METHODOLOGY

Computational Social Science: Discovery and Prediction. Edited by R. Michael Alvarez. New York: Cambridge University Press, 2016. 337p. $99.99 cloth, $34.99 paper.
doi:10.1017/S1537592716003327

— Scott de Marchi, *Duke University*
— Scotte E. Page, *The University of Michigan–Ann Arbor*

We still read Tocqueville in part because he had the good fortune to observe and write about two revolutions soon after they occurred in his native France and the United States. While the big data and computational revolution may not be as calamitous, it represents the largest transformation of the social sciences since the game-theoretic revolution that began with John Von Neumann and Oskar Morgenstern's 1944 book, *Theory of Games and Economic Behavior*. For that reason alone, R. Michael Alvarez deserves enormous credit for producing an impressive volume on these topics in the middle of the revolution.

In describing the big data revolution, the contributors emphasize the heterogeneity of new sources and types of data. BIG Data, or what might more accurately be called BIG Complex Data (BCD), includes real-time tweets during uprisings, multitudes of online polls and surveys, precinct-level voting results, texts of political speeches and hearings, and geocoded property values, incomes, and political contributions. These data can be fat—with many variables per observation—or tall—with millions of observations. We can embed the data in networks with links capturing cosponsors of bills, references in speeches, or citations in legal rulings.

The hydrant of BCD overwhelms traditional tools and has led to a bestiary of new techniques, ranging from Latent Dirichlet Analysis (LDA, used to analyze text), to random forests (used to identify features in fat data), to community-detection algorithms (on networks). These techniques originated in a variety of disciplines and are tuned to the peculiarities of those domains. Genetic data differ from Twitter data, which in turn differ from the text in bills before Congress. Yet these all get tossed into the BCD bin.

R. Michael Alvarez, the volume editor, set out with ambitious goals: i) to explain current best practices and methods, ii) to do so accessibly with a minimum of jargon, iii) to provide examples of how new data + new statistical techniques change the practice of social science, and iv) to be honest about challenges and shortcomings. We applaud the organization, the choice of topics and authors, and the extensive framing. The chapters within the volume largely succeed.

Space limitations preclude discussing each chapter in appropriate depth, and so we discuss a few exemplars. The chapter on LDA by Margaret Roberts, Brandon Stewart and Dustin Tingley succeeds on all four criteria. They invoke the mixture model of height in a population with gender as a latent variable to develop intuition for the way in which the LDA separates mixtures of topics. Latent variables and the number of parameters involved (where each topic is a vector over the words in the dictionary for a given corpus) cause the maximum likelihood surface to resemble a rugged landscape with many local optima. Failing to recognize this fact makes it difficult to substantively interpret the answers one gets from topic models. Put another way, every time one hits the enter key, one could get different topics. Their approach to solving this problem with the Structural Topic Model (STM) is useful not only for generating reliable results but also for educating researchers about challenges in understanding topic models.

The value of new forms of data—texts, tweets, blogs, events—and so on—is also a central theme of the volume. As Justin Grimmer notes in his chapter, with recent innovations in using text as data, one can examine the link between the rhetoric and presentational styles of legislators and their constituents.

All of the chapters demonstrate ingenuity in exploiting new sources of data with new statistical methods. Many chapters also demonstrate how data science can improve public policy. Brian Griepentrog et al. employ cluster analysis and model averaging to improve detection of pests introduced at U.S. borders and estimate the number of U.S. citizens abroad to increase their participation in elections; Ines Levin, Julia Pomares and R. Michael Alvarez apply machine learning algorithms to detect election fraud.

The revolution outlined in *Computational Social Science* has just begun, though, and it is difficult to judge its ultimate influence. One issue is that many chapters examine data sources that are of marginal interest to social science. For example, Betsy Sinclair's chapter on networks uses examples from a karate club at a Midwestern university in the early 1970s. Daniel Conn and Christina Ramirez apply random forest models to a California Health Interview Survey. Phillip Price and Andrew Gelman's chapter describes research on radon exposure, an important public health issue but not a central concern of the discipline.

While the challenges discussed in these chapters are in fact general—actors influence others via a network, data sets exist where the number of features exceeds the number of observations, and data often come from multiple sources—they would have a larger impact if the authors had identified analogous problems in mainstream political science where using new methods might improve existing answers.

Even when the chapters deal directly with social science data, the relevance to theory is tenuous because

this volume comments little on the relationship between these techniques and theory. Do they allow us to better test existing theories? Or do they enable us to construct richer theories and ask big new questions?

Grimmer's chapter, an *excellent* introduction to topic models, exemplifies this lack of engagement with theory. While the data set used in this chapter contains 170,000 press releases from 2005 to 2010, the test of the model depends on a much smaller sample and much less sophisticated analysis where Grimmer looks at 20 "spikes" in three of the 44 granular topics recovered by his model, and whether or not some of them correspond to real-world events. He concludes that in 2005, the Republican members claimed credit for spending on various projects in 23.2% of press releases, and by 2010 this had dropped to 9%. During this time, Republicans also increased the frequency of their attacks on the Obama administration, particularly on the issue of Obama's health-care plan. BIG Data, new techniques, but nothing new or even that interesting theoretically.

Similarly, the chapter by Joshua Tucker et al., takes on a fascinating topic: the connection between social media use and protest movements in Turkey and Ukraine. Again, we have a huge data set of ~30 million tweets in the case of Turkey and ~11 million for Ukraine. But the goal of the chapter is not to establish a causal link between participation in protests and social media use or to look at how social media use changed these protests. Rather, when one boils down the standards these authors present to validate their model, the dual goal is to determine whether or not people tweet more when there is a protest and whether important or violent events that occur midprotest cause a spike in tweets. Unsurprisingly, people do tweet as expected, but the proof in both cases is whether three (for Turkey) or four (for Ukraine) events correspond to a subset of spikes in Twitter usage during each protest.

Tucker et al. avoid the more interesting question of how Twitter and Facebook impacted the dynamics of the protests. In passing at the end of the chapter, the authors note that they contacted 16 people involved in the protest and found that 14 of them heard about the protests on social media. Ironically, they apply BCD to small questions and small data to the big question. We intend this not as a criticism but as a comment on the opportunity before us. The granularity of BCD may well allow for testing theories of the role of social media in political uprisings. For that to occur, we need more interaction between empiricists and theorists.

We believe that empirically minded researchers should read, teach, and engage this volume. We applaud the editor and authors for their creative use of previously unexplored sources of data and for mapping out the edges of the frontier of a new social science. We also encourage more theoretical-minded researchers to contemplate how new data connect to long-standing theoretical questions in social science about why people participate in politics, why nation-states engage in conflict, or how we might maintain the commons. Alvarez and the other authors are reaching out to you, the broader community. We hope you respond in kind. Without communication between empiricists and theorists, the revolution of BIG data could lack significance.

Process Tracing: From Metaphor to Analytic Tool.
By Andrew Bennett and Jeffrey T. Checkel. New York: Cambridge University Press, 2014. 344p. $99.00 cloth, $36.99 paper.
doi:10.1017/S1537592716003339

— Alexander Lee, *University of Rochester*

Social scientists spend much of their time making statements about cause and effect, and developing complex theories of causal relationships. The most basic way to test such theories is comparison of cases, whether a small number of case studies (in a qualitative setting) or a larger number of observations (in a quantitative setting). Practitioners of comparison techniques have tended to discourage making causal inferences within single cases (Gary King, Robert O. Keohane, and Sidney Verba, *Designing Social Inquiry: Scientific Inference in Qualitative Research*, 1994), due to problems of generalizability. However, there are many instances where a single case study may be the only viable research design.

In dealing with problems of causal inference in single cases settings, the "central" (p. 4) technique in political science is process tracing, a term first developed in cognitive psychology and appropriated by Alexander George (Alexander L. George, "Case Studies and Theory Development: The Method of Structured, Focused Comparison" in Paul Gordon Lauren, ed., *Diplomatic History: New Approaches*, 1979). Where multi-case studies attempt to infer causation from the correspondence of cause and effect, process tracing seeks to use the mechanism itself as evidence, to examine "whether the causal process a theory implies is in fact evident in the sequence and values of the intervening variables" (Alexander L. George, and Andrew Bennett, *Case Studies and Theory Development in the Social Sciences*, 2005, p. 6). Quite commonly, this involves examining the statements of decision makers involved in the process, but the technique can potentially be applied to the actions of individuals and groups as well.

As process tracing has become more popular in recent years, it has suffered something of a "buzzword problem," with the term being promiscuously applied to a wide variety of qualitative techniques with little link to the original idea. Moreover, much of the literature on the topic has been polemical in tone, advocating process tracing's efficacy relative to other techniques, particularly quantitative ones, rather than distinguishing good and bad examples and providing advice on techniques.