Proceedings of Statistics Canada Symposium 2016
Growth in Statistical Information: Challenges and Benefits

Big Data: A Survey Research Perspective

Reg Baker[1]

Abstract[2]

Big data is a term that means different things to different people. To some, it means datasets so large that our traditional processing and analytic systems can no longer accommodate them. To others, it simply means taking advantage of existing datasets of all sizes and finding ways to merge them with the goal of generating new insights. The former view poses a number of important challenges to traditional market, opinion, and social research. In either case, there are implications for the future of surveys that are only beginning to be explored.

1. Introduction

In 1987, to commemorate its 50th anniversary, *Public Opinion Quarterly* asked 16 well-known scholars and survey practitioners to offer their visions of the future of public opinion research (Bogart 1987). In his response, "James Beniger wrote that "a host of new technologies will . . . make possible the real-time mass monitoring of individual behavior . . . Survey research will increasingly give way to more direct measures of behavior made possible by new computer-based technologies."

Almost three decades later this future, if not now, is at least more clearly visible. This is the world of big data, and depending on where you sit and the type of research you do it is either a dream come true or the Apocalypse.

This paper begins by defining what we mean by the term "big data." It then explores some of the challenges we face when working within the big data paradigm. It concludes with some musings about how big data may impact the future of surveys.

2. Defining Big Data

What exactly do we mean when we talk about "big data?" The most often heard definition is the 3Vs—volume, variety, and velocity (Laney 2001). While that may be a neat summary of the challenges posed by big data, it is hardly a definition. Ward and Barker (2013) reviewed the definitions of big data most often used by various players in the big data ecosystem of IT consultants, hardware manufacturers, software developers, and service providers. They noted that most definitions touch on three primary attributes: size, complexity, and tools. They proposed this definition:

> "Big data is a term describing the storage and analysis of large and or complex data sets using a series of techniques including, but not limited to: NoSQL, MapReduce, and machine learning."

Put another way, we might simply say that:

[1] Marketing Research Institute International, 5073 Red Fox Run, Ann Arbor, MI, USA, 48105

[2] This paper is based on a chapter of the same title in Paul Biemer, Edith de Leeuw, Stephanie Eckman, Brad Edwards, Frauke Kreuter, Lars Lyberg, Clyde Tucker, and Brady West (Editors) *Total Survey Error in Practice* due to be published by Wiley in early 2017. Because of a copyright agreement with the publisher we are unable to present that chapter here, only this brief summary.

> Big data is a term that describes datasets so large and complex that they cannot be processed or analyzed with conventional software systems.

Dutcher (2014) asked 40 different thought leaders for their definitions and got roughly 40 different responses. Those responses are summarized in the word cloud in Figure 1.

Figure 1
Word Cloud Based on Dutcher (2014).

We also might further elaborate on that by noting three principal sources:

1. **Transaction data** that describe some event, such as when a person interacts with a business or government entity. Customer data and government administrative data are two common examples.
2. **Social media data** drawn (or scraped) from social media networks, blogs, ecommerce reviews, web searches, and other sources of free form text and non-traditional data such as photographs and video recordings.
3. **The Internet of Things (IoT)** meaning data collected from interconnected devices covering a wide range of objects including autos, household appliances, scanners, traffic lights, security cameras, wearable sensors, GPS locators, and so forth. IoT is the least deployed but potentially most far reaching of the three sources of big data.

3. Some Challenges

Researchers interested in exploiting the potential of all these data face at least four major challenges.

3.1 Technology

Big data is a world of terabytes, petabytes, exabytes, and zettabytes. It is Walmart capturing more than a million customer transactions each hour and uploading them to a database in excess of 3 petabytes (SAS 2102). It is the Weather Channel gathering 20 terrabytes of data from sensors all around the world each and every day (Henchen 2013). It is the astonishing number of data transactions being generated every minute on social media and interconnected smart devices such as scanners and wearable technologies. By comparison, market, opinion, and social researchers still mostly work in a world of gigabytes. We simply are not accustomed to working at big data's scale.

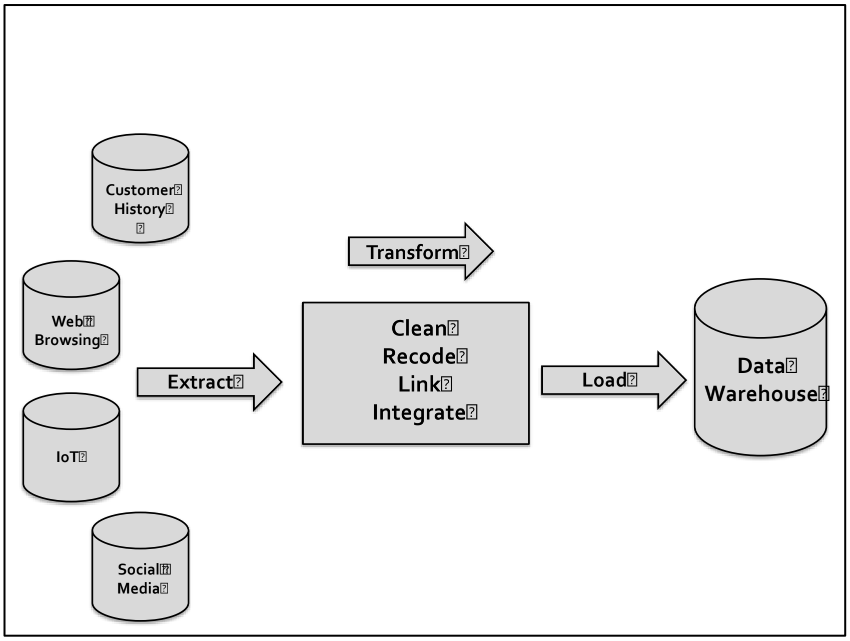
Taking big data seriously also requires a significant investment in people as well as technology. There is more to big data than hiring a data scientist. The AAPOR Report On Big Data (2015) has a useful summary of the skills and technology needed to do big data. While the report does not put a price tag on the investment, it likely is well beyond what all but the very largest research organizations can afford.

3.2 Data Quality

Researchers working in big data often are quick to point out that the quality of the data is not what we are accustomed to. More often than not the data were collected for some purpose other than research and the attention paid to the accuracy of individual items, their overall completeness, their consistency over time, their full documentation, and even their meaning pose serious challenges to their reuse. Readers familiar with the Total Survey Error model (TSE) as described by Groves (1989) will recognize that big data is vulnerable to all of the same deficiencies as surveys—gaps in coverage, missing data, poor measurement, etc. The key difference is that survey researchers, at least in theory, design and control the data making process in a way that users of big data do not.

Much of the value of big data lies in the potential to merge multiple data sets together (e.g. customer transaction data with social media data or IoT data). That is an expensive and difficult process (See Figure 2), but it also is a point at which errors are easily introduced. The heart of this merging process is bits of computer code called ETLs (for extract, transform, and load) that specify what data are extracted from the source databases, how they are edited and transformed for consistency, and then merged to the output database, typically some type of data warehouse.

Figure 2
Simplified view of Data Linkage



Take a moment and consider the difficulty of specifying all of those rules. If you have ever written editing specs for a survey dataset then you have some inkling of the difficulty. Now consider that in a data merge from multiple sources you can have the same variable with different coding; the same variable name used to measure different things; differing rules for determining when a item is legitimately missing and when it is not; detailed rules for matching a record from one data source with a record from another; different entities (customers, products, stores, GPS coordinates, tweets) that need to resolved; and so on. This is difficult, tedious, unglamorous, and error-prone work. Get it wrong, and you have a mess.

3.3 Analytics

And then there is the issue of tools. Most of the software survey researchers routinely use grinds to a halt with big data. It's just not built to process files at the petabyte scale. There is a whole suite of tools, virtually all of which relay on massive parallel processing, that are well beyond what most of us are even thinking about.

Market, opinion, and social researchers—including those working in official statistics--are doing some interesting and worthwhile things with what is probably more accurately described as "secondary data." It is legitimate to ask whether this is really big data. And even if it is, most still have not grasped the importance of the analytic shift required to really exploit big data's potential.

Consider Chris Anderson's famous 2008 *Wired* editorial, "The end of theory: The data deluge makes the scientific method obsolete."

> "Faced with massive data, this approach to science — hypothesize, model, test — is becoming obsolete. . . Petabytes allow us to say: 'Correlation is enough.' We can stop looking for models. We can analyze the data without hypotheses about what it might show. We can throw the numbers into the biggest computing clusters the world has ever seen and let statistical algorithms find patterns where science cannot."

This is a fair statement of the data science perspective and its faith in machine learning--the use of algorithms capable of finding patterns in data unguided by a set of analytic assumptions about the relationship among data items. To paraphrase Vasant Dhar (2013), we are used to asking the question, "Do these data fit this model?" The data scientist asks the question, "What model fits these data?"

The research design and analytic approaches that are at the core of survey researchers were developed at a time when data were scarce and expensive, when the analytic tools at our disposal were weak and under powered. The combination of big data and rapidly expanding computing technology has changed that calculus.

This may sound heretical to many of us in the social sciences. But there also is a longstanding argument within the statistical profession about the value of these algorithmic analysis methods. For example, in 2001 the distinguished statistician Leo Breiman described two cultures within the statistical profession.

> "One assumes that the data are generated by a given stochastic data model. The other uses algorithmic models and treats the data mechanism as unknown. . .If our goal as a field is to use data to solve problems, then we need to move away from exclusive dependence on data models and adopt a more diverse set of tools."

One can find similar arguments from statisticians going back to the 1960s (See, for example, Tukey 1962).

There are dangers, of course, and arguments about correlation versus causality (and endogeneity) are to be taken seriously. There is even a website (http://www.tylervigen.com/spurious-correlations) devoted to some of the more entertaining but completely wrong conclusions one can reach from correlations run amok. But any serious data scientist will be quick to note that doing this kind of analysis requires more than good math skills, massive computing power, and a library of machine learning algorithms. Domain knowledge and critical judgment are essential. Or, as Nate Silver (2012) reminds us, "Data-driven predictions can succeed—and they can fail. It is when we deny our role in the process that the odds of failure rise"

3.4 Ethics

Two of the most important pillars of the ethical foundation of market, opinion, and social research are consent and confidentiality. Consent means that prospective research participants are provided with a description of the purpose of the research and how their data will be used. Confidentiality means that the researcher will provide the level of data protection needed to ensure that the identity of an individual participant is never disclosed to any party without the explicit consent of that participant, and that any data released for analysis will be anonymized.

The evolving world of big data is problematic on both fronts. When we reuse data collected for some other purpose we have an obligation to determine whether our planned use of the data is consistent with the terms under which individuals agreed to provide them. Further, recent research has demonstrated that the amount of data now available on individuals and the processing power to combine and analyze it has rendered traditional approaches to anonymizing obsolete (See, for example, Sweeny 2013; de Montjoye et al. 2015). For further discussion see Lane et al (2014), AAPOR Report on Big Data (2015).

4. Big Data and the Future of Surveys

Market researchers, in particular, spend a great deal of time these days thinking about and forecasting the future of the broader research industry (See, for example, Kaden et al. 2012). There is some consensus forming around three transitions that over time will transform what we do.

First, it is widely believed that we are moving from a world where data are scarce and expensive to one in which they are plentiful and cheap. Put simply, this is the difference between surveys (expensive) and the world created by the embedding of technology in all aspects of our daily lives.

Second, research will focus less on gathering data by asking questions and more on observation and listening. The increasing and widespread acceptance of principles of cognitive psychology such as dual process theory (Kahneman 2011) has lead many to argue that behavior is a more reliable predictor of the choices people make than asking about attitudes and intentions as is typically done in surveys.

Third, as the data sources available for studying a topic multiply, researchers will emphasize synthesis of data from multiple sources and methods over analysis of a single dataset.

In this world, surveys become just one of many ways to do research. When Joan Lewis, Consumer and Market Knowledge Officer at Proctor and Gamble was asked whether social media would replace surveys she responded, “We need to be methodology agnostic.” (Neff 2011) By that she meant that there are multiple methods and data sources that might be used to study a problem. We already see this playing out in official statistics as national institutes around the world migrate away from censuses based on classic survey data collection techniques to the use of administrative data.

The long-term viability of surveys may well rest on their designed character, their ability to target a specific population, to specify the data of interest, and to design a collection process that minimizes error. There is an element of taking what you can get with big data and issues of coverage, meaning, and accuracy are a constant concern. The future of big data, and perhaps that of the survey profession, rests on how well we solve these problems.

For now, the most prevalent view seems to be one that stresses the complementarity of big data and surveys (Forsyth and Boucher 2015; Macer 2015). As comforting as that may be, it does not free us from the responsibility to broaden our understanding of data sources and methods beyond surveys. To draw on Couper’s (2013) analogy, surveys are like screwdrivers with different uses and capabilities. But, they are not the only tools in our kit. There are also hammers, pliers, wrenches, and drills with each designed for a task that screwdrivers cannot do. To do a job right we use them in combination. The challenge to our profession is to learn when and how to use the right tool(s) for the job at hand. That may be a tall order for those of us accustomed to a survey as the always-obvious choice.

References

AAPOR (2015) AAPOR Report on Big Data. Retrieved on March 30, 2015 from http://www.aapor.org/AAPORKentico/AAPOR_Main/media/Task-Force-Reports/BigDataTaskForceReport_FINAL_2_12_15.pdf

Anderson, C. (2008) The end of theory: The data deluge makes the scientific method obsolete. *Wired*. 16(7). Retrieved on June 19, 2014 from http://www.wired.com/science/discoveries/magazine/16-07/pbtheory.

Bogart, Leo. 1987. “The Future of Public Opinion: A Symposium,” in *Public Opinion Quarterly*, Vol. 51, Supplement, pp. S173-S191.

Brieman, Leo. (2001). Statistical Modeling: The Two Cultures**.** *Statistical Sciences***.** 16(3) 199-231.

Couper, M.P. 2013. Is the Sky Falling? New Technology, Changing Media, and the Future of Surveys. *Survey Research Methods* 7(3): 145-146.

de Montjoye, Y., Radelli, L., Singh, V.K., and Pentland, A. 2014. Unique in the shopping mall: On the reidentifiability of credit card metadata. *Science*. 347(6221), 536-539.

Dhar, Vasant. 2013. Data Science and Prediction. *Communications of the ACM*. 56, 12, 64-73.

Dutcher, J. 2014. What is Big Data? Retrieved on June 14, 2015 from http://datascience.berkeley.edu/what-is-big-data/

Forsyth, J., and Boucher, L. (2015). “Why Big Data is Not Enough.” *Research World*. 50, January/February, 26-27.

Goves, R. M. (1989). *Survey Errors and Survey Costs.* New York: Wiley and Sons.

Henschen, D. (2013) “Big Data Reshapes Weather Channel Predictions” *Information Week* Retrieved on June 21, 2015 from http://www.informationweek.com/big-data/software-platforms/big-data-reshapes-weather-channel-predictions/d/d-id/1112776

Kaden,, R.J., Linda, M. and Prince, M. (2012) eds. *Leading Edge Marketing Research: 21st Century Tools and Practices*, Washington, DC: Sage.

Kahneman, Daniel (2011) *Thinking Fast and Slow*. New York: Farrar, Straus, and Giroux.

Lane, J., Stodden, V., Bender, S. and Nissenbaum, H. (2014) *Privacy, Big Data, and the Public Good: Frameworks for Engagement*, New York: Cambridge University Press.

Laney, Douglas (2001). 3D Data Management: Controlling Data Volume, Velocity and Variety. Retrieved on February 5, 2015 from http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf.

Macer, Tim. (2015). “Big Data Plus Research Means More Accurate Results.” *Research World*. 52, May/June, 13-15.

Neff, Jack. (2011). “Will Social Media Replace Surveys as a Research Tool?” *Advertising Age*. Retrieved on April 12, 2015 from http://adage.com/article/news/p-g-surveys-fade-consumers-reach-brands-social-media/149509/

SAS (2012). Big data meets big data analytics. Retrieved on December 28, 2104 from http://www.sas.com/resources/whitepaper/wp_46345.pdf.

Silver, N. (2012) *The Signal and the Noise*. New York: The Penguin Press.

Sweeny, L. 2013. Matching Known Patients to Health Records in Washington State Data. Retrieved on July 17, 2015 from http://dataprivacylab.org/projects/wa/1089-1.pdf.

Tukey, J.W. (1962). “The Future of Data Analysis.” *The Annals of Mathematical Statistics*. 33(1): 1-67.

Ward, Jonathan Stuart, and Adam Barker (2013). Undefined by data: a survey of big data definitions. arXiv: 1309.5821vi. Retrieved on November 12, 2014 from http://arxiv.org/pdf/1309.5821v1.pdf.