CHAPTER 9

Alternative Data in FinTech and Business Intelligence

Lin William Cong, Beibei Li, and Qingquan Tony Zhang

9.1 Introduction

Alternative data is transforming the financial industry in insurance, crowdfunding, investment management processes, etc. Most asset managers, including hedge funds, mutual funds, foundations, and pension funds, start to realize the complex forces driving this digital transformation. Investment managers that do not follow this seismic shift and update their investment processes are increasingly facing strategic risks and disintermediation: they may very well be outmaneuvered by existing and new competitors who build their processes around alternative data. Understanding the value created from alternative data and participation in this trend provides strategic opportunities for both industry and academia. Big data is generally characterized by high volume, velocity, and variety; hence, they often require specific technology and analytical tools, e.g., Machine Learning and Natural Language Processing, for transformation into value (De Mauro et al. 2016). The recent advances in

L. W. Cong (✉)
Cornell University, Ithaca, NY, USA
e-mail: will.cong@cornell.edu

B. Li
Carnegie Mellon University, Pittsburgh, PA, USA
e-mail: beibeili@andrew.cmu.edu

Q. T. Zhang
University of Illinois Urbana Champaign, Champaign, IL, USA
e-mail: tony.zhang@chicagobooth.edu

M. Pompella and R. Matousek (eds.), *The Palgrave Handbook of FinTech and Blockchain*, https://doi.org/10.1007/978-3-030-66433-6_9

data storage, cloud computing, and statistical tools have gradually reduced costs of gathering data, spurring numerous third-party data aggregators and collectors. This gives rise to alternative data that are not from standard statements or reports, or are unstructured in terms of format. While alternative data have been actively explored in computer science and engineering fields (e.g., voice recognition and machine translation), researchers in finance and business economics have only started to devote attention to them in the past decade.

Given the large number of studies on the emergent field of alternative data and the lack of well-established frameworks for analysis, this chapter provides a brief introduction to a few major types of alternative data, as well as the methods and examples of analyzing or utilizing them, in academic research or practice.

We start with textual data and analytics, which have been used in finance and accounting since the dawn of the century. We discuss the various approaches, the data sources, and recent developments. We then move on to examine images, another form of unstructured data that is available in abundance before touching on audio and video data.

Another non-mutually exclusive major category of alternative data entails digital footprints. The ubiquitous adoption and usage of smart and connected mobile, web, and sensor technologies today have completely changed the way individuals behave and make decisions. These smart technologies have led to the pervasive digitization of individual behavior across digital and physical environments at a very fine-grained level (e.g., social media activities and digital word-of-mouth, online search and clickstream, online and mobile shopping, mobile app activities, and location trajectories), all of which we term as "digital footprints." This information can provide a new lens through which practitioners in the financial industry can better monitor, understand, and optimize human decision-making in the market. By looking into these digital footprints of human beings and their interactions with technologies, managers and policymakers can design more effective strategies for financial platforms to improve the profitability and economic welfare of institutions.

Finally, we discuss the Internet of Things (IoT), which has become prominent in tech innovations and represents a dominant source of alternative data. Widely regarded as a breakthrough in improving consumer lives and retail industry efficiency, the IoT is prevalent in business activities such as manufacturing, logistics, personalized recommendation, etc. With its development comes data collected from decentralized crowds. IoTs can track customers' real-time location to better understand their behavior, generating micro-level information to better predict the future performance of corporations. New technological solutions developed based on the IoT for retailers enable the exploration of authentic customer' behaviors and cheaper marketing opportunities across the world. Whether these innovations take the form of customer experience improvements or business process optimization, the possibilities for IoT are endless and not yet fully understood. We intend to provide

some insights into IoT's huge potential by illustrating several examples of IoT-powered data applications.

The aforementioned alternative data exhibit several common features. First, they are ad hoc, non-standard, with large volumes and large dispersion in terms of data quality. Therefore, we need new tools such as neural-network-based natural language processing and cloud computation. We also need to be very careful in the collection and pre-processing of data for meaningful information retrieval.

Second, alternative data are also often generated jointly by large crowds. In that regard, they are hard to manipulate because individuals all have limited influence on the process. For example, it is easy for someone to fake a personal phone number, but the location data collected by mobile service providers are hard to tamper with. Even if one manipulates his or her location, it constitutes just one data point in a data set with millions of observations, and would thus hardly affect any aggregate analysis.

Finally, alternative data are more diverse and available compared to mainstream numerical data. This means small firms and new entrants may utilize them to have an edge in this nascent stage of industry evolution. This encourages competition and facilitates financial inclusion. Moreover, these data enable them to fill missing markets and better serve the unbanked and historically disadvantaged populations (e.g., thin-filed users, low-income or less-educated people), who otherwise may not receive access to financial services due to no/low historical financial credits in a traditional setting. Therefore, the emergence of alternative data has impacts on the real economy with welfare consequences.

The remainder of the article is organized as follows: Sect. 8.2 discusses a few common forms of alternative data; Sect. 8.3 introduces research and uses cases on digital footprints; Sect. 8.4 surveys applications of data generated from the Internet of Things; finally, Sect. 8.5 summarizes promising future directions for research and for industry development.

9.2 Texts, Images, Voices, and Videos

9.2.1 Textual Data and Analyses

Texts are perhaps the most salient alternative data used in finance and business economics. News articles remain a rich source of information in textual formats. The Wall Street Journal's data are widely used in academic studies, as are The New York Times and the Financial Times. News not only conveys information through each article, but also reveals hidden structures of corporate networks (Schwenkler and Zheng 2019).

Beside news in general, firm-specific news from Factiva could complement corporate filings such as 10K and 10Q (Management Discussion and Analysis

[MD&A], Risk Factor Discussions, etc.) for cross-sectional analysis. Conference call transcripts, analyst reports, IPO prospectus, patent data, and tweets are all alternative data sources.

Earlier studies using textual data are typically count-based and rely on the researchers to predefine a relevant dictionary or word list. Antweiler and Frank (2004), Tetlock (2007), and Loughran and McDonald (2011) are notable pioneering studies. Given the maturity of textual analysis in finance and business economics, for survey articles on text-based analysis in economics, sociology, and political science, we refer the readers to Gentzkow et al. (2017), Evans and Aceves (2016), and Grimmer and Stewart (2013). In particular, Gentzkow et al. (2017) point out that new techniques are needed to deal with the large-scale and complex nature of textual data.

Machine learning is such a technique that is increasingly used in textual analysis. One unsupervised learning tool, Topic Modeling (typically implemented using Latent Dirichlet Allocation [LDA] first introduced by Blei et al. [2003]), has gained popularity in economic and finance studies (Huang et al. 2017; Jegadeesh and Wu 2017). The algorithm lets data self-generate topics and themes. Word embedding from the natural language processing (NLP) literature presents an alternative machine learning tool. Such neural networks language models preserve the syntactic and semantic structure well while maintaining computational tractability. Cong et al. (2019) develop a textual-factor framework to allow projections of numerical or textual information onto a space spanned by a set of interpretable textual factors; Cong et al. (2019), and Hanley and Hoberg (2019) further combine LDA and word2vec to measure corporate governance and systemic risks in the economy.

The various textual analysis tools originated in economics, statistics, and computer science each have advantages and limitations, as illustrated in Fig. 9.1. Cong et al. (2019) contain general discussions of tradeoffs involved, as well as recent developments on word embedding, customized and dynamic count-based methods, and other cutting-edge statistical tools.

Reproduced from Cong et al. (2019), Fig. 9.1. Textual analysis in economics and finance is traditionally count-based, whereas statistical models for analyzing texts often involve inference and regression models that are transparent. More recently, advances in machine learning and natural language processing, especially in deep learning, allow researchers to use black-box machine learning tools to extract information from texts.

9.2.2 Images

Another popular alternative data related to finance is satellite imagery. Image recognition techniques, particularly deep learning algorithms like Convolutional Neural Networks (CNNs), are adopted to process the satellite images. In recent years, many innovative usages of satellite imagery to cross-validate the business metrics, e.g., revenue and store traffic, have emerged.

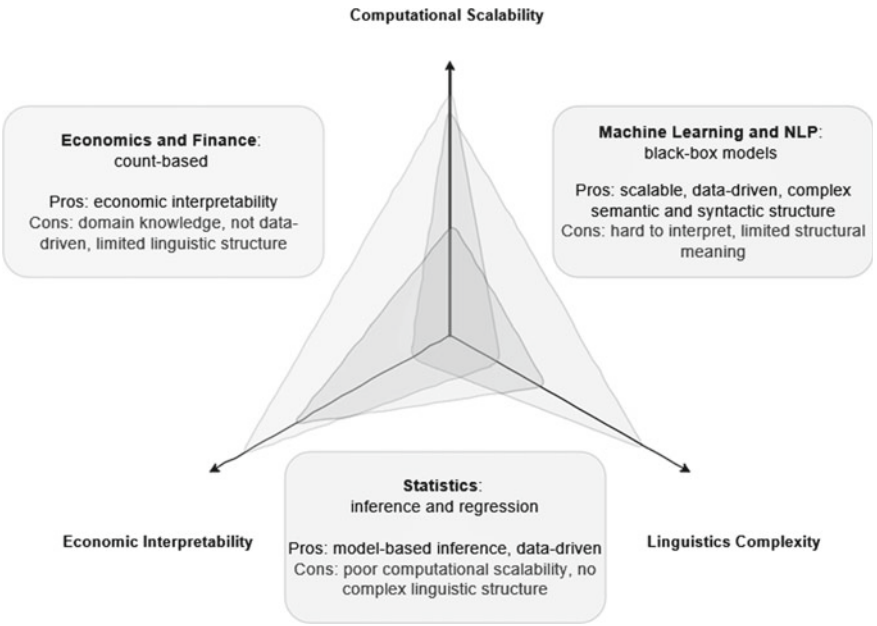

Fig. 9.1 Tradeoffs in various approaches to analyzing texts

Satellite firms provide preprocessing datasets for many sectors. Orbital Insight, founded in 2013, offers satellite data with broad sector applications (OrbitalInsight 2020). Kpler, founded in 2009, offers global oil & gas cargo flow data using satellite, government, and other public registry sources (Kpler 2020). Rsmetrics focuses on metals and commodities, real estate and industrial sector applications (Rsmetrics 2020). For example, RS Metrics is using satellite images of Tesla's production lots to gauge how many cars are being produced and shipped. Machine learning algorithms not only are able to discern between various Tesla vehicle models, but also are able to tell which cars are still parked in the same space and have not been moved between images.

Spire Global, founded in 2010, offers AIS Data providing 3 years of past and real-time positional data for the worlds shipping fleet, satellite collected ADS-B data for aircraft tracking, and GPS-RO Profiles which enable more precise weather forecasting (Spireglobal 2020). Umbra Lab and ICEYE generate raw satellite data using micro-satellite technology that captures imagery regardless of weather conditions (Umbralab 2020; ICEYE 2019).

SpaceKnow, founded in 2014, provides ultra large-scale planetary analysis through the use of satellite imagery (Spaceknow 2019). It parses satellite imagery and has machine learning algorithms that automatically identify objects like cars, boats, trees, and even swimming pools. The usefulness of this tool is evident in examples like SpaceKnow's Satellite Manufacturing Index (SMI) which is said to be superior to China's Purchasing Managers Index

(PMI). The PMI is an index compiled by the National Bureau of Statistics of China which compiles the results of a monthly survey of enterprises' purchasing managers. SpaceKnow's SMI uses an algorithm that compares satellite images of more than 6,000 industrial facilities, and produces a result that's remarkably correlated to the index China produces.

Researchers in finance and accounting have recently started to exploit satellite images. For example, Zhu (2019) examines whether the availability of alternative data improves price informativeness and helps discipline corporate managers. Specifically, the author uses satellite images that provide normalized car counts in parking lots of retailers to find out if a reduction in formation acquisition through alternative data increases long-run price informativeness. To the extent that alternative data provide more information about future profitability of projects available, they can also discipline the manager to make better real investment decisions. The author finds evidence for both phenomena.

Beside satellite images, profile photos are also used in business-related studies. Willis and Todorov (2006) show that people typically make up their minds after a 100-ms exposure to a face. Graham et al. (2016) show that perceived competence plays a more important role than "beauty" in CEO selection and compensation. Bai et al. (2019) find that mutual fund managers who appear "confident" outperform their peers. X. Huang et al. (2018) find that the likelihood of funding increases with entrepreneurs' apparent competence.

Other literature uses facial features as a proxy for testosterone level to study the biological foundation of economic decision-making. For example, Jia et al. (2014) find that male CEO's facial width to height ratio (fWHR) positively correlates with the propensity of financial misreporting. He et al. (2019) document that the fWHR of Chinese male sell-side analysts is associated with higher forecast accuracy. Teoh et al. (2019) apply social psychology models and machine learning techniques to the LinkedIn profile pictures of U.S. sell-side analysts to study how facial traits affect analyst's behavior and performance.

9.2.3 Voices and Videos

The tools for analyzing images in finance and economics research are typically simple, with some exceptions using cutting-edge machine learning tools. Compared to images that are static, voices and videos are much harder to analyze and require more advanced analytics. Research in this area is just starting.

Mayew and Venkatachalam (2012) measure managerial affective states during earnings conference calls by analyzing conference call audio files using vocal emotion analysis software. They find that, when managers are scrutinized by analysts during conference calls, positive and negative affects displayed by managers are informative about the firm's financial future. Analysts do

not incorporate this information when forecasting near-term earnings. When making stock recommendation changes, however, analysts incorporate positive but not negative affects. Their study demonstrates that managerial vocal cues or expressions in conference calls contain useful information about a firm's fundamentals, incremental to both quantitative earnings information and qualitative "soft" information conveyed by linguistic content during the question and answer portion of earnings conference calls.

Another voice data application is the automation of call center operations through voice-enabled automation which provides a massive cost-cutting opportunity for insurance companies. Insurance companies can conduct sentiment analysis to identify certain customer traits and needs based on the emotion and tone in the customer's voice. Accenture developed a systematic method to detect the emotion of customers from voice signal data (Petrushin 2007). These types of improvements in voice processing not only offer cost reduction but also introduce game-changing innovations.

9.3 Digital Footprints

In recent years, the high penetration of mobile devices and internet access has offered new and unparalleled sources of fine-grained user-behavior data such as individuals' cellphone usage, online and mobile activities (e.g., web browsing, click-stream and tap-stream, shopping, and payment), social media and social network activities, GPS locations, and movement trajectories. We term these data "digital footprints" of users. In this section, we take financial credit risk assessment as an example, and discuss how such new sources of alternative data can be leveraged to improve financial predictions, profitability, and social welfare.

9.3.1 Motivation

Conventional data typically cover data from a credit bureau, a credit application, or a lender's own records on existing consumers. Alternative data, instead, come from public social media sites or private applications and devices, and might not directly relate to a consumer's credit behavior. Nevertheless, such new, rich sources of data could show significant potential to complement the conventional data in enhancing the accuracy of existing credit risk assessment (Carroll and Rehmani 2017). Moreover, recent studies have found that credit risk prediction suffers from unintended biases due to potential correlations between input (observed) features and sensitive attributes (such as race, gender, or income) (e.g., Barocas and Selbst 2016; Dobbie et al. 2018; Fu et al. 2019). To some extent, such correlations are due to a lack of control over unobservable factors. Leveraging alternative new sources of behavioral data can enable better control for individual features previously omitted from models, thus reducing biases in credit risk prediction.

Furthermore, prior work on financial credit risk prediction (e.g., Serrano-Cinca et al. 2015) mostly used training data heavily biased toward successfully approved loan applicants whose credit risks had been perceived to be low enough for loan approval ("approved samples" hereafter), as applications initially perceived to be high risk, on the other hand, tend to be immediately rejected, with the result that no further loan payment data on these applicants is to be recorded or included in model training later. Obviously, running credit risk models using approved samples alone can be rather problematic. Approved samples, compared with a true population of loan applicants, tend to have lower probabilities of default and may have significantly different socio-economic characteristics (e.g., higher income, better educated). The patterns or relationships learned from such biased samples might have limited generalizability, and hence may lead to poor predictive performance for new applicants. Moreover, if initially approved samples are biased (intentionally or unintentionally) toward certain sensitive attributes, such errors could be further amplified when training with such samples.

Motivated by the current challenges facing financial service markets, it is important for financial platforms to explore whether and how this new source of users' digital footprint data can help alleviate these concerns. Can digital footprint data help improve predictive performance in microloan credit risk assessment? Moreover, which type of information is the most valuable? Besides, can such digital footprint data help alleviate concerns about training-sample bias (i.e., using approved samples only for model training)? How can we leverage this new type of alternative data to achieve more accurate risk assessment, better financial performance, and, ultimately, higher social welfare for financial platforms?

Note that it remains costly for financial service providers to acquire, store, and process information (Loufield et al. 2018). To obtain an individual's information from multiple sources, financial service providers have to establish close relations with third-party data providers such as social media providers, telecommunication companies, and mobile network operators, as well as other specialized data vendors. Moreover, the increasing size and complexity of alternative, and mostly semi-structured or unstructured, information often requires sophisticated techniques and multiple players to turn it into something of value. Last but not least, financial service providers might face potential information privacy concerns and security regulations. Therefore, the ability to evaluate the credit risk of borrowers with minimally accessible information is key to the burgeoning microloan market.

In other words, given a plethora of structured and unstructured individual behavioral data available across various channels, what information is most valuable to the financial credit market? This is a major challenge for many financial platforms today. In a recent study (Lu et al. 2020), the authors examined and compared the values from various types of digital footprint data for credit risk assessment. The authors provided, for microloan platforms, important managerial insights into what information is the most valuable, and hence,

should be efficiently combined with conventional data to maximize profits and minimize potential prediction bias. We will discuss this in more detail in the following subsections.

9.3.2 Recent Progress

In the past, financial risk assessment focused on conventional features such as loan characteristics, borrower characteristics, credit history, and social capital (Mersland and Strøm 2010) argued that a larger loan amount is associated with a higher probability of loan default. Everett (2015) found a positive relationship between interest rates and default risks. Serrano-Cinca et al. (2015) compared the default risks of 14 loan purposes and ranked them from most risky (e.g., small businesses) to least risky (e.g., weddings). Based on a field experiment in India, Field and Pande (2008) found that the type of repayment schedule (i.e., weekly or monthly repayment) had no effect on delinquency or default. Getter (2003) showed that the size of a household's payment burden (i.e., monthly payments relative to monthly income) had an insignificant effect on delinquency and only a very small effect on default behavior.

Regarding personal (borrower) characteristics, both hard and soft information showed effectiveness for evaluation of default risks (Emekter et al. 2015). Hard information refers to structured and quantifiable information. Examples include the borrower's credit scores and demographic information (Gross and Souleles 2002; Iyer et al. 2015; Lin et al. 2013; Ravina 2007). Specifically, Ravina (2007) discovered that low credit scores and low incomes were related to high default rates. Gross and Souleles (2002) revealed a significantly positive correlation between borrowers' ages and default risks. On the other hand, soft information covers unstructured information, such as loan histories, current circumstances, and social networks (Collier and Hampshire 2010; Iyer et al. 2015). For example, Lin et al. (2013) observed that friendship within a social network was associated with a lower ex-post default rate.

With more access to digital footprint data such as cellphone usage and social media information in recent years, several scholars have studied default risk prediction. Tan et al. (2016) and Mehrotra et al. (2017) utilized phone usage data, browsing logs, and mobility traces to evaluate borrowers' credit risks. Their empirical findings suggested that the accuracy of default prediction increased by approximately 4% after incorporating users' cellphone call and SMS (text message) network data. Lu et al. (2020) and Ma et al. (2018) found that phone usage patterns, including telecommunication patterns, mobility patterns, and app usage patterns, offered predictive capability for loan defaults. In the study of Björkegren and Grissen (2017), individuals in the highest quantile of risk as indicated by behavioral signatures in mobile phone data were 2.8 times more likely to default than those in the lowest quantile. Regarding the usage of social media information, Tan and Phan (2018) showed that incorporating social network information could improve creditworthiness prediction in microfinance by up to 300%. Yuan et al. (2018)

proposed a parallel topic modeling method for user-behavioral pattern mining on microblog data, having found that it outperformed traditional credit scoring methods. Ge et al. (2017) examined the predictive values of borrowers' self-disclosure on social media accounts and a more active social media presence (e.g., having larger social networks and posting more messages).

9.3.3 A Case Study of Microloan Risk Management

In this section, we discuss in detail a case study by Lu et al. (2020) on leveraging user digital footprint data to improve microloan risk management. In this study, the authors cooperated with a major microloan company in an Asian country to conduct a large field experiment from December 2 to 22, 2017.

One critical challenge in designing and evaluating financial risk models is that the counterfactual scenarios are completely unobserved—when someone's loan application is rejected, platforms do not observe any further loan repayment behavior of this applicant in the future. This can cause at least two issues. First, platforms cannot evaluate the "what-if" scenarios in the real-world setting—what if we approved a different set of loan applications? Would that lead to a lower default rate and better profitability? These counterfactual scenarios are impossible to observe because platforms simply do not record these alternative applicants' loan repayment behavior if their applications got rejected in the first place. Second, an even deeper issue is that the training data used for model training only contain the "approved sample" from the previous practice. Those counterfactual cases (i.e., applicants who got rejected) will never enter the training data. This may lead to serious problems if the approved sample is systematically different from the counterfactual cases in certain "sensitive" dimensions such as race or gender. Risk assessment models based on partially biased training data may lead to unexpected financial bias or service inequality.

To alleviate these concerns, Lu et al. (2020) partnered with the financial platform and designed a novel "mega-experiment." During the experimental period, the platform approved loan applications from all applicants (as opposed to the usual situation wherein only 40–45% of applicants are approved based on the personal experience of platform staff). It is worth noting that by approving all loan applications and tracking borrowers' repayment behaviors over time, the authors are able to recover all possible counterfactual cases—those applicants whom, under normal circumstances, would be rejected. This unique "mega-experimental" setting enables the authors to form an unbiased sample for model training by including behavioral patterns from the entire loan applicant population, and also allows for evaluation of the risk assessment model under various counterfactual scenarios that otherwise would go unobserved. The authors then collected a fine-grained dataset with detailed user digital footprint records from all loan applicants during the experimental period.

Furthermore, when calculating the profits of a microloan platform, the authors consider not only the losses from defaults but also the revenues from delinquent fine payments. Therefore, unlike previous studies with default indicators only (e.g., Duarte et al. 2012), they define a multiclass categorical credit risk indicator that captures the following borrowers' repayment behaviors: being delinquent, delinquent but not in default, and in default. The authors also consider the repayment rate and profit per loan (or loan profit) as alternative numerical credit indicators.

Given that their data cover multiple information sources, they construct and extract, as inspired from the existing literature, more than 100 features covering four main categories: commonly adopted conventional data (e.g., borrower demographic and socio-economic characteristics, credit history, and loan attributes), online activities (e.g., shopping), mobile activities (e.g., cellphone usage and location mobility traces), and social media activities. Those features were applied to the training of different state-of-the-art machine learning models and identify the values of different sources of information for credit risk assessment in the contexts of delinquent and default cases. For comparison, similar analyses were conducted using approved samples collected from the same platform. This comparison between the approved samples and the full applicant sample enables the authors to identify the potential financial impact of training-sample bias.

This empirical analysis yields several interesting findings. First, the prediction results show that among the four sets of features constructed, mobile activities, particularly cellphone usage and mobility trajectory features, present the highest predictive power, followed by online shopping activities. For social media users, social media presence and sentiment are also valuable in predicting users' repayment behavior. Interestingly, at a more granular level, among all of the alternative data-related features, consumption of gaming-related products (e.g., game app usage, amounts spent on game cards) ranks at the top.

Second, a platform welfare analysis indicates that, when predicting borrowers' credit risks with cellphone usage and mobility trace information, the corresponding loan permission strategy yields 15% more revenue gains to the microloan platform than does the case with conventional features only. The platform can achieve a further 7% revenue gain when making loan approval decisions based on credit risk prediction with all of the feature sets. In addition, under certain loan approval rates, loan permission strategies based on the predicted delinquent-but-not-in-default probabilities or numerical repayment rates and loan profits can bring higher revenue gains than the current industry practice that is based primarily on the predicted default probabilities. This finding confirms that on the premise of accurate risk prediction with alternative data, lending to borrowers with a certain level of delinquency risk, despite a relatively high default risk, can also yield positive economic gains.

Third, this study demonstrates that bias indeed exists if only approved samples are used or only conventional data are used for model training, which

can lead to significant losses of not only prediction accuracy but also economic gains for microloan platforms. Interestingly, these existing approaches tend to favor higher income and more-educated applicants from areas with a more developed economy. By leveraging alternative data, microloan platforms are more likely to include lower income and less-educated loan applicants from less-developed geographical areas—those historically disadvantaged populations that have been largely neglected in the past. This case study thus demonstrates the tremendous potential of leveraging alternative data to alleviate such inequality in the financial service markets while achieving higher platform revenues in the meantime.

The contributions of this case study are multifold. First, it is the first study to investigate the predictive power and financial value of multidimensional alternative data (including cellphone and mobile app usage, mobility trajectories, shopping behavior, and social media information) for borrowers' credit risk assessment and microloan platforms' revenue enhancement. Second, while previous studies simply focused on default probability, this study contributes to the literature with more sophisticated credit risk indicators. This extra information allows us to examine the trade-off between profits from delinquency and losses from default. Third, the unique field-experimental setting can examine "what-if" counterfactual scenarios under different loan permission strategies. By comparing the final rankings of loan applicants based on the predicted risk scores (i.e., the recommended approved loans) generated by different models, data or training sets, financial platforms can interpret not only "what" strategies but also "why" these strategies perform better and lead to higher economic returns to platforms. Such interpretability is critical and can help institutions understand where potential prediction bias and economic loss may come from, and how to address them. Fourth, such an approach enables microloan platforms to easily adopt cost-effective solutions based on what is easier to implement in practice. For example, training-sample bias has been a major challenge in both prior research and industry practice, due to practical data limitations. Incorporating alternative data can largely offset potential economic losses caused by training-sample bias and can lead to a significant improvement in platform revenues even when platforms have no access to the unbiased full sample of loan applicants during model training.

9.4 Applications of IoT-Based Data

9.4.1 IoT-Based Alternative Data

Advances in the Internet of Things (IoT) have empowered almost every industry to become more efficient and smart. Due to the large amount of alternative data produced, IoT adoption has opened up a completely new landscape in many sectors, including finance. For example, contemporary farming uses LIDAR technology (a surveying method that measures distance to a target by illuminating the target with laser light and measuring the reflected light with

a sensor), typically used in autonomous driving cars, to identify insects while robots pick weeds with the aid of computer vision. Videos, images, and voice capture technology can help farmers monitor the growing process of crops. Construction technology startups, using artificial intelligence and the IoT, have made construction work more like a manufacturing process. Versatile Natures, an Israeli company, offers a holistic view of a construction project by mounting IoT sensors under the hook of a crane (Versatile 2020). The sensors constantly collect and analyze data, with the goal of giving site managers actionable insights such as information on materials, redundancies, construction progress, and crane utilization. Inspirit IoT, an IoT startup from Illinois, aims to reduce the impact of on-site environments on workers' safety and construction schedules by implementing an AI-based algorithm over a traditional monitoring system to detect safety concerns (InspiritIoT 2020). Inspirit IoT makes sensors that measure environmental metrics, including temperature, humidity, carbon monoxide, etc. IoT's penetration into industries such as retail and wholesale, and hence a sustainable growing opportunity in finance, can be attributed to the following advantages by IoT.

9.4.1.1 Improved Customer Experience

Today, many retailers have increased their interaction with customers, but the IoT will bring a more personalized and meaningful experience. As ordinary "objects" become smart devices, the customer experience becomes fully digital, creating a growing trend of personalization. Relying on this interconnected environment, companies can design and create products and services centered on each consumer with data rendered from IoT.

9.4.1.2 Optimized Supply Chain Operations

"Industrial Internet" describes how companies can use cloud computing, mobile telecommunication, big data, and other technologies to closely integrate digital space with the real world, thereby improving operational efficiency and fostering innovation. It is expected that by 2030, the combination of industrial Internet and IoT devices will create an additional value of more than $14 trillion for the global economy.

In the face of increasingly complex supply chains, the growing importance of digital channels, and rising customer requirements, connected devices and products provide an opportunity for retailers to optimize operations. For example, wireless RF technology can improve the accuracy of inventory tracking, while data visualization technology makes it easier for employees to track the location of products in the supply chain. Merchants can even offer this service to customers, for example, to support customers in reviewing the progress of orders in the production and distribution process.

Store managers can also use online smart price tags to adjust pricing in real time, such as lowering the price of a promotional product or a poorly selling product, or increasing the price of a sought-after product. A fully integrated pricing system will help retailers better achieve price synchronization

between shelves, checkouts, and various channels, ensuring that online stores and physical stores are priced consistently.

In addition, merchants can integrate other IoT devices in the supply chain to further improve store operations and reduce costs. For example, sensors based on IoT technology can help store managers monitor and adjust lighting brightness and temperature to achieve energy savings and cost reductions while improving customer comfort.

Sensors can automate many of the tasks that currently need to be done manually, such as tracking inventory of individual items or adjusting prices, which will give salespeople more time to communicate with customers and further enhance in-store services.

As the above have clearly indicated, IoT technology helps firms to better understand once fragmented scenarios, leading to an improvement of business as a whole. From a FinTech perspective, the broad applications of IoT remain in the retail industry in which firms have the direct desire and incentives to push forward. The IoT has been maturing such that there are currently enough IoT sensors and devices that firms can start experimenting at a scale showing what the technology is truly capable of in various industries. As such, an enormous scale of alternative data is produced, intentionally or unintentionally, offering opportunities to study corporate business from multiple angles. This was utilized in the postcrisis period that was characterized by a low-interest rate environment such that investors spent large amounts of resources and capital in identifying anomalies through the alternative data of the IoT and rapid funding of their new discoveries.

We will discuss how IoT-based data is created and utilized in multiple business settings.

9.4.2 The Advance of the IoT-Driven Retail Industry

The retail industry caters to hundreds of millions of people each year. It also gathers and maintains multitudes of data—point of sales transactions, customer details like addresses, reviews on e-commerce websites, browsing history, vendor details, product details, etc. Given the proven effectiveness of the use of data to create sophisticated and accurate systems that learn through experience, it makes sense that retailers, with all the data in their possession, make use of this data and current technology to create vastly personalized buying experiences for customers, more efficient inventory and delivery processes, and increasingly secure environments for purchasing products.

E-commerce dramatically shifts the strategy and structure of firms that are active in domestic and international markets as companies race toward the digitization of their business processes (Koh et al. 2006). These shifts create new opportunities for small- and medium- sized enterprises (SMEs) that want to compete with the major incumbent players in markets. Most of them heavily rely on the technical assistance from large high-tech firms or market places, e.g., Google or Amazon, where customer relationships are nourished

and supported by digital tools. Retailers may have a lot of issues—ranging from inventory to location to customer service—but one of the largest challenges arises from unnecessary marketing failures that are fully self-inflicted. For instance, the brand is often "lost" from the moment of a product entering into the sales channel.

In the four key aspects of the retail business—product, efficiency, store, and sales—online brand promotion and e-commerce have gradually visualized the effective marketing of products and their impacts. In online stores and marketing, due to the complex and diverse sources of store traffic, it is often difficult to effectively precipitate user assets, while the effect of offline promotion is hard to track, resulting in the separation of online and offline data information. From a financial planning and marketing budgeting perspective, the question of who are the consumers at the other end of the product, often becomes a blind spot for the brand to perceive the user, making it extremely challenging to convert the sales into non-switching or long-term consumers. That, coupled with the problems of frauds, low-quality replica, and other issues, alongside the interference from certain unlicensed middlemen make the marketing cost of brand investment out of the real value of the target end users and service providers. Researchers (e.g., Peng 2012) have classified the factors tied to marketing failures into three major groups, including competition-specific, institution-specific, and resource-specific factors, that condition online retail companies' online strategies. These failures, unfortunately, though retail involves unlimited exogenous factors and multiple issues, stand out to become problems that arise from a failure to construct a clear and aligned story, strategy, and system as well as an inability to embrace the desire of customers.

Success in the retail industry has always been tough, but the current battleground in globalization or deglobalization presents new challenges and opportunities in a faster manner to all of the participants. Advertisements have been deemed "smart" as the internet with wide-bandwidth communication powers up the fast customization and deployment of ads with precision targeting of customers given their preferences and behaviors, learned from historical personal data or personal network research. Every company in every sector, including retail, is essentially advertising their dependence on big data. When constructing any transaction there are several steps that must be taken, either in a specific order, or in parallel, so a snag in one step tends to snowball into more problems down the line. Merchants, manufacturers, advertising agencies, logistics companies, and IT innovators hope that by adopting IoT solutions to cut costs, trace transportation, and use limited sale and marketing resources more efficiently, they can turn the capricious, fragmented, and spatially distributed world of e-commerce and retail into something more closely resembling what it is supposed to be—a service process for individuals. The focus is not only on how to sell goods or deliver faster, but also on turning retail and e-commerce into a regimented process that can be better understood and optimized. Amazon, for example, has a reputation for operating on a large scale of online presence and it has facilitated such a presence

through the emphasis on offline merchants and supply chain optimization with reinforcement from IoT solutions since 2014.

Like other industries which have undergone a digital revolution, thanks to the fast advance of IoT technology and blockchain, many aspects of the retail industry are also being revolutionized. Today, home appliances, home security and comfort products, and even health care products are becoming part of the IoT ecosystem. Retailers in home décor or consumer electronics can not only increase the sales of these connected devices, such as Home Depot, which has more than 600 "smart" products, but also leverage the data provided by these devices to extend the business scope to consumers' homes.

Some retailers are taking advantage of various interconnected products by becoming an integrated platform. The basic idea of these platforms is to make it easier for customers to communicate to each other's home devices. For example, Lowe's launched the "Smart Home Hub," the Iris platform, which can communicate with any device via networking technologies such as WIFI, ZigBee, or Z-Wave. The platform also has an open interface so manufacturers can interface with their products. Iris has enabled Lowe's to compete directly with telecom providers such as AT&T and Verizon, while also creating new opportunities for the company—working with manufacturers to integrate products into the Iris platform. In addition, Home Depot's Wink and Staples' Connect as well as other platforms are also being released.

Other types of retailers, such as grocery stores, can build or collaborate with such platforms. Connected platform provides retailers with another channel for direct interaction with customers, opening up a hidden treasure trove of customer data. This information covers almost every aspect of home life—from electricity use to consumption trends.

Under this context, the remaining chapter will focus on how IoT data is used by retailers and wholesalers, utilizing machine learning and deep learning algorithms to identify potential business locations, the creation of personalized recommendations on e-commerce websites and mobile applications, and how the data are used to identify and track both products and customers.

This practice of leveraging existing models (and/or creating newer ones) and algorithms to explore data to learn from experience has manifested in many ways in finance applications. The applications of machine learning, and more recently deep learning, have come a long way from targeting using predictive analytics in 2002 to targeting customers with emails about products it believed they would want next (Coussement and Van den Poel 2009), to Amazon using computer vision to create a frictionless grocery-buying experience for its customers (Grewal et al. 2017).

9.4.3 Categories of Data from IoT Ecosystem

In general, data from the IoT can be categorized based on the properties of the sensing, including, but not limited to, geolocation data from GPS, imaging data from video sensors, and data generated from other devices.

9.4.3.1 Geolocation-Based IoT Applications

Due to the wide adoption of smartphones in consumers throughout every country, the once seemingly impossible-to-acquire information on consumers' geolocation data can now be easily collected through either GPS, WiFi, or other wireless signals (Zhang et al. 2020). Advanced techniques, including machine learning, are then applied to geolocation data to extract insights which may be valuable for businesses and investors. It is estimated that there are 2 billion smartphone users in the world. The smartphone that people carry everywhere is in fact a tracking device that knows more about where people go and daily habits than even they do. Tracking people's smartphone locations is just one way that companies can acquire analytical insights.

Many companies in this sector are currently focusing on tracking bundle traffic in and around store locations. There are direct and indirect ways of collecting this geolocation data. The direct way collects data by tracking the location of users' cellphones. This kind of data can typically be purchased from mobile service providers, e.g., T-Mobile or Verizon, China Mobile, etc. The indirect way involves placing mobile advertisements on goods, e.g., bar codes or QR codes, so that a consumer's location can be instantly reported when they are triggered or scanned. Firms using the direct way include AirSage and Advan Research, while examples of the latter include Tencent and Walmart.

9.4.3.2 Case in Focus: AirSage

AirSage specializes in collecting and analyzing anonymous location data, such as cell phone and GPS data, to identify patterns (Smith et al. 2005). It does so by tracking mobile phone data using patented technology to capture and analyze mobile phone signal tower data. It has secured location data from various sources, including smartphone SDKs, fleet, and navigation systems. The data provided include both real-time and historical data.

AirSage distinguishes data based on transportation, travel and tourism, and commercial real estate. The company processes more than 15 billion mobile device locations everyday with the widest coverage of any location-based service provider in the United States. Note that data features a group breakdown on anonymous origin/destination matrix with time stamps.

In travel and tourism applications, AirSage's data will help identify visitor demographics, behaviors, and build seasonality trends with historical data in destination markets. AirSage covers most of the metro areas in the United States, so that anonymous devices in almost every city can be retrieved.

The GPS coordinates of cell phones collected over the course of a week, a month, etc., allow analysts to get an estimate of the number of visitors in a certain season. Analysts can then improve the accuracy of predictions for top-line revenue by combining the geolocation intelligence data as a proxy. Such cases include Six-flags, Disney, and Lululemon, all of which are publicly traded companies.

One outstanding firm for utilizing the indirect approach to collect consumer location and behavior profiling data is China's Tencent via its

Code system. Unlike the direct way, in which the location information is directly retrieved from the apps on smartphones attached to users, the indirect approach records the location of goods, through which the end-user profiles and locations can be acquired.

9.4.3.3 Case in Focus 2: Tencent Code Solution

With years of development in the consumer market of China, Tencent has evolved into one of the largest Internet-based value-added services providers in China. By adopting its latest cloud technology and IoT platform, Tencent has established large-scale, stable, and robust infrastructure and capabilities, complemented by online security, artificial intelligence, big data analytics, location-based services, and other proprietary technologies, to support ecosystem partners across various industries (Rong et al. 2015).

Like Amazon, Tencent has accumulated a presence in the Consumer Internet ecosystem over the years, building its strength in developing the largest consumer market in the world. The massive Weixin and QQ user bases serve as the "digital gateway" for industries, while official accounts, mini programs, mobile payments, marketing solutions, and WeChat Work serve as the "digital tools" that connect developers and enterprises to potential customers. One such example is the implementation of code tracking systems in the retail industry.

Tencent Smart Retail introduced the full-code digital marketing package which helped the retail industry to "seek people by goods" and better connect users. Though seemingly simple at first look, it involves a very sophisticated system. The core concept is that Tencent's products are digitized at the core, so each product has a unique digital ID, which will then allow the merchants to track the life cycle of each individual good.

In marketing—despite the inability to establish direct connections with consumers, the difficulty in managing channel terminals, and the lack of long-term operation mechanisms for digital assets—the application of Tencent Optima in different scenarios will help brands build full-chain digital management and solve the above problems. "No Field Verification" provides a completely new solution: goods to connect people. Through Tencent, every bottle of select beverages is printed with a QR code. The code can be entered into the official code system to make coupons. In this way, the brand realizes the visualization of offline users, allowing target consumers to connect, acquire insight, and operate. At the same time, goods have also broken through the original single consumer goods' attributes, becoming a direct communication medium between brands and consumers. Regardless of whether consumers buy online or offline, they can use the products themselves to achieve further connection with the brand. This is also an important activity to make brand marketing activities no longer strongly dependent on the field.

9.4.3.4 Image-Based IoT Applications

Other than geolocation data, image sensors have been widely used in collecting alternative data. In this section, different from the geospatial image data we discuss earlier, we mainly focus on the image data from IoT devices typically in home appliances and retail business.

The most popular motivation involved in creating such a dataset is the anticipation that information exploited from the dataset analytics will help to make recommendations to consumers to drive revenue or to better understand customers. Almost every retailer has a website or mobile app that utilizes recommendation systems to suggest products to consumers. Most of these systems use text-based data to provide such recommendations. This data primarily includes customer details like their demographics (age, gender, address, etc.) and their purchase history. For these algorithms to work, each product has data tags for its category. Using the data from each consumer, scores are created for products, then products with the highest scores are recommended to each consumer. These values would only exist for products that a customer had already bought and were created using the consumer's information online. However, offline stores will not be able to monitor this type of traffic without image-based sensors. So why can not traditional retailers be enabled with a similar technology? Thanks to the advancement of IoT technology and deep learning neural networks, image-based sensing data can be captured and analyzed with relatively low costs and high efficiency. Some companies started investigating into this domain. One retail analytics startup called Nomi developed their sensor platform that tracks customer behavior in traditional brick-and-mortar retail stores. Each arriving person in the store is assigned a tracking identifier using its advanced video camera. The cloud-based software analytic system then links the person's movements across Brickstream sensors, following the person wherever they go. The 3D sensors on the Nomi platform can see past overlapping objects to provide a truly accurate measurement of what people are actually doing in the store. With more than 140,000 sensors being used in stores located in more than 60 countries, Nomi's image data has a truly remarkable value proposition for retailers.

9.4.3.5 Other IoT Data Analytics

In retail industry, one easily accessible category of data is Point of Sale (POS or ePOS) data. Retailers today collaborate with suppliers and share sales and inventory information in order to increase profits. The most common source of shared data is driven from the UPCs scanned at checkout registers. POS data is typically sent electronically from retailers and distributors in transactions known as EDI 852 and EDI 867 or through vendor portals in files generated from their internal data warehouses. By summing up the Point of Sale (POS) data of approximately 2,000 American supermarket stores from 2001 to 2012 for every company, Ishikawa et al. (2016) compared the growth rate of the POS sales data with each company's actual sales. They discovered that the growth rates in quarterly sales for companies whose anchor products are daily

necessities in the United States were strongly related to the POS data's growth rate, thus demonstrating that nowcast (real-time observation of company sales) is possible, at least for this type of business enterprise.

Recent progress in the retail industry includes autonomous shopping centers powered by advances in computer vision technique. AmazonGo is a prime example of this. AmazonGo stores do not have any human staff or cash registers. Consumers enter these stores, pick up the groceries that they need and leave. Many aspects of this seemingly simple operation require the use of computer vision:

1. Customers need to be identified using facial recognition as soon as they enter the store.
2. Every product that is removed from the shelf needs to be accounted for. This operation has two aspects: the customer picking up the product needs to be identified, and the correct amount needs to be added to the customer's bill.
3. The product removed from the shelf must be accounted for and replaced with an identical item from the inventory.

Data from the purchase can be used to recommend products to the consumer in the future. Identification and tracking of customers and products require computer vision algorithms and fusion sensors to work in perfect conjunction to achieve accurate results. Every time a product is picked up, sensors need to detect the reduction in weight and pressure on the shelf, and the vision algorithms at work need to identify which product has been taken from the shelf.

Other retailers also use images and videos to create better shopping experiences for use in stores. Candy retailer Lolli & Pops leverages facial recognition to identify loyalty program members as they enter the store and proceeds to provide them with personalized recommendations. Walmart uses video data to monitor missed scans during checkouts and potential thefts. Schnuck Markets uses robots to monitor shelves and take stock of inventory.

Since there are many retailers that operate on-ground stores, there are many variations of technology being used to simplify product tracking and checkout. Many retailers use bar-code scanners at self-checkout counters. However, that still requires the consumer to individually scan each item. Redmon et al. (2016) propose a method called YOLO (You Only Look Once), which uses shape detection and categorization to identify all the products in a consumer's cart. This method consists of two CNNs. The first CNN is a GoogleNet-inspired network that classifies products into 17 predefined shapes, and then an R-CNN is used to classify the shapes into categorized products. The time taken to detect and classify the objects is approximately 69.3 ms per frame and is done with approximately 75% test accuracy.

IBM, partnering with Tesco, implemented a project that focused on monitoring products on shelves and using images to differentiate between similar products placed close to one another (Marder et al. 2015). It focuses on addressing two common problems encountered while detecting objects on shelves:

1. Images used in training sets are usually high-quality studio photos compared to the real-time lower quality images that need to be classified in stores.
2. Many products of the same type look alike, shape detection and categorization can be difficult for such products.

The model proposes a complex method that takes images from shelves and performs an initial classification of the products on the shelves. However, these classifications are not specific classifications, but similarity groups. Once products are grouped, features are extracted from the images and are used to classify the products more specifically.

9.5 Concluding Remarks and Future Directions

To conclude, we summarize the key takeaways of the discussion in this overview of research in economics and business-related fields utilizing alternative data. We reviewed the merits and scope of the different categories of alternative data and the methodologies that have been considered. In particular, we highlight textual analysis in corporate finance, image processing in financial markets and governance, digital footprints from social media and mobile devices, and IoT-based data retrieval and applications.

- Textual analysis increasingly requires advanced tools. Dynamic dictionaries/word lists and methodologies that effectively integrate domain expertise with effective information extraction from data are likely to become widely used.
- Dynamic alternative data such as videos and audios could provide valuable interaction to researchers and practitioners. Machine learning tools for analyzing them are available and represent rich information sources for social scientists to explore. While computer vision has not yet reached the same accuracy as the human eye, it has proved to be a viable automated alternative to traditional methods of engaging with customers. This has further resulted in cutting-edge research being carried out to improve existing applications and create new ones.
- Alternative data and their associated processing tools could prove fundamental in explainable AI and interpretation of complex, black-box machine learning models.

- Crowd-sourced data and individuals' "digital footprints" such as users' cellphone usage, online and mobile activities, social media, and social network activities, can be leveraged to improve financial predictions, profitability, and social welfare, for example, in credit risk assessment.
- IoT, as one of the breakthrough techniques in retail and wholesale industries, has been a powerful venue for financial analytics. The geolocation, image and transaction data streams from over 400 retailers and distributors have only been part of the alternative data that have been utilized. Within five years, the consensus view is that IoT data will become the largest volume of alternative data for finance analytics. Both new challenges and opportunities will emerge as more dynamic and advanced IoT devices are developed.
- It remains open how regulators and institutions can best address data privacy issues. More generally, it is a holy grail in data science to have multiparty usage of data while preserving privacy. Related are tools for merging traditional data with alternative data.
- In the spirit of the Lucas Critique, researchers should examine how the use of alternative data and research findings affect the data-generating process itself, together with subsequent socio-economic implications.

This article by no means illustrates all possibilities provided by the potential and large scale of alternative data. Given the pace of development in blockchain technology, deep learning techniques, and IoT technology, we expect research in this area would also evolve quickly. That said, the general trend and utility of using alternative data are here to stay and are likely to significantly impact the world of FinTech and business intelligence.

Acknowledgements We thank Samuel Petruzzi for excellent research assistance. This research was funded in part by the Ewing Marion Kauffman Foundation and R. C. Evans Fellowship. The contents of this publication are solely the responsibility of the authors.

References

Antweiler, Werner, and Murray Z. Frank. 2004. "Is all that talk just noise? The information content of internet stock message boards." *The Journal of Finance* 59 (3): 1259–1294.

Bai, John Jianqiu, Linlin Ma, Kevin A. Mullally, and David H Solomon. 2019. "What a difference a (birth) month makes: The relative age effect and fund manager performance." *Journal of Financial Economics* 132 (1): 200–221.

Barocas, Solon, and Andrew D. Selbst. 2016. "Big data's disparate impact." *California Law Review* 104: 671.

Björkegren, Daniel, and Darrell Grissen. 2017. "Behavior revealed in mobile phone usage predicts loan repayment." *arXiv preprint* arXiv:1712.05840.

Blei, David M., Andrew Y. Ng, and Michael I Jordan. 2003. "Latent dirichlet allocation." *Journal of Machine Learning Research* 3 (January): 993–1022.

Cao, Sean, Vivian Fang, and Lijun Lei. 2019. Negative peer disclosure. Technical report. *Working Paper*.

Carroll, Peter, and Saba Rehmani. 2017. "Alternative data and the unbanked." *Oliver Wyman Insights*.

Collier, Benjamin C., and Robert Hampshire. 2010. "Sending mixed signals: Multilevel reputation effects in peer-to-peer lending markets." In *Proceedings of the 2010 ACM conference on computer supported cooperative work*, 197–206. ACM.

Cong, Lin William, Pouyan Foroughi, and Nadya Malenko. 2019. "A textual factor approach to measuring corporate governance." *Work in Progress*.

Cong, Lin William, Tengyuan Liang, Baozhong Yang, and Xiao Zhang. 2019. "Analyzing textual information at scale." *Working Paper*.

Cong, Lin William, Tengyuan Liang, and Xiao Zhang. 2019. "Textual factors: A scalable, interpretable, and data-driven approach to analyzing unstructured information." *Working Paper*.

Coussement, Kristof, and Dirk Van den Poel. 2009. "Improving customer attrition prediction by integrating emotions from client/company interaction emails and evaluating multiple classifiers." *Expert Systems with Applications* 36 (3): 6127–6134.

De Mauro, Andrea, Marco Greco, and Michele Grimaldi. 2016. "A formal definition of Big Data based on its essential features." *Library Review* 65 (3): 122–135.

Dobbie, Will, Andres Liberman, Daniel Paravisini, and Vikram Pathania. 2018. *Measuring bias in consumer lending*. Technical report. National Bureau of Economic Research.

Duarte, Jefferson, Stephan Siegel, and Lance Young. 2012. "Trust and credit: The role of appearance in peer-to-peer lending." *The Review of Financial Studies* 25 (8): 2455–2484.

Emekter, Riza, Yanbin Tu, Benjamas Jirasakuldech, and Min Lu. 2015. "Evaluating credit risk and loan performance in online Peer-to-Peer (P2P) lending." *Applied Economics* 47 (1): 54–70.

Evans, James A, and Pedro Aceves. 2016. "Machine translation: mining text for social theory." *Annual Review of Sociology* 42.

Everett, Craig R. 2015. "Group membership, relationship banking and loan default risk: The case of online social lending." *Banking and Finance Review* 7 (2).

Field, Erica, and Rohini Pande. 2008. "Repayment frequency and default in microfinance: Evidence from India." *Journal of the European Economic Association* 6 (2–3): 501–509.

Fu, Runshan, Manmohan Aseri, Param Vir Singh, and Kannan Srinivasan. 2019. "'Un'fair Machine Learning Algorithms." Available at SSRN 3408275.

Ge, Ruyi, Juan Feng, Bin Gu, and Pengzhu Zhang. 2017. "Predicting and deterring default with social media information in peer-to-peer lending." *Journal of Management Information Systems* 34 (2): 401–424.

Gentzkow, Matthew, Bryan T Kelly, and Matt Taddy. 2017. *Text as data*. Technical report. National Bureau of Economic Research.

Getter, Darryl E. 2003. "Contributing to the delinquency of borrowers." *Journal of Consumer Affairs* 37 (1): 86–100.

Graham, John R., Campbell R. Harvey, and Manju Puri. 2016. "A corporate beauty contest." *Management Science* 63 (9): 3044–3056.

Grewal, Dhruv, Anne L. Roggeveen, and Jens Nordfält. 2017. "The future of retailing." *Journal of Retailing* 93 (1): 1–6.

Grimmer, Justin, and Brandon M. Stewart. 2013. "Text as data: The promise and pitfalls of automatic content analysis methods for political texts." *Political analysis* 21 (3): 267–297.

Gross, David B., and Nicholas S. Souleles. 2002. "An empirical analysis of personal bankruptcy and delinquency." *The Review of Financial Studies* 15 (1): 319–347.

Hanley, Kathleen Weiss, and Gerard Hoberg. 2019. "Dynamic interpretation of emerging risks in the financial sector." *Review of Financial Studies* Forthcoming.

He, Xianjie, Huifang Yin, Yachang Zeng, Huai Zhang, and Hailong Zhao. 2019. "Facial structure and achievement drive: Evidence from financial analysts." *Journal of Accounting Research*.

Huang, Allen H., Reuven Lehavy, Amy Y. Zang, and Rong Zheng. 2017. "Analyst information discovery and interpretation roles: A topic modeling approach." *Management Science* 64 (6): 2833–2855.

Huang, Xing, Zoran Ivković, J. Jiang, and I. Wang. 2018. "Swimming with the sharks: Entrepreneurial investing decisions and first impression." In *Presentation at the American Economic Association Annual Meeting*.

ICEYE. 2019. "ICEYE website." Accessed 2019. http://www.iceye.com.

InspiritIoT. 2020. "InspiritIoT website." Accessed 2020. http://www.inspirit-iot.com/.

Ishikawa, A., S. Fujimoto, and T. Mizuno. 2016. "Nowcast of firm sales using POS data toward stock market stability." In *2016 IEEE international conference on Big Data (Big Data)*, 2495–2499. December. https://doi.org/10.1109/bigdata.2016.7840887.

Iyer, Rajkamal, Asim Ijaz Khwaja, Erzo F.P. Luttmer, and Kelly Shue. 2015. "Screening peers softly: Inferring the quality of small borrowers." *Management Science* 62 (6): 1554–1577.

Jegadeesh, Narasimhan, and Di Andrew Wu. 2017. "Deciphering fedspeak: The information content of FOMC meetings."

Jia, Yuping, Laurence Van Lent, and Yachang Zeng. 2014. "Masculinity, testosterone, and financial misreporting." *Journal of Accounting Research* 52 (5): 1195–1246.

Koh, Chang E, Hae Jung Kim, and Eun Young Kim. 2006. "The impact of RFID in retail industry: Issues and critical success factors." *Journal of Shopping Center Research* 13 (1): 101–117.

Kpler. 2020. "Kpler website." Accessed 2020. http://www.kpler.com.

Lin, Mingfeng, Nagpurnanand R. Prabhala, and Siva Viswanathan. 2013. "Judging borrowers by the company they keep: Friendship networks and information asymmetry in online peer-to-peer lending." *Management Science* 59 (1): 17–35.

Loufield, Ethan, Dennis Ferenzy, and Tess Johnson. 2018. "Accelerating financial inclusion with new data." *Center for Financial Inclusion*.

Loughran, Tim, and Bill McDonald. 2011. "When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks." *The Journal of Finance* 66 (1): 35–65.

Lu, Tian, Yingjie Zhang, and Beibei Li. 2020. "Financial risk assessment with alternative data: Prediction, profit, and equality." *Working Paper*.

Ma, Lin, Xi Zhao, Zhili Zhou, and Yuanyuan Liu. 2018. "A new aspect on P2P online lending default prediction using meta-level phone usage data in China." *Decision Support Systems* 111: 60–71.

Marder, Mattias, Sivan Harary, Amnon Ribak, Y. Tzur, Sharon Alpert, and Asaf Tzadok. 2015. "Using image analytics to monitor retail store shelves." *IBM Journal of Research and Development* 59 (2–3): 3–1.

Mayew, William J, and Mohan Venkatachalam. 2012. "The power of voice: Managerial affective states and future firm performance." *The Journal of Finance* 67 (1): 1–43.

Mehrotra, Rishabh, Prasanta Bhattacharya, Tianhui Tan, and Tuan Phan. 2017. "Predictive power of online and offline behavior sequences: Evidence from a micro-finance context."

Mersland, Roy, and R. Øystein Strøm. 2010. "Microfinance mission drift?" *World Development* 38 (1): 28–36.

OrbitalInsight. 2020. "Orbital Insight website." Accessed 2020. http://www.orbitalinsight.com.

Peng, Mike W. 2012. "The global strategy of emerging multinationals from China." *Global Strategy Journal* 2 (2): 97–107.

Petrushin, Valery A. 2007. *Detecting emotions using voice signal analysis*. US Patent 7,222,075, May.

Ravina, Enrichetta. 2007. "Beauty, personal characteristics, and trust in credit markets." *Personal Characteristics, and Trust in Credit Markets (December 2007)*.

Redmon, Joseph, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. "You only look once: Unified, real-time object detection." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 779–788.

Rong, Ke, Guangyu Hu, Yong Lin, Yongjiang Shi, and Liang Guo. 2015. "Understanding business ecosystem using a 6C framework in Internet-of-Things-based sectors." *International Journal of Production Economics* 159: 41–55.

Rsmetrics. 2020. "Rsmetrics website." Accessed 2020. http://www.rsmetrics.com.

Schwenkler, Gustavo, and Hannan Zheng. 2019. "The network of firms implied by the news." Available at SSRN 3320859.

Serrano-Cinca, Carlos, Begoña Gutiérrez-Nieto, and Luz López-Palacios. 2015. "Determinants of default in P2P lending." *PloS One* 10 (10): e0139427.

Smith, Cyrus W., IV Clayton Wilkinson, Kirk Carlson, Michael P. Wright, and Rahul Sangal. 2005. *System and method for providing traffic information using operational data of a wireless network*. US Patent 6,842,620, January.

Spaceknow. 2019. "Spaceknow website." Accessed 2019. http://www.spaceknow.com.

Spireglobal. 2020. "Spireglobal website." Accessed 2019. https://maritime.spire.com/.

Tan, Tianhui, Prasanta Bhattacharya, and Tuan Phan. 2016. "Credit-worthiness prediction in microfinance using mobile data: A spatio-network approach."

Tan, Tianhui, and Tuan Q. Phan. 2018. "Social media-driven credit scoring: The predictive value of social structures." Available at SSRN 3217885.

Teoh, Siew Hong, Lin Peng, Yakun Wang, and Jiawen Yun. 2019. "Face value: Do perceived-facial traits matter for sell-side analysts?" *Working Paper*.

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

Umbralab. 2020. "Umbralab website." Accessed 2020. http://www.umbralab.com.

Versatile. 2020. "Versatile nature website." Accessed 2020. https://www.versatile.ai/.

Willis, Janine, and Alexander Todorov. 2006. "First impressions: Making up your mind after a 100-ms exposure to a face." *Psychological Science* 17 (7): 592–598.

Yuan, Hui, Raymond Y.K. Lau, Wei Xu, Zhaokang Pan, and Michael C.S. Wong. 2018. "Mining individuals' behavior patterns from social media for enhancing online credit scoring." In PACIS, 163.

Zhang, Qingquan, Xiyuan Zhang, Jiayue Wang, Yanjun Li, Beibei Li and Danxia Xie. 2020. "Economic impact analysis of the Coronavirus, an alternative data perspective." *Work in Progress.*

Zhu, Christina. 2019. "Big data as a governance mechanism." *The Review of Financial Studies* 32 (5): 2021–2061.