Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 3, Issue. 4, April 2014, pg.1155 – 1161

RESEARCH ARTICLE

Opportunities and Challenges of Big Data in Economics Research and Enterprises

Tushar M. Chavan[1], S. P. Akarte[2]

[1]ME (CSE) ,First Year, Department of CSE, Prof. Ram Meghe Institute Of Technology and Research, Badnera, Amravati

Email ID: chavan.3491.tushar@live.com

[2]Proffesor, Department of CSE, Prof. Ram Meghe Institute of Technology and Research, Badnera, Amravati

Email ID: s_akarte25@rediffmail.com

ABSTRACT

Many believe that "big data" will transform business, government and other aspects of the economy. In this article we discuss how new data may impact economic policy and economic research. Large-- scale administrative datasets and proprietary private sector data can greatly improve the way we measure, track and describe economic activity. They also can enable novel research designs that allow researchers to trace the consequences of different events or policies. We discuss some of the challenges in accessing and making use of these data. We also consider whether the big data predictive modeling tools that have emerged in statistics and computer science may prove useful in economics.

The term big data draws a lot of attention, but behind the hype there's a simple story. For decades, companies have been making business decisions based on transactional data stored in relational databases. Beyond that critical data, however, is a potential treasure trove of non-traditional, less structured data: weblogs, social media, email, sensors, and photographs that can be mined for useful information. Decreases in the cost of both storage and compute power have made it feasible to collect this data - which would have been thrown away only a few years ago. As a result, more and more companies are looking to include non-traditional yet potentially very valuable data with their traditional enterprise data in their business intelligence analysis.

1. INTRODUCTION

The media is full of reports about how big data will transform business, government and other aspects of the economy. As economists who happen to live and work in the epicenter of the data revolution, Silicon Valley, we have wondered for some time about how these developments might affect economics, especially economic research and policy analysis. In this article, we try to offer some thoughts. We start by trying to describe what is meant by big data, and what about it is new from the perspective of economists, who have been sophisticated users of data for a long time. From an economic policy perspective, we highlight the value of large administrative data sets, the ability to capture and process data in real time, and the potential for improving both the efficiency of government operations and informing economic policy---making. From an economic research perspective, we emphasize how large, granular datasets can enable novel research designs. We consider whether the big data tools being developed in statistics and computer science, such as statistical learning and data mining techniques, will find much application in economics.

While the potential benefits of Big Data are real and significant, and some initial successes have already been achieved (such as the Sloan Digital Sky Survey), there remain many technical challenges that must be addressed to fully realize this potential. The sheer size of the data, of course, is a major challenge, and is the one that is most easily recognized. However, there are others. Industry analysis companies like to point out that there are challenges not just in Volume, but also in Variety and Velocity [Gar2011], and that companies should not focus on just the first of these. By Variety, they usually mean heterogeneity of data types, representation, and semantic interpretation. By Velocity, they mean both the rate at which data arrive and the time in which it must be acted upon. While these three are important, this short list fails to include additional important requirements such as privacy and usability.

Big data typically refers to the following types of data:

 Traditional enterprise data – includes customer information from CRM systems, transactional ERP data, web store transactions, and general ledger data.

 Machine-generated /sensor data – includes Call Detail Records ("CDR"), weblogs, smart meters, manufacturing sensors, equipment logs (often referred to as digital exhaust), trading systems data.

 Social data – includes customer feedback streams, micro-blogging sites like Twitter, social media platforms like Facebook

The McKinsey Global Institute estimates that data volume is growing 40% per year, and will grow 44x between 2009 and 2020. But while it's often the most visible parameter, volume of data is not the only characteristic that matters. In fact, there are four key characteristics that define big data:

 Volume. Machine-generated data is produced in much larger quantities than non-traditional data. For instance, a single jet engine can generate 10TB of data in 30 minutes. With more than 25,000 airline flights per day, the daily volume of just this single data source runs into the Petabytes. Smart meters and heavy industrial equipment like oil refineries and drilling rigs generate similar data volumes, compounding the problem.

 Velocity. Social media data streams – while not as massive as machine-generated data – produce a large influx of opinions and relationships valuable to customer relationship management. Even at 140 characters per tweet, the high velocity (or frequency) of Twitter data ensures large volumes (over 8 TB per day).

 Variety. Traditional data formats tend to be relatively well defined by a data schema and change slowly. In contrast, non-traditional data formats exhibit a dizzying rate of change. As new services are added, new sensors deployed, or new marketing campaigns executed, new data types are needed to capture the resultant information.

 Value. The economic value of different data varies significantly. Typically there is good information hidden amongst a larger body of non-traditional data; the challenge is identifying what is valuable and then transforming and extracting that data for analysis.

To make the most of big data, enterprises must evolve their IT infrastructures to handle these new high-volume, high-velocity, high-variety sources of data and integrate them with the pre-existing enterprise data to be analyzed.

2. Importance of Big Data

Even twenty or thirty years ago, data on economic activity was reatively scarce. In just a short period of time, this has changed dramatically. One reason is the growth of the internet. Practically everything on the internet is recorded. When you search on Google or Bing, your queries and subsequent clicks are recorded. When you shop on Amazon or eBay, not only every purchase, but every click is captured and logged. When you read a newspaper online, watch videos, or track your personal finances, your behavior is recorded. The recording of individual behavior does not stop with the internet: text messaging, cell phones and geo--- locations, scanner data, employment records, and electronic health records are all part of the data footprint that we now leave behind us.

A specific example may be illustrative. Consider the data collected by retail stores. A few decades ago, stores might have collected data on daily sales, and it would have been considered high quality if the data was split by products or product categories. Nowadays, scanner data makes it possible to track individual purchases and item sales, capture the exact time at which they occur and the purchase histories of the individuals, and use electronic inventory data to link purchases to specific shelf locations or current inventory levels. Internet retailers observe not just this information, but can trace the consumer's behavior around the sale, including his or her initial search query, items that were viewed and discarded, recommendations or promotions that were shown, and subsequent product or seller reviews. And in principle these data could be linked to demographics, advertising exposure, social media activity, offline spending, or credit history.

There has been a parallel evolution in business activity. As firms have moved their day---to---day operations to computers and then online, it has become possible to compile rich datasets of sales contacts, hiring practices, and physical shipments of goods. Increasingly, there are also electronic records of collaborative work efforts, personnel evaluations and productivity measures. The same story also can be told about the public sector, in terms of the ability to access and analyze tax filings, social insurance programs, government expenditures and regulatory activities.

Obviously, this is a lot of data. But what exactly is new about it? The short answer is that data is now available faster, has greater coverage and scope, and includes new types of observations and measurements that previously were not available. A key aspect of such modern datasets is that they have much less structure, or more complex structure, than the traditional cross--sectional, time--series or panel data models that we teach in our econometrics classes.

When big data is distilled and analyzed in combination with traditional enterprise data, enterprises can develop a more thorough and insightful understanding of their business, which can lead to enhanced productivity, a stronger competitive position and greater innovation – all of which can have a significant impact on the bottom line.

For example, in the delivery of healthcare services, management of chronic or long-term conditions is expensive. Use of in-home monitoring devices to measure vital signs, and monitor progress is just one way that sensor data can be used to improve patient health and reduce both office visits and hospital admittance.

Manufacturing companies deploy sensors in their products to return a stream of telemetry. In the automotive industry, systems such as General Motors' OnStar ® or Renault's R-Link ®, deliver communications, security and navigation services. Perhaps more importantly, this telemetry also reveals usage patterns, failure rates and other opportunities for product improvement that can reduce development and assembly costs.

The proliferation of smart phones and other GPS devices offers advertisers an opportunity to target consumers when they are in close proximity to a store, a coffee shop or a restaurant. This opens up new revenue for service providers and offers many businesses a chance to target new customers.

Retailers usually know who buys their products. Use of social media and web log files from their ecommerce sites can help them understand who didn't buy and why they chose not to, information not available to them today. This can enable much more effective micro customer segmentation and targeted marketing campaigns, as well as improve supply chain efficiencies through more accurate demand planning.

Finally, social media sites like Facebook and LinkedIn simply wouldn't exist without big data. Their business model requires a personalized experience on the web, which can only be delivered by capturing and using all the available data about a user or member.

3. Analysis of Big Data

The analysis of Big Data involves multiple distinct phases as shown in the figure below, each of which introduces challenges. Many people unfortunately focus just on the analysis/modeling phase: while that phase is crucial, it is of little use without the other phases of the data analysis pipeline. Even in the analysis phase, which has received much attention, there are poorly understood complexities in the context of multi-tenanted clusters where several users' programs run concurrently. Many significant challenges extend beyond the analysis phase. For example, Big Data has to be managed in context, which may be noisy, heterogeneous and not include an upfront model. Doing so raises the need to track provenance and to handle uncertainty and error: topics that are crucial to success, and yet rarely mentioned in the same breath as Big Data. Similarly, the questions to the data analysis pipeline will typically not all be laid out in advance. We may need to figure out good questions based on the data. Doing this will require smarter systems and also better support for user interaction with the analysis pipeline. In fact, we currently have a major bottleneck in the number of people empowered to ask questions of the data and analyze it [NYT2012]. We can drastically increase this number by supporting many levels of engagement with the data, not all requiring deep database expertise. Solutions to problems such as this will not come from incremental improvements to business as usual such as industry may make on its own. Rather, they require us to fundamentally rethink how we manage data analysis.

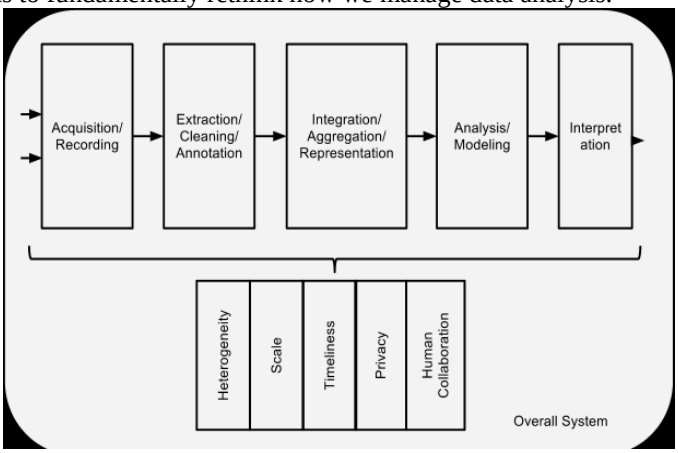

Fig. The Big Data Analysis Pipeline. Major steps in analysis of Big Data are shown in the flow at top. Below it are Big Data needs that make these tasks challenging.

Fortunately, existing computational techniques can be applied, either as is or with some extensions, to at least some aspects of the Big Data problem. For example, relational databases rely on the notion of logical data independence: users can think about

what they want to compute, while the system (with skilled engineers designing those systems) determines how to compute it efficiently. Similarly, the SQL standard and the relational data model provide a uniform, powerful language to express many query needs and, in principle, allows customers to choose between vendors, increasing competition. The challenge ahead of us is to combine these healthy features of prior systems as we devise novel solutions to the many new challenges of Big Data.

In this paper, we consider each of the boxes in the figure above, and discuss both what has already been done and what challenges remain as we seek to exploit Big Data. We begin by considering the five stages in the pipeline, then move on to the five cross-cutting challenges, and end with a discussion of the architecture of the overall system that combines all these functions.

Since data is not always moved during the organization phase, the analysis may also be done in a distributed environment, where some data will stay where it was originally stored and be transparently accessed from a data warehouse. The infrastructure required for analyzing big data must be able to support deeper analytics such as statistical analysis and data mining, on a wider variety of data types stored in diverse systems; scale to extreme data volumes; deliver faster response times driven by changes in behavior; and automate decisions based on analytical models. Most importantly, the infrastructure must be able to integrate analysis on the combination of big data and traditional enterprise data. New insight comes not just from analyzing new data, but from analyzing it within the context of the old to provide new perspectives on old problems.

For example, analyzing inventory data from a smart vending machine in combination with the events calendar for the venue in which the vending machine is located, will dictate the optimal product mix and replenishment schedule for the vending machine.

3.1Solution Spectrum

Many new technologies have emerged to address the IT infrastructure requirements outlined above. At last count, there were over 120 open source key-value databases for acquiring and storing big data, while Hadoop has emerged as the primary system for organizing big data and relational databases maintain their footprint as a data warehouse and expand their reach into less structured data sets to analyze big data. These new systems have created a divided solutions spectrum comprised of:

- Not Only SQL (NoSQL) solutions: developer-centric specialized systems
- SQL solutions: the world typically equated with the manageability, security and trusted nature of relational database management systems (RDBMS)

NoSQL systems are designed to capture all data without categorizing and parsing it upon entry into the system, and therefore the data is highly varied. SQL systems, on the other hand, typically place data in well-defined structures and impose metadata on the data captured to ensure consistency and validate data types.

3.2 Interpretation

Having the ability to analyze Big Data is of limited value if users cannot understand the analysis. Ultimately, a decision-maker, provided with the result of analysis, has to interpret these results. This interpretation cannot happen in a vacuum. Usually, it involves examining all the assumptions made and retracing the analysis. Furthermore, as we saw above, there are many possible sources of error: computer systems can have bugs, models almost always have assumptions, and results can be based on erroneous data. For all of these reasons, no responsible user will cede authority to the computer system. Rather she will try to understand, and verify, the results produced by the computer. The computer system must make it easy for her to do so. This is particularly a challenge with Big Data due to its complexity. There are often crucial assumptions behind the data recorded. Analytical pipelines can often involve multiple steps, again with assumptions built in. The recent mortgage-related shock to the financial system dramatically underscored the need for such decision-maker diligence -- rather than accept the stated solvency of a financial institution at face value, a decision-maker has to examine critically the many assumptions at multiple stages of analysis.

In short, it is rarely enough to provide just the results. Rather, one must provide supplementary information that explains how each result was derived, and based upon precisely what inputs. Such supplementary information is called the provenance of the (result) data. By studying how best to capture, store, and query provenance, in conjunction with techniques to capture adequate metadata, we can create an infrastructure to provide users with the ability both to interpret analytical results obtained and to repeat the analysis with different assumptions, parameters, or data sets.

4. Opportunities for Economic Policy

The potential uses of big data for economic policy roughly parallel the uses in the private sector. In this section, we start by describing the data resources available to the government, and also how private sector data might be used to better track and forecast economic activity. We then describe how big data might be used to inform policy decisions or to improve government services, along the lines of some of the information products and services described in the prior section.

4.1. Making Use of Government Administrative Data

Through its role in administering the tax system, social programs, and regulation, the federal government collects enormous amounts of granular administrative data. Examples include the rich micro---level datasets maintained by the Social Security Administration, the Internal Revenue Service, and the Centers for Medicare and Medicaid. Although there is less uniformity, state and local governments similarly generate large amounts of administrative data, particularly in areas such as education, social insurance and local government spending.

Government administrative data are almost certainly under---utilized, both by government agencies and, because of limited and restricted access, by researchers and private data endors who might use this data to uncover new facts. The major datasets also tend to be maintained separately, unlike in many European countries, which may have datasets that merge individual demographic, employment and in some cases health data, for the entire population.

4.2. New Measures of Private Sector Economic Activity

Government agencies also play an important role in tracking and monitoring private sector economic activity. Traditionally much of this has been done using survey methods. To measure price inflation, the Bureau of Labor Statistics sends surveyors out to stores to manually collect information on the posted prices and availability of approximately 80,000 appropriately selected items. These data are aggregated into various inflation indices such as the Consumer Price Index. Measures of unemployment, consumer expenditure, and wages and benefits rely on similar survey---based methodologies.

5. Opportunities for Economic Research

We now take up the question of how the data revolution might affect economic research, in terms of the scope and quality of the results, the methods used, and the required training of empirical economists. Since economic research is primarily retrospective analysis, some of the most obvious impact of big and granular data – providing a detailed snapshot of economic activity (almost) in real time – may not be as important for research. Yet, other aspects may. The first, and most obvious, effect of big data on economic research will be to allow better measurements of economic effects and outcomes. More granular and comprehensive data also can help to pose new sorts of questions and enable novel research designs that can inform us about the consequences of different economic policies and events. We will provide some examples below, all of which are in the spirit of empirical economics continuing to look more or less the same as it does now, only with more and better data. A less obvious possibility is that new data may end up changing the way economists approach empirical questions and the tool they use to answer them. As an example, we consider whether economists might end up embracing some of the statistical data---mining tools described earlier. Why is this less obvious? To begin, it would mean something of a shift away from the single covariate causal effects framework that has dominated much of empirical research over the last few decades. In the minds of many economists, there is a sharp distinction between predictive modeling and causal inference, and as a result statistical learning approaches have little to contribute. Our view is that such a distinction is not always so sharp, and we think that this type of work will be increasingly used in economics as big datasets become available for researchers and as empirical economists gain greater familiarity and comfort with machine learning statistical tools.

6. CHALLENGES

Several challenges confront economists wishing to take advantage of large new datasets. These include gaining access to data, developing the data management and programming capabilities needed to work with large---scale datasets, and finally and most importantly!) thinking of creative approaches to summarize, describe and analyze the information contained in these data.

Data Access. Research on topics such as labor economics, productivity and household consumption traditionally have relied on government survey data such as the U.S. Census, the Panel Study of Income Dynamics(PSID) and the National Longitudinal Survey of Youth (NLSY). For many of these data, there are well---established protocols or accessing and making use of the data. In some cases, 25 such as the U.S. sensus Data Research Centers, these protocols are cumbersome and probably discourage a fair number of researchers, but at least they reflect a conscious effort to tradeoff between research access and confidentiality concerns. These systems are still being worked out for the large---scale administrative data that recently has been used for economic research: from the IRS, Medicare, or Social Security Administration. He privacy issues associated with the increased amount of data are important, and have been already discussed in this publication just year ago Goldfarb and ucker, 2012). But as Card et al. (2010) point out, many European countries, such as Norway, Sweden and Denmark, have gone much farther to facilitate research. The experience in these countries suggests that broader access is possible, and that as may be expected, reducing the barriers to accessing such datasets profound effect on the amount of researchers using it. Many of the novel data we have discussed above belongs to private companies. Accessing private company data creates several issues for researchers. First and most obviously, not every company wants to work with researchers. While many view it as potentially beneficial, and a useful way to learn from outsiders, others ay view it as a distraction or focus on the publicity risks. Researchers who collaborate with companies generally need to enter into contracts to prevent disclosure of confidential information, and may face some limits on the questions they can study. Our experience has been that the benefits of working with company data generally far outweigh the costs, put that a fair amount of effort on both sides is required to develop successful collaborations.

With Big Data, the use of separate systems in this fashion becomes prohibitively expensive given the large size of the data sets. The expense is due not only to the cost of the systems themselves, but also the time to load the data into multiple systems. In consequence, Big Data has made it necessary to run heterogeneous workloads on a single infrastructure that is sufficiently flexible to handle all these workloads. The challenge here is not to build a system that is ideally suited for all processing tasks. Instead, the need is for the underlying system architecture to be flexible enough that the components built on top of it for expressing the various kinds of processing tasks can tune it to efficiently run these different workloads.

The very fact that Big Data analysis typically involves multiple phases highlights a challenge that arises routinely in practice: production systems must run complex analytic pipelines, or workflows, at routine intervals, e.g., hourly or daily. New data must be incrementally accounted for, taking into account the results of prior analysis and pre-existing data. And of course, provenance must be preserved, and must include the phases in the analytic pipeline. Current systems offer little to no support for such Big Data pipelines, and this is in itself a challenging objective.

7. CONCLUSION

We have entered an era of Big Data. Through better analysis of the large volumes of data that are becoming available, there is the potential for making faster advances in many scientific disciplines and improving the profitability and success of many enterprises. However, many technical challenges described in this paper must be addressed before this potential can be realized fully. The challenges include not just the obvious issues of scale, but also heterogeneity, lack of structure, error-handling, privacy, timeliness, provenance, and visualization, at all stages of the analysis pipeline from data acquisition to result interpretation. These technical challenges are common across a large variety of application domains, and therefore not cost-effective to address in the context of one domain alone. Furthermore, these challenges will require transformative solutions, and will not be addressed naturally by the next generation of industrial products. We must support and encourage fundamental research towards addressing these technical challenges if we are to achieve the promised benefits of Big Data.

There is little doubt that over the next decades "Big Data" will change the landscape of economic policy and economic research. As we emphasized throughout, we don't think that big data will substitute for common sense, economic theory or the need for careful research designs. Rather, it will complement them. How exactly remains to be seen. In this article we tried to lay out what we see as the vast opportunities, as well as challenges that come with the ongoing data revolution. We look forward to seeing how it will play out.

REFERENCES

[1] Belloni, Alexandre, D.Chen, Victor Chernozhukov and Christian Hansen (2012a). "Sparse Models and Methods for Optimal Instruments With an Application to Eminent Domain. "*Econometrica* 80(6), 2369---2429..

[2] Folowing the Breadcrumbs to Big Data Gold. Yuki Noguchi. *National Public Radio*, Nov. 29, 2011. http://www.npr.org/2011/11/29/142521910/the-digital-breadcrumbs-that-lead-to-big-data.

[3] The Search for Analysts to Make Sense of Big Data. Yuki Noguchi. *National Public Radio*, Nov. 30, 2011. http://www.npr.org/2011/11/30/142893065/the-search-for-analysts-to-make-sense-of-big-data.

[4] Big data: The next frontier for innovation, competition, and productivity. James Manyika, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, and Angela Hung Byers. McKinsey Global Institute. May 2011.

[5] Pattern-Based Strategy: Getting Value from Big Data. Gartner Group press release. July 2011. Available at http://www.gartner.com/it/page.jsp?id=1731916

[6] Drowning in numbers -- Digital data will flood the planet—and help us understand it better. *The Economist*, Nov 18, 2011. http://www.economist.com/blogs/dailychart/2011/11/big-data-0

[7] Goel, Sharad, Jake Hofman, Sebastien Lahaie, David Pennock and Duncan Watts (2010). "Predicting Consumer Behavior with Web Search." *Proceedings of the National Academy of Sciences* 107(41), 17486---17490.

[8] Finkelstein, Amy, Sarah Taubman, Bill Wright, Mira Bernstein, Jonathan Gruber, Joseph P. Newhouse, Heidi Allen, Katherine Baicker, and the Oregon Health Study Group (2012). "The Oregon Health Insurance Experiment: Evidence from the First Year." *Quarterly Journal of Economics* 127(3), 1057---1106.

[9] Using Data for Systemic Financial Risk Management. Mark Flood, H V Jagadish, Albert Kyle, Frank Olken, and Louiqa Raschid. Proc. Fifth Biennial Conf. Innovative Data Systems Research, Jan. 2011

[10] The Age of Big Data. Steve Lohr. *New York Times*, Feb 11, 2012. http://www.nytimes.com/2012/02/12/sunday-review/big-datas-impact-in-the-world.html

AUTHORS

Mr. Tushar M. Chavan, ME (CSE) ,First Year, Department of CSE, Prof. Ram Meghe Institute Of Technology and Research, Badnera, Amravati. Sant Gadgebaba Amravati University, Amravati, Maharashtra, India – 444701

Prof. S. P. Akarte, Assistant Professor, Department of CSE, Prof. Ram Meghe Institute Of Technology and Research, Badnera,Amravati. Sant Gadgebaba Amravati University, Amarvati, Maharashtra, India - 444701

