ArkPSA
Arkansas Political Science Association

Book Review *Computational Social Science: Discovery and Prediction*
Author(s): Yan Gu
Source: *The Midsouth Political Science Review*, Volume 18, 2017, pp. 81-84
ISSN: 2330-6882 [print]; 2330-6890 [online]
Published by: Arkansas Political Science Association
Website: http://uca.edu/politicalscience/midsouth-political-science-review-mpsr/

Book Review

Computational Social Science: Discovery and Prediction (Analytical Methods for Social Research). Alvarez, R. Michael, ed. 2016. New York, NY: Cambridge University Press.

Yan Gu
Henry M. Jackson School of International Studies
University of Washington

> Big data is like teenage sex: everyone talks about it, nobody really knows how to do it, everyone thinks everyone else is doing it, so everyone claims they are doing it.[1]

As one of the famous big data quotes from Dan Ariely states, big data in social science research has been a subject of much confusion: the term is broad, and its applications and directions are unclear. In this sense, this volume of *Computational Social Science: Discovery and Predictions* stands out as a pioneering effort to fill the hole. The editor, R. Michael Alvarez, presents an excellent introduction to the trends, strengths and challenges of the ongoing big data and computational revolution in social science, especially political science and public policy.

At first glance, the big data revolution is featured by data in new forms and of huge amounts, such as social media posts. For example, Joshua Tucker et al. (chapter 7) scraped 40,820,341 tweets related to two protests in Turkey and Ukraine to understand how social media affects individuals' decision to participate in protests. However, this book also shows that data of more traditional forms, for example text data, have also become rich sources of big data, due to the increasing availability of textual data in electronic format. In chapter 8, Justin Grimmer develops a collection of 170,000 House press releases to reveal what topics legislators discussed and how they expressed their priorities when communicating with constituents after the 2008 election. Moreover, big data is also about "identifying, compiling, merging, and prioritizing the large amount of data available" from different sources (248). Brian Griepentrog et al. (chapter 9) aggregate data from the U.S. government, the Organization for Economic Cooperation and Development, reports from national statistical agencies in other countries, etc. to create huge datasets.

[1] Dan Ariely. 2013. Facebook post. https://www.facebook.com/dan.ariely/posts/904383595868 (accessed May 29, 2017)

Nevertheless, one point that the editor highlights is that data itself, no matter how large it is, is not the core of the big data revolution. Instead, innovative methodological and analytical tools are the key for big data analysis. As Gary King puts it, "data is often easy to obtain and cheap, and more so every day" (vii). It is only in recent years that social scientists have been armed with advancements in statistical, computational, and machine learning tools that important questions can be tackled with big data as sources.

This book presents inspiring examples on how these tools help social scientists to make progress in answering questions that were previously difficult to study. Roberts, Stewart, and Tingley (chapter 2) detail topic modeling methods, specifically the structural topic model (STM), can be used to analyze large datasets of text and deal with the issue of multimodality. Supervised and unsupervised machine learning algorithms are employed to detect election manipulation, which are superior to simple linear models or distribution tests because they can detect more subtle forms of election fraud (chapter 10). Network modeling, ranging from visualization of networks to quantification of networks, and to community detection algorithms, help researchers discover interesting facts from computationally complex networks (chapter 4). Finally, open-source software, the trend of code-sharing, and the development of open-source platforms, enable researchers to access code easily. For instance, Tucker et al. developed a python package to scrape Twitter and Facebook, and this package is now available publicly at Github.

One major goal of this volume is to demonstrate how methods developed by statisticians and computer scientists can be adapted to social science research. The editor succeeds in this mission by breaking the volume into two parts. The first half volume introduces new tools, models, and initiatives developed by some of the leading scholars. The second part showcases some exploratory research projects as examples of how some new tools are applied to tackle important problems in social science. These projects cover a variety of topics in political science and public policy, demonstrating the potential of big data tools. For instance, the piece by Justin Grimmer (chapter 8) is an application of the STM that Roberts, Stewart, and Tingley introduce in chapter 2. This framing serves as an effective guidance for readers to learn.

A growing trend of cross-disciplinary collaboration in this emerging area of computational social science, as well as the challenges it faces, are

demonstrated in this book. Some chapters are contributed by teams of computationally-minded social scientists, while others are collaborations between social scientists, statisticians and computer scientists. It is reasonable to argue that innovations have been fueled by inter-disciplinary collaborations. However, since we are still in the beginning of this trend, scholars are calling for more efforts from academia, government, and the private sector toward this collaborative future. John Beieler et al. (chapter 3) appeal to establish the Open Event Data Alliance which collects and opens real-time political event data for scholars, analysts and practitioners. The differences between computer scientists and social scientists could be an obstacle to this future. Hanna Wallach, as a socially-minded computer scientist, characterizes some of the differences in the conclusion chapter. For example, scholars from different disciplines do not share the same norms, incentive structures, and research goals. Studies on social behaviors through online platforms conducted by computer scientists often aim for improving user experience and generating additional revenue, while social scientists use social media data to answer big-picture "why" questions. These issues need to be better addressed in the future.

Readers may remain unsatisfied by the claim made by many authors on integrating big data methods with existing social science methods in the volume. Many chapters touch on it but do not defend it soundly. Social science inquiry has an emphasis on causal inference, while many of the big data tools scholars have used so far are best designed for exploratory and observational analyses. For instance, network modelling can produce simple summary statistics about the characteristics of a network, but not directly draw casual inference. In chapter 8, Justin Grimmer explores large amount of text data and finds that Republican House members abandoned credit claiming after Obama's election, but what causal effects that this shift may have had on constituent response over time are not clear. To move forward, Grimmer suggests to pair random assignment with machine learning methods.

Another illustration is chapter 7, in which Joshua Tucker et al. study the relationship between social media and protests. Though they gather a huge amount of data, much of their article is not to discover how social media affects peoples' decision on participating protests. Instead, what they find is that increases in social media activities are associated with a number of key events during the protest. To further examine the causal effect, the authors conduct surveys of protesters. But they merely interviewed 16 individuals and only 2 of them initially heard about the protests on social media. It's not

convincing to argue that "it at least points in the direction of social media playing an important role in informing people about protests" as the authors suggest (217). The argument that big data methods can enhance traditional causal inference could have been stronger if the authors had better integrated the two approaches or the editor had chosen articles with stronger research designs.

In sum, this book is a useful resource to social science students, scholars, and practitioners who are interested in big data. It illustrates what and how big data methods can be used in social science research, and points out the opportunities and where it is possibly heading. It is honest in its shortcomings and shows the gap for futures researchers to fill.