The Review of Financial Studies

Big Data in Finance

Itay Goldstein
University of Pennsylvania and NBER

Chester S. Spatt
Carnegie Mellon University and NBER

Mao Ye
University of Illinois at Urbana-Champaign and NBER

Big data is revolutionizing the finance industry and has the potential to significantly shape future research in finance. This special issue contains papers following the 2019 NBER-RFS Conference on Big Data. In this introduction to the special issue, we define the "big data" phenomenon as a combination of three features: large size, high dimension, and complex structure. Using the papers in the special issue, we discuss how new research builds on these features to push the frontier on fundamental questions across areas in finance—including corporate finance, market microstructure, and asset pricing. Finally, we offer some thoughts for future research directions. (*JEL* G12, G14, G3)

1. The "Big Data" Revolution

The digital age has created mountains of data that continue to grow exponentially. The International Data Corporation estimates that the world generates more data every two days than all of humanity generated from the dawn of time to the year 2003. This "big data" revolution is reshaping the financial industry. As the *Wall Street Journal* wrote, "Today, the ultimate Wall Street status symbol is a trading floor comprising Carnegie Mellon Ph.D.s, not Wharton M.B.A.s."[1] This industry transition has already started to affect

This introduction is written for a special issue of the *Review of Financial Studies* focused on big data in finance. The authors thank Ken French, Harrison Hong, Wei Jiang, Andrew Karolyi, and Jim Poterba for comments. We thank Jim Poterba and Carl Beck for help with the NBER Workshops on Big Data. Ye acknowledges support from National Science Foundation grant 1838183 and the Extreme Science and Engineering Discovery Environment (XSEDE).

[1] G. Rogow, "Meet the New Kings of Wall Street," *Wall Street Journal*, May 21, 2017, https://www.wsj.com/articles/the-quants-meet-the-new-kings-of-wall-street-1495389163.

The Review of Financial Studies 34 (2021) 3213–3225

doi:10.1093/rfs/hhab038 Advance Access publication April 5, 2021

the way we teach students. Along with the drop in the number of Master of Business Administration (MBA) programs, as well as the decline in applications and enrollment in MBA programs,[2] we see a surge of new programs such as Master of Business Analytics (also MBA).

The impact of big data on academic research in finance is also starting to reveal itself, but with it many questions emerge. The classical definition of big data as encompassing three V's (volume, velocity, and variety) has a strong relation to engineering and computer science, but it does not fully reflect the opportunities and challenges that big data poses to research and practice in finance. What does big data in finance actually mean? How can financial economists benefit from the big data revolution? Does big data open new research topics for financial economists or allow us to answer traditional questions in novel and more revealing ways? Is this really a revolution for finance research or just a continuation of a gradual change? After all, large datasets always have been a feature of research in finance.

In October 2018, the National Science Foundation (NSF) provided a joint grant to the National Bureau of Economic Research (NBER) and the National Center for Supercomputing Application (NCSA) at the University of Illinois at Urbana–Champaign that aimed to explore answers to these questions. Part of the grant is dedicated to education and outreach and support a series of NBER conferences to explore the future of big data research in finance. The summer conferences, organized by Toni Whited and Mao Ye, focus on tutorial sessions on big data techniques and presentations of early ideas on big data. The winter conferences, organized by Itay Goldstein, Chester Spatt, and Mao Ye, focus on completed papers using big data and related methodologies.

This special issue of the *Review of Financial Studies* (RFS) on big data in finance includes four papers from the first NBER-RFS Winter Conference on Big Data held on March 8, 2019, and two other papers that are closely related to this theme. The RFS has the tradition of encouraging scholars to pursue risky projects that have the potential to push the frontiers of research in finance. The NBER-RFS Conferences on Big Data and this special issue reflect the RFS's efforts to encourage the use of big data in finance studies and provide a natural complement to the RFS FinTech initiative that was featured in the May 2019 special issue (see Goldstein, Jiang, and Karolyi 2019 for an introduction).

In this introduction, we try to define what "big data" encompasses in the context of finance research. We then review the six papers included in the special issue, discussing how they are related to each other and to the general theme. Finally, we provide some thoughts for future research directions.

[2] C. Cutter, "Elite MBA Programs Report Steep Drop in Applications," *Wall Street Journal*, October 15, 2019, https://www.wsj.com/articles/elite-m-b-a-programs-report-steep-drop-in-applications-11571130001.

2. What Is Big Data in Finance Research?

It is fairly clear that a definition of big data in finance research should be different from ones that are used in engineering and statistics. Researchers in these disciplines focus on providing facilities and tools to capture, curate, manage, and process data. Financial economists, on the other hand, focus on applying these tools to address interesting economic questions. While it is risky to give a broad-based definition at this stage, we think it is important to try. The definition may be imprecise or incomplete, but it will provide a starting point for future iterations and corrections.

We thus propose three properties that together can potentially define big data in finance research: large size, high dimension, and complex structure. This definition combines the characteristics of the data with possible new research questions that cannot be addressed using "small data." We used this definition in our call for papers for the 2019 NBER-RFS Winter Conference. "Big data" papers can feature different combinations of these three properties. We now elaborate on what each of these properties captures.

Large size: As the term "big data" suggests, it would be impossible to avoid a reference to size. This feature means that data are large in an absolute or relative sense. A natural example for absolute size is transaction-level market microstructure data.[3] In a relative sense, big data is defined relative to the best existing "small data." Many datasets are small simply because they are a subset of a larger dataset. By subsampling or aggregating observations into categories or taking snapshots of activities in time series, large datasets are made smaller. Using the underlying larger dataset is important if it overcomes the sample selection bias in the small dataset, or if it captures important economic activities not depicted in the small dataset.

High dimension: "Big data" is not just about size. The second feature means that the data have many variables relative to the sample size. Machine learning, which is often thought of as a hallmark of big data research, is a common solution to the dimension challenge, and it is increasingly used in finance research. Machine learning techniques become economically meaningful if they satisfy, but are not limited to, the following criteria: (i) the actual economic problem involves lots of variables; (ii) the impact of the variables is highly nonlinear or involves interaction terms among the variables (high dimensionality of function class); and (iii) prediction is more important economically than statistical inference. The most natural research questions

[3] One day of current option trading data alone is roughly two terabytes. In the 2019 NBER-RFS Summer Conference on Big Data supported by the same NSF grant, the chief economist of the U.S. Securities and Exchange Commission (SEC), S. P. Kothari, pointed out that one of the biggest data collection efforts in finance is the Consolidated Audit Trial (CAT), which provides a single, comprehensive database enabling regulators to track more efficiently and thoroughly all trading activity in equities and options throughout the U.S. markets. https://www.sec.gov/news/speech/policy-challenges-research-opportunities-era-big-data.

occur when the decision-makers are machines, such as algorithmic traders or robo-advisors.

Complex structure: Finally, another important feature is that data are not in the traditional row-column format. Unstructured data include text, pictures, videos, audio, and voice. Unstructured data create value if they can measure economic activities that cannot be captured using structured data. Unstructured data are often high-dimensional by nature. The first step to analyze the data is usually to extract features from the unstructured data, often with help from deep learning and computer science. For example, researchers may extract semantic information from text using natural language processing (NLP), identify tone information from voice and audio using speech recognition, and recognize geographic or facial information from images and videos using computer vision (CV).

Overall, as these features reveal, big data is not only about the size of the data, but also about other characteristics. Developments in all three categories—increased availability and capability of handling large datasets, developments in methodologies to deal with high dimensionality, and the emergence of complex datasets with new methods for processing them—have led to the increased prominence of big data in finance research.

Each of the six papers in this special issue fits into one or more of these three categories. Anand et al. (2021) analyze the agency conflicts between brokers and their customers using a particularly large dataset established by the Financial Industry Regulatory Authority (FINRA) called the Order Audit Trail System (OATS). The dataset is big also in the relative sense because the OATS data include publicly unavailable information on broker identity and do not suffer from attrition and sample selection bias from self-reported data. Easley et al. (2021) also analyze large market microstructure data and, due to high dimensionality, apply machine-learning techniques to evaluate the effectiveness of traditional market microstructure measures after machines started dominating trading. The dataset in Giglio, Liao, and Xiu (2021) is distinctive not for its size, but for its high dimensionality. They also use machine-learning techniques to develop a new framework to deal with data snooping, a major concern in empirical asset pricing. Unlike the study by Giglio, Liao, and Xiu (2021), where high dimensionality comes from a large number of hypothesis tests that may lead to false positives in multiple testing, the high dimensionality in the paper by Erel et al. (2021) comes from the interaction terms and nonlinearity. Erel et al. (2021) show that machines can dominate humans in choosing directors, perhaps because machines suffer less from biases or agency conflicts. Papers by Benamar, Foucault, and Vega (2021) and Li et al. (2021) both use unstructured data. Benamar, Foucault, and Vega (2021) measure information demand and uncertainty using clickstream data provided by a vendor that transforms unstructured data into structured data. Li et al. (2021) transform unstructured data themselves

and develop a measure of corporate culture from textual data based on earnings calls.

These six papers cover topics in asset pricing, corporate finance, and market microstructure, demonstrating the broad scope of big data techniques in finance research. We now turn to describe these papers in more detail, their relation to one another, and to the broader theme.

3. What Is Included in This Special Issue?

In the first paper in the special issue, Erel et al. (2021) show that machine learning can outperform the actual selection of new board members, currently done by humans. They demonstrate that directors who algorithms predict will perform poorly indeed do, compared to a realistic pool of candidates in out-of-sample tests.[4] Relative to algorithm-selected directors, management-selected directors who later receive predictably low shareholder approval are more likely to be male, have larger networks, and sit on more boards. One possibility is that firms that nominate predictably unpopular directors tend to be subject to homophily, while the algorithm selects a more diverse board. The authors also find that firms that nominate predictably poor directors suffer from worse corporate governance structures, which suggests that agency conflicts could be a driver for the distortion in selecting directors.

The analysis in this paper is among the first applications of machine-learning methods in corporate finance, demonstrating the broad appeal of these methods across areas of finance. The authors demonstrate the usefulness of these methods by showing that traditional OLS results are unable to adequately predict director performance. They attribute these findings to nonlinearity and interactions among variables being key in predicting future performance. These results raise interesting questions for future research, trying to understand why the interaction among variables and/or the nonlinearity in the effects of different variables are so important.

Machine learning in a corporate finance context is a key characteristic of the second paper in the special issue, written by Li et al. (2021). The authors try to quantify the notion of corporate culture and understand its implications across firms. Corporate culture is important because it is perceived to be a key factor behind many business successes and failures (Graham et al. 2018), and it is thought to be able to solve problems that cannot be regulated properly ex ante (Guiso, Sapienza, and Zingales 2015). Data challenges have always made studying corporate culture a formidable task. Despite the boom in empirical studies since the mid-1980s,[5] variables of economic interest may not

[4] The task of measuring the performance of an individual director is challenging because directors generally act collectively on the board. The authors' main measure of director performance is the level of shareholder support in annual director reelections, because Hart and Zingales (2017) emphasize that directors' fiduciary duty is to represent the interests of the firm's shareholders.

[5] See Einav and Levin (2014).

be measured perfectly with structured data. Indeed, in the interview evidence by Graham et al. (2018), corporate executives suggest 11 sources of data to measure corporate culture, most of which are unstructured data.

Li et al. (2021) make progress by using NLP models to extract key features of corporate culture from earnings call transcripts, which is one source of data suggested by corporate executives. They use a semi-supervised machine-learning approach with word embedding for textual analysis instead of the traditional "bag of words" approach (Loughran and McDonald 2011). The "bag of words" approach is good at predicting the tone of a document by counting positive or negative words, but it is hard to capture important semantic information in an earnings call. The authors provide a method to decompose corporate culture onto a five-dimensional space of innovation, integrity, quality, respect, and teamwork, which are the five most-often mentioned values by the S&P 500 firms (see Guiso, Sapienza, and Zingales 2015). Guiso, Sapienza, and Zingales (2015) find that the culture-performance link is more significant during periods of distress, and that corporate culture is shaped by major corporate events, such as mergers and acquisitions. They show that firms scoring high on the cultural values of innovation and respect are more likely to be acquirers, and firms closer in cultural value are more likely to merge.

Another area where machine-learning methods have much unexploited potential is market microstructure. The third paper in the special issue, by Easley et al. (2021), explores an application for analyzing whether machine-based trading affects the efficacy of market microstructure measures that were developed before machines dominated trading volume. Specifically, Easley et al. (2021) examine whether six extant market microstructure measures—the Roll measure, the Roll impact,[6] volatility (VIX), Kyle's λ, the Amihud measure, and the volume-synchronized probability of informed trading (VPIN)—can still predict the future values of price and liquidity.

The authors find that the answer is still positive after the rise of high-frequency and machine-based trading. The functional form to make such predictions, however, depends on the application. For example, for making predictions within the same asset, a simple logistic regression performs almost as well as complex machine-learning techniques. One explanation is that there is already a deep understanding of the market structure for a single asset. For making predictions across assets, however, machine learning strictly dominates simple logistic regression.[7] Although the rise of high-frequency and machine-based trading has made cross-asset trading more the norm, few market microstructure theories show how these cross-asset effects should, or even could, occur. Easley et al. (2021) provide strong evidence that the interactions

[6] Roll impact is the Roll measure divided by the dollar value traded over a certain period.

[7] The cross-asset effects in their paper mean using market microstructure measures in one asset, such as equity futures, to predict price and liquidity dynamics of another asset, such as fixed-income futures.

among assets can predict market outcomes and that machine learning helps address challenges from high dimensions in cross-asset market microstructure.

Thinking about big data in the context of market microstructure research more broadly, it is often noted that large datasets were the norm in this literature for a long time. Yet, the fourth article in the special issue, by Anand et al. (2021), pushes the boundary in this sense, analyzing a particularly large dataset to identify agency conflicts between institutional traders and their brokers. To find such agency conflicts, it is very instructive to know the brokers' identities, which are missing in the publicly available TAQ data. Self-reported data would suffer from attrition or sample selection bias issues. Anand et al. (2021) use OATS data to surmount these two challenges, as it is comprehensive regulatory data from FINRA.

The authors find that brokers, who route more orders to affiliated alternative trading systems (ATSs), offer lower execution quality (lower fill rates and higher implementation shortfall costs) for their customers. Therefore, these brokers take the private benefit by increasing the market share and fee revenues of their own ATSs, but do not necessarily satisfy their fiduciary responsibilities to achieve the best execution for their customers. As Anand et al. (2021) use a large and comprehensive dataset, a subsample of the dataset can still generate enough statistical power, which allows the authors to establish causality using a unique controlled experiment that overlaps with their sample period: the SEC Tick Size Pilot (TSP). Based on a triple-difference analysis, the authors find that execution quality improves for TSP-treated stocks for orders handled by brokers who prefer affiliated ATSs since the TSP imposes constraints on brokers to route orders to ATS venues.

The fifth paper in the special issue, written by Benamar, Foucault, and Vega (2021), also analyzes a large dataset in the context of trading in financial markets. Another important feature of this paper is the processing of unstructured data. Here, unlike in Li et al. (2021), who process such data themselves, Benamar, Foucault, and Vega rely on commercial data vendors that preprocessed the raw and unstructured data into structured data. This is part of the trend in the era of big data: along with the boom of data availability, the data vending industry has grown as well. J. P. Morgan's *Big Data and AI Strategies* report provides a 78-page summary of available data vendors.[8] Benamar, Foucault, and Vega (2021) measure information demand with webpage clickstream statistics from Bitly, a URL-shortening service provider.[9] They use this to understand the role of uncertainty in financial market trading, a topic that has long occupied academics studying financial markets.

Benamar, Foucault, and Vega (2021) show that information demand is a good proxy for uncertainty because, based on their theory, an exogenous increase in

[8] Kolanovic and Krishnamachari (2017).

[9] A shortened URL is a compressed link to certain webpages. For example, https://academic.oup.com/rfs/advance-articles can be shortened to https://bit.ly/3mS7yDv.

an asset's uncertainty motivates investors to search for more information on it. The search for information, however, cannot fully neutralize the increase in uncertainty. Thus, a stronger information demand about future interest rates ahead of macroeconomic and monetary policy announcements (MMPAs) implies that U.S. Treasury yields exhibit both higher uncertainty and stronger sensitivity to MMPAs. They find that a one-standard-deviation increase in the number of Bitly clicks on the news related to nonfarm payroll (NFP) in the two hours preceding NFP announcements raises the sensitivity of U.S. Treasury note yields by 4 to 6 basis points (bps), depending on maturity. The increase is economically significant because the unconditional sensitivity of U.S. Treasury note yields to NFP announcements varies between 3. 5 bps and 7 bps (depending on maturity) during their sample period. They also find that such predictability mostly comes from clicks within two hours before the announcement, which highlights the usefulness of high-frequency data for measuring information demand and uncertainty.

Finally, closing the special issue is the paper by Giglio, Liao, and Xiu (2021). This paper belongs to the asset pricing literature, in which machine-learning methods have already been explored in some depth. A recent special issue of the *Review of Financial Studies* featured some of this research in the context of new methods for the cross-section of returns (see Karolyi and Van Nieuwerburgh 2020 for an introduction). Giglio, Liao, and Xiu (2021) show how machine learning can be applied by proposing a new framework to rigorously perform multiple hypothesis testing in linear asset pricing models, with a focus on addressing data snooping.

The dimension challenge in Giglio, Liao, and Xiu (2021) comes from multiple testing—that is, when trying to identify which factors in the "factor zoo" add explanatory power for the cross-section of returns or to identify which funds among thousands of funds can produce positive alpha. If the number of tests is high due to a large number of factors or funds, a potentially large fraction of the tests will be positive purely by chance and lead to a high false discovery rate. Giglio, Liao, and Xiu (2021) solve data snooping and false positives using a combination of matrix completion, wild bootstrap, screening, and false discovery control. Matrix completion, a machine-learning technique, helps them to interpolate missing data and latent factors. The latent factors constructed from machine learning correct correlation among alpha test statistics. Bootstrap and screening improve the robustness of multiple testing in a finite and skewed sample. The authors illustrate their framework using a hedge fund dataset, but their toolbox can be applied in other asset pricing research as well.

4. Where Does Big Data Research Go from Here?

The six papers in this special issue can provide a starting point for discussing big data in finance. As a burgeoning field, big data and machine learning raise

many new questions. We discuss several promising lines of research. We believe the list will continue to grow and be refined over time.

4.1 Machine learning and learning machines

To date, most research using machine learning, including papers in this special issue, use machine learning to understand human behavior. One promising area of machine learning in finance is when the decision-makers are machines. For example, most existing machine-learning research in asset pricing uses monthly return data from CRSP or quarterly holding data from 13F filings. Yet traders who apply machine learning techniques often operate at a horizon that is much less than a month. Hedge funds such as Renaissance, Two Sigma Investments, D. E. Shaw Group, PDT Partners, and TGS Management Company make thousands of trades and manage tens of billions of dollars in investor assets.[10] These firms, which are faster than most traditional funds but slower than high-frequency traders, are largely outside the radar of the academic finance literature. One exception is Chinco, Clark-Joseph, and Ye (2019), who find that machine learning aims to predict news at the minute-by-minute horizon. A promising new line of research is to bridge the gap between studies that focus on the monthly horizon or above and the studies on high-frequency traders, which focus on horizons below a second. In this underexplored territory, applying machine learning is not only natural but also necessary. Just as insights into human behavior from the psychology literature spawned the field of behavioral finance, so can insights into algorithmic behavior (or the psychology of machines) spawn an analogous blossoming of research in algorithmic behavioral finance.

4.2 Feedback effects of the big data revolution

Once machines become decision-makers, will corporations change their behavior? The widespread application of machine learning in the investment community and the feedback effects between the secondary market and corporate decisions (Bond, Edmans, and Goldstein 2012) imply that firms should respond to the big data revolution. While no papers in this special issue examine feedback effects, we saw related studies at the 2020 NBER-RFS Winter Conference on Big Data. Cao et al. (2020) find that firms adjust their 10-Ks and 10-Qs to cater to machine readers. The next step following their research is perhaps to examine whether firms react to the big data revolution when making real decisions. For example, as investors increasingly become machines, will firms increasingly pursue shorter-term projects? Does the advent of "big data" reduce managers' incentives to learn from market prices because firms now have more information sources, or does it increase incentives because prices aggregate more information from the "big data" collected by investors?

[10] G. Zuckerman and B. Hope, "The Quants Run Wall Street Now," *Wall Street Journal*, May 21, 2017, https://www.wsj.com/articles/the-quants-run-wall-street-now-1495389108.

4.3 Heterogeneous impact of the big data revolution

Although big data provides more information for sophisticated players such as institutional investors and firms, the impact of big data may not always be positive. Chawla et al. (2019) show that social media, which allows enthusiasm for the market to spread much more widely than it would have otherwise (Shiller (2015)), can push price away from fundamentals. In Chawla et al. (2019), the price pressures led by retail traders quickly revert, probably because sophisticated arbitragers rapidly jump in and trade against retail behavioral bias. We witnessed a much more significant impact of social media during the GameStop episode in January 2021. Retail traders coordinated using social media, resulting in the hedge fund Melvin Capital losing 53%.[11] The interaction between retail and sophisticated investors leads to extreme market volatility. The impact of big data on different types of agents and its aggregated effect on society will be an interesting new direction to explore.

4.4 More complex data

Big data in finance starts from analyzing large-size data such as trades and quotes. More recent development allows researchers to use natural language processing (NLP) to extract information from unstructured data such as text (Gentzkow, Kelly, and Taddy 2019). A promising research line is to analyze data of more complex structures, such as audio, video, and images if these more complex data provide additional insights. For example, Li et al. (2021) use the transcripts of earnings call as input for their analysis in this special issue. The earnings call transcripts are small data when we compare them with the audio file that generates the transcripts. Mayew and Venkatachalam (2012) show that managerial vocal cues contain information about a firm's fundamentals, incremental to information conveyed by linguistic content. As the NBER-RFS Big Data Conference evolves, we see submissions using more complex datasets, such as satellite images (Gerken and Painter 2020). More complex datasets create value for finance researchers if they measure economic activities that cannot be captured using simpler data.

4.5 Regulations

As machines start to be major players in many areas such as trading (Angel, Harris, and Spatt 2015), it will be interesting to examine whether existing regulations, which are designed mostly for humans, need to be adapted to an environment with machines. O'Hara, Yao, and Ye (2014) provide one example for such need. Regulators used to consider trades of less than 100 shares to come from retail traders, and would exempt these odd lots from the reporting requirement. Yet informed traders later became major sources of odd lots by

[11] J. Chung, "Melvin Capital Lost 53% in January, Hurt by GameStop and Other Bets," Wall Street Journal, January 31, 2021, https://www.wsj.com/articles/melvin-capital-lost-53-in-january-hurt-by-gamestop-and-other-bets-11612103117.

using algorithms to slice and dice their orders to less than 100 shares to escape the reporting requirement. While much of our financial regulatory system focuses on actual realized transactions, assessing problematic aspects of the underlying algorithms is arguably more fundamental and cuts to the heart of such issues as the possibility of front running by market markers, whether brokers have satisfied their best execution responsibilities, and whether insiders are exploiting informational advantages. Spatt (2020) discusses how regulations designed years ago need to be adapted to modern reality. The traditional focus of regulators has not emphasized biases in specific algorithms.

The other promising line of research on big data will be on privacy regulations and the fairness of algorithms and data (e. g., Kearns and Roth 2020). The question becomes extremely important because algorithms and data increasingly became a major resource for the economy, particularly for finance. Back in 2017, the *Economist* published a story titled "The World's Most Valuable Resource Is No Longer Oil, but Data," which called for new regulations for the data economy.[12] Who owns the data, what is the price of the data, and what is the impact of unfair access to data? Easley, O'Hara, and Yang (2016) provide a theoretical analysis of the issue. It would be interesting to explore this topic empirically.

4.6 Theory

The papers in this special issue are predominantly empirical, but theoretical work is also important for big data in finance. Although high-dimensional data are often defined as when the number of variables is larger than the number of observations (Martin and Nagel 2019), the dataset frequently used in finance research is typically large enough to cover the number of variables. The success of machine learning often comes from high-order interaction terms between variables (Mullainathan and Spiess 2017). Indeed, the success of machine learning for the papers in this issue also comes from nonlinear terms and interactions between variables. Such high-order interactions are a natural place to develop new theoretical models to explain why one economic variable's impact depends on its interaction with another variable. The nonlinearity also motivates theory models to explain why a variable's impact depends largely on its value. Machine learning is one way to describe the world, and we also need theory to explain the world.

Theory may become more important in the era of machine learning and artificial intelligence for one simple reason. Human judgment can be inconsistent, whereas machines tend to make consistent decisions based on their model. Li and Ye (2020) find that their theory model can generate quantitatively accurate predictions for market liquidity in cross-section and after corporate events such as stock splits, probably because liquidity providers

[12] "The World's Most Valuable Resource Is No Longer Oil, but Data," *Economist*, May 6, 2017, https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data.

are now algorithms, and these algorithms probably make decisions using similar models to the theoretical models in Li and Ye (2020).

4.7 Interdisciplinary collaborations

Future work on big data in finance may involve more scholars from other fields. We believe such collaborations will expand the tools and scope of research in finance and economics and help researchers overcome big data challenges.

Researchers can overcome the large-size challenge by collaborating with supercomputing centers. The NSF's Extreme Science and Engineering Discovery Environment Project (XSEDE) provides computing resources and staff support to manage and store large datasets free of charge. NBER has posted videotaped lectures for researchers in economics and finance on the application process for such free resources on the webpage for the 2018 Summer Conference on Big Data.[13]

Researchers can overcome the high-dimension challenge and the complex-structure challenge by collaborating with scholars from the fields of math, statistics, and computer science. The recent development in deep-learning models like natural language processing (NLP), speech recognition, and computer vision (CV) helps researchers parse textual, verbal, and visual data. Researchers can also choose to work with data vendors. J. P. Morgan's *Big Data and AI Strategies* report provides a list of vendors for alternative data, such as satellite photos, sentiment measures, and credit card usages.

The NSF lists big data as one of its 10 big ideas and provides funding to support innovative, interdisciplinary research in data science. We hope this special issue is only a starting point, and that we will see more research at the intersection of big data, finance, and public policy for many years.

References

Anand, A., M. Samadi, J. Sokobin, and K. Venkataraman. 2021. Institutional order handling and broker-affiliated trading venues. *Review of Financial Studies* 34:3364–402.

Angel, J. J., L. E. Harris, and C. S. Spatt. 2015. Equity trading in the 21st century: An update. *Quarterly Journal of Finance* 5:1–39.

Benamar, H., T. Foucault, and C. Vega. 2021. Demand for information, uncertainty, and the response of US Treasury securities to news. *Review of Financial Studies* 34:3403–55.

Bond, P., A. Edmans, and I. Goldstein. 2012. The real effects of financial markets. *Annual Review of Financial Economics* 4:339–60.

Cao, S., W. Jiang, B. Yang, and A. L. Zhang. 2020. How to talk when a machine is listening: Corporate disclosure in the age of AI. NBER Working Paper 27950.

Chawla, N., Z. Da, J. Xu, and M. Ye. 2019. Information diffusion on social media: Does it affect trading, return, and liquidity? Working Paper.

Chinco, A., A. D. Clark-Joseph, and M. Ye. 2019. Sparse signals in the cross-section of returns. *Journal of Finance* 74:449–92.

[13] http://www2.nber.org/si2018_video/bigdatafinancialecon/.

Easley, D., M. Lopez de Prado, M. O'Hara, and Z. Zhang. 2021. Microstructure in the machine age. *Review of Financial Studies* 34:3316–63.

Easley, D., M. O'Hara, and L. Yang. 2016. Differential access to price information in financial markets. *Journal of Financial and Quantitative Analysis* 51:1071–1110.

Einav, L., and J. Levin. 2014. Economics in the age of big data. *Science* 346(6210):715.

Erel, I., L. Stern, C. Tan, and M. S. Weisbach. 2021. Selecting directors using machine learning. *Review of Financial Studies* 34:3226–64.

Gentzkow, M., B. Kelly, and M. Taddy. 2019. Text as data. *Journal of Economic Literature* 57:535–74.

Gerken, W. C., and M. Painter. 2020. The value of differing points of view: Evidence from Financial Analysts' Geographic Diversity. Working Paper.

Giglio, S., Y. Liao, and D. Xiu. 2021. Thousands of alpha tests. *Review of Financial Studies* 34:3456–96.

Goldstein, I., W. Jiang, and G. A. Karolyi. 2019. To FinTech and beyond. *Review of Financial Studies* 32:1647–61.

Graham, J. R., J. Grennan, C. R. Harvey, and S. Rajgopal. 2018. Corporate culture: The interview evidence. Working Paper.

Guiso, L., P. Sapienza, and L. Zingales. 2015. The value of corporate culture. *Journal of Financial Economics* 117:60–76.

Hart, O., and L. Zingales. 2017. Companies should maximize shareholder welfare not market value. *Journal of Law, Finance, and Accounting* 2:247–74.

Karolyi, G. A., and S. Van Nieuwerburgh. 2020. New methods for the cross-section of returns. *Review of Financial Studies* 33:1879–90.

Kearns, M., and A. Roth. 2020. *The ethical algorithm: The science of socially aware algorithm design*. Oxford: Oxford University Press.

Kolanovic, M., and R. Krishnamachari. 2017. *Big data and AI strategies: Machine learning and alternative data approach to investing*. J. P. Morgan. Available at https://www.cfasociety.org/cleveland/Lists/Events%20Calendar/Attachments/1045/BIG-Data_AI-JPMmay2017.pdf.

Li, K., F. Mai, R. Shen, and X. Yan. 2021. Measuring corporate culture using machine learning. *Review of Financial Studies* 34:3265–315.

Li, S., and M. Ye. 2020. The share price that maximizes liquidity: A tale of two discretenesses. Working Paper.

Loughran, T., and B. McDonald. 2011. When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *Journal of Finance* 66:35–65.

Martin, I., and S. Nagel. 2019. Market efficiency in the age of big data. NBER Working Paper 26586.

Mayew, W. J., and M. Venkatachalam. 2012. The power of voice: Managerial affective states and future firm performance. *Journal of Finance* 67:1–43.

Mullainathan, S., and J. Spiess. 2017. Machine learning: An applied econometric approach. *Journal of Economic Perspectives* 31:87–106.

O'Hara, M., C. Yao, and M. Ye. 2014. What's not there: Odd lots and market data. *Journal of Finance* 69:2199–236.

Shiller, R. J. 2015. *Irrational exuberance*. Princeton, NJ: Princeton University Press.

Spatt, C. S. 2020. Is equity market exchange structure anti-competitive? Working Paper.