Big Data in Finance: Highlights from the Big Data in Finance Conference Hosted at the University of Michigan October 27-28, 2016

Summary Paper
Released: February 23, 2018

The University of Michigan Center on Finance, Law, and Policy
Professor Michael S. Barr, Faculty Director
Brian T. Koziara, Research Assistant

Mark Flood, U.S. Office of Financial Research, Panel Moderator
Professor Alfred Hero, University of Michigan, Panel Moderator
Professor H.V. Jagadish, University of Michigan, Panel Moderator
Co-Rapporteurs

Abstract

How can financial data be made more accessible and more secure, as well as more useful to regulators, market participants, and the public? As new data sets are created, opportunities emerge. Vast quantities of financial data may help identify emerging risks, enable market participants and regulators to see and better understand financial networks and interconnections, enhance financial stability, bolster consumer protection, and increase access to the underserved. Data can also increase transparency in the financial system for market participants, regulators and the public. These data sets, however, can raise significant questions about security and privacy; ensuring data quality; protecting against discrimination or privacy intrusions; managing, synthesizing, presenting, and analyzing data in usable form; and sharing data among regulators, researchers, and the public. Moreover, any conflicts among regulators and financial firms over such data could create opportunities for regulatory arbitrage and gaps in understanding risk in the financial system.

The Big Data in Finance Conference, co-sponsored by the federal Office of Financial Research and the University of Michigan Center on Finance, Law, and Policy, and held at the University of Michigan Law School on October 27-28, 2016, covered a number of important and timely topics in the worlds of Big Data and finance. This paper highlights several key issues and conference takeaways as originally presented by the contributors and panelists who took part.

Big Data in Finance
Table of Contents

What is Big Data?

In today's complex world, hundreds of millions of transactions and events occur between parties each day. With the advent of technologies that allow mankind to record and analyze individual data points associated with these events and transactions more easily than ever before, new methods, techniques, and uses for such information and analysis have become readily apparent. This mass of information has been generally termed "Big Data" in recognition of its tremendous scope and future potential to reshape fundamental elements of society. Big Data has already found applications in weather, healthcare, crime prevention, transportation and infrastructure, policymaking, and the world of finance, as individuals and organizations in each of these fields seek to harness the insights and power of this information to their benefit.

The term "Big Data" therefore, describes the massive quantities of data generated and collected in commercial transactions and other events. Big Data is a case of a difference in scale becoming a difference in kind. Legacy processes may be overwhelmed by the scale of Big Data in any of the "Four V" dimensions (i.e., volume, velocity, variety, and veracity), rendering familiar tasks impossible in practice. On the other hand, certain insights that are unavailable at smaller scales may emerge in Big Data. While each individual data point created in financial transactions or events might not be particularly valuable on its own, when taken and analyzed in aggregate, these combined data points can yield powerful insights for firms and policymakers. Until recently, society lacked the tools or technology to generate, capture, and analyze this information and apply its insights to decision-making processes. Now that such tools and technology have evolved, many questions have arisen, including many addressed at the Big Data in Finance Conference and summarized in this paper.

So, who has an interest in Big Data? Scientists, engineers, financiers, and policymakers all have an interest in gathering, securing, and protecting this information and using it in different ways. Those who analyze Big Data, however, are not the only ones who can understand and have a voice in how it is handled. Everyday consumers are increasingly becoming aware of how data from their transactions and daily lives is generated, recorded, stored, moved, and analyzed.

A particularly relevant field in which Big Data has come to the forefront is the financial services industry. Multiple players and parties, ranging from individual consumers, asset managers, traders and brokers, to banks and policymakers, are now utilizing and harnessing the power of Big Data in new and unprecedented ways to predict financial market trends, tailor products to consumers, and monitor and manage idiosyncratic and systemic risk. The world of finance also poses interesting questions for applications of Big Data and offers a case study in the interactions between consumers, firms, and regulators. Indeed, sensitive consumer data from the financial services industry can provide great benefits, helping to manage risk and support economically efficient solutions for individuals and firms. Finance is also a sector in which the risks and problems of Big Data are apparent—particularly when it comes to consumer privacy, data security, and questions about who should own, have access to, and be able to use and analyze data. Finance thus provides us with a helpful framework for analyzing issues about Big Data that might also be relevant in other fields.

What Are We Using Big Data For?

Big data has already found many uses and applications within and outside of the financial services industry. It has been utilized to better predict and model medical diagnoses and prognoses, to approximate and model optimal routing for transportation and logistics, and to

better determine resource allocations for a whole host of public policy concerns, including crime prevention, public health, and education. In the financial services sector, firms such as hedge funds leverage powerful algorithms to comb through data to discern small mispricings that might support profitable trading strategies. Foreign exchange traders use large quantities of data to set exchange rates more efficiently, and financial regulators and policymakers analyze massive amounts of historical data to predict trends in their relevant markets. Credit card companies rely on an individual's credit score, a function of past financial data, in determining his or her suitability for a card and general creditworthiness. Investment banks use past trading price history in valuation methods for corporate transactions such as mergers and acquisitions, along with a host of other financial metrics around the profitability and performance of individual business divisions within a valuation target. In short, big data is already being used and leveraged by a number of players within the financial services industry, and the ways and speed with which it is being incorporated into everyday decision making are growing quickly.

One of the most interesting areas in which big data can serve a societal benefit with regard to the financial services industry has to do with financial stability. Both prudential regulators and risk managers within firms can leverage the better understanding of market trends and performance afforded to them by big data to predict, prevent, and manage risk within their firms or financial markets more broadly. These regulators and risk managers can examine firm-specific data as well as data on broader market trends and patterns to determine the relative health and risk inherent in different lines of business, firms, industries, or the macro-economy, and can then use these analyses to better craft policy at the micro or macro level.[1]

[1] For a discussion of systemic risk management, *see* Marco Espinosa, "Systemic Risk and the Redesign of Financial Regulation," *Global Financial Stability Report*, International Monetary Fund, Chapter 2.

With the rise of big data comes a range of issues surrounding consumer access and use of such data. Questions of data ownership, licensing, and permissive use give rise to different regulatory and contractual approaches to handling consumer data, and in some cases limiting or expanding consumer access to it. These questions and concerns continue to evolve with our understanding and use of big data and will be addressed later in this paper.

Data Quality, Data Gaps, and Information Arbitrage

To unpack many of the applications for and concerns around big data, it is important first to understand how data is collected, manipulated, modeled, and analyzed by various parties, from financial institutions to researchers to governmental bodies and regulatory agencies. A fundamental understanding of key considerations in the data modeling and analytics space will assist in this regard.

Data Availability

Data availability is often determinative with regards to the usefulness or applicability of analysis on a particular set of data. While it follows that the quality of data analysis flows from the availability of high-quality data from which to draw such analysis, there are several components to high-quality data: Such data must be easily available and accessible, and data integrity is also important: high-quality, usable data will have fewer gaps, and will be consistent in quality, readability, and usability.

Information is the most precious asset in financial markets. In his Moderator Note for Panel 2 of the Big Data in Finance Conference, Professor Amiyatosh Purnanandam points out that financial regulation has focused much on the timeliness and accuracy of information available to market participants. But these regulations around information present challenges: what kind and how much information should be collected, and to whom should it be

disseminated? Where do gaps exist as to information and data availability in the financial system? How can we improve the quality of financial data, and is more data necessarily better?[2]

Purnanandam finds it helpful to group information issues around the two broad themes of opacity and asymmetry. Whereas information opacity is the absence of information with every party in the financial system, information asymmetry logically refers to differences in the information available to different parties to a transaction or event.[3] Using an example from the Global Financial Crisis to illustrate, if no one knew the true value of a mortgage backed security, then it would be a case of information opacity. An information asymmetry would exist, however, if the originators of these securities knew more about their true value than their passive (prospective) buyers. Purnanandam believes that the origins of many market failures and breakdowns can be attributed to issues related to information opacity and asymmetry, and notes that in extreme cases, information asymmetries can lead uninformed parties to simply cease from transacting with informed parties altogether out of fear for potential unknown risks involved. Private contracts and disclosure requirements, among other market-based regulations, have helped the financial services industry make progress toward eliminating friction caused by information asymmetry, but much work remains to be done in filling data gaps and raising the quality of data available to market participants, as highlighted by the Global Financial Crisis.[4]

Gaps in data can themselves create information asymmetries. As with any information asymmetry, these gaps can create a disadvantage for some parties in the financial services industry while providing a key advantage to others. This advantage can be a strategic or competitive one, or it can foster and promote investment in data and information creation and

[2] Purnanandam, Amiyatosh. *Moderator Note, Panel 2: Data Quality, Data Gaps, and Information Arbitrage*. Big Data in Finance Conference, UNIVERSITY OF MICHIGAN CENTER ON FINANCE, LAW, AND POLICY (Oct. 27-28, 2016).
[3] *Id.*
[4] *Id.*

generation. There is also a valuable arbitrage opportunity for those with access to such data to profit on these data gaps, ostensibly with the outcome of a more efficient market in the long run.

Purnanandam notes that information disclosure and data quality have not kept pace with the increasing complexity of financial securities and transactions.[5] Securitization and structuring of many financial products have created supposedly "risk-free" assets that are also opaque, or in some cases, provide little information to buyers. With little data available, valuation becomes increasingly time-consuming and difficult, and concerns around information asymmetry and opacity become more relevant. While recent initiatives such as the Legal Entity Identifier (LEI), a unique code used to identify legally distinct units in financial transactions, are a step in the right direction, much work remains to be done toward removing data gaps that lead to this information friction. With fewer data gaps and asymmetry, one would then hope to see higher liquidity and better functioning of markets, particularly in stressful market conditions;[6] however, it is difficult to predict parties' future behavior in a world where such disclosure is stronger and information asymmetry is smaller. Much in the same way that many traditional banking activities shifted to the shadow banking sector before the Global Financial Crisis, some parties might try to change the fundamental nature of certain products, or move them to a less-regulated sector. Too much information could also lead some parties to decline to participate in certain markets, whether due to an overwhelming amount of information or the reduced opportunities for information arbitrage that closures in data gaps present. Effective policies will therefore consider these possibilities in a proactive and prospective way, instead of just looking to past trends and patterns to predict and influence future market effects and events.[7]

[5] Purnanandam, Amiyatosh. *Moderator Note, Panel 2: Data Quality, Data Gaps, and Information Arbitrage*. Big Data in Finance Conference, UNIVERSITY OF MICHIGAN CENTER ON FINANCE, LAW, AND POLICY (Oct. 27-28, 2016).
[6] *Id.*
[7] *Id*.

Data Accessibility and Ownership

There is an inherent tension between keeping data both private and secure, yet also usable.[8] While this topic will be discussed later in this paper, and was the subject of much examination and discussion at the Big Data Conference, the optimal level of security and privacy around big data for consumers is not always the level at which firms and other parties in the financial services industry can optimize innovation; that is, privacy and security can often present roadblocks to more efficient ways of using financial data for innovation.

In a related vein, there is the question of data ownership. Who owns big data? While the answer varies widely based upon the parties and interactions involved in the creation of such data, the context of their relationship, and what ownership rights may already have been signed away, determining who should own big data—whether it is simply collected or intentionally created—carries enormous implications for the ability of firms to maintain or achieve strategic and competitive advantages in their use of big data.

One individual's data has limited value on its own; greater value lies in the aggregate when firms or policymakers assemble a more comprehensive dataset for general analysis and insights.[9] At the same time, there may be conflicts regarding who owns or should own the underlying data. Strong arguments can be made for protection of a consumer's personally identifying data, but transaction data resulting from an interaction between two or more parties may give rise to rights and responsibilities for both parties, and any intermediaries involved in the transaction. One might also distinguish between naturally occurring data, which might be

[8] *Id*. For further discussion of data privacy and security, see *infra* pg. 36.
[9] Barr, Michael. *Moderator Note, Panel 3: Big Questions About Big Data*. Big Data in Finance Conference, UNIVERSITY OF MICHIGAN CENTER ON FINANCE, LAW, AND POLICY (Oct. 27-28, 2016).

owned by one or the other or both parties to the transaction, compared with designed data,[10] which might be owned by the party that creates or commissions it. Trickier questions arise when designed data is derivative of naturally occurring data owned by other parties to a transaction. Should these parties be able to assert some control over the creation of such data? Should there be disclosure requirements? Should there be guidelines or regulations around the use of the derivative data? Should consumers and firms be able to be reimbursed if and when their data is used? Additionally, there may be conflicts between economic efficiency and respect for individual consumer empowerment, although such empowerment might enhance competition.

The old adage "Knowledge is Power" suggests that given many of the competitive advantages to be gained by collecting, analyzing, and maintaining information about consumers and markets in the form of big data, firms will be reluctant to share any information that might result in competitive gain for another party. Can the financial services sector cooperate and share data in a way that is mutually beneficial and unlocks value for multiple parties involved, consistent with concerns about data security, privacy, and consumer rights?

Challenges in Financial Data Modeling and Analytics

In addition to challenges that arise from data access, gaps, and quality, data scientists face significant problems in the areas of data integration and visualization. For decision makers to exploit the opportunities created by big data, they must first be able to organize, clean, and integrate the data into a usable format. While the details of data integration and visualization vary across domains, the need for data cleaning, integration, and analytics is universal, regardless

[10] For an explanation and further discussion of the difference between naturally occurring data and designed data, see *infra*, pg. 20.

of the field of application, as Professor H.V. Jagadish points out.[11] Jagadish has highlighted the need for technical advancements in multiple areas of the "Big Data pipeline" in his 2014 paper.[12]

As Mark Flood of the U.S. Office of Financial Research indicates in his Moderator Note for Panel 6 of the conference, scalability poses challenges for all disciplines where big data is utilized, and data integration and visualization are two distinct but interrelated phases in the process of marshalling financial information and delivering it to the user.[13] Flood explained that *data integration*, drawing on the work of Bernstein and Haas, is the task of massaging raw data inputs into a common conceptual framework for use,[14] whereas *data visualization* involves a robust set of conceptual abstractions such that data can be rendered and compared in a useful way on a two-dimensional page or computer screen.[15]

According to Flood, data scalability issues tend to exhibit common facets, irrespective of the application domain. These he calls the "Four Vs," volume, velocity, variety, and veracity. These are all present in the application of big data to the world of finance, and financial data also have their own set of special concerns.[16] Raw data must ultimately be integrated and abstracted

---

[11] *See* H.V. Jagadish, *Moderator Note, Panel 5: Research Challenges in Financial Data Modeling and Analytics.* Big Data in Finance Conference, UNIVERSITY OF MICHIGAN CENTER ON FINANCE, LAW, AND POLICY (Oct. 27-28, 2016).

[12] Jagadish, H.V. et al. (2014) "Big Data and Its Technical Challenges," *Communications of the ACM*, 57(7), July, 86-94.

[13] Flood, Mark. *Moderator Note, Panel 6: Data Integration and Visualization*. Big Data in Finance Conference, UNIVERSITY OF MICHIGAN CENTER ON FINANCE, LAW, AND POLICY (Oct. 27-28, 2016).

[14] Bernstein, P.A., and Haas, L.M., "Information integration in the enterprise," *Communications of the ACM*, 51(9), September 2008, 72-79.

[15] Sarlin, P., "Mapping Financial Stability," PhD thesis, Abo Akademi University, 2013. http://tucs.fi/publications/view/?pub_id=phdSarlin_peter13a; Sarlin, P. and Peltonen, T.A., "Mapping the State of Financial Stability," *Journal of International Financial Markets, Institutions and Money*, 26, 2013, 46-76; Flood, M.D., Lemieux, V., Varga, M.J. and Wong, W., "The Application of Visual Analytics to Financial Stability Monitoring," *Journal of Financial Stability*, 2016, http://dx.doi.org/10.1016/j.jfs.2016.01.006.

[16] Brose, M., Flood, M., Krishna, D., and Nichols, W. (eds.), *Handbook of Financial Data and Risk Information*, Cambridge University Press, 2014, http://www.cambridge.org/us/academic/subjects/economics/finance/handbook-financial-data-and-risk-information.

in the middle before it can be visualized in the user-facing part of the informational exchange.[17]

Data Integration

The integration of new data into older, existing data sets or frameworks poses a perennial challenge, as much data must be refreshed and/or updated consistently to continue being useful, particularly in the financial services industry.[18] The format in which data sets exist and methods for updating them vary widely, and challenges arise in handling legacy business units that are the result of mergers and acquisitions. While artificial intelligence and machine learning have made great strides in their ability to integrate new data into existing data sets seamlessly and efficiently, simply having the right systems, governance, and standards is rarely enough by itself.[19] Risks are still inherent in ensuring the quality of the data integration process, and the ensuing usability of the updated data set, and application-specific and situation-specific technology will have to be employed to overcome these barriers.[20]

The world of finance is driven to a large degree by fast-changing information technology; indeed, few industries are so dependent on information technology. At the same time, systemic financial crises are relatively rare, and it is difficult to adapt such fast changing technology to the analysis of systemic trends and events. The difficulties presented by new types of data sets and analysis often do not map nicely onto the previous formats employed, and thus the data analyst must make a choice: create a new standard that includes the new type of data set, or try to translate the new data set into the old system.[21] While the first approach can be cumbersome, the second approach can also diminish the integrity of the new data set in its organic form. Thus, as

[17] Flood, *supra* note 13.
[18] Jagadish, *supra* note 11.
[19] Flood, *supra* note 13.
[20] Jagadish, *supra* note 11.
[21] *Id*.

Computer Science Professor Zachary Ives pointed out at the Big Data in Finance Conference, the process of integrating new types of data into existing or legacy data sets must be iterative, as must the process of setting standards for data use and integration.

Professor Claire Monteleoni highlighted a number of challenges vis-à-vis machine learning and data integration. Even with machine learning and artificial intelligence, there is simply an explosion of data: vast, high-dimensional, noisy, raw, sparse, streaming, time-varying, and sensitive or private. Often, machine learning and artificial intelligence must be able to process data as it arrives in real time, and then examine neighboring or similar data sets to determine where new data sets fit best, or whether data should be included in an analysis at all. Older data may be revealed outliers only after newer data becomes integrated and available for analysis, implying a need for new processes to determine how best to go back and scrub such outlier data from the universe of integrated data sets without destroying the integrity of prior analysis.

The problem of data integration manifests itself in two key areas of finance. The first is at the firm level. As Professor Jagadish points out in his panel Moderator Note, individual financial firms collect data on individual transactions in many, many separate data systems. Firm-wide consistency is difficult to enforce, particularly owing to the fact that many of these separate data systems were created at different times, for different business divisions, and often for varying purposes or modes of data analysis. The process of aggregating all of these different types of data from across the firm – a necessary step in getting a picture of the firm as a whole – is hampered by these inconsistencies and the high costs and inflexibility they present for aggregation.[22]

[22] *Id*.

The second area where this problem can also be observed is on the macro level, where a macro-prudential regulator or supervisory authority often suffers not from a lack of data, but from a lack of *actionable* data; that is, data in a usable format that can be useful in helping identify risk and craft policy.[23] If the process of aggregation and proper integration of big data is difficult at the firm level, then these same difficulties are multiplied at the macro level, where numerous firms participating in various markets record and report their data in at least as many different ways. If regulators and economists can better aggregate and integrate these data sets, however, they should be able to identify more accurately – and hopefully mitigate – systemic risks in the broader financial sector and public markets. Selecting the right model for such data integration poses an enormous challenge. For policymakers, the incentives to get these models "right" are very high, because false discovery rates are a real concern.[24] Jagadish points out that to make sense of big data, we need not only faster hardware and models, but also new approaches to data analysis altogether.[25] Data visualization is one answer to this challenge.

Data Visualization

Data visualization addresses a key question arising from the exponential increase in the availability of big data and scientists' ability to integrate new information into existing data sets. At what point do data sets become so complex that machines – much less humans – have difficulty analyzing or presenting the data and analysis in such a way that makes sense for purposes of application in the real world? In other words, a challenge arises when advances in data collection and analysis are not followed by an increased understanding or ability to also visualize that which is being collected and analyzed.

---

[23] *Id*.
[24] Fan, et al. (2014) "Challenges of Big Data Analysis," *National Science Review*, 1(2), June, 293-314.
[25] Jagadish, *supra* note 11.

Finance professor Sanjiv Das emphasized the importance of making theory the basis of our data handling; data handling is not only a solution, but is also itself a problem. According to Das, simple theoretical models are important, because questions around data handling can take precedence over the data itself when visualizing big data. When disparate data is the norm and dimension reduction is necessary to process and visualize data, a simple and elegant understanding of the theoretical underpinnings of the processes and analysis is a necessary prerequisite to performing these tasks.

Dimension reduction will continue to be important in our visualization and analysis of data.[26] With machine learning and artificial intelligence quickly outpacing human capacity to understand and process large data sets, we must formulate means by which to understand and visualize this automated analysis.[27] If we cannot see potential errors, weaknesses, or blind spots in artificial intelligence's analysis of large data sets, then we will not be able to correct them. Thus, functional data visualization is necessary to choose between allowing machine learning and artificial intelligence to continue on its course and knowing when we might need to add additional human insight or guidance.

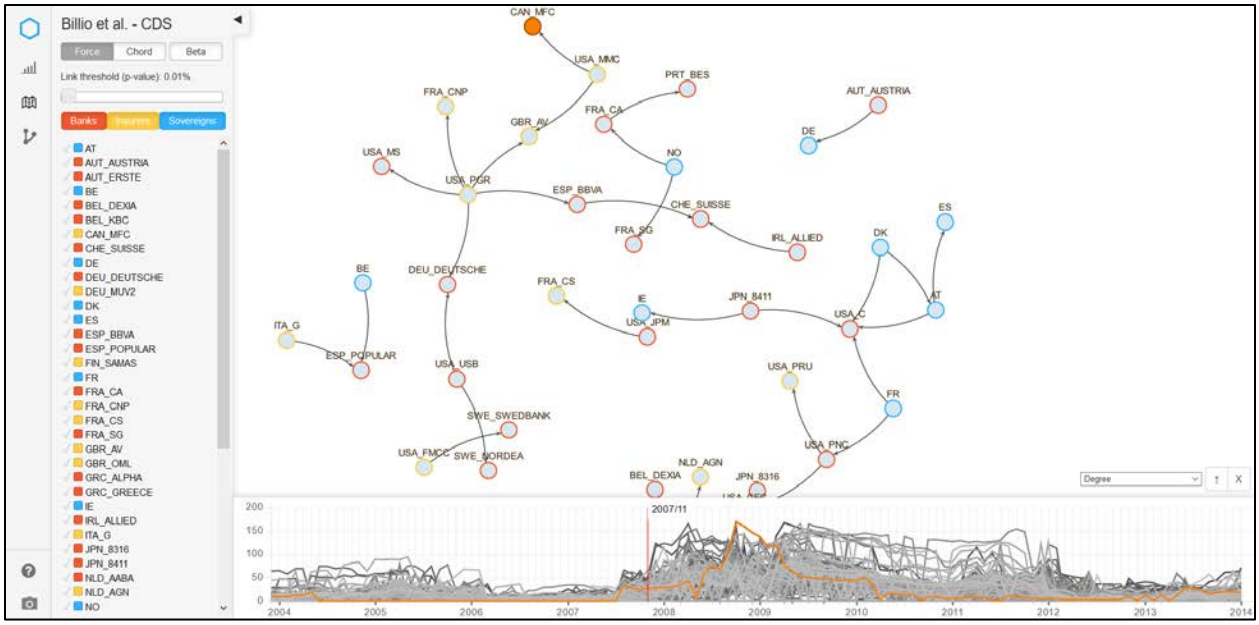
Financial economist Mark Flood of the U.S. Office of Financial Research highlighted some of the additional considerations surrounding big data applications in the financial sector. Because policymakers use data for decision-making and rulemaking, they cannot have an unstable picture from such data. Getting the data abstractions "right" in the visualization process is imperative, because these abstractions cannot be revised after the fact. He specified four types of financial stability data: numeric, geospatial, network, and text, and also gave four dimensions of financial data. *Coverage* refers to the scope of institutions or markets being measured,

[26] Flood, *supra* note 13.
[27] *Id*.

frequency refers to the temporal frequency of data collected and analyzed, *granularity* is the degree of aggregation across the entities measured, and *detail* goes to the "number of columns in the database," or how many different relevant characteristics are available for each measured entity.

Aurel Schubert drew upon his experience at the European Central Bank (ECB) to illustrate challenges in data integration and visualization from a practical perspective. Citing the Eurozone crisis as an example, he explained that Euro-wide aggregated data no longer tells as useful a story for policymakers as it did when the Eurozone was thought to be more economically and financially homogeneous prior to the Global Financial Crisis. Instead, the ECB must now rely on more granular and micro-data on a country-by-country and "security by security" basis, not only to determine security issues and holdings (Who issues what? Who holds what?), but also concentration of such issuances and holdings in specific areas of the market.[28] These are questions that country-by-country aggregated data could not answer (or might even mask), but that can pose key considerations for policymakers looking to ensure regional economic and financial stability.

[28] *Id*.



An example of one form of interactive data visualization for data on credit default swaps, provided courtesy of conference speaker Peter Sarlin.

Experts are also considering new ways to explore data through interactive visualization. Peter Sarlin, co-founder of Silo.AI, explained progress on annotatable dashboards, which allow users to track work around data. This focuses more on exploration of data sets, as opposed to communicating a story from the data. The approach focuses on the more human side of data analysis. Ultimately, human visual constraints are the biggest problem driving the need for simplification and dimension reduction of data. That is not to say, however, that visualization tools are currently solving this challenge.[29] Sarlin posited that data analysts often build isolated knowledge silos due to the lack of proper human tools for interaction, but that because analysis is a social process, allowing for interactions across different visualization platforms will allow for

[29] *Id.*

enhanced analysis.[30] An example application of this would be taking news information and using it to predict stock prices thereafter.

But how do we visualize data most effectively so as to gain insight from it? Dr. Margaret Varga of the North American Treaty Organization (NATO) Exploratory Visual Analytics Research Task Group spoke about her work in developing visual analytics for NATO member states and how some of the progress being made by NATO with regard to data visualization could find similar applications in the world of finance. NATO's Exploratory Visual Analytics group works across many domains, supporting situation awareness for financial, aviation, healthcare, maritime, and cyber applications, and is extremely multi-disciplinary and cross-domain in its approach to analyzing data sets. Varga emphasized the importance of visualizing data in ways that align with how we as humans process information: enabling users to make decisions based on data, forming or changing opinions on issues presented, and learning and discovering new phenomena. Additionally, she pointed out that traditional approaches are often inadequate for understanding big data. Visual analytics can offer solutions to problems such as analyzing, integrating, and presenting massive amounts of dynamic, complex, and multi-source information in a usable and tractable manner. Effective situation awareness is similarly critical for financial stability, where important context around how a situation has developed, as well as how it may progress or change, are crucial to predicting what may happen and responding appropriately.[31] This context can consist both of traditional financial data and other seemingly

[30] *See generally* Peter Sarlin, *Macroprudential oversight, risk communication and visualization*. Journal of Financial Stability (2016), 10.1016/j.jfs.2015.12.005.
[31] For further discussion of visual analytics, *see* M. Flood, V.L. Lemieux, M. Varga, and B.L. Wong, *The Application of Visual Analytics to Financial Stability Monitoring*, Journal of Financial Stability (2016).

non-financial data that would not otherwise appear to have an impact on financial and economic trends.[32]

In sum, too much information can lead to a poverty of attention. Perhaps the biggest challenge that big data presents is how to make sense of it in a manner that facilitates understanding and insight, rather than simply collecting data for collection's sake.[33] We have standard methods of presenting data already, but as these data systems and our ability to analyze them grow increasingly complex and sophisticated, we must devise better ways to integrate and present the insights gleaned from big data.

Data Sharing and Transparency

One of the most important considerations around big data is the manner in which it is shared, and with whom. Given so many of the other questions and concerns around handling of big data, the paradigms for sharing such information with other users can have a tremendous or deleterious effect upon the efficacy and usability of such data. During the Big Data in Finance Conference at the University of Michigan in October 2016, Panel 4 Moderator Matthew Shapiro discussed some of the key considerations and questions around data sharing.

As previously mentioned, it is useful to distinguish between naturally occurring data and designed data. Naturally occurring data is created by businesses or individual consumers in the course of their daily activities, transactions, and interactions. Naturally occurring data is important for a host of purposes, as Shapiro points out, including measurement of key economic and social indicators; research concerning behavior of individuals, households, and firms and how they interact in the market; evaluation of effects of public policy; improving the efficiency

[32] Flood, *supra* note 13.
[33] Purnanandam, *supra* note 2.

and stability of financial markets; informing the public, businesses, and policymakers; and general regulation of economic activity, among others. Designed data, by contrast, involves information that is intentionally created and collected by governments, researchers, or firms for a specific purpose, in a format such as a survey, as opposed to simply collecting data that already previously existed on its own.[34]

Naturally occurring financial data also has a large scope. Examples include government administrative records such as tax filings, registries of real estate transactions, filings with financial regulators, private financial records including account balances, debt records, composition and performances of financial portfolios, mortgage applications, credit histories and scores, credit card and bank loan records, and the prices and volumes of securities transactions.[35] This naturally occurring financial data can range from simple details such as the balance in an account to the complex information around a one-time private transaction in a derivatives market.[36]

Economics professor Matthew Shapiro highlights the fact that naturally occurring data in finance is shared; that is, because there are at least two parties to every financial transaction (and sometimes more), naturally occurring data are shared at least among the multiple parties to a transaction, and often they are also shared with a financial intermediary or agent, such as an exchange or broker-dealer (which itself may have a business interest in understanding information around the types of transactions it facilitates). Examples include banks and depositors knowing balances and transactions on credit card and checking accounts, buyers and sellers in securities markets knowing each other's identities in a private transaction or sharing

[34] Shapiro, Matthew. *Moderator Note, Panel 4: Data Sharing and Transparency.* Big Data in Finance Conference, UNIVERSITY OF MICHIGAN CENTER ON FINANCE, LAW, AND POLICY (Oct. 27-28, 2016).
[35] *Id*.
[36] *Id.*

this information with a third party such as a broker if trading on an exchange, information exchanged between the buyer and seller of a home, and companies issuing securities such as bonds or equities sharing information to investors at the time of issuance and on an ongoing basis. Because both sides of a transaction will (although in many contexts, this is itself changing) have access to such information, questions of data sharing focus, at the present time, on the costs and benefits of sharing data with third parties for a variety of purposes including statistical, research, and regulatory.[37]

Designed data, on the other hand, is not always available to more parties than just the one who "created" and/or "collected" it. The question of access to this data becomes thornier when one considers the fact that many designed data sets are derived from naturally occurring data. For example, should consumers whose information is used to create a designed data set have some level of access to or disclosure about the derivative data? Many of these consumers do not even realize that such derivative designed data exists, much less how to access it.[38]

Regardless of whether it is naturally occurring or designed data, there are costs and benefits to sharing such data with third parties, be it for regulatory, statistical, research, or commercial purposes. The degree to which parties in the financial services industry already share data varies widely and depends in large part on the type of transaction, product, or sector in question. For example, a good deal of data is publicly available in the form of 10K filings with the SEC. Similarly, information on real estate transactions is widely accessible. And there is a broad consensus that making such data available in instances such as these has purpose in

[37] *Id*.
[38] *Id.*

promoting transparent markets and fair valuations for corporate securities and property tax calculations, respectively.[39]

Unfortunately, much financial data is also private, or semi-private, depending on applicable jurisdiction-specific disclosure requirements and/or market practice. In many instances, financial data is voluntarily shared only between select parties who have an interest in pooling the data to allow for more efficient markets. For example, consumers and financial institutions voluntarily share data with credit bureaus to allow for broader and cheaper extension of credit than would be possible without such sharing; this pooling is still limited and exclusive to select parties.

There are enormous potential benefits to increased data sharing, and numerous applications for such additional data. These must be weighed, however, against the agency costs, loss of confidentiality, and potential loss of privacy that could result from such data sharing. Indeed, as Amiyatosh Purnanandam pointed out, there is an inherent tension between keeping data both private and secure, yet also shareable and usable.[40] Many are confident, however, that the costs and benefits of data sharing can be balanced by legal, institutional, and technological solutions that allow such data to be shared while still protecting privacy, confidentiality, and commercial interests, including various competitive advantages of firms.[41]

Data sharing can take many forms, and proper systems and an understanding of the scope of data are first required before data transparency can be achieved. Often, useful patterns in data are missed because the data itself is not available for analysis, or because it is maintained in an

[39] Some of the issues of public transparency and data privacy are further explored in M. Flood, J. Katz, S. Ong, and A. Smith, "Cryptography and the Economics of Supervisory Information: Balancing Transparency and Confidentiality", OFR Working Paper, No. 0011, September 2013. http://financialresearch.gov/working-papers/files/OFRwp0011_FloodKatzOngSmith_CryptographyAndTheEconomicsOfSupervisoryInformation.pdf.
[40] Purnanandam, *supra* note 2.
[41] Shapiro, *supra* note 34.

awkward format. Speaker David Bholat of the Bank of England brought up an example of this found in Bank of England data dating back to 1844. The Bank has maintained information of every loan on its balance sheet since that year and has published this information weekly. When examining some of this information in a study of the Panic of 1847, new information and patterns about financial crises that had largely been overlooked up until the present day have been uncovered. This information, largely pertaining to the Bank's status as a lender of last resort, can still have applications and be useful today as modern-day regulators and prudential policymakers chart optimal responses to potential financial crises.[42]

Bholat further described several types of data sharing, the first of which is within organizations. In this context, an optimal scenario would be one where there is a lack of ownership of data, or in other words, that it is part of an ecosystem and responsive to the entirety of the organization. This is because any attempt to determine ownership of data can lead to contentious disagreement on questions of rights of control and privacy. There are advantages to moving from the mindset of a data "log" or "repository" to that of an open portal, where information is accessible by everyone. This openness is important because it promotes efficiency in data sharing and use.

The second type of data sharing has to do with sharing between organizations. As an example, Bholat cited the mismatch between the global scope of the financial system and the national scope of regulatory apparatuses. He suggested that this incompatibility can be overcome with data sharing. While he acknowledged that there will undoubtedly be challenges with regard to politics and questions of national sovereignty and national financial autonomy, Bholat

[42] For further treatment of central banks and their use of Big Data, *see* Bholat, David M., *The Future of Central Bank Data*, Journal of Banking Regulation, Macmillan Publishers Ltd., pp. 1-10 (Jul. 3, 2013), https://ssrn.com/abstract=2298979.

believes that policymakers should not allow these challenges to distract from the benefits of data sharing. He further asserted that greater data sharing between regulators is necessary to formulate a systemic view and respond to issues of systemic risk. He pointed to examples of arrangements already in place between the Bank of England and the U.S. Office of Financial Research as cases of efforts that are already underway and cited the Bank for International Settlements as one player in sharing data between national financial systems and national regulators. He also cited trade repositories that allow national regulators to see all the derivatives trades occurring; however, he qualified this by reminding us that each national regulator only sees their country's part of the puzzle and suggested that greater sharing and transparency among all G20 countries would be a significant improvement and step in the right direction.

Bholat identified the third type of data sharing as sharing outside of organizations; that is, data sharing with the public at large. He used the example of crowdsourcing innovative ideas for data visualization from the broader public as a particular application of this type of data sharing which could aid policymakers in their analysis of such data. Ultimately, this type of data sharing promotes accountability, openness, and transparency, and Bholat believes that such data sharing should be seen as a public good.

The differences between naturally occurring and designed data can be seen in a practical context of data sharing between institutions.[43] Institutions are less inclined to protect naturally occurring data, which comes from multi-party interactions (and are accessible by other parties as a result), as a matter of competitive advantage. Designed data and the insights and analysis from it, on the other hand, could be considered proprietary in a sense, and such analysis and insights might be central to an institution's competitive advantage. In fact, these advantages of keeping

[43] Shapiro, *supra* note 34.

designed data private (versus sharing publicly) contribute to institutions' motivation for investing in the creation of designed data. A requirement to share this sort of data (whether with consumers, other financial institutions, or a governmental agency) might erode the incentives for commissioning such data sets.[44]

Bholat also believes that larger private financial institutions should be challenged to share more of their data in the same way that fintech companies do; that is, regulators and the general public need to see more than just their financial statements to understand what is going on at a more granular level within these institutions. Collecting additional granular information from such institutions will only be helpful to the extent that it can be collected, identified, and analyzed in a useful way. Matthew Reed, Chief Counsel of the Office of Financial Research in the U.S. Treasury Department, pointed out that while we have identified the need to compel different institutions and individuals to provide such data, the policy methods for achieving that goal often get muddled in the complex regulatory framework within the United States. Because the institution or government agency collecting data is only empowered to do so within its own regulatory authority and mandate, data is largely collected according to the same regulatory patchwork that exists for monitoring the financial system and various players in it, instead of all of this data being assembled in a central location for easier access. This exacerbates issues of coordination and policymaking between federal regulatory agencies when attempting to generate insights from the data.[45]

Business Professor Deborah Lucas raised the example of the United States government as a financial institution (and a large one at that) that must tackle issues related to data sharing at

[44] *Id*.
[45] For a discussion of some of the problems raised by the current regulatory "patchwork" in the U.S., *see* Ryan Calo, *Consumer Subject Review Boards: A Thought Experiment*, 66 STAN. L. REV. ONLINE 97 (2013).

many levels. The trillions of dollars of federal government credit outstanding includes federal deposit insurance, the Fannie Mae and Freddie Mac programs, pension plans, student loans, and the like, as this credit is spread across over 100 separate credit programs throughout the federal government. The quality and content of data collected on these loans varies widely across programs, and the only universally collected information is minimal, such as size, fees, interest rate, and sometimes demographics of the borrower and/or loan performance. There is no uniform standard for default, nor a consistent calculation of default rates across different agencies or with the private sector. While interest rates on different types of loans ostensibly reflect the costs the government bears due to defaults, there is no government-wide data set available that could assist policymakers and program administrators in making better decisions regarding such loan programs.

Such data, if made available to program administrators and the public at large, could be important for transparency and efficiency. Ultimately, if taxpayers absorb credit risk on federal loans, then the public should have information available to understand those risks. Greater transparency around federal loans and their effect on the federal debt could also change minds and influence spending priorities and could also be used to evaluate the efficacy of different programs, as well as the value proposition provided by such government programs compared to the private sector. Such data sharing by the federal government could also be helpful to prudential authorities in revealing systemic risk and to instill in individual government agencies a greater awareness of the systemic risk (or incremental systemic risk) that those agencies might be contributing to the overall financial system. The limited resources of the federal government create internal resistance to additional data sharing programs, however. It is often difficult to make data collection a high priority for an agency running on a tight budget. While there is less

concern for protecting proprietary interests as there would be in the private sector, there could still be negative consequences to sharing internal data between individual government agencies.

Framing Big Questions About Big Data

While big data can provide value to firms, consumers, policymakers, and other decision makers, many experts in the field continue to wrestle with important questions about big data. For example, how should we treat issues of data security and privacy? Who should own, access, and be able to analyze big data? Can and should big data be used to either reduce or exacerbate societal ills such as discrimination and inequality? Given the complex considerations and context surrounding big data discussed thus far, it is important to examine several key areas where the best path forward is still up for debate, and where the future is still unclear. Moderator Michael Barr raised some of these questions in Panel 3 of the Big Data in Finance Conference.[46]

Portability and Ownership of Consumer Financial Data

One of the most pressing issues around big data today has to do with consumer ownership of data. What belongs to consumers, and what belongs to financial institutions such as banks and brokers? As highlighted by Law and Public Policy Professor Barr and others, there is a fierce debate over whether customers should be empowered to own their own financial data in a way that might foster competition in the financial services industry.[47] Should consumers be able to take this information and share it with another financial service provider, and if so, should banks and other financial intermediaries be able to do likewise? Examples of this already exist with regard to importing trading and brokerage data into tax preparation services, as well as online payment systems and other fintech services.

[46] Barr, *supra* note 9.
[47] Karl Antle, *The Looming Battle Over Customer Data*, THE CLEARING HOUSE, BANKING PERSPECTIVE Q1 (2016).

One of the most important questions about big data centers around data portability. Should consumers be able to take their data with them when they switch to a different financial institution? What, if anything, should the prior financial institution be allowed to retain and/or maintain? How should such data be moved, and how can the integrity, usability, privacy, and security of said data be protected in such a move? The paradigm for such decisions and balancing may shift based on new uses and applications for different types of consumer data and the costs and benefits associated with different proposed policies.[48]

Methods of sharing data between financial institutions and other parties are fiercely debated. As noted by the panel, in his 2016 Annual Letter to shareholders, J.P. Morgan Chase's CEO Jamie Dimon suggested that the bank and many other established players in the financial services industry should follow a model of "pushing" financial data to consumers with their consent, as opposed to allowing it to be "pulled" by fintech companies and other disruptive players in the financial services industry.[49] The practical difference between these two methods lies primarily in the fact that "pushing" would require affirmative consent from a consumer, given to a large bank such as J.P. Morgan Chase, before that bank would share any of that consumer's information with another party (e.g., a fintech company). Practically speaking, such consent would be more difficult to obtain, and this difficulty favors large, established financial institutions who will then slow the flow of necessary consumer data to disruptive fintech players.

Data access is important for continued innovation, from fintech startups using open application programming interfaces (APIs), to credit rating agencies needing accurate and relevant information about an individual's or a company's past financial history to assess the

[48] Barr, *supra* note 9.
[49] *See* 2016 Annual Report and Letter to Shareholders, J.P. MORGAN CHASE & CO. https://www.jpmorganchase.com/corporate/investor-relations/document/ar2016-ceolettershareholders.pdf.

party's creditworthiness.[50] Sharing big data is not always a zero-sum game that erodes one party's competitive advantage to benefit another's. In many cases, value can be unlocked for many different parties involved in the financial services industry by sharing useful data.

In the interest of allowing a flow of basic, everyday financial data to consumers, governments and financial regulatory agencies have begun to move to protect consumers' access to and ability to use such data. For example, in the United States, the Consumer Financial Protection Bureau (CFPB) has the authority under Section 1033 of the Dodd-Frank Wall Street Reform and Consumer Protection Act to require banks and other financial institutions to provide customers with access to account information in a readable and usable format.[51] Improved data protocols could also contribute to instant payments and deposit availability, reducing the need for expensive and often unanticipated overdrafts. At the same time, there are significant concerns as to anti-money laundering and fraud issues that could be mitigated with increased analysis of big data around a customer's account and transaction history.

In addition to worries about losing consumer wallet share, banks are concerned about account security and customer privacy, including risks of impersonation, which present a challenge from a compliance and know-your-customer (KYC) rules. As noted by Barr, when third-party fintech applications use a customer's access data to log in to his or her account, the third-party fintech application is essentially impersonating the customer.[52] This is the same approach used in a phishing attack, a common type of cyberattack used by hackers; however, the customer's bank currently has no way to verify whether the access is being undertaken by an authorized third-party fintech application with the customer's permission, or a malicious party

[50] Barr, *supra* note 9.
[51] 12 U.S.C. § 5533.
[52] Barr, *supra* note 9.

with criminal intent. This creates headaches for financial institutions trying to monitor customer accounts for fraudulent activity. Furthermore, these banks also face a strategic disadvantage when they lose control over their customers' data to new fintech startups, as disruptive players in the fintech space can erode the legacy firms' established advantage built up over years of monopoly access to their customers' financial information. Many third-party fintech applications will "scrape" all of a customer's data from their account upon logging in, even without express permission, and this increased access by third-party data aggregators can overload the servers of established financial institutions.[53]

Fintech companies, too, have a distinct set of concerns regarding data ownership and portability. Many fintech startups are viewed as disruptors, and large, established financial institutions are seeking either to limit disruptive competition through acquisitions or squeezing these fintech companies out of the market. Many firms rely on accessibility of their customers' financial data from other financial institutions (i.e., data that is either "pushed" to fintech companies, or that they receive permission to access on behalf of their customers).[54] As a result, fintech companies must balance data accessibility with erosion of their competitive position as innovators and disruptors in the financial services industry.

There may be a tradeoff between security and innovation.[55] How much consumer data should fintech companies share to maximize security? Innovation? Where should the line be drawn between balancing these two interests, which are both vital to the long-term viability of any fintech firm handling such consumer data yet still trying to innovate in the broader financial

[53] Sidel, Robin, *Big Banks Lock Horns with Personal-Finance Web Portals.* Wall St. J. (Nov. 4, 2015), http://www.wsj.com/articles/big-banks-lock-horns-with-personal-finance-web-portals-1446683450.
[54] Barr, *supra* note 9.
[55] For a discussion of this cost-benefit analysis, *see* Adam Thierer, *A Framework for Benefit-Cost Analysis in Digital Privacy Debates*, 20 GEO. MASON L. REV. 1055, 1056 (2013).

services industry? Questions of consumer choice also come into play. Should customers be willing to give up some of their security with regard to financial technology and their financial data in exchange for continued or increased innovation in the sector? Many consumers unwittingly (or in some cases, wittingly but carelessly)[56] sign away their rights to a certain level of security and protection around their data already, effectively allowing a financial services provider to determine subjectively what the optimal balance between security and innovation might be.[57] Should consumers have more of a choice in this regard?

Data Privacy

Although the concept of information privacy has been around at least since the beginning of the 20th Century and the dawn of mass circulation of newspapers with photos,[58] consumers often give little thought to their information privacy.[59] Today's information privacy policy is fundamentally based on the concept of access control; that is, giving individuals control over who has access to their information, and how.[60] There is such a volume and variety of information that it becomes nearly impossible for an individual consumer to know and monitor how his or her information is being used and what permissions he or she has consented to, along with the possible implications for such access. Can individuals effectively control the access and flow of their information? Should these considerations be viewed through the lens of consumer

[56] Mary Madden and Aaron Smith, *Reputation Management and Social Media, Pew Internet & American Life Project* (2010). http://www.pewinternet.org/Reports/2010/Reputation-Management.aspx.
[57] Barr, *supra* note 9.
[58] Glancy, Dorothy, *The Invention of the Right to Privacy*, 21 ARIZ. L. REV. 1, 8 (1979); *see* Samuel Warren & Louis Brandeis, *The Right to Privacy*, 4 HARV. L. REV. 193, 195 (1890).
[59] *See* Mary Madden & Aaron Smith, *supra* note 56.
[60] Lalana Kagal & Hal Abelson, *Access Control is an Inadequate Framework for Privacy Protection* 1(MIT Computer Science and Artificial Intelligence Lab).

protection? As a matter of personal dignity or civil liberty? Can we treat data as a commodity or as a property right? Or perhaps all or none of the forgoing?[61]

Many believe that regardless of the lens chosen, the focus of privacy protection should shift from controlling the access and flow of information to controlling its use, thus also fundamentally shifting the burden of privacy protection to the firms which use personal data.[62] This would allow various privacy controls to be built-in at different points along the big data value chain and could vary in type and scope between different uses for consumer data. Once data has been accessed and the consumer loses control over it, is it more vulnerable to abuse, including systemic unaccountability and increased likelihood of financial crimes?[63]

Discrimination and Big Data

Big data may permit discrimination to occur far more effectively and on a much broader scale.[64] With behavioral data and advanced algorithms to process it, artificial intelligence is capable of making decisions that may automatically and systematically exclude certain minority or socioeconomic demographic groups from participating in certain financial services or target them with more aggressive pricing and/or penalties. While the basic methodology behind such sophisticated data analysis and processing can be used correctly to draw granular distinctions between individuals and "score" them for non-discriminatory purposes (even helping to alleviate the effects of inappropriate human biases), such technology can also have a dark side if left

[61] Samuelson, Pamela, *Privacy as Intellectual Property*, 52 STAN. L. REV. 1125, 1170 (1999).
[62] *Privacy by Design in Big Data* (report by the European Union Agency For Network and Information Security), 1, 5 (2015).
[63] Hoofnagle, Chris, *How the Fair Credit Reporting Act Regulates Big Data*, Future of Privacy Forum, BIG DATA AND PRIVACY: MAKING ENDS MEET DIGEST 86 (Stan. Law Sch. Center for Internet and Society ed., 2013).
[64] Solon Baracas & Andrew D. Selbst, *Big Data's Disparate Impact*, 104 CAL. L. REV. 671, 719 (2016).

unchecked, and advanced algorithms may systematically (even if "unintentionally") deepen discriminatory practices and effects on certain communities.[65]

Algorithmic scoring of individual consumers can benefit firms through price discrimination and more targeted advertising, which can improve overall economic efficiency and the operation and functioning of efficient markets.[66] But these same algorithms can fuel faulty insights, inadvertently and disproportionately lumping people into the "wrong" group, particularly if not properly designed or monitored.[67] Theoretically, price discrimination and targeted advertising can benefit historically marginalized groups, which are often more price-sensitive;[68] however, price discrimination can also disadvantage these groups, especially if marginalized markets lack practical competition. Price discrimination and the data that supports it is ripe for abuse by less ethical firms and individuals.

Similarly, big data has allowed many firms to develop "scoring" processes that lead, in effect, to a "scored society." Examples of these "scores" include traditional credit scores and also newer consumer buying power scores, all of which can yield tremendous insights and value to firms seeking to target the right consumers for various product offerings and to price such products efficiently.[69] While innovative scoring processes can support greater inclusion in some instances (e.g., extending credit to a financially underserved consumer based on information

[65] O'Neill, Cathy, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*, Crown Publishing Group, New York (2016).
[66] *Big Data and Differential Pricing*, 4-5, EXEC. OFFICE OF THE PRESIDENT (2015).
[67] *See* Newman, Nathan, *How Big Data Enables Economic Harm to Consumers, Especially to Low-Income and Other Vulnerable Sectors of the Population*, 6 (Public Comment to the Federal Trade Commission, 2014).
[68] Bhagwan Choudhry, Sanjiv Das, and Barney Hartman-Glaser, *How Big Data Can Make Us Less Racist*, Zocala Public Square (Apr. 28, 2016).
[69] Pam Dixon & Robert Gellman, *The Scoring of America: How Secret Consumer Scores Threaten Your Privacy and Your Future* 27 (2014).

unavailable in a traditional credit report),[70] scoring with big data can be ethically questionable. One example of such a futuristic system might be found in China's development of a system of "social credit scores," which comprises files on each citizen from public and private sources (identifiable by biometric characteristics) and will be mandatory by the year 2020.[71]

Ultimately, big data analysis is still designed in part by humans, and human biases, even if unintentional, will still taint the various steps of the big data process, from collection all the way to interpretation. Artificial intelligence, too, has its own logic, and may result in discrimination in ways that human observers cannot readily detect or predict. Increased reliance on big data and related scoring processes also threatens individual autonomy and case-by-case decision-making. The sheer volume and variety of big data blurs the line between measurement and manipulation, and also between privacy and probability.[72]

Unfortunately, big data and advances in data processing technology make it harder to detect possible violations of equal opportunity laws, particularly as the level of sophistication in data analysis increases over time.[73] Seemingly innocuous data points such as location and buying habits or other behavioral habits can become a proxy for socioeconomic standing or other demographic factors. Our increased capacity to process and analyze data, however, may also aid policymakers in detecting such discriminatory effects and give them some tools to combat it.

---

[70] Robinson & Yu, *Knowing the Score: New Data, Underwriting, and Marketing in the Consumer Credit Marketplace*, 2 (Ford Foundation, 2014).
[71] *See generally* Celia Hatton, *China's 'social credit': Beijing sets up huge system*, BBC NEWS (Oct. 26, 2015), www.bbc.com/news/world-asia-china-34592186; *see also* Mara Hvistendahl, *Inside China's Vast New Experiment in Social Ranking*, WIRED (Dec. 14, 2017), www.wired.com/story/age-of-social-credit.
[72] *See* Viktor Mayer-Schonberger & Kenneth Cukier, *Big Data: A Revolution that Will Transform How We Live, Work, and Think* (2013).
[73] Barocas & Selbst, *supra* note 64.

Data Privacy and Data Security

Data privacy and security are among the more pressing questions in the realm of big data. In the first panel at the Big Data in Finance Conference, Moderator and computer science professor Alfred Hero prefaced the discussion by noting that storage of sensitive information used to be frowned upon. Now, as storage of such information becomes increasingly necessary, the methods and manner in which it is securely stored have come under much scrutiny, and suggestions for revisions and improvements in this area abound. Given the benefits of big data and the enormous role that it now plays in financial services, we must find a way to safeguard data privacy to protect consumer rights and firms' proprietary secrets, while not unduly burdening those with legitimate uses for the information and whose data analysis of such data could prove beneficial.[74]

We first unpack several key concepts and terms around data handling. Perhaps most important is the distinction between data privacy and security. Data privacy refers to the question of what information (or aspects of information) someone should have, who should have it, and for how long. Data security, on the other hand, refers to protecting against unauthorized access outside of such parameters. With more data may come more risk, if for no other reason than the simple fact that a greater body of data presents a larger target, and greater challenges for storage and security.[75] Similarly, data privacy is a tool that can be both used and abused. Cyber attackers are becoming increasingly sophisticated, leveraging powerful hacking programs against financial networks and systems, which are particularly vulnerable to interception and disruption due to their high value to such attackers and the high volume of web traffic. In essence, increased

[74] Hero, Alfred. *Moderator Note, Panel 1: Data Privacy and Security*. Big Data in Finance Conference, UNIVERSITY OF MICHIGAN CENTER ON FINANCE, LAW, AND POLICY (Oct. 27-28, 2016).
[75] *Id*.

volume hampers the ability to detect attacks or breaches of security, particularly at a time when attacks are increasing in complexity and sophistication.

Big data has enabled banks and financial institutions to protect their customers through advanced analytics and fraud detection techniques. These techniques range from analysis of consumer spending patterns and transaction locations to leveraging information from phone location, and social media, such as location check-ins and other data. There is still low consumer confidence in such assurances of privacy and security; Americans' trust in their credit card companies, for example, while still higher than their trust in their search engine providers (16%) and social media sites (11%), still registers at only 38%.[76] Efforts by financial services companies to gain consumers' trust is an uphill battle, and it only takes one major, widely-publicized breach of security in the financial services sector to lower consumer confidence and diminish consumers' trust in the ability of companies to securely handle their data.[77]

Hero points out that it is important to recognize that data privacy, in and of itself, is not monolithic. There are various ways in which privacy can be protected and abused, and various points in the life of data in which privacy protectors and abusers might focus their efforts.[78] Borrowing from Michael Steinmann, Hero highlights "The Four Rs" as four challenges around handling of big data: reuse, repurposing, recombination, and reanalysis.[79] Essentially, at any point during these steps in handling and manipulating data, privacy can be particularly vulnerable or at risk, and efforts to protect data privacy should thus be concentrated accordingly.

---

[76] M. Madden & L. Rainie, *Americans' attitudes about privacy, security, and surveillance*, Pew Research 6 (May 20, 2015) http://www.pewinternet.org/2015/05/20/americans-attitudes-about-privacy-security-and-surveillance.
[77] Hero, *supra* note 74.
[78] *Id*.
[79] Steinmann, Michael, et al., *A Theoretical Framework for Ethical Reflection in Big Data Research*, Ethical Reasoning in Big Data, 11-27 (Springer International, 2016).

There is an inherent tension between the ease of access to big data and facilitating its analysis and use,[80] and the need to protect such data from being too easy for attackers to access. Law professor Peter Swire used the analogy of an M&M candy to describe one approach to data security and privacy. Swire asserted that there is a big temptation to create a large reservoir or "lake" of big data, which would enable consumers, brokers, and even potentially regulators to access and analyze data easily and in one place. In the M&M analogy, this data is protected by a hard "shell" of security, but the inside is "squishy" and less protected once the secure shell is broken. According to Swire, this analogy demonstrates why *data segregation* is important: the same centralization that benefits ordinary data users also assists cyber attackers.

Other protection strategies include data masking or anonymization. Swire pointed out that in many cases, data analysts do not need to know the names or other potential identifying information of the individuals whose data is being analyzed. Under the Gramm-Leach-Bliley Act, anonymized data can even be released publicly for analysis and use by third parties. Unfortunately, re-identification or de-anonymization of data is possible and made much easier by the ubiquitous and comprehensive nature of big data, as data points include customer funds, social media posts, location, and other public records. Largely because of powerful search tools in today's day and age, a few data points can quickly limit a particular piece of information to a single individual, effectively leading to de-anonymization. Swire concluded that anonymized data is, therefore, not fully anonymized or "safe."

Data minimization, an effort to preserve only what is necessary, important, and valuable so as to better protect such data, can have costs and benefits as well. On one hand, the adding of new data has costs, and decision-makers must determine which new information is worth adding;

[80] Purnanandam, *supra* note 2.

minimization of data also limits what the data can be used for and the extent to which it can be analyzed. On the other hand, the benefit of data minimization is the smaller target it presents for attackers and the reality that it is easier to protect a smaller set of information.

Big data can also target products and services to specific consumers. Information ranging from an individual's prior account history to his or her social media data can help determine what products or services might be more appealing to him or her.[81] From the firm's perspective, this promotes economic efficiency in marketing its products; however, firms must be careful that targeting does not become illegal discrimination.[82] According to Swire, this is of particular concern in the context of fair lending, as targeted advertising based on race and other characteristics may violate federal law. An important question remains as to how best to allow firms to target potential customers efficiently without compounding the effects of racial disparities found in past lending practices and violations.[83]

Speaker Jonathan Katz of the University of Maryland raised the issue of differentiating between data security and privacy in greater detail. While security involves safeguarding against those who should not have access to data, he asserted that privacy is a much fuzzier issue, largely due to finer distinctions between the types of data accessible by individuals or firms, and how limited that access may or may not be. Katz framed the questions on this topic around which computations and analysis can be done without compromising privacy and how such computations can be done privately. He asserted that data anonymization is insufficient to protect privacy, because of the increasing capacity to de-anonymize data based on other sources.

---

[81] *See* Wei Y, Pinar Yildirim, Christophe Van den Bulte, and Chrysanthos Dellarocas, "Credit Scoring with Social Data," *Marketing Science*, 1-25 (October 2015).
[82] *See* Hero, *supra* note 74.
[83] *See* discussion on "Discrimination and Big Data" *supra* pg. 33.

Differential privacy, according to Katz, is one leading solution to the problem; however, he questioned how well we can differentiate certain functions and treatment of data.[84]

Katz also posited that if a data holder can be trusted, then that trust extends to that data holder's publishing and use of the function of such data when making decisions. The problem, according to Katz, is that there is no single trusted entity in the United States to hold and secure consumer data. This is a result of several factors, including incompatible trust assumptions among consumers and organizations, legal and regulatory constraints, liability and cost, and the reality that collecting data in one single location presents a convenient target for attackers.

So then, how can computations and analysis of data be done privately without such an unfeasible ideal single, trusted entity holding all the data? In the old method of computation, all parties would send their data to a central entity, which would then create a data function that could be used for computation and analysis. Given the risks in such a centralized system, however, Katz instead proposed the idea of secure multiparty computation, in which all parties agree to a cryptographic protocol that securely shares federated inputs to a calculation while publicly sharing the output, avoiding the need ever to centralize the input data. Katz asserted that such a method would ensure both computational correctness and data privacy.

Forter CEO Michael Reitblat provided additional perspective around data security and privacy, but encouraged decision makers to approach the issue from the point of view of a cyber attacker or criminal. He breaks down cyberattacks into four types: (1) amateur or "prank" attacks, (2) criminals who make a living off of attacks, (3) government-sponsored cyberattacks,

[84] For a survey and discussion on differential privacy, *see* Cynthia Dwork, *Differential Privacy: A Survey of the Results*, Theory and Applications of Models of Computation, Vol. 4978, pp. 1-19 (Apr. 2008), https://www.microsoft.com/en-us/research/publication/differential-privacy-a-survey-of-results/.

usually aimed at disrupting services in another country or to achieve some political end, and (4) hackers who are seeking recognition or "street cred" in the cyber community.

The first three types of cyberattacks, Reitblat pointed out, can be disincentivized by reducing economic benefits of attacks or increasing the costs. He noted that the second group will generally cause the most damage due to their sophistication and economic incentives for launching cyberattacks with ransomware or obtaining financially valuable information that can be used or sold for illicit gains. Also, unlike a government actor or state-sponsored attack, individuals can be harder to track and identify, and are less likely than government sponsors to face repercussions from their targets. Reitblat warned that while he believes we are moving in the right direction, with many more tools now available to track and assess potential damage caused by cybercriminals among data systems and financial assets, no company, individual, or institution is immune from potential attack, and nothing is ultimately completely impenetrable.

Conclusion

Big Data will become increasingly important in the world of finance, and increasingly critical to efforts by financial institutions, firms, and policymakers to learn more about consumers and broader market trends. As society grows in its ability to collect, integrate, process, analyze, visualize, and act on such data, questions of systemic risk, human accountability and control, fair and proper use, discrimination, privacy, security, and consumer protection will continue to percolate. We hope our Big Data in Finance conference and this summary paper will serve as a useful guidepost as these debates continue.