View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Redfame Publishing: E-Journals

RedFame

Applied Economics and Finance
Vol. 1, No. 1; May 2014
ISSN 2332-7294 E-ISSN 2332-7308
Published by Redfame Publishing
URL: http://aef.redfame.com

Big Data in Firms and Economic Research

Sofia Berto Villas-Boas[1]

Correspondence: Sofia B. Villas-Boas, Department of Agricultural and Resource Economics, 207Giannini Hall, Berkeley, CA 94720-3310, USA.

Received: February 26, 2014 Accepted: March 20, 2014 Available online: March 25, 2014

doi:10.11114/aef.v1i1.375 URL: http://dx.doi.org/10.11114/aef.v1i1.375

Abstract

This paper discusses how firms use "big data" and the role and challenges for economists when getting involved in big data research. The firms' success stories have taken advantage of building the biggest databases, using the best extraction tools, and using the fastest algorithms for data analysis and management. Although there are already great examples of economists entering big data research, analysts involved at present are mostly statisticians and computer scientists. The challenges economists face lie in computation, publication, and replication when using proprietary big data. The opportunities for economists lie in modeling and designing frameworks for analyzing large observational data panels, as well as developing empirical designs and strategies that are motivated by a model framework; in this way, economists can guide large-scale big data experiments toward identifying causal effects, rather than just correlations.

Keywords: big data, economic empirical research, firms and academia relationships, data sharing

1. Introduction

Big data is becoming a common term in industry, media and academia. Since the 1990s, the so called three V's - Volume, Velocity, and Variety - have defined big data, as more and more data are created, originating from a large variety of sources, such as virtual data, that have been exploding. As of 2012, according to IBM, the speed of data generation is 2.5 exabytes per day, and that speed is expected to double roughly every three years. One exabyte (EB) is equal to 1018 bytes, which equals 1 billion gigabytes. To have an idea of what this number means in terms of usage, according to IBM, most 64 bit computers can "address" 16 exabytes, and an exabyte corresponds to between 500 and 3,000 times all the content of the Library of Congress (Johnston, 2012).

Firms and consumers are faced with an explosion of data before making decisions. Firms have to use discretion about how much information to release to consumers; because of information processing costs, if too much information is provided, consumers may decide not to process information and may end choosing not to purchase (e.g., Branco et al., 2013). There are differing points of view on whether consumers and firms have been successfully able to process the relevant information originating from big data to make decisions. One open question is whether and how incorporating big data into firms' decisions leads to significant improvements in firm outcomes or efficiency within the organization. Another important question is whether there is a return on the investment for big-data-centered initiatives.

This paper begins with successful examples of how big data has helped firms understand their customers – both end consumers and intermediate firms - and use that knowledge in their strategy to gain a competitive advantage. The success stories have taken advantage of building the biggest databases, using the best extraction tools, and using the fastest algorithms for data analysis and management.

Then it makes the point that economists should start getting involved in big data research, whereas the data analysts involved at present are mostly statisticians and computer scientists. One advantage of economists is that they can model frameworks for analyzing large observational data panels, as well as developing empirical designs and strategies that are motivated by a model framework; in this way, economists can guide large-scale big data experiments toward identifying causal effects, rather than just correlations.

This paper then turns to examples of successful examples where academic research has already used big data originating from firms. It discusses the resulting research possibilities, and how data-sharing agreements have evolved into creating unique data and empirical settings to answer research questions of interest to economics. It highlights the limitations of using datasets originating in firms, and then discusses the open challenges researchers deal with when using proprietary big data; these challenges include computation, publication, and replication.

2. How Big Data Has Been, and Can Be, Used by Firms

Using data for management is not a new managerial insight. What is new when using big data to gain useful insights for management are the issues pertaining to the size of the data, the speed at which new data becomes available, and the variety of available data at the firm's disposal at any point in time. Firms aim to incorporate the insights based on big data that originate from within their information technology (IT) divisions, or they team up with established or start-up IT firms (see e.g., the "Big Data" series in The Wall Street Journal, 2013,) to help collect, combine, process, extract, and use big data effectively.

With the rise of big data, the traditional role of IT in producing reports evaluating existing and past management decisions has shifted dramatically. Now, an IT division can be extremely valuable to managers before decisions are made. The approach and challenges depend on what kind of data are gathered and then used in decision making. If the data are structured (for example, data that fit into row and column storage or other standard management formats), big data analysts need to master the best extraction tools and develop the fastest algorithms for data analysis to help managers use big data insights before making strategic decisions. If the data are unstructured (for example, originating from video, or consumer feedback, or comments), other challenges for big unstructured data management emerge before those data can be used to gain insights useful for strategy.

2.1 Usage of Big Data

There is some multi-industry evidence (TCS, 2013 and Ramaswamy, 2013) that big data usage is still low. According to an executive survey of twelve hundred companies in Asia, Latin America, North America, and Europe, 47% of the companies had no big data initiatives and 40% stated they still used a hunch or their "gut" feeling to make decisions. Among those who do engage in big data initiatives, the average reported return on investment (ROI) on big data for 2012 was 46%. Interestingly, the executives from the companies that had the biggest reported ROI stated that they believed the biggest benefits from turning to big data initiatives included gaining a better understanding of their consumer base, being able to better predict consumer loyalty, and evaluating product performance. In sum, the core benefit lies not in having turned to big data, but in initiatives that lead to finding novel ways to use insights from big data in the design of business strategies. Another study found that there is also a shortage in the labor force of workers who have analytical skills and who are able to make decisions based on the available data (McKinsey Global Institute, 2012)

2.2 Successful Usage of Big Data

Several examples of such initiatives are now discussed; these are not by any means exhaustive of the cases when firms have been able to successfully analyze and use big data. In the examples mentioned, firms used big data to gain consumer loyalty, discover new patterns of demand for new products, and implement big data market experiments to test strategies before implementing them. In addition, some firms' success stories combined their own big data with additional third party big data to gain a managerial advantage over competitors or to provide an innovative service.

2.2.1 Understand Consumers and Gain Market Share

Probably the Amazon case is the best known and most discussed case of a firm strategically using big data (see e.g., Madden, 2012). Online book sales are recorded, as are browsing and click through patterns for each personal computer accessing Amazon.com, which enable Amazon to understand demand much better than the competition. Amazon does this by tracking not only what customers bought, but also what else they looked at; how they navigated through the site; how much they were influenced by promotions, reviews, and page layouts; and similarities across individuals and groups. Amazon also used big data based tools to produce recommendations for consumers; these tools incorporate responses to recommendations to improve future personal recommendations.

2.2.2 Retain Consumers

It is estimated that Walmart collects more than 2.5 petabytes of data every hour from its customer transactions, and other retailers sit on similarly massive data sets. A big apparel brick and mortar and online retailer, like Macy's, uses software developed by a third party to project a predicted customer's path by combining data from multiple sources within the company, such as past purchase data, online purchase and search data, online banking systems, returns, or contacts with service call centers, with data from outside the company, such as Twitter and Google Trends. The idea behind the software is to use all available data on a topic (say, a particular good) and then make predictions on the outcome of that topic for a particular type of consumer. This big data initiative has increased revenue by making predictions based on correlations about what a consumer is likely to buy and when the consumer is likely to make the purchase, and then to guarantee availability of the product for purchase or shipment. It also predicts, via correlations, when a consumer is marginally not likely to buy a particular good, and then suggests a strategy to induce this customer to make a purchase, via a targeted promotional email, for example.

2.2.3 Increase Consumer Loyalty

A recently implemented loyalty program by the retailer Safeway, called Just for You, personalizes discounts to each consumer based on a consumer's historic purchase records collected through Club Cards. A consumer signs up for this program and receives personalized offers at the point of purchase, as well as manufacturer's coupons for products the consumer buys regularly. According to the first quarter earnings report in 2013, this big data initiative has successfully increased consumer loyalty and consumer spending per shopping occasion.

2.2.4 Increase Customer Satisfaction

Sprint used a big data innovation to integrate data from across its channels, online store, brick and mortar retail, catalogs, and telesales, to get a better picture of why its customers were not happy with the service provided. Given those insights, Sprint made the necessary changes to its business and rose from last to first among carriers in the American Customer Satisfaction Index, while also reducing its call center budget by half.

2.2.5 Collapse Big Data into Manageable Data for Predictions

Opera is a company that offers applications based on "signals," which are repeatable patterns that can be found in sets of big data. Opera has built up a library of signals, such as customer price sensitivity. Using these signals is helpful to a firm when making predictions about an outcome of interest. The signals vary by industry. The idea behind using a signal is very clever: it projects the big data available to a firm onto a smaller set of manageable data, the signals. For instance, Opera enabled a major Japanese car company to increase the selling price of cars returned at the end of their lease period. It did so by breaking the cars' attributes into components and setting a value for each component based on information originating from the relevant signals.

2.2.6 New Product Launch and Bargaining Power

The internet company Netflix recently invested in content, using internal big data on video streaming and DVD viewing to select which shows to produce and how to promote them. According to industry analysts, this shift into content production not only demonstrates a successful strategic use of big data to increase subscriptions, but also illustrates how Netflix may be able to increase its bargaining power vis-a-vis traditional content producers - a strategy analogous to that used by retailers when displaying their own store brands vis-a-vis national brands.

2.2.7 Combining Third Party Real Time Data

Mayer-Schonberger and Cukier (2013) discuss an example based on Google that uses correlations from big data to help institutions such as the Centers for Disease Control (CDC) track the flu season. Based on the location of people who are searching for information related to the flu or flu symptoms, Google Flu Trends gives the CDC an instantaneous idea of where people are getting sick, so that it can design strategies accordingly. As another example, using external GPS location data from cell phones originating in Macy's parking lots, analysts from the Massachusetts Institute of Technology's Media Lab were able to predict, in real time, how many people were likely to shop at Macy's during Black Friday (McAfee and Brynjolfsson, 2012). This made it possible to estimate the retailer's sales on that critical day and assign sale representatives accordingly.

2.2.8 Relationships with Firms in the Supply Chain

Retail Solutions Inc. uses big data tracking based technology to help firms improve efficiency in transactions with suppliers (upstream firms) and retailers (downstream firms). As an example, Kimberly-Clark was able to use those tools to launch a new product in half the time it usually takes to stock all their distributors (The Wall Street Journal, 2011) and to get sales data specific to products, distributors, and retail stores in real time.

2.2.9 Non Structured Data, Social Networks, and Human Resources

The data-driven approach to recruitment is a reality in business– think about the film Moneyball. Big data based developments in human resource recruiting have been led by LinkedIn. LinkedIn uses data to match patterns between job transfers, characteristics of members, and characteristics of firms where they work and have worked, to recommend a pool of potential hires to potential recruiters, given the recruiters' own attributes (Anders, 2013). We are familiar with receiving emails soliciting us to join a network of individual X. Firms receive similar recommendations in terms of "candidates for you." Some patterns identified by LinkedIn originate from structured data, such as biographical information, while other patterns originate from unstructured data, such as photos and videos in a member's social network page on LinkedIn or Facebook.

3. Big Data and Academic Research

Most prestigious universities, their business schools in particular, are engaging in research and instruction about how firms can make use of big data. The emphasis is on how to (i) gather and collect information from inside and outside a company in real time; (ii) develop new insights from that information, for example, discovering patterns of customer

loyalty or disloyalty; and (iii) use these insights to improve outcomes. For instance, following the previous example, a firm can use a strategic marketing strategy to target a consumer who is about to leave. These teaching and research activities result in IT and empirical analysis playing a key role in a firm's industrial organization and internal organization, in industries as diverse as retailing, banking, and health.

University graduates who become successful big data analysts are mostly computer science and statistics majors. The skills needed include the ability to computationally use all available data on a topic and then make predictions on the outcome of that topic, based on correlations found in the big data, rather than based on a small and manageable data set (Mayer-Schonberger and Cukier, 2013). The more data analysts have, the more correlations, and the more accurate the pattern and the prediction about a particular topic. However, a correlation does not mean that one thing caused the other. Moreover, one can find all kinds of correlations, as there are many possible combinations, and it is easy to get lost instead of using the insights effectively.

3.1 The Case for Economists

Economists can and should jump on this big data trend. Economists are trained to develop theoretical insights to guide data analysts in where to look in the big data. Economists also are suited to develop model frameworks to derive counterfactual testable predictions to analyze causation, rather than simple correlations, whether using big data panel data sets or experimental big data. These skills are well-developed in graduate economics programs, and economics undergraduates have begun to develop them as well.

Economists can ultimately contribute by formulating questions that are based on a model, which will help analysts understand the findings. It is very easy to get lost in the big data or to stop at finding interesting correlations, given observations of several variables of interest. I believe that universities should develop a way to get undergraduate economics students and most definitely graduate economics students involved in big data analysis. One big advantage from a research point of view, given the massive amount of data, is that a model may need fewer assumptions and be more data driven once a research question has been formulated. This is always appealing to researchers.

While academic researchers have always been able to access and purchase small data sets directly from firms to answer a particular research question, now the aim of researchers is to access big data that goes beyond the traditional small data set of an individual firm. This can be done by developing mutually beneficial partnerships with proprietary big data owners to conduct innovative research. Economists work at the frontier of the newest statistical and data analytics methods, backed up with theoretical models, and thus are uniquely positioned to approach firms and analyze big data. Based on the successful examples in existing retailing data agreements, one advantage of initiating such academic-firm partnerships is that data owners also allow partner-researchers to use the big data for research questions of interest beyond the firms' managerial objectives.

There are leading efforts to bring big data-based research to economics, e.g., by Athey et al. (2013) and by Jeziorski and Segal (2013). On the one hand, the authors in this stream of research expose the profession to methods used by IT companies, such as the machine learning that Athey used in her own empirical research based on big data (Ito, 2013). On the other hand, and possibly more importantly, this research illustrates how to apply carefully designed theoretical insights to big data in order to gain a better understanding of mechanisms underlying the industrial organization of an industry.

3.2 The Evolving Research Possibilities

Historically, data originating from data-sharing agreements between individual researchers and grocery retailers have allowed practitioners to tackle questions of firm and consumer behavior. Typically, the researcher signs a non-disclosure agreement, or the affiliated university signs an industry alliance data-sharing agreement with the retail data provider. The original datasets are scanner-based datasets, for a subset of the retail chain and for a subset of years, where scanner data consist of observations originated every time an item gets scanned at the cash register. One such successful and widely used example is the "Dominick's Database," at the James M. Kilts Center of the University of Chicago Booth School of Business. This data set originated from a research collaboration that consisted of in- store price and shelf experiments conducted by university researchers at the retail stores, so it is limited in products, time, and store space to focus on the data related to the experimental design.

There are several advantages of an industry alliance data-sharing agreement. Not only is there access to existing unique big (and growing) data, but developing a relationship between the parties makes it possible to create data variation in the form of in-store experiments. Participation in the original experimental design, familiarity with the experimental setting, and potential confounding factors open up possibilities to answer new research questions. Another advantage is having access to auxiliary data needed in a research project, as the firm-retailer relationship allows researchers to follow up with additional data requests beyond the original and traditionally used scanner datasets.[1]

More recently, the data agreements that have been most successful in boosting economic research have been those that

do not focus on a single collaboration and the resulting limited datasets. Rather, researchers involved have been able to develop relationships that result in broad data-sharing agreements pertaining to the whole existing big data set of a large grocery retailer. One such example is the multi-university data-sharing agreement with a large U.S. national retail chain. The proprietary data involved are large and growing every minute, as records consist of the scanned product transaction records of every purchase recorded by cashiers or online payment methods, detailing products, prices and purchasers' masked identifiers. This requires the researchers involved to face new challenges of dealing with a large and growing dataset that is proprietary in nature.

3.3 The Challenges for Research

There are many challenges ahead when dealing with large, growing, proprietary datasets used in economic research. The main challenges include developing the necessary technical savvy, dealing with a growing data collection, and publishing replicable work based on proprietary data. These challenges also present opportunities for universities to get involved and be part of the big data "revolution."

First, while important work is being done in several disciplines, such as computer science, marketing, economics, statistics, applied mathematics, and operations research, I do not see anyone pulling together the efforts dispersed in those many fields. I believe the institutions able to leverage and combine those efforts will be the leaders in research and teaching. They will train the work force that can make sense of the big data now available and develop new jobs to use the insights for things that have not yet been thought of.

Second, broadly used econometric methods and tools may not be appropriate for indirect noisy big data (Cho and Judge, 2013). It may be that answers to important economic questions may have been restricted by methods, by data, or by both. Appropriate econometric and computational methods are available. Economists need to become familiar with these new tools and new data sets. Moreover, it is imperative to integrate research questions, methods, and the acquisition of the relevant big data to tackle the questions of interest.

Third, universities need to think carefully about data, and to develop storage, documentation procedures, and infrastructure to ensure the integrity, confidentiality and research access to this growing collection. Universities also need to be able to guarantee the long-term availability of raw and transformed data, as well as methods and results, for purposes of replication and extension of the published research. Consider this hypothetical scenario: "Authors XYZ published a paper in a top peer-reviewed social science journal, using proprietary data from a source called from now on the "company," and the journal requires the ability to replicate the results. However, the company will not make the data public. What is the right protocol to guarantee both confidentiality and replicability?"

The need for clarity on this point grows as more and more papers use proprietary data. Proprietary data issues reach across disciplines; the problem is not limited to the social sciences. A recent review of papers published in the academic year 2009 by Alsheikh-Ali et al. (2011) found that 44 of the 50 leading scientific journals, which have the highest impact factors, have data-sharing policies, but less than 30 percent of the papers they published fully adhered to the instructions. The data-sharing policy of the journal Science, for instance, states that "all data necessary to understand, assess and extend the conclusions of the manuscript must be available to any reader of Science," but occasionally there are exemptions granted for proprietary data that need to be judged by the editor on a case-by-case basis.

The key challenge is to establish a platform to deal with the use of proprietary data for both research and publication of the results in peer-reviewed journals, protecting the data's proprietary nature and, at the same time, satisfying "data availability for replication" policies. Much work needs to be done here, involving data providers, researchers, journal editorial boards, and data centers collecting proprietary big data.

4. Conclusion

Big data are here to stay and will only become bigger in the future. The key to a firm's success in switching to big data does not lie in figuring out how to collect and store large and growing amounts of data. Rather, success lies in being able to derive useful insights in strategic decision-making processes. Those insights are helpful to a firm when interacting with its competitors, suppliers, and consumers, as well as when the firm wants to change its internal organization to become more efficient.

Academic researchers are uniquely positioned to analyze big data by helping analysts understand where to look and how to formulate questions. The main challenge ahead is to train researchers to computationally deal with large and growing proprietary data in academic research. The second is to solve the issue of property when publishing the findings. This is not just an issue for the social sciences, as proprietary big data issues reach across disciplines. The first step toward a solution is to be able to work across disciplines and combine their efforts, thus leveraging existing teaching and research initiatives using big data

Acknowledgements

The Author thanks the Editor and an anonymous referee for their comments and the Giannini foundation for support. I also thank several industry interview participants for sharing their experience and Cyndi Berck and George Judge for helpful suggestions.

References

Alsheikh-Ali, A. A., Qureshi, W., Al-Mallah, M. H., & Ioannidis, J. P. A. (2011). Public Availability of Published Research Data in High-Impact Journals. *PLoS ONE, 6*(9), e24357. http://dx.doi.org/10.1371/journal.pone.0024357.

Anders, G. (2013). Who Should You Hire? LinkedIn Says: Try Our Algorithm, Forbes, April 4, 2013.

Athey, S., Calvano, E., & Gans, J. S. (2013). The Impact of the Internet on Advertising Markets for News Media, SSRN working paper, research.joshuagans.com.

Branco, F., Sun, M., & Villas-Boas, J. M. (2013). Too Much Information? Information Gathering and Search Costs, working paper, http://groups.haas.berkeley.edu/marketing/PAPERS/VILLAS/TMI2-2013.pdf

Cho, W., & Judge, G., (2013). An Information Theoretic Approach to Network Tomography, working paper, http://cho.pol.illinois.edu/wendy/research.html

Mayer-Schonberger, V., & Cukier, K. (2013). Big Data: A Revolution that Will Transform How We Live, Work and Think, Eamon Dolan/Houghton Mifflin Harcourt; 1 edition (March 5, 2013).

IBM, O/S Concepts, "A Brief History of Virtual Storage and 64-bit Addressability," http://publib.boulder.ibm.com/infocenter/zos/basics/index.jsp?topic=/com.ibm.zos.zconcepts/zconcepts_102.htm.

Ito, A. (2013). Stanford Economist Musters Big Data to Shape Web Future, Bloomberg, June.

Jeziorski, P., & Segal, I. (2012). What Makes Them Click: Empirical Analysis of Consumer Demand for Search Advertising, working paper., http://faculty.haas.berkeley.edu/przemekj/ads_paper.pdf

Johnston, L. (2012). How Many Libraries of Congress Does it Take? Library of congress blog. http://blogs.loc.gov/digitalpreservation/2012/03/how-many-libraries-of-congress-does-it-take/.

Madden, S. (2012). How Companies Like Amazon Make You Love Them, Fast Company, May 2, 2012, blog: http://www.fastcodesign.com/1669551/how-companies-like-amazon-use-big-data-to-make-you-love-them.

McKinsey Global Institute. (2012). Big Data: The Next Frontier for Innovation, Competition, and Productivity, June.

McAfee, A., & Brynjolfsson, E. (2012). Big Data: The Management Revolution, Harvard Business Review, 9 pages. Oct. 01, 2012.

Ramaswamy, S. (2013). What the Companies Winning at Big Data Do Differently, Bloomberg, June: http://www.bloomberg.com/news/2013-06-25/what-the-companies-winning-at-big-data-do-differently.html.

TCS. (2013). The Emerging Big Returns on Big Data, Tata Consultancy Services, http://www.tcs.com/big-data-study/Pages/download-report.aspx

The Wall Street Journal. (2011). Big Data Blog Series, Big Data- Venture Capital Dispatch- WSJ, http://blogs.wsj.com/venturecapital/tag/big-data/.

The Wall Street Journal. (2012). Big Data Blog Series, Big Data- Venture Capital Dispatch- WSJ, http://blogs.wsj.com/venturecapital/tag/big-data/.

The Wall Street Journal. (2013). Big Data Blog Series, Big Data- Venture Capital Dispatch- WSJ, http://blogs.wsj.com/venturecapital/tag/big-data/.

Notes

Note 1. While the focus of this subsection is on direct data agreements between a researcher and a retailer, academic researchers can purchase manageable sized data directly from companies. In the retail setting, traditional providers are Information Resources, Inc. (IRI) or AC Nielsen. Recently, efforts by the INFORMS Society for Marketing Science (ISMS) have been developed to maintain a dataset of unprecedented size, allowing possibilities for research at the micro and macro level, taking advantage of multi-product, multi-market and multi-retail data over several years, covering multiple business cycles.

(cc) BY

This work is licensed under a Creative Commons Attribution 3.0 License.