We Are All Social Scientists Now: How Big Data, Machine Learning, and Causal Inference Work Together

Justin Grimmer, *Stanford University*

Information is being produced and stored at an unprecedented rate. It might come from recording the public's daily life: people express their emotions on Facebook accounts, tweet opinions, call friends on cell phones, make statements on Weibo, post photographs on Instagram, and log locations with GPS on phones. Other information comes from aggregating media. News outlets disseminate news stories through online sources, and blogs and websites post content and receive comments from their readers. Politicians and political elites contribute their own messages to the public with advertising during campaigns. The federal government disseminates information about where it spends money, and local governments aggregate information about how they serve their citizens.

The promise of the "big data" revolution is that in these data are the answers to fundamental questions of businesses, governments, and social sciences. Many of the most boisterous claims come from computational fields, which have little experience with the difficulty of social scientific inquiry. As social scientists, we may reassure ourselves that we know better. Our extensive experience with observational data means that we know that large datasets alone are insufficient for solving the most pressing of society's problems. We even may have taught courses on how selection, measurement error, and other sources of bias should make us skeptical of a wide range of problems.

This statement is true; "big data" alone is insufficient for solving society's most pressing problems—but it certainly can help. This paper argues that big data provides the opportunity to learn about quantities that were infeasible only a few years ago. The opportunity for descriptive inference creates the chance for political scientists to ask causal questions and create new theories that previously would have been impossible (Monroe et al. 2015). Furthermore, when paired with experiments or robust research designs, "big data" can provide data-driven answers to vexing questions. Moreover, combining the social scientific research designs makes the utility of large datasets even more potent.

The analysis of big data, then, is not only a matter of solving computational problems—even if those working on big data in industry primarily come from the natural sciences or computational fields. Rather, expertly analyzing big data also requires thoughtful measurement (Patty and Penn 2015), careful research design, and the creative deployment of statistical techniques. For the analysis of big data to truly yield answers to society's biggest problems, we must recognize that it is as much about social science as it is about computer science.

THE VITAL ROLE OF DESCRIPTION

Political scientists prioritize causal inference and theory building, often pejoratively dismissing measurement—inferences characterizing and measuring conditions as they are in the world—as "mere description" or "induction." Gerring (2012) showed, for example, that 80% of articles published in *American Political Science Review* focus on causal inference. The dismissal of description is ironic because much of the empirical work of political scientists and theories that they construct are a direct product of description. Indeed, political scientists have developed a wide range of strategies for carefully measuring quantities of interest from data, validating those measures, and distributing them for subsequent articles. Therefore, although descriptive inference often is denigrated in political science, our field's expertise in measurement can make better and more useful causal inferences from big data.

The VoteView project is perhaps the best example of political science's expertise with measurement and why purely descriptive projects affect the theories we construct and the causal-inference questions we ask (McCarty, Poole, and Rosenthal 2006; Poole and Rosenthal 1997).[1] VoteView is best known for providing NOMINATE scores—that is, measures of where every representative to serve in the US House and Senate falls on an ideological spectrum. The authors are emphatic that NOMINATE measures only low-dimensional summaries of roll-call voting behavior. Like other measurement techniques, these summaries are a consequence of both the observed data and the assumptions used to make the summary (Clinton and Jackman 2009; Patty and Penn 2015). Extensive validations suggest, however, that the measures are capturing variation in legislators' expressed ideology (Clinton, Jackman, and Rivers 2004; Poole 1984; Poole and Rosenthal 1985; 1997).

The impact of the VoteView project is broad and substantial. NOMINATE measures appear in almost every paper about the US Congress and in much of the work of other scholars related to US politics. These findings have fueled numerous debates. Perhaps one of the most famous findings

doi:10.1017/S1049096514001784

is that polarization in Congress—that is, the ideological distance between the two parties—has grown substantially in the past 40 years (McCarty, Poole, and Rosenthal 2006; Poole and Rosenthal 1984). This basic descriptive insight, which characterizes the state of the world rather than explaining why, has led to a large literature on the origins of polarization (McCarty, Poole, and Rosenthal 2006; 2009; Theriault 2008) and its consequence for governance (e.g., Krehbiel 1998). The findings on polarization also have reached the media, providing evidence for claims about the historic distance between the two parties. They even have been extended to include all candidates and donors across all levels of government (Bonica 2014) as well as all users of massive social networking websites (Bond and Messing 2014).

Social scientists know that large amounts of data will not overcome the selection problems that make causal inference so difficult.

The opportunities for important descriptive inferences abound in big data. For example, census data and social media posts can contribute to an important developing literature about how some of the fastest growing demographic groups (e.g., biracial Americans) reconcile their competing social and political identities (Davenport 2014). Aggregated newspaper articles can provide unprecedented accounts of the media's agenda (Boydstun 2013). Online discussions can answer broad questions about how often the public talks about politics during daily life. Each descriptive inference is important on its own and, if linked to broader populations (Nagler and Tucker 2015), would facilitate causal inferences and theoretical advances.

Each example also demonstrates the distinctive way that social scientists use machine-learning algorithms. Social scientists typically use machine-learning techniques to measure a certain characteristic or latent quantity in the world—a qualitatively different goal than computer scientists, who use the measures for prediction (Chang et al. 2009; Grimmer and Stewart 2013; Quinn et al. 2010). To measure latent quantities, social scientists must make consequential and untestable assumptions to compress data into some measure, similar to the assumptions necessary for causal inference. To assess how those assumptions affect the inferences made, social scientists developed a suite of methods for validating latent measures. These tools are invaluable in making descriptive inferences from big data that are useful for the most vexing problems—which provides our first example of how the analysis of big data is best viewed as a subfield of the social sciences.

RESEARCH DESIGN IN LARGE DATASETS

Descriptive inferences tell us about the world as it is. Big data proponents, however, argue that it also can tell us about the world as it could be. Big data, we often are told, will facilitate "data-driven" decision making. Companies and policy makers are told that they can use the large collections of information to be aware of the consequences of their actions before they are taken. Academics are told they can use the massive datasets to test causal theories that would be impractical in smaller datasets.

Of course, social scientists know that large amounts of data will not overcome the selection problems that make causal inference so difficult. Instead, a large literature has emerged to argue that causal inferences require a rigorous research design, along with a clear statement of the assumptions necessary for that design to yield accurate causal estimates (Imai, King, and Stuart 2008; Sekhon 2009). The best studies then will provide an argument about why those assumptions are satisfied and an analysis of what happens if they are violated.

Big data alone is insufficient to make valid causal inferences; however, having more data certainly can improve causal inferences in large-scale datasets. Consider, for example, using matching methods and the characteristics of observations to make treatment and control units comparable (Ho et al. 2007; Rosenbaum and Rubin 1983). A challenge in matching methods is that there may be few units similar on a wide range of characteristics; therefore, there may be potential discrepancies on observable characteristics, let alone differences on unobserved traits. However, massive datasets may provide ideal settings for matching, wherein the multitude of units ensures that the matches are close or that the treatment and control units are similar (Monroe et al. 2015).

Other research designs used to estimate causal effects also could benefit from a massive number of observations. For example, numerous papers use regression-discontinuity designs to estimate a valid local estimate of an intervention's effect (Lee 2008; Lee, Moretti, and Butler 2004). One limitation of the design is that there often are too few units very close to the discontinuity; therefore, units farther away must be used to obtain precise estimates. If there is a discontinuity in a large dataset, however, it is necessary to borrow information from units that are far from the discontinuity.

Massive datasets and social networking sites provide opportunities to design experiments on a scale that was previously impossible in the social sciences. Subtle experiments on a large number of people provide the opportunity to test social theories in ecologically valid settings. The massive scale of the experiments also provides the chance to move away from coarse treatments estimated at the population level to more granular treatments in more specific populations. The result will be a deeper understanding of the social world. Designing experiments and developing robust observational research designs requires more than computational tools. Social science is necessary, then, for big data to provide data-driven decision making.

COMBINING MACHINE LEARNING AND CAUSAL INFERENCE

Large collections of data not only improve the causal inferences we make. The computational tools that often are associated with

the analysis of big data also can help scholars who are designing experiments or making causal inferences from observational data. This is because many problems in causal inference have a close analogue in machine learning. Indeed, scholars who recognize this connection already have improved how experiments are designed and analyzed.

Consider, for example, blocking observations in an experiment—that is, grouping together observations before random assignment to improve the precision of estimated effects. Higgins and Sekhon (2014) leveraged insights from graph theory to provide a blocking algorithm with guarantees about the similarity of observations assigned to the same block. Moore and Moore (2013) used tools to provide a blocking algorithm for experiments that arrive sequentially. Machine-learning tasks also are helpful for the closely related task of matching. Hazlett (2014) used a kernel method to create a flexible matching method to reduce imbalances between treatment and control units.

Machine-learning methods also can improve what we learn from experiments and the types of experiments that are conducted. Not only are effect estimates interesting for the entire population of units in the experiment; we also might be interested in how the treatment effects vary across units. Furthermore, machine-learning methods are effective at identifying actual differences in response. For example, Imai and Ratkovic (2013) extended variable selection methods to estimate treatment-effect heterogeneity, whereas Green and Kern (2012) used Bayesian additive regression trees to capture systematic heterogeneity in treatment effects.

Indeed, combining machine learning to make causal inferences is one of the fastest growing and most open fields in political methodology. There is much work to be done in estimating causal effects in texts (Roberts et al. 2014) and political networks (Fowler et al. 2011). There also are numerous opportunities to combine experimental design with machine-learning algorithms to learn how high-dimensional treatments affect response. This area presents an opportunity for leveraging the insights from social science, the computational tools from machine learning, and the big data sources that now are abundant.

WE ARE ALL SOCIAL SCIENTISTS NOW

The big data revolution has been hailed as a triumph of computation and, indeed, it is. Computational advances have led to monumental changes in the tools that everyday people use to live their life, immense progress in how the data are stored, and unprecedented tools to analyze large collections. The results are the largest and most detailed datasets in the history of the world. However, the big data revolution also is a recognition that the problems addressed by quantitative social scientists—measuring quantities of interest from noisy data and inferring causal effects—are abundant. Therefore, for big data to be useful, we must draw on the substantial knowledge base that social scientists have amassed about how to most effectively use quantitative tools to solve social scientific problems. Recognizing the value of social science will lead to fruitful collaboration. Although social scientists have little experience with massive datasets, we have extensive experience with causal inference. Data scientists have significantly more experience with large datasets but they tend to have little training in how to infer causal effects in the face of substantial selection.

Computational advances have led to monumental changes in the tools that everyday people use to live their life, immense progress in how the data are stored, and unprecedented tools to analyze large collections.

Social scientists must have an integral role in this collaboration; merely being able to apply statistical techniques to massive datasets is insufficient. Rather, the expertise from a field that has handled observational data for many years is required. For "big data" to actually be revolutionary, we must recognize that we are all social scientists now—regardless of in which field our degree is. ■

NOTE

1. Of course, there are other important data collections in the study of the US Congress that have many of the same characteristics, including the Policy Agendas Project (Jones, Wilkerson, and Baumgartner 2009) and the Congressional Bills Project (Adler and Wilkerson 2014).

REFERENCES

Adler, E. Scott, and John Wilkerson. 2014. "Congressional Bills Project." Available at www.congressionalbills.org. Accessed August 1, 2014.

Bond, Robert, and Solomon Messing. "Quantifying Social Media's Political Space: Estimating Ideology from Publicly Revealed Preferences on Facebook." Stanford University Unpublished Manuscript.

Bonica, Adam. 2014. "Mapping the Ideological Marketplace." *American Journal of Political Science* 58 (2): 367–87.

Boydstun, Amber. 2013. *Making the News: Politics, the Media, and Agenda Setting.* Chicago: University of Chicago Press.

Chang, Jonathan, Jordan Boyd-Graber, Sean Gerrish, Chong Wang, and David Blei. 2009. "Reading Tea Leaves: How Humans Interpret Topic Models." In *Neural Information Processing Systems Proceedings*, 288–96.

Clinton, Joshua D., and Simon Jackman. 2009. "To Simulate or NOMINATE?" *Legislative Studies Quarterly* 34 (4): 593–621.

Clinton, Joshua, Simon Jackman, and Douglas Rivers. 2004. "The Statistical Analysis of Roll Call Data." *American Political Science Review* 98 (02): 355–70.

Davenport, Lauren. 2014. "Politics between Black and White." Redwood City, CA: Stanford University Unpublished Manuscript.

Fowler, James H., Michael T. Heaney, David W. Nickerson, John F. Padgett, and Sinclair Betsy. 2011. "Causality in Political Networks." *American Politics Research* 2: 437–80.

Gerring, John. 2012. "Mere Description." *British Journal of Political Science* 42 (4): 721–46.

Green, Donald P., and Holger L. Kern. 2012. "Modeling Heterogeneous Treatment Effects in Survey Experiments with Bayesian Additive Regression Trees." *Public Opinion Quarterly* 76 (3): 491–511.

Grimmer, Justin, and Brandon M. Stewart. 2013. "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts." *Political Analysis* 21 (3): 267– 97.

Hazlett, Chad. 2014. "Kernel Balancing (KBAL): A Balancing Method to Equalize Multivariate Distance Densities and Reduce Bias without a Specification Search." Cambridge, MA: MIT Unpublished Manuscript.

Higgins, Michael J., and Jasjeet S. Sekhon. 2014. "Improving Experiments by Optimal Blocking: Minimizing the Maximum Within-Block Distance." Berkeley: University of California Unpublished Manuscript.

Ho, Dan, Kosuke Imai, Gary King, and Elizabeth Stuart. 2007. "Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference." *Political Analysis* 15 (3): 199–236.

Imai, Kosuke, Gary King, and Elizabeth Stuart. 2008. "Misunderstandings between Experimentalists and Observationalists about Causal Inference." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 171 (2): 481–502.

Imai, Kosuke, and Marc Ratkovic. 2013. "Estimating Treatment Effect Heterogeneity in Randomized Program Evaluation." *Annals of Applied Statistics* 7 (1): 443–70.

Jones, Bryan, John Wilkerson, and Frank Baumgartner. 2009. "The Policy Agendas Project." Available at http://www.policyagendas.org. Accessed August 1, 2014.

Krehbiel, Keith. 1998. *Pivotal Politics: A Theory of US Lawmaking*. Chicago: University of Chicago Press.

Lee, David, Enrico Moretti, and Matthew Butler. 2004. "Do Voters Affect or Elect Policies? Evidence from the US House." *Quarterly Journal of Economics* 119 (3): 807–59.

Lee, Frances. 2008. "Dividers, Not Uniters: Presidential Leadership and Senate Partisanship, 1981–2004." *Journal of Politics* 70 (4): 914–28.

McCarty, Nolan, Keith Poole, and Howard Rosenthal. 2006. *Polarized America: The Dance of Inequality and Unequal Riches*. Cambridge, MA: MIT Press.

———. 2009. "Does Gerrymandering Cause Polarization?" *American Journal of Political Science* 53 (3): 666–80.

Monroe, Burt L., Jennifer Pan, Margaret E. Roberts, Maya Sen, and Betsy Sinclair. 2015. "No! Formal Theory, Causal Inference, and Big Data Are Not Contradictory Trends in Political Science." *PS: Political Science and Politics* 48 (1): this issue.

Moore, Ryan T., and Sally A. Moore. 2013. "Blocking for Sequential Political Experiments." *Political Analysis* 21 (4): 507–23.

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

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

Poole, Keith. 1984. "Least Squares Metric, Unidimensional Unfolding." *Psychometrika* 49 (3): 311–23.

Poole, Keith, and Howard Rosenthal. 1984. "The Polarization of American Politics." *Journal of Politics* 46 (4): 1061–79.

———. 1985. "A Spatial Model for Legislative Roll Call Analysis." *American Journal of Political Science* 29 (2): 357–84.

———. 1997. *Congress: A Political-Economic History of Roll Call Voting*. Oxford: Oxford University Press.

Quinn, Kevin et al. 2010. "How to Analyze Political Attention with Minimal Assumptions and Costs." *American Journal of Political Science* 54 (1): 209–27.

Roberts, Margaret E. et al. 2014. "Structural Topic Models for Open-Ended Survey Responses." *American Journal of Political Science*. DOI 10.1111/ajps.12103.

Rosenbaum, Paul R., and Donald R. Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70 (1): 41–55.

Sekhon, Jasjeet S. 2009. "Opiates for the Matches: Matching Methods for Causal Inference." *Annual Review of Political Science* 12: 487–508.

Theriault, Sean M. 2008. *Party Polarization in Congress*. Cambridge: Cambridge University Press.