SYMPOSIUM

Big Data, Causal Inference, and Formal Theory: Contradictory Trends in Political Science?

Introduction

William Roberts Clark, *Texas A&M University*
Matt Golder, *Pennsylvania State University*

Political science is experiencing two methodological "revolutions." On the one hand, there is the "credibility revolution," a movement that emphasizes the goal of obtaining secure causal inferences (Angrist and Pischke 2010). On the other hand, there is the "big data revolution," a movement that emphasizes how our increasing ability to produce, collect, store, and analyze vast amounts of data is going to transform our understanding of the political world "and even ameliorate some of the most important, but previously intractable, problems that affect human societies" (King 2014, 166). These "revolutions" have largely separate origins and have generally developed in isolation to one another. Despite this, they are both having an enormous impact on political science—they are changing how we think in terms of the methods we use to learn about the political world, the types of questions we ask, the research that makes it into our journals, and the way we train our graduate students. This symposium grew out of a desire to examine the potential tensions and complementarities between these two trends, and how each revolution interacts with the role of formal theory in political science research. We asked the symposium contributors whether big data, causal inference, and formal theory were contradictory trends in political science.

On the face of it, there would appear to be nothing contradictory about these trends. More data is better than less, and the attention to research design that accompanies the renewed interest in causal inference is integral to the maturation of our science, as is the clarity and rigor inherent in formal theory. Monroe et al. (2015) provide strong support for this point of view, highlighting the considerable complementarities that exist between data, theory, and research design. Specifically, they offer numerous examples of political scientists using big data to design innovative experiments, to gain empirical leverage on new research questions, and to develop new theoretical insights. The claim, though, that more data, better theory, and better research design is always good, while almost certainly true, is somewhat problematic.

Compelling formal models in economics long ago taught us that in a world of scarce resources, the conclusion that "if all three things are good, we should just do more of all three" is not always helpful. It may or may not be an optimal bargaining strategy for trying to convince others (the NSF, Congress, deans, and provosts) to enlarge the pie devoted to political science research. Either way, the resources that we have at our disposal at any time (such as hiring lines, courses offered in graduate programs, and dollars to spend on visiting speakers) are more or less fixed. At the margins, we have to decide how to invest our resources.

At this point, one might reason that "if more of *A* means less of *B* and *C*, then we must simply decide which of these three things is 'most important' and invest more in that." Such an approach might make good theater—who wouldn't want to see Rocío Titiunik, John Patty, Burt Monroe, and others engaged in a Battle Royal? Once again, though, that compelling formal model from economics tells us that our reasoning here might be misleading. There are circumstances in which the consumption of good *A* may actually increase the demand for good *B*. In other words, it may be possible that the "income gains" from progress in one area, say big data, can actually allow us to purchase more in other areas.

In our introduction, we use insights from a simple microeconomic model to frame the ensuing symposium discussion. To what extent are big data, causal inference, and formal theory substitutes or complements? Before we begin, we clarify some terminology. Perhaps unsurprisingly, the symposium contributors fail to agree on what is meant by *big data, formal*

doi:10.1017/S1049096514001759

theory, and *causal inference*. For us, "big data" refers to the idea that technological innovations such as machine learning have allowed scholars to gather either new types of data, such as social media data, or vast quantities of traditional data with less expense. "Formal theory" refers to the use of formal logic or mathematics to state and interrogate the implications of propositions about political behavior or institutions, predominantly in the form of noncooperative game theory. "Causal inference" refers to a recent trend in political science that questions the use of regression-based observational studies and advocates the use of alternative research designs, such as experiments, regression discontinuities, instrumental variables, and matching, to address threats to causal inference.

DOES BIG DATA CREATE OPPORTUNITIES FOR FORMAL THEORY AND CAUSAL INFERENCE?

One can think of big data as a process of technological change that results in a change in the price of data collection relative to other research activities such as theory building and research design. The standard microeconomic model tells us that a change in relative prices leads to changes in the way that scarce resources (time or money) are allocated between alternative uses. This change in the allocation of resources can be decomposed into a *substitution effect* and an *income effect*. The substitution effect resulting from increased efficiency in data collection leads to a bigger share of resources being allocated toward data collection. When technological change results in a dollar purchasing—or a unit of labor producing—more data than previously, resources may be shifted out of the production of formal theory or research design toward data collection. This substitution effect is already quite visible, with big data articles appearing with growing frequency in political science journals, and universities devoting increasing resources to programs in data analytics and the like.[1]

Arguably, recent technological change has made the collection of certain types of data cheaper than others. It seems to us that the differential in the ease with which different types of data can be collected has grown. While technology has given us almost immediate and virtually costless access to certain sources of voluminous data such as Facebook posts and Twitter tweets, it has not significantly altered the costs of collecting other types of data. For example, it is hard for us to see how data on, say, electoral rules or central bank independence is that much easier to collect today than in the recent past. In their contribution to the symposium, Nagler and Tucker (2015) make a strong case that social media data, and Twitter data in particular, can be used to address important substantive questions. The concern, however, is that the increasing demands on scholars to publish early and often, coupled with the ease and cheapness with which certain types of data can be collected, will drive the types of questions that scholars ask rather than substantive import.

The income effect may be less intuitive to some. For any given budget, a reduction in the price of a datum increases one's purchasing power with the result that we can spend more resources on other things. For example, we might spend more on formal theory and research design than on data collection, even though their relative prices have increased. In general, the relative magnitudes of the substitution and income effects depend on preferences and the status quo allocation of resources. The important thing to realize is that the efficiency gains in data collection could, in principle, free up resources to be spent on formal theory and research design. One positive thing, therefore, about technological change is that it means an increase in purchasing power, which means that the competition for scarce resources need not be zero-sum.

BIG DATA AND CAUSAL INFERENCE

Popular accounts of big data often laud the transformative power of big data and suggest that it can replace the scientific method, with theory and research design essentially becoming obsolete. The editor-in-chief of *Wired Magazine*, for example, writes that "Petabytes allow us to say: 'Correlation is enough.'...We can throw the numbers into the biggest computing clusters the world has ever seen and let statistical algorithms find patterns where science cannot" (Anderson 2008). This view implies that big data and causal inference are effectively substitutes for each other. As the contributors to this symposium all point out, these breathless accounts of big data almost always come from the mass media and computer science, not political science.

Political scientists have typically had more reservations about the promise of big data and have more quickly recognized potential pitfalls than their computer scientist counterparts. This is largely because of their training in statistical theory, which makes them aware of issues related to things like sampling populations, confounders, over fitting, and multiple hypothesis testing. Jeff Leek, a biostatistics professor, highlights this point on his blog, *Simply Statistics*, when he notes that many of the recent failures of big data in areas such as genomic science, economics, and public health have resulted from a lack of statistical expertise.[2] This concern is explicitly recognized by Grimmer (2015) in his contribution when he argues that it is time for big data scientists to become social scientists, not just computer scientists.

As Titiunik (2015) makes clear in her contribution, political scientists and computer scientists need to recognize that big data and causal inference are not substitutes. There is nothing about increasing either the number of observations or variables in a data set that solves the basic problem of causal inference. To identify a causal effect, we want to know how the outcome of interest is expected to change if the causal variable of interest were to change *while everything else* stayed constant. Credibly identifying such a source of exogenous variation requires both creative insight about, and substantive knowledge of, the phenomenon under study. As Ashworth, Berry, and Bueno de Mesquita (2015) note in their contribution, "no existing computational technique can come remotely close to [providing] either of these. For the time being, good empirical research will continue to rely on actual people doing actual thinking about social phenomena."

While big data is not a substitute for causal inference and good research design, it can be a complement. Proponents of matching argue that observational data can be used for causal inference when we can show that our observations are balanced on a host of observables and differ only in terms of

our purported cause and effect.[3] With small samples, it may not be possible to achieve good balance across all of the theoretically relevant covariates. To the extent that larger sample sizes are more likely to contain the necessary observations to achieve balance, big data can help with casual inference (Monroe et al. 2015).

Some scholars, of course, claim that "selection on observables," either through matching or regression analysis with controls, is problematic for drawing valid causal inferences. Researchers may fail to consider all of the possible confounding variables, or they may not be able to include them in their models because they are, in some sense or to some degree, intrinsically unobservable. One solution to the problem of potential unobservable confounding variables is to find an "instrument"—a variable that is correlated with the causal variable of interest, *X*, and which only affects the outcome variable, *Y*, through *X*. Big data can help to the extent that it makes previously unobservable variables observable, thereby reducing the need for an instrument, or by making new potential instruments available.[4]

As these examples illustrate, while scholars concerned about valid causal inference are generally more concerned about the *quality* of data than its *quantity*, big data can be helpful either by making new measures available or by increasing the sample size so that the number of observations with the desired quality increases. Big data is likely to contribute along these lines so long as the desired qualities of the data are not negatively correlated with the quantity of data.

Despite the potential for these positive complementarities, our contributors do raise some concerns about the interaction between big data and causal inference. Ashworth, Berry and Bueno de Mesquita (2015), for instance, suggest that there is a danger for "familiar mistakes" in empirical analyses to be exacerbated as the quantity of data increases. As an example, they point to how machine learning can automate and accelerate multiple hypothesis testing, thereby facilitating "the industrial scale production of spurious results." In his contribution, Keele (2015) reminds us "how bias can grow as sample size increases with observational data," with the possibility that we end up with a "very precisely estimated highly biased effect where the confidence intervals may no longer include the true value." Issues related to multiple hypothesis testing, omitted variables, and the like are obviously not new. The concern, however, is that they are likely to be more problematic in an era of big data.

BIG DATA AND FORMAL THEORY

What about the relationship between big data and formal theory? While Titiunik (2015) argues that big data is no substitute for theory or research design, she sees an important role for big data when it comes to description, exploration, and hypothesis generation. Grimmer (2015) makes a similar point when he argues that big data is particularly well-placed to help us with measurement issues and that "opportunities for important descriptive inferences abound in big data."

And so with data science, as with all science, the choice confronting the researcher is not between induction or deduction, or description and explanation, but between observation guided by conscious theorizing and observation guided by unconscious theorizing.

Whereas we agree with these contributors when it comes to the potential benefits that big data have for description and measurement, we believe that these benefits can only be realized to the extent that it is combined with theory. As Manski (2013) notes, and Keele (2015) reminds us, conclusions result from a combination of data and theoretical assumptions—without assumptions, data is just data. The importance of theory for description was recognized long ago by Popper ([1959] 2003), who claimed that induction was not so much wrong as impossible. Without a theoretical understanding of the world how would we even know what to describe? And so with data science, as with all science, the choice confronting the researcher is not between induction or deduction, or description and explanation, but between observation guided by conscious theorizing and observation guided by unconscious theorizing.

Let us consider two examples. In his contribution, Grimmer (2015) points to the *VoteView* project and its *Nominate* scores (Poole and Rosenthal 1997) as an example of how "purely descriptive projects" using big data can "affect the theories we construct and the causal inference questions we ask." We take issue with the claim that *Nominate* scores are purely descriptive. While *Nominate* scores technically reflect the *votes* of legislators, they are arrayed along an ideological dimension, thereby giving the impression that it is possible to infer something about legislator *preferences*. However, the jump to preferences requires an underlying theoretical assumption that it is preferences alone that *cause* legislator votes.

In their contribution, Nagler and Tucker (2015) claim that social media data such as Twitter provide access to the "unfiltered" opinions of individuals. Tweets or Facebook posts, however, like legislative votes, are forms of public behavior. As such, one needs to consider that individuals are likely to act strategically in what they reveal about themselves and to whom they reveal it. A moment's introspection should reveal that individuals regularly self-censor their posts or tailor them to achieve their goals with respect to particular audiences. Without a theory to delineate the strategic incentives in a given context, it would be difficult if not impossible to make accurate descriptive inferences from social media data.

As these examples illustrate, theory should play an important role when big data is used to make descriptive inferences. In their contribution, Patty and Penn (2015) show how formal theory can also be helpful when it comes to measurement,

before we even attempt to draw inferences. Many of the areas in political science associated with big data, such as those dealing with networks, text analysis, and genetics, require researchers to reduce high-dimensional data into lower-dimensional measures that can be used in empirical analyses. There are often many ways in which this can be done. As Patty and Penn (2015) illustrate in the context of social networks, formal theory, in particular social choice theory, can play an important role in identifying the most appropriate data reduction technique. Indeed, they conclude that "formal theory is the heart of measurement."

CAUSAL INFERENCE AND FORMAL THEORY

Although we are generally optimistic about the potential complementarities that exist between big data and causal inference, as well as between big data and formal theory, we finish by noting some potential tensions that exist between causal inference and formal theory. In their contribution, Ashworth, Berry, and Bueno de Mesquita (2015) argue that formal theory and causal inference are inherently complementary enterprises. This is because formal theory generates all-else-equal claims and causal inference tests all-else-equal claims. We see a tension, however, in the *type* of all-else-equal claims made by formal theory and causal inference.

Scholars emphasizing the requirements for valid causal inference take the randomized control trial as the gold standard. In this framework, a potential cause is a treatment that can be randomly assigned. The treatment is considered a cause if our manipulation of it results in a substantive change in our outcome variable. Thus, to demonstrate causality we manipulate something in an experiment or, failing that, we look for a natural experiment, a naturally occurring example of variation in our purported cause where all other potentially relevant variables are held constant. This view of causality can lead to a radical empiricism. "Theory" is superfluous. *Why* did the outcome variable change? Because our research design allowed us to demonstrate *what* happened when we manipulated the treatment.

The tension of this approach with formal theory becomes more evident when we realize that the all-else-equal claims that come out of formal models are almost always conditional in nature. The causal mechanism identified in a formal model is, by necessity, dependent on a particular context defined by the model's simplifying assumptions. Moreover, the propositions that are derived from these models tend to be conditional as well, something like: "above some cut point in variable Z, the propensity of the actors to engage in behavior Y is increasing in variable X; below that cut-point, Y and X are unrelated" (Brambor, Clark, and Golder 2006).

If these are the kinds of comparative statics that come out of our theoretical models, then it is not enough to hold "all other variables constant," which is essentially what randomization does in causal inference studies. It really matters at what values these other variables are being held. Our basic point is that theory, formal and informal alike, often calls our attention to the fact that context matters, and it is sometimes hard to read the causal inference literature as anything other than an almost single-minded attempt to use research design as a tool for stripping observations of as much context as possible.[5] Given this, it is little surprise that causal inference studies often seem to be more interested in cleanly identifying unconditional "causal effects" than testing the implications of existing theories (Huber 2013).

To some extent, the issue here can be seen in terms of the distinction between internal and external validity (Shadish 2010; Shadish, Cook, and Campbell 2002). Internal validity refers to our ability to determine whether the observed variation between our treatment variable and our outcome variable reflects a causal relationship. In contrast, external validity is about determining whether our treatment effect differs across people and settings, something we have referred to more generally as context. While randomization is key to ensuring internal validity, it does little, if anything, to address external validity. Our concern then is with how causal inference studies tend to neglect external validity in favor of maximizing internal validity.

If theory points to the importance of conditional effects, then this needs to be explicitly incorporated into one's research design. Yet, it is often difficult to evaluate the modifying effect of context in a strict causal inference framework. Perhaps the best strategy is to use a factorial design, because this allows researchers to examine the interactive effect of multiple treatments. However, an informal survey that we conducted of the articles published from 2009 through 2013 in the *American Journal of Political Science*, the *American Political Science Review*, and the *Journal of Politics* reveals that the use of factorial designs is relatively rare. Significantly, the studies that did use factorial designs tended to report only unconditional treatment effects and not whether these effects differed in significant ways across the various treatment contexts. As is well-known, the fact that unconditional treatment effects might look different, or indeed be statistically significant in one context but not in another, is not necessarily evidence of a significant causal modifying effect (Berry, Golder, and Milton 2012). More important, the degree to which factorial designs can take account of context depends on our ability to randomize contextual variables, and it is all but impossible to randomize key theoretical variables such as gender, race, geography, or institutions. It is also difficult for these designs to deal with continuous contextual variables.

A much more common way that experimental and quasi-experimental studies attempt to take account of context is by interacting a randomized treatment variable with some observable contextual variable. The addition of an observable variable in this way, however, immediately compromises the internal validity of the analysis.[6] Whereas these studies allow one to cleanly identify the causal effect of the treatment in each distinct context, they do not allow one to identify the *causal* modifying effect of context. And it is this causal modifying effect, as we have argued, that is often central to our theoretical claims.

An additional area of tension between formal theory and causal inference has to do with the kinds of randomized treatments that are used in experimental studies. It is often difficult to imagine the game-theoretic model that would have produced the kinds of treatments observed in some studies. What, for example, are we supposed to learn about the effect of government programs that theory tells us should only be

targeted to political supporters or adversaries from a study where groups of citizens are randomly exposed to these programs? These studies are looking at how individuals respond to out-of-equilibrium behavior. If our formal models do not place restrictions on out-of-equilibrium behavior, what can we learn about the empirical relevance of these models from such experimental studies? This particular tension between causal inference and formal theory can easily be resolved by keeping theory firmly in mind when choosing experimental treatments. Here again the point is that theory has a central role to play in research design; it should not be ignored in our search for clean causal effects.

Theory has a central role to play in research design; it should not be ignored in our search for clean causal effects.

We finish by returning to the definition of causal inference with which we began. We defined causal inference with respect to a particular set of methods, such as experiments, regression discontinuities, instrumental variables, and matching. We deliberately chose this particular definition because we suspect that this is how many people in our field view causal inference. We actually think that this view of causal inference is both potentially dangerous and misleading. It is dangerous because defining causal inference in terms of a particular set of methods that we accept as being indicative of good work and worthy of publication in our top journals is likely to result in us privileging certain theoretical questions over others in disregard to their substantive import (Huber 2013). This is because certain questions are more amenable to experimental analysis than others. It is rare, for example, for scholars to be able to randomly assign important institutions, such as electoral rules, to particular contexts.[7] To illustrate this, we point again to the informal survey we conducted of the articles published in *AJPS*, *APSR*, and *JOP* from 2009 through 2013. Of the 813 articles published during this time, 175 used a method commonly associated with causal inference. Of these, the overwhelming majority focused on behavior (160, 91%) as opposed to institutions (15, 9%).

To define causal inference in terms of a particular set of methods is also misleading because there are many ways of determining causality. Ordinary least squares and other regression-based models are, after all, causal inference methods—they allow us to draw causal inferences when their assumptions hold. For us, the value of the credibility revolution is in reminding us of the importance of research design in general, what Keele (2015) refers to as the "discipline of identification," and not in what some might consider a methodological fetishism where a clear distinction is drawn between "good" and "bad" methods without consideration of the particular trade-offs that are involved in addressing the research question at hand.

CONCLUSION

This brief introduction probably raises more questions than it answers. Our central goal has been to show that one could, as we do, see real value in big data, causal inference, and formal theory, and yet still be troubled about the way that these fit together. Our intention has been to highlight complementarities, while keeping an eye open for possible tensions between these approaches. We do not claim to have identified all such tensions, let alone resolved the ones that we have identified. If we have encouraged the reader to delve into the contributions offered by the wonderful scholars participating in this symposium, then we have accomplished our goal. Please read on, and help us keep the conversation going.

ACKNOWLEDGMENTS

We thank Charles Crabtree, Christopher Fariss, Sona Golder, and Luke Keele for their helpful discussions when writing this introduction. ■

NOTES

1. As one example, Pennsylvania State University is investing heavily in a big data social science initiative and is developing both a dual-title PhD program and an interdisciplinary undergraduate major in social data analytics. For more information, see bdss.psu.edu.
2. See http://simplystatistics.org/2014/05/07/why-big-data-is-in-trouble-they-forgot-about-applied-statistics/.
3. This strategy is essentially a large *N* application of Mill's method of difference (Sekhon 2010).
4. At this point, we hasten to add that "good instruments come from institutional knowledge and your ideas about the process determining the variable of interest" (Angrist and Pischke 2009, 117; Deaton 2010).
5. Conversations with Rob Franzese have repeatedly hammered this point home.
6. Our point here also highlights the fact that the distinction between experimental and observational studies is not as large in practice as is often claimed. These types of studies are better thought of not as discrete categories, but rather as points along an internal validity continuum.
7. Even if we could randomly assign institutions, there is always our previous concern that such random assignment would make little sense because such institutions may well be out-of-equilibrium choices for the context in which we are interested.

REFERENCES

Anderson, Chris. 2008. "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete." *Wired Magazine*. http://archive.wired.com/science/discoveries/magazine/16-07/pb_theory.

Angrist, Joshua D., and Jörn-Steffen Pischke. 2010. "The Credibility Revolution in Empirical Economics: How Better Research Design Is Taking the Con out of Econometrics." *Journal of Economic Perspectives* 24 (2): 3–30.

———. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton, NJ: Princeton University Press.

Ashworth, Scott, Christopher Berry, and Ethan Bueno de Mesquita. 2015. "All Else Equal in Theory and Data (Big or Small)." *PS: Political Science and Politics* 48 (1): this issue.

Berry, William, Matt Golder, and Daniel Milton. 2012. "Improving Tests of Theories Positing Interaction." *Journal of Politics* 74: 653–71.

Brambor, Thomas, William Roberts Clark, and Matt Golder. 2006. "Understanding Interaction Models: Improving Empirical Analyses." *Political Analysis* 14: 63–82.

Deaton, Angus. 2010. "Instruments, Randomization, and Learning about Development." *Journal of Economic Literature* 48 (June): 424–55.

Grimmer, Justin. 2015. "We Are All Social Scientists Now: How Big Data, Machine Learning, and Causal Inference Work Together." *PS: Political Science and Politics* 48 (1): this issue.

Huber, John. 2013. "Is Theory Getting Lost in the 'Identification Revolution'?" *The Political Economist*. Summer: 1–3.

Keele, Luke. 2015. "The Discipline of Identification." *PS: Political Science and Politics* 48 (1): this issue.

King, Gary. 2014. "Restructuring the Social Sciences: Reflections from Harvard's Institute for Quantitative Social Science." *PS: Political Science and Politics* 47 (1): 165–72.

Manski, Charles F. 2013. *Public Policy in an Uncertain World: Analysis and Decisions*. Cambridge, MA: Harvard University Press.

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.

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

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

Poole, Keith T., and Howard Rosenthal. 1997. *Congress: A Political-Economic History of Roll Call Voting*. New York: Oxford University Press.

Popper, Sir Karl. [1959] 2003. *The Logic of Scientific Discovery*. New York: Routledge.

Sekhon, Jasjeet. 2010. "The Neyman-Rubin Model of Causal Inference and Estimation Via Matching Methods." In *The Oxford Handbook of Political Methdology*. Eds. Janet M. Box-Steffensmeier, Henry E. Brady, and David Collier. New York: Oxford University Press.

Shadish, William R. 2010. "Campbell and Rubin: A Primer and Comparison of Their Approaches to Causal Inference in Field Settings." *Psychological Methods* 15 (1): 3–17.

Shadish, William R., Thomas D. Cook, and Donald T. Campbell. 2002. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Belmont, CA: Wadsworth.

Titiunik, Rocío. 2015. "Can Big Data Solve the Fundamental Problem of Causal Inference?" *PS: Political Science and Politics* 48 (1): this issue.

SYMPOSIUM CONTRIBUTORS

Scott Ashworth *is associate professor in the Harris School of Public Policy Studies at the University of Chicago. His research uses game theoretic models to study a variety of issues in political science, with a special emphasis on campaigns and elections. He can be reached at sashwort@uchicago.edu.*

Christopher R. Berry *is associate professor and faculty director of the Center for Data Science and Public Policy in the Harris School of Public Policy Studies at the University of Chicago. His research interests include metropolitan governance, the politics of public finance, and intergovernmental fiscal relations. He can be reached at crberry@uchicago.edu.*

Ethan Bueno de Mesquita *is professor, deputy dean, and faculty director of the Center for Policy Entrepreneurship in the Harris School of Public Policy Studies at the University of Chicago. His research is focused on applications of game theoretic models to a variety of political phenomena including political violence and accountability in elections. He can be reached at bdm@uchicago.edu.*

William Roberts Clark *is the Charles Puryear Professor of Liberal Arts and head of the department of political science at Texas A&M University. His research focus is on comparative and international political economy, with a particular emphasis on the politics of macroeconomic policy in open economy settings. He can be reached at wrclark@tamu.edu.*

Matt Golder *is associate professor in the department of political science at Pennsylvania State University. His research focus is on how political institutions affect democratic representation, with a particular emphasis on electoral rules and party systems. He can be reached at mgolder@psu.edu.*

Justin Grimmer *is associate professor in the department of political science at Stanford University. His research examines the role of communication in political representation, using machine learning techniques to study large collections of text. He can be reached at jgrimmer@stanford.edu.*

Luke Keele *is associate professor of American politics and political methodology at Pennsylvania State University. His research examines how causal inferences can be drawn from statistical evidence with applications in elections and voter turnout. He can be reached at ljk20@psu.edu.*

Burt L. Monroe *is associate professor of political science, as well as director of the Quantitative Social Science Initiative and principal investigator of the NSF-funded Big Data Social Science IGERT at Pennsylvania State University. His research is in comparative politics, examining the impact of democratic institutions on political behavior and outcomes, and social science methodology. He can be reached at burtmonroe@psu.edu.*

Jonathan Nagler *is professor of politics at New York University, and a codirector of the NYU Social Media and Political Participation (SMaPP) laboratory (smapp.nyu.edu). His research focuses on voting and elections, with an emphasis on the impact of the economy and issue positions of candidates on voter choice. He can be reached at jonathan.nagler@nyu.edu.*

Jennifer Pan *is a PhD candidate in government at Harvard University. Her research focuses on Chinese politics and the politics of authoritarian regimes, using automated content analysis and experiments to examine the interactions between autocrats and citizens in the absence of electoral competition. She can be reached at jjpan@fas.harvard.edu.*

John W. Patty *is professor of political science and director of the Center for New Institutional Social Sciences at Washington University in St. Louis. His research focuses generally on mathematical models of politics, and his substantive interests include the US Congress, the federal bureaucracy, American political development, and democratic theory. He can be reached at jpatty@wustl.edu.*

Elizabeth Maggie Penn *is associate professor of political science at Washington University in St. Louis. Her research focuses on mathematical models of politics and social choice theory, and her current interests include modeling the links between political institutions and political behavior. She can be reached at penn@wustl.edu.*

Margaret E. Roberts *is assistant professor of political science at the University of California, San Diego. Her research interests lie in the intersection of political methodology and the politics of information, with a specific focus on methods of automated content analysis and the politics of censorship in China. She can be reached at meroberts@ucsd.edu.*

Maya Sen *is assistant professor in Harvard University's John F. Kennedy School of Government. Her research interests include statistical methods, law, race and ethnic politics, and political economy. She can be reached at maya_sen@hks.harvard.edu.*

Betsy Sinclair *is associate professor of political science at Washington University in St. Louis. Her research focuses on the social foundations of participatory democracy—the ways in which social networks influence individual political behavior. She can be reached at bsinclai@wustl.edu.*

Rocío Titiunik *is assistant professor of political science at the University of Michigan. Her research focuses on political methodology, with an emphasis on the development and application of experimental and quasi-experimental methods to the study of political behavior and institutions. Substantively, her research focuses on incumbency advantage, political participation, and legislative behavior. She can be reached at titiunik@umich.edu.*

Joshua A. Tucker *is professor of politics and (by courtesy) Russian and Slavic studies at New York University, a codirector of the NYU Social Media and Political Participation (SMaPP) laboratory (smapp.nyu.edu), and a coauthor of the Monkey Cage blog at* The Washington Post. *His research focuses on mass political behavior in East-Central Europe and the former Soviet Union, with a particular emphasis on elections and voting, the development of partisan attachment, public opinion formation, mass protest, and social media. He can be reached at joshua.tucker@nyu.edu.*