3b2 Version Number : 11.0.3184/W Unicode (Apr 10 2014)
File path : P:/Santype(JATS)/Journals/TandF_Production/MIJS/v46n1/MIJS1130410/MIJS_A_1130410_J.3d
Date and Time : 7/1/16 and 16:39

International Journal of Sociology, 46: 1–13, 2016

ISSN: 0020-7659 print/1557-9336 online
DOI: 10.1080/00207659.2016.1130410

Routledge
Taylor & Francis Group

The Potential of "Big Data" for the Cross-National Study of Political Behavior

Russell Dalton

University of California, Irvine

The U.S. presidential elections of 2008 and 2012 were a coming out party for Big Data applications in electoral studies. The Obama campaign developed a model of identifying individual voters that guided their campaign strategy to a successful outcome. This essay considers whether this model is exportable to elections and other aspects of political behavior in cross-national research. There is uneven development to date, but a growing awareness of the potential of Big Data electoral campaigns. More broadly, the essay discusses the current potential and limits of Big Data for the cross-national study of political behavior. The potential exceeds the actual applications, and there are major challenges for academic, theory-testing research using Big Data methods. It is unclear whether Big Data can successfully address these challenges.

Keywords big data; cross-national and comparative survey research

One of the most dramatic methodological changes in political behavioral research during the past several decades has been the explosion of empirical resources available to scholars. The landmark study in American voting behavior, *The American Voter* (Campbell et al. 1960) was based on two surveys of the 1952 and 1956 electorates. The seminal work on political participation in the United States, *Political Participation in America* (Verba and Nie 1972) was similarly based on a single national opinion survey. The landmark political behavior studies in many countries began with a single, nationally representative, high quality, academic public opinion survey. Almond and Verba's (1963) cross-national analyses in *The Civic Culture* were considered a major advance because they compared citizens from five countries.

Today, the empirically oriented researcher can virtually carry a sampling of the world's population around on a flash drive with a copy of the World Values Survey (WVS) or the collection of regional barometer studies. Large international projects in sociology (International Social Survey Program [ISSP] and European Social Survey [ESS]) or political science

Russell J. Dalton is a research professor at the Center for the Study of Democracy at the University of California Irvine. He has received a Fulbright Professorship at the University of Mannheim, a Barbra Streisand Center fellowship, German Marshall Research Fellowship, and a POSCO Fellowship at the East/West Center. He has written or edited over 20 books and 150 research articles that reflect his scholarly interests in comparative political behavior, political parties, social movements, and empirical democratic theory.

Address correspondence to Russell Dalton, University of California, 3151 Social Science Plaza, Irvine, CA 92697-5100. E-mail: rdalton@uci.edu

(Comparative Study of Electoral Systems [CSES]) expand our database on individuals and their actions. The cross-national breadth and longitudinal reach of our contemporary data sources were probably nearly unimaginable to the political behavior scholars of the 1960s–1970s—or even to scholars more recently.

The most recent potential advance is linked to the possibility that Big Data can provide even more evidence on the actions of the public and the context in which these actions are taken. For example, several organizations have developed databases from voter registration lists that include virtually every registered U.S. voter. Election consulting firms merge official voting records with other information from government databases, economic data from commercial vendors, and various social data. These Big Data methods are used to predict each individual's likelihood of voting and even his/her probable voting choice (Hersh 2015). An article in the *Harvard Business Review* thus proclaimed "2012: The First Big Data Election" (Hellweg 2012; also Hersh 2015). 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?

There are several possible definitions of Big Data and the essays in this collection show there is not a consensual definition. So, I will briefly describe my understanding before continuing. To me Big Data does not mean just a data collection with a very large *N*. Government agencies and research projects have been collecting massive amounts of social science data for decades. These projects confront many of the challenges noted in the article by Reith, Paxton, and Hughes in this collection, and there are several large data collection projects related to the study of political behavior. Below I discuss some of the problems such aggregation efforts face to complement other works in this issue.

Rather, Big Data typically involves the collection of indirect evidence of massive size from diverse and often unconventional sources (Jenkins et al. in this collection). The data were collected for other purposes, such as government recording requirements or business activities, and are repurposed in a Big Data collection. It is the merging of distinct data—voter records, economic records, online activity, and other—that produces a massive database extending beyond any single component. Such data are often unstructured or require structure before use by conventional analysis methods, or might call for the application of new analytic methods of data processing. Another common trait is that Big Data can often cover an entire population of interest—all voters, or all Twitter users—rather than a sample of the population. Thus, a very large survey of citizen participation cross-nationally or aggregating several surveys is not Big Data in my interpretation. Merging voting records from the Registrars of Voters (or commercial aggregators) with data on each voter's consumer activity, social media/online information, contributions to political campaigns, and perhaps surveying subsets of this population, is, however, Big Data.

This essay considers the potential value of Big Data to academic social science from the perspective of someone who has been involved with the collection and analysis of large cross-national public opinion surveys. What are the challenges that Big Data techniques face in studying the political behavior of individual citizens? What are the opportunities that these developments offer? Like any exploration into new territory we will find things we do not expect, and not find things we expect. But if we begin with clearer expectations our search might be more fruitful.

WHAT DO WE WANT TO EXPLAIN?

The first question we confront when considering Big Data options is to determine what we want to explain. To me it seems that studying political behavior is more complicated than explaining consumer behavior, which has been a productive field for Big Data (Mayer-Schönberger and Cukier 2014). This section describes past advances in this research, and some of the questions that arise depending on what we want to explain.

Voting and Political Participation

The potential value of Big Data in explaining political behavior received a large boost from the experiences of recent U.S. elections (Hersh 2015; Issenberg 2012). In 2012 the Obama campaign turned to the consulting firm, Catalyst, to use its massive database of millions of individuals that merged voter records for past elections (and information on past turnout), campaign finance records, and hundreds of variables from commercial sources (Hersh 2015). Figure 1 provides a simplified overview of the Big Data logic for the 2012 Obama campaign. In the first step, Catalyst had a real-time database of the universe of U.S. voters drawn from voter registration rolls, which were updated as the campaign progressed. This already provided a rich store of data on who was registered, often with a party affiliation (and/or race) as part of registration, how often they voted, and in which type of elections. The second step added a diverse array of sources to enrich these individual records, which is listed as process data in the figure. This might be information on campaign contributions from the FECC, gun registration records, consumer spending and other economic data, household composition, magazine subscriptions, neighborhood characteristics, and other indicators culled from diverse sources. Hundreds of variables from diverse sources in varied formats were included in the database.

An iterative data analysis process, which is termed "analytics" in the figure, is used to predict various outcomes. Analytics is a general term for collecting, organizing, and analyzing large sets of data that are often generated for other purposes and in varied formats. Because most Big Data are not well-structured (i.e., are "messy") given their diverse origins and their collection for other purposes, this requires considerable data manipulation to merge and coordinate data sources. Some information is lacking, but can be interpolated from other information. Then predictive analytics develops models to explain the desired outcome or behavior. The goal is to discover patterns and useful information to predict outcomes for the individuals in the collection.

In the case of the 2012 U.S. presidential election, the campaign analytic consultants ranked people on their probability of voting on Election Day, and their probability of voting for either Obama or Romney. The processed data were supplemented by direct-contact canvassing and telephone interviews for a subset of voters in the database.[1] The results of these contacts were fed back into the database. Here they can be used to further validate and refine the predictive analyses from the previous iteration of the model. Direct contact evidence of intention to vote or candidate preference can be compared to the results of the previous model (Hersh 2015). These data analytics also became an integral part of the campaign, especially in the ten battleground states, allowing the Obama campaign to customize the messages it sent to supporters, target

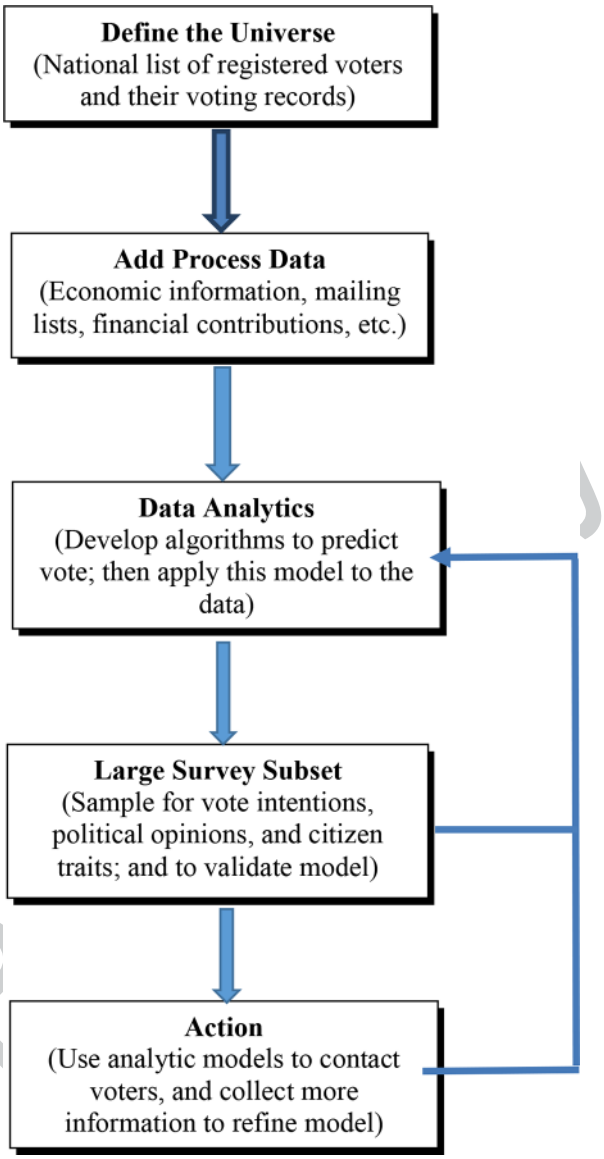

FIGURE 1 An illustrative flowchart of Big Data collection for voting predictions.

likely donors on a massive scale, keep real-time data on contributions, and mobilize potential supporters on Election Day. Less successfully, Republican consulting firms and the Romney campaign attempted to replicate these efforts.

From the perspective of an electoral research scholar, this multilevel Big Data collection marks a dramatic advance in research potential—it might be considered the "gold standard"

for Big Data research on political behavior.[2] But we should also recognize that this is an exceptional case. Researchers are trying to predict a behavior that is known to occur on a certain date in the future, in a known location, and with a limited choice set. Even the starting point—government records on who has registered to vote—is exceptional for political participation generally.

These tasks become much more complex when applied to voting in most other democracies, moreover. Past experience and my informal poll of election researchers found that individual turnout records are not as accessible in most democracies as they are in the United States, and in many cases are simply not available. Strict national privacy laws or government bureaucratic resistance often protects voting information that is widely available from commercial sources in the United States. For example, a French attempt to use big data approaches in the 2012 French presidential election faced a series of impediments (Pons 2013). Registered voter lists in France are accessible only from individual municipalities, and there are limitations on how political parties may use these records. The record of who voted is available for public viewing at the municipalities for only a 10-day period after an election, and again there are limitations on party usage of these data.[3] There are restrictions on access to individual voter records across most affluent democracies, even in Belgium and Australia where the government maintains turnout records because of compulsory voting.[4] So the massive information base on *registration and voting turnout* available from American electoral registers may be an anomaly among democratic countries.

There may be ways to approach the Big Data model through other methods. Norway, for example, is consolidating its voter registration records, and these may be available to nongovernmental entities. In Britain, Canada, Australia, and several other democracies, the parties or candidates have access to electronic voter registration records to allow them to more effectively campaign in elections. The parties can then add other data from party records to develop a richer profile of individual citizens. Some reports of the October 2015 Canadian parliamentary elections thus claimed it was Canada's first Big Data campaign (Ormiston 2015).

Access to consumer data is more restrictive outside the United States, however, especially in Europe where the open collection and sharing of consumer and personal information is restricted by national and European Union (EU) privacy laws. But the pattern is varied—for example, individual tax records are publicly available in Norway—something that would be unthinkable in the United States. In other cases, campaigns may find that localized data collection circumvents the impediments to creating large national databases. Such localized collection might be especially valuable in countries with district-based electoral systems where analysts can focus on critical swing districts, such as in the UK, Australia, and Canada. Local data collections also have valuable potential in conducting campaign field experiments (John nd; Pons 2013). The Big Data path from U.S. elections will be more difficult to follow in other countries. However, there is mounting evidence that some political parties (and election consulting firms) are moving in this direction, often avoiding public visibility because of the sensitivity of collecting personal information on the voters.[5]

Elections are distinctive political acts; how does one systematically measure and explain other forms of political participation, such as who writes a letter to a political official, works with a political group, or attends a protest? While voting is a public act in an institutionalized setting with regular timing and record keeping, most other forms of participation are more

episodic, noninstitutionalized, and citizen-initiated.[6] There have been major advances in measuring aggregate levels of some of these political acts using media sources, such as the *World Handbook* measures of protest and collective action. But these are aggregate statistics addressed to macro-level research questions, and not individual-level data that allow us to study the factors stimulating individual political behavior. The most effective way to measure these other forms of participation in a representative manner is to ask people directly, which is a reversion to traditional survey research methods. Thus, one of the efforts in the data collection project represented in this issue of the journal is to merge cross-national surveys measuring individual protest activity (see Powałko and Kołczyńska in this issue; Tomescu-Dubrow and Slomczynski in this issue).

The exception to this limitation may be online forms of political expression. Some of the most fertile fields of research involve analyses of Twitter, Facebook, and social media postings. Advocates suggest that Big Data analytics have the potential to forecast election outcomes and track other social and political phenomena in the same way that tracking polls might forecast outcomes (Anstead and O'Loughlin 2014; Sang 2012; Tumasjan et al. 2010; Wang et al. 2014). There is valuable information in charting the flow of information in such postings, the content analyses of these postings, the networks of interaction, and their shifting currents over time. Especially in terms of organizing activities among potential participants, whether they are in Brussels or in Cairo, social media are an important new form of communication (Christensen 2011; Khondker 2011). While this is an intriguing new research area, such data may face limits in representativeness and the depth of information that would allow us to test general theories of the causal forces influencing individual-level participation (Hargittai 2015). Online and social media's primary value seems to be in description, research on target groups, exploratory analysis, and process tracing—rather than generalizations to mass political behavior.

Political Attitudes

The research challenges increase if we want to describe and predict citizens' political attitudes. Because opinions lack the firm behavioral element of political participation, they are more subjective. There have been some attempts to infer policy positions through social media and Big Data sources, such as membership in social groups or financial contribution records. This may produce massive amounts of data from online aggregators, but the reliability and validity of such measures seems unclear at least for the foreseeable future. Most survey researchers readily admit that opinions are more difficult to measure than behavior because they involve what people think and not just how they act.

In summary, assessing citizens' political behavior via Big Data appears much more challenging than studying economic behavior or other sorts of online activity if our goal is to analyze the behavior of the general public. Analytic models to determine online advertising for an individual Web surfer is more amenable to Big Data analytics than to track Americans' changing policy preferences over time. Building models of the consumer preferences of Amazon customers narrows the data universe to Amazon customers. But the social sciences are typically concerned with a larger universe of people. Thus, studying political behavior beyond voting or political attitudes may touch at the limits of Big Data's current analytic potential.

ISSUES OF MEASUREMENT

One of the premises of "Big Data" is that a massive collection of diverse data can generate new analytic tools. We can collect rich information on individuals: records of past behavior, social ties, and social and economic characteristics beyond what is normally available in an opinion survey. Then these process data can be updated to provide a dynamic analysis. Sounds intriguing, but is it feasible?

Let me discuss protest activity as an example because it epitomizes the episodic, noninstitutionalized, and citizen-initiated forms of political participation and is closely related to the interests of this collection of articles. While we have relatively firm statistics on turnout, researchers cannot agree on the level of protest in the United States or other democracies today, or whether it is increasing or decreasing (Dalton 2013: ch. 3; Putnam 2000; Zukin et al. 2006). Protest comes in many forms, and there are debates on what constitutes protest as well as what constitutes political participation more broadly (van Deth 2014).

Researchers have attempted to study protest comparatively using two general methods. One approach is cross-national survey research. The most prominent example is the World Values Survey (WVS), which now has six waves of data beginning in 1981. The WVS asks respondents to report their participation in various protest or contentious political activities (Jakobsen and Listhaug 2014). The survey also includes a rich variety of personal characteristics, political attitudes, and social values that might be useful in developing individual-level models to predict protest activity. Several other large cross-national surveys—the International Social Survey Program, the Regional Barometer projects (e.g., Afrobarometer, Eurobarometer, etc.), and the European Social Survey—have similar data on protest activity, although typically with a shorter time period or fewer countries.

One complication of survey projects is their uneven national coverage. The World Values Survey, for instance, has surveyed almost 100 different countries across its multiple waves. But most countries were surveyed in only one or two waves, and less than a fifth of the WVS countries were interviewed in four of the initial five waves. The timing of interviews is also variable across waves and countries. I vigorously applaud efforts to merge different surveys together to expand coverage (Schoene and Kołczyńska 2014). However, I worry about the validity of the merged protest statistics if we must convert these data to a single metric; for example, to combine all the differently worded and scaled questions on boycotting to a single, comparable boycott variable.[7] The irregular spacing of surveys and the lack of synchronicity across projects makes the aggregation of different surveys even more problematic. Then, there is the inconsistency in the other individual-level questions asked in each survey.

Schoene and Kołczyńska (2014; also Kohler 2008) are making an admirable effort to identify the quality of surveys as a first step toward aggregation. Some of my research on protest and environmental group membership in affluent democracies found that the aggregate correlations can sometimes reach .70–.80 across projects, but sometimes significantly less. A more systematic comparison of survey projects by Inglehart and Welzel (2010) had more sanguine results. Ideally, we would need split-half survey experiments in several countries to recalibrate differently worded questions. All of this is a big task; but the field should make the efforts. In the end, however, the aggregation of separate surveys are big data collections, but not Big Data. All of the materials are collected using traditional survey research methods.

A second general method of studying protest is to collect information on individual protest events from media sources. Craig Jenkins used the Reuters wire service reports to assemble a cross-national (97 countries) and cross-temporal (1994–2004) database on protest events (Maher and Peterson 2008).[8] These data include rich information on the types of protest in each country, the major actors involved in each protest, the government response, and other characteristics. Similarly, Hanspieter Kriesi led a team that used media sources to collect longitudinal data on specific protest actions for a set of European democracies (Kriesi et al. 2012). These data provide an exceptional resource for tracking the ebb and flow of protest activity over time and across countries, and can provide unique information on the nature of protest that is missing from survey-based data. But they can only indirectly study the behavior of individual citizens.

If these are the traditional modes of research, can Big Data enrich or extend the analyses? There seem to be several possibilities. One possibility, so far untried to my knowledge, is to merge survey data on protest into the Big Data collections created for electoral studies. Such a data merger would provide a much richer environment to study who protests and why. A similar strategy would be to use Big Data methods to identify protestors or protest supporters through social media posts and other online sources. Network analysis models can uncover interpersonal dynamics that are difficult to study with general population studies. This might be the application of Big Data methods to small data samples. In social media terms, one might call this #protest research.

However, a complication is the aforementioned limitation on Big Data collections as a cross-national enterprise. Data protection laws and restrictions on data access limit the information available in many countries. And even if individual-level consumer or social data were available, it would be different data across countries because of national variations in privacy restrictions. Thus it is very challenging, at the present, to see cross-national Big Data options for studying political behavior by upscaling traditional survey research methods.

Another possible approach is to think of integrating different types of protest data. The media-based protest data tell us little about the individual-level causes of political action; the survey-based protest data lack a context or specificity—we know people protested but not the climate in which protest occurred. Researchers might merge these two types of data to produce a multilevel model of protest activity, with the aggregate data collections helping to define the political context.[9] Adding additional national characteristics or geographic data from other data sources may provide even richer measures of the context for individual action (Dalton, van Sickle, and Weldon 2010). Such multilevel contextual analyses represent important steps forward in the immediate future of comparative behavior research (Dalton and Anderson 2011).

In short, this approach might become a Big Data project not by adding more individual data, but by adding more information on the context for political protest; what are the characteristics of the country in cross-national studies, or the characteristics of the community within countries. It is relatively easy to add national aggregate data because other research projects are assembling a wealth of national-level political indicators. The *World Handbook of Political Indicators IV* is one notable example. Even more ambitious is the Quality of Government project at the University of Gothenburg in Sweden (qog.pol.gu.se). Since 2004 the QoG team of researchers has merged data from numerous international social science projects that have relevance for the quality of governance. Projects such as these provide models of the potential of cross-national data aggregation, and the Big Data element might be richer measures on the context for individual action rather than data on the individual.

Another option is to embrace new online data collection methods (Ackland 2013; Cantijoch, Gibson, and Ward 2014). Systematic Internet polling, social media polls, Twitter posts, and other online content, like self-selecting online questionnaires all provide a way to collect information from very large numbers of people at very modest cost. Analyses of Twitter posts have provided interesting data on information flows and social networks. This information could be linked to the data collected by online retailers, information aggregators, and database firms—depending on national privacy legislation. This would be Big Data in that it is a very large collection of data, from diverse sources, messy in its structure and content, and imprecise in many details. The logic of Big Data is that analytic models can use the bulk of the information to identify underlying relationships. If this works for Big Data in U.S. election campaigns, maybe an online equivalent can work for cross-national political behavior research.

I am skeptical about whether this last option has enduring potential for academic scholarship (see Hargittai 2015; Hesse, Moser, and Riley 2015). The click-through poll online is good for drawing the user's attention to a Web page, or feeding the beast of Big Data collection. Often such data are used for marketing information or for news accounts claiming to tap public opinion. There remain fundamental questions about the representativeness of such data, especially if we delve deeper than overall marginal distributions. Even reputable online survey research polls struggle to be sufficiently representative for scholarly research that is concerned about population subgroups. Thus, the deeper question is whether such information is appropriate for academic scholarship, which has higher standards of representativeness, measurement and validity than marketing research or media outlets generating online content. I address these topics in the next section.

THEORY AND CAUSALITY

I am intrigued by the potential that Big Data approaches have shown for some applications and some research questions in the social sciences. Many applications of Big Data analytics set the goal of successfully predicting outcomes. The Obama campaign wanted to predict exactly who would turn out on Election Day; Google wants to predict who might buy certain products if presented with an online ad (and what the ad should look like to generate more traffic). It is intriguing when researchers claim they can successfully predict election outcomes based on voluntary poll information collected from X-box users before the 2012 U.S. presidential election (Wang et al. 2014). The attempt to predict election outcomes through Twitter and other online data streams has spread cross-nationally (Sang 2012; Tumasjan et al. 2010).[10] So much for expensive random samples, extensive interviewer training, sophisticated survey questions, and the like. Q10

One might expect Google or Amazon to be primarily concerned with outcomes, rather than with developing theory-based analytic algorithms to explain consumer behavior. If there is an algorithm with hundreds of variables that successfully predicts outcomes, do we need to know why this works and what it says about the people making the decisions? But this is not just in the business world. Sasha Issenberg (2012: 247ff) describes this very process for the 2008 Obama campaign, the poster child for Big Data political behavior analysis. The campaign hired Ken Strasma, an outside Big Data consultant to manage the campaign analytics. Strasma built and revised analytic algorithms to capture the strategy of the Obama campaign, and made successful predictions. But the campaign itself did not have access to the algorithms inside Strasma's black box. Almost like visiting the Wizard of Oz, the campaign had met someone Q11

who could predict the future and they did not need to know how. This is the lure of Big Data analytics for many users.

One factor that differentiates social science from journalism is that we need to know how and why things happen. Deductive methods, theory testing, and causality are central to our endeavors. At least from the outside, Big Data methods often seem to emphasize outcomes over understanding these outcomes. Thus, Sudulich et al. (2014: 14–17) worry that the lack of methodological consistency and positivist rigor is a shortcoming of inductive and ad hoc methods in political studies using Internet resources. In contrast, other researchers are more sanguine (Hesse, Moser, and Riley 2015; Nagler and Tucker 2015; Monroe et al. 2015). This debate will continue, and it depends on the questions we are asking and the evidence we are using. However, Twitter cannot predict elections—at least not yet—and even if it could, the reason for election outcomes would remain unexplained.

In one sense, this is the most important challenge for Big Data research in terms of academic social science and adding to human knowledge. If left in the hands of market researchers and economists, it might not be addressed. But this is also the challenge that can be most easily addressed if scholars use Big Data to develop theoretical driven analytics, and meaningfully test alternative models.

CATCHING A MOVING TARGET

The development of the Internet has had a transformative effect on many aspects of our lives. Commerce is deeply affected by the online marketplace. Once rare information is now available with a few keystrokes, or a request to Siri. Political campaigns that were once based on meeting halls, and then televisions screens, are now shifting toward the Internet as well (Rommele and Schneidemesser 2015). Citizens who used to sit by their typewriters to write elected officials now dash off e-mails en masse. There is no denying the political importance of the Internet as a vehicle for commerce, entertainment, and politics.

I mention this because I am not claiming that the Internet and Big Data research based on online sources are irrelevant to scholarship. This is far from the case, because I study online political action in my own research. There is real value in studying the political content of this new medium. Analyses of Twitter exchanges can give us a new tool to study conversations and networks, analyzing online content on social media or even party campaign Web sites is a new and potentially insightful way to study political communication. My task here, however, was to consider the potential of Big Data approaches to understand comparative political behavior, which is a distinct research goal.

The experience of recent U.S. elections demonstrates that Big Data can be a very powerful tool in predicting electoral behavior. But to be a valuable scholarly research tool for understanding political behavior, Big Data faces three challenges: representativeness, measurement issues, and the potential for theory testing.

If our goal is to describe and perhaps explain activities online, or explain the activities of a well-defined subgroup, this is a feasible and valuable research objective. However, if our goal is to understand the mass public and the reasons for their actions, online sources are inevitably problematic because the Web is not a representative medium by nature, and there are limited ways to ensure that results can be generalized to a known population. Even as Sudulick et al. introduced a research collection on social media and Web-based data, they observed that

"almost all of the chapters in this volume make it clear the limitations of their analysis in relying solely on online-generated data, when one is seeking to make generalizations about wider populations of interest" (Sudulick et al. 2014: 11; also Anstead and O'Loughlin 2014). The issue of representativeness remains a sincere challenge for most online methods embedded in Big Data research projects when applied to mass political behavior.

Issues of measurement are a second challenge. Philip Converse used to have a favorite saying: "what's important we cannot measure, and what we can measure isn't important." Then his tag line was "Nevertheless … ". Converse was speaking of the difficulty of measuring public opinion and political behavior with mass opinion surveys, where the researchers can ask the questions they believe are most relevant to the topic at hand. Big Data tend to work the other way around, more like the second part of Converse's refrain. Using data collected for other purposes as surrogates for what we want to measure can lead to statistically significant correlations and even high predictive power. But we know that measuring opinions reliably is a challenging task, and measuring them indirectly presents another challenge for Big Data research projects. Often I have seen Big Data analyses that produce fascinating charts and high levels of prediction, but the analyses lack a clear dependent variable because what is important cannot be measured by Big Data sources.

Finally, the complexity of Big Data approaches in commerce lean toward validation by successful prediction. However, the goal of academic research is to understand causal processes and this inevitably requires deductive theory testing. There is nothing inherently contradictory between Big Data methods and theory testing, if the effort is made (Monroe et al. 2015; Nagler and Tucker 2015). Complex algorithms try to test models of the world against the real world of empirical evidence. Here it seems that academic applications can make real strides by emphasizing theory more in the development of these analyses.

In the end, we should be positive about the future of Big Data to address old political behavior research questions in new ways, and address new questions that emerge from this process. Many of the essays in this collection contribute to this goal. Big Data can be a powerful tool that dwarfs past methodologies in many areas. Like prospectors sifting through real mountains, Big Data can give us the tools to find the veins of gold in the mountains of social science data that the Internet and other sources are now producing. The challenge is to do this correctly, and not be distracted by the method itself.

ACKNOWLEDGMENTS

The author would like to thank many colleagues who helped with current information on the accessibility of election data cross-nationally and provided other advice: Bernt Aardal, Christian Collet, Patrick Dumont, David Farrell, Diego Garzia, Rachel Gibson, Peter John, Oddbjorn Knutsen, Ian McAllister, Vincent Pons, Andrea Rommele, Rüdiger Schmitt-Beck, and Christian Welzel. J. Craig Jenkins and a journal reviewer also provided valuable feedback on earlier versions of this essay.

NOTES

1. One of the most promising areas is the blending of these methods with field experiments to test alternative methods of voter mobilization and persuasion (Green and Gerber 2008; John 2013).

2. It is noteworthy that this methodology is generally unavailable for scholarly research. Instead it is controlled by political organizations or commercial enterprises for their own interests, and even the campaigns do not have access to the proprietary data collections.

3. Personal communication with Vincent Pons, March 12, 2015.

4. I consulted with election specialists in nine countries (Australia, Belgium, Canada, France, Germany, Japan, Italy, the Netherlands, United Kingdom). In most countries the records are decentralized to the local governments and access is restricted in various ways. In most instances the voting records list who is registered, but not past voting behavior. For example, Canada has a national list of registered voters, but use is restricted to members of parliament, electoral commissions, and political parties. In two cases, Britain and Italy, recent legislation has further restricted access to voter records. Only in the case of Norway do new rules suggest greater access in the future.

5. Evidence of this international interest is the widespread consulting activities of U.S. Big Data experts from the Obama and Romney campaigns, who have advised several major campaigns in Europe since 2012.

6. One exception is contributing to a U.S. campaign that is recorded and publicly available. But this information is not readily available nor as politically important in other democracies.

7. For example, three large cross-national projects use substantially different questions to measure boycott behavior, and recoding into a fully comparable dependent variables seems problematic:

World Values Survey: Now I'd like you to look at this card. I'm going to read out some different forms of political action that people can take, and I'd like you to tell me, for each one, whether you have actually done any of these things, whether you might do it or would never, under any circumstances, do it [joining in boycotts]?

International Social Survey Program: Here are some different forms of social and political action that people can take. Please indicate whether you have done any of these things in the past year, in the more distant past, whether you have not done it but might do it, or have not done it and would never, under any circumstances, do it [boycotted, or deliberatively bought, certain products for political, ethical or environmental reasons]?

European Social Survey: There are different ways of trying to improve things in [country] or help prevent things from going wrong. During the last 12 months, have you done any of the following [boycotted certain products]?

In addition to these differences in the stem question, the wording of the response codes varies across surveys, and the WVS changed its question format over time. Each survey is a reliable database, but merging them into a single variable seems very challenging.

8. These data are available from https://sociology.osu.edu/worldhandbook.

9. I see less potential to add the survey-based protest data to the aggregate media-based data series because of varying methodologies and time frames for each data source. However, the cross-referencing of both aggregate data sources would be valuable for validating both estimates (Dalton, van Sickle, and Weldon 2008).

10. It is not surprising if the term "Literary Digest Poll" popped into your thinking as you read the last paragraph. The *Literary Digest* was a U.S. magazine that correctly predicted the 1920 to 1932 elections through a straw poll. In 1936, however, it predicted a landslide victory for Republican Alf Landon over Democrat Franklin Roosevelt, with the predictable outcome for the future of the *Literary Digest* and straw polls.

REFERENCES

Ackland, Robert. 2013. *Web Social Science: Concepts, Data and Tools for Social Scientists in the Digital Age*. Thousand Oaks, CA: Sage.

Almond, Gabriel, and Sidney Verba. 1963. *The Civic Culture*. Princeton, NJ: Princeton University Press.

Anstead, N., and B. O'Loughlin. 2014. "1936 and All That: Can Semantic Polling Dissolve the Myth of Two Traditions of Public Opinion Research?" Pp. in *Analyzing Social Media Data and Web Networks: New Methods for Political Science*, edited by Marta Cantijoch, Rachel Gibson, and Stephen Ward. Basingstoke: Palgrave Macmillan. Q16

Campbell, Angus, et al. 1960. *The American Voter*. New York: Wiley.

Cantijoch, Marta, Rachel Gibson, and Stephen Ward, eds. 2014. *Analyzing Social Media Data and Web Networks: New Methods for Political Science*. Basingstoke: Palgrave Macmillan.

Christensen, Christian, ed. 2011. *Twitter Revolutions? Addressing Social Media and Dissent.* Special Issue of *Communication Review* 14.

Dalton, Russell. 2013. *Citizen Politics: Public Opinion and Political Parties in Advanced Industrial Democracies*. 6th ed. Washington, DC: Congressional Quarterly Press.

Dalton, Russell, and Christopher Anderson, eds. 2011. *Citizens, Context and Choice*. Oxford: Oxford University Press.

Dalton, Russell, Alix van Sickle, and Steven Weldon. 2010. "The Individual-Institutional Nexus of Protest Behavior." *British Journal of Political Science* 40:51–73.
Green, Donald, and Alan Gerber. 2008. *Get Out the Vote: How to Increase Voter Turnout*. 2nd ed. Washington, DC: Brookings Institution 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* 659:63–76.
Hellweg, Eric. 2012. "2012: The First Big Data Election." *Harvard Business Review* (November).
Hersh, Eitan. 2015. *Hacking the Electorate: How Campaigns Perceive Voters*. Cambridge: Cambridge University Press.
Hesse, Bradford, Richard Moser, and William Riley. 2015. "From Big Data to Knowledge in the Social Sciences." *Annals of the American Academy of Political and Social Science* 659:16–32.
Inglehart, Ronald, and Christian Welzel. 2010. Changing Mass Priorities: The Link between Modernization and Democracy." *Perspectives on Politics* 8:551–67.
Issenberg, Sasha. 2013. *The Victory Lab: The Secret Science of Winning Campaigns*. Q17
Jakobsen, Tor Georg, and Ola Listhaug. 2014. "Social Change and the Politics of Protest." Pp. ■ in *The Civic Culture Transformed*, edited by Russell Dalton and Christian Welzel. Cambridge: Cambridge University Press. Q18
John, Peter. "Field Experiments in Political Science Research." *Oxford Bibliographies Online*. Oxford: Oxford University Press (http://www.oxfordbibliographies.com/).
Khondker, Habibul Haque. 2011. "The Role of New Media in the Arab Spring." *Globalizations* 8:675–89.
Kohler, Ulrich. 2008. "Assessing the Quality of European Surveys: Towards an Open Method of Coordination for Survey Data." Pp. ■ in *Handbook of Quality of Life in the Enlarged European Union*, edited by Jens Albers, Tony Fahey, and Chiara Saraceno. London: Routledge. Q19
Kriesi, Hanspeter et al. 2012. *Political Conflict in Western Europe*. Cambridge: Cambridge University Press.
Maher, Thomas, and Lindsey Peterson. 2008. "Time and Country Variation in Contentious Politics: Multi-Level Modeling of Dissent and Repression." *International Journal of Sociology* 38(3):58–81.
Mayer-Schönberger, Viktor, and Kenneth Cukier. 2014. *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. New York: Houghton Mifflin Harcourt.
Monroe, Burt et al. 2015. "No! Formal Theory, Causal Inference, and Big Data Are Not Contradictory Trends in Political Science." *PS: Political Science and Politics* 48:71–74.
Nagler, Jonathan, and Joshua Tucker. 2015. "Drawing Inferences and Testing Theories with Big Data." *PS: Political Science and Politics* 48:84–88.
Pons, Vincent. 2014. "Does Door-to-door Canvassing Affect Vote Shares? Evidence from a Countrywide Field Experiment in France." Paper presented at CREST, Harvard Business School, BGIE group, Boston. Q20
Putnam, Robert. 2000. *Bowling Alone: The Collapse and Renewal of American Community*. New York: Simon and Schuster.
Rommele, Andrea, and Dirk von Schneidemesser. 2015. "Election Campaigning Enters a Fourth Phase: From Professionals to Citizens." Paper presented at the annual meeting of the American Political Science Association, San Francisco, California.. Q21
Sang, Erik, Tjong Kim, and Johan Bos. 2012. "Predicting the 2011 Dutch Senate Election Results with Twitter." Proceedings of the Thirteenth Conference of the European Chapter of the Association for Computational Linguistics.. Q22
Schoene, Matthew and Marta Kołczyńska. 2014. "Survey Data Harmonization and the Quality of Data Documentation in Cross-National Surveys." Ohio State University, CONSIRT Labs: Methodology of Survey Data Harmonization (http://consirt.osu.edu/wp-content/uploads/2014/11/CONSIRT-Working-Papers-Series-3-Schoene-and-Kolczynska.pdf).
Sudulich, Laura et al. 2014. "Introduction: The Importance of Methods in the Study of the 'Political Internet.'" Pp. ■ in *Analyzing Social Media Data and Web Networks: New Methods for Political Science*, edited by Marta Cantijoch, Rachel Gibson, and Stephen Ward. Basingstoke: Palgrave Macmillan. Q23
Tumasjan, Andranik, Timm Sprenger, Philipp G. Sandner, and Isabell M. Welpe. 2010. "Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment." Paper presented at the Fourth International AAAI Conference on Weblogs and Social Media. Q24
van Deth, Jan. 2014. "A Conceptual Map of Political Participation." *Acta Politica* 49:349–67.
Verba, Sidney, and Norman Nie. 1972. *Participation in America*. New York: Harper and Row.
Wang, Wei et al. 2014. "Forecasting Elections with Non-representative Polls." *International Journal of Forecasting*. Q25
Zukin, Cliff, Scott Keeter, Molly Andolina, Krista Jenkins, and Michael X. Delli Carpini. 2006. *A New Engagement? Political Participation, Civic Life, and the Changing American Citizen*. New York: Oxford University Press.