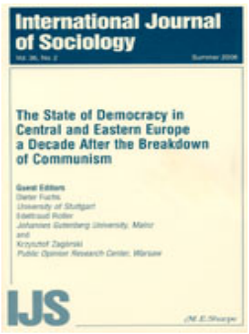

International Journal of Sociology

ISSN: 0020-7659 (Print) 1557-9336 (Online) Journal homepage: http://www.tandfonline.com/loi/mijs20

Political Behavior and Big Data

J. Craig Jenkins, Kazimierz M. Slomczynski & Joshua Kjerulf Dubrow

To cite this article: J. Craig Jenkins, Kazimierz M. Slomczynski & Joshua Kjerulf Dubrow (2016) Political Behavior and Big Data, International Journal of Sociology, 46:1, 1-7

To link to this article: http://dx.doi.org/10.1080/00207659.2016.1130409

International Journal of Sociology, 46: 1–7, 2016

ISSN: 0020-7659 print/1557-9336 online
DOI: 10.1080/00207659.2016.1130409

Routledge
Taylor & Francis Group

GUEST EDITORS' INTRODUCTION

Political Behavior and Big Data

J. Craig Jenkins
Department of Sociology, Ohio State University
Kazimierz M. Slomczynski and Joshua Kjerulf Dubrow
Institute of Philosophy and Sociology, Polish Academy of Sciences

Interest in the use of "big data" in the social sciences is growing dramatically. Yet, adequate methodological research on what constitutes such data, and about their validity, is lacking. Scholars face both opportunities and challenges inherent in this new era of unprecedented quantification of information, including that related to political actions and attitudes. This special issue of the International Journal of Sociology addresses recent uses of "big data," its multiple meanings, and the potential that this may have in building a stronger understanding of political behavior. We present a working definition of "big data" and summarize the major issues involved in their use. While the

J. Craig Jenkins is a professor of sociology, political science, and environmental science at Ohio State University. He has received a Fulbright Professorship at the Peace Research Institute of Oslo (PRIO), a Leiv Eiriksson Mobility Fellowship from the Norway Research Council, and the Robin M. Williams Jr. Award for Distinguished Contributions to Scholarship, Teaching and Service from the Section on Peace, War and Social Conflict of the American Sociological Association. He has written or edited 4 books and over 100 research articles addressing social movements, protest theory, event data methods, and political economy.

Kazimierz M. Slomczynski is a professor at the Institute of Philosophy and Sociology at the Polish Academy of Sciences. He is also director of the Cross-National Studies: Interdisciplinary Research and Training program (CONSIRT) of Ohio State University and the Polish Academy of Sciences. He is principal investigator of the Polish Panel Survey POLPAN, conducted every five years since 1988, and directs other studies, including a project on ex post harmonization of cross-national surveys. He has published numerous articles in various international scientific journals and authored or coauthored several books in Polish and English. His main interest is in social stratification and methodology of social sciences.

Joshua Kjerulf Dubrow received his Ph.D. from Ohio State University and is an associate professor at the Institute of Philosophy and Sociology, Polish Academy of Sciences, and Program Coordinator for CONSIRT of Ohio State University and the Polish Academy of Sciences. He is on the Executive Board of the Committee on Political Sociology (ISA and IPSA). His research on interdisciplinarity has appeared in *Sociologias*, the *American Sociologist*, and *Quality and Quantity*.

Address correspondence to J. Craig Jenkins, Department of Sociology, Ohio State University, 238 Townshend Hall, 1885 Neil Ave. Mall, Columbus, OH 43210. Email: jenkins.12@osu.edu.

papers in this volume deal with various problems – how to integrate "big data" sources with cross-national survey research, the methodological challenges involved in building cross-national longitudinal network data of country memberships in international nongovernmental organizations, methods of detecting and correcting for source selection bias in event data derived from news and other online sources, the challenges and solutions to ex post harmonization of international social survey data – they share a common viewpoint. To make good on the substantive promise of "big data," scholars need to engage with their inherent methodological problems. At this date, scholars are only beginning to identify and solve them.

Keywords big data; harmonization; methodology; political behavior; social sciences

The wealth of quantitative data—including data from cross-national survey projects, official governmental and nongovernmental organization (NGO) statistics, newspapers and electronic newswires, and a variety of Internet-based websites, blogs, and social media sites—has generated a large and growing empirically based literature on political behavior. Yet, social scientists have only begun to use this wealth to its fullest capacity, as advances in computing infrastructures, methods, and Internet communication technologies create new opportunities for developing and integrating diverse types of information into social science data. Social science faces the challenge of "big data," a new era of the quantification and analysis of political behavior on an unprecedented scope and scale. Will it rise to this challenge?

BIG DATA AND THE SOCIAL SCIENCES

"Big data" has become a buzzword of no common definition (boyd and Crawford 2012; Mayer-Schönberger and Cukier 2014). In general, big data refers to any data set that has an unusually large number of cases and is composed of a diversity of sources. The number of variables can be very small, there is no limit on time or space coverage, and there is no specification of how diverse "diversity of sources" should be.[1]

In academia, big data research can be found almost everywhere, but the bulk of it is located in a handful of disciplines and published within the past few years. Using the Web of Science database, we searched for "big data" as a topic in any academic product, from the beginning to 2015. This returns 4,083 products, including 1,580 articles. One can find big data academic products almost everywhere, including business and management, health and medicine, library science, statistics, economics, psychology, physics, geography, law, education, and philosophy, among others. The majority of these products are in computer science, engineering, and telecommunications. As Figure 1 shows, the majority of big data academic products are from 2011 to the present. In sociology and political science, 88 percent were published since 2014.[2] In 2015 alone, two major social science journals, *PS: Political Science and Politics* (vol. 48, no. 1) and *Annals of the American Academy of Political and Social Science* (vol. 659, no. 1), published special sections on big data. As of this writing, neither the *American Sociological Review* nor the *American Political Science Review* has published an article on the implications of big data for sociology or political science.

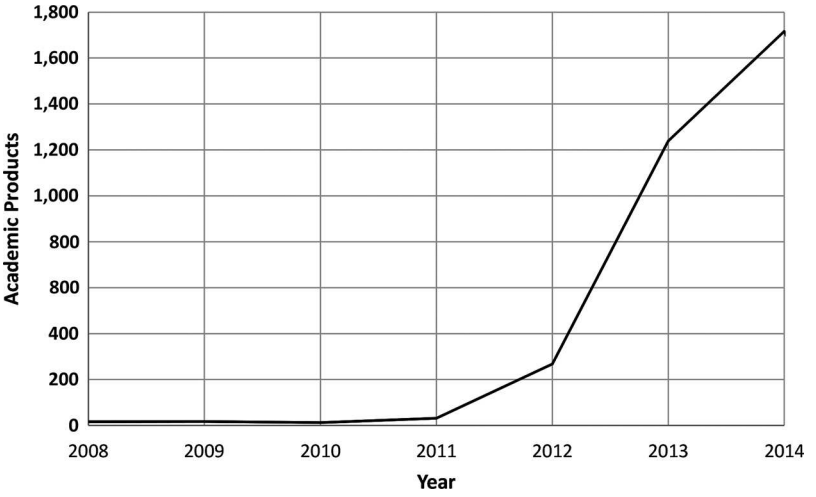

FIGURE 1 "Big Data" as a Web of Science topic in all academic products, 2008–2014. *Note*: Academic products include article, book, book review, book chapter, editorial material, proceedings paper, review, and letters.

METHODOLOGICAL IMPLICATIONS OF BIG DATA

A diverse set of big data proponents means a diverse set of approaches. The book by Mayer-Schönberger and Cukier (2014), *Big Data: A Revolution That Will Transform How We Live, Work, and Think*, an Amazon bestseller, is an attempt to both cheerlead and set forth principles of big data (some of these are also found in Lin 2015). They identify three main principles: $n =$ all, messiness, and correlations. Big data proponents argue that the era of collecting samples to obtain representativeness will, in a big data world, give way to the idea of collecting entire populations, that is, moving from "$n =$some" to "$n =$ all." Bulk data collection in an "$n =$ all" world produces an unprecedented scope and diversity of error in the data. Mayer-Schönberger and Cukier (2014) argue that this error, which they call "messiness," is not a problem. As they write:

> In dealing with ever more comprehensive datasets, which capture not just a small sliver of the phenomenon at hand but much more or all of it, we no longer need to worry so much about individual data points biasing the overall analysis. Rather than aiming to stamp out every bit of inexactitude at increasingly high cost, we are calculating with messiness in mind. (Mayer-Schönberger and Cukier 2014: 39–40)

They advocate for the search for correlations rather than causation (55). This approach to big data requires data mining and, hence, the end of hypothesis testing: "No longer do we necessarily require a valid substantive hypothesis about a phenomenon to begin to understand our world." (ibid.). Correlations, they argue, will take the place of theoretically driven hypotheses (ibid.). Behind all this is the argument that looking for correlations is cheaper and faster than searching for cause: correlations are easy, causation is hard.

As research on big data in sociology and political science is so new—the vast majority of publications appeared in the past year and a half—there is no discernible trend in how social science

talks about, or uses, big data. Without much empirical work to draw upon, sociology and political science tend to speculate on what big data means for their academic disciplines and for society in general. Hesse, Moser, and Riley (2015: 18) argue that big data, with its peculiar demands of quantitative and technological skills, will cause all social scientists to question the foundations of their methodology and how they produce knowledge. Boyd and Crawford (2012) are concerned with the ethical implications of an $n =$ all approach on the issues of privacy, consent, and anonymity.

Some social scientists who study political behavior see an advantage in more and bigger data. Nagler and Tucker, with some reservations, "argue that having more data is an opportunity for, not a constraint on, testing theories of political behavior" (2015: 84). The biggest big data source for studying political behavior comes from the Internet and, at this early date, two of the most prominent are search engines (e.g., Google) and social media. Articles on the substantive promise and methodological perils of social media point to problems of sample bias—Twitter users are not evenly distributed in the population, and thus are not fully representative—and the inherent ambiguities of their data. These ambiguities stem primarily from Silicon Valley companies' refusal to provide academics with the necessary information to understand the who, what, when, and where of their data (Hargittai 2015; Nagler and Tucker 2015: 85–88). An issue that has received little attention concerns how to integrate existing social science data that have a known range of reliability with big data sources—such as social media—that have an unknown range of reliability. Social media and the Internet have become main ways of communicating and acting politically; their integration with existing social science data has become inevitable.

The extent to which big data opportunities constitute a "disruptive innovation" that requires a paradigm shift is debatable, according to Kitchen:

> it seems likely that the data-driven approach will eventually win out and over time, as Big Data becomes more common and new data analytics are advanced, will present a strong challenge to the established knowledge-driven scientific method. To accompany such a transformation the philosophical underpinnings of data-driven science, with respect to its epistemological tenets, principles and methodology, need to be worked through and debated to provide a robust theoretical framework for the new paradigm. (Kitchen 2014: 10)

In this new paradigm the issue of significant statistical results and randomness should be appropriately addressed. As Efron (2010) explains, it is easy to go wrong with huge data sets and thousands of questions to answer at once. When the number of variables grows, the number of meaningless correlations also grows (Taleb 2012). Thus, much research is needed with theoretical and practical analysis to provide useful methods to contend with data that are huge, diversified, and evolving. In addition, the issue of the quality and suitability of data becomes more and more important. Bigger data are not always better data. It depends on whether the data are noisy or not, and whether they are representative of what we are looking for. Thus, if big data requires a new paradigm for scientific research, we must create innovative solutions to the innovative problems that these data create.

ARTICLES IN THIS ISSUE

This issue of the *International Journal of Sociology* is based on discussions initiated in "Interdisciplinary Studies of Political Behavior and the Use of 'Big Data,'" a conference and

workshop in May 2013, held at the Mershon Center for International Security Studies, the Ohio State University and co-organized by Cross-National Studies: Interdisciplinary Research and Training Program, the Ohio State University and the Polish Academy of Sciences (CONSIRT). The articles published in this issue come from political sociology, cross-national methodology, and computer science, and focus on the methodology of analyzing political behavior.

The opening article by Russell J. Dalton, "The Potential of Big Data for the Cross-National Study of Political Behavior," serves as a detailed introduction to the social science perspective of big data hinted at in this introduction. Dalton asks a provocative question: "If Google can effectively predict what article we are most likely to read and Amazon can predict what product we want to buy, can these or comparable tools be used to describe (and explain) the political behavior of contemporary publics with a causal understanding of their behavior?" (p. 9, in this issue). Dalton's article suggests that the answer is in the uses and the methodology of big data. From a business perspective, à la Mayer-Schönberger and Cukier (2014), a big, messy data set full of errors and lots of correlations, if it provides profitable results, is not only good enough, but it may be more useful than a painfully constructed statistically representative data set. Pragmatic results are what matters. Academics have a different set of concerns: they want to reduce errors and test theories with causal mechanisms. Big data has great potential for improving on what we know about political behavior. Academia's contribution is not to follow the lead of businesses, but to set scientifically rigorous standards for the methodological and theoretical approach to big data.

The next three articles, each with a different substantive problem, contend with an academic approach to the methodology of big data. The first of these, "Building Cross-National, Longitudinal Data Sets: Issues and Strategies for Implementation," by Nicholas E. Reith, Pamela Paxton, and Melanie M. Hughes, is about the methodological challenges involved in building a cross-national, longitudinal, network data set of country memberships in international nongovernmental organizations (INGO) that spans from 1950 to 2008. Using the Union of International Associations *Yearbook of International Organizations*, and funded by a National Science Foundation grant, they started with 17,000 INGO names. Through a rigorous process of matching and identifying, they whittled it down to 5,200 unique organizations. Their article provides a step-by-step guide to solving the challenge of creating a large numerical data set from textual data that can be applied to many substantive areas.

Turning news text into numerical event data that can describe contentious politics that can later be merged with other cross-national data sets fulfills the promise of big data diversity as Reith et al. have described. In "What Should We Do about Source Selection in Event Data? Challenges, Progress, and Possible Solutions," J. Craig Jenkins and Thomas V. Maher argue that the Internet and the growing availability of online news archives, blogs, activist Web sites, and the like, coupled with the development of automated methods of reliably coding large amounts of text into event data, promises to create a rich new basis for constructing event data that can be used for conflict monitoring, political forecasting, and scientific analysis. But before this big data possibility will be fully accepted in the scientific community, the problem of representativeness and potential selection bias needs to be addressed. By its nature, event data cannot meet conventional standards of representativeness in the sense that the full universe of real world events can never be fully identified. What is the implication of this? Drawing on the logic of "capture/recapture" and a relative

inference strategy for assessing source selection, Jenkins and Maher outline a set of methods that can be developed to assess and control for the risk of selection bias. Without accepting Mayer-Schönberger and Cukier (2014), they argue for a relative standard of representativeness that allows us to gauge improvements in coverage, comprehensiveness, and reliability and incorporate big data information into conventional scientific discussions of causality and inference.

In "Harmonization of Cross-National Survey Projects on Political Behavior: Developing the Analytic Framework of Survey Data Recycling," Irina Tomescu-Dubrow and Kazimierz M. Slomczynski refer to the concept of big data in a specific manner: basic units of observation (from different projects) in millions, and variables of interest (from different sources) in hundreds, without a reference as to whether researchers produce data or "just gather/combine" them. They describe the challenges and solutions to ex post harmonization of survey data from 22 well-known international survey projects into a data set of 2.3 million respondents, covering a total of 142 countries and territories, and spanning almost 50 years, to construct common measures of political behavior, social attitudes, and demographics. Substantively, this project engages with the relationship between democracy and protest behavior in comparative perspective, by proposing a theoretical model that explains variation in political protest through individual-level characteristics, country-level determinants, and interactions between the two. The work of the authors' team became the springboard for the analytic framework of Survey Data Recycling (SDR) that facilitates the reuse of information from extant cross-national survey projects. The goal of the SDR approach is to minimize the "messiness" of data built into original surveys, expand the range of possible comparisons over time and across countries, and improve confidence in substantive results. A key aspect of the proposed framework deals with control variables for the quality of the source data and ex post harmonization controls, facilitating validity and reliability assessments of the target variables.

Przemek Powałko and Marta Kołczyńska, working on the harmonization project described by Tomescu-Dubrow and Slomczynski, have a difficult interdisciplinary job: they must build the digital bridge over the divide that separates social scientists interested in big data with the computer science needed to work with it. In their article, "Working with Data in the Cross-National Survey Harmonization Project: Outline of Programming Decisions," Powałko and Kołczyńska describe the harmonization process with regard to data access and storage. To deal with the large and diverse data, and seeking replicable, nonproprietary tools to work with, Powałko and colleagues created an environment based on freeware and open-source software. In plain language, they describe the process and the computing environment so that future researchers can use, critique, and build upon it.

Our purpose for this issue of the *International Journal of Sociology* is to meaningfully join the upsurge in popularity of big data in the social sciences by calling for more methodological research. An upsurge in popularity is not necessarily a revolution. As the authors reveal, to make good on the substantive promise of big data we would first need to solve big data's inherent methodological problems. At this early date, we are only now beginning to identify and solve these problems. We hope future researchers will heed the warnings, add to the growing body of methodological studies, and chart their research course with care.

FUNDING

This project was conducted by the Institute of Philosophy and Sociology of the Polish Academy of Sciences (IFiS PAN) and the Mershon Center for International Security Studies at Ohio State University (MC OSU), with funding from the National Science Centre, Poland (grant number 2012/06/M/HS6/00322). Cross-national Studies: Interdisciplinary Research and Training program (consirt.osu.edu) hosts the project.

NOTES

1. Some argue that the term is misleading. In years past, the average social scientists' computer could not handle the size of large data sets. Today, data that were once considered too large are more efficiently compressed and off-the-shelf computers can make extensive use of them. Due to technological advances, data that were once considered big are big no longer (boyd and Crawford 2012: 663). This trend will continue.

2. We used the Web of Science database, with "sociology" and "political science" as research areas. In 2013, there were two articles with "big data" as the article topic. In 2014, there were 27, and halfway through 2015, there were 29.

REFERENCES

boyd, danah, and Kate Crawford. 2012. "Critical Questions for Big Data." *Information, Communication and Society* 15(5):662–79.

Efron, Bradley. 2010. *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*. New York: Cambridge University Press.

Hargittai, Eszter. 2015. "Is Bigger Always Better? Potential Biases of Big Data Derived from Social Network Sites." *Annals of the American Academy of Political and Social Science* 659(1):63–76.

Hesse, Bradford W., Richard P. Moser, and William T. Riley 2015. "From Big Data to Knowledge in the Social Sciences." *Annals of the American Academy of Political and Social Science* 659(1):16–32.

Kitchin, Rob. 2014. "Big Data, New Epistemologies and Paradigm Shifts." *Big Data and Society* April–June:1–12.

Lin, Jimmy. 2015. "On Building Better Mousetraps and Understanding the Human Condition: Reflections on Big Data in the Social Sciences." *Annals of the American Academy of Political and Social Science* 659(1):33–47.

Mayer-Schönberger, Viktorm and Kenneth Cukier. 2014. *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. New York: Mariner Books.

Nagler, Jonathan and Joshua A. Tucker. 2015. "Drawing Inferences and Testing Theories with Big Data." *PS: Political Science and Politics* 48(1):84–88.

Taleb, Nassim N. 2012. *Antifragile: How to Live in a World We Don't Understand*. London: Penguin Books.