SYMPOSIUM

No! Formal Theory, Causal Inference, and Big Data Are Not Contradictory Trends in Political Science

Burt L. Monroe, *Pennsylvania State University*
Jennifer Pan, *Harvard University*
Margaret E. Roberts, *University of California, San Diego*
Maya Sen, *Harvard University*
Betsy Sinclair, *Washington University in St. Louis*

As with any high fashion, the beauty and horror of "big data"[1] is in the eye of the beholder. The question that prompted the present symposium—"Are formal theory, causal inference, and big data contradictory trends in political science?"—is representative of the concerns that big data has raised in political science. Indeed, this is representative of discussions underway in every area of social science about how big data interacts with existing modes of inquiry as well as its potential benefits (Lazer et al. 2009; Varian 2014) and potential pitfalls (boyd and Crawford 2012; Lazer et al. 2014).

A review of these discussions does not yield any consensus on even what is meant by the term "big data." For us, the concept is broad and simultaneously captures several ideas. For us, the intuitive criterion—"lots of data"—is both unnecessary and insufficient. When referring to data *qua* data, "big" is defined relative to the computational or informational capacities of conventional approaches. Data can be big in many dimensions: observations (e.g., the US Census), covariates (e.g., gene sequences), file size (e.g., images), or network bandwidth (e.g., video) (Monroe 2013). Even "small" social data may contain interdependencies (e.g., networks, space-time, or hierarchy) that imply big models. The term also may describe the body of computational innovation that has become associated with such data, roughly equivalent to "data science" but need not imply the application of "data mining." Among these innovations, we might list data collection (e.g., GPS tracking), data manipulation (e.g., web scraping), management (e.g., Hadoop), information extraction (e.g., natural language processing), or efficient computation (e.g., MapReduce), as well as the body of inferential techniques referred to variously as "statistical learning" or "machine learning."

Taken together, these new types of and approaches to data are enabling new forms of data-intensive political science, some of which in isolation appear to challenge established models of inquiry in political science and science more generally. It is undeniable that prominent popular examples of big data and data science bear little resemblance to the datasets and research designs of traditional social scientific inquiry. Tweets are not collected experimentally or via random sample. Netflix does not require a theory or causal framework to turn correlations into "you-might-also-like" recommendations. An increasing number of applications within political science have the same flavor as Netflix-style problems, with success measured by out-of-sample prediction. Prominent examples include election forecasting (Linzer 2013) and conflict forecasting (Brandt, Freeman, and Schrodt 2011). Schrodt (2014) goes so far as to label models that do not risk the failure of out-of-sample tests as "pre-scientific." One data scientist asserted provocatively that, in prediction tasks, "specialist knowledge is useless and unhelpful"—a direct challenge to the role of theory (Aldhous 2012). So, to be sure, big data research taken as a whole includes a more pluralistic range of scientific tasks and inferential strategies than in the conventional social science toolkit.

We argue, however, that none of this means that big data is fundamentally incompatible with formal theory, causal inference, or social science research methods in general. To the contrary, big data already is interacting with formal theoretic and causal inference approaches in ways that are not only consistent with these approaches but that also enhance them by enabling us to answer new questions. Perhaps more important, social science is beginning to shape the world of big data. Much of big data is *social* data—that is, data about the interactions of people: how they communicate, how they form relationships, how they come into conflict, and how they shape their future interactions through political and economic institutions. It is the responsibility of social scientists to assume their central place in the world of big data, to shape the questions we ask of big data, and to characterize what does and does not make for a convincing answer. In the discussion that follows, we describe examples in political science in which big data helps us to (1) design better experiments, (2) make better comparisons between more precise populations of

doi:10.1017/S1049096514001760

interest, and (3) observe theoretically relevant social and political behavior that previously was difficult to detect.

BIG DATA CAN HELP US DESIGN BETTER EXPERIMENTS

We begin with what initially appears to be an obvious contradiction between causal inference and big data. Within the causal-inference community, the randomized experiment is generally considered the "gold standard" for the valid identification of treatment effects. Within the big data community, social media such as Twitter often comprise a default example—being familiar, ubiquitous, and undeniably big. Because data generated as the "exhaust" of social media are in every way observational and in no way experimental—that is, no manipulation of treatment, no random assignment of treatment and control—they appear irrelevant for scholars interested in causality. One way to utilize these data for causal inference is to treat them inductively, as a source of suggestive hypotheses that can be tested experimentally.

A recent pair of studies (involving two of this paper's authors) to determine censorship mechanisms in China illustrates this possibility. In the first study, King, Pan, and Roberts [KPR] (2013) analyzed a corpus of more than 11 million Chinese social media and blog posts that were collected before and after the Chinese government censored them via content filtering.[2] KPR determined whether the posts were censored and then used supervised learning methods to compare the content of censored and uncensored texts. It was surprising that the authors found that Chinese censorship is focused on stopping discussion of collective action while allowing criticism of the state. In other words, Chinese citizens can view vitriolic online criticism of policies and lower-level officials, but posts that praise the government are censored if they discuss issues such as ongoing protests.

In the second study, KPR (2014) conducted a randomized experiment to validate this result, noting the limitations of interpreting the previous result as causal. In this experiment, 1,200 social media posts—which discussed ongoing collective-action and noncollective-action events supportive and critical of the government—were written and submitted with random assignment to 100 of the top social media platforms across China. That is, using another type of big data analysis, the authors validated the theory generated from their observational study with an experiment: they found a significant causal effect on censorship of posts related to collective action and no causal effect on censorship of those supportive or critical of the government. Both KPR studies used big data and big data technology in the service of social science research-design principles.

BIG DATA CAN PROVIDE EMPIRICAL LEVERAGE THROUGH PRECISE SUBPOPULATIONS

Big data is typically diverse data and often advantageous not for enormous sample sizes but rather for providing sufficient sample size on small subpopulations or even individuals. This is crucial in the implementation of methods for causal inference in observational data in which experiments are impractical or unethical. The central principle of methods such as matching is to compare treated and untreated observations that were as similar as possible before treatment. Achieving balance, however, requires pruning the data of observations, creating a tradeoff between bias and variance (King, Lucas, and Nielsen 2014). Higher standards for balance will reduce bias but may leave data too sparse for useful inference. The access in big data to fine-grained subpopulations can alleviate the bias–variance tradeoff.

That is, using another type of big data analysis, the authors validated the theory generated from their observational study with an experiment: they found a significant causal effect on censorship of posts related to collective action and no causal effect on censorship of those supportive or critical of the government.

In a recent example, Hersh (2013) leveraged millions of individual-level observations to understand how people close to the victims of September 11 were influenced by the terrorist attack. From detailed individual-level data, he isolated two treatment groups—families of September 11 victims and neighbors of September 11 victims—and compared them to precisely matched control individuals. He found that, relative to the control group, those who were close to the victims of September 11 became more involved with politics in the years following the terrorist attack and also exhibited a conservative shift in their voting behavior.

In a similar way, the availability of large-scale, fine-grained data can enable tests of complex, rare, or subtle phenomena predicted by formal theory, in the manner of the National Science Foundation–led Empirical Implications of Theoretical Models (EITM) initiative (Granato and Scioli 2004). Consider, for example, Osorio's (2013) formal model of localized drug violence in Mexico. Among other phenomena, the model implies that increased democratization will lead to greater law enforcement and that greater law enforcement will lead to greater violence among drug-trafficking organizations. Testing these predictions requires very detailed data about relatively uncommon events—by municipality, by day, by actor—data that Osorio generated from Spanish-language newspapers and that described a quarter-million events in the Mexican war on drugs during 10 million municipality-days.

In a similar way, Ansolabehere, Hersh, and Shepsle [AHS] (2012) used data about almost 2 million individuals to test subtle predictions about voter registration. A positive correlation between voter registration and age has long been known

but attributed to habit or other social psychological explanations. Instead, the AHS model accounts for this through the dynamics of voter mobility. The model can be differentiated from alternative explanations in predictions about the shape of the relationship—a distinction that can be detected only in massive individual-specific data. As in the Osorio study, we observe a theory-driven use of big data, in the spirit of EITM, which otherwise would not have been possible.

Formal theory can point to the conditions under which a selection process may render a behavior unobservable in conventional data but observable in less-constrained data. Much of social choice theory, for example, is built on the fundamental ubiquity and instability of multidimensional preferences. The empirical relevance of this for the study of the US Congress, for example, appears to be muted by ideal point estimates that find voting to be well summarized in the modern era by one dimension (Poole and Rosenthal 1991). Conversely, there are theoretical models that suggest mechanisms (e.g., negative agenda control) will be used by leaders to suppress the consideration of bills along higher dimensions, with the empirical side effect of suppressing the detection of multidimensional preferences in legislative voting data (Dougherty, Lynch, and Madonna 2014). Recent work argues that legislative-speech records are less constrained by majority agenda-setters and can reveal higher dimensions, such as distributive preferences induced by electoral systems defined by geographic districts (Monroe, Colaresi, and Quinn 2008) or by multidimensional political phenomena such as heresthetical maneuvers of opposition parties (Tzelgov 2012). Similar to the findings of Patty and Penn (2015), in these examples, we observe measurement from big data, informed by and in the service of theory.

Formal theory can point to the conditions under which a selection process may render a behavior unobservable in conventional data but observable in less-constrained data.

BIG DATA CAN REVEAL BEHAVIOR THAT PREVIOUSLY WAS DIFFICULT TO OBSERVE

Another feature of the KPR studies discussed previously is the use of big data to reveal behaviors—that is, censored communications—that are difficult to observe because they are actively hidden. In some cases, "Data" initially becomes "Big" through less strict selection and censoring mechanisms than other data. When this is true, there is potential to observe new behaviors of theoretical interest.

Racism and other forms of out-group hostility are famously difficult to research directly; however, more indirect approaches can be fruitful. Recent work by Stephens-Davidowitz (2014a), for example, examined Google searches conducted during the 2008 US presidential race. He found that certain parts of the country were more likely to use racial epithets in conjunction with searches on Barack Obama's name—patterns that standard social science survey techniques failed to detect. Linking these search results with voter returns, Stephens-Davidowitz found strong evidence that racial hostility cost Barack Obama significant vote shares. In a similar search-based strategy calibrated with Craig's List ads and social media data, Stephens-Davidowitz (2014b) estimated that 5% of American men are gay and that social intolerance keeps 50% to 80% of gay men "in the closet," a previously unknown statistic. Of course, social media and similar data reflect only the population from which they were extracted (DiGrazia et al. 2013; Nagler and Tucker 2015). As with nonresponse to telephone surveys and opt-in to Internet surveys, the validity of these estimates depends on our ability to convincingly model and calibrate processes of selection to the population of interest. This is one strength of a social science approach to big data. A landmark example of this strength is provided by Lazer et al. (2014), who demonstrated that "Google Flu Trends" is not the valid leading indicator of flu incidence that it initially appeared to be.

Indirect effects in social contexts are invisible in conventional data but can be observed in big data. For example, when studying how friends influence one another, experimental design possibilities for estimating direct effects are well understood; however, it is more difficult to design experiments to estimate indirect effects without partnering with big data. Suppose we want to estimate the effect of a friend casting a ballot on whether another voter will cast a ballot. Whereas it is possible to directly deliver a stimulus about voting to the friend, the researcher must rely on the friendship structure and the size of big data to estimate the indirect effect. In a recent example, Bond et al. (2012) conducted such an experiment by partnering with Facebook to include 61 million individual users. Their experiment found that, indeed, a Facebook friend may influence another to cast a ballot. Other social experiments have also relied on big data—for example, a large experimental population gleaned from voter registration lists or geographic

Political scientists have spent decades to hone, test, and perfect the appropriate methods for asking and understanding these types of questions, and we have a central role in integrating these insights into the world of big data.

mapping—to estimate indirect social effects (Nickerson 2008; Sinclair 2012).

CONCLUDING REMARKS

It is obvious that big data is here to stay. We believe that the strongest contribution to knowledge will come when we harness both the power of big data *and* rigorous methods and theories from social science. Given the theme of this symposium, we focus on big data's ability to help us (1) design better experiments, (2) make better comparisons between precise populations of interest, and (3) observe theoretically relevant social and political behavior that was previously difficult to detect. Beyond the relationship with formal theory and causal inference, big data offers other possibilities to enhance what we have done before or to enable us to do new things (Monroe 2013; Varian 2014).

Of course, we are fully aware of the strong critiques suggesting that current big data analyses are driven by atheoretical inquiries, or that many causal claims are being made solely by invoking "big data" without rigorous analysis or an understanding of the underlying assumptions. This criticism undoubtedly is true. It also is true that the public and policy makers have been swept up in the excitement about big data and that many are convinced by claims generated from big data research without sufficient intellectual scrutiny. As social scientists, this obviously is frustrating.

However, rather than dismissing these trends or disengaging, social scientists should view this juncture as a significant opportunity. Big data has the power to transform and expand the universe of answerable social science questions; as social scientists, we can and should shape the direction of big data analysis. Ultimately, much of big data is simply a richer version of social science data—data about how humans behave and interact—for which we already have developed a conceptual understanding. Political scientists have spent decades to hone, test, and perfect the appropriate methods for asking and understanding these types of questions, and we have a central role in integrating these insights into the world of big data.

ACKNOWLEDGMENT

This material is based in part on work supported by the National Science Foundation under IGERT Grant DGE-1144860, "Big Data Social Science." ■

NOTES

1. The standard usage affectation would be "Big Data," which we do not use here When referenced as a concept or scientific approach, however, we do treat "big data" as a singular noun.
2. The "biggest," most computationally complex aspect of data collection here is not the number and variety of posts captured but rather the speed with which they had to be captured after posting and before removal—a fundamentally social scientific feature of the research design.

REFERENCES

Aldhous, Peter. 2012. "Specialist Knowledge Is Useless and Unhelpful." *New Scientist*, December 7. Available at http://www.newscientist.com/article/mg21628930.400-specialist-knowledge-is-useless-and-unhelpful.html.

Ansolabehere, Stephen, Eitan Hersh, and Kenneth Shepsle. 2012. "Movers, Stayers, and Registration: Why Age Is Correlated with Registration in the U.S." *Quarterly Journal of Political Science* 7 (4): 333–63.

Bond, Robert M., Christopher J. Fariss, Jason J. Jones, Adam D. I. Kramer, Cameron Marlow, Jaime E. Settle, and James H. Fowler. 2012. "A 61-Million-Person Experiment in Social Influence and Political Mobilization." *Nature* 489: 295–8.

boyd, danah, and Kate Crawford. 2012. "Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon." *Information, Communication & Society* 15 (5): 662–79.

Brandt, Patrick T., John R. Freeman, and Philip A. Schrodt. 2011. "Real Time, Time Series Forecasting of Inter- and Intra-State Political Conflict." *Conflict Management and Peace Science* 28 (1): 48–64.

DiGrazia, Joseph, Karissa McKelvey, Johan Bollen, and Fabio Rojas. 2013. "More Tweets, More Votes: Social Media as a Quantitative Indicator of Political Behavior." *PLoS ONE* 8 (11): e79449.

Dougherty, Keith L., Michael S. Lynch, and Anthony J. Madonna. 2014. "Partisan Agenda Control and the Dimensionality of Congress." *American Politics Research* 42 (4): 600–27.

Granato, Jim, and Frank Scioli. 2004. "Puzzles, Proverbs, and Omega Matrices: The Scientific and Social Significance of Empirical Implications of Theoretical Models (EITM)." *Perspectives on Politics* 2 (2): 313–23.

Hersh, Eitan D. 2013. "Long-Term Effect of September 11 on the Political Behavior of Victims' Families and Neighbors." *Proceedings of the National Academy of Sciences* 110 (52): 20959–63.

King, Gary, Christopher Lucas, and Richard Nielsen. 2014. "The Balance-Sample Size Frontier in Matching Methods for Causal Inference." Available at http://gking.harvard.edu/files/gking/files/frontier_2.pdf

King, Gary, Jennifer Pan, and Margaret E. Roberts. 2013. "How Censorship in China Allows Government Criticism but Silences Collective Expression." *American Political Science Review* 107 (2): 326–43.

King, Gary, Jennifer Pan, and Margaret E. Roberts. 2014. "Reverse Engineering Chinese Censorship: Randomized Experimentation and Participant Observation." *Science* 345 (6199): 1–10.

Lazer, David M., Ryan Kennedy, Gary King, and Alessandro Vespignani. 2014. "The Parable of Google Flu: Traps in Big Data Analysis." *Science* 343 (6176): 1203–5.

Lazer, David, Alex (Sandy) Pentland, Lada Adamic, Sinan Aral, Albert-László Barabási, Devon Brewer, *et al.* 2009. "Life in the Network: The Coming Age of Computational Social Science." *Science* 323 (5915): 721–3.

Linzer, Drew. 2013. "Dynamic Bayesian Forecasting of Presidential Elections in the States." *Journal of the American Statistical Association* 108 (501): 124–34.

Monroe, Burt L. 2013. "The Five Vs of Big Data Political Science: Introduction to the Virtual Issue on Big Data in Political Science." *Political Analysis*, Virtual Issue 5.

Monroe, Burt L., Michael P. Colaresi, and Kevin M. Quinn. 2008. "Fightin' Words: Lexical Feature Selection and Evaluation for Identifying the Content of Political Conflict." *Political Analysis* 16 (4): 372–403.

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

Nickerson, David. 2008. "Is Voting Contagious? Evidence from Two Field Experiments." *American Political Science Review* 102 (1): 49–57.

Osorio, Javier. 2013. "Democratization and Drug Violence in Mexico." Paper presented to the American Political Science Association, Chicago, August 31.

Patty, John W., and Elizabeth Maggie Penn. 2015 "Analyzing Big Data: Social Choice, & Measurement." *PS: Political Science and Politics* 48 (1): this issue.

Poole, Keith T., and Howard Rosenthal. 1991. "Patterns of Congressional Voting." *American Journal of Political Science* 35 (1): 228–78.

Schrodt, Philip A. 2014. "Seven Deadly Sins of Contemporary Quantitative Political Analysis." *Journal of Peace Research* 51 (2): 287–300.

Sinclair, Betsy. 2012. *The Social Citizen*. Chicago: University of Chicago Press.

Stephens-Davidowitz, Seth I. 2014a. "The Cost of Racial Animus on a Black Presidential Candidate: Evidence Using Google Search Data." *Journal of Public Economics*. 118: 26–40.

Stephens-Davidowitz, Seth I. 2014b. "Estimating the Closeted Gay Male Population." Presentation to the Centers for Disease Control and Prevention, STD Prevention Conference, Atlanta, June 10.

Tzelgov, Eitan. 2012. "Damned If You Do and Damned If You Don't: Rhetorical Heresthetic in the Israeli Knesset." *Party Politics*. Available at doi:10.1177/1354068812462926.

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