Introduction

R. Michael Alvarez
California Institute of Technology

This volume has its origins in the rapid and staggering changes occurring in computational social research. As one of the editors of *Political Analysis* (an academic journal that publishes research articles in political methodology) and *Analytical Methods for Social Research* (a book series), I know I am witnessing a major shift in social science research methodology. Researchers have vast (and complex) arrays of data to work with; we have incredible tools to sift through the data and recognize patterns in that data; there are now many sophisticated models that we can use to make sense of those patterns; and we have extremely powerful computational systems that help us accomplish these tasks quickly.

When I was in graduate school in the late 1980s and early 1990s, those of us who worked with survey and public opinion polling data were considered "big-N" researchers in the social sciences. When I teach introductory research methods in my graduate seminars at Caltech, I will often have students read the 1978 *American Political Science Review* paper by Steven J. Rosenstone and Raymond E. Wolfinger, "The Effect of Registration Laws on Voter Turnout." Today this paper seems straightforward to students: Rosenstone and Wolfinger simply collected information on state-by-state voter registration and administrative practices, and merged that with the November 1972 U.S. Census Bureau's Current Population voting supplement, which the authors report as having more than 93,000 respondents.[1] They then tested, using a relatively simple binary probit model, for the effects of various registration and election administration procedures on whether the survey respondent reported having voted in the 1972 federal general elections.

Most students of statistics, methodology, or econometrics today are familiar with the binary probit model and its near-cousin, binary logit. These are techniques that model the probability that an outcome is met (here, did a voter turn out in an election) based on the covariates or regressors on the right-hand side of the model. The parameters of the probit and logit model are typically fit via maximum-likelihood optimization. Today a student could use an off-the-shelf statistics software package and replicate the original Rosenstone-Wolfinger analysis, literally in the blink of an eye, on his or her laptop computer.

Because students are accustomed to having such powerful tools as probit and logit, they often do not understand just how far applied social scientific research has advanced in the past few decades. For example, most of my students never even notice a critical sentence in the Rosenstone-Wolfinger paper: "Because of the tremendous cost of estimating a series of equations using all respondents in our sample, we took a subsample for the probit analysis" (page 28). Sometimes students see that statement and, scratching their heads, ask "What cost"?

If students are really interested, they may then read footnote 31 of the Rosenstone-Wolfinger article for further detail, which usually confuses them more: "Using the entire sample would have required about 50 *minutes* of computer time for each probit equation estimated" (page 28). Few students today realize that back when Rosenstone and Wolfinger were estimating the models for their paper, when probit models would take 50 minutes to run, they were likely using a mainframe computer. Accessing computer time usually required payment in the form of university research funds. These computational barriers seriously limited social science research.

I entered graduate school at a time when applied social science research computing was transitioning from mainframe computing to the personal computer. Some of my early applied research used what then would have been considered "big-N" American National Election Study (ANES) data sets, and employed maximum-likelihood techniques like probit and logit on mainframe computers at Duke University. A research project then required considerable thought and planning: if the university did not have the tapes containing the particular ANES data set that I wanted, we had to order the tapes from the ICPSR. Once the tapes were on campus and ready to be accessed, I had to write and submit jobs that would use IBM Job Control Language (JCL) to access the data set and to initiate some analytics routine using SAS language that followed the JCL. If I submitted a simple job in the afternoon, on the way to my office the next morning I could usually stop by the computing center for my printout. I could always tell, when entering the computer center, which jobs I had executed correctly: if there were boxes of green-and-white computer printouts with my name on them, that was a bad sign and usually implied thousands of pages of error messages; if there was a single sheet that usually meant that my JCL was a problem. A printout of a few dozen pages was usually a welcome sight.

Thus, technology limited what you could do as an applied researcher. Interactive and exploratory data analysis was impractical and in most cases impossible. Conducting extensive specification searches was time consuming. Running robustness tests, testing for the validity of important assumptions, and running other forms of model validation were seldom done. Integration, merging, or appending new data was difficult and time consuming, and typically not of interest. Estimation results were largely limited to the coefficient estimates and standard errors produced from the job's output file; producing interesting counterfactual estimates or fancy and colorful graphics was generally impossible and seldom done.

By the time my professional career began, it was possible to use computationally intensive iterative techniques like simple maximum-likelihood routines; yet much of the applied social science toolkit was based on tools that worked well in a computational environment of limited physical storage and limited memory. Tools like analysis of variance and simple ordinary least squares regression were the predominant methodological approaches in social science research, not because they were the best tools, but because they could be used in the computational facilities we had access to.

My point is that, because we could not combine interesting data sets easily, test the robustness of our assumptions readily, or produce informative secondary analyses to explain our results, the development of computational social science was hindered considerably. The computational resources available to researchers even shaped researchers' approaches, while limitations of physical storage and computer memory inhibited the use of large data sets and memory-intensive methods.

CHANGE IS GOOD

However, in the years since my graduate training, things have changed dramatically. Applied social science research has been greatly shaped by technological and computational innovations, and those changes are rapidly altering how we train our students, how we do research, and the types of problems we can study. Although there are many factors driving these changes, among the most important ones are the following:

- Easy and fast access, via the Internet, to data resources and databases. Whether through databases like the ICPSR, the Roper Center, or the Pew Research Center or from individual researchers themselves, today we have access to vast quantities of data that, through a speedy Internet connection, can be downloaded nearly instantly.
- Inexpensive computational power, including large amounts of inexpensive memory and physical storage. Cloud-based file storage means that large data arrays can be stored and accessed without straining local computer systems, and with parallel processing and cloud-based computation, very large-scale estimation problems can be made easily tractable.
- New forms of data (especially text) that can be easily obtained from many sources. The social media revolution – blogs, Facebook, Twitter, Tumblr, Instagram, and so on – is producing huge amounts of readily available data each minute.
- Open-source software and a culture of code-sharing. Most data scientists today use open-source platforms for data manipulation and modeling; for

example, R, Python, and Perl; code-sharing through platforms like GitHub makes such code easily accessible for researchers. It has also become "cool" to know how to write code, and with more researchers and practitioners writing sophisticated code using powerful analytical software languages, we will continue to be able to tackle ever more complex data sets and problems.

- More emphasis on and acceptance of multidisciplinary research. It is increasingly common for social scientists to work on data science teams with computer scientists, engineers, and those with other academic and professional backgrounds; research is more readily transported across traditional academic boundaries than in the past, facilitated by the increasing number of publication outlets and online communication of research products.

Of course, other transformations are occurring, but the aforementioned are some of the most significant forces. We see the results in articles in the leading journals of social sciences. For example, the first issue of 2015 for *Political Analysis* has three papers utilizing estimation methods that are computationally intensive: either via Bayesian methods (Traunmuller et al. 2015; Si et al. 2015) or using copula functions to model multiple political processes simultaneously (Chiba et al. 2015). Two papers focus on large-scale data generated from online activities, either Twitter status updates (Barbera 2015) or experiments conducted online (Guess 2015). And two papers undertake innovative causal modeling, one to model the incumbency advantage in Brazil (De Magalhaes 2015) and the other using voter registration data that has been merged with property sales records (Keele and Titiunik 2015). These studies would have been impractical or impossible to implement when I started my career: the data was not available, the files would have been difficult to store and manipulate, and the estimation approaches would have been computationally difficult or infeasible.

Some may grumble that all of this is just hoopla, that there is little going on other than researchers using increasingly sophisticated tools and data sets to answer the same old questions. Papers from the most recent issue of *Political Analysis* suggest that this is far from the case – the new data sets and analytical opportunities allow researchers to examine questions that could not be easily studied before and to conduct interesting and important research in cost-effective ways. And as readers will see in each of the chapters in this book, new data and new tools are allowing researchers to look at new questions and to examine the social world in ways that were impossible just a decade ago.

WHAT THIS BOOK IS ABOUT

The chapters in this book focus on some of the extraordinary work being conducted in *computational social science* – in academia, government, and the private sector – while highlighting current trends, challenges, and new directions. But it is important to tell readers what this book is about and the

Cambridge University Press
978-1-107-10788-5 - Computational Social Science: Discovery and Prediction
R. Michael Alvarez
Excerpt
More information

intentions behind putting this book together. This book is not about "big data." As Gary King notes in the Preface, the methodological and analytical changes focused on in this volume are really not about *the data*. Instead, the focus is on the methodological innovations driven by the availability of new types of data or by data of a different scale than has been previously possible. In other words, the chapters in this book all hinge on innovative computational tools used to answer lasting and important questions in social science and public policy, and thus I've titled the book *Computational Social Science*. Again, it's not the data, itself nor the "size" of the data that is critical to the innovations in this volume; the emphasis rather is on the cross-disciplinary applications of statistical, computational, and machine learning tools to the new types of data, at a larger scale, to learn more about politics and policy.

These innovations are difficult to compartmentalize and categorize; it is not easy to determine where computational social science is heading. In a recent symposium in *PS: Political Science & Politics*, a primary publication of the American Political Science Association, symposium editors William Roberts Clark and Matt Golder begin their introductory essay noting that the "big data revolution" is changing political science (Clark and Golder 2015). I suspect that one could easily find similar remarks from prominent economists, sociologists, epidemiologists, statisticians, and so on. In an effort to better understand developments in this field, I reached out to researchers developing new methodologies and innovative ways of looking at the new data that is proliferating at a rapid rate. Those contributions are included in the first part of this volume, which is titled "Computational Social Science Tools." The other set of contributions are from researchers using these new tools and methodologies to tackle important problems in social science and public policy. Those chapters are in the second part of the volume, "Computational Social Science Applications." The primary intention of the book is to showcase the new methodological tools being developed and applied by some of the leading researchers, and then to explain how academics and the private sector are using many of these tools to solve problems in social science and public policy.

The contributions are mainly focused on political science and public policy. I had originally cast a wide net, seeking contributions from economics, sociology, social and cognitive neuroscience, and psychology. Yet I realized that, by focusing on politics and policy, a sharper definition could emerge from the exciting new computational work in these areas. The focus on political science and policy also helps identify important new research opportunities (I'll return to those later in this Introduction), which are also touched on in many of the chapters and in the book's conclusion.

The first chapter is by Christopher Warshaw: "The Application of Big Data in Surveys to the Study of Elections, Public Opinion, and Representation." Warshaw is a leading scholar in the development of methods to link multiple public opinion polls or surveys, having already published a series of innovative and important papers in this emerging area of computational social science

(Caughey and Warshaw 2015; Tausanovitch and Warshaw 2013; Warshaw and Rodden 2012). Much of Warshaw's research agenda has focused on linking multiple polls or surveys, usually polling data collected by commercial survey research firms, and then using the linked data sets to study historical trends in public opinion after making statistical adjustments to ensure that the data analyses provide reasonable and valid inferences. Other components of Warshaw's work, including the research he presents in Chapter 1, use large-scale and big-N surveys and innovative statistical techniques to study political representation. As he argues, by using these new techniques, researchers can answer questions about representation and public opinion that they could not answer before. He points to a number of ways that this new approach to the study of public opinion and political representation can be improved, especially through the use of new causal inference methods in the large-scale public opinion polling framework.

The second chapter, by Margaret Roberts, Brandon Stewart, and Dustin Tingley, "Navigating the Local Modes of Big Data: The Case of Topic Models," focuses on text data. Their chapter looks closely at important analytic methods for studying text data, a topic these authors have a great deal of experience with.[2] The analysis of large-scale databases of text has become a growth area in recent years, in particular because so much textual information is readily available in electronic form, either through databases such as Lexis-Nexis, where data is produced by scraping text from websites, or from social media tools like Twitter or Facebook. Many methodologies have been developed to search through these large databases of text; one recent tool is the Latent Dirichlet Allocation (LDA) model, which is a type of "topic model" (Blei, Ng, and Jordan 2013; Blei 2012, 2014). Topic models are popular techniques because they search for underlying or "latent" commonalities across textual data based on word patterns reappearing in that data. Another way to think about topic models is that they are really data reduction techniques for very large arrays of textual data; by looking for repeating patterns of text they can identify patterns or themes that help classify or categorize textual data.

In Chapter 2, the authors examine a particular issue that can plague topic models like the structural topic model (STM, a type of LDA model that these authors have written about elsewhere; Roberts et al. 2014). The issue is multimodality and how it presents a problem for the estimation of topic models and the analysis of textual data. The problem of multimodality is a general one in optimization. Say that the underlying data-generating process in the real world really comprises two different phenomena. Consider Hillary Clinton. Assume that we had set up a Twitter tool some time in 2010 to track tweets associated with Clinton, and that tool has been collecting data since that time. We would have coverage that spans her time as secretary of state and then the early stages of her potential presidential candidacy (I'm writing this Introduction in the summer of 2015). It is likely that analyses of this data (for example, a topic model or some type of sentiment analysis) would encounter the problem of multimodality; the way in which people tweeted about Clinton while she was the chief diplomat for the United States is quite likely very different from the way people tweet about Clinton now. Roberts and her colleagues discuss these problems in textual data and how topic models like the STM deal with them.

The third chapter is by John Beieler, Patrick Brandt, Andrew Halterman, Philip Schrodt, and Erin Simpson, "Generating Political Event Data in Near Real Time: Opportunities and Challenges." This team focuses on methods to collect "event data," information that records an occurrence at a particular point in time. These event data are popular in the study of international relations, where for example scholars wish to collect data about international events, such as conflict or cooperation among nations, and then analyze these data from different perspectives – exploiting the dynamic nature of such data or the ways in which data like these can yield important patterns about networks or network effects. In the recent past, event data was typically collected and "coded" (translated into information that is machine readable) by human researchers; although human coding can often allow complex nuances or contexts to be deduced from the original source information (for example, a newspaper), such coding can be inaccurate, time consuming, and costly.

Machine coding of events data is a project that one of the authors, Philip Schrodt, has been engaged in for a considerable time. In particular, his work on the Kansas Event Data System (KEDS) is an early example of the utility of machine coding of large quantities of textual information into event data (Schrodt 1998). In Chapter 3, Schrodt and his colleagues discuss how the changing mass media environment, where media sources themselves are now fluid and dynamic, is altering how scholars approach the machine coding of event data; in particular they discuss how their Open Event Data Alliance (OEDA) approach provides a model for the development of large event data sets. The OEDA model, based on the principles of open data, open source code, and standardization of approaches for machine coding of event data, is an interesting one that might serve as the foundation for similar machine-driven data collections initiatives that will undoubtedly appear in the near future. Scholars and practitioners considering the merits of open data, open source code, and standardization of coding approaches in other areas where large arrays of textual data are collected and coded – for example, the study of social media – will find this chapter particularly interesting.

Next, in Chapter 4 Betsy Sinclair discusses modeling networks, a rapidly growing question in a number of academic fields, in the private sector, and in many different areas of public policy. Sinclair's recent book on political networks provides examples of the importance of studying networks and also demonstrates how difficult studying networks can be (Sinclair 2013). Much of the existing theoretical literature in political science and most of our methodological tools focus on individuals, organizations, candidates for office, politicians, nations, and so on; they are considered independent entities. For example, most of our theories and models of individual voting behavior assume that voters are making their decisions independently of the decisions of their family, colleagues, or friends. A typical empirical model of presidential voting behavior will include such factors as issues, partisanship, ideology, and demographic controls as explanatory variables, but will not typically include variables that might account for how voting decisions might depend on information from voters' friends, neighbors, family, or coworkers.

Sinclair's chapter, "Network Structure and Social Outcomes: Network Analysis for Social Science," provides an excellent overview of why studying networks is important for social science and gives some important background for how social scientists approach network analysis. As she points out, whether the networks are small or large, they are all computationally complex, which is no doubt why the study of networks was neglected in social science until we got to a point where we had sufficient computing power. Sinclair notes that if there are N individuals under study, they might be connected to 2^N other individuals in a network. In a small population of ten individuals, that's 1,024 possible connections. As Sinclair discusses, there are two primary approaches to studying these large and complex patterns of interconnections: spatial statistical approaches that assume that connections are more likely the closer that individuals are in an unobserved space, and exponential random graph models.[3] Sinclair's chapter is a fantastic primer for readers interested in learning more about network modeling.

In Chapter 5, Peter Foley writes about the estimation of political ideology in "Ideological Salience in Multiple Dimensions." Measuring political ideology has a long history in political science, because it is a central concept in the theoretical literature, in the study of elections and voting behavior, and in research on legislators and political representation.[4] The study of political ideology continues today, with scholars producing new and innovative ways to estimate the concept for voters and representatives (see, for example, Bonica 2014 or Barbera 2015).

Foley builds on these various literatures, starting with the foundation of formal political theory. He uses the formal conception of political ideology, which represents ideology spatially, and generalizes to allow for two spatial issue dimensions and a weighting of each dimension to reflect the importance of the issue dimension to the decision maker. The model that Foley develops allows for the simultaneous estimation of ideological placement and issue dimension weights; unsurprisingly this turns out to be computationally complex. This approach to the simultaneous estimation of placements and dimensional weights has great promise in both academia and other applied settings; for example, political campaigns and estimating simultaneously how a candidate or campaign's message might influence both the importance of an issue to a voter and the voter's preference on the issue.

In the sixth chapter, Daniel Conn and Christina Ramirez discuss the application of machine learning techniques to another important area of public and policy interest: biomedical research. Their chapter, titled "Random Forests

Cambridge University Press
978-1-107-10788-5 - Computational Social Science: Discovery and Prediction
R. Michael Alvarez
Excerpt
More information

and Fuzzy Forests in Biomedical Research," provides an excellent introduction to random forests for the interested reader. The random forest approach is best understood through the intuition behind classification and regression trees (CARTs). These "tree" methods proceed by breaking a high dimensional covariate space into smaller spaces and then producing something that can be depicted as a "tree." As Conn and Ramirez discuss in their chapter, a random forest model is then really a method that can estimate many trees (the "forest"). These models are very helpful in situations where the analyst has a data set with many covariates and where there are many potential interactions between the covariates that are not necessarily well known a priori by the researcher. Random forest models have seen some use in social science settings (see, for example, Grimmer and Stewart 2013), and as Conn and Ramirez argue, they will be increasingly used in the future.

The second part of the book, "Computational Social Science Applications," highlights ways in which today's computational social science is being applied to important problems.

It begins with a chapter that looks at mass political participation and political representation in a novel way. The team from NYU's Social Media and Political Participation Laboratory (Joshua Tucker, Jonathan Nagler, Megan MacDuffie Metzger, Pablo Barbera, Duncan Penfold-Brown, John Jost, and Richard Bonneau) uses a large trove of Twitter data to study recent mass protests in Turkey and Ukraine. As Tucker et al. argue in their chapter, social media tools like Twitter have become part of the political protest landscape, and some have argued that social media now plays an important role in the dissemination of information during political protests. Others, probably most famously Malcolm Gladwell (2010), have been skeptical about whether social media has altered the course of recent protest movements. Thus the question of whether social media tools help resolve collective action problems (e.g., Olson 1971) and help protest organizers better mobilize potential supporters is an important and open one.

Tucker and colleagues posit five necessary conditions that, if met, would support the hypothesis that social media plays an important role in facilitating political protest. They then utilized a python tool to collect all tweets that use keywords or hashtags relating to protest in the two nations under study; this tool yielded more than 40 million tweets! The authors then used the data to document support for their necessary conditions and to argue that social media tools play an important role in the development of political protest – and thus in potentially changing the nature of political representation in each nation. This is powerful stuff; these researchers provide evidence that supports the hypothesis that social media tools might be changing the basic calculus of collective action. In their conclusion they present a research agenda for further analysis to document and examine how these changes may be reshaping political participation throughout the world.

"Measuring Representational Style in the House: The Tea Party, Obama, and Legislators' Changing Expressed Priorities" (Chapter 8) comes from Justin Grimmer, one of the leading scholars using textual data sets – and developing tools to analyze those data sets – to study political representation. His recent book uses textual data in a number of interesting ways to build on the seminal work of Fenno (1978), providing an important new way to study the question of how political representatives communicate with their constituents (Grimmer 2013). In his chapter, Grimmer employs a large database of press releases from members of the U.S. House of Representatives, which he analyzes using a topic model like the one that Roberts and colleagues discussed in Chapter 2. Grimmer's focus in his use of topic models, however, is on estimating the nesting of topics, developed from topic modeling approaches discussed by Li and McCallum (2006). The method Grimmer presents in his chapter partitions a set of narrower topics within broader, more general topics. Methodologically, this approach helps advance the literature on topic models. But this contribution also has interesting substantive implications, because Grimmer is able to demonstrate both the general and specific ways that representatives communicate with their constituents.

Chapter 9 is written by a team of authors from the Fors Marsh Group: Brian Griepentrog, Sean Marsh, Sidney Carl Turner, and Sarah Evans. In "Using Social Marketing and Data Science to Make Government Smarter" they note that governments today are looking to use their resources more effectively and to engage in the development of data-driven policy. Simultaneously, increasing amounts of useful data are now available to governments, typically in two forms. Grienpentrog et al. distinguish between "tall" data (with a very large number of rows or observations) and "fat" data (with a very large number of covariates or predictors).

The authors discuss two government-sponsored research projects that illustrate how "tall" and "fat" data are being used by U.S. agencies to improve their decision making. The "tall" analytics example comes from a project for the U.S. Department of Agriculture, where they used data collected from agricultural inspections of travelers entering the United States, augmented by other data. The analysis helped the department improve outreach and marketing efforts, so as to prevent further risks to U.S. agriculture from diseases and pests originating abroad. The second example in their chapter looks at "fat" data through how their team has worked with the U.S. Department of Defense to estimate the population of American citizens abroad. This study involved the collection of a large (i.e., "fat") array of data arranged by country-years, and a model-based attempt to estimate, country-by-country, the number of American citizens residing, working, or studying in each country outside the United States. These estimates are being used by the Department of Defense in marketing and outreach efforts to help improve voter enfranchisement of American citizens abroad.