Financial Innovation

REVIEW Open Access

Alternative data in finance and business: emerging applications and theory analysis (review)

Yunchuan Sun[1*], Lu Liu[1], Ying Xu[1], Xiaoping Zeng[1], Yufeng Shi[2], Haifeng Hu[3], Jie Jiang[3] and Ajith Abraham[4]

*Correspondence:
yunch@bnu.edu.cn

[1] International Institute of Big Data in Finance, Business School, Beijing Normal University, Beijing 100875, China
[2] Institute for Financial Studies, Shandong University, Jinan 250100, China
[3] Business School, Beijing Normal University, Beijing 100875, China
[4] Machine Intelligence Research Labs (MIR Labs), Scientific Network for Innovation and Research Excellence Auburn, Auburn, Washington 98071, USA

Abstract

In the financial sector, alternatives to traditional datasets, such as financial statements and Securities and Exchange Commission filings, can provide additional ways to describe the running status of businesses. Nontraditional data sources include individual behaviors, business processes, and various sensors. In recent years, alternative data have been leveraged by businesses and investors to adjust credit scores, mitigate financial fraud, and optimize investment portfolios because they can be used to conduct more in-depth, comprehensive, and timely evaluations of enterprises. Adopting alternative data in developing models for finance and business scenarios has become increasingly popular in academia. In this article, we first identify the advantages of alternative data compared with traditional data, such as having multiple sources, heterogeneity, flexibility, objectivity, and constant evolution. We then provide an overall investigation of emerging studies to outline the various types, emerging applications, and effects of alternative data in finance and business by reviewing over 100 papers published from 2015 to 2023. The investigation is implemented according to application scenarios, including business return prediction, business risk management, credit evaluation, investment risk prediction, and stock prediction. We discuss the roles of alternative data from the perspective of finance theory to argue that alternative data have the potential to serve as a bridge toward achieving high efficiency in financial markets. The challenges and future trends of alternative data in finance and business are also discussed.

Keywords: Alternative data, Behavioral data, Commercial data, Credit evaluation, Enterprise management, Finance innovation, Investment, Market efficiency, Price prediction, Risk evaluation, Sensing data

Introduction

The rapid development of data technologies, including data capturing, storage, management, and analysis, makes it possible to obtain a vast amount of data from different sources, such as individual behaviors, business processes, or various sensors. This provides alternative ways to understand the running states of enterprises, different from traditional data of financial statements (Froot et al. 2017; Bartov et al. 2018; Tang 2018). This type of data, called alternative data, are generally obtained unofficially but can

reflect business operations and can be used to investigate financial markets and business processes instead of traditional data. There are various sources of alternative data, such as social networks, online shopping, satellite imagery, and the Internet of Things. Although there is some noise from alternative data (Gholampour 2019; Goldstein and Yang 2019; Chen et al. 2020), many studies have demonstrated that alternative data benefit investment decisions and business management (Bernile et al. 2017; Cohen et al. 2020; Gao et al. 2021a, b).

Alternative data have gradually demonstrated their commercial value and prospects for practice and academic research. According to the prominent data provider Eagle Alpha, in practice, alternative data will continue to hold significant value because of their innovative nature and ability to tap into fresh and unique data sources.[1] This study also demonstrates that alternative data can be applied in different sectors, including healthcare, industry, and real estate, especially in the financial industries. The New York Times[2] reported that alternative data can help people without a credit history evaluate their creditworthiness. When used by a real company, such as TransUnion, alternative data help approve more than 20% of applicants. In academia, alternative data have been widely used for predicting stock prices and informing business decisions, encompassing theoretical domains such as traditional, behavioral, and cultural finance, as well as management. Alternative data have expanded the horizons of finance studies, aiding researchers in obtaining a comprehensive view of market dynamics and company status.

Focusing on investment, it has been shown that alternative data could refine revenue prediction accuracy, reducing the mean absolute error from 88 to 2.6% (Ekster and Kolm 2021). Challenges are also highlighted, such as the complexity of estimating a dataset's potential value and the presence of technical obstacles. Another survey provides more applications for using alternative data in practice and academics, but primarily focuses on research and cases before 2020 (Charoenwong and Kwan 2021). Given the rapid evolution of data technology, there is a palpable need for a survey to capture the advanced developments in this emerging field and to present the value of alternative data in finance and business. This article aims to answer the following questions: (1) What are the advantages of alternative data compared to traditional financial data? (2) What are the applications of different types of alternative data in diverse financial scenarios? (3) What is the impact of alternative data on market efficiency through the lens of financial theories? and (4) What are the future challenges and trends of alternative data applications?

To guarantee the comprehensiveness of our survey, we examined all articles published from 2015 to 2023 in highly ranked journals based on the 2018 Academic Journal Guide (AJG) ranking. The Academic Journal Guide (AJG) is from the Association of Business Schools, an influential guide that evaluates and ranks scholarly journals in business and finance. We included the Journal of Finance, Journal of Financial Economics, and the Review of Financial Studies. We also incorporated some exemplary articles that may not

[1] Eagle Alpha. (2020). Alternative Data Primer and 10 Thematic Case Studies for Investors.https://www.eaglealpha.com/2020/10/27/alternative-data-primer-and-10-thematic-case-studies-for-investors/

[2] Carrns, A. (2017). Little Credict History? Lenders Are Taking a New Look at You. New York Times. https://www.nytimes.com/2017/02/24/your-money/26money-adviser-credit-scores.html

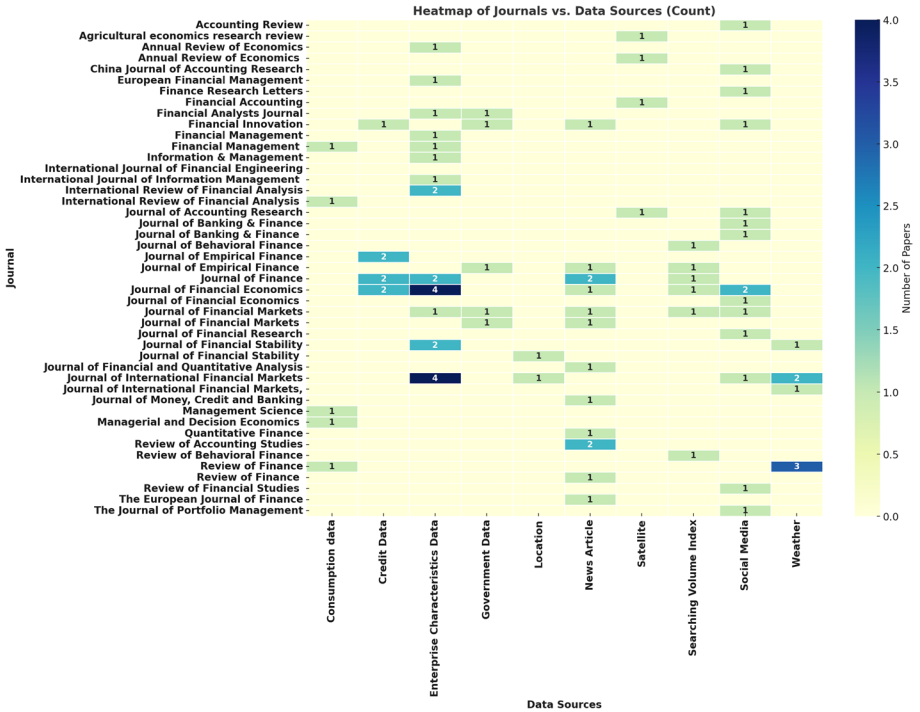

Fig. 1 Research distribution of journals from different data sources

be from high-level journals but are great examples of using alternative data. We collected approximately 100 relevant articles for our survey. More details are shown in Fig. 1.

These papers are classified along two dimensions: data sources and applications. We distinguish data sources as data from user behavior, business processes, and sensors. The applications of these alternative data include business returns, business risk management, credit evaluation, investment risk management, market prediction, and stock prediction. Figure 2 illustrates a literature overview of our classification from alternative data sources and applications, where the x-axis represents the types of applications and the y-axis shows the types of data sources. The numbers within the graph indicate the percentages of data used for each application. For instance, the first row of Fig. 2 shows that among all consumption data articles, 40% use data for business returns, 20% for business risk management, and 40% for stock prediction.

This survey provides a detailed overview of various types of alternative data and examines their specific applications in finance and business. We explore the significance of alternative data within the framework of financial theories and point out that alternative data are a bridge that helps the market become more efficient. Extant studies related to alternative data mainly focus on the specific applications of alternative data, while there is little analysis of the potential impact of alternative data on market efficiency through the lens of financial theories. Our theoretical analysis reveals that alternative data have the potential to address information asymmetry, reduce agency costs, enhance investor rationality, and mitigate limitations to arbitrage, thus improving market efficiency. Furthermore, we discuss the practical challenges and trends associated with the real-life use of alternative data, offering in-depth descriptions of the obstacles and opportunities

Fig. 2 Research distribution of alternative data in different applications

for researchers and practitioners in the field. Overall, this paper enriches the comprehension of the progress and prospective developments in financial research arising from applying alternative data.

The rest of the paper is organized as follows. Section "Alternative data versus traditional data" highlights the fundamental differences between alternative and traditional data and offers five major advantages of alternative data. Section ``Alternative data categories" categorizes alternative data into three categories and 10 subcategories and summarizes their features, such as publicity, cost, updated frequency, and providers. Section ``Financial applications of alternative data" details different alternative data types and their corresponding finance applications. Section ``Roles of alternative data from financial theoretical insights: a bridge leading the market to be more efficient" analyzes how alternative data enhance market efficiency through the lens of financial theories. Section "Discussion of challenges and future trends" discusses the challenges encountered in alternative data applications and provides a visionary outlook on the future of alternative data. Section "Conclusions" concludes the paper.

Alternative data versus traditional data

In this section, we explore the differences between alternative and traditional data and highlight the unique advantages of alternative data.

Traditional data are officially distributed to convey the comprehensive condition of an enterprise. They are periodically issued by financial institutions, such as banks, insurance, Internet finance, options, security, precious metals, etc., or by enterprises through financial statements and announcements. Benefiting from its authority, governments,

managers, investors, and academic experts rely on traditional financial data to develop policies, make decisions, and predict economic trends. Many well-known financial models, such as the Fama–French five-factor model, are built on exploiting traditional data (Fama and French 2015).

Financial data provide insights into what has happened, what is happening, and what will happen in the business world. They have been used to develop most financial models in academia and industry. However, there are several problems with traditional financial data. First, information released by enterprises is often incomplete due to proprietary requirements. Firms like to issue confidential data in ambiguous or subjective ways, which they are required to open; however, using this type of data makes it difficult to determine exactly what has happened to the company. Some companies leverage these circumstances to commit fraud. For example, Kangde Xin (Stock No: 002450), an A-share listed company in China, committed fraud by concealing its actual financial status, and Luckin Coffee (Nasdaq: LK) fabricated financial information after its IPO. In addition to earnings manipulation, some companies choose to disclose misleading indicators, such as EBITDA,[3] which could overestimate company performance (Fridson 1993). Second, official reports are always issued periodically (e.g., monthly CPI, retail sales reports, quarterly earnings, etc.). However, quarterly or annual reports are generally insufficient and not timely enough for government, business, and investor decision-making.

Fortunately, emerging alternative data can help address these limitations. Unlike traditional data, alternative data are gathered from nontraditional sources, mainly individual behaviors (media posts, product reviews, search trends, etc.), business processes (company exhaust data, commercial transaction, credit card data, etc.), and a variety of sensors (satellite image data, foot, car traffic, ship locations, etc.). Increasingly, alternative data can be captured electronically and in a timely manner through devices connected to the Internet. This, in principle, allows investors to access a broad range of market- and enterprise-related information in real time (Jagtiani and Lemieux 2019; Loder 2019). It also helps disclose the actual running status of the company or the reality of macroeconomics in a timelier manner than traditional data. For instance, millions of prices of products on Amazon and JD.com, gathered by web crawlers, can be used to assess inflation and, thus, are an alternative to the official monthly CPI; online analysis on customer traffic and transactions can provide estimations of real-time sales; and satellite images are suitable for evaluating agricultural production or economic conditions. Given alternative data's special and real-time nature, skilled quantitative investors are able to access additional information that cannot be obtained from traditional data sources.

In financial markets, both alternative and traditional data can reflect the running status of a company. While traditional data are more authoritative due to their official nature, alternative data provide real-time information, enabling managers and researchers to explore a more comprehensive understanding of financial markets and aiding in identifying emerging opportunities and risks. Alternative data have the following five advantages.

[3] Abbreviation for "Earnings Before Depreciation, Interest, Tax and Amortization", used to calculate company performance, with high degree.

(1) Multiple sources

The previous section summarized the most-used alternative data into 10 categories based on the literature. Due to the development of data science, the scope of alternative data ranges from textual content in online news articles, comments, and reviews to visual data captured through images and videos from unmanned aerial vehicles and cameras. We anticipate that more alternative data sources will be identified in the future. The multiple sources of alternative data improve the horizon of financial studies, enabling more robust conclusions from different angles.

(2) Heterogeneity

The multisourcing feature leads to heterogeneity of the data. Traditional data, such as numbers, texts, and dates, are usually structured, organized, formatted, and easy to store with the support of mature database systems, whereas alternative data encompasses not only structured but also semistructured and unstructured formats, including pictures, sentences, videos, or audio. The heterogeneity of alternative data reduces the possibility of manipulation, allowing them to exhibit real economic conditions more accurately.

(3) Flexibility

The improvement of cloud computing allows alternative data to be updated and retrieved within seconds. This rapid processing will enable data analysis from various online platforms, facilitating the examination of several topics, both retrospectively and in real time. The flexibility of alternative data means that investors and researchers no longer rely solely on traditional sources, such as periodic, annual, or monthly reports. Instead, they can access diverse and timely information, ranging from consumer sentiment captured through social media to transactional data and geospatial information, which provide a more dynamic and comprehensive view of the market, opening up new possibilities for informed decision-making and innovative research.

(4) Objectiveness

Alternative data are independent of any specific firm's purpose or influence because it is publicly generated, unfiltered, and continuously updated from diverse sources. It is a credible and objective research and operational source in finance and business.

(5) Constant evolution

From website browsing data to remote sensing capturing images, the emergence of current types of alternative data is driven by the advancement of data technologies. As technology progresses, we can anticipate uncovering more novel alternative data in the future. The constantly evolving alternative data enhance existing analytical approaches and open new avenues for innovation in business and finance.

Table 1 Alternative data types, features, users, and providers

Types	Subcategories	Features			Providers	Related research
		Publicity	Cost	Updated frequency		
Behavioral data	Social media	High	Low	Seconds	Facebook, Tweeter, Stocktwits.com, Yelp	Ge et al. (2019)
	New articles	High	Low	Daily	Online News Portals	Obaid and Pukthuanthong (2022)
	Searching volume index	High	Low	Seconds	Google, Baidu	Vasileiou and Tzanakis (2022)
Commercial data	Consumption data	Low	High	Seconds	Retailers, E-Commerce such as Amazon.com and Taobao.com	Niu et al. (2023)
	Credit data	Low	High	Monthly	Credit Bureaus, Financial Institution, Peer-to-Peer (P2P), Lending Platforms	Rozo et al. (2023)
	Enterprise characteristics data	Medium	Medium	Monthly	Job Posting such as Linkedin, Employee Reviews	Duréndez et al. (2023)
	Government data	High	Low	Monthly	Government Bureaus, Public Records and Databases	Gupta and Pierdzioch (2023)
Sensing Data	Satellite data	Low	High	Daily	Data vendors	Subash et al. (2018)
	Location data	High	Medium	Seconds	Financial Institutions, Data vendors	Geng et al. (2023)
	Weather data	High	Medium	Hourly	National Weather, Service (NWS), Open Weather	Gregory (2021)

Alternative data categories

In this section, we categorize commonly used alternative data into three main types: behavioral, commercial, and sensor-based, further delineating these into 10 subcategories. We have summarized them by features, providers, and related research (Table 1).

Behavioral data are from users' actions, habits, and interactions on various digital platforms, showing their attention and sentiments. They are short-term, real-time, and low-cost. This article divides behavioral data into three types according to sources: (1) data from social media (websites like Twitter, Facebook, Yelp, etc.), (2) data from news articles (articles from WTO, The New Times, The Wall Street Journal, etc.), and (3) Searching Volume Index (The number of search terms). These data are mostly correlated with fluctuations in stock prices and companies' future returns.

Commercial data are from business operations and processes. They are separated into four types. Consumption data are derived from buying or selling activities through

online payment or retail markets. When a product is sold, its transaction records, such as price, purchase time, number of items, purchase location, and even logistics, can all be considered consumption data, which could directly reflect enterprise sales performance. Credit data are obtained from credit bureaus. They are beneficial for banks or institutions in determining the credit scores of firms and individuals to reduce loan risks. Enterprise characteristic data are sourced from internal aspects of the company, such as culture, employee satisfaction, psychological factors, cultural background, and social responsibility. For example, CEO experience would affect the CEO's investment decisions (Bernile et al. 2017). Similarly, companies with high innovation costs are more likely to be noticed by investors (Francis et al. 2012). Enterprise characteristic data serve as valuable metrics for internal performance optimization and risk mitigation tools for investment strategies. Government data are macroscopic information for the country, such as policy, demographic census, and government statements, which greatly impact firms, stock prices, and markets. They provide a comprehensive view of the prevailing socioeconomic situation, implying the potential opportunities and risks in the current market.

Sensing data are collected from the sensors or cameras of satellites, mobile applications, or smart applications of the Internet of Things. Three types of sensing data are primarily used in the financial sector: satellite, location, and weather data. Satellite data include images collected from cameras set on the orbital spacecraft, which display the surface status of the Earth. Scholars use satellite data to track human activities and predict future trends in the retail market or factory and agricultural production. Location data display the geolocation of firms or users from WIFI signals and Bluetooth on mobile phones. The long distance between firms and institutions could cause information asymmetry (Mazur and Salganik-Shoshan 2017). Weather data are accumulated from specific instruments, such as thermometers or radars. The data are multifaceted, including temperature, humidity, perceived level, season, climate, and air quality. Weather can affect investors' sentiment levels and investment decisions (Kliger and Levy 2003).

Each type of alternative data contributes unique information, facilitating a comprehensive understanding of enterprises and markets from different perspectives. Behavioral data, for example, offer an immediate view of people's attitudes and preferences. Commercial data display a real-time snapshot of market conditions. Enterprise characteristic data reveal the internal management and condition of businesses. With its authoritative and comprehensive nature, government data encompass a wide range of sectors and offer insights into policy impacts and macroeconomic trends. Sensing data provide a macroscopic view of the environmental and geographical impacts on markets and operations. As we move further into the data-driven age, alternative data's dominance and ethical use will undoubtedly become pivotal in shaping business strategies, informing policy, and guiding investment decisions.

Financial applications of alternative data

Behavioral data

Behavior data have a public focus and provide viewpoints on different topics, products, stocks, and businesses, symbolizing the possible behaviors of consumers and investors. They can predict future sales, forecast stock prices with stock ranking, mitigate

management risks by monitoring enterprises' actions, reflecting market trends with public attention. Based on recent research, we divide behavioral data into social media, news, and search volume index (SVI). Data from social media are real-time, large-scale, global, and flexible, making them invaluable for tracking trends in stock prices and company returns, emerging sentiments, and real-time reactions to events. In contrast, news data are considered credible and professional, offering an objective view of market conditions and enterprise situations. The SVI quantitatively measures public interest and sentiments toward specific topics, helping investors grasp current market trends and forecasting future market dynamics. This section illustrates how to use social media data, news data, and the SVI to predict financial performance and improve the business decision-making process.

Social media data

The information exchange revolution enables people to freely share their views with a broad audience through social networks. Data from social media are large-scale, independent, and efficient (Bartov et al. 2018) and can be used to forecast stock prices, reduce management risks, and evaluate market values due to their popularity, internationalization, and open source nature (Rizkiana et al. 2017). Here, we discuss the top three most-used types of social media: connection-based, interest-based, and review-based.

Connection-based networks are designed for all users to openly share their thoughts and establish online connections without constraints on content or format, resulting in a broader user base and a richer dataset than other platforms, thereby providing a comprehensive representation of public opinion. These networks are primarily used to forecast stock prices through sentiment analysis of companies and to assess company returns by examining social media posts and connections.

First, investors and researchers employ connection-based networks to predict short-term price movements by analyzing the volume and sentiment of companies. For example, a study used lexicon-based sentiment analysis of Twitter discussions on financial products combined with bilateral Granger causality to predict the returns of the top cryptocurrencies (Kraaijeveld and De Smedt 2020). After analyzing tweets from President Trump's official Twitter accounts, it was found that the stock prices, trading volume, and institutional investor attention of firms mentioned by Donald Trump would increase (Ge et al. 2019). When connection-based networks are available in different countries, researchers apply the data to examine how sentiment influences stock markets differently in each nation. According to the investigation of Facebook, which is accessible in more than 20 countries, the US has the highest coefficient value when comparing the relationship between sentiment divergence and stock volatility. If the divergent sentiment score rises to 0.01, the trade volume will increase to 0.05 standard deviation, which is approximately 100 million daily shares (Siganos et al. 2017).

Second, connection-based networks are used to predict a company's profitability. They have an advanced recommendation engine and millions of users, attracting companies and CEOs to publish their activities and product information on Twitter to increase brand awareness and potential customers. Some firms adjust Facebook posts to increase attention to earnings news (Hasan and Cready 2019). By using text mining to recognize CEO's personalities based on their posts on Facebook and Twitter, it shows that these

Table 2 Investor groups on different opinions

	Agree before announcement	Disagree before announcement
Agree after announcement	AA (Benchmark)	DA (Convergence)
Disagree after announcement	AD (Divergence)	DD

posts affect cost efficiency, profitability, and employee productivity (Wang and Chen 2020). In addition, the aggregate opinions about companies' prospects on Twitter could positively correlate with the company's forthcoming quarterly earnings and abnormal stock price reaction (Bartov et al. 2018).

Third, connection-based networks provide a unique opportunity to investigate social ties between stakeholders and companies based on their interactions, such as "like," "follow," or "comment," which could be applied to project company performance. Previously, many studies relied on offline data sources. For example, a study used employee stock purchase plans and found that workplace networks played a role in employees making similar financial decisions (Ouimet and Tate 2019). With connection-based network data's availability, we now have a valuable resource for mapping online social connections among various stakeholders, including customers, employees, investors, and neighboring communities (Grover et al. 2019). These data enable real-time monitoring of dynamic interactions within online social networks among companies and assess their management effectiveness. A study used data from LinkedIn interactions and interviews with professionals finds that active participation in social media networks can lead to valuable business relationships, new business development, and improved business performance (Quinton and Wilson 2016).

Interest-based networks, which are focused on specific topics such as games, music, or finance, attract users who are passionate and knowledgeable about their specific fields, leading to conversations that are more concentrated and valuable than those on common social media. In the financial sector, interest-based networks for investors, such as Stocktwits.com or Scutify.com, allow users to discuss stocks, funds, and futures. Mostly investors, professional analysts, and industry insiders utilize these networks, and their discussions and insights offer valuable perspectives on market trends and stock performance.

Many scholars use interest-based network conversations to predict stock prices and investment risks (Bartov et al. 2018; Sun et al. 2021a, b). For instance, after collecting 487,193 specific discussion posts from CMoney, it was found that stock rumors could predict abnormal trading behavior (Cheng et al. 2023). More specifically, we can analyze how, when, and which investor discussions impact investment decisions and stock prices. There is a great example of using data from Stocktwits.com, which investigates whether investors' changed attitudes increase trading volume around announcement day. It compares four groups based on whether investors agree or disagree before or after the announcement, as shown in Table 2.

In Table 2, AA means Agree before Announcement and Agree after Announcement. It sets AA as the benchmark, AD as Divergence, and DA as convergence. When investors agree before the announcement (AD), the average abnormal return is 0.33%; in contrast,

when investors disagree (DA), their average return is −0.05%. The conclusion is that investors with divergent opinions are associated with higher earnings, and investors with convergence of opinions are associated with lower earnings (Giannini et al. 2019). Another example is Seeking Alpha, one of the most influential social websites for investors, which is used to determine how different comments affect the price by analyzing their number of views, whether read-to-end or different tones (Chen et al. 2014).

Review-based networks are online platforms designed for users to evaluate and rank products, services, or businesses. They usually provide an anonymous mode to reduce the effect of herding behavior and peer pressure, ensuring that results are close to reality and trustworthy. Such specificity makes researchers and investors more efficient in predicting future sales of certain products, the performance of individual firms, and the price of particular stocks.

First, studies demonstrate that review-based networks affect sales volume and customer attention. In the Apple Store, apps with positive comments or appearing in best-seller lists will be sold more, whereas those with negative comments will be sold less (Sorensen 2017).

Second, review-based networks can predict the future returns and risks of companies. For instance, online reviews of 297,933 employees from 11,975 US tourism and hospitality firms in Glassdoor demonstrate that an increase in job satisfaction decreases the chances of an employee leaving the company by approximately 14.87% and correlates with a 1.2–1.4 boost in return on assets (Stamolampros et al. 2019).

Third, review-based networks mirror investor expectations and predict future stock prices. In the Forcerank app, when investors do not have access to other participants' choices or the current rankings, their expectations for future stock performance are influenced by past stock returns rather than by systematic or fundamental analysis, especially in cases of past negative returns. Moreover, such rankings based on past returns can predict stock prices for the following week (Da et al. 2021). An analysis of 14.5 million reviews on Amazon indicates that abnormal reviews positively predict stock returns and earnings of firms (Huang 2018).

News data

News data are from articles or reports published by third-party professionals, and the high reputation of the publishers and authors makes news credible and objective to investors and companies (Hillert et al. 2018). Additionally, news is published more frequently and contains a wider range of topics than financial reports, such as business news (staff reduction plan, merge and acquisition, product launches), economic data (like import and export, unemployment rates, fertility), and policy decisions (such as interest rate changes or industry adjustment). Numerous studies have shown that news data influence stock prices by altering investor behavior, mitigating potential management risks by monitoring companies.

First, news content can be an essential factor in stock return prediction (Gherghina and Simionescu 2023). For example, a Fear Index based on WHO's disease-related articles (DRN) demonstrates that negative sentiment positively impacts pharmaceutical companies following DRNs (Donadelli et al. 2017). Other than diseases, radical news on matters such as policies, significant events, or wars can have a similar effect (Kaplanski

and Levy 2010, 2012; Horváth and Huizinga 2011). After studying the percentage of Photo Pessimism (news photos having negative sentiment) with CNN, it was revealed that Photo Pessimism is inversely correlated with the next day's market returns and positively correlated with the following trading week's market returns. Specifically, the average impact of a one standard deviation shift in sentiment derived from article photos on the next day's market return is 4.2% (Obaid and Pukthuanthong 2022). Furthermore, different types of news with various exposure rates and transmission speeds have diverse impacts, such as geopolitical news, which is more valuable than economic news when predicting equity premiums (Adämmer and Schüssler 2020). A higher exposure rate to news and faster news transmission attract greater investor attention and lead to more robust stock returns (Hirshleifer et al. 2009; Tao et al. 2020).

Moreover, news could act as external monitoring for managers to construct the right investment strategies. Managers may make overinvestment or underinvestment decisions due to risk aversion, information asymmetry, cost constraints, or overconfidence. News could foster a celebrity culture, boost CEO confidence, and increase spending on capital, research and development (R&D), and acquisition to solve underinvestments (Gao et al. 2021a, b).

Searching volume index

The SVI provides a timely, cost-efficient, and globally quantitative measurement of public interest and sentiment. When investors lose an information resource such as Google, it increases stock price crash risk due to information asymmetry (Xu et al. 2021). A study found that Google covers 90% of the market search volume, providing researchers and investors with less subjective information freely and instantly (Chiu et al. 2020). This causes many investors to choose to search for their stocks on search engines. Thus, many studies have realized the potential value of the SVI in predicting stock prices. The SVI[4] is defined as

$$SVI_t^j = \frac{Searches_t^j}{TotalSearches_t \times Constant^j},$$

$Searches_t^j$ represents the number of searches for term j in period t. $TotalSearches_t$ is the number of total searches in period t and $Constant^j$ is a scaling constant (Reyes 2018).

The SVI significantly affects investors' investment decisions. Higher stock searching volume induces investors to make similar decisions on the stock market, causing herding behavior (Nofsinger and Sias 1999). An analysis of Google search volume from 2018 to 2021 with Wavelet Coherence Analysis showed that the searches lead to AMC returns, and a high AMC stock price increases the number of searches (Vasileiou and Tzanakis 2022). Also, the rising Google search volume of a firm's most popular product robustly predicts a positive surprise return around the time of the announcements (Da et al. 2011).

In addition to stocks, the SVI of other content can indirectly change investor behaviors. An analysis of Google search terms found that investors will trade out risky assets

[4] Google Trends is available at http://www.google.com/trends/

when the fear level is high (Kostopoulos et al. 2020). Similarly, during the COVID-19 pandemic, the increasing number of COVID-19 searches spreads fear sentiment, causing a strong negative impact on the stock market (Subramaniam and Chakraborty 2021).

Commercial data

Commercial data, including consumption, enterprise characteristics, credit, and government data, can reflect or affect the economic activities of enterprises and identify uncertainties in business activities. Commercial data can be used to anticipate sales and returns of companies, mitigate investment risks, improve business management, reduce loan risks, and predict market trends. Specifically, consumption data reflect enterprise revenue and can be used to hedge investment risk; enterprise characteristics data focus on exposing the quality of senior managers, corporate culture, employee management, and other internal conditions of enterprises; credit data optimize the accuracy of the current credit risk assessment system; and government data disclose systemic risks. This section elaborates on the specific relationship between commercial data and business returns.

Consumption data

Consumption data are collected from consumers' shopping behaviors and provide a wealth of information, including product prices, purchase volumes, geographical locations, and payment methods. For companies that rely heavily on consumer spending, such as retailers, consumption data and their corresponding statistics, such as anomalies in purchasing patterns, purchase frequency, and customer segmentation (e.g., based on demographics, geography, or shopping habits), serve as powerful predictors of sales, enabling investors and researchers to anticipate excess returns and mitigate investment risks.

First, consumption data, such as real-time quarterly retail sales, have strong predictive power for unexpected revenue and excess returns (Froot et al. 2017). The availability of e-commerce sales data can provide analysts with more accurate enterprise value evaluation and earnings forecasts (Niu et al. 2023). Transaction-level credit card expending data have reached similar conclusions. After accounting for unexpected changes in earnings and sales, for every 20% increase in a company's adjusted customer spending, there is a corresponding 1.5% increase in cumulative abnormal returns observed 60 days following the release of the company's financial report (Agarwal et al. 2021).

Additionally, consumption data allow researchers and investors to analyze in-depth which type of customers produce more significant company returns, and segment them into categories based on spending capability and loyalty levels. These characteristics serve as indicators for forecasting the sustainability of future demand for a company's offerings. The most significant returns are achieved from customers with high FICO scores, substantial liquidity, and intense loyalty (Agarwal et al. 2021).

Second, consumption data availability could reduce information asymmetry and minimize potential investment risks. Leveraging online sales data from major e-commerce sites, such as T-mall, Taobao, JD.com, and YHD.com, demonstrates that disclosing sales data lowers the possibility that managers will conceal bad news and improves the accuracy of market expectations, mitigating the stock price crash risk (Li and Liu 2023).

Consumption data provide investors with a reliable assessment of business retail performance in a timely and direct manner. However, collecting consumption data may raise privacy concerns. When confronted with this problem, privacy computing is a viable solution that allows for data analysis without exposing the actual data.

Enterprise characteristics data

In addition to traditional financial data, such as ROA, Tobin Q, or dividends, enterprise characteristics data correlated with company performance should also be considered, including executive psychological and cultural traits, diversity of managers, and employee happiness. Such nonfinancial indicators can significantly influence managerial decisions and employees' overall productivity and morale, contributing to a company's operations and returns. Furthermore, variables such as research and development (R&D) prowess, CSR initiatives, environmental, social, and governance (ESG) ratings, and supply chain can shape a company's public perception, affecting its stock market performance and mitigating its managerial risks.

First, senior managers' personal backgrounds and characteristics significantly influence corporate governance and development strategy (Duréndez et al. 2023). CFOs with uncertainty avoidance cultural backgrounds could reduce price crash risk for companies, especially when there is greater information asymmetry, increased risks, and a more influence of the CFO on the firm's strategic decisions (Fu and Zhang 2019). CEOs with more masculine faces or increasing perceptions of bad luck in the Chinese Zodiac are more sensitive to corporate risks (Kamiya et al. 2019; Li et al. 2021). Additionally, overconfident CEOs who are optimistic about their firms' future tend to attract more stable supplier relationships with greater investment, longer duration, more substantial labor commitments, and lower turnover rates (Phua et al. 2018). Board directors with athletic backgrounds have better physical fitness, mental resilience, leadership, and teamwork skills, leading to better firm performance (Dong et al. 2019).

Managers' diverse backgrounds and characteristics can also affect company returns. Using data on board working experience, it has been found that the diversity of board tenures can effectively reduce corporate risk, and gender-diverse boards tend to make sound and safe decisions, reducing corporate risks (Mascia and Rossi 2017; Ji et al. 2021; Mohsni et al. 2021). However, not all prior experiences or features benefit the company's development. Managers who survived fatal disasters successfully tend to be more aggressive, whereas those with trauma act more conservatively (Bernile et al. 2017).

Second, when employees have high satisfaction and a sense of belonging to the company, they tend to be more productive, generating higher company returns. Glassdoor is an employer review and recruiting website where current and former employees review their companies, salaries, interview experience, senior management, and corporate benefits. Employee satisfaction represented by Glassdoor ratings is highly associated with performance, leading to higher sales volumes and positive changes in corporate economic fundamentals (Green et al. 2019).

Third, given the strict demands imposed by regulatory authorities on social responsibility disclosure, scholars focus more on the performance of social responsibility and its relationship with firm development. Firms' social responsibility ratings are mainly defined by the national legal system (Liang and Renneboog 2017), including corporate

governance, cultural background, and corporate environmental performance (Lu and Wang 2021). When a company has a higher environmental, social, and governance (ESG) level, it can induce positive changes in public sentiment, which in turn can positively affect market pricing for a firm's sustainable development activities with big ESG data (Serafeim 2020).

Fourth, companies' innovation capability could reflect the enterprise's future development, especially in high-technology industries. The level of enterprise innovation ability affects many factors, such as enterprise loans, credit, and investment. Improving innovation ability can reduce the information gap between enterprises and lending banks, bolster enterprises' borrowing capacity, and facilitate their growth and development (Francis et al. 2012). In addition, enterprises that engage in iterative innovation to gain a competitive advantage in the market could decrease stock price volatility and minimize the risk of bankruptcy. This suggests that a company with high innovation capability faces lower credit constraints, which promotes enterprise expansion (Eisdorfer and Hsu 2011).

Fifth, a company's supply chain can also be a predictive tool for stock prices and risks. A previous study demonstrated that adding the supply chain to stock selection in a multifactor model could increase the model's profitability and reduce its risks (Sun et al. 2023). An analysis of the supply chain networks of 2115 Chinese A-share listed manufacturing companies found that network centrality significantly negatively impacts the risk of stock price crashes (Shi et al. 2022).

Finally, the attention of external participants, such as analysts and investors, in the market is highly correlated with corporate stock performance. Different durations of holding stocks by institutions have different effects on company stock liquidity, i.e., the increase in long-term/short-term institutional stock holdings is related to negative/positive company liquidity (Wang and Wei 2021). Moreover, shared analyst coverage has been confirmed to represent firm connections, and the momentum factor of connected firms can generate significant monthly alpha (Ali and Hirshleifer 2020).

Based on these findings, we conclude that enterprise characteristic data provide valuable insights into various facets of company operations, enabling researchers and managers to make informed decisions, enhance overall corporate performance, and effectively manage risk and engage in strategic planning.

Credit data

Financial institutions and firms typically construct credit ratings using traditional data to assess the ability of businesses or individuals to repay loans (Vanini et al. 2023); however, this limits the opportunities for some capable individuals and companies to obtain loans, especially in the case of start-ups. With credit data, such as behavioral loan tracking, location-based information, and mobile application data, researchers and institutions could broaden the range of factors considered in credit ratings, increasing the original credit model's accessibility and accuracy, reducing loan risk, and identifying current market conditions.

First, many studies have shown that credit data can enhance the accuracy of credit scoring models. Adding variables such as purchase and repayment information, email usage, psychometric data, and demographic data can improve traditional credit model

accuracy (Djeundje et al. 2021; Rozo et al. 2023). With borrowers' online shopping habits, payment histories, locations, device usage, and web browsing data, the credit model enhances its predictive accuracy by 18.4% over traditional internal scoring systems, better identifying individuals' risk of loan default (Jiang et al. 2021). When comparing loans produced by Lending Club to identical loans originating from banks, the rating grades assigned by private credit companies with alternative data are comparable to those assigned by traditional banks (Jagtiani and Lemieux 2019).

Second, credit data could identify potential misconduct, reducing loan risk. For example, legal expenses could evaluate the presence of dishonest activity in bank loans and internal controls (McNulty and Akhigbe 2017). In addition, people with close relationships tend to have similar credit scores. Using a sample of approximately five million investor-loan-hour data from a Chinese peer-to-peer website, individuals associated with specific relationships were found to have comparable credit records (Caglayan et al. 2021). Likewise, researchers have examined the influence of coworkers' misconduct records on the chance of employees committing financial fraud using data on people's job histories and misbehavior records from US financial advisers. Advisors are likelier to engage in misconduct if they are paired with new coworkers with a previous misconduct record during a merger (Dimmock et al. 2018).

Credit data can also be used to identify market conditions and fluctuations. For instance, loan-level data from millions of used-car transactions show that loan maturity is an important lever affecting durable goods prices by influencing consumer responses. Specifically, extending loan maturities by one year could increase a car's price by approximately 2.8% (Argyle et al. 2020).

Overall, alternative data sources offer valuable insights into individual or company loan decisions and credit assessments, improving accuracy and providing a comprehensive understanding of credit behavior and market dynamics.

Government data

Government data mainly conveys macroscopic information such as public policy (government spending, regulations, policies) and demographics (population, age distribution), which is collected and published by official agents, making it credible and authoritative. These indicators can affect the price of commodities, the future of industries, and investors' sentiments; therefore, investors, researchers, and policymakers rely on these data to understand the economy's overall condition and make informed investment decisions.

First, government data affect the price of commodities, such as crude oil, reducing investment risks. Crude oil prices usually reflect the overseas CPI, inflation risk, and commodity pricing level. Using an emotional score to analyze a wide range of global news, it is found that news related to macro fundamentals, such as public finance, foreign exchange, and housing, could affect oil prices in the short term and significantly predict oil returns in the long term (Brandt and Gao 2019). Similarly, researchers clarify the long-run covariability between economic policy uncertainty and oil prices, revealing varying degrees of correlation between oil prices and policy uncertainty across different countries (Shahbaz et al. 2021). It has also been proven that U.S. economic conditions at the state level, including mobility measures, labor market indicators, real economic

activity, expectations measures, financial indicators, and household indicators, can predict future realized volatility of oil price returns (Gupta and Pierdzioch 2023).

Second, government data are often used to determine the value of companies and market conditions. For instance, a study examined the resident population data collected from customs and port authorities and found a greater concentration of residents near a company's headquarters positively correlates with a higher level of trade value for the firm (Cohen et al. 2017).

Third, government data can influence investor sentiments, financial markets, and stock returns. For example, political climate change data suggest that people invest more aggressively when their preferred party is in power (Bonaparte et al. 2017). In addition, recent research indicates that when using the Consumer Confidence Index as a proxy for sentiment, a negative correlation appears between investor sentiment and stock returns (Wang et al. 2021).

Government data are crucial to understand commodity pricing, market performance, and investor sentiment. It contributes to a comprehensive understanding of financial market economic trends and conditions.

Sensing data

The rapid progress of technology, such as mobile computing and the Internet of Things, has significantly facilitated implementing smart applications, including smart transportation, smart industries, smart agriculture, and smart cities. Millions of sensing devices, including various sensors (temperature, air pressure, pollutants, etc.) and cameras, are deployed in satellites, smartphones, automobiles, and transportation to capture the status of their surroundings. After being precisely computed and examined, some of these statistics are assessed as reliable economic trend indicators, while others can be used to predict market prices. On a macro level, these data are effective for analyzing the status of the entire economy, whereas, on a micro level, they reflect business performance. Given that sensing data are collected instantaneously, the influence of their analytical results on forecasting financial market developments is crucial. Currently, the main sensing data used for academic analysis and prediction are satellite, location, and weather. This section discusses the financial uses of sensing data.

Satellite data

Satellite data are high-resolution images of the Earth's surface. Some scholars consider that the limited availability of satellite data may only benefit skilled investors and worsen the information asymmetry between institutional and ordinary investors. After examining the difference-in-difference experiences with parking lot fill rates, it was discovered that expanding access to satellite data may result in increased short-selling activity, decreased trading, and decreased stock liquidity (Katona et al. 2018). However, more studies indicate that such data could provide advantages. For example, satellite data have propelled research on the relationship between price efficiency and corporate governance, potentially enhancing price efficiency, reducing managerial rent extraction, and decreasing the likelihood of trading based on private knowledge (Zhu 2019). Satellite data provide a comprehensive and up-to-date view of the planet, including information about land cover, population density, transportation networks, and infrastructure

development. Many studies have concluded that satellite data can be employed to predict stock and commodity prices and convey economic conditions.

First, satellite data can be used to predict stock returns. Research has found that store-level data across more than 67,000 stores for 44 major retailers can anticipate retailers' performance for the next quarter. Abnormal changes in parking lot rates can positively predict stock returns (Katona et al. 2018).

Satellite data can also be applied to forecast the production and prices of commodities to reduce investment risks. A vendor of satellite data, RS Metrics, uses imagery to capture changes in metal inventories and forecast metals and commodity prices.[5] Research applications include predicting crop yields (Kleshchenko et al. 2012).

Due to the global coverage of satellite data, researchers can also explore economic performance within and variation across societies worldwide. The global luminosity map drawn from nighttime satellite images clearly discerns political boundaries, which aims to balance the merits and drawbacks of macro cross-country studies. By combining this information with history, archeology, and archival information, economists can assess issues and clarify the key drivers of economic activity worldwide (Michalopoulos and Papaioannou 2018). More specifically, satellite night light data could predict common economic problems, such as whether the number of people living in poverty is increasing (Subash et al. 2018).

In summary, we anticipate that satellite data will become more widely available as technology advances and supports more innovative research.

Location data

The location of a business is often pivotal role in determining its success, particularly in consumer-centric industries. A good location can attract more consumers, leading to high revenue. In addition to projecting company revenue, recent studies have highlighted that geographical proximity could significantly improve credit evaluation and business risk management.

Usually, the largest challenge for credit evaluation is information asymmetry, but when firms are situated closer to the credit agency, information gathering and transmission efficiency are significantly enhanced. The close geographical distance between a company and a credit agency could alleviate its financial burden by mitigating information asymmetry. For example, advanced railway systems in China could increase pricing efficiency by reducing travel time (Gao et al. 2021a, b). Face-to-face interactions allow credit centers to have the latest information and establish contractual relationships with borrowing firms. Building upon this understanding, a subsequent study illustrated that developing high-speed railways can significantly reduce credit costs for local private enterprises, particularly when they encounter difficulties in accessing or dealing with sensitive information (Geng et al. 2023).

Conversely, close proximity between a company and its institutional investors could enhance the quality of external corporate oversight, resulting in more robust executive compensation arrangements that incentivize managers to work diligently and assume

[5] Using satellites to forecast metals and commodity prices. Nanalyze. Retrieved June 7, 2022, https://www.nanalyze.com/2019/02/satellites-forecast-commodity-prices/

risks (Mazur and Salganik-Shoshan 2017). This could help align manager and shareholder interests, mitigating agency costs.

Weather data

Psychological researchers have discovered that weather also influences people's moods (Cunningham 1979). Scholars in the financial field have been intrigued about how meteorological conditions impact investor sentiment, subsequently affecting investment choices and market performance. The study of weather incorporates a range of aspects, including pollution level, temperature, humidity, perception levels, spring or summer, and even lunar phase.

Numerous studies indicate that variations in weather impact investors' trading decisions. For example, on overcast and rainy days, investors are more pessimistic regarding future uncertainty (Kliger and Levy 2003). Another study used more variables, including wind speed, daylight hours, humidity, and air pressure, and showed that investors prefer to trade more when the weather is good (Schmittmann et al. 2015). Heterogeneous characteristics of enterprises, such as size, BE, and momentum-related factors, are susceptible to hurricane events, resulting in strong abnormal stock returns (Lanfear et al. 2019). The market performance on the day after transitioning from Daylight Saving Time (summer time) to Standard Time (winter time) is lower than that on the same weekday that did not undergo this change (Mugerman et al. 2020).

Researchers have also demonstrated that weather, such as severe flood events, affects market conditions, causing loan risks (Garbarino and Guin 2021). Many researchers have included weather as a new factor in improving economic models, such as the ARFIMA Seasonal GARCH model (Taştan and Hayfavi 2017). Similarly, global temperature shocks related to increasing CO_2 emissions are systemic priced factors in the Arbitrage pricing theory model, since there is a significantly positive risk premium for global temperature shocks with an estimated 1.8% annual impact on the US cost of equity (Gregory 2021). Using data on CO_2 emissions, scholars can identify the usage of durable goods and develop a CCAPM model, with strong cross-sectional pricing power (Chen and Lu 2018).

Roles of alternative data from financial theoretical insights: a bridge leading the market to be more efficient

Advancements in mobile technology, data storage, cloud computing, and machine learning are making alternative data more accessible and affordable. As a result, alternative data availability has gradually improved, increasing stock price informativeness and enhancing market efficiency (Zhu 2019). In the following section, we further analyze the role of alternative data through the lens of both traditional and behavioral finance theories. We determine that alternative data serve as a bridge to the market to become more efficient.

Alternative data and traditional finance theory

Traditional finance theory assumes that investors are rational utility maximizers and markets are efficient, where security prices synchronously reflect all market information. Fama (1970) proposed the Efficient Markets Hypothesis (EMH), pointing out the

trichotomy of (i) strong-form informational efficiency, where all public and private information in the market is reflected in security prices; (ii) semistrong-form informational efficiency, where investors with private information can beat the market but investors with no more than public information cannot; and (iii) weak-form informational efficiency, where investors with no more than a subset of public information, historical asset prices, cannot beat the market.

A prerequisite for strong-form informational efficiency is that the costs of obtaining prices that reflect information are always zero (Grossman and Stiglitz 1980). Strong-form informational efficiency is impossible since there are always information acquisition costs and continual trading costs (Fama 1991). Moreover, information asymmetry among different investors and agency problems between investors and managers would also affect market efficiency (Zhu 2019).

Decreasing information asymmetry

Before alternative data became widely available, some exclusive information, such as consumer transactions, was only known by specific market participants, such as privately informed corporate managers and institutional investors, who could obtain private information at high costs. The information imbalance among different parties results in serious information asymmetry between investors and managers and between retail and institutional investors.

Currently, with detailed and high-frequency alternative data increasingly conceivable, investors are likely to obtain incremental information that they cannot obtain from the disclosed financial reports of firms at a lower cost and with greater precision. For example, with the emergence of e-commerce platforms (e.g., Taobao, http://www.taobao.com/; JD, https://www.jd.com/; Tiktok, https://www.douyin.com/; Amazon, https://www.amazon.com), investors can acquire enterprises' statistical data on online sales with more real-time performance and finer granularity than through quarterly or annual enterprise financial reports. These enterprise-oriented alternative data sources could convey their internal business operation knowledge and be used to predict future performance, fundamentals, or stock prices. Alternative data have become a vital source of available private information. When investors incorporate incremental information from alternative data into their investments, they will be more informed and make more effective decisions.

Statman (2018) divided EMH into price-equals-value and hard-to-beat market hypotheses. Price-equals-value markets are those where security prices always equal their fundamental values, whereas the price-equals-value market hypothesis claims that investment prices always equal their intrinsic values. Hard-to-beat markets are those in which some investors can beat the market and obtain excess returns, but most investors cannot. Similar to Fama's three forms, Statman proposed three forms of the hard-to-beat market hypothesis: 1) exclusively available information, where even investors with exclusively available information cannot beat the market; 2) narrowly available information, where investors with exclusively available information can beat the market, but investors with no more than narrowly available information cannot beat it; and 3) widely available information, where investors with exclusively or narrowly available information can beat the market, while those with no more than widely available information cannot.

Stateman's three forms of hard-to-beat markets represent different degrees of information asymmetry: widely available information indicates the greatest information asymmetry, whereas exclusively available information indicates the least asymmetry. From the insights of alternative data and their role in moderating information asymmetry, we could interpret the above three forms of hard-to-beat markets as follows: (1) when all investors can acquire exclusively available alternative information, it is almost impossible for any investor to beat the market; (2) when some investors can acquire exclusively available alternative information and the rest can acquire only narrowly available alternative information, the former investors can beat the market while the rest cannot; and (3) when some investors can acquire exclusively or narrowly available alternative information, and the rest can acquire only widely available information, investors with exclusively or narrowly available information can beat the market.

It is reasonable that the emergence and application of alternative data could diminish information asymmetry, especially between managers and sophisticated investors. For example, using alternative data, such as online reviews and patents, much research has examined constructing novel pricing factors such as sentiments (Liew and Budavari 2017), technological links (Lee et al. 2019), and employee reviews (Green et al. 2019). Some sophisticated investors, such as hedge funds, use these innovative pricing factors to earn returns. They take long and short positions on stocks with high-value factors and stocks with low-value factors. Pursuing these strategies in the real market may explain why the predictive power of some novel characteristics has weakened over time (Barberis 2018), which is consistent with the increase in price informativeness due to the availability of alternative data.

With increasingly sophisticated investors acquiring valuable alternative information, information asymmetry between these investors and managers has decreased, contributing to more efficient markets. Since there are always differences in data availability and the abilities of different investors (Sun and Zeng 2022), not all market participants obtain the same information or react the same way. Similarly, not all investors benefit from these alternative data, especially when the data acquisition cost is high. Information asymmetries between sophisticated and retail investors may be more extreme in the short term. In the long term, however, this type of information asymmetry would decrease when more alternative data are available at low costs, further promoting market efficiency.

Reducing agency costs

Agency problems resulting from the separation of ownership and control of enterprises are the cause of investment inefficiencies by managers, which are sometimes misaligned with shareholder interests (Richardson 2007; Zhu 2019). The availability of alternative data can reduce information asymmetry and benefit the transfer of information advantages, helping to reduce agency costs.

First, managers could optimize their business measures and focus more on performance targets. Specifically, alternative data such as consumer transactions and online company reviews can help managers better understand consumer preferences and expectations. Acquiring alternative data about competitors can also help enterprises develop response strategies in advance. Hence, managers could actively develop more

rational and promising business strategies to meet consumer demand and maintain market competitiveness. Additionally, the transfer of information advantages within an organization brought about by real-time and widely available alternative data is a positive signal of decision-making and accounting centralization. As the world's largest retailer, Walmart has devised a "continuous real-time benchmarking" with automatic alerts for abnormal performance, making acceptable target setting possible (Marr 2018). In US manufacturing companies, the authority of headquarters is attributed to predictive analytics using massive and high-frequency data, particularly in the centralized decision-making related to marketing and human resources (Labro et al. 2023). Meanwhile, the authority could help headquarters understand the performance outcomes of plant managers and promote greater use of performance-based incentives.

Furthermore, with more internal and external information, managers will engage in fewer improper securities-related transactions. As stated above, the availability of alternative data could reduce managers' opportunities to trade on private information about the company. Evidence has shown that managers do not disclose all their private information (Froot et al. 2017), and managers have an advantage in private information over personal trading earnings (Rogers 2008). Such insider trading damages the interests of other stakeholders. When investors obtain more alternative data containing information not publicly disclosed by the company, they can incorporate incremental information into their investment behaviors and impact stock prices more quickly. Thus, insiders have fewer opportunities to earn excess returns. Empirically, Zhu (2019) found that managers exploit private information about future earnings through personal trades less when stock prices reflect more information from alternative data. In addition to insider trading, there are many other market manipulation activities, including but not limited to corporate fraud and stock price manipulation. Multisourced alternative data are valuable indicators for detecting such manipulation activities. For example, in 2020, Muddy Waters demonstrated that Luckin Coffee had committed fraud by exploiting alternative data such as WeChat records in the operating store managers group, Luckin customer receipts, the Luckin app, and videos.

In sum, alternative data have incremental information content compared with traditional data, which could be used to predict firm performance and stock prices, increasing price informativeness. Simultaneously, applying state-of-the-art technologies could reduce data acquisition costs, and gradually scale up applying alternative data. Using more information derived from alternative data in financial markets can mitigate information asymmetry and agency problems, improving market efficiency in the long term.

Alternative data and behavioral finance theory

Although the efficient market hypothesis is far-reaching for classical financial research and ample research has provided supportive empirical evidence, many findings do not reconcile with this hypothesis, i.e., the so-called "anomalies," such as calendar patterns (Cooper et al. 2006), idiosyncratic volatility (Ang et al. 2006), value premium (Lakonishok et al. 1994), momentum (Jegadeesh and Titman 1993), and postearnings announcement drift (Ball and Brown 1968)). Violations of the efficient market hypothesis have resulted in some researchers to focus on investors' irrational behaviors, leading to a growth in behavioral finance research. Regarding the two main assumptions of the

traditional finance framework that people are rational and seek to maximize expected utility (Barberis 2018), behavioral finance challenges them from the perspective that individuals have biased beliefs resulting from psychological heuristics (Kahneman and Tversky 1973). Investors are faced with a bounded rationality that they cannot promptly gather and react to all information related to their investment (Huberman and Regev 2001). Behavioral finance argues that investors' irrational behaviors driven by cognitive biases result in security prices deviating from their fundamental values, and sophisticated investors cannot eliminate this deviation because of arbitrage limits (Shiller and Fischer 1984). Similar to the earlier analysis of how alternative data will promote more efficient markets, we next analyze how alternative data could help react to the challenges to efficient markets posed by behavioral finance.

Improving investor rationality

Efficient markets assume that investors are fully intelligent and rational enough to adequately identify and process complex information. They can distinguish all potential arbitrage opportunities and trade, returning the price to a fair value level where investors no longer obtain excess returns. However, it should be noted that this rational man hypothesis is too strong for real markets to realize, especially when there is a high proportion of retail investors. Retail investors have an insufficient and biased understanding of professional output (such as financial reports, policy issues, etc.), even in the case of some institutional investors. Behavioral finance is summarized as cognitive bias and decision deviation. Because of these traits, investors exhibit irrational trading behaviors when confronted with traditional information, earning them the label of "noise traders".

Multisource alternative data, such as the volume of commuters, investor sentiment, and geographical location, could be regarded as raw data compared with traditional data such as financial reports (Dugast and Foucault 2018), reducing investors' comprehension barrier. We have discussed that such information can predict stock prices by various paths; therefore, once deviation is effectively corrected, investors can quickly obtain proper message signals to identify potential arbitrage opportunities, even if they know less about financial knowledge than professional investment institutions.

Moreover, financial technology innovation may aggravate capital income inequality, resulting in unaffordable private information and reducing the involvement of uninformed investors in the market (Mihet 2022). With more alternative data emerging, there will be fewer completely uninformed traders and more quantitative investment in the market. Empirical evidence has shown that alternative data, such as satellite coverage data, lead to more informed short selling and fewer informed personal transactions (Katona et al. 2018). Alternative data can provide many effective stock selection and timing factors for quantitative trading that are practically unaffected by subjective noise but are close to perfect rationality. When noise traders tend to be more rational, the markets will move closer to the rational man hypothesis.

Mitigating limits of arbitrage

In efficient markets, monopolists cannot operate freely due to the assumptions of sufficient competition, freedom from friction, and the absence of arbitrary costs. However, these assumptions are difficult to meet. Real markets are always filled with insufficient

competition, high information acquisition costs, and borrowing constraints, resulting in market manipulation behaviors and mispricing. Consistent with the inability to eliminate information asymmetry, delayed arbitrage recognition is inevitable; however, it can be relieved with available and real-time alternative data.

Until now, almost all capital markets in the world, including developed markets in the US and Europe and emerging market, such as the Chinese stock market and Indian markets, are still weak or semistrong (Cajueiro and Tabak 2004; Kristoufek and Vosvrda 2013; Liu et al. 2020). Consequently, investors' temptation to seek excess returns will never weaken; thus, seeking exclusive information is an eternal topic in all markets. Traditionally, those who can access exclusive information include corporate insiders such as directors, executives, and professional brokers in stock exchanges. However, some newcomers have recently gained insightful information about companies by analyzing alternative data and obtaining huge excess profits in financial markets. Alternative data are regarded as a new source of fat alpha, and investors are undoubtedly willing to pay significant costs for excess returns. Compared with traditional data, alternative data are characterized by availability and legitimacy, making it possible to market alternative data trading.

The popularity of mobile devices, low-cost sensors, and other technology has reduced the cost of data collection, leading to the creation of some start-ups that collect alternative data. With the development of data technologies, internal information will be circulated in the market, and information costs will decrease. Meanwhile, multisource alternative data enable researchers to obtain information about brands, enterprise activities, and commodity transportation on the financing side and analyze market sentiment orientation and consumption trends on the investment side. This can prevent fragmentation and localization problems in information processing and improve noise traders' valid cognition of the overall capital market. The availability of these third-party data sets reduces the cost of investors' access to information and decreases information asymmetry between firm insiders and outside investors (Blankespoor et al. 2014, 2018). Thus, alternative data could provide a bridge to a more efficient market through cheaper information for arbitrage and more competitive activities. In fact, professional investors have begun to use alternative data in their investment strategies.[6]

It is also assumed that price discovery is prompt in efficient markets, and the emergence of new information and arbitrage opportunities is relative, considering market dynamics. Arbitrage opportunities will be quickly identified and traded, bringing prices to a level without excess returns. In addition, new arbitrage opportunities emerge relatively slowly. If the arbitrage process is also slow or arbitrage opportunities have been relatively prompt, then arbitrage opportunities would exist for a long time and the market would be ineffective. In real markets, mainstream information is often released with lags, resulting in delays in the judgment of the company's fundamental factors.

However, by benefiting from real-time alternative data, investors could engage in fundamental analysis ahead of mainstream data to grasp arbitrage opportunities quickly. Differentiating effective decision information from noise information about companies

[6] A Q4 2016 Data Sets Market Survey from Bank of America Merrill Lynch conveys this finding.

is crucial for shrewd investors who can discover prices. Empirically, alternative data leads to a significant improvement in short-term price efficiency, promoting the incorporation of fundamental information related to long-term performance of the enterprise. Through the difference-in-difference experiment with matching samples, Zhu (2019) finds that the market reaction to earnings announcements becomes relatively calm when alternative data are applied in financial markets.

In conclusion, there is a growing possibility that competitive trading will become a reality as the alternative data market grows. Simultaneously, market improvements will facilitate identifying high-quality alternative data that are distinct from noise data. Therefore, we speculate that alternative data can potentially improve capital market efficiency.

Discussion of challenges and future trends

Challenges

Alternative data help managers and investors understand enterprise conditions more clearly and efficiently; however, we cannot ignore the many challenges to utilizing alternative data in business administration and finance analytics.

Challenges in new data sources and acquisition technologies

First, it is necessary to look for the next untapped data source. As huge returns brought about by alternative data in financial markets attract many investors, current alternative data will gradually become mainstream, and its extra information advantage may continuously diminish. Only new data sources can bring excess returns when they are reflected in the stock price.

Meanwhile, alternative data may be biased. Any data influenced or generated by personal opinions inherently carry a degree of bias. For instance, data on employee satisfaction can be exaggerated due to fear of negative repercussions, such as job loss, fewer opportunities, or unfavorable treatment by supervisors or colleagues. Similarly, data pulled from social networks can be inherently skewed. Users may only share favorable content, avoid expressing extreme views to fit societal norms, or make comments influenced by partial or one-sided information. Selection bias is another challenge. Each type of alternative data cannot encompass all individuals and companies. In addition, large gigabytes of data are typically examined as a subset rather than in their entirety, which may lead to incorrect or incomplete interpretations. Thus, conclusions drawn from such data might not represent the entire population. There are two main approaches to solving the bias problem. First, we could use statistical methods, such as robustness checks, with other data sources, stratified random sampling to select representative samples, or anonymous users to avoid the herding effect. Second, we could use technology, such as bias detection in text with NLP techniques. However, more methods to solve data bias are still needed.

Challenges in the data process and analytics

Alternative data are multisource, heterogeneous, and contain unstructured data such as texts, images, audio, and videos. Compared to traditional data, which can be processed directly with analytic tools such as Stata, Eview, or Matlab, alternative data generally

require more technologies, such as machine learning and deep learning, to accomplish the mission.

Noisy information is always unavoidable when extracting valuable messages from alternative data, which might lead to skewed errors of type I and type II.[7] Moreover, social media may contain false, exaggerated, commercial, and unofficial information. While studying the impact of tweets on the price of top cryptocurrencies, about 1–14% of tweets are from "bots" accounts (Kraaijeveld and De Smedt 2020). Dealing with noise hidden in texts, audio, videos, and images is undoubtedly challenging for researchers and investors. This challenge is further amplified with the emergence of new NLP technologies such as ChatGPT and LLaMA.

Challenges in financial applications

It is difficult to interpret the value of alternative data in the market. Each type of alternative data requires prospecting and assetization (Hansen and Borch 2022). Alternative data analysis requires finance and data science knowledge to determine the potential relationship between the data and financial markets. Researchers and investors must have deep insights into financial markets and enterprises.[8]

Moreover, making decisions based on the analysis results of alternative data is challenging. More than one type of alternative data could be collected and used to discover what happened to the company. The problem is merging data of different types (texts, images, audio, or videos) and different periods to make decisions or design a more accurate prediction model without overfitting. Sometimes, the analysis results from different data may be inconsistent and contradictory.

Future trends

With the continuous development of technology, new data technologies are emerging, leading to breakthroughs in the volume of available alternative data. For example, as a primary type of alternative data, personal data, including shopping information, online comments, and company opinions, is greatly valuable in reflecting the company's actual status. However, most individuals are unlikely to share this data for privacy protection. To solve this problem, privacy computing technology would play a role in applying sensitive data. Privacy computing is expected to be critical for extending alternative data application scenarios. Encryption mechanisms can enhance data protection and reduce the risk of data leakage compared with traditional data usage. Using privacy computing, owners of alternative data can freely share data.

New research in this area is also booming with the constant emergence of alternative data. Next, we discuss future research trends.

First, existing research generally focuses on text data, such as social media, news, and policies. However, there is a greater volume of alternative data in the form of images,

[7] In the financial models, if we want to test whether there's relationship between two variables, we usually set null hypothesis and alternative hypothesis. Type I Error is to inaccurately reject the true null hypothesis which accidentally show there's relationship between variables. Type II Error is to inaccurately accept the null hypothesis that should be rejected, which wrongly believes that there's no relationship between variables.

[8] Marko, K., & Rajesh, K. T. (2017, May 18). Big Data and AI Strategies Machine Learning and Alternative Data Approach to Investing. Available at: https://cpb-us-e2.wpmucdn.com/faculty.sites.uci.edu/dist/2/51/files/2018/05/JPM-2017-MachineLearningInvestments.pdf.

audio, and videos (Instagram, TikTok, YouTube, and Douyin are examples of sources), attracting more people and containing more information about sentiment. The latest study proved that images and music have fewer barriers between languages and understanding bias and could be more effective in revenue prediction compared to text (Edmans et al. 2022; Obaid and Pukthuanthong 2022). For example, after collecting the daily average sentiment of the top 200 songs by the total number of streams from Spotify, the change in average sentiment of music was positively correlated with the same week's return and negatively associated with the next week's return (Edmans et al. 2022). Therefore, in the future, it will be beneficial to use new analysis techniques, such as image recognition and video analysis, to extract the implied information.

Second, alternative data could alleviate information asymmetry by optimizing credit evaluation and improving lending efficiency. Usually, startups frequently encounter information asymmetry problems when seeking funding support from banks or other financial institutions, since they lack official financial documentation to prove their potential. Alternative data could provide feasible solutions for these firms to establish risk assessment and warning systems. Currently, there are few studies on this topic. In the future, more researchers will be attracted to this area, especially with the encouragement of governments.

Third, different alternative data could be merged for a comprehensive analysis of the company. For example, satellite data may have limitations such as low availability, high cost, low temporal frequency, and limited information from a bird's eye view. Surveys on natural disaster detection mention that satellite images provide only vague descriptions of the disaster (Said et al. 2019). In addition, satellite images are too vague to determine how a disaster may have affected people's lives and sentiments. To obtain more detailed information, social media can be used to supplement satellite data by providing comments, images, or videos of the scene. By combining different types of alternative data, researchers can expand the scope of their investigations.

In summary, we expect increasingly more research on alternative data and the extension of their applications in finance and business. Alternative data will be a crucial factor for investment and management in the future.

Conclusions

Over the past decade, alternative data have opened a new window for the practice of and research on the financial sector, fueled by growing computational capabilities and the proliferation of new data sources. This survey examines more than 100 papers in this area from the past five years, providing systematic descriptions and analyses of applying alternative data in finance and business.

Compared to conventional data sources, alternative data are multisource, heterogeneous, flexible, objective, and constantly evolving. In academic research, alternative data have been applied in credit assessments, company management, risk analysis, investment decision-making, stock predictions, and market forecasting. The business and investment sectors have embraced their potential for extracting valuable insights, generating business opportunities, analyzing stock markets, and guiding investment strategies.

Furthermore, we argue that with sufficient and accessible alternative data sources, a bridge could be built to more efficient capital markets in the future. Its appearance could help market participants gain a more comprehensive and timely understanding of market dynamics, company performance, and risk factors, even in the case of individual investors. Information asymmetry could then be alleviated, and the market could respond faster and more efficiently.

This study delineates the myriad applications of alternative data and their impacts on holistic enterprise evaluation, well-informed investment decisions, and the symbiotic relationship between markets and information. However, research on alternative data is still in its early stages. Increasingly, more alternative data are being integrated into market analysis and investment processes, which will reshape the financial industry's future.

Acknowledgements
The authors thank the editor and the reviewers for their invaluable comments and suggestions and the International Institute of Big Data in Finance members for their suggestions.

Author contributions
SY provided the main idea and the framework of the work. SY, LL, XY contributed to enhancing the manuscript's idea and writing. ZX, HH, and JJ contributed to discussing the paper's content, and ZX developed the role of alternative data from the view of financial theories. SY, JJ, HH, and AA contributed to improving the text of the manuscript-script. All authors read and approved the final manuscript.

Funding
This research is sponsored by the National Natural Science Foundation of China (72371032) and the National Key Research and Development Program of China (2023YFC3305401).

Availability of data and materials
Data sharing is not applicable to this article as no datasets were generated or analysis during the current study.

Declarations

Competing interests
The authors declare that they have no competing interests.

Received: 7 June 2023 Accepted: 7 April 2024
Published online: 02 December 2024

References

Adämmer P, Schüssler RA (2020) Forecasting the equity premium: mind the news!*. Rev Financ 24(6):1313–1355. https://doi.org/10.1093/rof/rfaa007
Agarwal S, Qian W, Zou X (2021) Disaggregated sales and stock returns. Manage Sci 67(11):7167–7183. https://doi.org/10.1287/mnsc.2020.3813
Ali U, Hirshleifer D (2020) Shared analyst coverage: unifying momentum spillover effects. J Financ Econ 136(3):649–675. https://doi.org/10.1016/j.jfineco.2019.10.007
Ang A, Hodrick RJ, Xing Y, Zhang X (2006) The cross-section of volatility and expected returns. J Financ 61(1):259–299
Argyle B, Nadauld T, Palmer C, Pratt R (2020) The capitalization of consumer financing into durable goods prices. J Financ 76(1):169–210. https://doi.org/10.1111/jofi.12977
Ball R, Brown P (1968) An empirical evaluation of accounting income numbers. J Account Res. https://doi.org/10.2307/2490232
Barberis N (2018) Psychology-based models of asset prices and trading volume. In: Handbook of behavioral economics: applications and foundations 1, Vol 1. Elsevier, pp. 79–175.
Bartov E, Faurel L, Mohanram PS (2018) Can twitter help predict firm-level earnings and stock returns? Account Rev 93(3):25–57. https://doi.org/10.2308/accr-51865
Bernile G, Bhagwat V, Rau PR (2017) What doesn't kill you will only make you more risk-loving: early-life disasters and ceo behavior. J Financ 72(1):167–206. https://doi.org/10.1111/jofi.12432
Blankespoor E, Miller BP, White HD (2014) Initial evidence on the market impact of the xbrl mandate. Rev Acc Stud 19:1468–1503
Blankespoor E, Dehaan E, Zhu C (2018) Capital market effects of media synthesis and dissemination: evidence from robo-journalism. Rev Acc Stud 23:1–36
Bonaparte Y, Kumar A, Page JK (2017) Political climate, optimism, and investment decisions. J Financ Mark 34:69–94. https://doi.org/10.1016/j.finmar.2017.05.002

Brandt MW, Gao L (2019) Macro fundamentals or geopolitical events? A textual analysis of news events for crude oil. J Empir Financ 51:64–94. https://doi.org/10.1016/j.jempfin.2019.01.007
Caglayan M, Talavera O, Zhang W (2021) Herding behaviour in p2p lending markets. J Empir Financ 63:27–41. https://doi.org/10.1016/j.jempfin.2021.05.005
Cajueiro DO, Tabak BM (2004) Evidence of long range dependence in asian equity markets: the role of liquidity and market restrictions. Phys A 342(3–4):656–664
Charoenwong B, Kwan A (2021) Alternative data, big data, and applications to finance. In: Fintech with artificial intelligence, big data, and blockchain, Vol Springer, pp. 35–105.
Chen Z, Lu A (2018) Seeing the unobservable from the invisible: the role of co2 in measuring consumption risk*. Rev Financ 22(3):977–1009. https://doi.org/10.1093/rof/rfx027
Chen H, De P, Hu Y, Hwang B-H (2014) Wisdom of crowds: the value of stock opinions transmitted through social media. Rev Financ Stud 27(5):1367–1403. https://doi.org/10.1093/rfs/hhu001
Chen Y, Kelly B, Wu W (2020) Sophisticated investors and market efficiency: evidence from a natural experiment. J Financ Econ 138(2):316–341. https://doi.org/10.1016/j.jfineco.2020.06.004
Cheng L-C, Lu W-T, Yeo B (2023) Predicting abnormal trading behavior from internet rumor propagation: a machine learning approach. Financ Innov 9(1):3
Chiu P-C, Teoh SH, Zhang Y, Huang X (2020) Using google searches of firm products to assess revenue quality and detect revenue management. Available at SSRN 3252314. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3252314. Accessed 5 May 2023.
Cohen L, Gurun UG, Malloy C (2017) Resident networks and corporate connections: evidence from world war II internment camps. J Financ 72(1):207–248. https://doi.org/10.1111/jofi.12407
Cohen L, Malloy C, Nguyen Q (2020) Lazy prices. J Financ 75(3):1371–1415. https://doi.org/10.1111/jofi.12885
Cooper MJ, Mcconnell JJ, Ovtchinnikov AV (2006) The other january effect. J Financ Econ 82(2):315–341
Cunningham MR (1979) Weather, mood, and helping behavior: quasi experiments with the sunshine samaritan. J Pers Soc Psychol 37(11):1947
Da Z, Huang X, Jin LJ (2021) Extrapolative beliefs in the cross-section: what can we learn from the crowds? J Financ Econ 140(1):175–196. https://doi.org/10.1016/j.jfineco.2020.10.003
Da Z, Engelberg J, Gao P (2011) In search of fundamentals. AFA 2012 Chicago Meeting, Chicago
Dimmock SG, Gerken WC, Graham NP (2018) Is fraud contagious? coworker influence on misconduct by financial advisors. J Financ 73(3):1417–1450. https://doi.org/10.1111/jofi.12613
Djeundje VB, Crook J, Calabrese R, Hamid M (2021) Enhancing credit scoring with alternative data. Expert Syst Appl 163:113766
Donadelli M, Kizys R, Riedel M (2017) Dangerous infectious diseases: bad news for main street, good news for wall street? J Financ Mark 35:84–103. https://doi.org/10.1016/j.finmar.2016.12.003
Dong Y, Duan T, Hou W, Liu Y (2019) Athletes in boardrooms: evidence from the world. J Int Financ Mark Inst Money 59:165–183. https://doi.org/10.1016/j.intfin.2018.12.009
Dugast J, Foucault T (2018) Data abundance and asset price informativeness. J Financ Econ 130(2):367–391
Duréndez A, Dieguez-Soto J, Madrid-Guijarro A (2023) The influence of ceo's financial literacy on smes technological innovation: the mediating effects of mcs and risk-taking. Financ Innov. https://doi.org/10.1186/s40854-022-00414-w
Edmans A, Fernandez-Perez A, Garel A, Indriawan I (2022) Music sentiment and stock returns around the world. J Financ Econ 145(2):234–254. https://doi.org/10.1016/j.jfineco.2021.08.014
Eisdorfer A, Hsu PH (2011) Innovate to survive: the effect of technology competition on corporate bankruptcy. Financ Manage 40(4):1087–1117
Ekster G, Kolm PN (2021) Alternative data in investment management: usage, challenges and valuation. Available at SSRN 3715828
Fama EF (1970) Efficient capital markets: a review of theory and empirical work. J Financ 25(2):383–417
Fama EF (1991) Efficient capital markets: II. J Financ 46(5):1575–1617
Fama EF, French KR (2015) A five-factor asset pricing model. J Financ Econ 116(1):1–22. https://doi.org/10.1016/j.jfineco.2014.10.010
Francis B, Hasan I, Huang Y, Sharma Z (2012) Do banks value innovation? Evidence from us firms. Financ Manag 41(1):159–185
Fridson MS (1993) Financial shenanigans: How to detect accounting gimmicks and fraud in financial reports. Financ Anal J 49(3):87
Froot K, Kang N, Ozik G, Sadka R (2017) What do measures of real-time corporate sales say about earnings surprises and post-announcement returns? J Financ Econ 125(1):143–162. https://doi.org/10.1016/j.jfineco.2017.04.008
Fu X, Zhang Z (2019) Cfo cultural background and stock price crash risk. J Int Finan Mark Inst Money 62:74–93. https://doi.org/10.1016/j.intfin.2019.05.001
Gao H, Qu Y, Shen T (2021a) Geographic proximity and price efficiency: evidence from high-speed railway connections between firms and financial centers. Financ Manage. https://doi.org/10.1111/fima.12354
Gao X, Xu W, Li D, Xing L (2021b) Media coverage and investment efficiency. J Empir Financ 63:270–293. https://doi.org/10.1016/j.jempfin.2021.07.002
Garbarino N, Guin B (2021) High water, no marks? Biased lending after extreme weather. J Financ Stab 54:100874
Ge Q, Kurov A, Wolfe MH (2019) Do investors care about presidential company-specific tweets? J Financ Res 42(2):213–242
Geng C, Li D, Sun J, Yuan C (2023) Functional distance and bank loan pricing: evidence from the opening of high-speed railway in China. J Bank Financ 149:106810
Gherghina ŞC, Simionescu LN (2023) Exploring the asymmetric effect of covid-19 pandemic news on the cryptocurrency market: evidence from nonlinear autoregressive distributed lag approach and frequency domain causality. Financ Innov 9(1):1–58

Gholampour V (2019) Daily expectations of returns index. J Empir Financ 54:236–252. https://doi.org/10.1016/j.jempfin.2019.10.004
Giannini R, Irvine P, Shu T (2019) The convergence and divergence of investors' opinions around earnings news: evidence from a social network. J Financ Mark 42:94–120. https://doi.org/10.1016/j.finmar.2018.12.003
Goldstein I, Yang L (2019) Good disclosure, bad disclosure. J Financ Econ 131(1):118–138. https://doi.org/10.1016/j.jfineco.2018.08.004
Green TC, Huang R, Wen Q, Zhou D (2019) Crowdsourced employer reviews and stock returns. J Financ Econ 134(1):236–251. https://doi.org/10.1016/j.jfineco.2019.03.012
Gregory RP (2021) The pricing of global temperature shocks in the cost of equity capital. J Int Financ Mark Inst Money. https://doi.org/10.1016/j.intfin.2021.101319
Grossman SJ, Stiglitz JE (1980) On the impossibility of informationally efficient markets. Am Econ Rev 70(3):393–408
Grover P, Kar AK, Ilavarasan PV (2019) Impact of corporate social responsibility on reputation-insights from tweets on sustainable development goals by ceos. Int J Inf Manage 48:39–52. https://doi.org/10.1016/j.ijinfomgt.2019.01.009
Gupta R, Pierdzioch C (2023) Do us economic conditions at the state level predict the realized volatility of oil-price returns? A quantile machine-learning approach. Financ Innov 9(1):24
HaO L, Renneboog LUC (2017) On the foundations of corporate social responsibility. J Financ 72(2):853–910. https://doi.org/10.1111/jofi.12487
Hasan R, Cready WM (2019) Facebook posting activity and the selective amplification of earnings disclosures. China J Account Res 12(2):135–155
Hansen KB, Borch C (2022) Alternative data and sentiment analysis: prospecting non-standard data in machine learning-driven finance. Big Data Soc 9(1):20539517211070701
Hillert A, Jacobs H, Müller S (2018) Journalist disagreement. J Financ Mark 41:57–76. https://doi.org/10.1016/j.finmar.2018.09.002
Hirshleifer D, Lim SS, Teoh SH (2009) Driven to distraction: extraneous events and underreaction to earnings news. J Financ 64(5):2289–2325
Horváth BL, Huizinga H (2015) Does the european financial stability facility bail out sovereigns or banks? An event study. J Money Credit Bank 47(1):177–206
Huang J (2018) The customer knows best: the investment value of consumer opinions. J Financ Econ 128(1):164–182
Huberman G, Regev T (2001) Contagious speculation and a cure for cancer: a nonevent that made stock prices soar. J Financ 56(1):387–396
Jagtiani J, Lemieux C (2019) The roles of alternative data and machine learning in fintech lending: evidence from the lendingclub consumer platform. Financ Manag 48(4):1009–1029. https://doi.org/10.1111/fima.12295
Jegadeesh N, Titman S (1993) Returns to buying winners and selling losers: implications for stock market efficiency. J Financ 48(1):65–91
Ji J, Peng H, Sun H, Xu H (2021) Board tenure diversity, culture and firm risk: Cross-country evidence. J Int Financ Mark Inst Money. https://doi.org/10.1016/j.intfin.2020.101276
Jiang J, Liao L, Lu X, Wang Z, Xiang H (2021) Deciphering big data in consumer credit evaluation. J Empir Financ 62:28–45. https://doi.org/10.1016/j.jempfin.2021.01.009
Kahneman D, Tversky A (1973) On the psychology of prediction. Psychol Rev 80(4):237
Kamiya S, Kim YHA, Park S (2019) The face of risk: CEO facial masculinity and firm risk. Eur Financ Manag 25(2):239–270. https://doi.org/10.1111/eufm.12175
Kaplanski G, Levy H (2010) Exploitable predictable irrationality: the FIFA World Cup effect on the US stock market. J Financ Quant Anal 45(2):535–553
Kaplanski G, Levy H (2012) The holiday and yom kippur war sentiment effects: the Tel Aviv Stock Exchange (TASE). Quant Financ 12(8):1283–1298
Katona Z, Painter MO, Patatoukas PN, Zeng J (2018) On the capital market consequences of big data: evidence from outer space. J Financ Quant Anal. https://doi.org/10.1017/S0022109023001448
Kleshchenko A, Goncharova T, Naidina T (2012) Using the satellite data in dynamic models of crop yield forecasting. Russ Meteorol Hydrol 37(4):279–285
Kliger D, Levy O (2003) Mood and judgment of subjective probabilities: evidence from the us index option market. Rev Finance 7(2):235–248
Kostopoulos D, Meyer S, Uhr C (2020) Google search volume and individual investor trading. J Financ Mark. https://doi.org/10.1016/j.finmar.2020.100544
Kraaijeveld O, De Smedt J (2020) The predictive power of public twitter sentiment for forecasting cryptocurrency prices. J Int Financ Mark Inst Money. https://doi.org/10.1016/j.intfin.2020.101188
Kristoufek L, Vosvrda M (2013) Measuring capital market efficiency: global and local correlations structure. Phys A 392(1):184–193
Labro E, Lang M, Omartian JD (2023) Predictive analytics and centralization of authority. J Account Econ 75(1):101526
Lakonishok J, Shleifer A, Vishny RW (1994) Contrarian investment, extrapolation, and risk. J Financ 49(5):1541–1578
Lanfear MG, Lioui A, Siebert MG (2019) Market anomalies and disaster risk: evidence from extreme weather events. J Financ Mark. https://doi.org/10.1016/j.finmar.2018.10.003
Lee CMC, Sun ST, Wang R, Zhang R (2019) Technological links and predictable returns. J Financ Econ 132(3):76–96. https://doi.org/10.1016/j.jfineco.2018.11.008
Li Q, Liu S (2023) Does alternative data reduce stock price crash risk? Evidence from third-party online sales disclosure in china. Int Rev Financ Anal 88:102695
Li J, Guo JM, Hu N, Tang K (2021) Do corporate managers believe in luck? Evidence of the chinese zodiac effect. Int Rev Financ Anal. https://doi.org/10.1016/j.irfa.2021.101861
Liew J, Budavari T (2017) The 'sixth'factor—a social media factor derived directly from tweet sentiments. J Portfolio Manag 43(3):102–111
Liu X, Zhou X, Zhu B, Wang P (2020) Measuring the efficiency of china's carbon market: a comparison between efficient and fractal market hypotheses. J Clean Prod 271:122885

Loder A (2019) Goldman rolls out new etfs focused on artificial intelligence. Wall Street Journal. https://www.wsj.com/articles/goldman-rolls-out-new-etfs-focused-on-artificial-intelligence-11551978432. Accessed 5 May 2023.
Lu J, Wang J (2021) Corporate governance, law, culture, environmental performance and csr disclosure: a global perspective. J Int Financ Mark Inst Money. https://doi.org/10.1016/j.intfin.2020.101264
Marr B (2018) Walmart: big data analytics at the world's biggest retailer. Bernard Marr & Co, Milton Keynes
Mascia DV, Rossi SPS (2017) Is there a gender effect on the cost of bank financing? J Financ Stab 31:136–153. https://doi.org/10.1016/j.jfs.2017.07.002
Mazur M, Salganik-Shoshan G (2017) Teaming up and quiet intervention: the impact of institutional investors on executive compensation policies. J Financ Mark 35:65–83. https://doi.org/10.1016/j.finmar.2016.12.001
Mcnulty JE, Akhigbe A (2017) What do a bank's legal expenses reveal about its internal controls and operational risk? J Financ Stab 30:181–191. https://doi.org/10.1016/j.jfs.2016.10.001
Michalopoulos S, Papaioannou E (2018) Spatial patterns of development: a meso approach. Annu Rev Econ 10:383–410
Mihet R (2022) Financial innovation and the inequality gap. Available at SSRN 3474720. https://papers.ssrn.com/sol3/Papers.cfm?abstract_id=3474720. Accessed 6 Feb 2024.
Mohsni S, Otchere I, Shahriar S (2021) Board gender diversity, firm performance and risk-taking in developing countries: the moderating effect of culture. J Int Financ Mark Inst Money. https://doi.org/10.1016/j.intfin.2021.101360
Mugerman Y, Yidov O, Wiener Z (2020) By the light of day: the effect of the switch to winter time on stock markets. J Int Financ Mark Inst Money. https://doi.org/10.1016/j.intfin.2020.101197
Niu R, Xie G, Chen L, Zhao L, Wu M (2023) Information gain in alternative data: evidence from e-commerce sales and analyst earnings forecasts. Manag Decis Econ. https://doi.org/10.1002/mde.3863
Nofsinger JR, Sias RW (1999) Herding and feedback trading by institutional and individual investors. J Financ 54(6):2263–2295
Obaid K, Pukthuanthong K (2022) A picture is worth a thousand words: measuring investor sentiment by combining machine learning and photos from news. J Financ Econ 144(1):273–297. https://doi.org/10.1016/j.jfineco.2021.06.002
Ouimet P, Tate G (2019) Learning from coworkers: peer effects on individual investment decisions. J Financ 75(1):133–172. https://doi.org/10.1111/jofi.12830
Phua K, Tham TM, Wei C (2018) Are overconfident ceos better leaders? Evidence from stakeholder commitments. J Financ Econ 127(3):519–545. https://doi.org/10.1016/j.jfineco.2017.12.008
Quinton S, Wilson D (2016) Tensions and ties in social media networks: towards a model of understanding business relationship development and business performance enhancement through the use of linkedin. Ind Mark Manage 54:15–24
Reyes T (2018) Limited attention and m&a announcements. J Empir Financ 49:201–222. https://doi.org/10.1016/j.jempfin.2018.10.001
Richardson BJ (2007) Do the fiduciary duties of pension funds hinder socially responsible investment? Bank Financ Law Rev 22:145
Rizkiana A, Sari H, Hardjomijojo P, Prihartono B, Yudhistira T (2017) Analyzing the impact of investor sentiment in social media to stock return: Survival analysis approach. In: 2017 IEEE international conference on industrial engineering and engineering management (IEEM)
Rogers JL (2008) Disclosure quality and management trading incentives. J Account Res 46(5):1265–1296
Rozo BJG, Crook J, Andreeva G (2023) The role of web browsing in credit risk prediction. Decis Support Syst 164:113879
Said N, Ahmad K, Riegler M, Pogorelov K, Hassan L, Ahmad N, Conci N (2019) Natural disasters detection in social media and satellite imagery: a survey. Multimed Tools Appl 78:31267–31302
Schmittmann JM, Pirschel J, Meyer S, Hackethal A (2015) The impact of weather on german retail investors*. Rev Financ 19(3):1143–1183. https://doi.org/10.1093/rof/rfu020
Serafeim G (2020) Public sentiment and the price of corporate sustainability. Financ Anal J 76(2):26–46. https://doi.org/10.1080/0015198x.2020.1723390
Shahbaz M, Sharif A, Belaid F, Vo XV (2021) Long-run co-variability between oil prices and economic policy uncertainty. Int J Financ Econ. https://doi.org/10.1002/ijfe.247
Shi J, Liu X, Li Y, Yu C, Han Y (2022) Does supply chain network centrality affect stock price crash risk? Evidence from chinese listed manufacturing companies. Int Rev Financ Anal 80:102040
Shiller RJ, Fischer S, Friedman BM (1984) Stock prices and social dynamics. Brook Pap Econ Act 2:457–510
Siganos A, Vagenas-Nanos E, Verwijmeren P (2017) Divergence of sentiment and stock market trading. J Bank Financ 78:130–141. https://doi.org/10.1016/j.jbankfin.2017.02.005
Sorensen AT (2017) Bestseller lists and the economics of product discovery. Annu Rev Econ 9(1):87–101. https://doi.org/10.1146/annurev-economics-080614-115708
Stamolampros P, Korfiatis N, Chalvatzis K, Buhalis D (2019) Job satisfaction and employee turnover determinants in high contact services: insights from employees' online reviews. Tour Manag 75:130–147
Statman M (2018) Behavioral efficient markets. J Portf Manag 44(3):76
Subash SP, Kumar RR, Aditya KS (2018) Satellite data and machine learning tools for predicting poverty in rural India. Agric Econ Res Rev 31(2):231–240
Subramaniam S, Chakraborty M (2021) Covid-19 fear index: does it matter for stock market returns? Rev Behav Financ 13(1):40–50
Sun Y, Wu M, Zeng X, Peng Z (2021a) The impact of covid-19 on the chinese stock market: sentimental or substantial? Financ Res Lett 38:101838
Sun Y, Zeng X, Zhou S, Zhao H, Thomas P, Hu H (2021b) What investors say is what the market says: measuring china's real investor sentiment. Pers Ubiquit Comput 25:587–599
Sun Y, Liu L, Fang J, Zeng X, Wan Z (2023) Musu: A medium-term investment strategy by integrating multifactor model with industrial supply chain. Int J Financ Eng 10(02):2250034
Sun Y, Zeng X (2022) Efficient markets: information or sentiment? Available at SSRN 4293484. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4293484. Accessed 5 May 2023.

Tang VW (2018) Wisdom of crowds: cross-sectional variation in the informativeness of third-party-generated product information on twitter. J Account Res 56(3):989–1034. https://doi.org/10.1111/1475-679x.12183
Tao R, Brooks C, Bell A (2020) Tomorrow's fish and chip paper? Slowly incorporated news and the cross-section of stock returns. Eur J Finance 27(8):774–795. https://doi.org/10.1080/1351847x.2020.1846575
Taştan B, Hayfavi A (2017) Modeling temperature and pricing weather derivatives based on temperature. Adv Meteorol. https://doi.org/10.1155/2017/3913817
Vanini P, Rossi S, Zvizdic E, Domenig T (2023) Online payment fraud: from anomaly detection to risk management. Financ Innov 9(1):66. https://doi.org/10.1186/s40854-023-00470-w
Vasileiou E, Tzanakis P (2022) The impact of google searches, put-call ratio, and trading volume on stock performance using wavelet coherence analysis: the AMC case. J Behav Financ. https://doi.org/10.1080/15427560.2022.2100384
Wang S, Chen X (2020) Recognizing ceo personality and its impact on business performance: mining linguistic cues from social media. Inf Manag 57(5):103173
Wang X, Wei S (2021) Does the investment horizon of institutional investors matter for stock liquidity? Int Rev Financ Anal. https://doi.org/10.1016/j.irfa.2020.101648
Wang W, Su C, Duxbury D (2021) Investor sentiment and stock returns: global evidence. J Empir Financ 63:365–391. https://doi.org/10.1016/j.jempfin.2021.07.010
Xu Y, Xuan Y, Zheng G (2021) Internet searching and stock price crash risk: evidence from a quasi-natural experiment. J Financ Econ 141(1):255–275. https://doi.org/10.1016/j.jfineco.2021.03.003
Zhu C (2019) Big data as a governance mechanism. Rev Financ Stud 32(5):2021–2061

Publisher's Note