Big Data in Survey Research

Article *in* Public Opinion Quarterly · December 2015

DOI: 10.1093/poq/nfv039

CITATIONS
109

READS
7,225

9 authors, including:

Lilli Japec
Statistics Sweden
15 PUBLICATIONS **482** CITATIONS
SEE PROFILE

Julia Lane
New York University
227 PUBLICATIONS **4,815** CITATIONS
SEE PROFILE

Public Opinion Quarterly, Vol. 79, No. 4, Winter 2015, pp. 839–880

BIG DATA IN SURVEY RESEARCH
AAPOR TASK FORCE REPORT

LILLI JAPEC
Statistics Sweden
FRAUKE KREUTER*
Joint Program in Survey Methodology at the University of Maryland, University of Mannheim & Institute for Employment Research, Nuremberg, IAB
MARCUS BERG
Stockholm University
PAUL BIEMER
RTI International
PAUL DECKER
Mathematica Policy Research
CLIFF LAMPE
School of Information at the University of Michigan
JULIA LANE
Wagner School and Center for Urban Science and Progress, New York University
CATHY O'NEIL
Data Science Consultant
ABE USHER
HumanGeo Group

Lilli Japec is the director of the R&D Department at Statistics Sweden, Stockholm, Sweden. Frauke Kreuter is a professor in the Joint Program in Survey Methodology at the University of Maryland, College Park, MD, USA; professor at the University of Mannheim, Mannheim, Germany; and head of the Statistical Methods Research Department at the Institute for Employment Research (IAB) in Nürnberg, Germany. Marcus Berg is an adjunct lecturer in the Department of Statistics at Stockholm University, Stockholm, Sweden. Paul Biemer is a distinguished fellow at RTI International, Research Triangle Park, NC, USA, and associate director for survey research and director of the certificate program in survey methodology at the Odum Institute for Research in Social Sciences at the University of North Carolina, Chapel Hill, NC, USA. Paul Decker is president and CEO of Mathematica Policy Research, Princeton, NJ, USA. Cliff Lampe is an associate professor in the School of Information at the University of Michigan, Ann Arbor, MI, USA. Julia Lane is a professor in the Wagner School and Professor of Practice at the Center for Urban Science and Progress, New York University, New York, NY, USA; professor at the Melbourne Institute of Applied Economics and Social Research, University of Melbourne, Melbourne, Australia; professor at the BETA University of Strasbourg, Strasbourg, France; and institute fellow at the American Institutes for Research, Washington, DC, USA. Cathy O'Neil is a data science consultant, New York, NY, USA. Abe Usher is the chief technology officer at the HumanGeo Group, Washington, DC, USA. The authors are grateful for comments, feedback, and editorial help from Eran Ben-Porath, Jason McMillan, and the AAPOR council members. *Address correspondence to Frauke Kreuter, Joint Program in Survey Methodology, 1218 LeFrak Hall, College Park, MD 20742, USA; phone: 301-314-7911; e-mail: fkreuter@umd.edu.

doi:10.1093/poq/nfv039

Abstract Recent years have seen an increase in the amount of statistics describing different phenomena based on "Big Data." This term includes data characterized not only by their large volume, but also by their variety and velocity, the organic way in which they are created, and the new types of processes needed to analyze them and make inference from them. The change in the nature of the new types of data, their availability, and the way in which they are collected and disseminated is fundamental. This change constitutes a paradigm shift for survey research. There is great potential in Big Data, but there are some fundamental challenges that have to be resolved before its full potential can be realized. This report provides examples of different types of Big Data and their potential for survey research; it also describes the Big Data process, discusses its main challenges, and considers solutions and research needs.

What Is Big Data?

The term "Big Data" is an imprecise description of a rich and complicated set of characteristics, practices, techniques, ethical issues, and outcomes all associated with data.

Big Data originated in the physical sciences, with physics and astronomy early to adopt many of the techniques now called Big Data. Instruments like the Large Hadron Collider and the Square Kilometer Array are massive collectors of exabytes of information, and the ability to collect such massive amounts of data necessitated an increased capacity to manipulate and analyze these data as well.

More recently, large data sources have been mined to enable insights about economic and social systems, which previously relied on methods such as surveys, experiments, and ethnographies to drive conclusions and predictions. Below are some recent examples. Not all of these might immediately match what people have in mind when they think about Big Data; however, all of them share characteristics of Big Data, as presented below.

EXAMPLE 1: ONLINE PRICES

The MIT Billion Prices Projects, PriceStats,[1] is an academic initiative using prices collected daily from hundreds of online retailers around the world to conduct economic research. One statistical product is the estimation of inflation in the United States. Changes in inflation trends can be observed sooner in PriceStats than in the monthly Consumer Price Index (CPI). Some National Statistical Institutes in Europe are now using Internet robots to collect prices from the web or scanner data from retailers as part of their data collection for the CPI (Norberg, Sammar, and Tongur 2011; ten Bosch and Windmeijer 2014).

1. http://bpp.mit.edu/.

EXAMPLE 2: TRAFFIC AND INFRASTRUCTURE

Big Data can be used to monitor traffic or to identify infrastructural problems. For example, Statistic Netherlands uses traffic loop detection data to measure traffic intensity (Daas et al. 2013). Each loop counts the number of vehicles per minute that pass at that location, and measures speed and length. The City of Boston issued a smartphone application available to anybody, which is designed to automatically detect pavement problems.[2] Anyone who downloads the mobile app creates data about the smoothness of the ride. According to their website, these data provide the city with real-time information that it uses to fix problems and plan long-term investments.

EXAMPLE 3: SOCIAL MEDIA MESSAGES

A consumer confidence index is produced every month by Statistics Netherlands using survey data. The index measures households' sentiments on their financial situation and on the economic climate in general. Daas and Puts (2014) studied social media messages to see if they could be used to measure social media sentiment. They found that the correlation between social media sentiment (mainly Facebook data) and consumer confidence is very high (see figure 1).

Social media messages (in this case Twitter data) form the basis of the University of Michigan Social Media Job Loss Index,[3] with the goal of generating early predictions of initial claims for unemployment insurance. The predictions are based on a factor analysis of social media messages mentioning job loss and related outcomes (Antenucci et al. 2014).

CHARACTERISTICS OF BIG DATA

In order to know when and how Big Data can be an appropriate technique for social insight, it is important to know more about the different features of Big Data. While there is no singularly preeminent Big Data definition, one very widely used definition comes from a 2001 Gartner report (Laney 2001, 2012) describing several characteristics of Big Data:

Volume refers to the sheer amount of data available for analysis. This volume of data is driven by the increasing number of data-collection instruments (e.g., social media tools, mobile applications, sensors) as well as the increased ability to store and transfer those data with recent improvements in data storage and networking.

Velocity refers to both the speed at which these data-collection events can occur, and the pressure of managing large streams of real-time data. Across the means of collecting social information, new information is being added to the database at rates ranging from as slow as every hour or so to as fast as thousands of events per second.

2. http://bit.ly/1yrMKHB.
3. http://bit.ly/1meDoas.

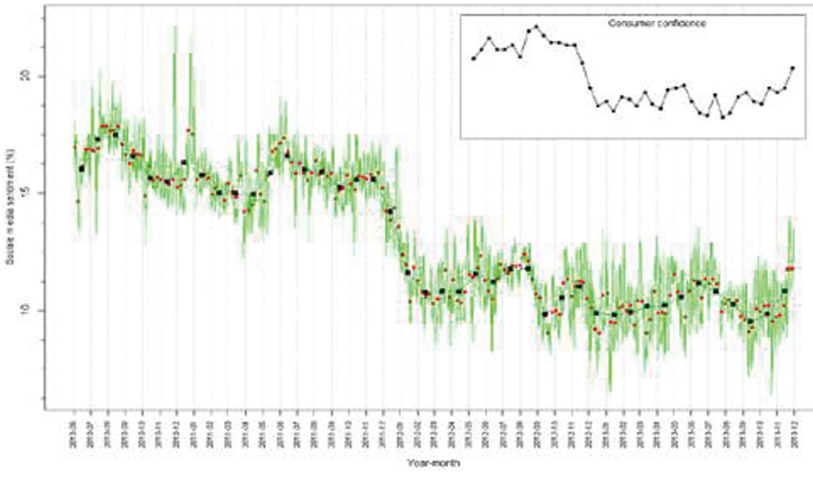

Figure 1. Social Media Sentiment (daily, weekly and monthly) in the Netherlands, June 2010–November 2013. The development of consumer confidence for the same period is shown in the insert (Daas and Puts 2014).

Variety refers to the complexity of formats in which Big Data can exist. Besides structured databases, there are large streams of unstructured documents, images, e-mail messages, video, links between devices, and other forms that create a heterogeneous set of data points. One effect of this complexity is that structuring and tying data together becomes a major effort, and therefore a central concern of Big Data analysis.

Others have added additional characteristics to the definition. These include *Variability* (inconsistency of the data across time), *Veracity* (ability to trust that the data are accurate), and *Complexity* (need to link multiple data sources).

There are many different types of Big Data sources; for example, *social media data*, *personal data* (e.g., data from tracking devices), *sensor data*, *transactional data*, and *administrative data*.

There are different opinions on whether administrative data should be considered to be Big Data or not. Administrative data are usually large in volume; they are generated for a different purpose and arise organically through administrative processes. Also, the content of administrative data is usually not designed by researchers. For these reasons, and because there is great potential in using administrative data, we will consider it to be in scope for this report.

There are a number of differences between administrative data and other types of Big Data that are worth pointing out. The amount of control a researcher has and the potential inferential power vary between different types of Big Data sources. For example, a researcher will likely not have any control of data from different social media platforms, and it could be difficult

to decipher a text from social media. For administrative data, on the other hand, a statistical agency can form a partnership with owners of the data and influence the design of the data. Administrative data are more structured and well defined, and more is known about the data than perhaps other Big Data sources.

BIG DATA AS "FOUND" DATA

A dimension of Big Data not often mentioned in the practitioner literature, but important for survey researchers to consider, is that Big Data are often secondary data, intended for another primary use. This means that Big Data are typically related to some non-research purpose and then reused by researchers to make a social observation. This is related to Sean Taylor's distinction between "found vs. made" data (Taylor 2013). He argues that a key difference between Big Data approaches and other social science approaches is that the data are not being initially "made" through the intervention of some researcher. When a survey researcher constructs an instrument, there are levels of planning and control that are necessarily absent in the data used in Big Data approaches. Big Data sources might have only a few variables, as compared with surveys that have a set of variables of interest to the researcher. In a 2011 Public Opinion Quarterly article and a blog post in his former role as director of the US Census Bureau, Robert Groves described a similar difference between organic and designed data (Groves 2011a, 2011b).

In the context of public opinion studies, a survey researcher could measure opinion by prompt responses about a topic that may never naturally appear in a Big Data source. On the other hand, the "found" data of social media are "nonreactive," or "naturally occurring," so that a data point, devoid of researcher manipulation, may be a more accurate representation of a true opinion or behavior. "Found" data may be a behavior, such as a log of steps drawn from networked pedometers or the previously mentioned recordings of travel patterns, which might be more accurate than what could be solicited in surveys given known problems with recall error (Tourangeau, Rips, and Rasinski 2000).

While the scale of data often used is what receives prominence, hence the name **Big** Data, it is actually this "found" nature of the data that is of concern to survey researchers. For example, since the data were not created for research, there often are no informed consent policies surrounding their creation, leading to ethical concerns. Additionally, there are statistical concerns with respect to the representative nature of the data. While these are serious concerns covered in more depth later in this report, they are not necessarily fatal to the proposition that Big Data can be used to construct social insights.

Data created through administering the tax systems, social programs, and regulation are also a form of "found" data. They are not created with a specific scientific research question in mind, but rather are the byproduct for the

respective administrative processes, just as (certain types of) paradata are created as a byproduct of survey data collections. In many instances, these administrative data are large in volume and share the unstructured nature of many other Big Data sources.

PARADIGM SHIFT

Before considering the usability and use of Big Data, it is worth exploring the paradigm shift happening in the presence of these new data sources. This change in paradigm stems from changes in many factors affecting the measurement of human behavior: the nature of the new types of data, their availability, and the way in which they are collected, mixed with other data sources, and disseminated. The consequences of these changes for public opinion research are fundamental in both the analysis that can be done and who the analysts might be. While the statistical community has moved beyond survey and even administrative data to begin understanding how to mine data from social media to capture national sentiment, from cell phone data to understand or even predict anti-government uprisings, and from financial data to examine swings in the economy, it is equally important to note that now some data are freely available and usable to anyone who wishes to mesh data points and series together and produce such analyses. With data readily accessible on the Internet, this creates opportunities for amateur, rather than professional, data analysts.

The change in the nature of the new type of data is transformative. Its characteristics—its velocity, volume, and variety—and the way in which it is collected mean a new analytical paradigm is open to statisticians and social scientists (Hey, Tansley, and Tolle 2009). The classic statistical paradigm was one in which researchers formulated a hypothesis, identified a population frame, designed a survey and a sampling technique, and then analyzed the results (Groves 2011a). The new paradigm means it is now possible to digitally capture, semantically reconcile, aggregate, and correlate data. These correlations might be effective (Halevy, Norvig, and Pereira 2009; Cukier and Mayer-Schoenberger 2013) or suspect (Couper 2013), but they enable completely new analyses to be undertaken—many of which would not be possible using survey data alone. For example, the new type of analysis might be one that captures rich environmental detail on individuals from sensors, Google Earth, videos, photos, or financial transactions. Alternatively, the analysis might include detailed information on unique and quite small subsets of the population (from microbiome data, or websearch logs), or the analysis could be on completely new units of analysis, like networks of individuals or businesses, whose connections can be captured only by new types of data (like tweets, cell phone conversations, and administrative records). As Kahneman (2011) points out, the new measurement can change the paradigm in its own right.

The change in paradigm also means changes in the production of public opinion research. The changes in the way data are processed and the type of skills needed to process the data are driven, in part, by the cost of converting data to usable information. The production process is very different in a Big Data world relative to a survey world. One of the most obvious Big Data advantages is that electronic data gathering is substantially cheaper than surveys. Surveys are inherently expensive, requiring a good deal of labor to collect the data. In contrast, Big Data, by relying on computer software and electronic data gathering, while requiring some upfront and maintenance costs, can be much more cost effective. But while Big Data are relatively cheap to collect, they can be expensive to clean and process, requiring a reallocation of the human capital that previously went into designing and sampling to structuring, linking, and managing the new types of data.

The change in data ownership has also transformed the way in which data are disseminated. The population of potential data analysts—trained and untrained—has dramatically expanded. This expansion can result in tremendous new insights, as the Sloan Digital Sky Survey and the Polymath project have shown (Nielsen 2012), and is reflected in Grey's Fourth Paradigm (figure 2) (Hey, Tansley, and Tolle 2009), but can also lead to the degradation of the quality of analysis that can be done and issues with the conclusions drawn and reported based on these data. AAPOR as an organization will need to find its place in giving guidance to the proper use of these data with respect to public opinion research.

Finally, the excitement of the change in research paradigm should be tempered by a recognition that our existing ways of protecting confidentiality are

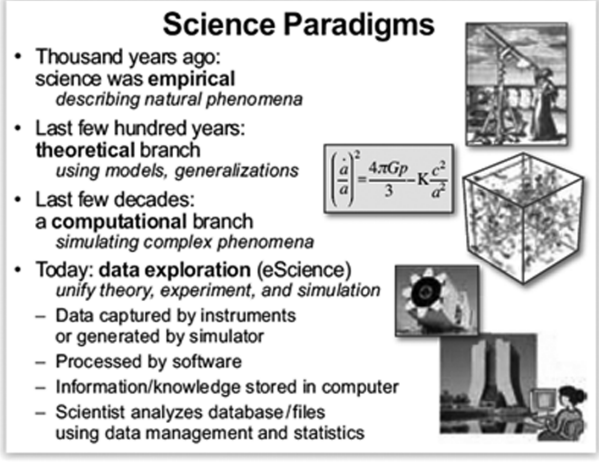

Figure 2. Science Paradigms from Hey, Tansley, and Tolle (2009).

no longer viable (Karr and Reiter 2014). As order and analytical rigor are hopefully brought to the new data frontier, we should ensure that the future structure of data access allows for good science to be attained while protecting the confidentiality of the unsuspecting contributors to this science. There is a great deal of research that can be used to inform the development of such a structure, but it has been siloed into disconnected research areas, such as statistics, cybersecurity, and cryptography, as well as a variety of different practical applications, including the successful development of remote-access secure data enclaves. We must piece together the knowledge from these various fields to develop ways in which vast new sets of data on human beings can be collected, integrated, and analyzed while protecting them (Lane et al. 2014). Here, too, AAPOR should extend the role it is currently playing, and involve itself in the discussion.

Why Big Data Matters

Personal data have been hailed as the "new oil" of the 21st century (Greenwood et al. 2014), with profound benefits to policy, society, and public opinion research. Detailed data on human beings can be used by policymakers to reduce crime, improve health delivery, and better manage cities (Keller, Koonin, and Shipp 2012). Society can benefit from these data as well. Recent work shows that data-driven businesses were 5 percent more productive and 6 percent more profitable than their competitors (Brynjolfsson, Hitt, and Kim 2011; McAfee and Brynjolfsson 2012). Using data with high volume, velocity, and variety, public opinion researchers can potentially increase the scope of their data-collection efforts while at the same time reducing costs, increasing timeliness, and increasing precision (Murphy et al. 2014).

The value of Big Data to each of these groups (policymakers, businesses, and researchers), and the balancing of the benefits and costs, including the risks of using these new data assets, differs for them because the calculus is different for each group. The Big Data benefits to policymakers have been well and often stated (Lohr 2012; Koonin and Holland 2014). The White House has noted that "Big Data technology stands to improve nearly all the services the public sector delivers" (Executive Office of the President 2014), but the costs of realizing these benefits are nontrivial. As mentioned earlier, even if data collection is cheap, the costs of cleaning, curating, standardizing, integrating, and using the new types of data can be substantial. Oftentimes federal, state, and local agencies do not have the internal capacity to do such analysis (Pardo 2014), and as a consequence they must make the data available either to consultants or to the research community, requiring the development of access protocols and modalities. Indeed, the federal government, many state governments, and some local governments have appointed Chief Data Officers to spearhead these many activities (Griffin 2008; Pardo 2014).

There are also substantial risks associated with replacing traditional data-collection methods, one of which is the misallocation of resources. For example, overreliance on Twitter data in deploying resources in the aftermath of hurricanes can lead to the misallocation of resources toward young, Internet-savvy people with cell phones and away from elderly or impoverished neighborhoods (Shelton et al. 2014). But all data-collection approaches suffer from similar risks. Poor survey methodology led the *Literary Digest* to incorrectly call the 1936 presidential elections (Squire 1988). Inadequate understanding of coverage, incentive, and quality issues, together with the lack of a comparison group, has hampered the use of administrative records, famously in the case of using administrative records on crime to make inferences about the role of death penalty policy in crime reduction (Donohue and Wolfers 2006; Levitt and Miles 2006). But, as is the case in traditional data-collection methods, it is also important to document these risks under the new approaches in order to address them. Another risk is the alienation of the people on whom the data are gathered. In some areas, there are no clear rules or guidelines governing privacy in this new world in which public and some private actions generate data that can be harvested (Ohm 2010; Executive Office of the President 2014; Strandburg 2014). Similarly, there are no clear data stewards or custodians who can be entrusted to preserving privacy and confidentiality (Lane and Stodden 2013).

The use of Big Data for research purposes also has substantial benefits to society. Commercial products can be effectively targeted to the right consumers, health interventions can be better designed, and taxpayers may pay less for government services (Lohr 2012). Commercial firms also benefit from lower expenses and greater efficiency (Brynjolfsson, Hitt, and Kim 2011; Tambe and Hitt 2012). Challenges include risks that are not well understood and quantified (Barocas and Nissenbaum 2014). Better crime data can help target police resources, but it can also exacerbate racial tensions (Gelman, Fagan, and Kiss 2007). More data on possible terrorists, like the Boston bomber, can aid in quick identification, but can also wrongly identify innocent citizens as terrorists (Tapia, LaLone, and Kim 2014). As Acquisti has noted:

> The mining of personal data can help increase welfare, lower search costs, and reduce economic inefficiencies; at the same time, it can be a source of losses, economic inequalities, and power imbalances between those who hold the data and those whose data is controlled. For instance, a firm may reduce its inventory costs by mining and analyzing the behavior of many individual consumers; however, the infrastructure needed to carry out analysis may require substantial investments, and if the analysis is conducted in manners that raise consumers' privacy concerns, those investments may backfire. Likewise, a consumer may benefit from contributing her data to a vast database of individuals' preferences (for instance, by sharing music interests with an online vendor, and receiving

> in turn targeted recommendations for new music to listen to); that same consumer, having lost control over that data, may end up suffering from identity theft, price discrimination, or stigma associated with the information unintended parties can acquire about her. (Acquisti 2014, 98)

The benefits for *public opinion* researchers are potentially extraordinary. The new type of data collection has been referred to as creating a "fourth paradigm" for science (Hey, Tansley, and Tolle 2009), and the importance of the intersection between social science and computer science represented by Big Data analysis has been recognized by major professional associations (Schenker, Davidian, and Rodriguez 2013). One clear benefit is that it adds to researchers' analytical tool kit. In addition to careful hypothesis-driven data collection, the new data have, as Robert Groves (2011a) has pointed out, four common and salient attributes that need to be incorporated into the research mindset: (1) they tend to measure behaviors, not internalized states like attitudes or beliefs; (2) they tend to offer near-real-time records of phenomena, and they are highly granulated temporally; (3) they tend to be lean in number of variables, many merely having some sort of an identifier and one other variable (e.g., a text tweet, a GPS coordinate); and (4) they rarely offer well-defined coverage of a large population (we don't know who isn't on Facebook, Twitter, or Google searches). Developing statistical techniques that exploit the richness of the data but preserve inference will be critical (Varian 2014), and so is the combination of data sources.

But most interestingly, the new data can change the way researchers think about behavior. For example, they enable the capturing of information on a subject's entire environment, offering the potential to understand the effects of complex environmental inputs on human behavior. In addition, some of the Big Data sources enable researchers to study the tails of a distribution in a way not possible with small data (assuming that the data sources do not suffer from self-selection). The tails of the distribution are often the more interesting and hardest to reach parts of the population being studied; consider healthcare costs for small numbers of ill people (Stanton 2006), or economic activity and employment by a small number of firms (Jovanovic 1982; Evans 1987).

The Big Data Process and Data-Quality Challenges

The massive amounts of very high-dimensional and unstructured data in Big Data bring both new opportunities and new challenges to the data analyst. Many of the problems with Big Data are well known, with some highlighted previously. Big Data is often selective, incomplete, and erroneous. New errors can be introduced downstream.

Big Data are typically aggregated from disparate sources at various points in time and integrated to form data sets. These processes involve linking records together, transforming them to form new variables, documenting the actions

taken, and interpreting the newly created features of the data. These activities also introduce errors that may be variable, creating noise and poor reliability, or systematic, leading to bias and invalidity. Thus, using Big Data in statistically valid ways is increasingly challenging, yet exceedingly important for quality inference.

The core issue confronting Big Data veracity is that these data are not generated from instruments and methods designed to produce valid and reliable data amenable to scientific analysis. Rather, as discussed earlier, these found data are often byproducts, also sometimes called *data exhaust*, from processes whose primary purposes do not always align with those of data analysts. There is also a risk of mischief; for example, automated systems can be written to generate content. Consequently, Big Data generators often have little or no regard for the quality of the data flowing from their processes. Therefore, it is the responsibility of Big Data analysts to be keenly aware of the data's many limitations and to take the necessary steps to limit the effects of Big Data error on their results.

A well-known example of the risks of Big Data error is provided by the Google Flu Trends series, which uses Google searches on flu symptoms, remedies, and other related keywords to provide "near-real-time" estimates of flu activity in the United States and 24 other countries worldwide. Compared to CDC data, the Google Flu Trends provided remarkably accurate indicators of flu incidence in the United States between 2009 and 2011. However, for the 2012–2013 flu seasons, Google Flu Trends predicted more than double the proportion of doctor visits for flu-like symptoms than the CDC (Butler 2013). Lazer et al. (2014) cite two causes of this error: Big Data hubris and algorithm dynamics. The former occurs when the Big Data researcher believes that the volume of the data compensates for any of their deficiencies, thus obviating the need for traditional, scientific analytic approaches. As Lazer et al. (2014, 1203) note, Big Data hubris fails to recognize that "quantity of data does not mean that one can ignore foundational issues of measurement and construct validity and reliability."

Although explanations vary, the fact remains that Google Flu Trends was too high, and by considerable margins for 100 out of 108 weeks starting in July 2012. Lazer et al. (2014) also blame "blue team dynamics," which occurs when the data-generating engine is modified in such a way that the formerly highly predictive search terms eventually failed to work. For example, when a Google user searched on "fever" or "cough," Google's other programs started recommending searches for flu symptoms and treatments—the very search terms the algorithm used to predict flu. Thus, flu-related searches artificially spiked as a result of these changes to the algorithm. In survey research, this is similar to the bias induced by interviewers who suggest to respondents who are coughing that they might have the flu, then ask the same respondents if they think they might have the flu.

Algorithm dynamic issues are not limited to Google. Platforms such as Twitter and Facebook are also frequently being modified to improve the user experience. A key lesson provided by Google Flu Trends is that successful analyses using Big Data today may fail to produce good results tomorrow. All these platforms change their algorithms more or less frequently, with ambiguous results for any kind of long-term study. Recommendation engines often exacerbate effects in a certain direction, but these effects are hard to tease out. Furthermore, other sources of error may affect Google Flu Trends to an unknown extent. For example, selectivity may be an important issue because the demographics of people with Internet access are quite different from the demographic characteristics related to flu incidence (see, for example, Thompson, Comanor, and Shay [2006]). This means that the "at risk" population for influenza and the implied population based on Google Internet searches do not correspond. This illustrates just one type of representativeness issue that often plagues Big Data analysis. In general, it is an issue that algorithms are not (publicly) measured for accuracy, since they are often proprietary. Google Flu is special in that it publicly failed. From what we have seen, most models fail privately, and often without anyone at all noticing.

Data deficiencies represent only one set of challenges for the Big Data analyst. Other challenges arise solely as a result of the massive size and dimensionality of the data. Fan, Han, and Liu (2014) identify three types of issues they refer to as (1) noise accumulation; (2) spurious correlations; and (3) incidental endogeneity. These issues should concern Big Data analysts even if the data could be regarded as error free. Nonsampling errors would only exacerbate these problems.

To illustrate noise accumulation (1), suppose an analyst is interested in classifying individuals into two categories—C1 and C2—based upon the values of 1,000 features (or variables) in a Big Data set. Suppose further that, unknown to the researcher, the mean value for persons in C1 is 0 on all 1,000 features while persons in C2 have a mean of 3 on the first 10 features and a value of 0 on the other 990 features. A classification rule based upon the first $m \leq 10$ features performs quite well, with little classification error. However, as more and more features are included in the rule, classification error increases because the uninformative features (i.e., the 990 features having no discriminating power) eventually overwhelm the informative signals (i.e., the first 10 features). In the Fan, Han, and Liu (2014) example, when $m > 200$, the accumulated noise exceeds the signal embedded in the first 10 features and the classification rule becomes equivalent to a coin-flip classification rule.

High dimensionality can also introduce spurious correlations (2) in that many unrelated features may be highly correlated simply by chance, resulting in false discoveries and erroneous inferences. For example, using simulated populations and relatively small sample sizes, Fan, Han, and Liu (2014) show that with 800 independent features, the analyst has a 50 percent chance of

observing an absolute correlation that exceeds 0.4. Their results suggest that there are considerable risks of false inference associated with a purely empirical approach to predictive analytics using high-dimensional data.

Finally, (3), a key assumption in regression analysis is that the model covariates are uncorrelated with the residual error. Endogeneity refers to a violation of this assumption. For high-dimensional models, this can occur purely by chance—a phenomenon Fan and Liao (2014) call "incidental endogeneity." Incidental endogeneity leads to the modeling of spurious variation in the outcome variables, resulting in errors in the model selection process and biases in the model predictions. The risks of incidental endogeneity increase as the number of variables in the model selection process grows large. Thus, it is a particularly important concern for Big Data analytics.

Fan, Han, and Liu (2014) as well as a number of other authors (see, for example, Stock and Watson [2002], Fan, Samworth, and Wu [2009], Hall and Miller [2009], Fan and Liao [2014]) suggest robust statistical methods aimed at mitigating the risks of (1)–(3). However, as previously noted, these issues and more are further compounded when nonsampling errors are introduced into the data. Biemer and Trewin (1997) show that nonsampling errors will bias the results of traditional data analysis and inflate the variance of estimates in ways that are difficult to evaluated or mitigate in the analysis process. Thus, the massiveness and high dimensionality of Big Data combined with the risks of variable and systematic errors require new, robust approaches to data analysis.

A TOTAL ERROR FRAMEWORK FOR BIG DATA

Dealing with the risks that nonsampling errors introduce in Big Data analysis can be facilitated through a better understanding of the sources and nature of the errors. Such knowledge is gained through in-depth knowledge of the data-generating mechanism, the data-processing infrastructure, and the approaches used to create a specific data set or the estimates derived from it. For survey data, this knowledge is embodied in a "total survey error (TSE)" framework that identifies all the major sources of error contributing to data validity and estimator accuracy (see, for example, Biemer [2010]). The TSE framework also attempts to describe the nature of the error sources and what they may suggest about how the errors could affect inference. The framework parses the total error into bias and variance components which, in turn, may be further subdivided into subcomponents that map the specific types of errors to unique components of the total mean squared error. It should be noted, that while our discussion of issues regarding inference has quantitative analyses in mind, some of the issues discussed here are also of interest to more qualitative uses of Big Data.

For surveys, the TSE framework provides useful insights regarding how the many steps in the data-generating and -preparation processes affect estimation

and inference and may also suggest methods for either reducing the errors at their source or adjusting for their effects in the final data products to produce inferences of higher quality. We believe that a Total Error framework is needed for Big Data (Biemer 2014). In this section, we offer a skeletal view of the framework for a Total Error approach for Big Data. We suggest an approach closely modeled after the TSE framework since, as we will see, a number of error sources are common to both. However, the Big Data Total Error (BDTE) framework necessarily will include additional error sources that are unique to Big Data and can create substantial biases and uncertainties in Big Data products. Like the TSE framework, the BDTE framework will aid in our understanding the limitations of the data, leading to better-informed analyses and applications of the results. It may also inform a research agenda for reducing the effects of error on Big Data analytics.

A typical survey data set is shown in figure 3 as a matrix consisting of some number of rows and columns. Data sets derived from Big Data may also be represented in this way and, thus, will share many of the same properties. In surveys, the rows may be sample or population elements, the columns may be the characteristics of the row elements, and the cells contain values of the characteristics for each element. The total error for this data set may be expressed by the following heuristic formula:

$$\text{Total error} = \text{Row error} + \text{Column error} + \text{Cell error}.$$

Row errors may be of three types; namely, *omissions*, where some population elements are not among the rows, *duplications*, where some population elements occupy more than one row, and *erroneous inclusions*, where some rows contain elements or entities that are not part of the population of interest.

For survey sample data sets, omissions include nonsampled elements in the population as well as population members deliberately excluded from the sampling frame. For Big Data, selectivity is a common form of omissions. For example, a data set consisting of persons who conducted a Google search in the past week necessarily excludes persons not satisfying that criterion. Unlike survey sampling, this is a form of nonrandom selectivity. For example, persons who do not have access to the Internet are excluded from the file. This exclusion may be biasing, in that persons with Internet access may have very different demographic characteristics than persons who do not have Internet

Record #	V_1	V_2	....	V_K
1				
2				
...				
N				

Figure 3. A Typical Rectangular Format for Traditional Data Analysis.

access. This problem is akin to non-coverage in sampling, depending on the population about which the researcher is attempting to estimate.

We can also expect that Big Data sets, such as a data set containing Google searches during the previous week, could have the same person represented many times. People who conducted many searches during that period would be disproportionately represented relative to those who conducted fewer searches. Other erroneous inclusions can occur when the entity conducting a search is not a person but another computer; for instance, via a web-scraping routine.

The most common type of column error is caused by inaccurate or erroneous labeling of the column data—an example of metadata error. For example, a business register may include a column labeled "number of employees," defined as the number of persons in the company that received a payroll check in the month preceding. Instead, the column contains the number of persons on the payroll whether they received a check last month or not, including persons on leave without pay. Such errors would seem to be quite common in Big Data analysis given the multiple layers of processing required to produce a data set. For example, data generated from a source, such as an individual Tweet, may undergo a number of transformations before it lands in a rectangular file such as the one in figure 3. This transformation process can be quite complex; for example, it may involve parsing phrases, identifying words, and classifying them as to subject matter and then further as to positive or negative expressions about the economy. There is considerable risk that the resulting features are either inaccurately defined or misinterpreted by the data analyst.

Finally, cell errors can be of three types: content error, specification error, or missing data. A content error occurs when the value in a cell satisfies the column definition but is still erroneous. For example, value satisfies the definition of "number of employees," but the value does not agree with the true number of employees for the company. Content errors may be the result of a measurement error, a data-processing error (e.g., keying, coding, editing, etc.), an imputation error, or some other cause. A specification error is just as described for the column error but applied to a cell. For example, the column is correctly defined and labeled; however, a few companies provided values that, although otherwise highly accurate, were nevertheless inconsistent with the required definition. Missing data, as the name implies, is just an empty cell that should be filled. As described in Kreuter and Peng (2014), data sets derived from Big Data are notoriously affected by all three types of cell errors, particularly missing or incomplete data.

EXTENDING THE FRAMEWORK FOR BIG DATA

The traditional TSE framework is quite general in that it can be applied to essentially any data set that conforms to the format in figure 3. However, in most practical situations it is quite limited because it makes no attempt to

describe the error in the processes that generated the data. In some cases, these processes constitute a "black box" and the best approach is to attempt to evaluate the quality of the end product. For survey data, the TSE framework provides a fairly complete description of the error-generating processes for survey data and survey frames (see, for example, Biemer [2010]). In addition, there has been some effort to describe these processes for population registers and administrative data (Wallgren and Wallgren 2007). But at this writing, very little effort has been devoted to enumerating the error sources and the error-generating processes for Big Data. One obstacle in this endeavor is that the processes involved in generating Big Data are as varied as Big Data are themselves. Nevertheless, some progress can be made by considering the generic steps involved.

In the *generate* step, data are generated from some source either incidentally or purposively. In the *extract/transform/load (ETL)* step, all data are brought together under a homogeneous computing environment in three stages. These stages are the *extract stage*, where data are harvested from their sources, parsed, validated, curated, and stored; the t*ransform stage*, where data are translated, coded, recoded, aggregated/disaggregated, and/or edited; and the *load stage*, where data are integrated and stored in the data warehouse.

In the last step, *analyze*, data are converted to information through a process involving two stages. The first stage is the *filtering (sampling)/reduction stage*, where unwanted features and content are deleted; features may be combined to produce new ones; and data elements may be thinned or sampled to be more manageable. The second stage is the *computation/analysis/visualization* stage, where data are analyzed and/or presented for interpretation and information extraction.

Figure 4 graphically depicts the flow of data along these steps. The severity of the errors that arise from these processes will depend on the specific data sources and analytic goals involved. Nevertheless, we can still consider how each stage might create errors in a more generic fashion.

For example, data-generation error is somewhat analogous to errors arising in survey data collection. Like surveys, the generic data-generating process for

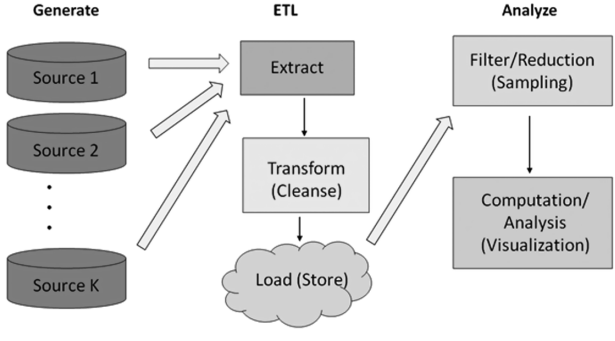

Figure 4. Big Data Process Map (graph created by Paul Biemer).

Big Data can create erroneous and incomplete data. In addition, the data-generating sources may be selective in that the data collected may not represent a well-defined population or one that is representative of a target population of interest. Thus, data-generation errors include low signal/noise ratio, lost signals, incomplete or missing values, non-random, selective sources, and meta-data that are lacking, absent, or erroneous.

ETL processes may be quite similar to various data-processing stages for surveys. These may include creating or enhancing meta-data, record matching, variable coding, editing, data munging (or scrubbing), and data integration (i.e., linking and merging records and files across disparate systems). ETL errors include specification error (including errors in meta-data), matching error, coding error, editing error, data-munging errors, and data-integration errors.

As noted above, the analysis of Big Data introduces risks for noise accumulation, spurious correlations, and incidental endogeneity, which may be compounded by sampling and nonsampling errors. Related to the former, data may be filtered, sampled, or otherwise reduced to form more manageable or representative data sets. These processes may involve further transformations of the data. Errors include sampling errors, selectivity errors (or lack of representativeness), and modeling errors.

Other errors that may be introduced in the computation stage are similar to estimation and modeling error in surveys. These include modeling errors, inadequate or erroneous adjustments for representativeness, improper or erroneous weighting, and computation and algorithmic errors.

Previously, we mentioned that all data collection suffers from error in the data-generating process. AAPOR is promoting the transparency of these processes. A similar effort will be very valuable for Big Data–driven research.

What Are the Policy, Technical, and Technology Challenges, and How Can We Deal with Them?

Public opinion research is entering a new era, one in which traditional survey research may play a less dominant role. The proliferation of new technologies, such as mobile devices and social media platforms, is changing the societal landscape across which public opinion researchers operate. As these technologies expand, so does access to users' thoughts, feelings, and actions expressed instantaneously, organically, and often publicly across the platforms they use. The ways in which people both access and share information about opinions, attitudes, and behaviors have gone through perhaps a greater transformation in the past decade than in any previous point in history, and this trend appears likely to continue. The ubiquity of social media and the opinions users express on social media provide researchers with new data-collection tools and alternative sources of qualitative and quantitative information to augment or, in some cases, provide alternatives to more traditional data-collection methods.

There is great potential for Big Data to generate innovation in public opinion research. While traditional survey research has a very important role, the addition of large-scale observations from numerous sources (e.g., social media, mobile computing devices) promises to bring new opportunities. To realize these potential advances, we must address numerous challenges in a systematic way. This section examines several policy challenges for Big Data (ownership, stewardship, collection authority, privacy protection), technical challenges (multidisciplinary skills required), as well as technology challenges (computing resources required).

POLICY CHALLENGE: DATA OWNERSHIP

Many individuals now produce data that are potentially useful for research as part of their everyday participation in the digital world. There has always been a lack of clarity in legal guidance stemming from a lack of clarity as to who owns the data—whether it is the person who is the subject of the information; the person or organization who collects that data (the data custodian); the person who compiles, analyzes, or otherwise adds value to the information; the person who purchases interest in the data; or society at large. The lack of clarity is exacerbated because some laws treat data as property and some treat it as information (Cecil and Eden 2003). The new types of data make the ownership rules even more unclear: data are no longer housed in statistical agencies, with well-defined rules of conduct, but are housed in businesses or administrative agencies. In addition, since digital data can be alive forever, ownership could be claimed by yet-to-be-born relatives whose personal privacy could be threatened by release of information about blood relations. For the AAPOR community, it will be important to stay informed about emerging rules and to be aware of differences in regulations across countries.

POLICY CHALLENGE: DATA STEWARDSHIP

An eloquent description of statistical confidentiality is "the stewardship of data to be used for statistical purposes" (Duncan, Elliot, and Salazar-Gonzalez 2011). Statistical agencies have been at the forefront of developing that stewardship community in a number of ways. First, on-the-job training is provided to statistical agency employees. Second, in the United States, academic programs such as the Joint Program on Survey Methodology, communities such as the Federal Committee on Statistical Methodology, and resources such as the Committee on National Statistics have been largely supported by the federal statistical community. In the past, the focus was almost exclusively on developing methodologies to improve the analytical use of survey data, and to a lesser extent, administrative data. It is important to expand efforts to train scientists in developing an understanding of such issues, such as identifying the relevant population and linkage methodologies. Around the United States,

several programs are emerging. However, it is important to integrate the training of these skills into the existing programs, in particular if the field is moving toward data integration from survey and non-survey data.

POLICY CHALLENGE: DATA-COLLECTION AUTHORITY

When statistical agencies were the main collectors of data, they did so under very clear statutory authority with statutory protections. For example, Title 26 (Internal Revenue Service) and Title 13 (Census Bureau) of the US Code provided penalties for breaches of confidentiality, and agencies developed researcher access modalities in accordance with their statutory authorization.

The statutory authorization for the new technology-enabled collection of data is less clear. The Fourth Amendment to the Constitution, for example, constrains the government's power to "search" the citizenry's "persons, houses, papers, and effects." State privacy torts create liability for "intrusion upon seclusion." Yet, the generation of Big Data often takes place in the open, or through commercial transactions with a business, and hence is not covered by either of these frameworks. There are major questions as to what is reasonably private, and what constitutes unwarranted intrusion (Strandburg 2014). Data generated by interacting with professionals, such as lawyers and doctors, or by online consumer transactions, are governed by laws requiring "informed consent" and draw on the Fair Information Practice Principles (FIPP). Despite the FIPP's explicit application to "data," they are typically confined to personal information, and do not address the large-scale data-collection issues that arise through location tracking and smart grid data (Strandburg 2014).

POLICY CHALLENGE: PRIVACY AND RE-IDENTIFICATION

The risk of re-identifying individuals in a micro data set is intuitively obvious. Indeed, one way to formally measure the re-identification risk associated with a particular file is to measure the likelihood that a record can be matched to a master file (Winkler 2005). If the data include direct identifiers, like names, Social Security numbers, and establishment ID numbers, the risk is quite high. However, even access to close identifiers, such as physical addresses and IP addresses, can be problematic. Indeed, the Health Insurance Portability and Accountability Act (HIPAA) regulations under the Privacy Rule of 2003 require the removal of 18 different types of identifiers, including other less obvious identifiers such as birthdate, vehicle serial numbers, URLs, and voice prints. However, even seemingly innocuous information makes it relatively straightforward to re-identify individuals, for example by finding a record with sufficient information such that there is only one person in the relevant population with that set of characteristics: the risk of re-identification has been increasing due to the growing public availability of identified data and rapid advances in the technology of linking files (Dwork 2011). With many

variables, everyone is a population unique. Since Big Data have wide-ranging coverage, one cannot rely on protection from sampling (Karr and Reiter 2014). Indeed, as Ohm (2010) points out, a person with knowledge of an individual's zip code, birthdate, and sex can re-identify more than 80 percent of Netflix users, yet none of those are typically classified as Personally Identifiable Information (PII).

POLICY CHALLENGE: MEANING OF "REASONABLE MEANS" NOT SUFFICIENTLY DEFINED

The statutory constraint on agencies such as the IRS and the US Census Bureau makes it clear that the agencies, as data producers, should take "reasonable means" to protect data, although these reasonable means are not defined. Trust clearly depends on people's views on privacy, but these views are changing rapidly (Nissenbaum 2011). Nissenbaum (2011, 34) also notes that it is increasingly difficult for many people to understand where the old norms end and new ones begin, as "Default constraints on streams of information from us and about us seem to respond not to social, ethical, and political logic but to the logic of technical possibility: that is, whatever the Net allows." Yet, there is some evidence that people do not require complete protection, and will gladly share even private information provided that certain social norms are met, similar to what Gerber reported in 2001. There are three factors that affect these norms: actors (the information senders and recipients or providers and users); attributes (especially types of information about the providers, including how these might be transformed or linked); and transmission principles (the constraints underlying the information flows).

WHAT WE CAN LEARN FROM CURRENT KNOWLEDGE

Kinney, Karr, and Gonzalez (2009) identify a variety of mechanisms for interaction between users and confidential data. As they note, in figure 5 (above), "there are three major forms of interaction: direct access, dissemination-based access (public data releases), and query-based access. Direct access imposes the least interference between the users and the confidential data. Dissemination-based access refers to the practice of releasing masked data in public files. In the query-based interaction mode, users cannot directly access individual data records, but are able to submit queries, either electronically or manually" (Kinney, Karr, and Gonzalez 2009, 127). Thorough reviews of different approaches are provided in Duncan, Elliot, and Salazar-Gonzalez (2011) and Prada et al. (2011).

The current statistical disclosure literature offers multiple ways of permitting access to microdata, but less relevant guidance about release.

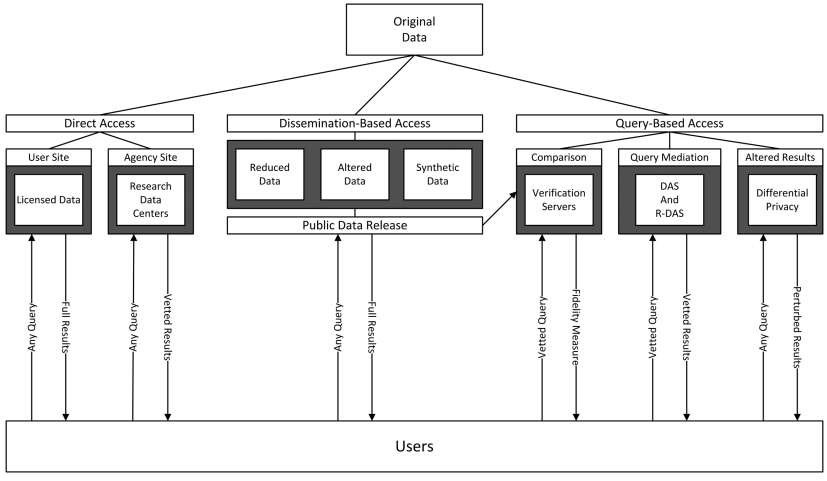

Figure 5. Models for User-Data Interaction, from Kinney, Karr, and Gonzalez (2009).

TECHNICAL CHALLENGE: SKILLS REQUIRED TO INTEGRATE BIG DATA INTO OPINION RESEARCH

Depending on the scale of the data being discussed, there can be significant challenges in terms of the skills and resources necessary to work with Big Data. In particular, most Big Data problems require a minimum of four roles: a *domain expert*, a *researcher*, a *computer scientist*, and a *system administrator* (see figure 6). The domain expert is a user, analyst, or leader with deep subject-matter expertise related to the data, their appropriate use, and their limitations. The researcher is a team member with experience applying formal research methods, including survey methodology and statistics. The computer scientist is a technically skilled team member with education in computer programming and data-processing technologies, and the system administrator is a team member responsible for defining and maintaining a computation infrastructure that enables large-scale computation. However, from our experience, many companies are trying to make do with only one person.

Domain expertise is particularly important with new types of data that have been collected without instrumentation, usually for purposes other than quantitative survey analysis. For example, looking at Big Data from social media sources requires an in-depth understanding of the technical affordances and user behaviors of that social media source. Posting to Twitter, as an example, involves norms and practices that could affect the interpretation of data from that source. This could refer to the use of handles and hashtags, certain terminology and acronyms used, or practices such as retweeting, modifying tweets, and favoriting. Additionally, it is important to understand to what degree

different forms of new media may underrepresent particular demographics (e.g., there may be a low number of citizens age 60 years and older using Twitter to express themselves).

Foundational research skills such as the application of classical survey methodology and the appropriate use of descriptive statistics remain critical for understanding Big Data. As the volume of digital data grows and the barrier to obtaining such data is continually lowered, there is an increasing risk of untrained engineers and computer programmers finding bogus associations in Big Data. To ensure that Big Data are appropriately integrated into public opinion research, there remains an ongoing requirement for classically trained researchers to be involved throughout the entire process.

From the *computer science skills* standpoint, baseline competencies can include the ability to work in command line environments, some capability with programming languages, facility with databases and database languages, and experience with advanced analytical tools. The larger the data set, the more important skills in databases and analytics become. Some researchers choose to partner with computer scientists, or skilled programmers, to cover these needed skills. While this has led to viable research partnerships, it creates a new need in terms of interdisciplinary collaboration. Major information technology components that are frequently used in the process of collecting, storing, and analyzing Big Data include *Apache Hadoop*, *Apache Spark*, *Java*, and *Python programming language*. *Apache Hadoop* is a system for maintaining a distributed file system that supports the storage of large-scale data (terabytes or petabytes of content), and the parallel processing of algorithms against large data collections. *Apache Spark* is a fast general-purpose engine for large-scale data processing that works in support of Hadoop or in-memory databases. *Java programming language* is a general purpose systems engineering language that supports the creation of efficient algorithms for data analysis, and *Python programming language* is a general-purpose systems engineering language that supports rapid prototyping and efficient algorithms for data analysis. Both Apache Hadoop and Apache Spark require a programming language such as Java or Python.

It is worth noticing that there are many different frameworks. Even though a framework such as, for example, Hadoop is commonly used today, given the fast development in this area, this may very well change soon. It could therefore be helpful to think in more general terms of clusters and parallel processing of unstructured data.

System administrators play an important role in defining, creating, and maintaining computing environments for the storage and analysis of Big Data. Working with Big Data often requires additional computing resources. Depending on the size of the data being considered, resources can range from hardware and server stacks that are manageable by non-specialist IT staff to very large-scale computing environments that include

high-powered computing stacks of hardware and software that often require specialist IT training. As an example, many universities offer High Performance Computing Centers (HPC) that include networked servers, structuring software like Hadoop, as well as database and analysis packages. System administrators responsible for maintaining Big Data computer platforms often use one of three strategies. For long-term storage of unique or sensitive data, it often makes sense to create and maintain an Apache Hadoop cluster using a series of networked servers within the internal network of an organization. Although expensive in the short term, this *internal compute cluster* strategy is often the lowest cost in the long term. There is a trend across the IT industry to outsource elements of infrastructure to "utility computing" service providers. Organizations such as the Amazon Web Services (AWS) division of Amazon.com make it simple for system administrators to rent prebuilt Apache Hadoop clusters and data-storage systems (see figure 7). This *external compute cluster* strategy is very simple to set up, but may be much more expensive than creating a long-standing cluster internally. Functional equivalents to Amazon Elastic Map Reduce Service are Microsoft HDInsight and Rackspace's Cloud Big Data Platform. Other alternatives include Hadoop on Google's Cloud Platform and Qubole. A common *hybrid compute cluster* strategy is to provision external compute cluster resources using services such as AWS for on-demand Big Data analysis tasks, and to create a modest internal computer cluster for long-term data storage.

TECHNOLOGY CHALLENGE: COMPUTATIONAL REQUIREMENTS

The formula "distance = rate x time" is well known by high school math students. This formula may be applied to simplify the understanding of why large-scale parallel processing computer clusters are a requirement for Big Data analysis. In the analysis of a very large data set, the volume of data to be processed may be considered the *distance* (e.g., 10 terabytes). Similarly, the number of available central processing units and magnetic hard drives for storing the media may be considered directly related to the *rate*.

All other factors being held equal, a system with 10 CPUs and 10 hard drives (10 computation units) will process a batch of data 10 times faster than a system with one CPU and one hard drive (1 computation unit). If an imaginary data set consists of 50 million records, and systems with 1 computation unit can process 100 records per second, then it will take approximately 5.7 days (50,000,000 records/100 records per second) to finish the analysis of data—potentially an unacceptable amount of time to wait. A system with 10 computation units can compute the same result in just 13.9 hours, a significant time savings. Systems like Apache Hadoop drastically simplify the process of connecting multiple commodity computers

into a cluster capable of supporting such parallel computations. (For many simple cases, parallel computation is not required—millions of records can often be processed on modern computers using scripting languages like R or Python.)

Although disk space may be relatively inexpensive, the cost of creating and maintaining systems for Big Data analysis can be quite expensive. In the past 30 years, the cost of storing data on magnetic storage media such as hard drives has decreased dramatically. A hard drive with 3 terabytes of storage capacity now costs less than $100 in the United States. However, the total cost of ownership of a Big Data analysis system is the sum of several components, including, at a minimum, the cost of *disk-based storage media*, cost of *active computation components* (computer central processing unit or CPU, Random Access Memory or RAM), and cost of *infrastructure elements* such as server farm rack space, electricity required, cooling costs, and network access and security fees.

When taken in aggregate, these components may cost tens or hundreds of thousands of dollars. It may not be feasible to create a permanent Big Data computer cluster to support a single study. Within AAPOR, there is the possibility to form public-private sector partnerships not only for sharing data but also for sharing analysis infrastructure.

How Can Big Data Be Used to Gain Insights?

The recent literature on developments in Big Data can give the reader the impression that there is an ongoing, head-to-head competition between traditional research based on data specifically designed to support research and new research methods based on more organic data or found data. Researchers who have created a career around the analysis of survey data are particularly anxious about the rise of Big Data, fearful that the skills they have developed throughout their career may become obsolete as Big Data begins to crowd out survey data in supporting future research.

We have seen similar debates on statistical methods. The predominant theory used in surveys emanates from the Neyman-Pearson framework. This theory states that survey samples are generated from a repeatable random process and governed by underlying parameters that are fixed under this repeatable process. This view is called the frequentist view and is what most survey researchers are most familiar with. An alternative theory is the Bayesian view that emanates from Bayes, Savage, deFinetti, and others. In this theory, data from a realized sample are considered fixed while the parameters are unknown and described probabilistically. Typically, a prior distribution of the parameter is combined with the observed data, resulting in a posterior distribution. The discussions of these views have successively moved from controversy to more pragmatic standpoints. A survey statistician's job is to make the most

valid inferences about the finite population, and therefore there is room for both views. Both frequentist and Bayesian statistics play key roles in Big Data analysis. For example, when data sets are so large that the analysis must be distributed across multiple machines, Bayesian statistics provides efficient algorithms for combining the results of these analyses (see, for example, Ibrahim and Chen [2000]; Scott et al. [2013]). Sampling techniques are key in gathering Big Data and for analyzing Big Data in a small computing environment (Leek 2014a, 2014b).

In general, framing the rise of Big Data as a competition with survey data or traditional research is counterproductive, and a preferred route is to recognize how research is enhanced by utilizing all forms of data, including Big Data as well as data that are designed with research in mind. Inevitably, the increased availability of the various forms of Big Data will supplant survey data in some settings. However, both Big Data and survey data have advantages and disadvantages, which we describe in more detail below. An effective and efficient research strategy will be responsive to how these advantages and disadvantages play out in different settings, and deploying blended research methods that maximize the ability to develop rigorous evidence for the questions of interest for an appropriate investment of resources.

Research is about answering questions, and the best way to answer questions is to start by utilizing all of the information that is available. The availability of Big Data to support research provides a new way to approach old questions as well as an ability to address some new questions that in the past were out of reach. However, the findings that are generated based on Big Data inevitably generate more questions, and some of those questions tend to be best addressed by traditional survey research. As the availability and use of Big Data increases, there is likely to be a parallel growth in the demand for survey research to address questions raised by findings from Big Data. The availability of Big Data liberates survey research, in the sense that researchers no longer need to generate a new survey to support each new research endeavor. Big Data can be used to generate a steady flow of information about what is happening—for example, how customers behave—while traditional research can focus instead on deeper questions about why we are observing certain trends or deviations from trends—for example, why customers behave as they do and what can be done to change their behavior.

In thinking about how to blend Big Data with traditional research methods, it is important to be clear about the relevant questions to be addressed. Big Data can be especially useful for detecting patterns in data or for establishing correlations between factors. In contrast, establishing causality between variables requires that data be collected according to a specific design in order to support models or research designs intended to isolate causality. Marketing researchers use Big Data for so-called A/B testing to establish causality,

though even this can be problematic, for example, since it relies on cookies. In the public sector, traditional research based on designed data is likely to continue to play a primary role in supporting policy development, particularly when customized data and research designs are necessary to ensure that we can identify causality between variations in public interventions and the outcomes that they affect. At the same time, research based on Big Data can be best utilized to meet the needs of program administrators, who are focused on monitoring, maintaining, and improving program operations within an ongoing policy regime. In this setting, measuring trends and correlations and making predictions may be sufficient in many cases—isolating causality is not essential—and the administrative data and related Big Data sources can best meet these needs. However, when causation is ignored and the focus is on predictions using models that are based on historical training data, there is a risk to perpetuate what happened in the past; for example, embedding racism, sexism, or other problematic patterns in the models.

RELATIVE ADVANTAGES OF SURVEY DATA AND BIG DATA TO SUPPORT RESEARCH

For many years, research has depended on data collected through surveys because there have been few alternatives. Even as alternative sources of data begin to proliferate, survey data retain some critical advantages in facilitating social science research. The primary advantage of basing research on survey data is the control it provides for researchers—the survey can be designed specifically to support the needs of the research. Use of a survey allows for customizing outcome measures to closely match the primary questions to be addressed by the research. For example, if a research project is designed to address hourly wage compensation as a key outcome of interest, the supporting survey can be designed to measure hourly compensation rather than use a proxy or impute hourly compensation from some pre-existing data source.

The control afforded by using a survey to support research also allows for generating estimates for samples that are representative of a specific population of interest. By using a specific population to create a probabilistic sample frame for a survey, researchers can use data from the survey sample to generate estimates that apply to the population with a known degree of precision. Researchers have fully developed the theory and practice of probability sampling and statistical inference to handle just this type of data collection and use these data effectively in addressing questions of interest.

In contrast, since most Big Data sources are organic and beyond the control of researchers, researchers using Big Data sources take what they get in terms of the population that is represented by the data. In many cases, the population represented by a Big Data source does not exactly match the population of interest. For example, databases based on Google searches are constrained to represent the searches conducted by Google users rather than the general

population or some other population of interest. It is difficult to assess the degree to which this may bias estimates relative to a given research question. Research on television audience measurement and viewing habits in the UK offers a choice between research based on a 5,100-household sample that is representative of the UK population, compiled by the Broadcasters' Audience Research Board (BARB), and research based on the SkyView 33,000-household panel, developed by Sky Media based on Sky Digital Homes (homes that subscribe to this particular service). While the SkyView panel is considerably larger than the BARB panel, the BARB panel can be used to generate estimates that are directly representative of the UK population. Another example of this is that sometimes people want to estimate TV viewership by twitter feeds. The problem is that people never tweet which news channel told them some piece of news, just the news itself, whereas people tweet which show they're watching if it's a drama like *House of Cards*. In this case, TV news will be underreported by Twitter analysis.

Regardless, Big Data have a number of advantages when compared with survey data. The clearest advantage of Big Data is that these data already exist in some form, and therefore research based on Big Data does not require a new primary data-collection effort. Primary data collection can be expensive and slow, which can either greatly delay the generation of new research findings or make new research prohibitively expensive. Problems may also arise with survey data collection as response rates trend down, particularly in research settings that would require lengthy surveys.

Compared with survey data, Big Data usually require less effort and time to collect and prepare for analysis. However, the effort associated with the creation and preparation of a Big Data set for analysis is not trivial. Even though Big Data already exist, it may still require substantial effort to collect the data and link digital data from various sources. According to expert estimates, data scientists spend 50 to 80 percent of their time collecting and preparing data to make it ready for investigation (Lohr 2014). The task of analyzing Big Data often involves gathering data from various sources, and these data—including data from sensors, documents, the web, and more conventional data sets—come in different formats. Consequently, startups are developing software to automate the gathering, cleaning, and organizing of data from different sources, so as to liberate data scientists from what tend to be the more mundane tasks associated with data preparation. There will, however, always be this type of routine work because you need to massage data one way for one study and another way for the next.

Big Data also are often available in high volumes, and with current technology, these high volumes of data are more easily processed, stored, and examined than in the past. For years, researchers have worked with data sets of hundreds or thousands of observations, which are organized in a relatively straightforward rectangular structure, with n observations and k variables. While these data sets are straightforward to deal with, the constrained volume

of the data created limitations with respect to statistical power. In contrast, Big Data come in many different forms and structures, and the potential for huge volumes of observations implies that statistical power is less of a concern than in the old days. As mentioned previously, however, huge volumes of data cause their own sets of problems. The varied structure (or lack of structure) and large volumes of observations in Big Data can be a challenge for processing and organizing the data, but the volume of observations in Big Data also translates into a more comprehensive and granular picture of the processes that are represented by the data. More granular and comprehensive data can help pose new sorts of questions and enable novel research designs that can inform us about the consequences of different economic policies and events. Finally, enhanced granularity allows researchers to examine behavior in greater detail, and also to examine much more detailed subgroups of the population with adequate statistical power. For example, traditional research may identify the impact of class size on student performance, but Big Data could allow us to investigate how it varies by grade, school, teacher, or student mix, assuming all other confounders can be removed. With Big Data, it is also possible to study the tails of a distribution, which is not possible with a small data set.

Big Data also are often available in real time, as it is created organically as individual behavior (for example, phone calls, Internet browsing, online shopping, etc.) is occurring. This characteristic of Big Data has made it particularly appealing in the private sector, where businesses can use data to support management decision making in a timely manner. Traditional research, which relies on primary data collection, is slow, and so it generally cannot support making decisions quickly. One analyst characterizes traditional research as being built for "comfort, not speed"—it generates sound findings that can instill confidence in the resulting decisions, but it generates them slowly and therefore cannot support quick decision making. In contrast, the timing of Big Data is more aligned with the cadence of decision making in a private or public sector management setting, where there is a premium on responding quickly to rapid changes in consumer demand or client need.

RESEARCH METHODS THAT EXPLOIT AVAILABILITY OF BIG DATA

As discussed above, Big Data are particularly advantageous in situations where decision makers want to use evidence to drive critical decisions. For a given organization interested in utilizing Big Data analysis to support effective operation of a program or set of programs, one can imagine at least three ways in which this would happen. First, Big Data can be used to match the right people to the right programs. For example, an employer engaged in a health management program to promote better employee health would want to be able to direct employees to the appropriate services given their needs, which

would require collecting, processing and analyzing data on individual health and behaviors. Second, Big Data can be used to facilitate better operations. In the case of an employee health management program, this might amount to using Big Data to support building and facilitating healthful interactions between employees, their interpersonal networks, care providers, and insurers. Third, Big Data can be used to measure the outcomes among participants, the impacts of the program on those outcomes, and the net value of the program. In the case of an employee health management program, this might entail measuring key health and work outcomes, extrapolating to future outcomes, estimating the impact of the program on these outcomes, and monetizing the impact estimates so as to estimate the net value of the program investment. Based on these estimates, managers could make informed decisions on how the program would evolve over time in order to best meet the needs of employees and the employer. Of course, any of these examples carry the risk that the information is used not in the employees' best interest, which gets back to the ethical challenges discussed before.

Given the potential benefit of Big Data in driving evidence-based decisions, private-sector organizations have quickly gravitated toward greater reliance on Big Data and have adopted research methods that exploit the advantages of these data. *Predictive analytics* and *rapid-cycle evaluation* are two of the Big Data–supported research methods that have become much more popular in the private sector in recent years. These methods allow managers to not only track ongoing activity, but also support decision making regarding how to respond tactically to a changing environment and customer base.

Predictive analytics refers to a broad range of methods used to predict an outcome. For example, in the private sector, predictive analytics can be used to anticipate how customers and potential customers will respond to a given change, such as a product or service change, a new marketing effort, establishment of a new outlet, or the introduction of a new product or service. Businesses can use predictive analytics to estimate the likely effect of a given change on productivity, customer satisfaction, and profitability, and thereby avoid costly mistakes. Predictive analytics can be conducted based on data that are collected as part of routine business operations and stored so as to support ongoing analytics, and these data can also be combined with other Big Data sources or survey data drawn from outside the organization.

Predictive analytics modeling also has been used to support new information products and services in recent years. For example, Amazon and Netflix recommendations rely on predictive models of what book or movie an individual might want to purchase. Google's search results and news feed rely on algorithms that predict the relevance of particular web pages or articles. Predictive analytics are also used by companies to profile customers and adjust services accordingly. For example, health insurers use predictive models to

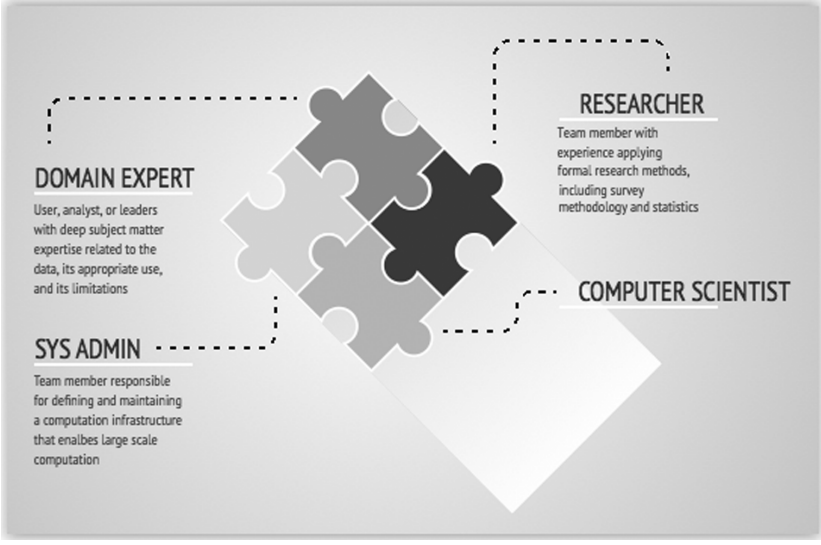

Figure 6. The Different Roles Needed in a Big Data Team (graph created by Abe Usher).

generate "risk scores" for individuals based on their characteristics and health history, which are used as the basis for adjusting payments. Similarly, credit card companies use predictive models of default and repayment to guide their underwriting, pricing, and marketing activities.

Rapid-cycle evaluation is the retrospective counterpart to predictive analytics—it is used to quickly assess the effect of a given change on the outcomes of interest, including productivity, customer satisfaction, and profitability. As with predictive analytics, rapid-cycle evaluation leverages the available operations data as well as other Big Data sources. The exact statistical methods used in rapid-cycle evaluation can vary according to the preferences and resources of the user. For example, rapid-cycle evaluation can be based on experimental methods in which a change is implemented in randomly chosen segments of the business or customers are randomly selected to be exposed to the change. In this way, the evaluation of a given change can be conducted by comparing outcomes among a "treatment group," which is exposed to the change, and a "control group," which is not exposed to the change.

Private businesses have begun to invest heavily in these capabilities. For example, Capital One has been a pioneer in rapid-cycle evaluation based on their transactions data to support business decisions, running more than 60,000 experiments and related analytics addressing a range of questions related to their operations or product offerings. Many other companies are moving in this direction as well (Manzi 2012).

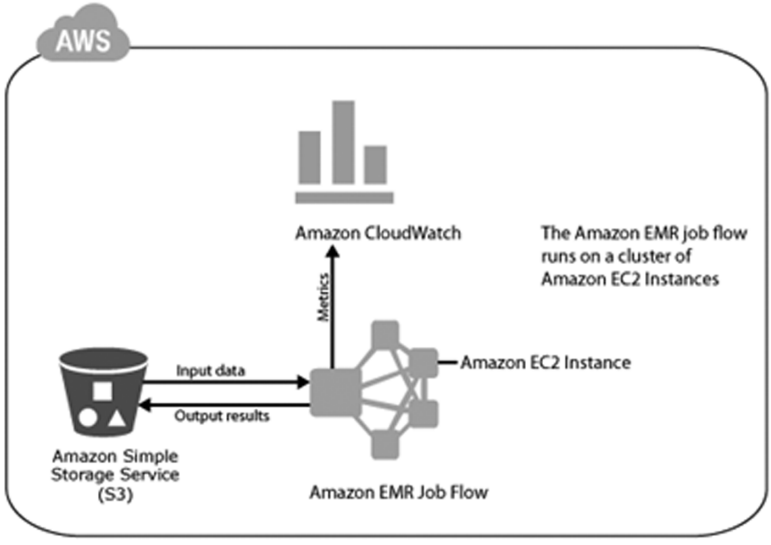

Figure 7. The Amazon Elastic MapReduce (EMR) Service Remains One of the Most Popular Utility Compute Cloud Versions of Hadoop (graph created by Abe Usher).

While the public sector is not moving as fast as the private sector in adopting Big Data and data analytics techniques, public administrators are beginning to appreciate the value of these techniques and experiment with their use in supporting administrative decisions and improving public programs (Cody and Asher 2014). At the broadest level, some government agencies at all levels are collecting available data and examining data patterns related to their operations, in the hope of generating insights. For example, a recent *New York Times* editorial (from August 19, 2014) highlights this trend in New York City by focusing on the ClaimStat initiative, which was begun recently by NYC comptroller Scott Stringer. ClaimStat collects and analyzes data on lawsuits and claims filed each year against the city. By identifying patterns in payouts and trouble-prone agencies and neighborhoods, city managers hope to learn from these patterns and modify operations so as to reduce the frequency and costs of future claims (*New York Times* Editorial Board 2014).

Predictive analytics can be used in the government sector to target services to individuals in need or to anticipate how individuals or a subset of individuals will respond to a given intervention, such as the establishment of a new program or a change in an existing program (Cody and Asher 2014). For example, program administrators can use administrative data and predictive analytics to identify clients who are at risk of an adverse outcome, such as unemployment, fraud, unnecessary hospitalization, mortality, or recidivism.

By knowing which participants are most likely to experience an adverse outcome, program staff can provide targeted interventions to reduce the likelihood that such outcomes will occur or reduce the negative effect of such an outcome.

With information from predictive analytics, administrators may also be able to identify who is likely to benefit from an intervention and identify ways to formulate better interventions. As in the private sector, predictive analytics can exploit the operational data used to support the day-to-day administration of a program, and the analytics may even be embedded directly in the operational data systems to guide real-time decision making. For instance, predictive analytics could be embedded in the intake and eligibility determination systems associated with a given program so as to help frontline caseworkers identify cases that may have eligibility issues or to help customize the service response to meet the specific needs of individuals. In some state unemployment insurance systems, for example, program administrators use statistical models to identify new applicants who are likely to have long unemployment spells and refer the applicants to reemployment services. With any of the predictive models, it is important that ethical and legal requirements are still met, which unfortunately is not always the case (for a discussion of unconstitutional sentencing, see http://bit.ly/1EpKt2j).

COMBINING BIG DATA AND SURVEY DATA

Despite the theoretical and practical advantages of Big Data analysis described above, a preferred strategy is to use a combination of new and traditional data sources to support research, analytics, and decision making, with the precise combination depending on the demands of a given situation. As described in the introduction, traditional research that relies on primary data can be deployed to address the questions that are not adequately or easily addressed using Big Data sources. In many cases, this will entail going beyond the observed trends or behaviors that are easily captured using Big Data to more systematically address questions regarding why those trends or behaviors are occurring. For example, imagine a large advertiser has constant, real-time monitoring of store traffic and sales volume. Traditional research designs, which probe survey panelists on their purchase motivation and point of sales behavior, can help a retailer better target certain shoppers. Alternatively, the analytic design can be expanded to bring in the data on store traffic and sales volume so that these data become the primary monitoring tool, and surveys are utilized to conduct deeper probing based on trends, changes in trends, or anomalies that are detected in the primary monitoring data.

Researchers recently have formulated ideas for blending Big Data with traditional research in the area of market research, which has traditionally

been heavily reliant on data collected through surveys. For example, Duong and Millman (2014) highlight an experiment based on the premise that the use of behavioral data collected online can be used in combination with survey data on brand recognition to enhance learning regarding advertising effectiveness. In their experiment, data collected on users' interactions with a website combined with data from a traditional online survey provided a clearer picture regarding the effect of different types of advertising than relying on the survey alone. Similarly, Porter and Lazaro (2014) describe a series of business case studies to illustrate how survey data can be blended with data from other sources to enhance the overall analysis. In one case study, the authors highlight the use of a blended data strategy to make comparisons by respondent. In the case study, consumer behavior data from website activity and transactions is combined with survey data capturing perceptions, attitudes, life events, and offsite behavior. By using respondent-level models to relate customer perceptions (from survey data) to behaviors for the same customers (from data on website activity), they were better able to understand the *whys* behind online behavior, and prioritize areas for improvement based on understanding the needs of different individuals.

Blending strategies are also being pursued by government agencies. For example, the National Center of Health Statistics (NCHS) is developing a record linkage program designed to maximize the scientific value of the Center's population based surveys.[4] The program has linked various NCHS surveys to administrative records from CMS and the Social Security Administration (SSA) under an interagency agreement among NCHS, CMS, SSA, and the Office of the Assistant Secretary for Planning and Evaluation, so these linked data can be used to support analysis of the blended data. Ultimately, linked data files should enable researchers to examine in greater detail the factors that influence disability, chronic disease, healthcare utilization, morbidity, and mortality.

Similarly, the US Census Bureau is identifying ways in which Big Data can be used to improve surveys and Census operations to increase the timeliness of data, increase the explanatory power of Census Bureau data, and reduce operational costs of data collection (Bostic 2013). For example, the Bureau is planning to use data on electronic transactions and administrative data to supplement or improve construction and retail and service statistics that the Bureau maintains. In construction, the agency is examining the value of using vendor data on new residential properties in foreclosure to aid analysis of data on new construction and sales. The agency is also looking at ways to incorporate the use of online public records that are maintained by local jurisdictions and state agencies. In retail, the agency is evaluating electronic payment processing to fill data gaps such as geographical detail and revenue measures by firm size. All the Nordic countries have a system

4. http://1.usa.gov/1IlwiLW.

of statistical registers that are used on a regularly basis to produce statistics. The system is shown in figure 8 and has four cornerstones: population, activity, real estate, and business registers (Wallgren and Wallgren 2014).

Conclusions and Research Needs

In this section, we revisit the questions in the task force mission.

Can/Should Big Data be used to generate population statistics related to knowledge, opinion, and behavior?

There are many different types of Big Data. In this report, we include administrative data as one of them. The different types of Big Data differ due to the amount of researcher control and the degree of potential inferential power associated with each type (Kreuter and Peng 2014). On one side of the spectrum, we have administrative data that has been used in some countries for many years to derive population estimates; for example, in the Nordic countries, their population censuses are based on administrative data. Statistical agencies form partnership with owners of administrative data and can influence the design of the data. On the other side of the spectrum, we have Big Data from social media platforms, where the researcher has no control of or influence on the data. During the past few years, we have seen examples of statistics based on social media data. We have also seen studies that compare estimates from Big Data sources to estimates from traditional surveys. At the moment, however, there is not enough research to allow best practices to be

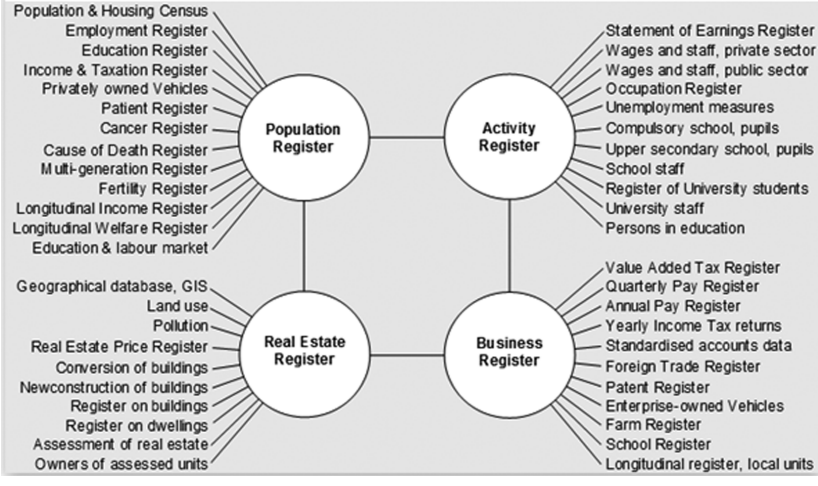

Figure 8. A System of Statistical Registers—by Object Type and Subject Field, from Wallgren and Wallgren (2014).

developed for deriving population estimates from these social media types of Big Data. In between, we have examples of Big Data sources where researchers actually can exert some control, for instance by positioning sensors in preassigned places to measure traffic flows and thereby also to some extent measure travel behavior.

One of the main criticisms regarding the use of Big Data is that there is no theory for making inference from it. The fact that Big Data is big is not enough, albeit some argue just that. The sampling theory that many statistical agencies rely on today was developed at a time when the only way to get data was to collect information from the total population. This was a very expensive endeavor, and the sampling theory came as the rescue. Today, we have a situation where a lot of Big Data is generated as byproducts from various processes or even the product from these processes. At the same time, it is difficult to get participation in surveys, costs for surveys are rising, and many of the assumptions from the sampling theory are violated due to nonresponse and other nonsampling errors. We are moving from the traditional survey paradigm to a new one, where multiple data sources might be used to gain insights. Big Data is one of these data sources that can contribute valuable information. It is essential that theory and methods be developed so that the full potential of Big Data can be realized, in particular for "found" data that lack purposeful design. We are not there yet.

The gathering or collection of Big Data contains errors that will affect any estimates made and each Big Data source will have its own set of errors. The potential impact of each error source will vary between different Big Data sources. Just as we do for "small data," we need to have a total error perspective when we consider using a Big Data source, and a Big Data Total Error framework would help guide research efforts (Biemer 2014).

How can Big Data improve and/or complement existing "classical" research methods such as surveys and/or censuses?

The availability of Big Data to support research provides a new way to approach old questions as well as an ability to address some new questions that in the past were out of reach. Big Data can be used to generate a steady flow of information about what is happening—for example, how customers behave—while traditional research can focus instead on deeper questions about why we are observing certain trends or deviations from them—for example, why customers behave the way they do and what could be done to change their behavior.

Administrative data are used in several countries as sampling frames, in the estimation process in order to improve precision, and in combination with surveys in order to minimize respondent burden. Other types of Big Data can be used in similar ways. Social media platforms can be used to get quick information about how people think about different concepts and to test questions.

Administrative data are also used as the gold standard in some methodological studies. For example, Day and Parker (2013) use data developed under the record linkage program to compare self-reported diabetes in the National Health Interview Survey (NHIS) with diabetes identified using the Medicare Chronic Condition Summary file, derived from Medicare claims data.

If we go beyond administrative data and look at other types of Big Data, we see the opposite. Now, survey data are used as a benchmark. There are a number of studies that look at estimates from a Big Data source and compare those results with estimates from a traditional survey. The correlation between the two sets of estimates is of interest in these types of studies. If the correlation is high (and does not suffer from unknown algorithmic changes), the Big Data statistics can be used as an early warning system (e.g., Google Flu) since they are cheap and fast. For this to work, transparency of algorithms is key, and agreements need to be found with the private sector to ensure that they are stable and known.

In the private sector, Big Data is used to manage work and to make decisions. Examples of research techniques used are predictive analytics and rapid-cycle evaluation.

Can Big Data outperform surveys? What if any current uses of Big Data (to learn about public knowledge, opinion, and behaviors) appear promising? Which types of applications seem inappropriate?

Big Data has a number of advantages when compared to survey data. An obvious advantage is that these data already exist in some form, and therefore research based on Big Data does not require a new primary data-collection effort. Primary data collection is usually expensive and slow, which can either greatly delay the generation of new research findings or even make new research prohibitively expensive.

As mentioned earlier, administrative data are being used in many countries. The Nordic countries have a system of statistical registers that are used on a regular basis to produce statistics about the population, businesses, as well as economic and real estate activities.

A useful strategy is to combine new and traditional data sources to support research, analytics, and decision making, with the precise combination depending on the demands of a given situation. Scanner data from retailers are one example of a type of Big Data source that combined with traditional survey methods can both increase data quality and decrease costs. Scanner data are, for example, used in the production of the Consumer Price Index (CPI) in several countries. Another example is Big Data obtained from tracking devices, such as a log of steps drawn from networked pedometers, which might be more accurate than what could be solicited in surveys given known problems with recall error. Other Big Data sources with a similar potential include sensor data and transactional data. As these examples show, so far the integration of the data sources is more straightforward if both small and Big Data are designed data. However, we are hopeful that

the work of AAPOR and others in this area will expand the integration to found data as well.

What are the operational and statistical challenges associated with the use of Big Data?

The current pace of the Big Data development is in itself a challenge. It is very difficult to keep up with the development, and research on new technology tends to become outdated very fast. Therefore, a good strategy for an organization is to form partnerships with others so that multidisciplinary teams can be set up in order to make full use of the Big Data potential.

Data ownership is not well defined, and there is no clear legal framework yet for the collection and subsequent use of Big Data. Most users of digital services have no idea that their behavior data may be reused for other purposes. Researchers must carefully consider data ownership issues for any content they seek to analyze. The removal of key variables as Personally Identifiable Information (PII) is no longer sufficient to protect data against re-identification. The combination of location and time metadata with other factors enables re-identification of "anonymized" records in many cases. New models of privacy protection are required.

Organizations seeking to experiment with Big Data computer cluster technology can reduce their initial capital outlays by renting prebuilt computer cluster resources (such as Apache Hadoop) from online providers. Systems such as Apache Hadoop drastically simplify the creation of computer clusters capable of supporting parallel processing of Big Data computations.

Although the cost of magnetic storage media may be low, the cost of creating systems for the long-term storage and analysis of Big Data remains high. The use of external computer cluster resources is one short-term solution to this challenge.

Appendix

Glossary on Big Data Terminology

Big Data: Data that are so large in context that handling the data becomes a problem in and of itself. Data can be hard to handle due to its size (volume) and/or the speed of which it's generated (velocity) and/or the format in which it is generated, like documents of text or pictures (variety)

Data-generating process: Also known as the likelihood function, the process from which the data are generated (i.e., where did the data come from)

Found data: Also known as organic data, data that are created as a byproduct of another process or activity (for example, sensor data from a production line or timestamps and geo-data created from a tweet)

Hadoop: An open-source distributed file system that can store both structured and unstructured data. Further, all data are duplicated so that no data are lost even if some hardware would break

Made data: Also known as designed data, data that are created with an explicit purpose (for example, survey data or data from an experiment)

Map-reduce: A divide-and-conquer data-processing paradigm that distributes a heavy computation between several computers, speeding up the total time of the computation (for example, having 10 computers searching 1 billion records each takes less time than having one computer searching 10 billion records by itself)

Structured data: Numerical and categorical data that fit into traditional relational databases. Most data that "feel natural" to work with can be considered structured data

Unstructured data: Data that do not follow a clear structure (for example, text in PDF files, sequences of video from security cameras, etc.) and that would need to be processed and organized in order to be worked with

References

Acquisti, Alessandro. 2014. "The Economics and Behavioral Economics of Privacy." In *Privacy, Big Data, and the Public Good: Frameworks for Engagement*, edited by Julia Lane, Victoria Stodden, Stefan Bender and Helen Nissenbaum, 98–112. New York: Cambridge University Press.

Antenucci, Dolan, Michael Cafarella, Margaret Levenstein, Christopher Ré, and Matthew D. Shapiro. 2014. *"Using Social Media to Measure Labor Market Flows."* National Bureau of Economic Research, Working Paper 20010. doi:10.3386/w20010.

Barocas, Solon, and Helen Nissenbaum. 2014. "Big Data's End Run around Anonymity and Consent." In *Privacy, Big Data, and the Public Good: Frameworks for Engagement*, edited by Julia Lane, Victoria Stodden, Stefan Bender, and Helen Nissenbaum, 44–75. New York: Cambridge University Press.

Biemer, Paul P. 2010. "Total Survey Error: Design, Implementation, and Evaluation." *Public Opinion Quarterly* 74:817–48.

———. 2014. *"Toward a Total Error Framework for Big Data."* Paper presented at the Annual Meeting for the American Association for Public Opinion Research, Anaheim, CA, USA.

Biemer, Paul P., and Dennis Trewin. 1997. "A Review of Measurement Error Effects on the Analysis of Survey Data." In *Survey Measurement and Process Quality*, edited by Lars Lyberg, Paul P. Biemer, Martin Collins, Edith de Leeuw, Cathryn Dippo, Norbert Schwarz, and Dennis Trewin, 603–32. New York: Wiley & Sons.

Bostic, William G. 2013. *"Big Data Projects at the Census Bureau."* Paper presented at the Council of Professional Associations on Federal Statistics, Washington, DC, USA.

Brynjolfsson, Erik, Lorin M. Hitt, and Heekyung Hellen Kim. 2011. *"Strength in Numbers: How Does Data-Driven Decision-Making Affect Firm Performance?"* Proceedings from the International Conference on Information Systems. http://ssrn.com/abstract=1819486.

Butler, Declan. 2013. "When Google Got Flu Wrong." *Nature News*, February 13.

Cecil, Joe, and Donna Eden. 2003. "The Legal Foundations of Confidentiality." Cited by Julia Lane, *Key Issues in Confidentiality Research: Results of an NSF Workshop*. National Science Foundation. http://1.usa.gov/1Eq58Df.

Cody, Scott, and Andrew Asher. 2014. "Smarter, Better, Faster: The Potential for Predictive Analytics and Rapid-Cycle Evaluation to Improve Program Development and Outcomes." *Mathematica Policy Research*. Available at http://brook.gs/1AMaW5t.

Couper, Mick P. 2013. "Is the Sky Falling? New Technology, Changing Media, and the Future of Surveys." *Survey Research Methods* 7:145–56.

Cukier, Kenneth, and Viktor Mayer-Schoenberger. 2013. "Rise of Big Data: How It's Changing the Way We Think about the World." *Foreign Affairs*, May–June. Available at http://fam.ag/1bKTv6t.

Daas, Piet J. H., and Marco J. H. Puts. 2014. "Social Media Sentiment and Consumer Confidence." *European Central Bank Statistics* Paper Series, No. 5.

Daas, Piet J. H., Marco J. H. Puts, Bart Buelens, and Paul A. M. van den Hurk. 2013. *"Big Data and Official Statistics."* Paper presented at the New Techniques and Technologies for Statistics Conference, Brussels, Belgium.

Day, Hannah R., and Jennifer D. Parker. 2013. "Self-Report of Diabetes and Claims-Based Identification of Diabetes among Medicare Beneficiaries." *National Health Statistics Reports*, No. 69.

Donohue, John J., and Justin Wolfers. 2006. "Uses and Abuses of Empirical Evidence in the Death Penalty Debate." *Stanford Law Review* 58:791–846.

Duncan, George T., Mark Elliot, and Juan Jose Salazar-Gonzalez. 2011. *Statistical Confidentiality, Principles and Practice*. New York: Springer.

Duong, Thao, and Steven Millman. 2014. *"Behavioral Data as a Complement to Mobile Survey Data in Measuring Effectiveness of Mobile Ad Campaign."* Paper presented at the Council of American Survey Research Organizations Digital Research Conference, San Antonio, TX, USA. http://bit.ly/1v4fFBc.

Dwork, Cynthia. 2011. "A Firm Foundation for Private Data Analysis." *Communications of the ACM* 54(1):86–95.

Evans, David S. 1987. "Tests of Alternative Theories of Firm Growth." *Journal of Political Economy* 95:657–74.

Executive Office of the President. 2014. *"Big Data: Seizing Opportunities, Preserving Values."* Washington, DC. http://1.usa.gov/1hqgibM.

Fan, Jianqing, Fang Han, and Han Liu. 2014. "Challenges of Big Data Analysis." *National Science Review* 1:293–314.

Fan, Jianqing, and Yuan Liao. 2014. "Endogeneity in Ultrahigh Dimension." *Annals of Statistics* 42:872–917.

Fan, Jianqing, Richard Samworth, and Yichao Wu. 2009. "Ultrahigh Dimensional Feature Selection: Beyond the Linear Model." *Journal of Machine Learning Research* 10:2013–2038.

Gelman, Andrew, Jeffrey Fagan, and Alex Kiss. 2007. "An Analysis of the New York City Police Department's 'Stop-and-Frisk' Policy in the Context of Claims of Racial Bias." *Journal of the American Statistical Association* 102:813–23.

Gerber, Eleanor R. 2001. "The Privacy Context of Survey Response: An Ethnographic Account." In *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*, edited by Pat Doyle, Julia Lane, Jules Theeuwes, and Laura Zayatz, 371–95. Amsterdam: Elsevier.

Greenwood, Daniel, Arkadiusz Stopczynski, Brian Sweatt, Thomas Hardjono, and Alex Pentland. 2014. "The New Deal on Data: A Framework for Institutional Controls." In *Privacy, Big Data, and the Public Good: Frameworks for Engagement*, edited by Julia Lane, Victoria Stodden, Stefan Bender, and Helen Nissenbaum, 192–200. New York: Cambridge University Press.

Griffin, Jane. 2008. "The Role of the Chief Data Officer." *Data Management Review* 18:28.

Groves, Robert M. 2011a. "Three Eras of Survey Research." *Public Opinion Quarterly* 75:861–71.

———. 2011b. "'Designed Data' and 'Organic Data.'" *Directors Blog*, May 31. US Census Bureau. Available at http://1.usa.gov/15NDn8w.

Halevy, Alon, Peter Norvig, and Fernando Pereira. 2009. "The Unreasonable Effectiveness of Data." *Intelligent Systems* 24(2):8–12. doi:10.1109/MIS.2009.36.

Hall, Peter, and Hugh Miller. 2009. "Using Generalized Correlation to Effect Variable Selection in Very High Dimensional Problems." *Journal of Computational Graphical Statistics* 18:533–50.

Hey, Tony, Stewart Tansley, and Kristin Tolle. 2009. *The Fourth Paradigm: Data Intensive Scientific Discovery*. Microsoft Research.

Ibrahim, Joseph, and Ming-Hui Chen. 2000. "Power Prior Distributions for Regression Model." *Statistical Science* 15:46–60.

Jovanovic, Boyan. 1982. "Selection and the Evolution of Industry." *Econometrica* 50:649–70.

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

Karr, Alan F., and Jerome P. Reiter. 2014. "Using Statistics to Protect Privacy." In *Privacy, Big Data, and the Public Good: Frameworks for Engagement*, edited by Julia Lane, Victoria Stodden, Stefan Bender, and Helen Nissenbaum, 276–95. New York: Cambridge University Press.

Keller, Sallie Ann, Steven E. Koonin, and Stephanie Shipp. 2012. "Big Data and City Living: What Can It Do for Us?" *Significance* 9(4):4–7.

Kinney, Satkartar K., Alan F. Karr, and Joe Fred Gonzalez Jr. 2009. "Data Confidentiality: The Next Five Years Summary and Guide to Papers." *Journal of Privacy and Confidentiality* 1:125–34.

Koonin, Steven E., and Michael J. Holland. 2014. "The Value of Big Data for Urban Science." In *Privacy, Big Data, and the Public Good: Frameworks for Engagement*, edited by Julia Lane, Victoria Stodden, Stefan Bender, and Helen Nissenbaum, 137–52. New York: Cambridge University Press.

Kreuter, Frauke, and Roger D. Peng. 2014. "Extracting Information from Big Data: Issues of Measurement, Inference and Linkage." In *Privacy, Big Data, and the Public Good: Frameworks for Engagement*, edited by Julia Lane, Victoria Stodden, Stefan Bender, and Helen Nissenbaum, 257–75. New York: Cambridge University Press.

Lane, Julia, and Victoria Stodden. 2013. "What? Me Worry? What to Do about Privacy, Big Data, and Statistical Research." *AMStat News*, December 1. Available at http://bit.ly/15UL9OW.

Lane, Julia, Victoria Stodden, Stefan Bender, and Helen Nissenbaum. 2014. "Editors' Introduction." In *Privacy, Big Data, and the Public Good: Frameworks for Engagement*, edited by Julia Lane, Victoria Stodden, Stefan Bender, and Helen Nissenbaum, xi–xix. New York: Cambridge University Press.

Laney, Douglas. 2001. "3-D Data Management: Controlling Data Volume, Velocity, and Variety." *META Group Research Note*, February 6. Available at http://gtnr.it/1bKflKH.

———. 2012. "The Importance of 'Big Data': A Definition." *Gartner Inc*.

Lazer, David M., Ryan Kennedy, Gary King, and Alessandro Vespignani. 2014. "The Parable of Google Flu: Traps in Big Data Analysis." *Science* 343:1203–1205.

Leek, Jeff. 2014a. "Why Big Data Is in Trouble: They Forgot about Applied Statistics." *Simplystats Blog*, May 7. Available at http://bit.ly/1fUzZO1.

———. 2014b. "10 Things Statistics Taught Us about Big Data Analysis." *Simplystats blog*, May 22. Available at http://bit.ly/S1ma4Z.

Levitt, Steven D., and Thomas J. Miles. 2006. "Economic Contributions to the Understanding of Crime." *Annual Review of Law and Social Science* 2:147–64.

Lohr, Steve. 2012. "The Age of Big Data." *New York Times*, February 11. Available at http://nyti.ms/1f7WKqh.

———. 2014. "For Big-Data Scientists, 'Janitor Work' Is Key Hurdle to Insights." *New York Times*, August 17. Available at http://nyti.ms/1Aqif2X.

Manzi, Jim. 2012. *Uncontrolled: The Surprising Payoff of Trial-and-Error for Business, Politics, and Society*. New York: Basic Books.

McAfee, Andrew, and Erik Brynjolfsson. 2012. "Big Data: The Management Revolution." *Harvard Business Review* 90:61–67.

Murphy, Joe, Michael W. Link, Jennifer Hunter Childs, Casey Langer Tesfaye, Elizabeth Dean, Michael Stern, Josh Pasek, Jon Cohen, Mario Callegaro, and Paul Harwood. 2014. "Social

Media in Public Opinion Research: Report of the AAPOR Task Force on Emerging Technologies in Public Opinion Research." *AAPOR Task Force* Report. Available at http://bit.ly/15V7coJ.

New York Times Editorial Board. 2014. "Better Governing Through Data." *New York Times*, August 19. Available at http://nyti.ms/1qehhWr.

Nielsen, Michael. 2012. *Reinventing Discovery: The New Era of Networked Science*. Princeton, NJ: Princeton University Press.

Nissenbaum, Helen. 2011. "A Contextual Approach to Privacy Online." *Daedalus* 140(4):32–48.

Norberg, Anders, Muhanad Sammar, and Can Tongur. 2011. "A Study on Scanner Data in the Swedish Consumer Price Index." Paper presented at the Statistics Sweden Consumer Price Index Board, Stockholm, Sweden.

Ohm, Paul. 2010. "Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization." *UCLA Law Review* 57:1701–1818.

Pardo, Theresa A. 2014. "Making Data More Available and Usable: A Getting Started Guide for Public Officials." Paper presented at the Privacy, Big Data, and the Public Good Book Launch. http://bit.ly/1Czw7u4.

Porter, Scott, and Carlos G. Lazaro. 2014. "Adding Big Data Booster Packs to Survey Data." Paper presented at the Council of American Survey Research Organizations Digital Research Conference, San Antonio, TX, USA.

Prada, Sergio I., Claudia González-Martínez, Joshua Borton, Johannes Fernandes-Huessy, Craig Holden, Elizabeth Hair, and Tim Mulcahy. 2011. "Avoiding Disclosure of Individually Identifiable Health Information. A Literature Review." *SAGE Open*. doi:10.1177/2158244011431279.

Schenker, Nathaniel, Marie Davidian, and Robert Rodriguez. 2013. "The ASA and Big Data." *AMStat News*, June 1. Available at http://bit.ly/15XAzX8.

Scott, Steven L., Alexander W. Blocker, Fernando V. Bonassi, Hugh A. Chipman, Edward I. George, and Robert E. McCulloch. 2013. "Bayes and Big Data: The Consensus Monte Carlo Algorithm." http://bit.ly/1wBqh4w.

Shelton, Taylor, Ate Poorthuis, Mark Graham, and Matthew Zook. 2014. "Mapping the Data Shadows of Hurricane Sandy: Uncovering the Sociospatial Dimensions of 'Big Data.'" *Geoforum* 52:167–79.

Squire, Peverill. 1988. "Why the 1936 Literary Digest Poll Failed." *Public Opinion Quarterly* 52:125–33.

Stanton, Mark W. 2006. "The High Concentration of US Health Care Expenditures." *Research in Action* 19:1–10.

Stock, James H., and Mark W. Watson. 2002. "Forecasting Using Principal Components from a Large Number of Predictors." *Journal of the American Statistical Association* 97:1167–1179.

Strandburg, Katherine J. 2014. "Monitoring, Datafication, and Consent: Legal Approaches to Privacy in the Big Data Context." In *Privacy, Big Data, and the Public Good: Frameworks for Engagement*, edited by Julia Lane, Victoria Stodden, Stefan Bender, and Helen Nissenbaum, 5–43. New York: Cambridge University Press.

Tambe, Prasanna, and Lorin M. Hitt. 2012. "The Productivity of Information Technology Investments: New Evidence from IT Labor Data." *Information Systems Research* 23:599–617.

Tapia, Andrea H., Nicolas LaLone, and Hyun-Woo Kim. 2014. "Run Amok: Group Crowd Participation in Identifying the Bomb and Bomber from the Boston Marathon Bombing." Proceedings from the International Conference on Information Systems for Crisis Response and Management, 265–74.

Taylor, Sean J. 2013. "Real Scientists Make Their Own Data." *Sean J. Taylor Blog*, January 25. Available at http://bit.ly/15XAq5X.

ten Bosch, Olav, and Dick Windmeijer. 2014. "On the Use of Internet Robots for Official Statistics." Paper presented at the Meeting on the Management of Statistical Information Systems, Dublin, Ireland.

Thompson, William W., Lorraine Comanor, and David K. Shay. 2006. "Epidemiology of Seasonal Influenza: Use of Surveillance Data and Statistical Models to Estimate the Burden of Disease." *Journal of Infectious Diseases* 194:S82–S91.

Tourangeau, Roger, Lance J. Rips, and Kenneth Rasinski. 2000. *The Psychology of Survey Response*. New York: Cambridge University Press.

Varian, Hal R. 2014. "Big Data: New Tricks for Econometrics." *Journal of Economic Perspectives* 28(2):3–27.

Wallgren, Anders, and Britt Wallgren. 2007. *Register-Based Statistics: Administrative Data for Statistical Purposes*. New York: Wiley & Sons.

———. 2014. *Register-Based Statistics: Statistical Methods for Administrative Data*. New York: Wiley & Sons.

Winkler, William E. 2005. "Re-Identification Methods for Evaluating the Confidentiality of Analytically Valid Microdata." Research Report Series, Statistics #2005–09. US Census Bureau.