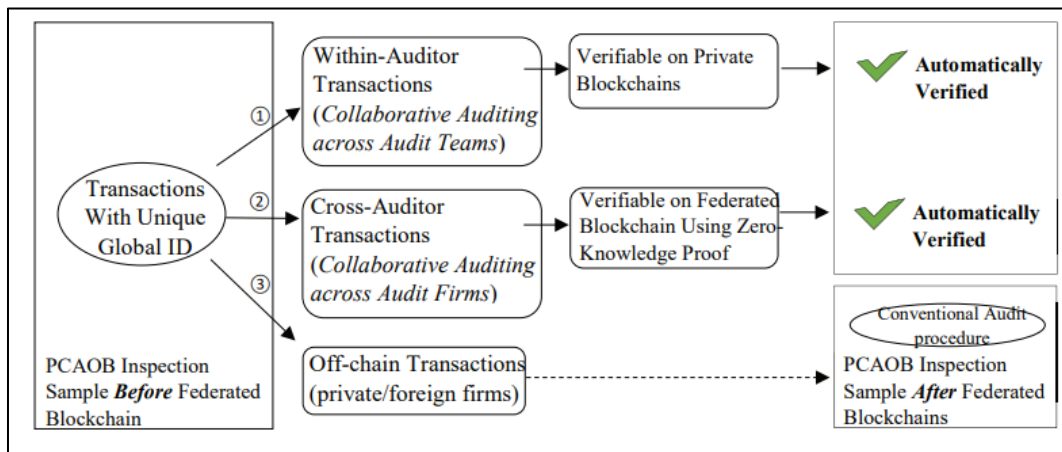
of its clients. In the base scenario, each node on the permissioned blockchain is administered by a team of the auditing firm. We note that permissioned blockchains considered for business applications typically only allow permissioned parties to join, use an efficient consensus mechanism such as majority voting, and may not need an intrinsic cryptocurrency/token, all of which differentiate them from public/permissionless blockchains like Bitcoin or Ethereum. These features offer more privacy, energy efficiency, and scalability, although they are not fully decentralized. Each client transaction is assigned a unique global ID to facilitate cross-party information verification. Transactions among clients of the same auditor are verified by the auditing teams working with the clients and recorded on the permissioned blockchain. Records on the permissioned blockchains are synchronized across all nodes to ensure immutability. On the permissioned blockchains, only permissioned nodes can manage records, and the nodes usually adopt a majority consensus that is efficient and scalable. Consequently, the costly mining process associated with public blockchains with proof-of-work protocols is avoided. Transactions between parties associated with different auditors utilize a cryptographic verification method that enables confirmation on the federated blockchain while maintaining the integrity of proprietary information.

Figure 9. Structure of the federated blockchain



As demonstrated in Figure 10, this kind of federated blockchain framework can facilitate two types of collaborative auditing. Type 1 concerns within-auditor transactions; the two parties in the transaction are audited by the same auditing firm but by different auditing teams. However, auditor teams may be located remotely in different audit offices, leading to high communication costs. A permissioned blockchain connecting the audit teams can automate the verification process. Type 2 entails collaborative auditing across firms, which could not happen without the federated blockchain system. In this case, the two parties in the transaction are audited by different audit firms, each residing in a separate blockchain ecosystem. The federated blockchain with encryption can facilitate automatic information sharing between auditors while protecting the privacy of clients' information. If a discrepancy is detected during the secure verification process, the auditors can reach out to the clients for the original records, or reach out to the counter-parties of the clients for authorization of verification. We note that once the blockchain is in place, any discrepancies are automatically detected, and firms will not have an incentive to misreport on the blockchain.

Figure 10. Auditing transactions within the blockchain



References

- Brav, A., Jiang, W., Li, T., and Pinnington, J. 2021. Picking friends before picking (proxy) fights: How mutual fund voting shapes proxy contests. Columbia Business School Research Paper (18-16).
- Cao, S., Cong, L., and Yang, B. 2023. Financial reporting and blockchains: Misreporting, auditing, and regulation. Working Paper,
- Cao, S., Jiang, W., Wang, J, and Yang, B. 2024. From man vs. machine to man+machine: The art and AI of stock analyses. *Journal of Financial Economics*, 160, 103910.
- Cao, S., Jiang, W, Yang, B, Zhang, A. 2023. How to talk when a machine is listening? Corporate disclosure in the age of AI. *Review of Financial Studies*, 36(9), 3603-3642.
- Cong, L., and He, Z. 2019. Blockchain disruption and smart contracts. *Review of Financial Studies*, 32(5), 1754-1797.
- Edelman, P., Jiang, W., and Thomas, R. 2019. Will tenure voting give corporate managers lifetime tenure? *Texas Law Review*, 97(5), 991-1030.
- Hong, H., Karolyi, G.A., and Scheinkman, J. 2020. Climate finance. *Review of Financial Studies*, 33(3), 1011-1023.
- Jiang, W., and Li, T. 2024. Corporate governance meets data and technology. Working paper.
- Karolyi, G.A., and Puente, J. 2023. Biodiversity finance: A call for research into financing nature. *Financial Management*, 52(2), 231-251.
- Krueger, P., Sautner, Z., and Starks, L. 2020. The importance of climate risks for institutional investors. *Review of Financial Studies*, 33(3), 1067-1111.
- Loughran, T., and McDonald, B. 2011. When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *Journal of Finance*, 66(10), 35-65.

- Myers, S., and Majluf, N. 1984. Corporate financing and investment decisions when firms have information that investors do not have. *Journal of Financial Economics*, 13, 187-221.
- Starks, L. 2023. Presidential address: Sustainable finance and ESG issues-value versus values. *Journal of Finance*, 78 (4), 1833-2411.
- Yermack, D. 2017. Corporate governance and blockchains. *Review of Finance*, 21(1), 7-31.
- Zhu, C. 2019. Big data as a governance mechanism. *Review of Financial Studies*, 32(5), 2021-2061.

Chapter 8 Analyzing Unstructured Data from Fund Managers

8.1. Mutual fund disclosure in Form N-CSR



[Mutual fund disclosure in form N-CSR](#)

Mutual funds registered with the SEC are obliged to furnish their shareholders with semiannual reports covering the first six months of the fund's fiscal year, along with an annual report summarizing the entire fiscal year's performance. While mutual funds must adhere to the SEC's mandated information disclosure criteria within the shareholder report, they have the flexibility to structure and present the information at their discretion. For a list of selected mandatory information to be included in mutual funds' shareholder reports, please refer to Table 1.

Table 1. Required information in Form N-CSR

	Annual report	Semiannual report
Expense example showing the cost in dollars for a hypothetical \$1,000 investment over the period covered by the report.	Yes	Yes
Graphical representation of holdings-table, chart, or graph of holdings by category.	Yes	Yes
Audited financial statements.	Yes	No
Unaudited financial statements.	No	Yes
Financial highlights.	Yes	Yes
Remuneration or compensation paid to directors, officers, and others.	Yes	Yes
Statement regarding the basis for approval of investment advisory contract.	Yes	Yes
Performance information	Yes	No
Management information about directors and officers	Yes	No
Availability of additional information about fund directors	Yes	No

The expense example is calculated based on a fund's expense ratio over the preceding six months, excluding any impact from sales loads, if applicable. This demonstration consists of two tables. The first table presents the real-dollar costs associated with a hypothetical \$1,000 investment in the fund over six months. The second table reports the corresponding costs in dollars for a hypothetical \$1,000 investment in the fund over the same period, assuming a 5% annual return rather than the actual return during that specific period. Utilizing this standardized hypothetical performance for expense calculation allows users to assess the fund's expenses relative to those of other funds. Figure 1 provides the shareholder expense example from the Fidelity Advisor Leveraged Company Stock Fund's 2020 Annual Report.

Figure 1. The shareholder expense example in Fidelity Advisor Leveraged Company Stock Fund's 2020 Annual Report

	Annualized Expense Ratio- A	Beginning Account Value February 1, 2020	Ending Account Value July 31, 2020	Expenses Paid During Period, ^B February 1, 2020 to July 31, 2020
Fidelity Advisor Leveraged Company Stock Fund				
Class A	1.08%			
Actual		\$1,000.00	\$964.90	\$5.28
Hypothetical- ^C		\$1,000.00	\$1,019.49	\$5.42
Class M	1.32%			
Actual		\$1,000.00	\$963.60	\$6.44
Hypothetical- ^C		\$1,000.00	\$1,018.30	\$6.62
Class C	1.86%			
Actual		\$1,000.00	\$961.00	\$9.07
Hypothetical- ^C		\$1,000.00	\$1,015.61	\$9.32
Class I	.81%			
Actual		\$1,000.00	\$966.00	\$3.96
Hypothetical- ^C		\$1,000.00	\$1,020.84	\$4.07
Class Z	.68%			
Actual		\$1,000.00	\$966.80	\$3.33
Hypothetical- ^C		\$1,000.00	\$1,021.48	\$3.42

^A Annualized expense ratio reflects expenses net of applicable fee waivers.

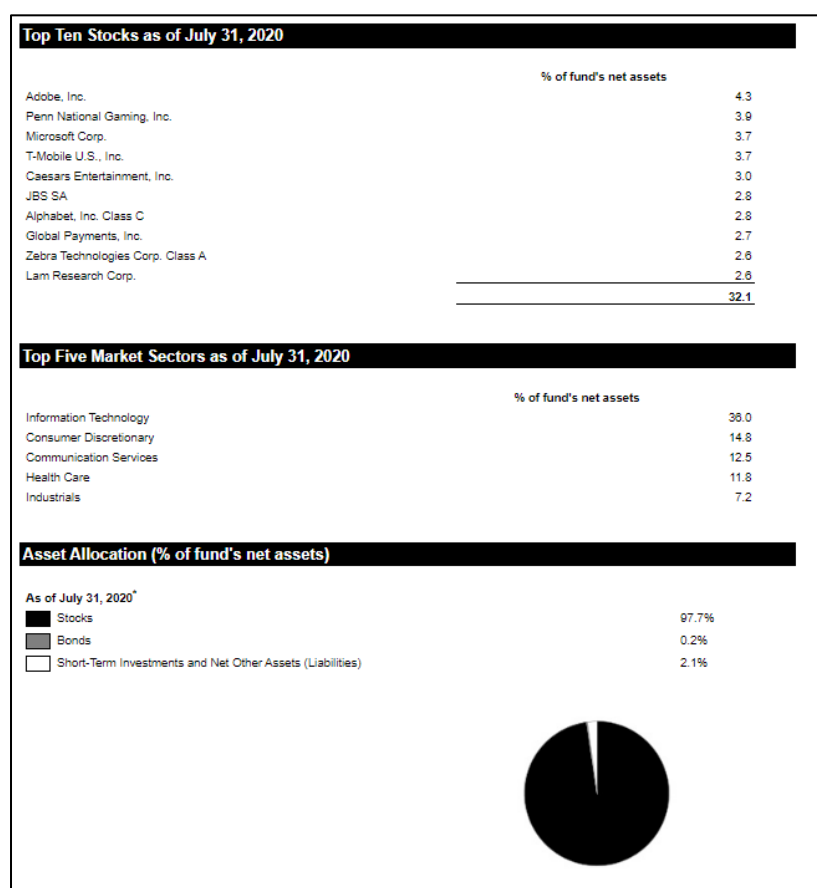
^B Expenses are equal to the annualized expense ratio, multiplied by the average account value over the period, multiplied by 182/366 (to reflect the one-half year period). The fees and expenses of any Underlying Funds are not included in each annualized expense ratio.

^C 5% return per year before expenses

The shareholder report includes a visual representation, whether it's a table, chart, or graph, depicting the fund's holdings categorized by various segments. Figure 2 offers an illustrative example. Additionally, the audited financial statements within the report provide either a comprehensive or summarized inventory of the portfolio holdings.

The financial highlights section includes a table summarizing financial data over the past five years or since the fund's inception if it is less than five years old. This summary table provides a breakdown of the changes in the fund's net asset value (NAV) from one year to the next, spanning a continuous five-year period.

Figure 2. The investment summary in Fidelity Advisor Leveraged Company Stock Fund's 2020 Annual Report



Performance data within the report usually encompasses three key elements:

- Management's comprehensive analysis of the fund's performance is often provided through shareholder reports.
- An informative line graph presenting a decade of performance (or the fund's entire existence if it is than ten years old) by contrasting a hypothetical initial investment of \$10,000 against an index.
- An essential table that delineates the fund's returns for varying periods, such as 1-year, 5-year, and 10-year (or its full existence if less than a decade).

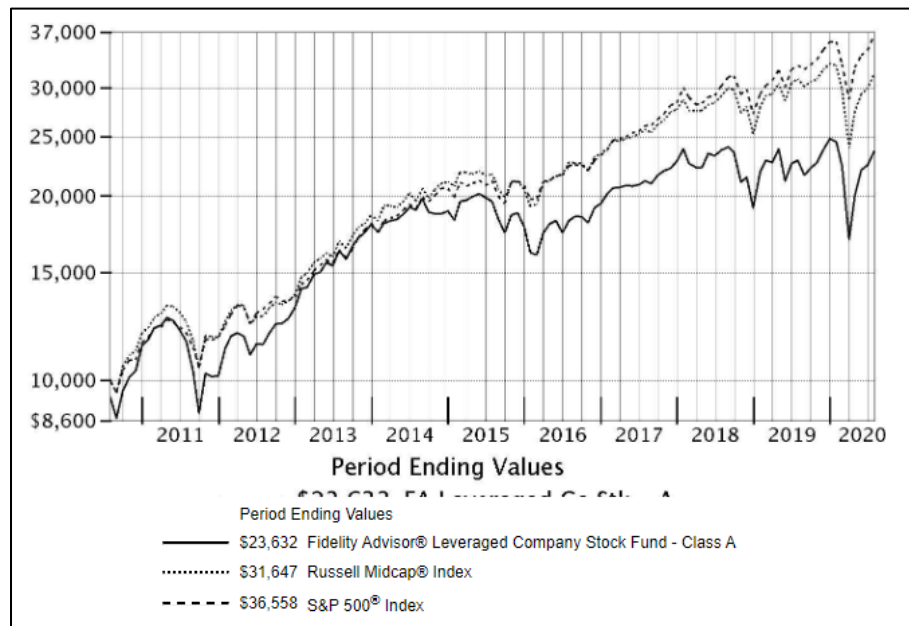
The Management's Discussion of Fund Performance is often positioned towards the report's beginning. This section offers management an opportunity to deliver a comprehensive review of the fund's performance throughout the fiscal year. This section analyzes the fund's performance in the context of prevailing market conditions and other factors that could have impacted the fund's returns. The annual report also includes a graphical representation, often in the form of a line graph, tracking the progress of a hypothetical \$10,000 initial investment in the fund over the past decade (or since the fund's inception if it is less than ten years old). This graph exhibits at least two lines, with one indicating the growth or decline in the value of the hypothetical \$10,000 investment over the past decade, and the other depicting the performance of a relevant, broad-based securities market index (e.g., the S&P 500 index) during the same period. Please refer to Figure 3 for an example of a performance graph. Below this graph, a table succinctly presents the fund's annualized (or average annual) returns over 1-year, 5-year, and 10-year periods (or for the fund's entire existence if it is less than ten years old). Figure 4 offers an illustrative example of a performance line graph.

Figure 3. The performance table in Fidelity Advisor Leveraged Company Stock Fund's 2020 Annual Report

Average Annual Total Returns			
For the periods ended July 31, 2020	Past 1 year	Past 5 years	Past 10 years
Class A (incl. 5.75% sales charge)	(2.40)%	2.63%	8.98%
Class M (incl. 3.50% sales charge)	(0.33)%	2.86%	8.98%
Class C (incl. contingent deferred sales charge)	1.79%	3.07%	8.81%
Class I	3.82%	4.12%	9.92%
Class Z	3.95%	4.26%	10.02%

In 2022, the SEC introduced amendments to modernize mutual fund shareholder reports. These amendments mandate that mutual funds deliver reports that are not only succinct but also visually engaging. Accordingly, funds are to provide tailored shareholder reports that emphasize key information, such as fund expenses, performance, and portfolio holdings. Furthermore, these revamped reports will come with instructions that promote the use of graphical and textual elements, enhancing their effectiveness. Additionally, funds will be obligated to structure the information in their reports in a data format that can be easily accessed. Funds will also be required to make certain information available online for those investors and financial professionals seeking more in-depth insights. This information will be accessible free of charge upon request. The purpose is to encourage transparency and fairness in how fees and expenses are presented in investment company advertisements (SEC 2022). The implementation of these rule amendments began on January 24, 2023.

Figure 4. The performance graph in Fidelity Advisor Leveraged Company Stock Fund's 2020 Annual Report



8.2. Empirical example: Extracting fund managers' private information and risk assessment from mutual fund shareholder reports



[Empirical example 1: Extracting fund managers' private information and risk assessment](#)



[Empirical example 2: Extracting fund managers' private information and risk assessment](#)

For mutual funds, shareholder reports serve a dual purpose beyond the legal requirement to disclose information such as portfolio holdings, fund performance, accounting statements, and voting policies. They provide an effective channel to communicate with shareholders and potential investors on various topics, including a dissection of wins and losses, comments on sector and fund performance, emphasis on the investment philosophy, and insights into the economy and market. Moreover, shareholder reports often delve into matters related to risk-taking, fiscal policy, politics, and global issues. The content of shareholder reports is discretionary and does not follow SEC form templates. This section discusses two empirical examples of extracting managers' private information and risk assessments from shareholder reports.

Private information

While almost all quantitative information from shareholder reports is drawn from portfolio holdings presented line by line under a template and extensively studied by investors, analysts, and researchers, the unstructured, qualitative data from the rich textual discussions remains underexplored. This data can potentially reveal managers' private information, such as forward-looking information and personal judgments.

There are several empirical challenges in extracting value-relevant information from mutual fund shareholder letters. The first hurdle is to extract intrinsic syntactic and semantic features from the unstructured text. The traditional bag-of-words approach in textual analysis relies on the meaning of individual vocabulary words and thus omits higher-order interactive features among words and sentences, which can contain important qualitative information. For example, the word "board" would have the same context-free representation in "welcome on board" and "board of directors." The second obstacle arises from decoding relevant features from shareholder letters, i.e., determining what features are likely to be associated with the private information of fund managers.

To address the initial challenge, Cao, Yang, and Zhang (2024) employ a cutting-edge advancement in natural language processing (NLP). They leverage Bidirectional Encoder Representations from Transformers (BERT), a language model introduced by Devlin, Chang, Lee, and Toutanova in 2019. Unlike conventional language models that process context in a unidirectional manner, BERT simultaneously considers context from both left to right and right to left. Thanks to its pre-training on extensive volumes of unlabeled text, BERT can capture intricate higher-order semantic and syntactic relationships within words and sentences while retaining the broader context.

To address the second challenge, Cao et al. (2024) develop a deep neural network model to understand the relationship between linguistic features derived from fund managers' shareholder letters obtained through a large language model (LLM) and subsequent fund performance. According to the Fama and French (1993) and Carhart (1997) four-factor model, fund performance is measured as alpha. To train and validate their model, they partition the sample of shareholder letters retrieved from the SEC's Electronic Data Gathering, Analysis, and Retrieval (EDGAR) system and divide it into two subsets: a training set and a test set. They then train the model using shareholder letters from 2006 to 2014, which constitute the training set. Subsequently, they use the trained model to predict future fund performance based on shareholder letters from 2015 to 2018, which comprise the test set.

After addressing these two challenges, they establish a measure of fund informativeness. *Textual Fund Information* is derived from predictions made by the neural network model empowered by an LLM. Funds are then categorized as either informed or uninformed based on this measure. The results reveal that funds identified as informed consistently outperform their uninformed counterparts. Considering the substantial influence of Morningstar ratings on funds and, more importantly, on investors, the study also explores the potential link between *Textual Fund Information* and Morningstar ratings and the likelihood of the fund subsequently receiving an improved rating. The findings suggest that informed funds tend to achieve higher Morningstar ratings. This indicates that *Textual Fund Information* is a dynamic metric, capturing information that is largely independent of observable fund characteristics and persistent, unobservable factors at the manager, fund, and company levels.

For fund managers, the ability to accurately evaluate and manage risk is directly relevant to their goal of achieving superior returns. Conventional risk-taking measures are calculated based

on historical numerical data of returns or holdings. Despite being easy to process and analyze, numerical data do not contain forward-looking information regarding managers' risk assessments. In contrast, while managers' qualitative discussions of portfolio decisions in shareholder reports convey rich information, these textual data are challenging to process due to their unstructured and high-dimensional nature.

Risk assessment

Cao, Yang, and Zhang (2023) utilize deep learning to extract syntactical relationships between words relevant to risk management from mutual fund managers' portfolio discussions in shareholder reports. Specifically, they construct pairs of dependent words using the neural-network dependency-parser described in 4.2. Each pair consists of a risk-related word and a sentiment word. For instance, in the excerpt "To avoid **excessive risks**, we chose to avoid companies that appeared to be at highly depressed valuations at the cost of short-term upside potential," "excessive" is the sentiment word and "risk" is the risk-related word. Unsurprisingly, this pair expresses negative risk assessment. On the contrary, in the excerpt "we maintain our focus on identifying businesses with idiosyncratic growth drivers that should power through a variety of economic or market scenarios and whose stocks present **attractive risk**/reward opportunities," "attractive" is the sentiment word and "risk" is again the risk-related word. This pair expresses positive risk assessment.

They then investigate the following questions: 1) Can deep learning effectively capture managers' narrative risk assessments that reflect their future risk-taking? 2) Are forward-looking risk assessments associated with fund skill and superior performance? 3) How do investors respond to managers' narrative risk assessments? Their results suggest that negative (positive) risk assessments strongly predict subsequent reductions (increases) in a manager's risk-taking

behavior, and that this measure outperforms the bag-of-words measures in predicting future risk-taking. Additionally, the study finds that risk-conscious managers who report negative risk assessments are able to generate higher future fund performance. These managers tend to be less overconfident and have better intra-quarter trading skills. Finally, the study finds that risk-conservative managers reduce both the systematic and idiosyncratic risks of their portfolios, indicating that they reduce both market exposure and individual bets. More specifically, they reduce their exposure to downside risk when they have negative risk assessments, suggesting that they may anticipate future market uncertainty of declines and avoid investing in risky stocks.

References

- Carhart, M. 1997. On persistence in mutual fund performance. *Journal of Finance*, 52(1), 57-82.
- Cao, S., Yang, B., and Zhang, A. 2023. Managerial risk assessment and fund performance: Evidence from textual disclosure. Working paper.
- Cao, S., Yang, B., and Zhang, A. 2024. Beyond the lines: Deciphering private information from fund managers' narratives. Working paper.
- Devlin, J., Chang, M., Lee, K., and Toutanova, K. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186.
- Fama, E., and French, K. 1993. Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, 33(1), 3-56.

Chapter 9 Analyzing Image Data

9.1. Images in corporate executive presentations



[Images in corporate executive presentations](#)

Corporate executive presentations take on a unique role in corporate disclosures. These events include non-deal road shows organized to spark investor interest and promote the company's image; initial public offering (IPO) road shows where executives engage with potential investors before going public; broker-hosted investor conferences which provide CEOs with a platform to connect with a broader investor audience; and capital market day events dedicated to outlining the company's long-term vision and strategy.

Corporate executive presentations exhibit two distinctive features. Firstly, due to the time constraints executives face during live presentations, they often rely heavily on visual and graphic elements in their slides. Executives recognize the need to convey complex ideas concisely, and visuals are crucial in achieving this goal. Charts, graphs, diagrams, images, and videos are commonly used to capture the audience's attention and facilitate a better understanding of the presentation's content. Figure 1 shows an example of charts from an executive presentation. Figure 2 presents images of production sites under construction in executive presentation.

Secondly, executive presentations differ from other forms of corporate disclosures by providing a wealth of visual insights into the firm's product designs and operational plans. While other corporate disclosures predominantly focus on quantitative data, such as financial statements and performance metrics, executive presentations provide a complementary perspective by showcasing the visual aspects of the company's offerings. This may include detailed product designs, prototypes, manufacturing processes, supply chain diagrams, and strategic plans. Figure 3 displays two examples of images of product designs and prototypes used in executive

presentations. Using these visual elements, executives aim to provide a holistic view of the company's future direction, growth strategies, and competitive advantage.

Figure 1. An example of a chart in executive presentation

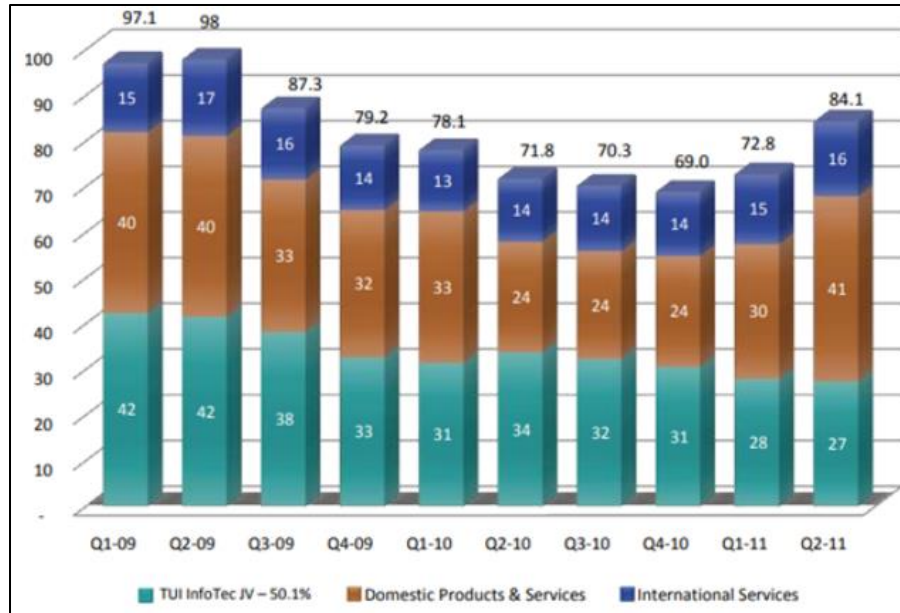


Figure 2. An example of images in executive presentations



Figure 3. Two examples of product images in executive presentations



9.2. Empirical example: Visual information in the age of AI

[Empirical example: Visual information in the age of AI](#)

Analyzing visual and graphic information has traditionally posed challenges due to its unstructured and high-dimensional nature. An image can contain tens of thousands of pixels, each with millions of possible colors that form complex patterns and objects. However, recent strides in machine learning and AI have empowered image recognition algorithms, which now boast

capabilities approaching human-level understanding. Cao, Cheng, Wang, Xia, and Yang (2023) leverage deep learning to extract essential features of firms' operations from corporate executive presentations.

In the initial phase, they undertake a manual review and categorization (labeling) of a randomly chosen subset of images, creating a training sample for the machine learning algorithms. Each image is assigned to one of three categories: Operations Summary, Operations Forward, or Others. To enhance accuracy and minimize human errors in the labeling process, they implement cross-validation and require consensus classification by at least three graduate research assistants. The construction of the training sample follows a two-step bootstrapping approach. Initially, they label a random sample of 3,000 images, using it to train the machine learning model and generate preliminary predictions for all images. Subsequently, they finalize the training sample by manually classifying 20,000 pre-labeled images with balanced representation across categories.

Various machine learning models, including random forests, gradient boosting, and neural networks, have been applied across diverse domains. Image recognition, which poses a significant challenge for deep learning, reached a notable milestone with Google's development of ImageNet (Li et al., 2009), which demonstrates performance on par with humans. The key deep learning model that ImageNet and other leading image recognition algorithms utilize is the Convolutional Neural Network (CNN). This multi-layer neural network features lower layers capturing finer details, while higher layers extract high-level information, such as identifying objects within the image.

Recognizing business images poses a challenge due to the absence of off-the-shelf models tailored for this purpose, and training a CNN model typically demands a sizable training dataset. Cao et al. (2023) utilize transfer learning (Pratt, 1993; Rajat et al., 2006) to build their deep learning

model based on pre-trained CNN models and train it with their business image sample. Specifically, they build a neural network on top of a pre-trained CNN neural network from a state-of-the-art image recognition model VGG16 (Simonyan and Zisserman, 2014). They keep the parameters of CNN layers, fine-tuning the model with the training sample. The resulting model is termed the Transfer CNN model. Transfer learning takes advantage of existing CNN models trained on extensive datasets and tailors the model to a specific business problem. Cao et al. (2023) also explore a model that integrates both image and text information from presentations (Transfer CNN + Text).

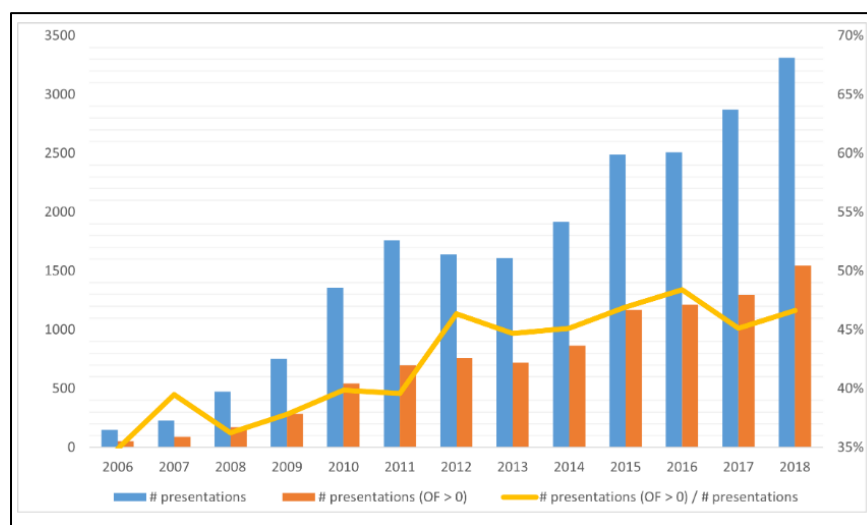
They train four different model architectures: 1) A CNN model built from scratch (CNN); 2) A deep learning model that processes both images and text (CNN + Text); 3) A transfer learning model utilizing a pre-trained CNN model for image processing (Transfer CNN); and 4) A transfer learning model that processes both images and texts (Transfer CNN + Text). For each model, they evaluate its out-of-sample performance using the ten-fold cross-validation approach, employing the stratified sampling method to split samples. The results are presented in Table 1. They use four measures to evaluate the out-of-sample performance of the models, including accuracy, precision, recall, and F1 score. Among the four architectures, Transfer CNN and Transfer CNN + Text exhibit the best F1 score and accuracy performance. Transfer CNN + Text outperforms other models with an accuracy of 80.0% and an F1 score of 79.3%. After fitting the transfer CNN + text model, they use the fitted model to obtain a final classification for the entire image sample.

Table 1. Performance of machine learning models

	CNN (%)	CNN + Text (%)	Transfer CNN (%)	Transfer CNN + Text (%)
<i>Accuracy</i>	75.0	76.5	77.0	80.0
<i>Precision</i>	77.3	70.7	77.5	78.9
<i>Recall</i>	75.0	76.5	77.0	80.0
<i>F1 Score</i>	73.5	69.8	78.7	79.3

Utilizing the classified categories of each slide page, they aggregate the number of pages under a specific category at the presentation level, scaling it by the total number of pages. On average, a presentation slide deck includes 3.6% Operations Forward slides and 11% Operations Summary slides. Figure 4 shows the time series of different types of information contained in presentations from 2006 to 2018. Alongside a clear increasing trend in the number of all presentation types, the ratio of presentations with Operations Forward images also rises over times. Notably, in 2006, only 35% of corporate presentations included Operations Forward images; this figure increased to 40% in 2010 and 47% in 2018, indicating that firms are more inclined to incorporate Operations Forward visual information in recent years.

Figure 4. Time series of corporate presentations



This figure plots the annual number of presentations (bar plot and left axis), and the ratio of presentations with Operations Forward images over all types of presentations (line plot and right axis) in our sample from 2006 to 2018. Presentations are classified as containing Operations Forward images if any slides in the presentation display figures with Operations Forward information.

Source: Cao et al. (2023)

References

- Cao, S., Cheng, Y., Yang, M., Xia, Y., and Yang, B. 2023. Visual information in the age of AI: Evidence from corporate executive presentations. Working paper.
- Deng, J., Dong, W., Socher, R., Li, L., Li, K., and Li, F. 2009. ImageNet: a large-scale hierarchical image database. *IEEE Conference on Computer Vision and Pattern Recognition*, 248–255.
- Pratt, L. Y., 1993. Discriminability-based transfer between neural networks, NIPS Conference: Advances in Neural Information Processing Systems 5, Morgan Kaufmann Publishers, 204–211.
- Rajat, R, A.Y., Ng, A., and Koller, D. 2006. Constructing Informative Priors using Transfer Learning, Twenty-third International Conference on Machine Learning.
- Simonyan, K., and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

Chapter 10 Analyzing the Balance Sheet

10.1. Data structure in balance sheets



[Data structure of the balance sheet](#)

A balance sheet provides a snapshot of a company's financial position at a given point in time. It outlines the company's resources (assets), namely, what it owns. The balance sheet also reports the sources of financing for these assets. There are two primary methods through which a company can finance its assets. It can raise funds from stockholders, known as owner financing, or it can acquire capital from banks, creditors, and suppliers, which is known as nonowner financing. This means both owners and nonowners hold claims on the company's assets. Owner claims on assets are referred to as equity, and nonowner claims are referred to as liabilities. As all financing is directed towards investments, we can establish the fundamental relationship: investing (assets) equals financing (liabilities + equity). This equality is called the accounting equation (Figure 1).

Figure 1. The accounting equation in the balance sheet

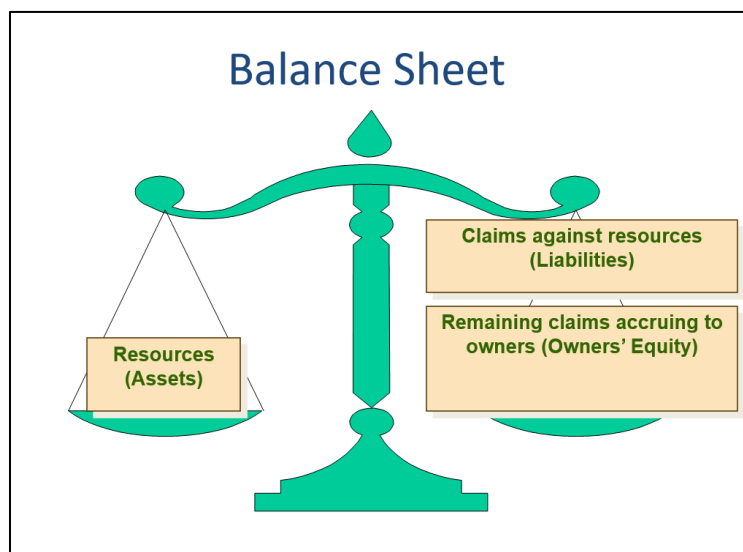


Figure 2 illustrates Los Gatos Corporation's balance sheet as of December 31, 2013. There are fourteen line items across the assets, liabilities, and owners' equity categories. For a larger

company with intricate business structures and models, the balance sheet could encompass over 250 variables. These variables offer valuable information for stakeholders seeking to understand the company's operating strategies and outcomes. For instance, managers can leverage the balance sheet to assess liquidity and solvency, while investors can use it to evaluate companies' operating performance.

Figure 2. The balance sheet of Los Gatos Corporation

LOS GATOS CORPORATION	
Balance Sheet	
At December 31, 2013	
Assets	
<i>Current assets:</i>	
Cash	\$ 20,000
Accounts receivable, net of allowance for uncollectible accounts of \$5,000	55,000
Inventories	<u>55,000</u>
Total current assets	130,000
<i>Investments:</i>	
Bond sinking fund	\$ 20,000
Note receivable	<u>20,000</u>
Total investments	40,000
<i>Property, plant, and equipment:</i>	
Machinery	190,000
Less: Accumulated depreciation	<u>(70,000)</u>
Net property, plant, and equipment	120,000
<i>Intangible assets:</i>	
Franchise	<u>30,000</u>
Total assets	<u><u>\$320,000</u></u>
Liabilities and Shareholders' Equity	
<i>Current liabilities:</i>	
Accounts payable	\$ 50,000
Interest payable	5,000
Note payable	<u>50,000</u>
Total current liabilities	105,000
<i>Long-term liabilities:</i>	
Bonds payable	110,000
<i>Shareholders' equity:</i>	
Common stock, no par value; 100,000 shares authorized; 50,000 shares issued and outstanding	\$ 70,000
Retained earnings	<u>35,000</u>
Total shareholders' equity	<u>105,000</u>
Total liabilities and shareholders' equity	<u><u>\$320,000</u></u>

Debate about fair value accounting



[Debate about fair value accounting](#)

Before delving into the analysis of a balance sheet, it is crucial to acknowledge that the values of the variables within the balance sheet are determined based on accounting standards and managerial judgment. Assets and liabilities are measured either at fair value or historical cost, following relevant accounting standards. When using the historical cost accounting method, the focus lies on the initial price paid by the company during the acquisition of an asset or the incurrence of a liability. The balance sheet reflects either the purchase price or a reduced value due to factors such as obsolescence, depreciation or depletion. For financial assets, the price remains unchanged until the security is liquidated. Historical cost accounting is considered more conservative and reliable since it is based on a fixed price that is fully known, namely the actual price paid by the company. While this eliminates uncertainty in the initial valuation decision, it introduces uncertainty in future periods regarding the true value of assets, which distracts from the relevance of historical cost.

The alternative approach measures assets and liabilities at fair value, which represents the price at which knowledgeable and willing parties would exchange or settle them. Fair value accounting entails adjusting the prices of certain assets on the balance sheet in each reporting period to reflect changes in market prices. Fair value accounting enhances the relevance of accounting information. However, determining the fair value of assets and liabilities is not always straightforward, as it involves subjective judgement. Given the pros and cons of both historical cost and fair value accounting, debates persist on how assets and liabilities on a balance sheet should be valued. When analyzing a balance sheet, it is essential to consider the values of the assets and liabilities and how they are measured.

Furthermore, the evolution of business models across various industries has led to a shift in value creation, with increasing emphasis on intangible assets such as ideas, knowledge, brands, content, data, and human capital rather than physical assets like machinery or factories. However, the accounting framework has not kept pace with this transformation. Existing accounting standards often fail to recognize the value generated by certain intangible assets, both in terms of their representation on the balance sheet or disclosure in footnotes. While tangible assets like property and equipment are typically included on a company's balance sheet, investments in internally generated intangibles are generally expensed as incurred. Consequently, a company's most valuable assets often remain unaccounted for on its balance sheet. When examining a balance sheet, it is thus crucial to consider the implicit value of these intangible assets. Due to the intricate nature of intangibles and the diversity in how companies manage and investors evaluate them, there is no universally applicable method for their measurement.

10.2. Empirical example: Analyzing data in the balance sheet



[Analyzing the balance sheet](#)



[An empirical example of analyzing the balance sheet](#)

Prior studies have investigated the implication of asset growth for shareholders. Fairfield, Whisenant, and Yohn (2003) find that asset growth exhibits negative associations with a one-year-ahead return on assets after adjusting for current profitability. On the other hand, Cooper, Gulen, and Schill (2008) document that asset growth rates are strong predictors of future stock returns. The question remains: is asset growth good or bad for shareholders?

While stockholders are owners of public companies, the day-to-day control of company resources lies in the hands of professional managers. This separation of ownership and control gives rise to what are known as agency problems. One of the typical agency problems is empire

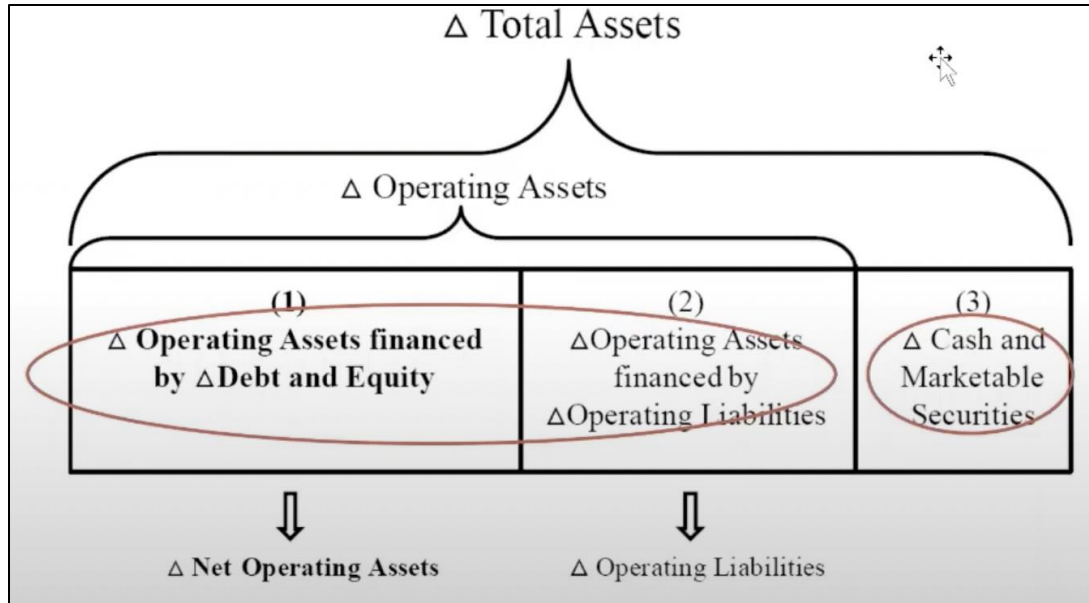
building. Managers are incentivized to grow companies aggressively to fulfill personal and career ambitions, but reckless expansions could result in inefficient usage of resources and decreases in shareholder wealth. Therefore, while asset growth may suggest healthy growth of the company's business, it might also indicate empire building. To fully understand the implication of asset growth, it is important to separate the growth of assets from normal business activities and the growth of assets caused by empire building.

Companies engage in various activities that can be categorized as operating or nonoperating. Operating activities encompass producing and selling company products and services to customers. Nonoperating activities involve nonstrategic cash investment in marketable securities and debt financing endeavors. Asset growth can stem from both operating and nonoperating activities. The growth of total assets can be broken down into two components: growth funded by operating liabilities and growth funded by debt and equity (refer to Figure 3). For instance, a company may negotiate favorable credit terms with suppliers, which essentially represents a loan from the suppliers to the company. Alternatively, the company could obtain a bank loan to finance purchases from suppliers. Both financing activities increase assets and liabilities as per the accounting equation.

There is no doubt that both suppliers and banks meticulously assess the financial standing of a company before making financing decisions. However, suppliers may possess a comparative advantage in terms of information as they have industry-specific knowledge, engage in daily business transactions, and have access to unbiased private information. Additionally, suppliers have stronger economic incentives as their credit risks are typically less diversified than those of banks. Consequently, we should consider whether growth financed by these more informed stakeholders, such as suppliers, may indicate better future performance, while growth financed by

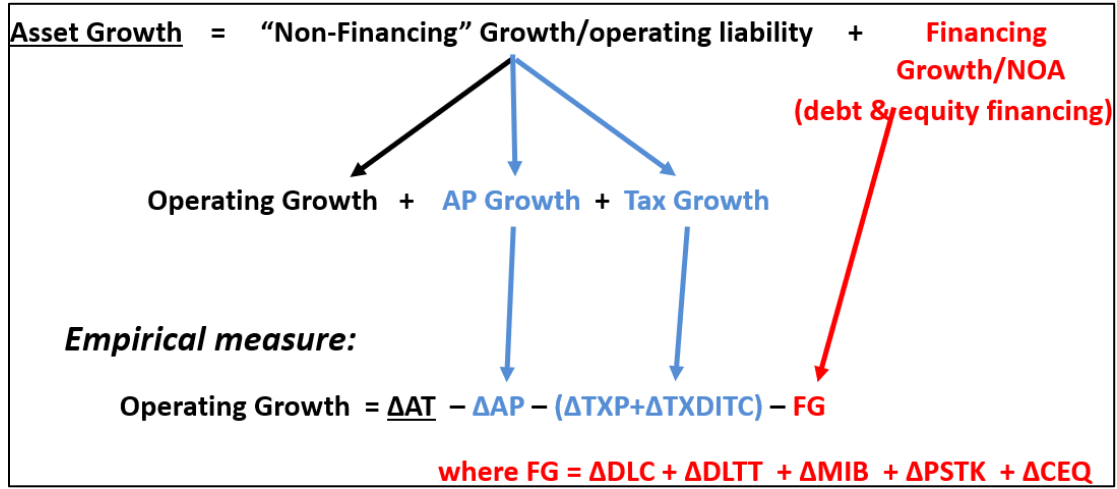
debt and equity could potentially predict worse future performance. This hypothesis can be explored by analyzing data from the balance sheet.

Figure 3. Decomposing balance sheet items



In order to examine the consequences of asset growth related to operating and nonoperating liabilities, Cao (2016) decomposes asset growth into non-financing growth and financing growth. Non-financing growth in assets is driven by increases in operating liabilities, such as accounts payable, while financing growth in assets arises from increases in debt or equity financing, such as bank loans. Non-financing growth can be further decomposed into operating growth, growth in accounts payable (representing financing from suppliers), and growth in tax payable (representing financing from tax authorities). The focus is to understand the distinct implications of operating growth, accounts payable growth, and tax payable for a company's operating performance and stock market performance, the two metrics that attract the most attention from investors.

Figure 4. Decomposing asset growth



Cao et al. (2016) measures operating performance with return on assets (ROA), computed as net income divided by average total assets at the beginning and the end of a year and reflects returns from the perspective of the entire company. This return includes both profitability (numerator) and total assets (denominator). To earn a high ROA, managers must earn profit and minimize the assets invested to the level necessary to achieve the profit. The explanatory variables are various components of asset growth. The regression model is stated below where ROA_{t+1} is one-year ahead ROA; ROA_t is the current ROA; and ΔROA_t is the change in ROA from $t-1$ to t .

$$ROA_{t+1} = \beta_0 + \beta_1 AssetGrowth_t + \beta_2 ROA_t + \beta_3 \Delta ROA_t + \varepsilon \quad (1)$$

Table 1 tabulates the regression results, including various types of asset growth as explanatory variables. Regression 1 includes growth in accounts payable ($OAgrowth_{AP}$) as one of the explanatory variables. Regression 2 includes growth in operating assets other than accounts payable and tax payable ($OAgrowth_{Other}$) as one of the explanatory variables. Regression 3 includes both in the same model. The results suggest that growth in accounts payable is negatively associated with future ROA, while growth in operating assets other than accounts payable and tax payable is positively associated with future ROA.

Regressions 4 and 5 add growth on net operating assets (*NOAgrowth*) as an additional explanatory variable. The results show that growth in net operating assets is negatively associated with future ROA. Regressions 6 and 7 further indicate that growth in both current and long-term net operating assets is negatively associated with future ROA. Interestingly, both the growth of accounts payable and growth of operating assets other than accounts payable and tax payable become positively associated with future ROA when controlling for the growth of net operating assets. This means that growth of net operating assets is a correlated omitted variable in Regressions 1, 2, and 3.

Regression 8 includes growth in various operating assets in one regression and “horse raced” against each other. The results confirm that the growth of operating assets financed by operating liabilities is positively associated with future ROA while growth of operating assets financed by debt and equity is negatively associated with future ROA.

Table 1. The implications of decomposition of *OAgrowth_OL* and *NOAgrowth* for one-year ahead ROA and stock returns

	<i>OAgrowth_AP</i>	<i>OAgrowth_Other</i>	<i>NOAgrowth</i>	<i>Current NOAgrowth</i>	<i>LT_NOAgrowth</i>	<i>CASHgrowth</i>	<i>ROA_t</i>	$\Delta ROA_{t, t-1}$	Adjusted <i>R</i> ²
Regression 1	-0.01*** (-3.43)						0.77*** (50.40)	-0.13*** (-5.92)	0.48*** (42.62)
Regression 2		0.01*** (6.89)					0.77*** (50.61)	-0.13*** (-6.11)	0.48*** (42.62)
Regression 3	-0.01*** (-5.04)	0.01*** (7.42)					0.77*** (50.56)	-0.13*** (-6.11)	0.48*** (42.71)
Regression 4	0.00 (0.47)	0.02*** (8.17)	-0.03*** (-11.98)			0.00 (1.14)	0.79*** (49.15)	-0.14*** (-6.99)	0.49*** (43.91)
Regression 5	0.00*** (2.93)		-0.03*** (-13.18)				0.79*** (49.29)	-0.14*** (-6.81)	0.49*** (43.25)
Regression 6				-0.02*** (-9.80)			0.78*** (49.89)	-0.13*** (-6.45)	0.49*** (43.01)
Regression 7					-0.03*** (-11.91)		0.79*** (49.74)	-0.13*** (-6.49)	0.49*** (42.83)
Regression 8	0.00 (0.59)	0.02*** (8.48)		-0.01*** (-6.67)	-0.03*** (-11.29)	0.00 (1.42)	0.79*** (49.00)	-0.14*** (-7.10)	0.49*** (44.20)

Cao (2016) measures stock market performance with a one-year ahead stock return and applies the Fama-MacBeth procedure to investigate the association between operating growth, growth in accounts payable, and growth in tax payable and a company's stock market performance.

Table 2 tabulates the regression results. Regression 1 includes growth in accounts payable (*OAgrowth_AP*) as one of the explanatory variables. Regression 2 includes growth in operating assets other than accounts payable and tax payable (*OAgrowth_Other*) as one of the explanatory variables. Regression 3 includes both in the same model. Similar to the findings on operating performance, growth in accounts payable is negatively associated with the future stock return, while growth in operating assets other than accounts payable and tax payable is positively associated with the future stock returns.

As growth in net operating assets could be an omitted correlated variable, Cao (2016) includes it as an additional explanatory variable in Regression 4 and Regression 5. The results of Regressions 4 through 7 show that growth in net operating assets is negatively associated with future stock returns. Furthermore, growth of operating assets other than accounts payable and tax payable becomes positively associated with future stock returns when controlling for growth of net operating assets. Regression 8 confirms that operating assets financed by operating liabilities are positively associated with future stock returns; growth of operating assets financed by debt and equity is negatively associated with future stock returns.

A natural follow-up question is that whether investors recognize the difference between the growth of different components of assets and use that information to develop profitable trading strategies. Growth of operating assets financed by operating liabilities should predict future stock returns only when investors do not recognize the differences among various types of asset growth. To test this possibility, Cao, Wang, and Yeung (2022) regress asset growth on a three-day stock returns around earnings announcements. Models (1) through (3) in Table 3 show that asset growth financed by operating liabilities (*OPERATING_GRWOTH*) is negatively associated with three-day earnings announcement return for future two quarters while asset growth financed by nonoperating

liabilities (*FINANCING_GRWOTH*) is positively associated with three-day earnings announcement return for future three quarters. This indicates that investors do not realize the difference between asset growth financed by operating liabilities and debt or equity initially; in fact, it takes them six months to figure it out.

Table 2. Fama-MacBeth regressions of subsequent stock returns on decompositions of *OAgrowth_OL* and *NOAgrowth*

	<i>OAgrowth_AP</i>	<i>OAgrowth_Other</i>	<i>NOAgrowth</i>	<i>Current NOAgrowth</i>	<i>LT_NOAgrowth</i>	<i>CASHgrowth</i>	Adjusted R^2
Regression 1	-5.38*** (-4.00)						0.00** (2.60)
Regression 2		2.14* (1.86)					0.00*** (3.29)
Regression 3	-6.11*** (-4.72)	3.42*** (3.16)					0.01*** (3.45)
Regression 4	-2.94** (-2.42)	4.90*** (4.89)	-11.49*** (-9.02)			-0.80 (-0.89)	0.01*** (5.97)
Regression 5	-2.10 (-1.66)		-10.97*** (-9.15)				0.01*** (4.95)
Regression 6				-7.30*** (-5.76)			0.01*** (3.31)
Regression 7					-9.27*** (-7.95)		0.01*** (4.74)
Regression 8	-3.01** (-2.47)	4.86*** (5.20)		-5.37*** (-4.52)	-8.43*** (-7.94)	-0.59 (-0.68)	0.02*** (5.69)

Table 3. Regressing quarterly earnings announcement return on growth variables

Model	Dep. Variable	OPERATING_GROWTH	FINANCING_GROWTH	EA_RET _t	EA_RET _{q-1}	EA_RET _{q-2}	EA_RET _{q-3}	EA_RET _{q-4}	RET (-6,-1)	SIZE	BM	R^2
1	EA_RET _{q,t+1} ($q = 1$)	0.474*** (4.35)	-0.506*** (-4.42)	0.214 (1.64)					0.373** (2.51)	0.553*** (4.33)	0.189 (1.20)	0.015
2	EA_RET _{q,t+1} ($q = 2$)	0.313*** (2.84)	-0.602** (-2.12)	0.555*** (3.13)					0.578** (2.60)	0.123 (0.84)	0.523*** (3.06)	0.017
3	EA_RET _{q,t+1} ($q = 3$)	0.211 (1.31)	-0.528*** (-3.24)	0.308** (2.63)					0.308** (2.58)	-0.052 (-0.32)	0.199 (0.80)	0.013
4	EA_RET _{q,t+1} ($q = 4$)	0.126 (1.19)	-0.240 (-1.49)	0.190* (1.79)					-0.011 (-0.09)	0.124 (0.80)	0.550*** (3.56)	0.009

Since average investors do not immediately recognize the difference between asset growth driven by operating liabilities versus growth driven by equity or debt, one could develop a profitable trading strategy by holding stocks with low growth in nonoperating assets and selling stocks with high growth in nonoperating assets. We can test the outcome of this strategy using the portfolio sort method.

Specifically, stocks are sorted based on total asset growth (*TAgrowth*) and nonoperating asset growth (*NOAgrowth*) into quintiles by year, forming 25 portfolios. Then, average returns on these portfolios are computed over a year. The equal-weighted portfolios show that the trading strategy based on net operating asset growth yields returns ranging from 6.69% to 12.29%, holding total asset growth constant. The value-weighted portfolios show that the trading strategy yields returns ranging from 7.67% to 9.05%, holding total asset growth constant. Meanwhile, a trading strategy based on total asset growth does not yield significant positive returns.

Table 4. Comparisons of one-year-ahead abnormal returns of portfolios based on *NOAgrowth* and *TAgrowth*

Panel A: Equal-weighted portfolios						
<i>TAgrowth</i>	<i>NOAgrowth</i>					Control hedge H-L
	Low	1	2	3	High	
Low	3.95 (4.54)	2.41 (2.98)	-1.07 (-1.00)	-1.83 (-0.93)	-8.26 (-1.86)	-12.29** (-2.59)
1	5.45 (4.59)	3.42 (5.68)	1.21 (1.37)	-0.58 (-0.38)	-3.64 (-1.15)	-9.08** (-2.47)
2	6.94 (5.08)	6.10 (6.33)	1.91 (2.29)	-0.03 (-0.03)	-7.44 (-3.95)	-14.38*** (6.27)
3	6.69 (2.49)	3.08 (2.59)	2.59 (3.15)	0.10 (0.12)	-1.64 (-1.96)	-8.33*** (-2.98)
High	-0.20 (-0.07)	3.13 (0.96)	3.56 (1.84)	-1.01 (-0.92)	-6.89 (-4.76)	-6.69*** (-2.73)
H-L Control hedge	-4.14 (-1.43)	0.72 (0.21)	4.64* (1.85)	0.83 (0.47)	1.63 (0.39)	
Panel B: Value-weighted portfolios						
<i>TAgrowth</i>	<i>NOAgrowth</i>					Control hedge H-L
	Low	1	2	3	High	
Low	3.36 (2.78)	1.01 (1.49)	1.23 (0.78)	1.24 (0.46)	-5.65 (-1.44)	-9.05** (-2.16)
1	3.10 (2.56)	2.91 (2.81)	1.06 (1.05)	0.20 (0.07)	-1.93 (-0.50)	-5.02 (-1.30)
2	4.44 (3.74)	2.33 (2.46)	0.82 (1.58)	-1.39 (-1.08)	-5.60 (-2.37)	-10.04*** (-3.62)
3	7.67 (2.33)	1.55 (1.34)	-1.03 (-0.83)	-1.85 (-1.72)	-4.18 (-3.44)	-11.85*** (-2.86)
High	3.88 (0.91)	3.85 (1.61)	0.01 (0.00)	0.50 (0.24)	-3.78 (-3.28)	-7.67* (-1.93)
H-L Control hedge	0.52 (0.13)	2.84 (1.15)	-1.22 (-0.39)	-0.74 (-0.19)	2.28 (0.56)	

Finally, the same argument suggests lower earnings volatility could be associated with lower trading frictions. This is because firms with low earnings volatility likely experience less information asymmetry between buyers and sellers of the stock due to less uncertainty about future earnings. This, in turn, reduces trading frictions since the possibility of adverse selection is lower.

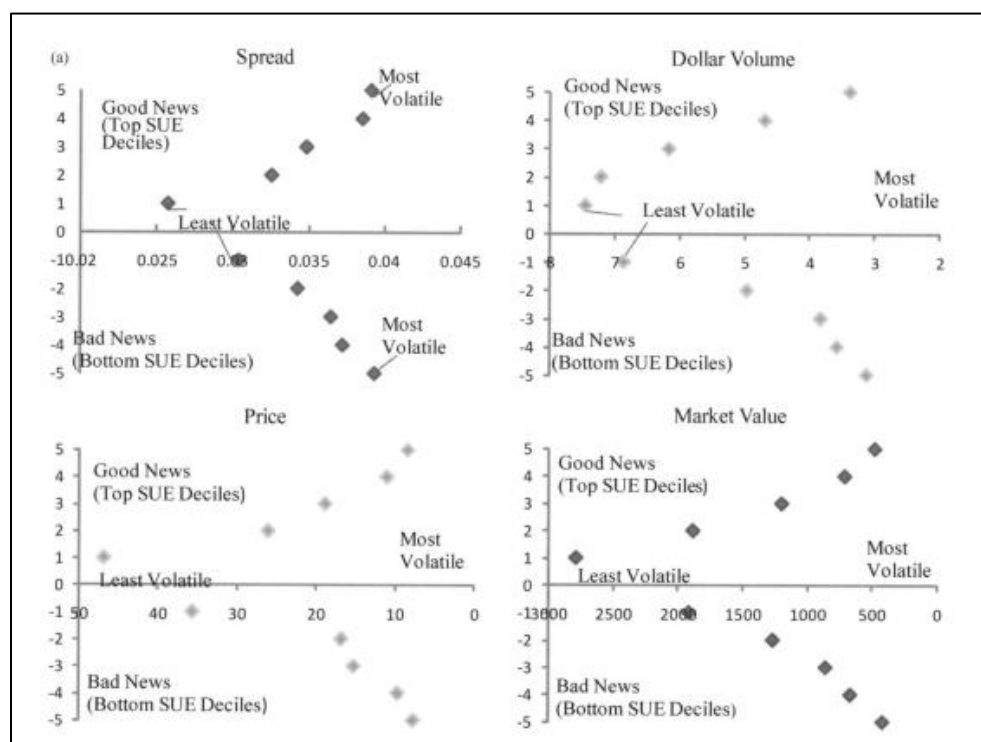
To examine the association between the volatility effect and trading frictions, Cao, Wang, and Yeung (2022) compare the average value of the trading friction proxy across volatility quintiles in the top and bottom standardized unexpected earnings (SUE) deciles to compare trading friction proxy values across different volatility quintiles while holding the SUE decile constant. Figure 5 displays the results. Transaction costs increase from left to right in each graph, which implies that the inverse proxies described above decrease from left to right. Spread is the transaction cost proxy in the top left panel. In the top two SUE deciles, the spread increases monotonically from 0.026 in the lowest volatility quintile to 0.039 in the highest volatility quintile. The bottom two SUE deciles display the same pattern, with Spread increasing from 0.030. This suggests that, when keeping the level of SUE constant, transaction costs as proxied by Spread increase in earnings volatility.

10.3. Machine learning applications for balance sheet data

The field of equity markets research has experienced significant growth and advancements due to the utilization of AI and big data. Initially, early studies in this domain focused on employing new methodologies and data to gain deeper insights into existing research questions, particularly in the areas of earnings and returns forecasting. Machine learning algorithms have advantages over traditional regression techniques, as they allow for non-linearity, incorporating high-dimensional and complex time-series data and implementing cross-validation techniques. Consequently, the adoption of AI and machine learning algorithms holds the potential to enhance forecasting performance. However, more recent studies have shifted their focus towards addressing emerging

questions that have arisen as a result of AI and big data, including the comparison between human and machine performance in various tasks. This transition reflects the evolving landscape of equity market research driven by technological advancements.

Figure 5. Earnings volatility and transaction cost proxies



In a recent study, Chen, Cho, Dou and Lev (2022) use decision tree methods to predict directions and signs of earnings, comparing them with a conventional logit model and financial analyst forecasts. They feed in more than 4,000 financial items identified through XBRL tags in corporate 10-K filings in current and lagged years and their annual changes, which together yield over 12,000 input variables. For every three-year period, they assign the first two years as a training period and the final year as the validation period for model selection and then conduct out-of-sample tests. They find that the machine learning approach demonstrates better out-of-sample predictive power and significant returns to portfolios formed on the basis of AI-generated

predictions. It should be noted, however, that the logit model is estimated based on a limited group of selected input variables; therefore, the superiority of the decision tree approach is attributable to both nonlinearity and the large group of inputs.

Cao and You (2024) examine the efficacy of machine learning in forecasting corporate earnings compared to conventional fundamental analysis models. They specifically examine three linear machine learning models, ordinary least squares regression (OLS), least absolute shrinkage and selection operator (LASSO), and Ridge regression (RIDGE), as well as three nonlinear machine learning models, random forest (RF), gradient boosting regression (GBR) and artificial neural networks (ANNs). These they compare with six conventional time-series and cross-sectional models. Feeding in a selection of 56 features from financial statements, they find that nonlinear machine learning models generate significantly more accurate and informative forecasts than the conventional forecasting models found in the literature. Notably, the superior forecasting capabilities of nonlinear machine learning models are attributable to both the nonlinearity and more disaggregated input features.

Chattopadhyay, Fang and Mohanram (2022) apply machine learning algorithms in earnings forecasts internationally, finding that the GBR and RF models perform the best in a global context compared to other simple linear models in the extant literature. The performance gain is particularly large for international firms with poorer information environments and more volatile earnings.

Binz, Schipper and Standridge (2022) apply a neural-network-based machine learning algorithm in a Dupont analysis framework to estimate Nissim and Penman's (2001) structure of decomposing accounting profitability and compare its out-of-sample predictability with random walk and linear regression models. Unlike the previous two studies, the inputs of their machine

learning algorithm are ratios based on Nissim and Penman (2001), and their focus is on the nonlinear relation between the input factors and the target profitability measures to be forecasted. They find that machine learning algorithms that incorporate nonlinearity perform better than the random walk or linear models and that investing strategies based on intrinsic values drawn from those forecasts generate significant abnormal returns. They further find that using a long time series of past information actually impairs forecasting performance.

AI and machine learning technologies have significantly influenced various aspects of our lives, including lifestyle, culture, economy, and environment. Accounting research has not been immune to this impact. Taking advantage of the advancements in AI and machine learning, accounting researchers have begun to harness AI technologies and new data in the realms of asset, liability, and equity.

Appendix 10. Regression Methods

A10.1. Linear Regression



[An overview of regressions](#)



[Three examples of regressions](#)

We use regressions to develop an understanding of relationships between variables. In regression and statistical modeling in general, we want to model the relationship between an output variable, or a response, and one or more input variables or factors. Depending on the context, output variables might also be referred to as dependent variables, outcomes, or simply Y variables, and input variables might be referred to as explanatory variables, effects, predictors or X variables. We can use regression and regression modeling results to determine which variables affect the response or help explain the response. This is known as explanatory modeling. We can also use regression to predict a response variable's values based on the important predictors' values. This is generally referred to as predictive modeling.

Simple linear regression is used to model the relationship between two continuous variables. The model below illustrates how Y changes for given values of X . Because the individual data values for any given value of X vary randomly about the mean, we need to account for this random variation, or error, in the regression equation. We add the Greek letter epsilon to the equation to represent the random error in the individual observations:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

Multiple linear regression models the relationship between a continuous response variable and a series of continuous or categorical explanatory variables. We add a slope coefficient for each explanatory variable when we fit a multiple linear regression model. Each coefficient represents

the average increase in Y for every one-unit increase in that explanatory variable, X_i , holding the other explanatory variable constant.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_i X_i + \varepsilon$$

A10.2. Fama-Macbeth Regression



[Fama-Macbeth regression](#)

The Fama-MacBeth procedure estimates consistent standard errors in the presence of cross-sectional correlation. The first step is a cross-sectional regression of the model to obtain the estimated beta factor of a stock at t over a period of T . The second step is to compute the overall estimate (λ) and standard errors (SE) using the time-series estimates of the beta-factor under the assumption that error terms are uncorrelated over time. A more modern approach is to run a standard panel regression and then cluster on the date variable.

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon$$

$$\lambda = \sum \beta_1 / T, \quad SE = \sqrt{\frac{\frac{1}{T} \sum (\beta_1 - \lambda)^2}{T}}$$

A10.3. Portfolio Sorts



[Two-way sorting](#)



[Risk-adjusted return sorting](#)

Economic theory, or empirical conjecture, often predicts that expected returns should be increasing (or decreasing) in some characteristic or feature. Portfolio sorts are very widely used to test theories or conjectures. One of the appeals of tests of the “top-minus-bottom” spread in portfolio returns is that they can be interpreted as the expected return on a trading strategy: short the bottom portfolio and invest in the top portfolio, reaping the difference in expected returns. A test based on a portfolio sort is usually conducted as follows:

- Individual stocks are sorted according to a given characteristic;

- These stocks are then grouped into N portfolios;
- Average returns on these portfolios over a subsequent period are then computed;
- The significance of the relationship is judged by whether the “top” and “bottom” portfolios have significantly different average returns.

References

- Binz, O., Schipper, K., and Standridge, K. 2022. What can analyst learn from artificial intelligence about fundamental analysis? Working paper.
- Cao, S. 2016. Reexamining growth effects: Are all types of asset growth the same? *Contemporary Accounting Research*, 33(4), 1518-1548.
- Cao, K., and You, H. 2024. Fundamental analysis via machine learning. *Financial Analysts Journal*, 80(2), 74-98.
- Chattopadhyay, A., Fang, B., and Mohanrm, P. 2022. Machine learning, earnings forecasting and implied cost of capital – US and international evidence. Working paper.
- Chen, X., Cho, T., and Dou, Y. 2022. Predicting future earnings changes using machine learning and detailed financial data. *Journal of Accounting Research*, 60(2), 467-515.
- Cao, S., Wang, Z., and Yeung, P. 2022. Skin in the game: Operating growth, firm performance, and future stock returns. *Journal of Financial and Quantitative Analysis*, 57(7), 2559-2590.
- Cooper, M., Gulen, H., and Schill, M. 2008. Asset growth and the cross-section of stock returns. *Journal of Finance*, 63(4), 1609-1651.
- Fairfield, P., Whisenant, S., and Yohn, T. 2003. Accrued earnings and growth: Implications for future profitability and market mispricing. *The Accounting Review*, 78(1), 353-371.
- Nissim, D., and Penman, S. 2001. Ratio analysis and equity valuation: From research to practice. *Review of Accounting Studies*, March, 109-154.

Chapter 11 Analyzing the Income Statement

11.1. Data structure in income statements



[Data structure of the income statement](#)

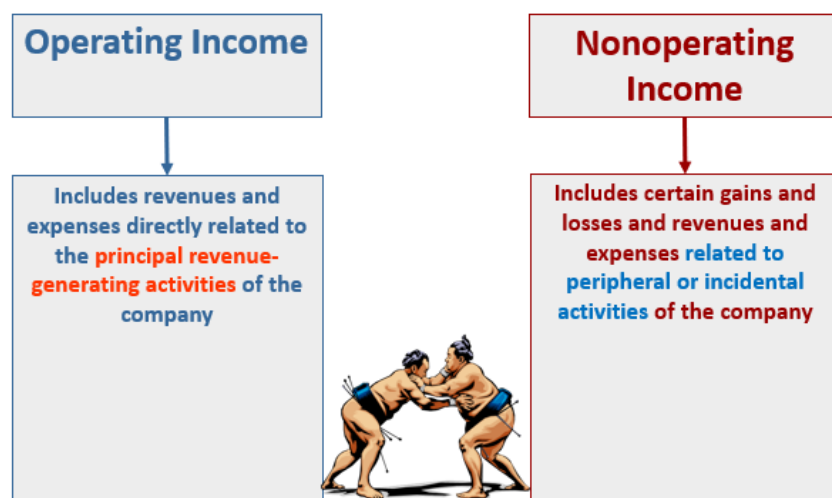
The income statement reports on a company's performance over a period of time and details its top-line revenue and its expenses. Revenue-less expenses equal the bottom-line net income. The income statement reports on both operating and non-operating activities. Operating activities relate to bringing a company's products or services to market and any after-sales support. The income statement captures operating revenues and expenses, yielding operating profit. Major operating line items in the income statement are revenues, costs of goods sold (COGS), and selling, general, and administrative expenses (SG&A). Nonoperating activities relate to items like borrowed money that creates interest expense and nonstrategic investments in marketable securities that yield interest or dividend revenue. Typical nonoperating line items on the income statement include interest expense on debt and lease obligations, loss of income related to discontinued operations, debt issuance and retirement costs, interest and dividend income on investments, and gains or losses on the sale of investments. Figure 1 provides an example of an income statement filed by Microsoft Corporation.

Operating profit less income tax on operating profit results in net operating profit after tax. This measure of a company's operating performance warrants special attention because it is the lifeblood of a company's value creation and growth. Total profit less total income tax results in net income. Net income is not equivalent to net operating profit after tax because nonoperating income could be transitory or irrelevant. Holding net income constant, a company with more operating income would have higher-quality earnings.

Figure 1. An income statement of Microsoft Corporation

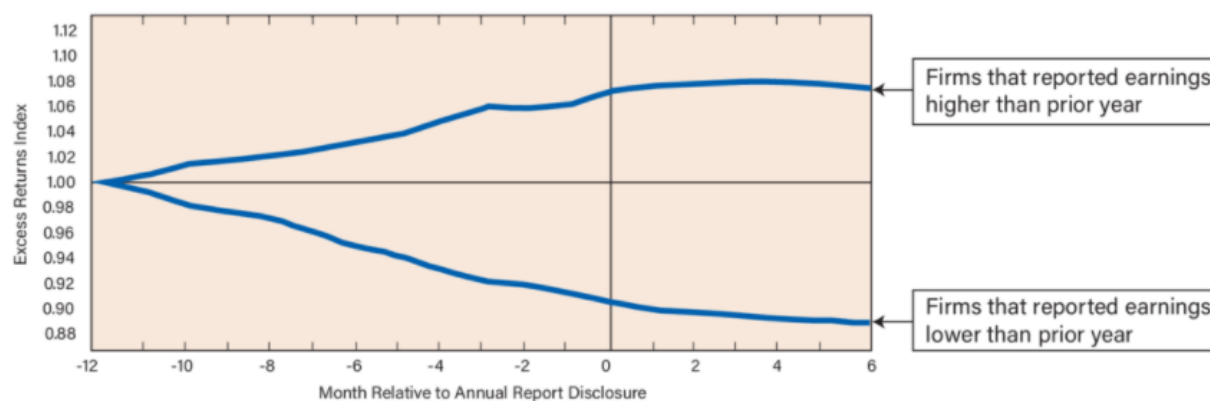
INCOME STATEMENT : MICROSOFT			
(In millions)	2012	2011	2010
Revenue	\$73,723	\$69,943	\$62,484
Cost of revenue	17,530	15,577	12,395
Research and development	9,811	9,043	8,714
Sales and marketing	13,857	13,940	13,214
General and administrative	4,569	4,222	4,063
Goodwill Impairment	6,193	0	0
Total operating expenses	51,960	42,782	38,386
Operating income	21,763	27,161	24,098
Other income (expense)		504	910
Income before income taxes		22,267	25,013
Provision for income taxes	5,289	4,921	6,253
Net income	\$16,978	\$23,150	\$18,760

The proportion of income attributable to a business's core operating activities of is one important aspect of earnings quality. If a business reports an increase in profits due to improved sales or cost reductions, earnings quality is considered high. Conversely, an organization can have low-quality earnings if changes in its earnings relate to other issues, such as the aggressive use of accounting rules, inflation, the sale of assets for a gain, or increases in business risk. In general, any use of accounting trickery to temporarily bolster earnings reduces the quality of those earnings. A key characteristic of high-quality earnings is that they are readily repeatable over a series of reporting periods rather than only being reported as the result of a one-time event.

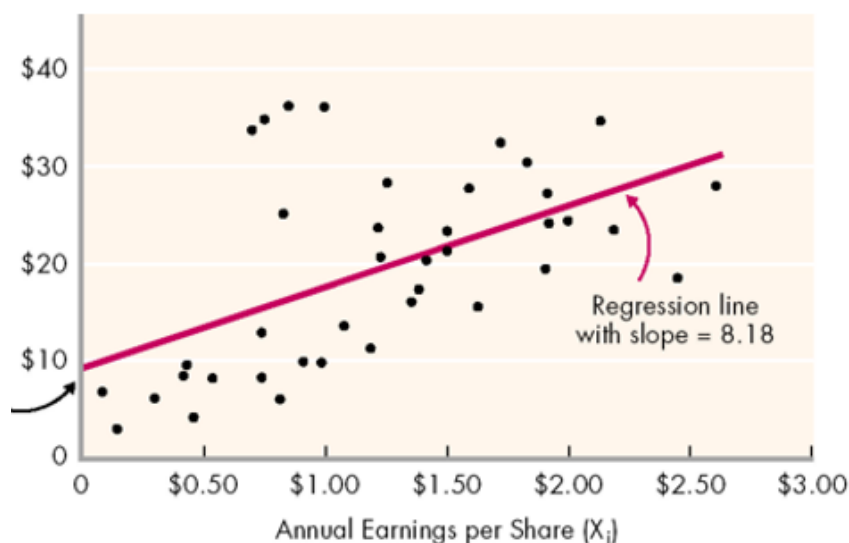
Figure 2. Operating income and non-operating income

11.2. Earnings and stock prices

Data in the income statement is closely related to stock markets. There is a natural positive relation between expected earnings and stock prices because investors expect dividends, which are paid out of earnings. Early research by Ball and Brown (1968) confirmed this expected relation.

Figure 3. Relation between expected earnings and stock prices

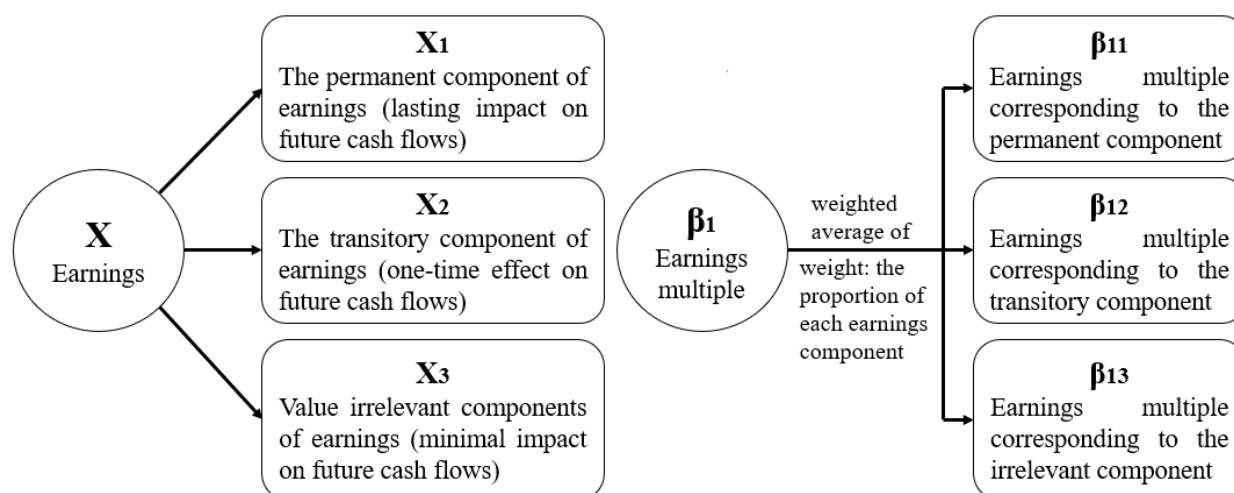
We can use regression to confirm the association between earnings and stock prices. In the equation $P = \beta_0 + \beta_1 X + \varepsilon$, P represents stock price, and X represents earnings per share. β_1 reflects the relation between earnings and stock prices, or earnings multiple.

Figure 4. 2002 P/E relationship for 40 restaurant companies

Earnings multiples often vary significantly across companies, industries, and business cycles. This is because stock price implies the current value of all expected future cash flows (CF_t), which depends on current and future earnings.

$$P = \beta_0 + \beta_{11}X_1 + \beta_{12}X_2 + \beta_{13}X_3 + \varepsilon$$

The permanent component of earnings (X_1) has a lasting impact on future cash flows, while the transitory component of earnings (X_2) only has a one-time effect on future cash flows. Value irrelevant earnings components (X_3) might even have minimal impact on future cash flows. If we regress stock price, P , on the three components of earnings, we will obtain the earnings multiple corresponding to each earnings component. The overall earning multiple, β_1 , is determined by the weight of each earnings component and the earnings multiple corresponding to each earnings component.

Figure 5. Components of earnings

Investors like to see high-quality earnings since these earnings tend to be repeated in future periods and provide more cash flows for investors. Thus, high-quality earnings entities are also more likely to have high stock prices. Conversely, those entities reporting lower-quality earnings will not attract investors, resulting in lower stock prices. For example, in Table 1, both firm A and firm B report earnings of \$10 per share. However, 60 percent of firm A's earnings are permanent, 30 percent are transitory, and only 10 percent are value-irrelevant. In contrast, only 50 percent of firm B's earnings are permanent, 20 percent are transitory, and the remaining 30 percent are value-irrelevant. Since firm A has a larger proportion of permanent components and a smaller proportion of value-irrelevant components than firm B, firm A's earnings quality is deemed better than firm B's. Firm A has an earnings multiple of 3.3, which is higher than firm B's earnings multiple of 2.7.

Table 1. Comparison of high and low earnings quality

	Firm A	Firm B
EPS as reported	\$10	\$10
Analyst's EPS decomposition		
Permanent component	60% of \$10 = \$6	50% of \$10 = \$5
Transitory component	30% of \$10 = \$3	20% of \$10 = \$2
Value-irrelevant component	10% of \$10 = \$1	30% of \$10 = \$3
Earnings multiple applied to each earnings component at cost of capital of $r = 20\%$		
Permanent component ($\beta_p = 5 = 1/.20$)	$5 \times \$6 = 30$	$5 \times \$5 = 25$
Transitory component ($\beta_T = 1$)	$1 \times \$3 = 3$	$1 \times \$2 = 2$
Value-irrelevant component ($\beta_0 = 0$)	$0 \times \$1 = 0$	$0 \times \$3 = 0$
Implied share price	\$33	\$27
Implied total earnings multiple (Share price/EPS as reported)	3.3	2.7

Differences in earnings components mix produces differences in P/E

11.3. Empirical example: Post-earnings announcement drift (PEAD) anomaly



[An empirical example of analyzing the income statement](#)

The post-earnings announcement drift is a phenomenon in which cumulative abnormal returns of stock tend to move in the same direction as the firm's earnings surprise for an extended period of time. The PEAD anomaly refers to the phenomenon whereby portfolios based on information from past earnings earn abnormal returns. Ball and Brown (1986) were the first to note that even after earnings are announced, estimated cumulative abnormal returns continue to drift up for “good news” firms and down for “bad news” firms. Foster, Olsen, and Shevlin (1984) estimate that over the 60 trading days after an earnings announcement, a long position in stocks with unexpected earnings in the highest decile, combined with a short position in stocks in the lowest decile, yields an annualized abnormal return of about 25 percent before transactions costs.

Some explanations for the PEAD anomaly suggests that at least a portion of the price response to new information is delayed. The delay might occur either because traders fail to assimilate

available information or because certain costs (e.g., transaction costs) exceed gains from the immediate exploitation of information for a sufficiently large number of traders. What is less clear is why a price response would be delayed.

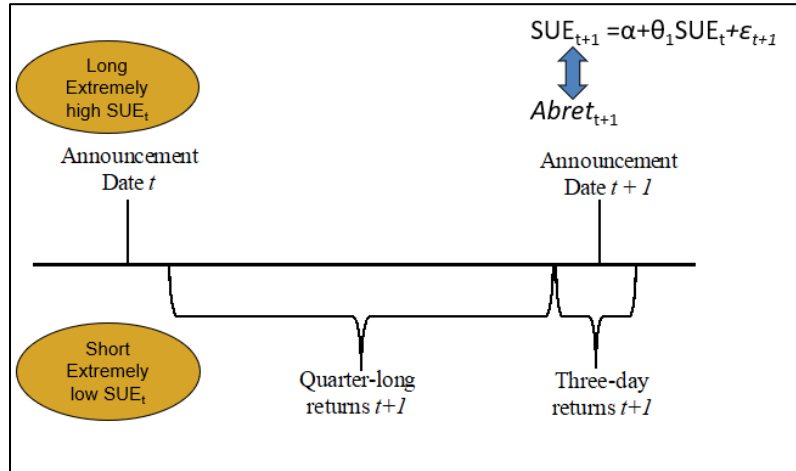
One possibility is that the market erroneously assumes a seasonal random walk for expected earnings and ignores the autocorrelations in earnings. For instance, if a company announces a new long-term contract with a customer that increases earnings at t , this should result in large positive seasonally adjusted unexpected earnings (SUE_t). This contract would also bring in a stream of future earnings, leading to large positive SUE in subsequent years (SUE_{t+1} , SUE_{t+2} , SUE_{t+3} , etc.). The value of this stream of earnings, which is the current value of all future SUEs, should be factored into the stock price at t . If all investors immediately factor the current value of all future SUEs into the stock price at time t , current SUE_t should not be associated with future abnormal returns. If investors fail to adjust stock price expectations immediately, then the current SUE_t should predict future abnormal returns. Specifically, assuming investors delay response to the stream of SUEs from t to $t+1$, current SUE_t should predict abnormal returns at $t+1$. As a result, informed investors can earn abnormal returns by longing stocks with high SUE and shorting stocks with low SUE at time t . This prediction can be tested by estimating the following regression model. If SUE_t predicts abnormal returns at $t+1$, θ_1 should be significantly positive (Figure 6).

$$Abret_{t+1} = \alpha + \theta_1 SUE_t + \varepsilon_{t+1} \quad (1)$$

For example, suppose Apple Inc. signed a contract with a client that yielded a \$1 million profit at time t , and this contract possesses a 40% likelihood of renewal. Hence, the overall estimated value of the contract would be \$1.4 million. If all investors were aware of the 40% renewal probability, the current SUE would have no relation to future returns, as the market has already incorporated all relevant contract information into stock prices on time. However, if investors

undervalued the contract, estimating it at \$1.2 million, the current SUE could predict future stock returns. This is because the remaining \$0.2 million information has not yet been reflected in stock prices at time t . In that case, informed investors can profit by buying stocks with high SUE and selling stocks with low SUE at a given time.

Figure 6. Seasonally adjusted unexpected earnings and announcement returns



Cao and Narayanamoorthy (2012) verify the existence of the PEAD anomaly. Panel A of Table 2 shows that SUE at t predicts future SUEs up to the subsequent three quarters. Panels B and C show the association between current SUE and future three-day earnings announcement abnormal returns up to two quarters and quarter-long abnormal returns up to three quarters. Variable definitions are listed in Appendix 11.

Figure 7 illustrates how, from 1988 to 2008, implementing a trading strategy based on the PEAD anomaly consistently resulted in positive abnormal returns. In practice, this strategy is commonly employed by investors.

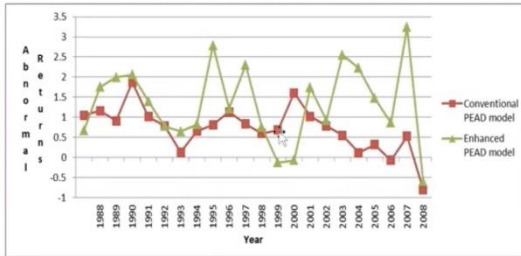
Table 2. SUE persistence and abnormal returns

Panel A: SUE (Dependent Variable $DSUE_{t+k}$)								
	$k = 1$		$k = 2$		$k = 3$		$k = 4$	
	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)
$DSUE_t$	0.365*** (79)	0.40***	0.203*** (57.51)	0.23***	0.066*** (11.56)	0.09***	-0.172*** (-27.58)	-0.17***
$Adj R^2$	0.133*** (39.43)		0.041*** (28.3)		0.005*** (6.06)		0.030*** (12.64)	
	$k = 1$		$k = 2$		$k = 3$		$k = 4$	
Panel B: 3-Day Abnormal Returns (Dependent Variable $AR_{3,t+k}$)								
$DSUE_t$	0.793*** (4.41)		0.261*** (3.17)		-0.231*** (-4.01)		-0.630*** (-10.49)	
$Adj R^2$	0.001*** (3.66)		0.000** (2.08)		0.000*** (2.89)		0.001*** (4.25)	
Panel C: Quarter-Long Abnormal Returns (Dependent Variable $AR_{9,t+k}$)								
$DSUE_t$	6.131*** (15.51)		2.137*** (7.58)		0.697** (2.2)		-0.450 (-1.1)	
$Adj R^2$	0.008*** (6.47)		0.001*** (4.72)		0.000*** (4.27)		0.000** (2.76)	
*** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.								

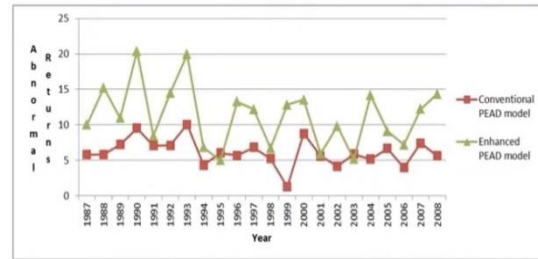
*** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

Figure 7. Trading strategy based on the PEAD anomaly

Three-day abnormal returns



Quarterly abnormal returns



Intuitively, the level of autocorrelation is lower for companies with more volatile earnings. As shown in Table 3, the autocorrelation coefficient is significantly smaller when earnings volatility ($EVOL$) increases.

To examine whether the effect of earnings volatility on the level of SUE autocorrelation passes on to the association between current SUE and future abnormal returns, we can augment the regression model above to include an interaction term between SUE_t and $EVOL_t$.

$$Abret_{t+1} = \alpha + \theta_1 SUE_t + \theta_2 EVOL_t + \theta_3 SUE_t * EVOL_t + \varepsilon_{t+1} \quad (2)$$

Table 3. Effect of earnings volatility on SUE persistence

	Model 1	Model 2	Model 3	Model 4	Model 5
$DSUE_t$	0.379*** (76.87)	0.369*** (76.6)	0.379*** (74.62)	0.395*** (61.95)	0.295*** (53.36)
$EVOL_t \times DSUE_t$	-0.136*** (-10.11)		-0.129*** (-10.25)	-0.100*** (-8.57)	-0.134*** (-10.06)
$EVOL_t$	0.027*** (4.37)		0.027*** (4.23)	0.035*** (6.64)	0.027*** (4.35)
$Size_t \times DSUE_t$		0.058*** (4.75)	0.022* (1.96)		
$Size_t$		-0.005 (-0.8)	0.003 (0.4)		
$Loss_t \times DSUE_t$				-0.092*** (-7.51)	
$Loss_t$				-0.032*** (-7.2)	
$I_t \times DSUE_t$					0.115*** (14.66)
I					0.008*** (4.33)
Adj R^2	0.136*** (40.83)	0.134*** (39.39)	0.137*** (40.71)	0.139*** (39.4)	0.139*** (39.86)

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

The results of this analysis are provided in Table 4. The coefficient of $DSUE$ in Model 1 of Panel A suggests that three-day earnings announcement abnormal returns in three days surrounding the earnings announcement date at $t+1$ is approximately 0.7 percent. Model 1 of Panel B suggests that abnormal stock returns from quarter t to quarter $t+1$ are about 6 percent. The negative coefficient of $EVOL \times DSUE$ suggests that abnormal returns are higher for stocks with less earnings volatility. Models 2 and 3 in both panels control for company size and potential losses at t . The effect of earnings volatility is robust to include both controls.

As noted above, the delay in investor response might occur because transaction costs are prohibitively high. The same regression includes earnings volatility and a proxy for transaction costs, $SPREAD$, to exclude this alternative explanation. If θ_3 continues to be significantly negative, then the effect of earnings volatility at least is not completely attributable to transaction costs.

Table 5 shows that θ_3 remains significantly negative, indicating a significant effect of earnings volatility on the association between SUE and abnormal returns and the effect of transaction costs.

$$Abret_{t+1} = \alpha + \theta_1 SUE_t + \theta_2 EVOL_t + \theta_3 SUE_t * EVOL_t + \theta_4 SPREAD_t + \theta_5 SUE_t * SPREAD_t + \varepsilon_{t+1} \quad (3)$$

Table 4. Effect of earnings volatility on PEAD returns

	Panel A: 3-Day Returns (Dependent Var $AR_{3,t+1}$)			Panel B: Quarterly Returns (Dependent Var $AR_{q,t+1}$)		
	Model 1	Model 2	Model 3	Model 1	Model 2	Model 3
$DSUE_t$	0.711*** (5.39)	0.689*** (4.24)	-0.036 (-0.2)	6.059*** (15.57)	6.393*** (16.08)	3.368*** (6.2)
$EVOL_t \times DSUE_t$	-0.625** (-2.74)	-0.583** (-2.19)	-0.583** (-2.53)	-5.016*** (-7.28)	-4.027*** (-4.95)	-4.819*** (-7.08)
$EVOL_t$	-0.497*** (-3.79)	-0.355*** (-3.2)	-0.499*** (-3.82)	0.302 (0.2)	0.695 (0.56)	0.304 (0.2)
$Size_t$	-0.563*** (-3.12)	-0.628*** (-3.67)	-0.566*** (-3.14)	-2.154* (-2.04)	-2.409** (-2.41)	-2.16* (-2.06)
$Size_t \times DSUE_t$	-2.012*** (-8.52)	-2.115*** (-8.5)	-1.976*** (-8.43)	-7.761*** (-10.39)	-8.391*** (-10.25)	-7.656*** (-9.98)
$Loss_t$		-0.357*** (-4.87)			-1.311** (-2.11)	
$Loss_t \times DSUE_t$		-0.421* (-1.89)			-2.694*** (-2.95)	
I_t			-0.234*** (-3.39)			-0.809 (-1.4)
$I_t \times DSUE_t$			1.017*** (6.24)			3.651*** (5.61)
Adj R^2	0.003*** (8.97)	0.004*** (9.87)	0.004*** (9.54)	0.016*** (6.45)	0.018*** (6.67)	0.019*** (7.21)

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

Table 5. Effect of earnings volatility on PEAD returns controlling for spread

	3-Day Returns (Dependent Var $AR_{3,t+1}$)	Quarterly Returns (Dependent Var $AR_{q,t+1}$)
$DSUE_t$	0.839*** (5.45)	6.865*** (12.48)
$EVOL_t \times DSUE_t$	-0.813** (-2.66)	-5.022*** (-6.7)
$EVOL_t$	-0.504*** (-3.45)	0.530 (0.33)
$Spread$	0.729*** (3.38)	2.151** (2.2)
$Spread \times DSUE_t$	2.036*** (7.23)	5.394*** (5.52)
Adj R^2	0.004*** (8.24)	0.017*** (6.89)

11.4. Machine learning application on income statement data

It is important to understand the patterns underlying income statements in order to make informed investment decisions. However, complexity arises from the substantial number of income-related variables. In general, it is challenging to extract valuable insights from extensive data. So how can we effectively gain insights from the vast amount of data present in income statements?

Item Description	Mnemonic	Field Size	Data Group	Data Type
Global Company Key - Company Annual Descriptor	GVKEY	6	co_adesind	VARCHAR2
Data Date - Company Annual Descriptor	DATADATE	8	co_adesind	DATE
Industry Format	INDFMT	12	co_adesind	VARCHAR2
Data Format - Company Annual Descriptor	DATAFMT	12	co_adesind	VARCHAR2
Level of Consolidation - Company Annual Descriptor	CONSOL	1	co_adesind	VARCHAR2
Population Source	POPSRC	1	co_adesind	VARCHAR2
Global Company Annual Key - Company Annual Fundamentals	GVKEY	6	co_afnd1/co_afnd2	VARCHAR2
Data Date - Company Annual Fundamentals	DATADATE	8	co_afnd1/co_afnd2	DATE
Industry Format - Company Annual Fundamentals	INDFMT	12	co_afnd1/co_afnd2	VARCHAR2
Data Format - Company Annual Fundamentals	DATAFMT	12	co_afnd1/co_afnd2	VARCHAR2
Level of Consolidation - Company Annual Fundamentals	CONSOL	1	co_afnd1/co_afnd2	VARCHAR2
Population Source - Company Annual Fundamentals	POPSRC	1	co_afnd1/co_afnd2	VARCHAR2
Accounting Standard	ACCTSTD	2	co_adesind	VARCHAR2
Acquisition Method	ACQMETH	2	co_adesind	VARCHAR2
ADR Ratio	ADRR	18,4	co_adesind	NUMBER
Adjustment Factor (Company) - Cumulative by Ex-Date	AJEX	24,12	co_adesind	NUMBER
Adjustment Factor (Company) - Cumulative byPay-Date	AJP	24,12	co_adesind	NUMBER
Actual Period End date	APDEDATE	8	co_adesind	DATE
Balance Sheet Presentation	BSPR	2	co_adesind	VARCHAR2
Comparability Status	COMPST	2	co_adesind	VARCHAR2
ISO Currency Code	CURCD	3	co_adesind	VARCHAR2
Native Currency Code	CURNCD	3	co_adesind	VARCHAR2
Currency Translation Rate	CURRTR	24,12	co_adesind	NUMBER
US Canadian Translation Rate	CURUSCN	18,4	co_adesind	NUMBER
Final Date	FDATE	8	co_adesind	DATE
Fiscal Year-end Month	FYR	2	co_adesind	NUMBER
Income Statement Model Number	ISMOD	4	co_adesind	NUMBER
OIL & GAS METHOD	OGM	8	co_adesind	VARCHAR2
Preliminary Date	PDATE	8	co_adesind	DATE
Period Duration	PDDUR	2	co_adesind	NUMBER
Cash Flow Format	SCF	4	co_adesind	NUMBER
Source Document	SRC	4	co_adesind	NUMBER
Status Alert	STALT	2	co_adesind	VARCHAR2
Utility - Liberalized Depreciation Code	UDPL	8	co_adesind	VARCHAR2
Update Code	UPD	4	co_adesind	NUMBER
Inventory Valuation Method	INVVAL	4	co_ainvval	NUMBER

There are two primary approaches collecting data from the broad range of income-related variables present in income statements. The first approach is theory-based, wherein we first

develop a hypothesis based on established theories and test its validity. By doing so, we can determine whether the patterns suggested by the theory can provide valuable guidance to investors.

Another approach is the application of machine learning techniques, which are particularly suited for handling high-dimensional data. In this approach, all income-related variables are incorporated into machine learning models. The machine learning models processes the data, continuously learning and adapting to identify patterns and relations between the variables and subsequent stock returns. The advantage of using machine learning techniques lies in their ability to handle vast amounts of data and identify complex patterns that might not be apparent through traditional analysis methods. Specifically, machine learning algorithms have the capacity to analyze the interactions among multiple income-related variables, thereby discovering hidden patterns and providing valuable insights.

Furthermore, income statement information could be used with other alternative data to generate more insights. Dichev and Qian (2022) find that NielsenIQ scanner data contain incremental information about manufacturers' revenue. Investors using scanner data thus could obtain an information edge over those only relying on income statement information.

Appendix 11. Key variable explanations***A11.1. Abnormal returns***

Abnormal returns, also known as “excess returns,” refers to the unanticipated profits (or losses) generated by a security/stock. Abnormal returns are measured as the difference between the actual returns that investors earn on an asset and the expected returns. Expected returns are estimated using, for example, the CAPM model. Abnormal returns can be positive or negative. Positive abnormal returns are realized when actual returns are greater than expected returns. Negative abnormal returns (or losses) occur when the actual return is lower than expected.

$$\text{Abnormal return} = \text{Actual return} - \text{Expected return}$$

For instance, in Cao and Narayanamoorthy (2012), $AR_{q, t+1}$ is calculated by subtracting the CRSP value-weighted index from the raw return for the period between two days after the quarter earnings announcement date and one day before the next announcement date. $AR_{s, t+1}$ is calculated by subtracting the CRSP value-weighted index from the raw return during the three-day window $(-1, +1)$ around quarter $t+1$'s announcement.

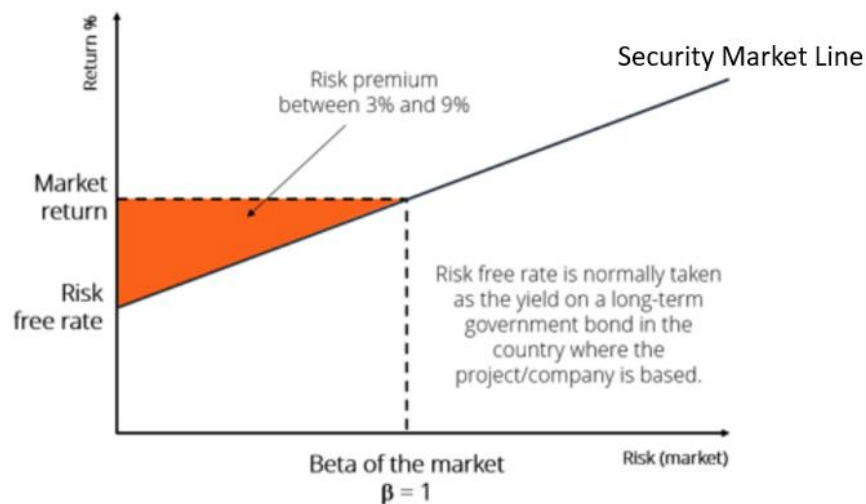
A11.2. Expected return

The expected return is equal to the risk-free return plus a risk premium. It is based on the idea of systematic risk (otherwise known as non-diversifiable risk), which dictates that investors need to be compensated for in the form of a risk premium. A risk premium is a rate of return greater than the risk-free rate. Investors typically look for a higher risk premium when taking on more risky investments.

$$\text{Expected return} = \text{Risk-free rate} + \beta * (\text{Market return} - \text{Risk-free rate})$$

The risk premium investors require is based on the β of that stock. β is a measure of a stock's risk reflected by measuring the fluctuation of its price changes relative to the overall market. In other words, it is the stock's sensitivity to market risk. For instance, if a company's β equals one, the expected return on a stock equals the average market return. A β of -1 means a stock has a perfect negative correlation with the market. Cumulative Abnormal Returns (CAR) are the sum of abnormal returns over a given period of time. It allows investors to measure the performance of an asset or security over a specific period of time, especially since abnormal returns over short windows tend to be biased.

Figure A1. Capital Asset Pricing Model



The Security Market Line (SML) graphically represents the relationship between expected returns and the associated risk levels (Figure A1). A security or portfolio that is in equilibrium lies on the SML, indicating that it is fairly priced, as its expected return equals the return required by the market at that level of risk. Assets lying above the SML are undervalued, as they offer a higher return than required for their risk level. Conversely, assets lying below the SML are overvalued, as they offer a lower return than required for their risk level.

A11.3. Standardized unexpected earnings

Standardized Unexpected Earnings (*SUE*) is calculated by taking the difference between the current earnings and the earnings from the same quarter in the previous year, divided by the closing market value of the preceding fiscal quarter.

DSUE is the *SUE* decile rank for each quarter transformed by dividing the rank by 9 and subtracting 0.5, resulting in values that range from -0.5 to $+0.5$.

A11.4. Earnings volatility

Earnings volatility (*VOL*) is the variance of the most recent eight quarterly earnings (including quarter *t*), scaled by average total assets.

EVOL is the earnings volatility (*VOL*) decile rank for each quarter transformed by dividing the rank by 9 and subtracting 0.5, resulting in values that range from -0.5 to $+0.5$.

References

- Ball, R., and Brown, P. 1968. An empirical evaluation of accounting income numbers. *Journal of Accounting Research*, 6(2), 159-178.
- Cao, S., and Narayanamoorthy, G. 2012. Earnings volatility, post-earnings announcement drift, and trading frictions, *Journal of Accounting Research*, 50(1), 41-74.
- Dichev, I., and Qian, J. 2022. The benefits of transaction-level data: The case of NielsenIQ scanner data. *Journal of Accounting and Economics*, 74(1), 101495.
- Foster, G., Olsen, C. and Shevlin, T. 1984. Earnings releases, anomalies, and the behavior of security returns. *The Accounting Review*, 59(4), 574-603.