23

The Role of Surveys in the Era of "Big Data"

Mario Callegaro and Yongwei Yang

Introduction: The Changing Definition of Big Data

The definition of "Big Data" is complex and constantly changing. For example, Dutcher (2014) asked 40 different thought leaders to define Big Data and obtained nearly 40 different definitions. However, there is some consensus in the literature on the main characteristics of Big Data as described by a widely cited Gartner report (Beyer and Laney 2012).

> In terms of ***Volume***, Big Data are those data that cannot be handled by traditional analytics tools.
> In terms of ***Velocity***, Big Data refers data that are coming in (almost) real-time.
> In terms of ***Variety***, Big Data are complex datasets and include very different sources of context such as unstructured text, media content such as images and videos, logs, and other data sources.

Adding to these three key characteristics of Big Data other authors have cited *variability* (how the data can be inconsistent across time), *veracity*

M. Callegaro (✉)
London, UK
e-mail: callegaro@google.com

Y. Yang
Boulder, CO, USA
e-mail: yongwei@google.com

D.L. Vannette, J.A. Krosnick (eds.), *The Palgrave Handbook of Survey Research*, https://doi.org/10.1007/978-3-319-54395-6_23

(accuracy and data quality) and *complexity* (how to link multiple databases properly). In practice, what is often called "Big Data" may not possess all six of these characteristics (e.g., you can have very large data of great complexity that may not come in with high velocity). We refer the reader to Baker (2016) for an extensive definition of Big Data in the context of survey research.

For the purpose of this chapter, we contrast Big Data with survey data. In this framework, a helpful definition of Big Data was proposed by Groves (2011) who, as a contrast to designed data (survey data), calls it organic data and described it as the data produced by "systems that automatically track transactions of all sorts" (p. 868).

In the survey literature, we find *Big Data thinking* in the emerging term of "Small Big Data" where the authors use multiple survey datasets to enable richer data analyses (Warshaw 2016; Gray et al. 2015). Small Big Data are more and more a reality thanks to the availability of social science data archives (e.g., the UK Data Service or the Roper Center for Public Opinion Research). Although we do not strictly classify them as Big Data per the aforementioned description, they are worth mentioning in this chapter.

Another way to contrast Big Data with survey data is to look at potential sources of Big Data that can answer research questions. Depending on the nature of the research questions, the answer will lie in a continuum of sources from Big Data on one side and survey data on the other side and the combination of the two in the middle – our thesis of this chapter. We identify the following main sources and subclasses. This list is not meant to be highly detailed and comprehensive, and some sets of data cannot be uniquely classified in one or another class:

- *Internet data: Online text, videos, and sound data*. It encompasses all online content relevant to a research question. Using such data is commonly referred to as Internet research methods (Hewson et al. 2016).
    - *Social media data*. Social media data are a subset of Internet data and include text, photos, and videos which are publicly available by mining social media networks such as Twitter and Facebook. Social media data are probably the first and most studied Big Data for public opinion measurement (Schober et al. 2016).
    - *Website metadata, logs, cookies, transactions, and website analytics*. These are data produced by websites and analytics tools (think about Google Analytics or Adobe Analytics) and used heavily in online advertisement, shopping analytics, and website analytics.

- *The Internet of Things.* Internet of Things (IOT) (Gershenfeld et al. 2004) refers to any device that can communicate with another using the Internet as the common transmission protocol. As more and more devices become connected via the Internet, more data are generated and can be used to answer research questions.
  - *Behavioral data* are a subset of the IOT. Behavioral data come from devices such as smartphones, wearable technology, and smart watches carried by subjects and passively recording data such as locations, physical activities, and health status (e.g., Swan 2013). Behavioral data can also be manually recorded by the users.
- *Transaction data*. In the business world, transaction data have been around since before electronic data formats existed. They are records of orders, shipments, payments, returns, billing, and credit card activities, for examples (Ferguson 2014). Transaction data are nowadays part of customer relationship management tools where the attempt is to capture every interaction a customer has with a company or product. The area is also called business intelligence (Hsinchun et al. 2012). The same applies to government and public sector where more and more user interactions are stored digitally.
- *Administrative data*. Administrative data and registers are a form of Big Data collected by public offices such as national health, tax, school, benefits, and pensions, or driver licenses databases. Administrative data have a long tradition of being used for statistical purposes (Wallgren and Wallgren 2014). Survey data can be linked to administrative data as shown by Sakshaug in this volume. Health data in some countries are collected and stored by private companies but, although they are of the same nature of public health data, they are usually not discussed as administrative data in the academic literature.
- *Commercially available databases*. More and more companies are collecting, curating, and storing data about consumers. By using publicly available records, purchasing records from companies, matching techniques (Pasek, this volume), and other algorithms such as imputations from other sources (e.g., census data), these companies create a profile for each individual in their database. They combine data from the previously mentioned sources just described. Examples are Acxiom, Epsilon, Experian Marketing Services, or, in the political domain, Catalist, Aristotle, and NationBuilder. These companies are often referred to as *data brokers* (Committee on Commerce, Science and Transportation 2013).

Finally, and related to survey data, we define *paradata* (Kreuter, this volume) as a source of Big Data. Paradata is data about the process of answering the survey itself (Callegaro 2013), including data collected by systems and third parties (e.g., interviewers) before, during, and after the administration of a questionnaire. Paradata often come in real time (think about collecting answer time per question on a web survey) and are in complex formats (e.g., user agent strings, time latency, mouse movements, interviewer observations).

The Perspectives About Error and Data Quality

Big Data does not necessarily mean good quality or without any error. Often Big Data comes with *Big Noise* (Waldherr et al. 2016). Within the survey research tradition, the concept of survey errors was developed in the early 1940s (Deming 1944) and has since evolved into the Total Survey Error (TSE) framework (Biemer 2010). Applying the concept of survey error to Big Data is a healthy data quality approach where cross-fertilization among the two disciplines is at its best. TSE "refers to the accumulation of all errors that may arise in the design, collection, processing, and analysis of survey data" (Biemer 2010, 817). More specifically *specification errors* occure when some concepts that we want to measure in a survey are actually measured differently. *Measurement errors* occur from the interviewers, the respondents, and the questionnaire itself including the data collection methods used to administered the questionnaire. *Frame errors* are errors related to the quality of the sampling frame. The frame can have missing units, duplications, units that are not supposed to be in the frame, and the records themselves can contain mistakes or be outdated. *Nonresponse errors* arise when some respondents (unit nonresponse) do not answer the questionnaire altogether and when some questions (item nonresponse) are not answered (e.g., income question). Finally, there are data *processing errors* stemming from the processes of tabulation, coding, data entry, and producing survey weights.

When applying the general framework of TSE to Big Data we obtain the Big Data Total Error (BDTE) (Japec et al. 2015). Errors in Big Data arise mostly during three steps used to create a dataset from Big Data (Biemer 2014, 2016):

1. *Data generation*. It is specifically the data generation process that differentiates Big Data from surveys, censuses, and experiments (Kreuter and Peng 2014). Data generation in Big Data is sometimes a "black box" and not

always well documented. Errors can take the form of missing data, self-selection, coverage, non-representativeness, and low signal to noise ratio.
2. *Extract, Transform, and Load (ETL)*. This is the process when the data are brought together in the same computing environment with the process of extraction (data accessed, parsed and stored from multiple sources), transformation (e.g., coding, recoding, editing) and loading (integration and storage). Errors in ETL can take the form of matching, coding, editing, and data cleaning errors.
3. *Analysis and visualization*. Here the errors can be of sampling, selectivity, modeling and estimation. Finally, errors might arise in the data visualization step.

It is important to note that the BDTE concept is relatively new, and outside the survey community (e.g., Biemer 2016) "very little effort has been devoted to enumerating the error sources and the error generating processes in Big Data" (Japec et al. 2015, 854). For an exception see Edjlali and Friedman (2011). We hope this chapter will provide a good starting point to conduct more research on how Big Data and surveys can safely and validly integrate.

Challenges and New Skills Needed for the Survey Researcher Working with Big Data

Gathering, analyzing, and interpreting Big Data requires technical expertise not traditionally gained from survey or social science research training. These may include database skills (NoSQL, relational DBMS), programming skills for mass data processing (e.g., MapReduce), data visualization expertise, as well as analytical techniques not commonly taught to students dealing with survey data (e.g., random forests). Foster et al. (2016) provides a timely discussion about this topic. Even among those who are proficient with Big Data applications, there might exist differing interests and strengths, such as the type A (analysis) versus type B (pipeline building) distinction of data scientists discussed by Chang (2015). Importantly, it will not be feasible to become proficient in all new tools and skills. Instead, a winning strategy is to collaborate with others who have different expertise and strengths.

Technical skills aside, when looking at Big Data as potentially providing substantive answers to what have been studied with surveys, two classes come to mind: Google Trends and social media listening tools. Google Trends (Stephens-Davidowitz and Varian 2015) provides an index of search

activities by keywords or categories as well as of interest on these keywords or categories over time. There are numerous examples of using Google Trends to forecast and approximate trends estimated from surveys. Choi and Varian (2012), for instance, show how Google Trends matches the survey-based Australian Consumer Confidence index and Scott and Varian (2015) reproduce the same results for the University of Michigan Consumer Confidence Index time series. Chamberlin (2010) explores Google trends correlations with U.K. Office of National Statistics official data on retail sales, property transactions, car registrations, and foreign trips. At the same time Google Trends does not answer more specific survey questions, such as demographic analysis (e.g., are female consumers more worried about the economy than male consumers?) or modeling questions (what are the drivers of the consumer sentiment in a particular country?).

Social media listening and monitoring tools perform two main tasks: locate social media content from a variety of social media sources and perform automated analysis (content analysis) of the text collected. These tools vary in the depth, range, and historical reach of the content aggregated. When it comes to content analysis, the most common classification of text is as positive, neutral, and negative. In order to do so, social media listening tools use different dictionaries to classify text (see González-Bailón and Paltoglou 2015). Another common usage of social media in the context of surveys is to use it as a a supplement to or replacement of pre-election polls. For example, the percent share of Twitter traffic messages mentioning the six political parties in the 2009 German election was very close to the actual election results (Tumasjan et al. 2010). Using social media tools to replace pre-election polls is not always successful as discussed in Jungherr et al. (2016). There are still many challenges from a methodological and technical point of view to be taken into account and researched.

Changes in the Survey Landscape

All the aforementioned tools and new types of data are making some wonder if surveys are eventually going to disappear because they will be replaced by Big Data. This is true, to some extent. Examples include Censuses in countries such as Finland and other Nordics countries (Statistics Finland 2004) being replaced by administrative data. Other countries go beyond the Census and use administrative data for other social statistics data collection, for example, the Netherlands (Bakker et al. 2014).

Two proponents of the rapid disappearance of surveys, Ray Poynter (2014) and Reginald Baker (2016), use ESOMAR Global Market Research (e.g., ESOMAR. 2015) reports over time to show a decline in the percent of budget spent by market research companies on surveys. For example, in comparison to 2013, the combined percent of online, telephone, face-to-face, and mail survey declined by 6 percent as compared to 2014 (ESOMAR. 2015, 20). The same report is also showing an increased trend of money spent in Automated/Digital and electronic data collection. This category refers to retail audits and media measurement. In other words, market research companies are investing more and more money in Big Data.

Although we do agree with the trend analysis of the ESOMAR reports, we disagree with the implications. The ESOMAR reports capture what is spent by market research companies around the world by contacting country market research associations. The same ESOMAR reports show a change in data collection methods moving more and more to online and smartphone surveys at the expense of other traditional data collection methods such as telephone and face-to-face surveys. What the report cannot capture are two other trends that show increased usage of:

- Do-it-yourself (DIY) web survey platforms
- In-house web survey tools

In the first case (DIY), companies such as SurveyMonkey reported generating 90 million survey completes per month worldwide (Bort 2015). This is not a small number. Qualtrics, another DIY survey tool, distributes one billion surveys annually (personal communication, February 11, 2016). Both survey platforms have major companies as clients in their portfolio.

In the second case, organizations are using in-house web surveys tools, without the need to outsource data collection to market research companies. For example, Google collects customer feedback at scale for all its products using probability-based intercept surveys called Happiness Tracking Survey (Müller and Sedley 2014). Other technology companies has followed suit (Martin 2016).

To summarize, we believe that the real trend in survey-based social and market research is the following:

- From offline data collection methods to web surveys
- From web surveys to mobile web surveys

- From outsourced market research to in-house market research using DIY web survey platforms
- From outsourced market research to in-house market research fully integrated with internal systems

How Surveys and Big Data Can Work Together

Answering the What and the Why

The most commonly shared view among researchers is that surveys and Big Data can and should be used together to maximize the value of each. This is, not surprisingly, one of the takeaways from the American Association for Public Opinion Research task force report on Big Data (Japec et al. 2015).

Looking ahead, the ideal case is to build on the strengths of both data collection methods. Big Data can measure behaviors and tell us the "what" while surveys can measure attitudes and opinions and tell us the "why." A good example of this view comes from a recent Facebook blog post written by two software engineers (Zhang and Chen 2016). The blog post explains the process Facebook used to redesign their News Feed.

> The goal of News Feed is to show you the stories that matter most to you. The actions people take on Facebook – liking, clicking, commenting or sharing a post – are historically some of the main factors considered to determine what to show at the top of your News Feed. But these factors don't always tell us the whole story of what is most meaningful to you. As part of our ongoing effort to improve News Feed, we ask over a thousand people to rate their experience every day and tell us how we can improve the content they see when they check Facebook – we call this our Feed Quality Panel. We also survey tens of thousands of people around the world each day to learn more about how well we're ranking each person's feed.

Surveys Are Just One of a Number of Tools

Sometimes market and survey researchers become so involved in surveys that they forget that surveys are not the *only* tool available to answer research questions (Couper 2013). An illustration can be found when looking at the level of precisions some surveys strive for when asking behavioral questions.

Despite the incredible advances in questionnaire design in past 50 years, asking behavioral questions is and will always be difficult because the answers rely on people's memories. For example, the U.S. Consumer Expenditure Survey used to ask the following questions about clothing purchases: *Did you purchase any pants, jeans, or shorts*? If the respondent said yes, a series of ancillary questions were asked such as: *Describe the item. Was this purchased for someone inside or outside of your household? For whom was this purchased? Enter name of person for whom it was purchased. Enter age/sex categories that apply to the purchase; How many did you purchase?; Enter number of identical items purchased; When did you purchase it?; How much did it cost?; Did this include sales tax?* (Dillman and House 2013, 84).

These questions were repeated for a series of items purchased in the reference month. As the reader can see, the question wording is stretching the limit of the survey tool by asking respondents to remember things with a level of precision that the human memory (in an interview setting) is not very well suited for (see also Eckman et al. 2014 for a discussion on this question wording). In fact, this specific question was an object of a redesign as the committee in charge described it: "This questionnaire structure creates […] *cognitive challenges"* (italics added) (Dillman and House 2013, 84).

We do not envision all behavioral questions being replaced by Big Data collection, but many could be, and there is already some work going in this direction (Sturgis, this volume). For example, Mastrandrea et al. (2015) compared diaries and surveys to wearable sensors and online social media to study social interactions among students in a high school in France. In another application of wearable sensors, Hitachi collected more than a million days' worth of data on employees' activities over the span of 9 years (Yano et al. 2015). The authors were able to correlate the sensor data with happiness measured via questionnaires.

Strengths and Challenges of Surveys and Big Data

Surveys have the advantage of being designed for the researchers to answer the question at hand. They also collect attitudes and opinion data which cannot be readily covered by Big Data. Challenges of surveys are encapsulated in the model of TSEs, and also by the size and coverage of survey data that, unlike few examples (census), are not meant to measure each single member of a particular population.

The most obvious advantage of Big Data collection is that it allows larger sample sizes that support more detailed analysis regarding space, time, and other

subgroups. Automated data is also better for measuring certain behaviors (e.g., avoiding recall bias), reducing respondent burden, and avoiding nonresponse bias in some settings. In addition, it can improve turnaround time and facilitate serendipitous findings about variables that no one thought to measure.

On the other hand, Big Data has important challenges. Researchers generally cannot choose what data are collected, or how to gather it. Second, much of Big Data generated is proprietary. Third, the availability of Big Data changes over time. For example, access to Twitter and Facebook changed over time concerning what could be downloaded, who could do it, and the extent of the data over a time period. Finally, as discussed before, Big Data come with Big Noise.

Privacy, Confidentiality, and Transfer of Data

Survey and market researchers have a long tradition and tools in place to handle collection, storage, and processing of survey data in order to guarantee the anonymity and confidentiality of the respondents (ESOMAR 2016; American Statistical Association 2016). Big Data are introducing new questions regarding the collection, storage, and transfer of personal information (Bander et al. 2016). For example, survey research organizations such as the University of Michigan Survey Research Center commonly use multiple databases to augment and enrich telephone and address based samples (Benson and Hubbard 2016). A more powerful example is what political campaigns can do by combining multiple databases and other sources starting from the voter registration databases that exist in each U.S. state. Already in 2008:

> Barack Obama's campaign began the year of his reelection fairly confident it knew the names of every one of the 69,456,897 Americans whose votes had put him in the White House. The votes may have been cast via secret ballot, but because Obama's analysts had come up with individual-level predictions, they could look at the Democrat's vote totals in each precinct and identify the people most likely to have backed him. (Issenberg 2012)

Four years later the Obama campaign created *Narwhal*, a software program that combined and merged data collected from multiple databases and financial sources. The Obama campaign began with a 10 Terabyte database that grew to 50 Terabytes by the end of the campaign (Nickerson and Rogers 2014). This accumulation of data using a census-like approach has privacy advocates worried describing it as "the largest unregulated assemblage of

personal data in contemporary American life" (Rubinstein 2014, 861; also Bennett 2013 for an international view).

If we think about the IOT or just about our smartphones, the implications for collecting, storing, and processing personal information are huge. For example, wearable activity bands and smart watches store a large amount of health and personal information that are transferred to apps and stored by the companies producing the devices. Questions such as: who owns these data, what happens to them when a company goes bust or is being acquired by another company? Do not have an easy answer. Privacy, ethical, and legal requirements for collecting, storing, and analyzing Big Data are questions that are here to stay (Lane et al. 2014).

Looking at the Future of Big Data and Surveys

The contemporary social researcher needs to look at Big Data as another source for insights together with surveys and other data collection methods. Unfortunately, at the time of this writing, there is little training available in survey and market research about Big Data. There are however some signs of growth such as the creation of the International Program in Survey and Data Science.[1]

What survey and market researchers can bring to the table is our ability to understand the research questions in greater depth. In the future, the answers to many research questions will not always come only from a survey or some qualitative data collection, but will be increasingly augmented and in some cases replaced by Big Data. The other main strengths that survey and market researchers can bring to the table are the concepts of TSE and BDTE. Shedding light on the limitations and challenges inherent in each data source is key to understanding a phenomenon and validly interpreting research findings. As we stated at the beginning of this chapter, Big Data does not necessarily imply high quality, and surveys can be used to check the quality of Big Data and vice versa.

We encourage survey researchers and practitioners to move the conversation from Big Data to *Rich Data*. We propose the term, rich data, to emphasize the importance of a mindset that focuses on not the mere size of data but their substance and utility. "Big" is never the end goal for research data collection. In fact, Big Data, when thoughtlessly collected and used, may lead to losses in both accuracy and efficiency (Poeppelman et al. 2013). Richness in the data,

[1] http://survey-data-science.net/

on the other hand, captures our methodological aims, namely to enhance and ensure the validity of the research conclusions and inferences as well as the utility of their applications. Specifically, richness means

- a comprehensive coverage of the constructs relevant to a research program.
- the inclusion of multiple complementary indicators that enable accurate and efficient quantification of the target constructs and their relationships.
- the application of appropriate tools to extract information from data, derive defensible and useful insights, and communicate them in compelling fashion.

The new and enhanced data sources and technologies discussed earlier in this chapter provide unprecedented opportunities for researchers and practitioners to improve the richness of their data – through tapping into hard to capture or previously not understood constructs, integrating a multitude of diverse signals (surveys, behavioral data, social media entries, etc.), and leveraging new analytic and visualization tools.

Examples include

- Using high-quality surveys to validate the quality of Big Data sources. This is the case of using surveys to validate the accuracy of voter registration records as reported by Berent et al. (2016).
- Using Big Data to ask better questions in surveys. Big Data can be used as validation data (true value) and different question wording can be tested to determine what is closer to the "true value." The idea is to extend the traditional validation data used in many medical studies such as physicians or nurse tests (e.g., Kenny Gibson et al. 2014) with validation data collected from wearables, or other IOT devices at scale.
- Augment Big Data with survey data such as the Google Local Guides.[2] This opt-in program asks its users to answer few "Yes, No, Not Sure" questions about locations such as restaurants, stores, or point of interest. For example, users can be asked if the restaurant they just visited is family friendly, or has Wi-Fi.

Big Data has opened the door for rich data. It is now time to move beyond the fixation on the size of data and take a more critical view of the new tools and opportunities to advance the science of measuring and influencing human thoughts, emotions, and actions.

[2] https://www.google.com/local/guides/

Acknowledgements We would like to thank Robert Bell (Google), Frauke Kreuter (University of Maryland) and David Vannette (Qualtrics) for their generous comments on the initial draft chapter.

References and Further Reading

American Statistical Association. 2016. "Committee on Privacy and Confidentiality." ASA. http://community.amstat.org/cpc/home.

Baker, Reginald P. 2017. "Big Data: A Survey Research Perspective." In *Total Survey Error: Improving Quality in the Era of Big Data*, edited by Paul P. Biemer, Edith De Leeuw, Stephanie Eckman, Brad Edwards, Frauke Kreuter, Lars Lyberg, Clyde Tucker, and Brady West, 47-70. Hoboken, NJ: Wiley.

Bakker, Bart F. M., Johan Van Rooijen, and Leo Van Toor. 2014. "The System of Social Statistical Datasets of Statistics Netherlands: An Integral Approach to the Production of Register-Based Social Statistics." *Statistical Journal of the IAOS* 30(4): 411–24. doi:10.3233/SJI-140803.

Bander, Stefan, Ron S. Jarmin, Frauke Kreuter, and Julia Lane. 2016. "Privacy and Confidentiality." In *Big Data and Social Science: A Practical Guide to Methods and Tools*, edited by Ian Foster, Rayid Ghani, Ron S. Jarmin, Frauke Kreuter, and Julia Lane, 299–311. Boca Raton, FL: CRC Press.

Bennett, Colin. 2013. "The Politics and the Privacy of Politics: Parties, Elections and Voter Surveillance in Western Democracies." *First Monday* 18(8). doi:10.5210/fm.v18i8.4789.

Benson, Grant, and Frost Hubbard. 2017. "Big Data Serving Survey Research: Experiences at the University of Michigan Survey Research Center." In *Total Survey Error: Improving Quality in the Era of Big Data*, edited by Paul P. Biemer, Edith De Leeuw, Stephanie Eckman, Brad Edwards, Frauke Kreuter, Lars Lyberg, Clyde Tucker, and Brady West, 478-486. Hoboken, NJ: Wiley.

Berent, Matthew K., Jon A. Krosnick, and Arthur Lupia. 2016. "Measuring Voter Registration and Turnout in Surveys. Do Official Government Records Yield More Accurate Assessments?." *Public Opinion Quarterly*, advance access. doi:10.1093/poq/nfw021.

Beyer, Mark A., and Douglas Laney. 2012. *The Importance of "Big Data": A Definition*. G00235055. Stamford, CT: Gartner.

Biemer, Paul P. 2010. "Total Survey Error: Design, Implementation, and Evaluation." *Public Opinion Quarterly* 74(5): 817–48. doi:10.1093/poq/nfq058.

Biemer, Paul P. 2014. "Dropping the 's' from TSE: Applying the Paradigm to Big Data." Paper presented at the 2014 International Total Survey Error Workshop

(ITSEW 2014), Washington, DC: National Institute of Statistical Science. https://www.niss.org/sites/default/files/biemer_ITSEW2014_Presentation.pdf.

Biemer, Paul P. 2016. "Errors and Inference." In *Big Data and Social Science: A Practical Guide to Methods and Tools*, edited by Ian Foster, Rayid Ghani, Ron S. Jarmin, Frauke Kreuter, and Julia Lane, 265–97. Boca Raton, FL: CRC Press.

Bort, Julie. 2015. "How ditching law school and quitting a bunch of good jobs led Dave Goldberg to tech fame and fortune – Business Insider." April 19. http://uk.businessinsider.com/the-incredible-career-of-david-goldberg-2015-4.

Callegaro, Mario. 2013. "Paradata in Web Surveys." In *Improving Surveys with Paradata: Analytic Use of Process Information*, edited by Frauke Kreuter, 261–79. Hoboken, NJ: Wiley.

Chamberlin, Graeme. 2010. "Googling the Present." *Economic & Labour Market Review* 4(12): 59–95. doi:10.1057/elmr.2010.166.

Chang, Robert. 2015. "Doing Data Science at Twitter: A Reflection of My Two Year Journey So Far. Sample Size N = 1." *Medium*. June 20. https://medium.com/@rchang/my-two-year-journey-as-a-data-scientist-at-twitter-f0c13298aee6#.t9wiz09mt.

Choi, Hyunyoung, and Hal Varian. 2012. "Predicting the Present with Google Trends." *Economic Record* 88(S1): 2–9. doi:10.1111/j.1475-4932.2012.00809.x.

Committee on Commerce, Science and Transportation. 2013. "A Review of the Data Broker Industry: Collection, Use, and Sale of Consumer Data for Marketing Purposes." United States Senate. http://educationnewyork.com/files/rockefeller_databroker.pdf.

Couper, Mick P. 2013. "Is the Sky Falling? New Technology, Changing Media, and the Future of Surveys." *Survey Research Methods* 7(3): 145–56. doi:http://dx.doi.org/10.18148/srm/2013.v7i3.5751.

Deming, Edward W. 1944. "On Errors in Surveys." *American Sociological Review* 9(4): 359–69.

Dillman, Don A., and Carol C. House. eds. 2013. *Measuring What We Spend: Toward a New Consumer Expenditure Survey. Panel on Redesigning the BLS Consumer Expenditure Surveys*. Washington, DC: National Academies Press.

Dutcher, Jennifer. 2014. "What Is Big Data? – Blog." September 3. https://datascience.berkeley.edu/what-is-big-data/.

Eckman, Stephanie, Frauke Kreuter, Antje Kirchner, Annette Jäckle, Roger Tourangeau, and Stanley Presser. 2014. "Assessing the Mechanisms of Misreporting to Filter Questions in Surveys." *Public Opinion Quarterly* 78(3): 721–33. doi:10.1093/poq/nfu030.

Edjlali, Roxanne, and Ted Friedman. 2011. Data Quality for Big Data: Principles Remain, But Tactics Change. G00224661. Stanford, CT: Gartner.

ESOMAR 2016. "ESOMAR Data Protection Checklist." ESOMAR. https://www.esomar.org/knowledge-and-standards/research-resources/data-protection-checklist.php.

ESOMAR. 2015. "Global Market Research 2015." ESOMAR.

Ferguson, Mike. 2014. "Big Data – Why Transaction Data Is Mission Critical to Success." Intelligence Business Strategies Limited. https://public.dhe.ibm.com/common/ssi/ecm/im/en/iml14442usen/IML14442USEN.PDF.

Foster, Ian, Rayid Ghani, Ron S. Jarmin, Frauke Kreuter, and Julia Lane. eds. 2016. *Big Data and Social Science: A Practical Guide to Methods and Tools*. Boca Raton, FL: CRC Press.

Gershenfeld, Neil, Raffi Krikorain, and Danny Choen. 2004. "The Internet of Things." *Scientific American* 291(4): 76–81. doi:10.1038/scientificamerican1004-76.

González-Bailón, Sandra, and Georgios Paltoglou. 2015. "Signals of Public Opinion in Online Communication. A Comparison of Methods and Data Sources." *The ANNALS of the American Academy of Political and Social Science* 659(1): 95–107. doi:10.1177/0002716215569192.

Gray, Emily, Will Jennings, Stephen Farrall, and Colin Hay. 2015. "Small Big Data: Using Multiple Data-Sets to Explore Unfolding Social and Economic Change." *Big Data & Society* 2(1). doi:10.1177/2053951715589418.

Groves, Robert M. 2011. "Three Eras of Survey Research." *Public Opinion Quarterly* 75(5): 861–71. doi:10.1093/poq/nfr057.

Hewson, Claire, Carl Vogel, and Dianna Laurent. 2016. *Internet Research Methods*. 2nd ed. London: Sage.

Hsinchun, Chen, Roger H. L. Chiang, and Veda C. Storey. 2012. "Business Intelligence and Analytics: From Big Data to Big Impact." *Mis Quarterly* 36(4): 1165–88.

Issenberg, Sasha. 2012. "How Obama's Team Used Big Data to Rally Voters. How President Obama's Campaign Used Big Data to Rally Individual Voters." *MIT Technology Review*. https://www.technologyreview.com/s/509026/how-obamas-team-used-big-data-to-rally-voters/.

Japec, Lilli, Frauke Kreuter, Marcus Berg, Paul P. Biemer, Paul Decker, Cliff Lampe, Julia Lane, Cathy O'Neil, and Abe Usher. 2015. "Big Data in Survey Research. AAPOR Task Force Report." *Public Opinion Quarterly* 79(4): 839–80. doi:10.1093/poq/nfv039.

Jungherr, Andreas, Harald Schoen, Oliver Posegga, and Pascal Jürgens. 2016. "Digital Trace Data in the Study of Public Opinion an Indicator of Attention Toward Politics Rather Than Political Support." *Social Science Computer Review* doi:10.1177/0894439316631043.

Kenny, Gibson, William, Hilary Cronin, Rose Anne Kenny, and Annalisa Setti. 2014. "Validation of the Self-Reported Hearing Questions in the Irish Longitudinal Study on Ageing Against the Whispered Voice Test." *BMC Research Notes* 7(361). doi:10.1186/1756-0500-7-361.

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 Benden, and Helen Nissenbaum, 257–75. New York: Cambridge University Press.

Lane, Julia, Victoria Stodden, Stefan Benden, and Helen Nissenbaum. eds. 2014. *Privacy, Big Data, and the Public Good*. New York: Cambridge University Press.

Martin, Jolie M. 2016. "Combining 'small Data' from Surveys and 'big Data' from Online Experiments at Pinterest." In *ALLDATA* 2016: The Second International Conference on Big Data, Small Data, Linked Data and Open Data (includes KESA *2016)*, edited by Venkat Gudivada, Dumitru Roman, Pia Di Buono Maria, and Mario Monteleone, 33–34. Lisbon: IARA. http://toc.proceedings.com/29767webtoc.pdf.

Mastrandrea, Rossana, Julie Fournet, and Alain Barrat. 2015. "Contact Patterns in a High School: A Comparison Between Data Collected Using Wearable Sensors, Contact Diaries and Friendship Surveys." *PloS One* 10(9): e0136497. doi:10.1371/journal.pone.0136497.

Müller, Hendrick, and Aaron Sedley. 2014. "HaTS: Large-Scale in-Product Measurement of User Attitudes & Experiences with Happiness Tracking Surveys." In *Proceedings of the 26th Australian Computer-Human Interaction Conference (OzCHI 2014)*, 308–15. New York, NY: ACM.

Nickerson, David W., and Todd Rogers. 2014. "Political Campaigns and Big Data." *Journal of Economic Perspectives* 28(2): 51–74. doi:10.1257/jep.28.2.51.

Poeppelman, Tiffany, Nikki Blacksmith, and Yongwei Yang. 2013. "'Big Data' Technologies: Problem or Solution?." *The Industrial-Organizational Psychologist* 51(2): 119–26.

Poynter, Ray. 2014. "No More Surveys in 16 Years? NewMR." August 27. http://newmr.org/blog/no-more-surveys-in-16-years/.

Rubinstein, Ira S. 2014. "Voter Privacy in the Age of Big Data." *Wisconsin Law Review* 2014(5): 861–936.

Schober, Michael F., Josh Pasek, Lauren Guggenheim, Cliff Lampe, and Frederick G. Conrad. 2016. "Social Media Analyses for Social Measurement." *Public Opinion Quarterly* 80(1): 180–211. doi:10.1093/poq/nfv048.

Scott, Steve, and Hal R. Varian. 2015. "Bayesian Variable Selection for Nowcasting Economic Time Series." In *Economic Analysis of the Digital Economy*, edited by Avi Goldfarb, Shane M. Greenstein, and Katherine E. Tucker, 119–35. Chicago, IL: Chicago University Press.

Statistics Finland. 2004. *Use of Registers and Administrative Data Sources for Statistical Purposes. Best Practices of Statistics Finland*. Helsinki: Statistics Finland.

Stephens-Davidowitz, Seth, and Hal Varian. 2015. "A Hands-on Guide to Google Data." http://people.ischool.berkeley.edu/~hal/Papers/2015/primer.pdf.

Swan, Melanie. 2013. "The Quantified Self: Fundamental Disruption in Big Data Science and Biological Discovery." *Big Data* 1(2): 85–99. doi:10.1089/big.2012.0002.

Tumasjan, Andranik, Timm O. Sprenger, Philipp G. Sandner, and Isabell M. Welpe. 2010. "Predicting Elections with Twitter: What 140 Characters Reveal About Political Sentiment." In *Fourth International AAAI Conference on Weblogs and Social Media*. http://www.aaai.org/ocs/index.php/ICWSM/ICWSM10/paper/view/1441.

Waldherr, Annie, Daniel Maier, Peter Miltner, and Enrico Günther. 2016. "Big Data, Big Noise. The Challenge of Finding Issue Networks on the Web." *Social Science Computer Review*, advance access. doi:10.1177/0894439316643050.

Wallgren, Andrew, and Britt Wallgren. 2014. *Register-Based Statistics: Statistical Methods for Administrative Data*. 2nd ed. Chichester, UK: Wiley.

Warshaw, Christopher. 2016. "The Application of Big Data in Surveys to the Study of Public Opinion, Elections, and Representation." In *Computational Social Science. Discovery and Prediction*, edited by R. Michael Alvarez, 27–50. New York: Cambridge University Press.

Yano, Kazuo, Tomoaki Akitomi, Koji Ara, Junichiro Watanabe, Satomi Tsuji, Nabuo Sato, Miki Hayakawa, and Norihko Moriwaki. 2015. "Measuring Happiness Using Wearable Technology. Technology for Boosting Productivity in Knowledge Work and Service Businesses." *Hitachi Review* 64(8): 517–24.

Zhang, Chen, and Si Chen. 2016. "News Feed FYI: Using Qualitative Feedback to Show Relevant Stories." February 1. http://newsroom.fb.com/news/2016/02/news-feed-fyi-using-qualitative-feedback-to-show-relevant-stories/.

Mario Callegaro is Senior Scientist at Google, London, in the User Research and Insights team, Brand Studio. He focuses on measuring brand perception and users' feedback. Mario consults on numerous survey and market research projects.

Mario holds a MS and a PhD in Survey Research and Methodology from the University of Nebraska, Lincoln.

Prior to joining Google, Mario was working as survey research scientist for Gfk-Knowledge Networks. He is associate editor of *Survey Research Methods* and in the advisory board of the *International Journal of Market Research.*

Mario has published numerous books, book chapters, and presented at international conferences on survey methodology and data collection methods.

He published (May 2014) an edited book with Wiley titled *Online Panel Research: A Data Quality Perspective,* and his new book coauthored with Katja Lozar Manfreda and Vasja Vehovar: *Web Survey Methodology* is available from Sage as of June of 2015.

Yongwei Yang is a Research Scientist at Google. He works on brand and user research, as well as general research on measurement and survey methodology. Yongwei enjoys figuring out how to collect better data and make better use of data, as well as to implement evidence-based interventions and evaluate their business

impact. His research interests include survey and test development and validation, technology-enhanced measurement and data collection, survey and testing in multi-population settings, concepts and models for understanding consumer and employee behaviors, and utility analysis of organizational interventions. He holds a PhD in Quantitative and Psychometric Methods from the University of Nebraska-Lincoln.