Research Challenges in Financial Data Modeling and Analysis

- **Lewis Alexander**, Managing Director and U.S. Chief Economist, Nomura
- **Sanjiv R. Das**, William and Janice Terry Professor of Finance and Business Analytics, Santa Clara University
- **Zachary Ives**, Professor of Computer and Information Science, University of Pennsylvania
- **H.V. Jagadish**, Bernard A. Galler Collegiate Professor of Electrical Engineering and Computer Science, University of Michigan
- **Claire Monteleoni**, Assistant Professor of Computer Science, George Washington University

February 2017

Abstract: Significant research challenges must be addressed in the cleaning, transformation, integration, modeling, and analytics of Big Data sources for finance. This article surveys the progress made so far in this direction and obstacles yet to overcome. These are issues that are of interest to data-driven financial institutions in both, corporate finance and consumer finance. These challenges are also of interest to the legal profession, as well as regulators. The discussion is relevant to technology firms that support the growing field of FinTech.

Keywords: big data; finance; integration, analytics, applications

Introduction

In many fields of endeavor today, data provide the basis for informed decision-making. This is particularly true of macro-prudential analysis: determination of financial stability requires cleaning, integration, and analysis of multiple disparate large and complex sources of data in a timely way. In fact, the use of Big Data requires technical advances in multiple stages of the Big Data pipeline, as discussed by Jagadish et al (2014). These needs for data cleaning, integration, and analytics are universal, they span many domains, and there is considerable excellent research expanding the frontiers of what we are capable of doing in this regard. This paper will provide an overview of some of the successes we have had, and the challenges that lie ahead.

Nevertheless, many solutions are, of necessity, situational, and we are not investing enough in tools and algorithms specifically for financial data. Indeed, the macro-prudential supervisor today too often suffers from a lack of actionable data, rather than a surfeit. Recent work by public and private agencies, such as the Financial Stability Report of the Office of Financial Research (OFR, 2015), and that of the Banque de France (see Flood, Jagadish, and Rashid 2016), focus on managing these data issues. The difference between the large volumes of source data and the shortage of actionable data is precisely the means to transform, clean, integrate, model, and analyze. This is an area of intellectual inquiry that crucially deserves attention.

The essential problem for individual financial firms is that data on individual transactions are collected in many, many separate data systems. Typically, those systems were created at different times, with different goals. They are designed and maintained by the individual business silos that they serve. Firm-wide consistency is hard to enforce, and it was not high priority for many institutions.

To get a picture of a financial firm as a whole, data from those disparate systems has to be aggregated. The process of aggregation is hampered by inconsistencies in the way financial transactions are recorded. Such inconsistencies are an obstacle to automation. They make aggregation less flexible and more expensive.

These same issues apply with even greater force at the system level. Different firms report data differently. It is a challenge for supervisors to integrate, aggregate, and analyze these data (Flood et al 2011).

Modeling in finance must drive the specific implementation of data extraction and integration. Stein (2013) argues that, in the realm of systemic risk analysis, models and data need to be aligned. The systemic risks associated with the subprime lending market and the crash of the housing market in 2007 could have been modeled through a comprehensive integration and analysis of available public datasets. For example, the datasets relevant to the home mortgage supply chain include the following: (a) regulatory documents made available by MBS issuers, publicly traded financial institutions and mutual funds; (b) subscription-based third party datasets on underlying mortgages; (c) individual home transaction data such as sales, foreclosure and tax records; (d) local economic data such as employment and income-levels; (e) financial news articles. Integrating these datasets may have provided financial analysts, regulators and academic researchers, with comprehensive models to enable risk assessment.

This has in fact been tackled in many ways since the crisis. Dhar (2016) highlights the trade-off between predictability and cost per error, very much in the vein of quality control theory. But in general, the ability to create predictions at the system level is helpful, and requires resolving large data problems. Progress is being made in this direction by researchers focusing on the mortgage space: see the discussion on using public data such as the Case-Shiller indexes, FHFA index, the NCREIF NPI, and NAREIT time series to improve system-wide predictions for the mortgage market

(Wallace 2011). The Real Estate and Financial Markets (REFM) laboratory[1] at Berkeley is aimed at building a big data environment in which the real estate markets may be monitored, and will be an important test case for the various technical issues concerning the use of financial big data for market prediction. For an objective measure of systemic risk over time for the broad financial system, to identify and predict financial institutions that contribute most to this risk, see the recent work by Das, Kim, and Ostrov (2017), that uses public information to create a systemic risk index and identify risky firms.

Economists have been the leaders in creating longitudinal panel datasets and have had a successful history of using national datasets from the Census Bureau, the Department of Labor, etc., and global datasets from the UN, World Bank, etc. Here, too, there has been much less activity in modeling that integrate multiple heterogeneous datasets. While fusing information from multiple datasets may pose technical, policy and privacy challenges, the potential benefits are immense. For example, social media data often contains features that could enhance macroeconomic statistics derived from traditional survey-driven datasets. Enriching longitudinal panel datasets with social media could explore hypotheses with a different focus or level of granularity; for example, one could study the decision making of individuals whose social media profiles would reflect their beliefs, intent, interests, sentiments, opinions, and states of mind.

To address these pressing needs, work is required in at least three areas that we consider in turn in the following sections. The ensuing ideas will benefit financial institutions in both areas, corporate and consumer finance; legal practitioners and regulators; and also technology companies that provide tools for FinTech.

Data Integration

Evaluation of systemic risk requires integration of data from multiple sources to obtain

[1] http://groups.haas.berkeley.edu/realestate/research/REFM_lab.shtml; see also https://wiki.umiacs.umd.edu/clip/ngfci/images/9/93/BIDS.pdf

information about the financial system as a whole, and enough of its multiple aspects to permit meaningful analysis. Data integration is hard to do well, particularly at scale. The issue is not merely one of format conversion. Rather, each independently created data source makes its own data representation and modeling choices, with regard to schema, vocabulary, and even semantics (Halevy, et al., 2006). The solution to this problem, in broad strokes, is to standardize wherever agreement can be achieved, and to work to address the variety where standards are not possible. Since integrated data may not be uniformly reliable or relevant, its origins or provenance (Green, et al, 2007a, 2007b) can help assess its reliability (Karvounarkis, et al, 2009) and even be used to improve the quality of the integration (Talukdar, et al, 2010). While there are many technical solutions that can assist in managing the lack of standards, the ultimate solutions in any context are usually a combination of application-specific tools with some common building blocks.

Consider, for example, the standardization of legal entity identification schemes across a range of independently managed datasets (see Rosenthal and Seligman, 2011). The recently achieved agreement on a globally standardized legal entity identifier (LEI) system is a huge step towards better financial data integration (GLEIF, 2014). But the LEI alone is far from the end of the integration story. Inroads are being made to augment the identification of the first-generation LEI to include complex ownership relationships (see OFR, 2015, p. 70), and to map the LEI to other common identification schemes (NIST, 2016). More advanced techniques would resolve colloquial mentions of names of financial institutions in news and social media and reconcile them with the formal identifiers. Xu, et al. (2016) perform entity resolution of names from residential mortgage backed securities prospectuses with institution names from a vendor list of asset-backed securities.

For macroprudential monitoring, a public Financial Entity Identification and Information Integration (FEIII) Challenge has been developed by the OFR and NIST to research technologies for financial datasets (including text) using automated identifier alignment and entity resolution (NIST, 2016). This effort will create a reference knowledge base with prototype tools, connecting collections of heterogeneous entity identifiers from multiple sources to facilitate information integration, using structured data (e.g., regulatory filings), and unstructured data (e.g., news articles, blogs, and social media). In general, many records align trivially, but there are a number of factors that make certain cases complicated.

- The different regulators keep different data on each organization. For one, an address might be a single field, whereas for another, the address might be broken into three columns, and in another might only have a zip code.
- There are often inconsistencies in how entity names and addresses are entered, in addition to outright errors and typos.
- There is implicit semantic knowledge included in a name, e.g., a name may contain "National Association" or "State Bank of" in its name. This complicates matching based on a similarity score that is obtained using some edit distance metric.

A successful first-round challenge culminated in presentations at the Data Science for Macro Modeling (DSMM) workshop held in San Francisco in June 2016. A second FEIII challenge is now in process, further advancing the creation of a community interested in financial data integration.

The Unstructured Entity Integration Team at IBM's Almaden Labs has created Midas, a system for data extraction and integration for use with disparate financial data. They have undertaken extensive work in high-level entity resolution and integration over non-traditional data (this resulted in their high level language, or HIL). Nine published papers emanated from the team related to HIL. This research has resulted in 4 filed patents.

There are several attractive features of HIL that make a significant scientific contribution in addition to its practical value in applications. First, it combines extract-transform-load (ETL) operations with entity resolution (ER). Second, it does so at large-scale in big data environments such as Hadoop/Spark (handling volume). Third, it easily combines data from various sources, providing an effective means of handling variety through efficient data integration. Fourth, the accuracy of the approach is extremely high, lending veracity to the process; both precision and recall were over 90% in an exercise on FFIEC, SEC, LEI data (this was done successfully for the NIST data challenge 2016). Finally, the research is now embedded in products such as BigInsights and BigMatch. [See Balakrishnan et al (2010); Burdick et al (2011); Alexe et al (2012); Hernandez et al (2013a); Alexe et al (2013); Hernandez et al (2013b); Burdick et al (2014); Burdick et al (2015); and Burdick at al (2016). Patents: ARC820130036; ARC820130148; ARC820120144; and YOR820121699.]

A growing number of financial institutions are interested in applying text mining tools to their management of portfolios, and for risk management. For a broad survey of tools and academic and practitioner applications, see Das (2014). HIL is a front-end tool that can make this possible. The general applicability of HIL speaks to its scientific appeal and potential, at least in the field of finance.

In Burdick et al (2011), HIL was used to extract and integrate data from various types of public financial filings. Many of these filings are lengthy documents of unstructured text, including several numbers and tables. There is a fair bit of complex entity resolution undertaken, where for example, names of people are often confused names of financial firms (we have a large number of firms named after people, such as Goldman, Morgan, etc.) One would imagine that financial firms would report their data as required by regulation in standardized formats, but sadly, this is not the case, and as a result, careful engineering is needed to generate clean and useful data for further analysis. HIL has proved to be extremely helpful in this endeavor, and the paper shows how to extract data to create a network map of the linkages between banks in the US financial system, so as to analyze system-wide risk. This is the sort of big data application that has the potential to make a huge impact on regulators and the financial system. One may take this research further and propose more refined models for measuring systemic risk assuming that systems like HIL will generate the data to construct interbank networks. See for example, Das (2016). There are many financial institutions, academics and regulators in finance who are definitely interested in using HIL.

Data Quality Management

Data often have errors, arising due to a variety of reasons (Dong and Srivastava, 2013). These reasons include errors in data recording, both intentional and unintentional, errors in data extraction, such as from text document analysis, errors in entity matching, errors in interpreting under-documented values, and so on. Maintaining data quality is not easy, particularly for high volume granular data, as discussed in the context of bank stress tests by Hunter (2014). The Basel Committee on Banking Supervision (BCBS) found that half the 30 systemically important banks that they studied are materially non-compliant with Principle 3 (data accuracy and integrity) in their implementation of the BCBS (2013) principles on risk data aggregation. It appears that it will be difficult for many firms to be fully compliant with the Principles (BCBS, 2015, p.3).

Data quality is a critical practical issue as bad data can result in costly erroneous decisions (Osborne, 2012). The magnitude of the data cleaning and preparation burden is growing rapidly (Dasu and Johnson, 2003), and this has resulted in the launch of tools for automated data cleaning (Rahm and Do, 2000), quality assessment (Pipino, et al., 2000), and data integration (Bernstein and Haas, 2008). Adapting these tools for use with financial data is far from trivial, as pointed out by Burdick, et al. (2015), yet substantial progress has been made, as the forensics in IBM's Midas system picks up data errors seamlessly and IBM reported these back to the SEC as well. Commercial tools such as that developed by Paxata (www.paxata.com) are very useful in filtering, cleaning, and data preparation.

Data quality in financial reporting may be particularly prone to subversion because it benefits the recording agent to do so, as is the case with the well-known practice of window dressing (Munyan, 2014), or more complex schemes. It is also believed to be commonplace to place one-sided trades and then cancel them prior to settlement.[2] Any aggregates computed during the time window prior to cancellation can thus be manipulated.

One way to find data quality problems is to compare reports from two or more independent sources. For example, most contracts and trades have two parties, each of which may have some reporting requirements. Reconciling these reports can identify problems with the data, possible under-reporting by some party, and more (Burdick et al 2010; Alexe et al (2013)). But any such reconciliation requires first a step of data integration, which could be challenging in itself as discussed above. Similarly, when extracting data from social media, we know that the extraction results will be less than perfect, but techniques to do better are evolving, see Leskovec (2011).[3] Corroboration with other sources can reduce error rates.

Data quality has also been the focus of recent legislation. The Basel committee released a consultative paper on data quality, see BIS (2013). This paper (BCBS239), developed by the Task Force on SIB Supervision of the Standard Implementation Group of the BIS, enunciated 14 principles in four categories: data governance, risk data aggregation, risk reporting, and supervisory review. Data quality centers around some important attributes such as completenesss (minimize missing values), validity (accuracy and consistency), and accessibility and ease of use. Informatica[4] developed a multiple criteria approach for

[2] See https://qz.com/133695/96-8-of-trades-placed-in-the-us-stock-market-are-cancelled/

[3] See the entire session at KDD here: http://snap.stanford.edu/proj/socmedia-kdd/

[4] http://mitiq.mit.edu/IQIS/Documents/CDOIQS_200777/Papers/01_59_4E.pdf

assessing data quality that applies to the finance setting, broken down into data exploration (column profiling, relationship, redundancy) and data quality (completeness, conformity, consistency, accuracy, duplication, integrity, range). Many services firms such as SAS are engaged in the implementation of BCBS239. We are experiencing growing agreement on the definition of data quality, as well as increasing tools and services for implementation of data quality standards.

Data Analytics

Model selection is a huge challenge with big data. Feature selection on an unstructured dataset can generate an arbitrary number of potential independent variables. This is also true of structured data. Sala-i-Martin (1997), working with a traditional growth equation, generated two million separate specifications from just 62 possible explanatory variables. Donoho and Stodden (2006) point out that the number of variables can sometimes exceed the number of data points. Many big data sources, such as news archives, are novel to financial econometrics, and there are as yet few theoretical constraints to curtail the specification space. In the case of policy questions, an analyst is incentivized to get the “right” answer, thus false discovery rates are a serious problem (see Fan, et al., 2014; and Domingos, 2012). Dhar (2013) suggests using out-of-sample predictive power as a model-selection criterion to ameliorate some of these problems. The key point is that big data necessitates new approaches, not just faster hardware. Fan, Han, and Liu (2014) offer an overview of the challenges.

Within the field of machine learning, methods of "online learning with expert advice” (e.g. Littlestone and Warmuth, 1989, Herbster and Warmuth, 1998; see Cesa-Bianchi and Lugosi, 2006, for a survey) may prove promising for applications to financial stability and monitoring. Here, the learner has access to an ensemble of “experts,” where each expert is simply a time-series; it need not be a skillful predictor. For example, algorithm variants that specialize in learning from non-stationary data have advanced the state-of-the art in various problems in climate science (Monteleoni et al., 2011, DelSole et al., 2015, Strobach and Bel, 2015; 2016). Recent advances (McQuade and Monteleoni, 2012; 2013) in learning from time-series panel data that can vary over both time, and over the dimensions of the panel, can address problems such as financial monitoring over multiple markets (Flood et al., 2015). Recent work by McQuade and Monteleoni addresses data with multiresolution interactions in time, by providing an online multi-task learning approach, treating predictions at different time lags as the "tasks" (McQuade and Monteleoni, 2015; 2016). This approach showed promise in a recent application to financial volatility prediction (McQuade and Monteleoni, 2016).

It is interesting to ask if the increasing effectiveness of highly nonlinear methods such as deep learning neural nets also applies to financial data. Perlich, Provost, and Simonoff (2003) undertook a detailed analysis to compare a linear approach such as logistic regression with a popular inductive, nonlinear method such as decision trees (the C4.5 entropy-based classifier). Their analysis of learning curves showed that for small data sets, logistic regression was more accurate than trees, but this is reversed when moving to large data sets. These results contrast with the findings in Lim, Loh, and Shih (2000) where logistic regression was found to be better. Perlich et al found that bagging was effective in improving the results of decision trees so that they performed much better on large datasets. These studies used about 30 different data sets, but these were not in the finance domain. Therefore, whether the results transfer over to financial data is an interesting question that is beginning to be addressed. We are aware of one instance that confirms the findings of Perlich, Provost, and Simonoff (2003), in a paper on credit card default prediction, by Butaru et al (2016), where decision trees outperform logistic regression on a very large dataset from major credit card firms.

New Applications

Several areas of finance have had at least some limited success in obtaining value from big data. In the next few paragraphs we delineate some of these areas, and explore some of the issues.

A major area for data analysis in finance is the analysis of systemic risk. This is essentially a big data problem because one can only understand the behavior of a system when one has all its data. Sampling runs the risk of capturing a part of the system that does not represent the whole. Modeling a subsystem, especially when examining dynamics, may lead to spurious outcomes that do not come close to being faithful to what may occur for the entire system (for some discussion on biological systems, see Dantzig, et al). However, one may find data such as stock prices that are summary variables for much of the dynamic behavior in a complex system, and exploit these data to some extent. How successful are such approaches is still an open empirical matter. Systemic risk measurement has seen recent advances, described in papers by Espinosa-Vega (2010); Espinosa-Vega and Sola (2010); Billio, Getmansky, Lo, and Pelizzon (2012); Merton, Billio, Getmansky, Gray, Lo, and Pelizzon (2013); and Das (2016).

Consumer finance is a large area in which big data has come to play a role. Financial firms are adopting techniques from consumer marketing in order to improve their relationship with their customers, and also their profitability. Credit scoring with social data is now widely in vogue and the models are pretty sophisticated, see Wei, Yildirim, den Bulte, and Dellarocas (2015) for an application using social media interactions. Lin, Prabhala, and Viswanathan (2013) exploit friendship networks to model lending choice in peer-lending. Big data helps eliminate bias from small data, as argued in Choudhry, Das, and Hartman-Glaser (2016), where stereotyping substitutes for a good model, as loan officers often make decisions based on small data. We are all aware of the embedded biases in the long history of redlining loans in home mortgages, see Ghent, Hernandez-Murillo, and Owyang (2014).[5] We may now eliminate such biases using data that does not rely on "protected characteristics" such as race and gender. However, big

[5] AI may be used to redline:
https://motherboard.vice.com/en_us/article/ai-could-resurrect-a-racist-housing-policy

data in consumer finance also has the potential to result in models that attribute erroneous causality, leading to victimization of underprivileged groups in our society. Such ills are outlined in detail in O'Neill (2016).

Many firms are using big data to improve targeting of their consumer finance offerings. CapitalOne is a good example. It "... formulated its digital strategy on three key pillars – the use of analytics, investment in digital talent and restructuring the company's IT workforce to enable rapid development and deployment of new innovative services."[6] The company uses analytics to target customers and also for customer retention. Targeting helps in finding good customers who would otherwise be screened out under older, coarse metrics. Merrill Lynch is using big data to improve underwriting of loans and better collections. Companies like ZestFinance also access varied sources of data in order to improve loan decisions.[7] A huge area of focus is fraud detection, especially in credit cards with losses of $31BN a year (Srinivasan 2016). However, the use of big data in consumer finance is not without its critics, as the credit history data may be contaminated, see NCLC (2014).

"Nowcasting" is another application of analytics in economics. The latency of economic indicators renders them ineffectual for policy making (Higgins 2014). There is usually a delay of at least a quarter in the production of economic data on GDP, inflation, etc., with the result that data analytics practitioners are now attempting to produce predictors of these statistics using higher frequency data in the economy, both quantitative and textual, as well as poll data. Examples of work in this area is Evans (2005); Giannone, Reichlin, and Small (2008); and Babura, Giannone, Modugno, and Reichlin (2013). Nowcasting is a perfect example of drawing data from various sources and integrating it for predictive analytics.

[6] See "Doing Business The Digital Way: How Capital One Fundamentally Disrupted the Financial Services Industry" -- CapGemini Consulting https://www.capgemini.com/resource-file-access/resource/pdf/capital-one-doing-business-the-digital-way_0.pdf

[7] http://blog.syncsort.com/2014/08/big-data/big-data-can-transform-consumer-finance/

Text Analytics is the new frontier of financial analytics. There is hardly a hedge fund that has not made some attempt at incorporating a text analytics layer in their strategies.[8] Commercial vendors abound in providing text-based macro signals (such as Ravenpack), or provide stock signal information (e.g., StockTwits, iSentium). There is a vast plethora of text mining tools in finance, and for a detailed review, see Das (2014). See also Jegadeesh and Wu (2013); Loughran and McDonald (2014). Text analytics is moving from simple and somewhat ad-hoc word-mining to formal econometric approaches, both frequentist and Bayesian. A case in point is the widespread use of topic analysis in financial applications, using the methodology from the seminal work by Blei, Ng, and Jordan (2003); the paper develops Latent Dirichlet Allocation (LDA), a technique that may be seen to be analogous to principal components analysis of text, though undertaken in a Bayesian framework.

FinTech is a potentially disruptive paradigm related to big data in finance. Financial services remain expensive, either because of inefficiencies or monopoly position of major financial institutions. Thus, technology driven solutions are posing a threat to the traditional models of banking, insurance, and consumer finance. Phillipon (2015, 2016) finds that the unit cost of financial intermediation has been around 2% for the past 130 years! (His measure is obtained as the ratio of the income of the finance industry to the quantity of intermediated assets. As another data point, the share of finance income to GDP has gone from 2% in 1940 to about 8% today.) This is similar across countries, and is not a typically US phenomenon (Philippon 2013). Central FinTech innovations are cryptocurrencies and blockchains, digital advisory (robo) systems, automated trading, use of artificial intelligence and machine learning, peer-to-peer lending, equity crowdfunding,

[8] Graham Bowley - "Computers that Trade on News" (New York Times, 2010-12-22): http://www.nytimes.com/2010/12/23/business/23trading.html; Roy Kaufman - "How Traders are Using Text and Data Mining to Beat the Market" (2015-02-12), https://www.thestreet.com/story/13044694/1/how-traders-are-using-text-and-data-mining-to-beat-the-market.html ; Jen Weiczner - "How Investors are Using Social Media to Make Money" (2015-12-7), http://fortune.com/2015/12/07/dataminr-hedge-funds-twitter-data/

and payment systems, especially in the mobile space. All these new paradigms are based on big data and also generate data of wide-ranging variety and size.

High frequency trading (HFT) algorithms are based on high volume data, mostly streaming sources. These algorithms absorb huge quantities of data from many sources, which are then parsed, and fed to sophisticated algorithms that execute trades quickly and efficiently, either in open markets or dark pools. Data handling in this domain needs to be highly efficient, and in many cases performance requires that the algorithms be embedded in hardware, using special purpose chips, rather than in software. Firms like TradeWorx (http://www.tradeworx.com/) and Automated Trading Desk (ATD, bought by Citibank for $680M in 2007) were pioneers in the field. Algorithmic trading results in about 50% of executed trades in the equity markets (this is down from around 2/3 of stock trades in the late 2000s, mostly because the profits from algorithmic trading are under competitive pressure, and regulatory oversight.

Blockchain and cryptocurrencies are widely heard of, but much less understood. They of course are at the frontier of new payment systems, but are envisaged to have a huge role also in financial contracting. As such this technology is not a big data application, but does involve big computation. Indeed, much of financial innovation centers around big data and/or high performance computing. A blockchain is just a shared file. By definition it is a decentralized record, with copies of the blockchain being maintained by several entities, with (hopefully) comprehensive security and consensus updates. The features are summarized in the acronym DIST (standing for a file that is Distributed, Immutable, Secure, and Trusted), see Ben-Ami (2016). Various banks are experimenting with blockchains for automated settlement, and have formed consortiums such as R3 (https://r3cev.com/). Other similar efforts are USC (Utility Settlement Coin) from UBS and three other major banks, as well as SETL coin from Goldman Sachs. Because blockchains will potentially permeate much of the financial landscape, any assessment of big data in finance requires consideration of this fast-growing technology.

Finally, cybersecurity is largely a big data issue in finance. Financial firms are being increasingly hacked (Faulkner 2015), and are required to protect personally identifiable information (PII) much more than before.[9] Also, how this data is used for business purposes raises interesting ethical issues of data provenance and privacy. Adherence to the Critical Security Controls (CSCs)[10] is a key part of a large bank's security process. The SANS Institute and the Center for Internet Security (CIS) require implementation of protocols that are essentially algorithms running on big data, and are more than mere log analysis.

Conclusion

Financial analysis can greatly benefit from Big Data. Effective macroprudential supervision requires it. However, barriers remain with respect to performing the cleaning, integration, modeling, and analytics required to derive actionable data from a diversity of data sources. An active research agenda is underway to develop the tools and algorithms to address these needs. This article surveys many of these opportunities and initiatives in areas of data integration, data quality, and analytics.

Acknowledgments

This article was prepared for the OFR/U. Michigan conference: Big Data in Finance, Ann Arbor, October 27-28, 2016. We are grateful to participants at the conference and M. D. Flood and S. McQuade for helpful input.

[9] The huge hack of J.P. Morgan affected some 83 million people and businesses. See Matthew Goldstein, Nicole Perlroth and David Sanger, "Hacker's Attack Cracked 10 Financial Firms in Major Assault," (New York Times, 2014-10-03):
https://dealbook.nytimes.com/2014/10/03/hackers-attack-cracked-10-banks-in-major-assault/

[10] https://www.sans.org/media/critical-security-controls/critical-controls-poster-2016.pdf

References

Alexe, B., M. A. Hernández, K. Hildrum, R. Krishnamurthy, G. Koutrika, M. Nagarajan, H. Roitman, M. Shmueli-Scheuer, I. Stanoi, C. Venkatramani, R. Wagle. Surfacing Time-Critical Insights from Social Media. *SIGMOD Conference* 2012: 657-66 (System Demonstration)

Alexe, B. D. Burdick, M. A. Hernández, G. Koutrika, R. Krishnamurthy, L. Popa, I. Stanoi, R. Wisnesky. High-Level Rules for Integration and Analysis of Data: New Challenges. In "Search of Elegance in the Theory and Practice of Computation", *LNCS* (8000) 2013: 36-55

Babura M, Domenico Giannone, Michele Modugno and Lucrezia Reichlin (2013). "Now-casting and Real-Time Data Flow," Working Paper No 1564 (July), *European Central Bank.*

S. Balakrishnan, V. Chu, M. A Hernández, H. Ho, R. Krishnamurthy, S. X. Liu, J. H Pieper, J. S. Pierce, L. Popa, C. M. Robson, L. Shi, I. Stanoi, E. Ting, S. Vaithyanathan, H. Yang. "Midas: Integrating public financial data." *SIGMOD Conference* 2010: 1187-1190 (System Demo)

Basel Committee on Banking Supervision (2015). "Progress in adopting the principles for effective risk data aggregation and risk reporting," January, http://www.bis.org/bcbs/publ/d308.htm.

Basel Committee on Banking Supervision (2013). "Principles for effective risk data aggregation and risk reporting," January, http://www.bis.org/publ/bcbs239.htm.

Ben-Ami, D., (2016). "A Beginner's Guide: Blockchain," *Pensions and Investments Europe* (Special Report, Securities Services), July-August, 46-47.

Bernstein, P. A. and L. M. Haas (2008). "Information Integration in the Enterprise," *Communications of the ACM*, 51(9), September, 72-79.

Billio, Monica, Mila Getmansky, Andrew W. Lo, Loriana Pelizzon (2012). "Econometric measures of connectedness and systemic risk in the finance and insurance sectors," *Journal of Financial Economics*, 104, 535--559.

Blei, D., A. Ng and M. Jordan (2003). "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, 3, 993--1022.

Burdick D., S. Das, M. A. Hernández, C.T. Ho, G. Koutrika, R. Krishnamurthy, L. Popa, I. Stanoi, S. Vaithyanathan. "Extracting, linking and integrating data from public sources: A financial case study." *IEEE Data Eng. Bull*, 2011. http://ssrn.com/abstract=2666384.

Burdick, D., A. Evfimievski, R. Krishnamurthy, N. Lewis, L. Popa, S. Rickards, P. Williams. Financial Analytics from Public Data. International Workshop on Data Science and Macro-Modeling (*DSMM'14*), in conjunction with *ACM SIGMOD* 2014.

Burdick, D., L. Popa and R. Krishnamurthy. Towards High-Precision and Reusable Entity Resolution Algorithms over Sparse Financial Datasets. International Workshop on Data Science and Macro-Modeling (*DSMM'16*), in conjunction with *ACM SIGMOD* 2016.

Burdick, D. R. Fagin, P. G. Kolaitis, L. Popa, W.-C. Tan. A declarative framework for linking entities". Internal Conference on Database Theory (*ICDT*) 2015, 25-43. (Best Paper Award. Extended version invited to appear in *ACM TODS*, July 2016).

Butaru, F., Q. Chen, B. Clark, S. Das, A.W. Lo, and A. Siddique, (2016). "Risk and Risk Management in the Credit Card Industry," *Journal of Banking and Finance*, 72:218--239. http://www.sciencedirect.com/science/article/pii/S0378426616301340

Choudhry, Bhagwan, Sanjiv Das, and Barney Hartman-Glaser (2016). "How Big Data Can Make Us Less Racist," *Zocala Public Square*, April 28, 2016.

Cesa-Bianchi, N. and G. Lugosi (2006). *Prediction, learning, and games*. Cambridge University Press.

Dantzig, G., DeHaven, J.C., Cooper, I., Johnson, S.M., DeLand, E.C., Kanter, H.E., Sans, C.F., (1959). "A Mathematical Model of the Human External Respiratory System," *RAND Corporation*, RM-2519.

Das, Sanjiv (2014). "Text and Context: Language Analytics for Finance," *Foundations and Trends in Finance*, v8(3), 145--260.

Das, Sanjiv (2016). "Matrix Metrics: Network-Based Systemic Risk Scoring", *Journal of Alternative Investments*, Special Issue on Systemic Risk, v18(4), 33-51.

Das, Sanjiv., Seoyoung Kim, and Daniel Ostrov (2017). "Dynamic Risk Networks: A Note," Working paper, Santa Clara University.

Dasu, T. and T. Johnson (2003). Exploratory Data Mining and Data Cleaning, Wiley-Interscience.

DelSole, T., Monteleoni, C., McQuade, S., Tippett, M. K., Pegion, K. and Shukla, J. (2015). "Tracking seasonal prediction models." In: *Proceedings of the Fifth International Workshop on Climate Informatics*.

Dhar, V., (2013). "Data Science and Prediction," *Communications of the ACM*, 56(12), December.

Dhar, V., (2016). “When to Trust Robots with Decisions, and When Not To,” *Harvard Business Review*, 17 May.

Domingos, P. (2012). “A Few Useful Things to Know about Machine Learning,” *Communications of the ACM*, 55(10), October, 78-87.

Dong, X. L. and Srivastava, D. (2013). “Big data integration,” In: 29th International Conference on Data Engineering (ICDE), 1245-1248.

Donoho, D. L. and Stodden, V. C. (2006). “Breakdown Point of Model Selection When the Number of Variables Exceeds the Number of Observations,” *IJCNN* '06. International Joint Conference on Neural Networks, 2006, http://academiccommons.columbia.edu/item/ac:140168.

Evans, M. D. D. (2005). “Where Are We Now? Real-Time Estimates of the Macroeconomy," *International Journal of Central Banking*, v1(2).

Espinosa-Vega, Marco A., and Juan Sola (2010). “Cross-Border Financial Surveillance: A Network Perspective," IMF Working paper no 10/105;

Espinosa, Marco (2010). “Systemic Risk and the Redesign of Financial Regulation," *Global Financial Stability Report*, IMF, Chapter 2.

Fan, J., Han, F. and Liu, H. (2014). “Challenges of Big Data Analysis,” *National Science Review*, 1(2), June, 293-314.

Faulkner, A., (2015). “ThreatMetrix Q4 2015 Cybercrime Report,” *ThreatMetrix*, San Jose, CA.

Flood, M. D., J. C. Liechty and T. Piontek (2015). “Systemwide commonalities in market liquidity.” *Office of Financial Research*: Working Paper 15-11.

Flood, M. D., H. V. Jagadish, Albert Kyle, Frank Olken and Louiqa Raschid (2011). “Using Data for Systemic Financial Risk Management.” *CIDR*, 144-147.

Flood, M. D., H. V. Jagadish, and L. Rashid (2016). “Big Data Challenges and Opportunities in Financial Stability Monitoring,” Financial Stability Review (of the Banque de France), v20, 129-142.

Ghent, A., R. Hernandez-Murillo, and M. Owyang (2014). “Differences in Subprime Loan Pricing Across Races and Neighborhoods,” *Regional Science and Urban Economics*, 48, 199-215.

Giannone, D., L. Reichlin, D. Small (2008). “Nowcasting: The real-time informational content of macroeconomic data," *Journal of Monetary Economics*, 55(4), 665—676.

Global Legal Entity Identifier Foundation (2014). "Annual Report 2014," https://www.gleif.org/en/about/governance/annual-report#.

Green, T., G. Karvounarakis and V. Tannen (2007a). "Provenance Semirings," *Proceedings of the 26th ACM SIGMOD-SIGACT-SIGART* Symposium on Principles of Database Systems. 2007.

Green, T., G. Karvounarakis, Z.G. Ives and V. Tannen (2007b). "Update Exchange with Mappings and Provenance," *Proceedings of the International Conference on Very Large Data Bases* (VLDB) 2007.

Halevy, A., A. Rajaraman and J. Ordille (2006). "Data integration: the teenage years," *Proceedings of the 32nd International Conference on Very Large Data Bases* (VLDB '06), 9-16.

Herbster M and M. K. Warmuth. (1998). "Tracking the best expert." *Machine Learning*, 32:151–178.

M. A. Hernández, G. Koutrika, R. Krishnamurthy, L. Popa, R. Wisnesky. HIL: A High-Level Scripting Language for Entity Integration. *EDBT* 2013: 549-560.

M. A. Hernández, K. Hildrum, P. Jain, R. Wagle, B. Alexe, R. Krishnamurthy, I. R. Stanoi, C. Venkatramani. Constructing Consumer Profiles from Social Media Data. *IEEE BigData Conference* 2013: 710-716.

Higgins, P., (2014). "GDPNow: A Model for GDP Nowcasting", Federal Reserve Bank of Atlanta, Working Paper 2014-7.

Hunter, M. (2014). "Statement by Maryann F. Hunter, Deputy Director, Division of Banking Supervision and Regulation, Board of Governors of the Federal Reserve System before the Committee on Banking, Housing, and Urban Affairs, U.S. Senate, Washington, D.C.," http://www.federalreserve.gov/newsevents/testimony/hunter20140916a.pdf.

Jagadish, H. V., J. Gehrke, A. Labrinidis, Y. Papakonstantinou, J. M. Patel, R. Ramakrishnan and C. Shahabi (2014). "Big Data and Its Technical Challenges," *Communications of the ACM*, 57(7), July, 86-94.

Jegadeesh N and D. Wu (2013). "Word power: A new approach for content analysis," *Journal of Financial Economics*, 110(3), 712--729.

Karvounarakis, G., Z.G. Ives and V. Tannen (2010). "Querying Data Provenance," *Proceedings of the 2010 ACM SIGMOD Conference on Management of Data*.

Leskovec, J. (2011). “Social media analytics: tracking, modeling and predicting the flow of information through networks.” In Proceedings of the 20th international conference companion on World wide web (*WWW* '11). ACM, New York, NY, USA, 277-278. DOI=http://dx.doi.org/10.1145/1963192.1963309.

Lim, T. S., W. Y. Loh, and Y. S. Shih, (2000). “A comparison of prediction accuracy, complexity, and training time for thirty–three old and new classification algorithms,” *Machine Learning*, 40:203--228.

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

Littlestone, N. and M. K. Warmuth (1989). “The weighted majority algorithm.” *Proceedings of the IEEE Symposium on Foundations of Computer Science* (FOCS), 256–261.

Loughran, T. and W. McDonald W (2014). “Measuring readability in financial disclosures," *Journal of Finance*, 69, 1643--1671.

McQuade, S. and C. Monteleoni (2012). “Global climate model tracking using geospatial neighborhoods.” *Proceedings of AAAI*, pages 335–341.

McQuade, S. and C. Monteleoni (2013). “MRF-based spatial expert tracking of the 2010 ACM SIGMOD Conference multi-model ensemble.” *New Approaches for Pattern Recognition and Change Detection*, session at American Geophysical Union (AGU) Fall Meeting.

McQuade, S. and C. Monteleoni (2015). “Multi-task learning from a single task: can different forecast periods be used to improve each other?” *Proceedings of the Fifth International Workshop on Management of Climate Informatics*.

McQuade, S. and C. Monteleoni (2016). “Online Learning of Volatility from Multiple Option Term Lengths.” *DSMM'16: Proceedings of the Second International Workshop on Data Science for Macro-Modeling*. Article No. doi: 10.1145/2951894.2951902

Merton, Robert. C., Monica Billio, Mila Getmansky, Dale Gray, Andrew Lo, and Loriana Pelizzon (2013). “On a New Approach for Analyzing and Managing Macrofinancial Risks,” *Financial Analysts Journal*, 69(2), 22—33.

Monteleoni, C., G. A.Schmidt, S. Saroha and E. Asplund (2011). “Tracking climate models.” *Statistical Analysis and Data Mining*: Special Issue: Best of CIDU 2010, 4(4): 72–392.

Munyan, B. (2014). “Regulatory Arbitrage in Repo Markets,” working paper, December, http://www.bmunyan.com/.

National Consumer Law Center (2014). “Big Data: A Big Disappointment for Scoring Consumer Credit Risk”. https://www.nclc.org/images/pdf/pr-reports/big-data-study.pdf

National Institute of Standards and Technology (2016). “Financial Entity Identification and Information Integration (FEIII) Challenge: About the challenge,” web page, https://ir.nist.gov/dsfin/about.html.

Office of Financial Research (2015). “Financial Stability Report,” December, https://financialresearch.gov/financial-stability-reports/.

O’Neill, Cathy (2016). “Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy,” *Crown Publishing Group*, New York.

Osborne, J. W. (2012). “Best Practices in Data Cleaning: A Complete Guide to Everything You Need to Do Before and After Collecting Your Data,” *SAGE Publications*.

Perlich, Claudia., Foster Provost, and Jeffrey S. Simonoff (2003). “Tree Induction vs. Logistic Regression: A Learning-Curve Analysis,” *Journal of Machine Learning Research*, 4:211--255.

Philippon, T. and A. Reshef (2013). “An International Look at the Growth of Modern Finance,” *Journal of Economic Perspectives*, 27(2), 73-96.

Philippon, T. (2015). “Has the us finance industry become less efficient? on the theory and measurement of financial intermediation?” *The American Economic Review*, 105(4), 1408–38.

Philippon, T. (2016). “The FinTech Opportunity,” Working paper, NYU.

Pipino, L. L., Y. W. Lee and R. Y. Wang (2002). “Data quality assessment,” *Communications of the ACM*, 45(4), 211-218.

Rahm, E. and P. A. Bernstein (2001). “A survey of approaches to automatic schema matching,” *VLDB Journal*, 10(4), December, 334-350.

Rahm, E. and H. H. Do (2000). “Data cleaning: Problems and current approaches,” *IEEE Data Engineering Bulletin*, 23(4), 3-13.

Rosenthal, A. and L. Seligman (2011). “Data integration for systemic risk in the financial system,” Chapter 4, *Handbook for Systemic Risk*, Fouque, J.-P. & Langsam, J. A. (Eds.), Cambridge University Press, 93-122.

Sala-i-Martin, X. X. (1997). “I Just Ran Two Million Regressions,” *American Economic Review*, 87(2), 178-183.

Srinivasan, S., (2016). “Using Big Data to Detect Financial Fraud Aided by FinTech Methods,” Working paper, Texas Southern University.

Strobach E. and G. Bel (2015). “Improvement of climate predictions and reduction of their uncertainties using learning algorithms.” *Atmospheric Chemistry and Physics*, 15(15):8631–8641.

Stein, R. M., (2013). “Aligning models and data for systemic risk analysis,” in *The Handbook of Systemic Risk*. Oxford University Press.

Strobach E. and G. Bel (2016). “Decadal climate predictions using sequential learning algorithms.” Journal of Climate, 29(10):3787–3809.

Talukdar, P. P., Z. G. Ives and F. Pereira (2010). “Automatically Incorporating New Sources in Keyword Search-Based Data Integration,” *Proceedings of the 2010 ACM SIGMOD Conference on Management of Data*.

Wallace, N., (2011). “Real Estate Price Measurement and Stability Crises,” Working paper, UC Berkeley.

Wei Y, Pinar Yildirim, Christophe Van den Bulte, Chrysanthos Dellarocas (2015). “Credit Scoring with Social Data,” *Marketing Science*, October, 1--25.

Xu, Z., D. Burdick and L. Raschid (2016). “Exploiting Lists of Names for Named Entity Identification of Financial Institutions from Unstructured Documents,” working paper, forthcoming.