*PA*rtecipazione e *CO*nflitto
** The Open Journal of Sociopolitical Studies*
http://siba-ese.unisalento.it/index.php/paco
ISSN: 1972-7623 (print version)
ISSN: 2035-6609 (electronic version)
***PACO*, Issue 11(2) 2018: 484-496**
DOI: 10.1285/i20356609v11i2p484

Published in July 15, 2018

SYMPOSIUM/2

THE PROMISE AND PITFALLS OF BIG DATA AND COMPUTATIONAL STUDIES OF POLITICS

Jennifer Earl
University of Arizona

1. Introduction

A great deal of ink has been spilled over "big data" in the last several years—it has consumed computer science, engineering, information science, medicine, and the social sciences, among other areas, and has ushered in new fields such as data science. In the United States (US), big data has also been featured in national funding priorities and in the funding decisions of specific agencies, such as the National Science Foundation (NSF). While White and Breckenridge (2014) note that funding to proposals about big data was non-existent at NSF until 2009, when a handful of projects began to be funded exclusively by the Directorate for Computer and Information Science and Engineering. The torrent of proposals opened up around five to six years ago, with 214 "big data" projects funded by NSF in 2013. Now, big data is a common talking point, even if its meaning and relevance to social science is still less settled. As this special issue demonstrates, it is also working its way into sub-fields, such as social movement studies.

In this article, I do not intend to rehash the many arguments that have been made over the last five years about the role, opportunities, and risks of big data (e.g., boyd

and Crawford 2012; Burrows and Savage 2014; Hesse, Moser, and Riley 2015; Oboler 2012; Rojas 2015; Sandra 2013). Instead, I make two central claims and elaborate them in turn. First, I offer an important distinction between two often elided concepts: big data and computationally intensive data collection and/or analysis. As I argue below, big data is about scale—when one speaks of big data, they are speaking of data whose size creates storage, access, and analysis problems precisely because of its size, which is often in the petabytes. Computational intensity, on the other hand, is related to the scope of the programming skills and tools needed to handle massive, complex, and/or poorly structured data, and data of any size—big to quite "small" but complex—may be computationally intensive.[1] Getting swept up in the language of big data without distinguishing between scale and computational intensity shifts the terrain away from the research-frontier problems that social science projects more routinely face and places the emphasis on engineering frontiers. Second, I also examine some common dangers, which do not inherently flow from big data or computational intensity, but represent major usage problems that can lead to collective peril. In this critical view of what is becoming common, if not accepted, research practices, I introduce principles that could improve research moving forward. While these two points may read as curmudgeonly—challenging what is becoming common parlance and practice—both provide important insights about the opportunities and dangers that may lay ahead. Both are also broadly relevant to the social sciences, and not restricted to social movement studies, although I try to animate the discussion with social movement examples.

2. Distinguishing Big Data from Computational Intensity

Computer science and the engineering were among the first to worry about and treat big data as a research frontier. In those fields, many of the leading edge problems posed by big data derive from scale: how do researchers efficiently store vast troves of data (e.g., as a physical storage problem, as a distributed computing problem, etc.), make that data accessible to users (which can mean accessible to programmers or to end-data users), use computational tools to help sort so much data, and track the

[1] While I am not attempting to introduce a formal measure of computational intensity, one could imagine a continuum on which the scale, complexity, or characteristics of data render it possible to collect, store, access, and analyze without any computing tools to data that is impossible to collect, store, access, and/or analyze without significant computing power and programming skills and tools. The more significant computing and generalized programming skills and tools (beyond, for instance, standard statistical packages such as Stata), the more computationally intensive one could claim such data to be.

provenance, versioning, and other important meta-data that are connected to these datasets that may include previously unthinkable numbers of data points?

Since scale is so field specific, considering big data from outside the social sciences can be helpful. The Large Hadron Collider (LHC) at CERN produces about 25 gigabytes (GB) of data per second (CERN 2018), which would be over 2.16 petabytes of data in a 24 hour period. Of course, the LHC is not continuously producing this volume of data, but its total data volume is staggering. In October 2017, the LHC computing staff set a record, collecting 12.3 petabytes of data (Jarlett 2017) and by June of 2017, CERN had 200 petabytes of data on tape storage, in hopes of preserving each piece of data for at least 20 years (Gaillard 2017). It is easy to see how storing, accessing, or analyzing this volume of data would produce unique and difficult engineering and computational challenges. These scale related challenges have driven massive amounts of innovation bringing physicists, engineers, computer scientists, and data scientists, among others, together, requiring new physical facilities and devices to be developed and new computational tools to collect, access, and analyze it. The research frontier created by this scale of data is about what is possible, not what is routine, readily available, or affordable.

Computational concerns have been integral to frontier solutions, just as expanding physical memory capacities, chip speeds, and processing infrastructures, among others, have been. Indeed, big data is always computationally intensive. CERN would be ill-advised to collect petabyte after petabyte of data if these data are practically unusable because it will take decades to access them, let alone analyze that data, but advances around truly big data are helping to make these data accessible and analyzed much more quickly.

The relationship between computational needs and big data has contributed to an elision between the two—scholars often say big data when the real issue is computational difficulty. But, they are not the same: computational intensity can exist even in moderately sized data but the central defining characteristic of big data is in its name—it's all about scale! While it might be impossible to find a big data project that is not also computationally intensive, it is easy to find computationally intensive projects that don't involve big data, whether that programming expertise and/or computing power is used for data extraction, recombination and/or merging, and/or analysis.

Separating these concepts is important because even if scale doesn't make it intractable to store, access, or analyze, more moderate data sizes (by CERN standards), it can still be expensive, require specialized but available devices and tools, and require significant expertise to collect, access, and analyze. For example, White and Breckenridge (2014) point to four big data sets within US social sciences: the General Social Survey,

the American National Election Survey, the Panel Study in Income Dynamics, and Integrated Public Use Microdata Series–International. All four are expensive, have required research infrastructure advancements, have yielded important social sciences advances, and are not big data if scaled against CERN. The same can be said of research in which hundreds of thousands of websites or millions or tweets are collected and analyzed.

While these are big by social science standards, they are not when scaled against many other datasets and the challenges they raise are not typically about size, but rather characteristics that may also co-occur with big data. White and Breckenridge's (2014) review of social science "big data" projects argues that in the social sciences, data challenges also result because data are "relational, aggregated, multilevel, and merged (RAMM)," by which they mean:

> *"Relational data have embedded linkages between individuals and groups; aggregated data include data and variables that are combined from multiple sources and measures; multilevel data combine measures at both individual and group levels; and merged data are combined from different original sources" (White and Breckenridge 2014, : 333).*

Any size dataset can have RAMM properties: while it may be true that the bigger the data, the more likely it is to also have some or all of these RAMM properties, RAMM doesn't only happen "at scale."

This brings us back to computational intensity. Even though the scale of big data greatly contributes to its computational intensity in other fields, in the social sciences properties like RAMM or the more unstructured nature of raw data contribute to its computational intensity. This computational intensity may stem from the programming skills and actual tools (or recombination of tools) required to extract more structured data from less structured data, to encode and make use of the RAMM properties of data, and to analyze data using computer-driven or computer-facilitated techniques. In other words, a core argument of this article is that while the scale of big data in other fields (think petabytes) creates computational intensity in the engineering fields, in the social sciences the RAMM properties, degree of unstructured raw data, etc. contribute more to computational intensity than sheer data volume. If social scientists far more frequently face computationally intensive problems than they face truly big data, it is important that we stop eliding the problems we face around computational intensity with the size or scale of datasets for at least four reasons.

2.1 Social Scientists Need to Focus on the Challenges We More Often Face

First, at a practical level, this elision helps to cultivate the illusion that our central challenges are the same physical and engineering limitations associated with storing and accessing such truly massive datasets. This makes engineering the solution and shifts our focus away from computational intensity in smaller, but still complex or unstructured, data. This is a major missed opportunity because there are a range of major data infrastructure problems that the social sciences could be collectively focused on, but without a conversation that is more explicit about our specific computational challenges, versus a general crowd-pleasing concern for "big data," we cannot shift the attention to our computational needs as effectively.

For instance, the integration of divergent data may be one of our biggest, and most fruitful, research infrastructure and computational challenges. A massive amount of data on politics and social movements already exists, but it is organized by dataset and each dataset may be available to only a segment of the research community (e.g., to ICPSR member institutions, to members of specific research networks). This requires that researchers know that the data they want has already been collected by someone else and that they can access that data. Even when accessible, these data have to be harmonized with other existing data, which can involve actual merging but also recalculations across divergent units of observation, etc. All of these actually represent very high barriers to the reuse of existing data.

Imagine a different research infrastructure world in which data is a common community resource because publicly funded datasets are contributed to a social movement data repository that organizes data at the variable, unit of observation, and time level (e.g., public opinion support for abortion, US States, 1960-2000), not to dataset-focused organizations (e.g., ICPSR). Just as Google transformed the web from a welter of individual and largely unusable sites to one of the most accessible, relevant, and rich information retrieval systems in the world by focusing on the relevance of content within those sites, scholars could search for variables, not datasets, and make on-the-fly datasets from these, including relevant meta-data. ICPSR is working on the logic of early Yahoo and other directory sites circa the late 1990s where the self-classification of overall themes into what amounted to searchable indexes dominated. ICPSR is like a massive card catalog for data. I am arguing we need to push for research infrastructures that work like Google, searching segments of information—variables, for instance—and making aspects of larger datasets more accessible. While this is a fantasy at the moment, it could be a reality. But, we will only collectively conceive, cooperate, prioritize, fund, and build such a repository if we start to see data complexity, not data

size, as our key research infrastructure problem. We need to stop fetishizing size and begin working to unleash the power of all of the data that already exists but is sorely under-utilized because of how it is stored and made accessible.

2.2 What Creates Data Complexity and/or Computational Intensity?

When we recognize that computational intensity—net of scale—is a key issue, we can begin to usefully unpack and understand the characteristics that create computational intensity and the specific problems and solutions related to those characteristics. For instance, social scientists often want to use computational tools to take very unstructured raw data and transform it into more useful structured data; the more unstructured raw data, the more computational intensity we might expect. By unstructured, I mean data whose values are not easily categorized or easily initially parsed. For example, webpages are less structured then tweets. Open text is less structured than a closed survey question.

There are already several common ways of approaching unstructured textual data. First, one could decide that the analyst only needs the part of the text that is more structured. For instance, this would involve "scraping" only certain elements from online artifacts such as scraping links from webpages (Garrido and Halavais 2003) or @mentions from tweets. Second, one could reduce the complexity of text into counts of specific elements, such as the number of mentions of proper names or other highly searchable textual targets. Third, one could use machine learning and other computer-assisted analysis techniques to try to identify specific signals or extract coarse meanings from data as part of a data extraction and/or reduction strategies, which also requires specific training and expertise (e.g., Elanor, Alessandro, and Adam 2014).

Some kinds of tasks are also computationally intensive, even if on a moderate scale. For instance, careful comparisons between versions or across corpuses of text represents a tractable but clear computational challenge. For instance, one can trace on a word-for-word basis different versions of famous works (the Bible, drafts of Shakespearean plays), the diffusion of specific themes or frames (Bail 2012), or count the occurrence of particular parts of speech or other elements in large lexicons with sufficient generalized programming skills.

Because our focus has been so easily shifted to big data and away from computational complexity, we collectively have not worked to identify the range of characteristics that are likely to create computation challenge. I have given a few as examples—handling unstructured data or doing massive but precise character comparisons—but there are undoubtedly many more. Identifying these computationally intensive situa-

tions and working on tools and protocols for handling these situations is more productive than using the covering term of big data and pretending that scale is the problem.

2.3 Building Analysis Tools that Matter to the Social Sciences

Once we understand what creates computational complexity, we are in a better position to push forward on solutions to those different causes. For example, as discussed above, a common social science goal is to analyze, compare, or extract meaning from a broader segment of text. We have long honed ways for doing this with human readers, but computationally extracting anything other than course meaning from text has proven difficult, at least at levels that conform to standard social scientific standards.

Several projects have attempted to use automated tools to do this in social movement studies. For instance, Hanna's Machine-Learning Protest Event Data System (MPEDS) project attempted to replace human coding with machine coding of newspaper articles. However, it has proven difficult to do this at levels of accuracy and thoroughness that rival human coding. This does not necessarily mean that this is impossible, but it does imply that major advances would be needed for this kind of endeavor to succeed. When we are focused on simply big data, we can forget how complicated it can be to extract real meaning from relatively modest sets of text in an automated way.

2.4 Computational and Programming Tools in Graduate Programs

Understanding the unique problems the social sciences, and social movements studies in particular face, allows us to understand the kind of training that graduate programs need to provide future PhDs. We don't need to teach fundamentals of distributed computing to all PhDs, but teaching more generalized programming languages, like python (and to some extent R) is important. I wholly believe that there will be a time in my career when programming is as essential of a tool for qualitative and quantitative researchers as understanding how to read an OLS regression table is for all sociological researchers. The basics of this won't be around the scale of data, but around the programming skills needed to manage many different kinds, including very unstructured, data. While some students may need to learn about tools built especially for large data, many won't even though they will benefit greatly from general computational skills.

3. Using Data Responsibly

In the rush to study "big data," many social scientists, including social movement scholars, have been captivated by the accessibility of raw data, particularly data from digital and social media, and have forgotten that digital and social media data, as with any other kind of data, has production, maintenance, and meaning patterns that we must understand before we can usefully learn from them. With "small data" and data that doesn't require computationally intensive collection, our research communities habitually apply expectations and criteria that help to ensure that we understand our data, we understand the processes that may be producing our data, and hence affect its meaning, and we understand the limits of what we can learn from that data. But, the rush and accessibility of "big" and/or online data has led many scholars to be so star struck by its accessibility and scale that they forget to ask even very basic questions. Put simply, too many people with the skills to collect these data haven't thought enough about what these data are, what they mean, or the social processes that produced them, curate them, or maintain them.

For instance, we are collectively developing a tendency to scrape first and ask questions later. In early digital analyses, researchers often scraped html links between websites, creating broad social network maps of areas of online conversation or activity, even though there was/is very little known about who tends to decide on posting a link and what considerations go into a typical decision to link (if there is a "typical" decision), how often the average curator culls or adds links, how often links are actively managed to manicure an image, the extent to which links reflect competitive or cooperative connections, etc. In short, despite a large amount of research studying connections made by links, we don't really know what a typical link "means" as a tie. It is not like friendship or who you ask for advice, and yet we have tended to study these links because they are there and we can scrape them, not because we know they are socially meaningful or how they are socially structured.

Today, people may scrape tweets or Facebook activity instead, but the we are still in that scrape first, ask questions later mode. Research is in the early stages of understanding what people attend to on their Facebook feeds and Twitter. Despite some very high profile users, Twitter is still used by only about a quarter of the American population (or less depending on the estimate you use) and its usage is highly demographically structured. While scholars are eager to study @mentions and retweets, there is a lot about Twitter usage itself, and also the relevance of Twitter usage to other social life, that is still unknown.

In all data—big and small, digital or traditional—we need to remember that data have production, maintenance, and meaning patterns that we must understand before we can study them, and learn from them, properly. We should try to understand the social history and life of the things we scrape and collect so we can understand what they may actually tell us. When there is not an existing literature on a platform or use, scholars who want to scrape it need to build that literature. When there is an existing literature, scholars need to start from that work and build on it instead of assuming that digital and social media don't have particular structures or social lives that we need to consider in understanding them.

Our motto should be "Understand and question first, scrape second," but we are a far from that. While data drawn from digital and social media are becoming popular in a wide variety of sociological sub-fields, including the study of protest, few scholars have taken courses, taught courses, or self-studied the large literature on digital and social media usage. To me, this is shocking, but an analogy may help skeptical readers understand my concern: there is a broad literature on newspaper data and how protest comes to be covered in newspapers (Earl et al. 2004) precisely because it is popular data and thus important to understand on its own terms. While it would be difficult to succeed in having research using newspaper data published without at least discussing these concerns, scholars routinely attempt to and sometimes succeed at publishing on digital and social media data without discussing its production or curation in a comparable way.

This is a major oversight because the production, curation, or maintenance of digital materials should affect what we think we can learn from it. Again, an example may help. I have read and reviewed a number of articles that scrape online repositories of social movement materials—Occupy websites, press releases from social movement organizations, etc.—and many often treat these archives unproblematically even though what lays within them is almost certainly not a perfect reflection of historical events or statements. Perhaps that press release archive has been pruned to remove releases that were later seen as embarrassing or associated with losses or negative events. Perhaps the record of self-reported events only contains the most successful of these events, not all of the events that actually occurred.

If someone delivered a box of press releases to my office and claimed they were all of the releases from a set of SMOs, would I just believe them? No! It would be incumbent on me to try to validate and understand the contours of that data before using it, but too often, we scrape data because we can, even if we don't know, and often don't try to determine, how complete it is, understand its origins, its social roles, or other important characteristics.

3.1 Evaluating Our Way to Better Practices and Research

There is a surprisingly easy solution: we can apply the same kinds of standards that social scientists have long held for offline material to "big data" and data gleaned from digital or social media. This would involve reviewers, editors, and readers expecting that research on the social meaning, production, curation, maintenance, etc. of the kind of data being examined would be reviewed and considered. This would involve making sure that we don't just trust the veracity of big or digital data. More formally, I propose a few principles, that if followed, would collectively raise the quality of research dramatically.

First, we should not be willing as reviewers or readers to treat the accessibility of information as a rationale for studying it. We need specific theoretical rationales, built on the back of empirical research about the kinds of data we are using, to explain why and how we use big or digital data.

Second, as reviewers and readers, we should expect that any research project using big or digital data (or any other kind of data, for that matter) engage existing research on the production, curation, maintenance, and/or social meaning of those data. This is just applying the same kinds of standards that have been used for offline data—for instance, newspaper data on events—to big data and digital data.

Third, as reviewers and readers we should begin just as an archivist or a historical sociologist when thinking about a corpus of material to study: what is the material an instance or case of and how complete is the material at recording those cases? What kinds of cases may be intentionally or accidentally missing? Could items that were once in the corpus have been removed (i.e., deleted posts or links)? How is that selection and curation of remaining items structured by a range of factors? What possibilities exist for measuring this selection or at least understanding the possible theoretical consequences of this selection?

Fourth, just as we should not accept accessibility as a research rationale for what kinds of broad data to collect, reviewers and readers should beware of operationalizations that are convenient but not necessarily meaningful. For instance, in work on protest participation, we have long known that intention or discussion of protest doesn't mean that one will actually participate (Klandermans and Oegema 1987), but I have reviewed a wide range of papers that attempt to "understand" micromobilization by studying tweets, or other online posts, about intended (i.e., potential) participation. Likewise, do @mentions really reflect a person's whole social network or even the most influential parts of it? Do users Twitter connections provide a good portrait of any specific users ties across multiple platforms and/or offline or hybridized ties? Many re-

search papers would seem to think so, boiling down relational influences to those catalogued on a particular platform in stylized ways. While convenient to study, these examples, and many things like them, theoretically doesn't pass muster and yet happen with alarming frequency because of the temptation of accessibility.

Finally, as reviewers and readers we need to demand a frank assessment of the impacts of studying particular kinds of online artifacts on the generalizations we might be able to draw from them. Again, an analogy may help. In experimental fields, external validity is a major concern. How much does the artificial environment of an experiment affect what we can learn from it? How helpful will findings from artificial situations be in understanding organic situations? Because this is a reasonable concern, experimentalists expect to have to discuss it and be evaluated on it. I am arguing that we should apply a similar standard to the use of digital and social media. The platforms we often study—however part and parcel of some people's lives they are—are still artificial environments. How much can people's behavior in a discussion forum or on Twitter tell us about their behavior in other settings, even other online settings? Put differently, are we learning things, for instance, about talk about intentions on Twitter, but not about micro-mobilization generally, by using this kind of Twitter data? I am not supposing an answer but I am arguing that reviewers and readers should demand this be a standard discussion topic in research using data from specific platforms. We should always be discussing how much of findings from a specific artificial and stylized platform can be used to understand a different, and perhaps, less constrained social environment and whether it is reasonable to study behavior on only a single platform at a time.

These principles are not exhaustive of the changes that I believe researchers, reviewers, and readers need to make but these are important starts. Indeed, the quality of research using big and/or digital data would be dramatically and instantly improved by the use of these principles. The availability of data has encouraged a rush to study data that we may not truly understand. I am making a simple but powerful claim: we should be as careful in our use and consideration of big and digital data sources as we are with other sources for which we have longer research legacies.

5. Conclusion

In this article, I opined the elision between big data and computation intensity and also the less-than-careful use of widely accessible digital and social media. I have tried to lay out ways forward that could improve our collective practice, and where space has allowed, couch this in terms of research on social movements. I want to end by

noting these cautions are precisely that—cautions, not misgivings. I am not claiming that scale is never an issue, just that it is not the major issue we should be focused on. Likewise, I am not arguing that big data or digital data are inherently problematic but rather that some research practices that are becoming common are problematic. There is an amazing amount of promise in computational intense data and methods, but, realizing this promise will require substantial research infrastructure advancements and a great deal of discipline, both in terms of fortitude and theory.

References

Bail C. A. (2012), "The Fringe Effect: Civil Society Organizations and the Evolution of Media Discourse about Islam since the September 11th Attacks", *American Sociological Review* (77)6:855-879.

boyd D., K. Crawford (2012), "Critical Questions For Big Data", *Information, Communication & Society* 15(5):662-679.

Burrows R., M. Savage (2014), "After the Crisis? Big data and the Methodological Challenges of Empirical Sociology", *Big Data & Society* 1(1):1-4.

CERN (2018), "Processing: What to Record." Accessed July 9, 2018. https://home.cern/about/computing/processing-what-record

Earl J., Martin M., McCarthy J. D., S. A. Soule (2004), "The Use of Newspaper Data in the Study of Collective Action", *Annual Review of Sociology,* 30:65-80.

Colleoni E., R. Alessandro, A. Arvidsson (2014), "Echo Chamber or Public Sphere? Predicting Political Orientation and Measuring Political Homophily in Twitter Using Big Data", *Journal of Communication,* 64(2):317-332.

Gaillard M. (2018), "CERN Data Centre Passes the 200-petabyte Milestone." Accessed July 9. https://home.cern/about/updates/2017/07/cern-data-centre-passes-200-petabyte-milestone

Garrido M., A. Halavais (2003), "Mapping Networks of Support for the Zapatista Movement: Applying Social-Networks Analysis to Study Contemporary Social Movements." in McCaughey, M., Ayers, M. D., (eds), *Cyberactivism: Online Activism in Theory and Practice*, New York: Routledge, pp. 165-184.

Hesse B. W., Moser R. P., W.T. Riley (2015), "From Big Data to Knowledge in the Social Sciences", *The ANNALS of the American Academy of Political and Social Science,* (659)1:16-32.

Jarlett H. K. (2018), "Breaking Data Records Bit by Bit", Accessed July 9. https://home.cern/about/updates/2017/12/breaking-data-records-bit-bit

Klandermans B., D. Oegema (1987), "Potentials, Networks, Motivations, and Barriers: Steps Towards Participation in Social Movements", *American Sociological Review,* 52:519-531.

Oboler A. (2012), "The Danger of Big Data: Social Media as Computational Social Science", *First Monday,* 17(7).

Rojas F. (2015), "Big Data and Social Movement Research." Accessed JUly 9 https://mobilizingideas.wordpress.com/2015/04/02/big-data-and-social-movement-research/

González-Bailón S. (2013), "Social Science in the Era of Big Data", *Policy & Internet* 5(2):147-160.

White P., R. S. Breckenridge (2014), "Trade-Offs, Limitations, and Promises of Big Data in Social Science Research", *Review of Policy Research* 31(4):331-338.

Author's Information

Jennifer Earl is a Professor of Sociology and (by courtesy) Government and Public Policy at the University of Arizona. She is Director Emeritus of the Center for Information Technology and Society and Director Emeritus of the Technology and Society PhD Emphasis, both at University of California, Santa Barbara. Her research focuses on social movements, information technologies, and the sociology of law, with research emphases on Internet activism, social movement repression, and legal change. She is the recipient of a National Science Foundation CAREER Award for research from 2006-2011 on Web activism. She has published widely, including an MIT Press book, co-authored with Katrina Kimport, entitled Digitally Enabled Social Change, which examines how the use of Internet affordances are reshaping the basic dynamics of protest online and was awarded an Honorable Mention for the Communication and Information Technologies Section of the American Sociological Association's Book Award in 2013.