ANNUAL REVIEWS

Annual Review of Political Science

The Challenge of Big Data and Data Science

Henry E. Brady

Department of Political Science and Goldman School of Public Policy, University of California, Berkeley, California 94720, USA; email: hbrady@berkeley.edu

Annu. Rev. Political Sci. 2019. 22:297–323

First published as a Review in Advance on January 21, 2019

The *Annual Review of Political Science* is online at polisci.annualreviews.org

https://doi.org/10.1146/annurev-polisci-090216-023229

Keywords

big data, data science, artificial intelligence, cyberinfrastructure, causality, prediction, text analysis, internet, smart cities, cyber-warfare, automation

Abstract

Big data and data science are transforming the world in ways that spawn new concerns for social scientists, such as the impacts of the internet on citizens and the media, the repercussions of smart cities, the possibilities of cyber-warfare and cyber-terrorism, the implications of precision medicine, and the consequences of artificial intelligence and automation. Along with these changes in society, powerful new data science methods support research using administrative, internet, textual, and sensor-audio-video data. Burgeoning data and innovative methods facilitate answering previously hard-to-tackle questions about society by offering new ways to form concepts from data, to do descriptive inference, to make causal inferences, and to generate predictions. They also pose challenges as social scientists must grasp the meaning of concepts and predictions generated by convoluted algorithms, weigh the relative value of prediction versus causal inference, and cope with ethical challenges as their methods, such as algorithms for mobilizing voters or determining bail, are adopted by policy makers.

BIG DATA AND DATA SCIENCE

"Big data and data science are being used as buzzwords and are composites of many concepts," says the US National Institute of Standards and Technology (NIST) in a 2015 "framework" report on "big data" (NIST 2015, p. 2). The phrase "big data" appears frequently in the press and in academic journals, and "data science" programs have sprouted in academia over the last five years. On March 29, 2012, the White House Office of Science and Technology Policy announced the "Big Data Research and Development Initiative" (Kalil 2012) that builds upon federal initiatives "ranging from computer architecture and networking technologies to algorithms, data management, artificial intelligence, machine learning, and development and deployment of advanced cyberinfrastructure" (NITRD 2016, p. 6). "Big data" appeared about 560 times per year in JSTOR from 2014 through 2017 even though it was mentioned less than once a year in the century before 2000 and only an average of about eight times a year between 2001 and 2010. In the last five years, at least 17 Data Science programs have started at major American research universities (**http://msdse.org/environments/**), and the internet is replete with advertisements for data science books and courses, often with the come-on of "Become a Data Scientist." The phrases have certainly caught on, but they mean different things to different people, and some even doubt that they identify something very new or useful (e.g., boyd & Crawford 2012, Donoho 2017, Smith 2018).

Despite the imperfection of these terms and the hyperbole that often surrounds them, they point to real changes that are important for political science. Big data, data science, and the related ideas of artificial intelligence, cyberinfrastructure, and machine learning contribute to the following developments and trends discussed in this article:

- *Societal and political change from big data and data science*. The volume, velocity, variety, and veracity of data being generated by and available to governments, armies, businesses, nonprofits, and people have combined with the enormous increases in computing power and improvements in data science methods to change society in fundamental ways. Big data and data science are creating new phenomena and raising basic questions about the control and manipulation of people and populations, the future of privacy, the veracity of information, the future of work, and many other topics that matter for political scientists.
- *Increasing amounts of data available to all scientists, including political scientists*. All the sciences are being affected by these changes. The Thirty Meter Telescope coming online in 2022 will generate 90 terabytes every night; genomic data are doubling in quantity every nine months and are currently being produced at approximately 10 terabytes per day; the Large Hadron Collider at CERN generates 140 terabytes per day. The World Wide Web produces about 1,500,000 terabytes every day, and this flow of data offers social scientists a chance to study the "sinews of society" (Weil 2012) and the "nerves of government" (Deutsch 1963) in a way that could not be done in the past. Now political scientists can observe and analyze (sometimes in real time) the information that people choose to consume, the information produced by political actors, the environment in which they live, and many other aspects of people's lives.
- *New ways political scientists organize their work*. With this onslaught of data, political scientists can rethink how they do political science by becoming conversant with new technologies that facilitate accessing, managing, cleaning, analyzing, and archiving data.
- *New kinds of questions asked by political scientists*. Political scientists must ask what they are trying to accomplish with concept formation, description, causal inference, prediction, and projection into the future. In the process, new methods and insights will be developed about political behavior, and new designs will be put forth for political institutions.

- *Dealing with ethical issues regarding political science research*. Finally, political scientists must think about complicated ethical issues regarding access, use, and broadcasting of information, and the possible misuse of their models and results.

Before considering these five changes and their implications for political science, I describe the exponential growth in data and computing power that has led to the prominence of so-called big data and data science, followed by definitions of these untidy phrases.

INCREASING VOLUME, VELOCITY, AND VARIETY OF BIG DATA

Social scientists must come to grips with the current dramatic transformations in the communication of information, which parallel the striking changes in transportation in the nineteenth century. In 1816, using horse-driven stagecoaches, mule-driven canal boats, or sailing packets, a trip between Philadelphia and Quebec took more than four days. By 1860, with the advent of steam-driven trains and steamboats, the time and cost for travel dropped by over two-thirds, and the same trip took just over one day (estimated from Taylor 1951, p. 141). These changes created new trading networks, new opportunities for migration, new kinds of cities with commuter suburbs, and new understandings of the world, with enormous implications for politics, economics, and society.

Changes every 20 years in information technologies punctuated the history of the late nineteenth, twentieth, and early twenty-first centuries: telephones (1870–1890s), phonographs (1870–1890s), cinema (1890–1920s), radio (1900–1920s), television (1940–1950s), mainframe computers (1940–1950s), personal computers (1970–1980s), the internet and World Wide Web (1980–2000s), cell phones (1980–2000s), and smart phones (2000s–present). The most fundamental innovation came with the move from analog devices to digital ones, starting in the 1950s and proceeding dramatically in the 1990s and thereafter. These changes brought (*a*) extensive digital datafication, in which myriad events are now digitally recorded; (*b*) widespread connectedness, in which events and people are identified so that they can be linked up with one another; (*c*) pervasive networking, such that people are embedded in a community of interacting users who become nodes in larger networks; and (*d*) ubiquitous computer authoring, where computers create new information that becomes part of the social system and its culture.

Political scientists led the way in studying these changes. Harold Lasswell and Karl Deutsch were early students of communications and their impacts on societies. In 1983, MIT political scientist Ithiel de sola Pool looked at the production of words in the American mass media (e.g., radio, television, records, movies, newspapers, books) and point-to-point media (telephone, first-class mail, telegrams, facsimile, and data communication) from 1960 to 1977. Pool found that words in these media doubled every eight years, growing at about 9% per year. He also found that "print media are becoming increasingly expensive per word delivered while electronic media are becoming cheaper," so that "growth in both mass and point-to-point media has been greatest in the electronic ones." Furthermore, "although the largest flow of words in modern society is through the mass media, the rate of growth is now fastest in media that provide information to individuals, that is, point-to-point media." Finally, "the words actually attended to from those media grew at just 2.9% per year" so that "each item of information produced faces a more competitive market and a smaller audience on average" (Pool 1983, p. 609). Pool predicted much of what we know about modern communications: They are growing fast, they are increasingly electronic and point-to-point, and people experience information overload and fragmented information flows. Perhaps most presciently, Pool (1983, p. 611) also said, "Computer networking is for the first time bringing the costs of a point-to-point medium, data communication, down to the range of costs characteristic of mass media."

Subsequent studies by political scientists and others (Lyman & Varian 2003, Bohn & Short 2012) focused on the volume or stocks of information (e.g., the number of books in a bookstore) as well as on the flows or velocity (the daily sales of books) and the variety of information (subjects of books). They also measured information in digital bytes instead of words so that the measures reflect the proliferation of images, which communicate many more bytes per second than do words through text or speech (Bohn & Short 2012, p. 986). Hilbert & López (2011, p. 63, table 1) found that the world's storage capacity in bytes per capita doubled every 40 months between 1986 and 2007. The bulk of the world's flow of communications was still in broadcast communications, which grew at the rate of 6% per year per capita, but (point-to-point) telecommunications grew at the rate of 28% and could conceivably exceed broadcast communications within 10–15 years. Finally, they computed a new quantity—the growth in the world's computational power measured in millions of instructions per second (MIPS)—and they found that human-guided general-purpose computation grew at an impressive compound annual rate of 58% per capita between 1986 and 2007. Embedded applications-specific computation grew even faster, at 83%.

This research identifies four notable trends, briefly mentioned above, that have produced the big-data revolution: extensive digital datafication, widespread connectedness, networking, and computer authoring. First, there is a tsunami of data about societal events, and digital communications are overtaking analog. This extensive digital datafication (Cukier & Mayer-Schoenberger 2013, p. 29) creates data in a format that can be readily stored and processed by computers. "Recording" or "digitalization" might be used instead of the ugly neologism "datafication," but it seems too passive for processes that are transmogrifying human interactions into data. Even though some of these data are relatively unstructured (text, audio, networks, or images), data scientists are figuring out ways to analyze them. Second, there is widespread connectedness because point-to-point telecommunications can be, in principle, more easily tracked than broadcasting. For example, whereas broadcasters traditionally required elaborate survey operations (such as Nielson's media-use diaries) to track their audience, Netflix has immediate data on the download of its movies. More generally, we can now record and connect data on individual postings, purchases, police encounters, and even perambulations. Datafication and connectedness mean that once-ephemeral events can now be identified and studied.

A third feature of the changing information environment, networking, is especially important for social scientists. Whereas once communications were classified as either person-to-person (e.g., conversation, letters, or telephone) or mass communications from one source to many people (e.g., books, newspapers, cinema, radio, or television), modern communications involve mediated social networks that combine features of both modes (Neumann 2016, Schroeder 2018). Twitter involves individual communications sent to many followers using hashtags that define self-mediated areas of concern. Facebook involves individuals with customized profiles who have networks of "friends" and who have affiliations with common-interest user groups that share information. Google involves a query by an individual who is provided with a list of relevant websites. Amazon involves a search for a particular product that results in suggestions about other relevant products that can be bought online. In all these media, knowledge about people's characteristics and their search behaviors is used to suggest and sometimes impose particular actions or relationships. The implications of these new modes of communication are not clear, but they probably operate differently in the three important spheres of politics, markets, and culture (Schroeder 2018). They may also have important impacts such as increasing the chance for political polarization through the creation of networks that are closed to dissenting opinions (Neumann 2016).

Finally, whereas the communication of information traditionally involved sending messages in the most verisimilar fashion possible even when the message was transformed along the way (e.g., from voice into electrical signals in a telephone), an increasing fraction of information is partly, if

not entirely, computer authored. Computers use programs to produce new outputs that combine inputs in novel ways: A Google search takes a request and delivers plausible "answers" to that search; a computer game produces a fantasy virtual environment for entertainment; a Computer Automated Design program produces a design that meets certain specifications; and so forth. Nature and humans no longer have a monopoly on authoring. We now live in an era when computers can author, publish, and supply new forms of information. Another job of social science is to improve and understand these processes.

DEFINITIONS OF BIG DATA AND DATA SCIENCE

The growth of data and the creation of large databases in business, government, daily life, and scientific research launched many efforts to understand and utilize data. Data mining, knowledge discovery (Maimon & Roach 2005), and business intelligence and analytics (Chen et al. 2012) became popular terms in business describing statistical and logical rule-based efforts to extract knowledge from large databases. Within engineering, a 70-year tradition continues of building computers and robots with artificial intelligence (Russell & Norvig 2009) that can perform human-like tasks such as playing chess or driving cars. Some of the methods developed by artificial intelligence researchers have been combined with traditional methods of statistics to produce methods for pattern recognition (Ripley 1995), machine learning (Bishop 2011), and statistical learning (Hastie et al. 2016). During the first decade of the twenty-first century, the need for better ways to process and use data, especially in the sciences, was discussed under the rubric of cyberinfrastructure (Atkins et al. 2003, Berman & Brady 2005), but more recently big data and data science have become popular phrases.

Big Data

For those of us who remember when computer memories were measured in kilobytes instead of terabytes (a factor of a billion more), "big data" seems like a moving target, but the term has arisen despite the advances in computer power because data seem to be growing faster than our ability to process them. The total volume in bytes, the variety (text, images, audio, video, sensor, social media, and other forms), and the daily velocity (Laney 2001) of data are growing even faster than computing power. The large volume leads to problems of storing and managing data. The growth in variety adds the difficulties of translating data from one form to another, and the growth in velocity leads to the need to edit data on the run and to choose what is important. More recently a fourth concern, the veracity of data, adds another layer of complexity on top of volume, variety, and velocity.

Size, complexity, and technological challenges provide one definition of big data (National Research Council 2013, Ward & Barker 2013), but they do not seem a sufficient basis for heralding a sea-change in our data environment, since the race between data set size and computer capabilities goes back to the advent of computing. The National Institute of Standards and Technology has more usefully proposed that "fundamentally, the Big Data paradigm is a shift in data system architectures from monolithic systems with vertical scaling (i.e., adding more power, such as faster processors or disks, to existing machines) into a parallelized, 'horizontally scaled', system (i.e., adding more machines to the available collection in order to deal with volume, variety, and velocity) that uses a loosely coupled set of resources in parallel" (NIST 2015, p. 5). But the statistician David Donoho (2017, p. 747) objects that "the *new* skills attracting so much media attention are not skills for better solving the *real* problems of inference from data; they are coping skills for dealing with organizational artifacts of large-scale cluster computing." We also do not know whether this new architecture is permanent or transient.

Beyond the sheer amount of data, the truly distinguishing features of the big-data revolution are the new technologies for recording, connecting, networking, and creating information. Human interactions through phone calls, email, texts, tweets, social media posts, and other technological methods are now digitally recorded, time- and location-stamped, and attributable to nodes in networks in ways that go far beyond the much more ephemeral media of the past. Many business, governmental, social, and scientific tasks now have digital trails, such as Fed-Ex tracking services, Web searches and purchases, parking meter payments, automobile trips, tax payments, photographs of social gatherings, weather and environmental measurements, digital images from microscopes and telescopes, and much more. When these are combined with the facts that the World Wide Web is an excellent site for social networks and accessing information and that computers can now author information and interact with us—perhaps even producing artificial intelligence and autonomous robot-like entities and virtual realities—the impression is not merely of big data but of immersive data that surround us in our daily lives. The "decentralization of data" identified by NIST may also be more than just a set of techniques for dealing with large computing problems, but the future shape of computing and the internet is still not clear. Consequently, the real impact of the big-data revolution is not so much the amount of data as a change in our cognitive environment (Lugmayr et al. 2016, Neumann 2016, Schroeder 2018) that requires new perspectives to deal with datafication, connectedness, networking, and computer authoring. These phenomena stem from the invention of new technologies including innovative methods in data science.

Data Science

Big data's companion idea, data science, relies less on the scale of the data than on a definition of a way to discover new knowledge in an age when data have proliferated and cry out for analysis. In 2001, the statistician William S. Cleveland put forth a plan to "enlarge the major areas of technical work in the field of statistics" by providing more resources for "computing with data" (Cleveland 2001, pp. 21, 22) and to call the new field "data science." In an address to the Computer Science and Telecommunications Board of the National Research Council in 2007, computer scientist Jim Gray advocated for "data-driven science" as a new scientific paradigm that uses large collections of data to make scientific discoveries. Gray (2009, p. xxv) proposed that there was a "need for tools to help scientists capture their data, curate it, and then visualize it," and that the goal was to "unify all the scientific data with all the literature to create a world in which the data and the literature interoperate with each other."

Starting from these ideas, NIST (2015, p. 7) describes data science as "the extraction of actionable knowledge directly from data through a process of discovery, or hypothesis formulation and hypothesis testing." One well-known Venn diagram (Conway 2013) places data science at the intersection of three areas: computer programming skills, mathematics and statistics knowledge, and substantive expertise in a field of research. The diagram includes machine learning as an important aspect of data science because machine learning deals directly with data and discovers patterns within it. No doubt the surprising success of machine learning (especially deep learning) in making predictions is one reason for the popularity of data science, but we do not know why deep learning works so well (Knight 2017). This raises a question confronted later in this article: How much do we have to understand about the model's underlying predictions to feel comfortable with a method? The question reflects long-standing concerns with causality versus correlation, experimental versus observational data, structural equation models versus reduced forms, and explanation versus prediction.

But these characterizations of data science are not entirely new either. In a famous article in 1962, the statistician John Tukey averred that perhaps he was not a statistician because "I have

Table 1 The seven activities of data science[a]

Activities	Examples
Data gathering, preparation, and exploration	Survey data, experimental data, genomic data, textual data, administrative data, image data, web data, and sensor data Data cleaning and exploratory data analysis methods for checking on outliers and data quality
Data representation and transformation	Relational and nonrelational databases Networks and graphs Other mathematical structures for data
Computing with data	R and Python Programming packages, text manipulation languages Cluster and cloud computing Reproducible workflows
Data modeling	Determining or hypothesizing data generating probability functions, structural and predictive modeling
Data visualization and presentation	Types of visualizations and graphs Rules for labeling and presenting data Psychological impacts of various displays
Data archiving, indexing, and search and data governance	Standards for open data and reproducibility Determining rules for access and privacy protection where necessary
Science about data science	How people do data science Impacts of data science and big data on society

[a]The activities are quoted from Donoho (2017, p. 755) except for "Data archiving, indexing, and search and data governance," which is my addition. The examples are my own.

come to feel that my central interest is in data analysis, which I take to include, among other things: procedures for analyzing data, techniques for interpreting the results of such procedures, ways of planning the gathering of data to make its analysis easier, more precise or more accurate, and all the machinery and results of (mathematical) statistics which apply to analyzing data" (Tukey 1962, p. 2). Tukey's impact on statistics has been immense (Statistical Science 2003), and his concept of data analysis covers much of the same ground as data science.

Statistician David Donoho (2017, p. 748) argues that "today's popular media tropes about data science do not withstand even basic scrutiny," but building upon Tukey's work "there is a solid case for *some entity* called 'data science' to be created…." Donoho (2017, p. 755) proposes that data science should encompass six activities to which I've added one more in **Table 1**, where I have also added examples.

Judging from **Table 1**, data science borrows methods and techniques that go beyond the traditional core of statistics, which is largely encompassed in item 4, "data modeling." Techniques of data gathering and preparation are typically taught in subject matter disciplines even though statistics started as an endeavor to collect data on the state and its people through censuses and surveys. Computer science and other academic departments deal with data representation and transformation and with computation. Data visualization and presentation often involve media laboratories and psychology departments. Data archiving, indexing, and availability form the core of work in schools of library science and their modern incarnations as schools of information. In one subject matter area, bioinformatics, more than 100 colleges and universities now offer programs that focus on these tasks, and there are a few digital humanities, social sciences, and environmental science programs. But at the moment it seems that the most popular way to move forward in this area is to create "data science" programs, including computer science, information, and statistics, which

allow for relationships with subject matter disciplines. The unsolved problem is how the applied data science being done in these disciplines can be incorporated into data science programs. For example, in addition to benefiting from using data science and big data, the social sciences can provide fundamental help in understanding the social construction and meaning of data, the causal impact of new information technologies, the ethical issues of privacy and data ownership, and the best ways for social institutions to use cyberinfrastructure (Berman & Brady 2005). Data science must encompass these issues.

However universities organize themselves to deal with these seven tasks, the following seems clear to me. The explosion in the number of methods and techniques for undertaking the tasks means that universities need to bring together the people working on them to learn from one another and to teach the next generation of students and scholars how to use them. There must also be some way to help scholars, either through collaboration with other scholars or by having specialists akin to collections specialists in libraries or museums, to use the many kinds of data, software, and techniques that are now available. Gone are the days when someone could learn, as I did, about a few kinds of data collection (e.g., surveys, content analysis, and administrative data), some FORTRAN and subroutine libraries such as NAG and IMSL, a bit about dBase and SQL, some statistics through econometrics and psychometrics, some statistics packages such as BMDP-SPSS-SAS-STATA-GAUSS, a computational program such as MATLAB, and a few other things and be at the forefront of data science in their discipline. There is just too much to be learned.

Real Phenomena, Inadequate Language

Many of the developments related to big data and data science are not new, but they have achieved a scale and level of impact that require new ways of describing them. The right language is hard to find.

The nineteenth century's transportation revolution was not just about the steam engine—it also involved the discovery of new forms of energy (oil and electricity); the invention of new kinds of motors (internal combustion and electrical); the creation of networks composed of rails, roads, and rivers; and even the development of new social norms such as standard time zones. Similarly, the information revolution is more than just computers or any other single thing. It also involves sensors, databases, programming languages, artificial intelligence, telecommunications, machine learning, social media, the internet, and many other inventions. Neither "big data" nor "data science" nor any other labels encompass all these innovations. The term cyberinfrastructure might have been useful, but it has not caught on. One leading data science scholar (Jordan 2018) argues for the use of the term "intelligent infrastructure," which is broader than "artificial intelligence," but it also has its limitations. We are left with real phenomena but inadequate language.

SOCIETAL AND POLITICAL CHANGE FROM BIG DATA AND DATA SCIENCE

Many authors have provided overviews of areas affected by big data (Chen et al. 2012, Cukier & Mayer-Schoenberger 2013, Mayer-Schönberger & Cukier 2014, Mosco 2014, Evans 2018). This article cannot provide an exhaustive review of the possible societal impacts of big data and data science, but I list a few prominent examples to show that they deserve more scrutiny by political scientists. I have chosen cyberwarfare and homeland security, smart cities, medicine, the media, and robotics.

Several recent books propose that cyberwarfare exists and that it threatens international security (e.g., Clarke & Knake 2011, Kaplan 2017), but skeptics (Rid 2012, Libicki 2014) argue that

while cyber disruptions may be a problem, they do not constitute classical warfare like the Japanese attack on Pearl Harbor, which involved a purposeful and publicly claimed act of violence for political advantage. Some leading examples of cyberwarfare—such as the Stuxnet virus's introduction into Iranian centrifuges, which destroyed an essential part of Iran's nuclear fuels enrichment program, or the massive denial-of-service attack (presumably by Russian hackers) on Estonia in April 2007—were almost surely purposeful, but at most they caused lost productivity and perhaps property damage. Most importantly, no state claimed responsibility in order to achieve direct political advantage. Although the case for cyberwarfare may be weak, the Web has certainly been used for "sabotage, espionage, and subversion" (Rid 2012, p. 5), as recent events involving Russia and the 2016 American election make clear (Sanger 2018, Jamieson 2018). Moreover, the American military is collecting and processing a flood of sensor and digital information (Porche et al. 2014), which could change the face of conflict (Dunlap 2014). Obviously, these developments get at the heart of political science studies of international relations and security.

"Smart Cities" is a popular book title with subtitles such as "Big Data, Civic Hackers, and the Quest for a New Utopia," "A Spatialised Intelligence" and "The Internet of Things, People, and Systems" (Townsend 2013, Picon 2015, Dustdar et al. 2017). Three streams of big-data work come together in this area. First, there are large, digitized administrative data sets on people and their relationship to schools, social welfare agencies (Brady et al. 2001), medical care, and police, and there are similar data sets on physical structures and their relationship to streets, services, land use, and zoning. Second, the reduced costs of sensors, wireless networks, and video cameras, combined with the ability to connect them with an "internet of things," make it possible to monitor and sometimes remotely control air pollution, traffic, parking, usage of electricity and water, utilities, safety, police and firefighter deployments, and many other aspects of a modern city. Third, internet data such as Google Street View, Zillow, Airbnb, or Yelp can provide information about businesses, real estate, and the physical condition of the city (Glaeser et al. 2018). These data can be linked by geo-coding the location of each person's house (or place of work), each structure or business, and each sensor. Increasingly, we can go farther and link data through recognition of vehicles, faces, or radio frequency identification tags, which makes it possible to track movements throughout the city (Hashem et al. 2016).

Using these data, the city and its operations can be described, managed, and evaluated. Maps of traffic, air pollution, or poverty can provide useful descriptions for those trying to understand where to live, where to travel, or what to do. Conditions can be managed and improved in real time by involving citizens in constant feedback on services, changing the timing of traffic lights, deploying police to areas with disturbances, asking industries to "spare the air" by reducing some activities, and so forth. Finally, evaluation results can indicate what is working and what is not so that processes can be improved.

Because the decisions about what data are collected, how they are processed, and how they are used all involve choices, often influenced by who has power and who does not, these systems are inherently political. They can easily become technocratic, overly influenced by corporate interests, and perhaps most alarmingly, the basis for the "panoptic" city—the urban counterpart of Jeremy Bentham's circular Panopticon, a prison in which all inmates were constantly visible to a centrally located guard station (Kitchin 2014).

Precision medicine, according to a 2011 report by the National Research Council of the National Academy of Sciences, is "the tailoring of medical treatment to the individual characteristics of each patient" (National Research Council 2011, p. 125). To practice precision medicine, a physician would combine information about the individual with medical knowledge about how people vary in their response to illnesses and treatments (Dzau & Ginsburg 2016). Individual information would come from electronic medical records and genomic data. The 2011 report

suggested creating a new taxonomy of human disease based on molecular biology that would serve as the basis for classifying diseases and people's reactions to them. To do this, an "information commons" would be created that linked molecular data, medical histories, and health outcomes (Beachy et al. 2015), and these data would be used to explore clinical associations (Hanauer et al. 2009, Miller 2012). These data could be a great boon to medical researchers, but they raise significant questions about privacy, ownership of data, and their relationship to issues such as race in America (Hochschild & Sen 2015) that could become high-profile political issues.

Changes in the media from the rise of the internet are now manifestly important for politics, but political scientists have lagged in their awareness of them. In 2002, in the first examination of the mass media in the *Annual Review of Political Science*, Schudson (2002, p. 249) quite properly takes political science to task because it "has never extended to the news media the lovingly detailed attention it has lavished on legislatures, parties, presidents, and prime ministers." Yet he does not even mention the internet or World Wide Web. He focuses on the relative merits of state- versus commercial-controlled media, journalism as "the story of the interaction of reporters and government officials" (Schudson 2002, p. 255), and the cultural norms that shape coverage of topics such as homosexuality and crime. He concludes, "The news media have always been a more important forum for communication among elites (and some elites more than others) than with the general population" (Schudson 2002, p. 263), with never a hint of the anarchy of uncontrolled news sources and direct leader–follower communications now bedeviling a world with Facebook, Google, and Twitter.

Ten years later, Farrell's (2012) *Annual Review of Political Science* article recognizes the potential importance of the internet for exacerbating political polarization or facilitating the Arab Spring, and he argues that the internet could sort citizens into homogeneous groups seeking information to confirm their ideological biases, discourage preference falsification in authoritarian regimes by making available a broader array of opinions, and overcome the costs of collective action by allowing like-minded and politically intense people to find one another. Although Prior still concludes, in his 2013 *Annual Review of Political Science* article titled "Media and Political Polarization," that "[i]nternet use shows few signs of ideological segregation" (Prior 2013, p. 122), he takes the internet seriously. And communications theorists such as Bennett & Segerberg (2012), Neumann (2016), and Schroeder (2018) argue for developing new models to understand the new media on the internet. Among other things, these theories must explain how people seek out and obtain information, since this is such a big part of what people have been enabled to do on the internet.

These four examples illustrate the kinds of questions that political scientists might ask about the impacts of big data and data science. In *Seeing Like a State*, Scott (1999) chronicled how states have misused census and other information. What will it mean when societies, businesses, and governments have access to large data sets about their populations that go far beyond a census? Who will own these data? Who will define what data get collected and used? What happens when news and information (e.g., blogs, cellphone videos) can be authored and disseminated without the editing power of peer reviews, journalistic norms, and a concern for their context and veracity? What new problems are created when information can be hacked and digital systems are vulnerable to viruses? When medical diagnoses or city operations depend on algorithms that sometimes fail? What biases will be baked into the algorithms? How can people be brought into the systems at the right places to ensure their participation, their rights, and their welfare?

One final example is worth exploring, although it seems the work of science fiction. As robots get better at sensing the world, as they learn the rudiments of pattern recognition if not full cognition, as they become adept at speech recognition and talking, as they can communicate with each other and with us through wireless networks and the cloud, and as they become embodied in autonomous machines with their own lightweight power sources, to what degree do they acquire

rights and responsibilities (Pratt 2015)? If robots replace people at their jobs, what is left for people to do? And if a great deal of wealth is embodied in robots, who owns the robots and who gets the return to their effort (Albus 1984)? Already some authors are proposing universal basic incomes (Manjoo 2016) and guaranteed jobs (Tankersley 2018) to deal with the possibility of job loss due to robots. What kinds of political problems does this raise, or was a 1962 article right to conclude, "Artificial intelligence is neither a myth nor a threat to man" (Samuel 1962)?

INCREASING AMOUNTS OF DATA AVAILABLE TO ALL SCIENTISTS, INCLUDING POLITICAL SCIENTISTS

In a 2015 report, NIST surveyed 51 cases of uses of big data involving government and commercial operations, defense, health care and life sciences, social media, astronomy and physics, earth and environmental science, and energy. Every area involved producing or analyzing many terabytes of data and about one-third of them involved petabytes of data (NIST 2015, pp. 6–45, Appendix B)—sometimes petabytes per year. Scientists are now generating data at a prodigious rate in research involving every physical scale from the subatomic to the cosmic: analyzing the subatomic structure of matter in CERN's Large Hadron Collider, investigating the atomic and chemical structure of materials through intense X-ray and other light sources and through mathematical simulations that start from basic physical principles, sequencing DNA and mapping proteins rapidly and completely, using real-time three-dimensional microscopy of cells at many different wavelengths to understand their operations, scanning animal and human brains and bodies using functional magnetic resonance imaging (fMRI), monitoring the environmental conditions of cities and regions using multiple methods (fixed sensors, radar, and satellite imaging), and undertaking telescopic surveys of the solar system and the universe at multiple wavelengths and in real time. Some of these data sets could be useful to political scientists, such as fMRI data for those studying political psychology (Theodoridis & Nelson 2012) or satellite sensor data for those studying the impacts of climate change on politics (Hsiang et al. 2013).

Social scientists have benefited from many new data sources as well. As of roughly 1980, political scientists had available a limited number of data sets, mostly about the United States but also about other countries: Historical election statistics, usually by county but in a few cases by precinct; surveys from the 1930s onwards; census data; Federal Election Commission (FEC) data on political contributions; roll-call data from legislatures and the United Nations; data from the Correlates of War Project, the *World Handbook of Political and Social Indicators*, and a few other sources. In the past 30 years, the volume and variety of data have increased enormously beyond these areas, especially thanks to administrative data, internet data, textual data, and sensor-audio-video data.

Administrative Data

Before surveys, political scientists interested in voting used turnout and voting data aggregated by precincts, counties, and states. Recently there has been a return to this kind of data, but often disaggregated in the form of voter registration lists from administrative data. These lists do not report election choices, but they are the official record of turnout and in some states they include political party registration. Brady & McNulty (2011) geo-code the addresses and precinct locations of millions of registered voters in Los Angeles to take advantage of a natural experiment in 2003 where the number of precincts was reduced by two-thirds for the state-wide recall election. They show that changes in polling place location alone had a significant impact on turnout (a few percentage points) and that increased distance to polling place further decreased voting. Using

voting records over time (from 1998 to 2012) and data on the residential addresses of 9/11 victims, Hersh (2013) shows that the families and neighbors of these victims voted at significantly higher rates (a few percentage points) after the event than carefully constructed control groups, and they changed their party identification toward the Republican party. Using voter registration files for the city of Chicago, Enos (2016) examines the impact of perceived racial threat on voter turnout by using a natural experiment in which public housing buildings with over 25,000 African American residents were demolished. He categorizes each voter's race using a Bayesian classifier based on the voter's name, location, and related census data. He finds that white voters' turnout decreased by 10 percentage points after the exit of their African American neighbors presumably reduced their perceived sense of threat. Ansolabehere & Hersh (2012) use 50-state voter registration records from a commercial firm, Catalist, LLC, to match individuals interviewed in the 2008 Cooperative Congressional Election Survey to their voting records to determine the correlates of vote misreporting. They describe methods for ensuring the quality of matches and the quality of registration lists, and they find that the correlation between basic socioeconomic characteristics and voting is lower for validated voters than for self-reported voters.

The role of ideology and money in politics has been a long-standing concern of political scientists. Bonica (2013) starts with the classic FEC political-contributions data for the 1980–2010 congressional election cycles and develops a generalized item-response theory count model to estimate an ideal point model of the ideology of candidates and Political Action Committees that contribute money. In order to obtain usable results, he restricts the sample "to candidates who received money from 30 or more unique contributors and contributors that give to 30 or more unique candidates" (Bonica 2013, p. 298). The technique provides estimates for first-time candidates who have no roll-call records from which to estimate their political positions, and the author shows that using his ideological estimates for candidates provides only "a negligible reduction in predictive power of legislative voting behavior" (Bonica 2013, p. 308) compared to roll-call votes. In other papers he connects these data with contributions by doctors (Bonica et al. 2014) and lawyers (Bonica et al. 2016) by linking the contributions data set to listings of these professionals. He describes a massive database that uses candidate names as a key to combine campaign contribution data, legislative voting and bill sponsorship data, election data, and text "from bills and amendments, floor debates, candidate websites, and social media" (Bonica 2016, p. 14). This information is combined to get candidate ideology scores, and it can be used to study the impact of money in politics. In addition, Bonica (2016, p. 18) develops a three-stage process "for measuring preferences and expressed priorities across issue dimensions that combines topic modeling, ideal point estimation, and machine learning methods." The topic model organizes the text into issue categories by using automated statistical methods described in more detail below.

Using lobbying reports available under the Lobbying Disclosure Act of 1995, Kim (2017) identifies firms that lobby on trade policy, and he links this information, using the names of firms, with databases such as Compustat and Orbis on the characteristics of firms. He adds to this all bills in Congress that had been lobbied, and information about tariffs and trade (Kim 2017, p. 10). By focusing on firms instead of industries, Kim shows that lobbying is firm-specific. In a related paper, lobbying data are combined with sponsorship data on congressional bills to show that, unlike electoral politics networks structured according to ideology, there are distinct "political communities in the lobbying network, which is organized according to industry interests and jurisdictional committee memberships" (Kim & Kunisky 2018).

Recent controversies over police behavior have led to major efforts to collect data on police stops (Pierson et al. 2017) and police use of force (Goff et al. 2016). Each study involves substantial linking across jurisdictions with idiosyncratic formats and definitions of variables. Both conclude

that there are substantial racial disparities even after controlling for many relevant features of police encounters.

These examples illustrate several important features of studies using administrative data. Large-scale administrative data sets on voting, lobbying, campaign contributions, trade, tax, welfare, police reports, 311 calls, and many other areas often provide the (legally) definitive data on these activities, but the data sets can contain errors (Luks & Brady 2003). Moreover, in order to get a data set that represents different areas and that has enough cases for analysis, studies often require *extensive* linking of more people, organizations, or events across jurisdictions. Extensive linking often requires dealing with the problems of combining data with different formats and variables.

These administrative data studies also benefit from *intensive* linking, in which more data about individual people, organizations, or events are added, as in the work by Bonica and Kim. Brady et al. (2001, p. 226) show how state governments have greatly increased the value of their social program databases by linking across eight programmatic areas including Medicaid, foster care, food stamps, welfare, and other areas. Even with this linking, however, these data often lack useful ancillary information—unlike surveys, they do not automatically collect lists of socioeconomic characteristics such as education, income, age, and so forth on people or financial and historical information on firms or organizations. Moreover, even when this information is collected, it may be of low quality unless it is an essential part of the business purpose of the program (e.g., for welfare programs, income data are reliable because they are part of the application process, but education data are not). Intensive linking to other data sets can often expand their utility tremendously, but these matches are often precarious given the complexity of names, places, and other identifying information. Linkages using probabilistic matching techniques or geo-coding can help facilitate this process, but they still involve elements of uncertainty and incompleteness.

Administrative databases are also often better at providing samples of people who do or encounter things than at portraying the complete universe of those who might have done things. For example, data on police traffic stops tell us who was stopped but not who should have been stopped. Campaign contribution data tell us who gave money, but we know only the value of the numerator in the ratio of those who gave to those who could have given. One approach is to link these data to population data, such as census data or motor vehicle license data, but these linkages can present legal and practical problems (Brady et al. 2001), and they also may not give the best denominator data; for example, in the police-stops example, we want the number of people in each group who should have been stopped given their behavior, not the number of people in each group who drive.

Internet Data

Using proprietary data on over six million Facebook users who had two or more "likes" for 1,223 official political pages representing political candidates, Bond & Messing (2015) estimate candidate and individual ideologies. Because the average number of likes is slightly over three, the matrix of candidates by people is very sparse except for some rows (e.g., Barack Obama's and Mitt Romney's), necessitating steps to adjust for different base frequencies for liking candidates. For those candidates for whom there is an independent measure of ideology from Congress's roll-call data, the correlation between the two measures of ideology was 0.47 for Democrats and 0.42 for Republicans (Bond & Messing 2015, p. 68). Similarly, with a data set of Twitter users from six countries, Barberá (2015) identifies Twitter followers of three or more political actors and uses ideal-point estimation methods to recover the ideologies of the politicians and the Twitter users.

Employing various sources of baseline data for each group, he finds evidence that validates these measures. He also finds evidence for political polarization among these Twitter users.

The Web makes it possible to follow events through time. Tinati et al. (2014) develop a tool for following Twitter information flows and network formation over time, and they apply it to a protest of university tuition fees in England in November 2011. They show how networks grow through retweets and that a small number of people are key players. Gomez-Rodriguez et al. (2012) show how information diffuses in 170 million blogs and news articles over a one-year period by developing an algorithm to infer networks of influence and diffusion. They show that the algorithm recovers the structure of simulated data, and it appears to work well with real data. News topics and memes can also be tracked on the Web to characterize a news cycle. By tracking 1.6 million media sites with 90 million articles over three months in 2008 (August–October), Leskovec et al. (2009) find that phrases come and go over 24 hours and that blogs pick up phrases with an average lag of 2.5 hours. Two mechanisms explain much of the up-and-down dynamics: imitation, in which memes persist because sources imitate other sources, and recency, in which older memes are extinguished because new phrases are preferred.

Using Facebook data, Bond et al. (2012) study whether social networks can affect behavior. They randomly assigned encouragements to vote and information about the person's polling place to millions of people on the day of the 2010 midterm election. The "social message group" of 60 million people were also shown up to six faces of their friends who had reported on Facebook that they had voted that day. The "informational message group" of over 600,000 people received only the encouragement to vote and information about their polling place. The "control group" did not receive any message. Those in the social message group were two percentage points more likely to say that they had voted than those in the informational message group, and other significant effects were found.

King et al. (2013) study the motivation of Chinese internet censorship by following the fate of blog posts over time. By comparing the content of those that were censored versus those that were not, they conclude that "the censorship program is aimed at curtailing collective action by silencing comments that represent, reinforce, or spur social mobilization, regardless of content" and that "posts with negative, even vitriolic, criticism of the state, its leaders, and its policies are not more likely to be censored" (King et al. 2013, p. 326). The study is notable for its real-time effort to locate blogs before they were censored (which typically occurred within one day) and its use of automated content analysis methods to analyze the blogs.

To estimate how racial animus affected the vote for Barack Obama in 2008, Stephens-Davidowitz (2014) calculates for media markets the fraction of Google searches that use a well-known derogatory term for African Americans. He finds that racial animus cost Obama roughly 4% of the national popular vote. His paper provides numerous checks on the validity and reliability of his measures.

In addition to sharing many of the same problems as administrative data, internet data are typically highly selective in terms of socioeconomic characteristics (especially by having more young people, although older people are catching up), and they often depend on people's involvement with platforms such as Facebook, Twitter, or Google. Moreover, this involvement is enmeshed with constant efforts by the companies running these platforms to encourage participation, which can lead to subtle selection effects that may mislead the researcher (Lazer et al. 2014). Absence of data is also a problem, as in the studies estimating ideology using Facebook and Twitter data. The compensating merits are that internet data often provide fascinating network data that would otherwise be unavailable; events can be studied as they unfold in real time; and hidden information on behaviors (such as searches about culturally disapproved themes) can be revealed. Nagler & Tucker (2015) discuss what can be learned from Twitter.

Textual Data

Those of us who have put together teams of students to do content analysis of texts know how time consuming and error prone the process can be. Automated methods promise greater efficiency, increased replicability, and perhaps less error-prone coding. Textual data provides an element often missing in our analysis of politics: the words of citizens and politicians. For example, political scientists study the personal vote, in which citizens support politicians in exchange for government money spent in their districts. But how do citizens know about these expenditures? Grimmer et al. (2012) identify the missing ingredient, which is legislators' statements to their constituents. By analyzing all 170,000 US House of Representatives press releases issued between 2005 and 2010 and coding them into five categories that measure two kinds of credit-claiming and three kinds of non-credit-claiming behavior, they find that constituents are more responsive to the total number of messages they receive than the amount of in-district expenditure claimed. To analyze this large corpus of material, they used a supervised learning algorithm (Hopkins & King 2010) that requires a set of hand-coded documents that can be used to "train" the method.

Wilkerson & Casas (2017) and Grimmer & Stewart (2013) provide excellent overviews of the profusion of content analysis methods developed in the last 15 years. Two other articles explore how these methods can be used to study culture (Bail 2014) and to improve the practice of qualitative research (Wiedemann 2013). The methods include the search for particular words or phrases (e.g., Stephens-Davidowitz 2014, Leskovec et al. 2009); the determination of what fractions of text fit into predetermined categories (e.g., King et al. 2013, Grimmer et al. 2012); the classification of each text into predetermined categories using supervised learning; the classification of text into unknown categories using unsupervised clustering methods; and the ideological scaling of political texts such as party platforms (Laver et al. 2003).

These methods require careful use. Grimmer & Stewart (2013) advise, "all quantitative models of language are wrong—but some are useful" (p. 269), and "quantitative methods augment humans, not replace them" (p. 270), so "validate, validate, validate" (p. 271). In addition, the more the methods are automated or unsupervised, the more they typically use complex statistical methods: mixture models with many local minima, in which one cannot guarantee a globally correct solution; lasso or ridge regression, which strive for simplicity that might underfit the data; and models with many parameters that often try to estimate values for each document with small amounts of data. To perform these tasks, they often use estimation methods such as the expectation maximization (EM) algorithm or Bayesian Markov chain Monte Carlo (MCMC) that take a long time to converge and can be tricky to use (see Roberts et al. 2014). Despite all these complexities, the methods can accomplish tasks that could not be done with typical budgets and research teams. Text reduction and analysis have progressed to a point where quantifying large bodies of text is possible. Arguably, these methods improve on human coding if suitable precautions are taken to check the results with human coders and to recognize the limitations of the analysis.

Sensor, Audio, Video, and Other Data

Hsiang et al. (2011) connect sensor data (from gauges and satellite observations) on temperature and rainfall with information on conflict from the "Onset and Duration of Intrastate Conflict" data set to study the impact of weather on civil conflicts. They use the El Niño/Southern Oscillation (ENSO) in weather to identify their model, and they find that the probability of new civil conflicts doubles during El Niño years. The supplementary materials describe the complexities of linking geo-coded sensor data to the boundaries of individual countries over time.

Jennifer Eberhardt and her colleagues use body camera data from stops by police officers in Oakland, California, to uncover racial disparities in officer respect. Starting from human transcriptions and coding of the audio portion of these data, they develop machine learning methods for studying the degree of respect exhibited in the text of police utterances toward people they have stopped. They note: "Future research could expand body camera analysis beyond text to include information from the audio such as speech intonation and emotional prosody, and video, such as the citizen's facial expressions and body movement, offering even more insight into how interactions progress and can sometimes go awry" (Voigt et al. 2017, p. 6525).

These examples demonstrate the power of linking sensor, audio, video, and other kinds of data to events, but they also reveal the substantial processing that must be done to use them correctly. Moreover, they suggest that we still need to improve our ability to transform these data into usable forms for our research given, for example, the complexities of facial expressions or body language in a video and the modifiable areal unit problem in geography, which stems from the difficulty of matching geo-coded point-based measures from sensors to different geographic entities such as cities, counties, states, or nations.

NEW WAYS POLITICAL SCIENTISTS ORGANIZE THEIR WORK

New Courses

Political science professors must develop new courses and become conversant with the new technologies developed by data scientists. New courses should go in two directions. One course should deal with the societal challenges of big data and what they mean for politics. Mergel (2016) has developed a curriculum for schools of public affairs which contains some pertinent elements, including sections on big data in politics, government, public health, and smart cities, but it does not have a section on the media, and it does not directly focus on the political issues such as data ownership and use, privacy, and loss of jobs that stem from big data.

A second course must teach students data science methods. A check of methods courses taught in political science departments at major universities suggests that this is well under way. These courses include programming in R or Python, an emphasis on resampling approaches to understanding statistics, an overview of the data sources described above, and careful discussions of methods for making predictions and those for inferring causality. Moreover, at least one edited book (Alvarez 2016) summarizes a good selection of relevant topics.

Neither of these courses deals with deeper theoretical issues such as how our epistemological and ontological presuppositions might be affected by new methods, the new forms of connectedness in society, and the rise of artificial intelligence. One should be properly skeptical of such grand possibilities, but Rogers (2013), Mayer-Schönberger & Cukier (2014), Mosco (2014), Boullier (2015), and Salganik (2017) provide some food for thought about what will happen when we make "the world self-aware and self-describing" (Evans 2018, p. 141).

New Research Management Methods

A few political scientists working with Google, Facebook, or very large data sets might have to learn about big-data architecture and the new decentralized methods of processing large sets of data such as Hadoop, Hive, NoSQL, and Spark (Varian 2014, Oussous et al. 2018), but for most it would be a waste of time. Instead, political scientists might better focus on new software for data cleaning, data management, reproducible science, life-cycle management of data, and data visualization. Here I briefly discuss data cleaning and reproducible science.

A tweet (@BigDataBorat) parodies the common belief that data cleaning takes up most of the time in research by saying "In Data Science, 80% of time spent prepare data, 20% of time spent complain about need for prepare data." Certainly data preparation is tedious and time-consuming (Kandel et al. 2012). DataWrangler (Kandel et al. 2011) displays data in an interactive interface like a spreadsheet and allows the researcher to make changes to one line of the data that are reproduced in all other lines of data based on the program's inferences about the general transformations that are desired. As the user interacts with the system, it improves its inferences and even makes suggestions so that it helps the researcher make improvements. The system keeps track of what has been done to the data so that the researcher can make sure it has been successful. A free version of it is available as Trifacta Wrangler. Another approach to cleaning data is the Tidyverse, which is a free collection of R programs that can be used to create a tidy dataset (Wickham 2014).

Reproducible science aims to make it possible for a second investigator to "recreate the final reported results of the project, including key quantitative findings, tables, and figures, given only a set of files and written instructions" (Kitzes et al. 2017, p. 13). Kitzes et al. (2017) exemplify reproducibility through 31 case studies in different scientific areas, including social science, with a focus on data acquisition, data processing, and data analysis. Most of the studies use tools from either Python (17 studies) or R (13) to create a reproducible workflow. Because these tools make it easier to obtain and to recreate research results, because journals are increasingly requiring reproducibility, and because the federal government has been moving toward requiring it for grantees, learning these methods is very worthwhile.

NEW KINDS OF QUESTIONS ASKED BY POLITICAL SCIENTISTS

Where Does Data Science Come From?

Data science methods primarily come from computer science, statistics, and library or information sciences with some roots in the efforts of biologists to model the connections among neurons in the human brain and the work of cognitive scientists (such as polymathic political scientist Herbert Simon) to develop artificial intelligence. The blending of these streams produces confusion because similar methods (e.g., neural nets and logistic regression) have been called by different names in these disciplinary areas, and the use of names such as artificial intelligence or neural nets can lead to the mistaken belief that these methods actually mimic the way the human brain works. In fact, most of the methods can be straightforwardly translated into the language of statistics (Sarle 1994, Warner & Misra 1996), and the connection with human intelligence is more metaphorical than exact. Some of this confusion also comes from the fact that until recently computer scientists were trying to solve pattern recognition problems and to advance predictive machine learning with the fewest errors without much knowledge of or concern with statistical models, while statisticians (especially econometricians and political methodologists) focused on unbiased or consistent estimators of models and hypothesis testing for causal impacts with little concern for prediction or learning. Information scientists were also trying to produce quick and efficient ways to index and access documents and knowledge with an emphasis on prediction and little concern for statistical methods or models.

Because of their emphasis on pattern recognition, computer scientists typically speak of assigning cases to classes based on their features (e.g., predicting whether someone could be classed as a diabetic based on body mass, age, serum insulin), whereas statisticians talk about predicting the value of a dependent variable based on independent variables or predictors, even though they are often dealing with the same problems. Computer scientists talk about activation functions, training sets, and learning, whereas statisticians talk about functional forms, samples, and estimation. In addition, computer scientists talk about supervised and unsupervised learning problems;

the former refers to problems where there is information on the relevant classes (e.g., specimens already classified into separate species) and the latter refers to problems without this information. Supervised learning uses methods with a dependent variable such as discriminant analysis or logistic regression, whereas unsupervised learning uses clustering, factor analysis, or multidimensional scaling. Once the newcomer to the field of data science recognizes these differences in nomenclature, books on pattern recognition (Ripley 1995), artificial intelligence (Russell & Norvig 2009), machine learning (Bishop 2011), and statistical learning (Hastie et al. 2016) seem less arcane and more approachable. Newcomers can also benefit from articles that bridge the gaps (Nickerson & Rogers 2014, Varian 2014, Mullainathan & Spiess 2017, Yarkoni & Westfall 2017, Athey 2018).

Increased computing power has also accelerated the development of five innovations. First, the Bayesian paradigm is no longer an outcast in American statistics since the realization that many intractable classical models can be considered Bayesian models with vague priors and that these models can be estimated effectively and efficiently using MCMC and other methods. Second, smoothing or regularizing approaches that require the estimation of nonlinear ridge or lasso regressions or the repeated application of complicated kernel estimation methods have become feasible, providing greater flexibility in model specification. Third, resampling and averaging methods that improve predictions, such as the bootstrap, bagging, boosting, Bayesian model averaging, and random forests, have become commonplace because of computing power that allows repeated estimation using slightly different models or samples. Fourth, the Akaike, Bayesian, and Schwartz information criteria (AIC, BIC, SIC) and methods such as cross-validation are now commonly used to select a parsimonious model. Fifth, computational methods have been developed (e.g., EM and genetic algorithms, MCMC methods, back-propagation) to estimate models with complicated density mixtures, large numbers of parameters, multiple local maxima, and knotty nonlinearities and constraints. These innovations have greatly increased the flexibility and predictive power of statistical models.

One reason data science has become so popular is that one variant of machine learning, called deep learning, has succeeded at difficult pattern recognition tasks such as speech and image recognition, natural language processing, and bioinformatics (LeCun et al. 2015). Deep learning is a variant of the canonical feed-forward neural network, which involves multilayer classifiers that use stacks of logistic or similar regressions (Sarle 1994, Schmidhuber 2015) where the inputs are features of the items that are to be classified. For example, for animals being classified as either dogs or cats, the features might be large or not-large, bark or no-bark, meow or no-meow, docile or not docile, white or not-white, and tail or no-tail. These features are coded with a one if present and a minus one if not present. Some of these features are more useful for distinguishing between dogs and cats than others. For each animal for which we have data, M weighted linear combinations of these L features are calculated where the weights reflect the diagnostic value of the features. After each of these combinations is transformed by a sigmoid activation function such as a logistic, it constitutes a hidden-layer variable, also called a neuron. The first hidden layer contains M of these hidden-layer variables employing different weighted linear combinations of the input variables. The results of these hidden-layer variables in this first hidden level are then either combined into another weighted linear combination and transformed according to the sigmoid function to decide whether the animal is a dog or a cat (with, for example, values near one indicating a dog and values near zero indicating a cat), or a second hidden level of N variables is created that takes weighted linear combinations of the M hidden-layer variables in the first hidden layer. This process can continue with more and more hidden layers until the final sigmoid function is reached that predicts whether the animal is a dog or cat. The model is evaluated on whether it gets the right answer most of the time.

The model is successful when it has the right weights so that it correctly separates the dogs from the cats. For example, a large, docile creature that barks is almost certainly not a cat, so the weights on those characteristics should be large and positive to produce a value near one (indicating a dog) in the sigmoid function, but the weights on having a tail or being white should be near zero since they are not very diagnostic features. The weight on having a meow should be negative. To make the models work, there must be enough hidden layers and hidden variables to provide the flexibility needed to fit all possible permutations of dog and cat features, and there must be efficient learning algorithms to identify the right weights so that the difficult cases are correctly classified. Shallow machine learning models have just a few hidden layers, and those with no hidden layers are called perceptrons. Deep machine learning models have many hidden layers. The overall complexity of the model depends on the number of hidden layers and the number of hidden variables or neurons.

We have known for over 25 years that systems with at least one hidden layer are universal approximators (White 1992) that can, with relatively arbitrary activation functions, approximate to any degree nonlinear continuous functions as long as there are enough neurons (hidden independent variables) in the model. Once it is clear that machine learning is simply a novel method for fitting (complicated) curves, it becomes less magical, but some mysteries remain. Why does deep learning work with a total number of weights and variables that seems far short of what would be necessary to approximate all of the possible curves? Why do models with many hidden layers sometimes do so much better than those with just one, especially since only one layer is needed for a universal approximator? How can we interpret the complex pattern of weights yielded by deep learning models? These questions have led to speculations that deep learning works because its layers can match the kinds of physical constraints that exist in the real world (Lin et al. 2017), and this speculation evokes a famous paper by the physicist Eugene Wigner (1960) titled "The Unreasonable Effectiveness of Mathematics in the Natural Sciences." Whatever the reason, deep learning methods seem to work remarkably well for pattern recognition problems, but their interpretation is often difficult given their arcane complexity. They are better at yielding predictions than explanatory insights.

What Kinds of Problems Can Data Science Solve?

There is so much hyperbole about big data and data science that one might think that we have either solved or obviated four of the most basic problems of empirical research: (*a*) forming concepts and providing measures of them; (*b*) providing reliable descriptive inferences; (*c*) making causal inferences from past experience; and (*d*) making predictions about the future. Data science has, in fact, made some contributions to solving each of them, especially forming concepts and making predictions about the future, but they continue to be fundamental and difficult problems (Smith 2018). Let us consider each in turn.

Artificial intelligence researchers have used unsupervised machine learning methods so that computers learn concepts in much the same way as political scientists have historically used factor or cluster analysis to identify concepts, as in the study of texts described above. One of the most informative studies of concept formation (Thagard 1992) used artificial intelligence models to understand "conceptual revolutions" in science. Machine learning excels at finding patterns, so it can be helpful in concept formation, but the basic problems of the interplay between defining concepts inductively or deductively, phenomenologically or ontologically, and pragmatically or theoretically remain. We do have some better tools to deal with them, such as model-based clustering techniques (e.g., Ahlquist & Breunig 2012) that allow for the evaluation of uncertainty in typologies, but concepts such as an atom, species, democracy, or topic are still very deep ideas based

on a complicated interplay between theory and data that goes beyond mere pattern detection—and that is why conceptual revolutions in science (e.g., quantum theory, plate tectonics, evolution, relativity theory, or topic analysis) are such a big deal. They reflect a gestalt change in the way we see the world. It is also why users of these methods must proceed carefully, as pointed out in the discussion about analyzing texts and topics.

Data science methods can help us to explore and describe data, to find interesting patterns in them, and to display them effectively. The use of big data helps us with descriptive inferences because it often provides a complete list of arrests, registered voters, food stamps recipients, etc., but the problem of defining the proper universe remains, since we may care about crimes, potential voters, or those eligible for food stamps, respectively. Moreover, internet samples are especially problematic because it is hard to define what universe they represent and how they were sampled from that universe. Having a lot of data does not ensure that they represent in a statistically reliable way (e.g., a random sample) an interesting and definable universe.

Perhaps most interesting, and perhaps worrisome, is the degree to which some advocates of data science have ignored or even rejected the need for causal inferences and fastened upon a narrow notion of statistical prediction. There are three sources of this inclination. The first is the idea that the availability of lots of data (either many cases or many variables) automatically solves the inference problem, which is, of course, false. Inference requires that we choose cases in the right way (e.g., a random sample) and that available variables include the actual cause and allow us to control for the right things to avoid spurious correlations (see Lazer et al. 2014, Titiunik 2015). The second source is the idea that machine learning, perhaps especially deep learning, yields insights that would otherwise be buried. That idea founders on questions about whether deep learning is actually providing insights or just fitting curves. Cukier & Mayer-Schoenberger (2013, pp. 32, 39) seem to capture both of these naïve ideas when they say that "[a] worldview built on the importance of causation is being challenged by a preponderance of correlations" and "[w]e can learn from a large body of information things we could not comprehend when we used only smaller amounts." The third and more defensible notion is that making reliable causal inferences is so hard that we should focus on prediction. This idea led to vector autoregression methods in macroeconomics (Sims 1980, Christiano 2012) 40 years ago, and it is at the core of many textbooks on machine learning. Breiman (2001) presents an elegant, early argument for this approach; Berk (2008) provides a thoughtful book-length treatment; and Shmueli (2010) discusses the trade-offs.

There are certainly practical and technical problems for which achieving a good prediction using machine or statistical learning is a satisfactory, and perhaps optimal, solution. Kleinberg et al. (2015) give an example involving decisions about hip or knee surgery where the surgeries only make sense if the patients live long enough to get through their typically lengthy rehabilitation periods. Yarkoni & Westfall (2017) provide examples from psychology, such as inferring the "big five" personality traits from the "likes" on Facebook pages and inferring the accuracy of people's memories about faces from fMRI data. Nickerson & Rogers (2014) show how predictive scores regarding campaign contributions or voting turnout can be used to increase the efficiency of campaigns. In research problems, good predictive methods can assure acceptable covariate balance in matching methods, high-quality classification of documents according to some characteristic, accurate imputation for missing values, good fits for curves in regression discontinuity designs, powerful instruments for instrumental variables estimation, and so forth.

These methods rely on situations where, in the language of econometrics, reduced form equations solve a problem either because there are no (or only small) structural changes in the mechanism producing outcomes or because the best fit is really the ultimate goal. But social scientists have known at least since the classic work on supply and demand that getting at causal mechanisms requires that statistical methods take into account the identification of structural or

behavioral models. The positive correlations between police presence and crime, between higher quantities of a good and higher prices, and between greater education and higher income do not necessarily mean that more police cause more crime, greater quantities of a good create higher prices, or even that more education produces more income. The current emphasis on experiments and quasi-experiments attempts to ensure better identification of these causal effects, and Athey (2018, pp. 21, 22), in a paper that predicts many ways in which machine learning can help improve causal estimation in economics, unequivocally predicts "no fundamental changes to theory of identification of causal effects" and "no obvious benefit from ML in terms [of] thinking about identification issues." That is the conclusion of a political science symposium on big data (Clark & Golder 2015), and I concur based on my understanding of causality (Brady 2009).

At the same time, political scientists need to think harder about how to combine information about causal mechanisms from strongly identified research designs (such as experiments or quasi-experiments) with sophisticated prediction methods and formal modeling to improve our ability to make projections about the future. These projections should take into account behavioral responses, heterogeneity in causal impacts, and general equilibrium effects that occur when policies are scaled up from a small experiment. This requires combining models, causal estimates, and predictions in ways envisioned by the Empirical Implications of Theoretical Models movement (Granato & Scioli 2004) and in ways undertaken by economists who joined vector autoregressions with concerns about causal mechanisms and macroeconomic models (Christiano 2012). Athey (2018) discusses some ways to do this, and perhaps her most important claim is that data science methods make it possible to develop better systematic model selection methods based on the data instead of specification searches that often involve multiple estimations and repetitive parsing of models until one model is presented, somewhat disingenuously, as "the model." Data scientists and statisticians are also considering trading off model complexity versus parsimony as both the sample size and the number of available variables increase (Powell 2017). Data science methods now make possible data-driven model selection using cross-validation and other approaches, estimation and averaging over many models, and accounting for model uncertainty as well as data uncertainty.

Data science currently provides many useful tools for political scientists, but their primary contribution is to provide for automated pattern recognition and better methods for prediction. Much more work has to be done before we can confidently use models to project into the future.

DEALING WITH ETHICAL ISSUES REGARDING POLITICAL SCIENCE RESEARCH

A separate article could be written about the ethical issues related to big data and data science. One contentious issue is the possibility of algorithmic injustice (Noble 2018), especially in the field of criminal justice. A number of writers (Harcourt 2007, Mbadiwe 2018, Williams et al. 2018) have worried that algorithms used to assign bail, decide on sentences, or place prisoners in various levels of detention rely on predictions that are not causal, that reproduce stereotypes, and that exacerbate racial biases. The result will be the reinforcement of existing forms of discrimination. But the problem is not easy, and "there is tension between improving public safety and satisfying the prevailing notions of algorithmic fairness" (Corbett-Davies et al. 2017, p. 797). To take another area, political campaign algorithms try to mobilize those voters who can be brought to the polls at least cost per vote, but this typically means that underrepresented voters become even more underrepresented because it costs more to mobilize them (Brady et al. 1999).

Athey (2018) notes that predictive algorithms can not only be unfair but may also be manipulable. For example, if someone knows that credit scores are improved when people shop at certain

stores, they may shop at those stores to increase their scores. The political and normative implications of these ethical issues must be studied by political scientists and taken into account when designing algorithms.

CONCLUSIONS

Big data and data science provide extraordinary new sources of data and methods for doing research. They are also changing the world in ways that spawn new kinds of political issues. They broaden the kind of quantitative work that can be done, and they bring political scientists into the middle of societal events in new ways through work on political campaigns, on the impacts of the media, on the operation of cities, on terrorism and cyberwarfare, on the design of voting and political systems, and many other areas. As this happens, political scientists will certainly do more and better research, but they will also have to think about the intellectual and practical value of their role as system designers when they find themselves or their work used to create new policies or social mechanisms. Just as engineers, lawyers, and increasingly economists use their knowledge about society to design social institutions, political scientists are now developing the tools to redesign political systems. How will this role be valued in the academy? What ethical and intellectual issues does it raise? From my perspective, becoming involved in developing new policies and social mechanisms would be a useful turn back toward the "policy sciences" advocated by Harold Lasswell (1951; see also Turnbull 2008), but political scientists will undoubtedly find themselves taking on new roles that will require debate and discussion within the profession.

DISCLOSURE STATEMENT

The author is not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

My thanks to Karen Chapple, Avi Feller, and Anno Saxenian for very helpful comments.

LITERATURE CITED

Ahlquist JA, Breunig C. 2012. Model-based clustering and typologies in the social sciences. *Political Anal.* 20(1):92–112

Albus JS. 1984. Robots and the economy. *Futurist* 18(6):38–44

Alvarez RM, ed. 2016. *Computational Social Science: Discovery and Prediction (Analytical Methods for Social Research)*. Cambridge, UK: Cambridge Univ. Press

Ansolabehere S, Hersh E. 2012. Validation: what big data reveal about survey misreporting and the real electorate. *Political Anal.* 20(4):437–59

Athey S. 2018. The impact of machine learning on economics. Draft chapter, Natl. Bur. Econ. Res., Cambridge, MA. **http://www.nber.org/chapters/c14009.pdf**

Atkins DE, Droegemeier KK, Feldman SI, Garcia-Molina H, Klein M, et al. 2003. *Revolutionizing science and engineering through cyberinfrastructure: report of the National Science Foundation blue-ribbon advisory panel on cyberinfrastructure*. Rep. Natl. Sci. Found., Washington, DC. **https://stewardshipgap.net/node/17**

Bail CA. 2014. The cultural environment: measuring culture with big data. *Theory Soc.* 43(3/4):465–82

Barberá P. 2015. Birds of the same feather tweet together: Bayesian ideal point estimation using Twitter data. *Political Anal.* 23:76–91

Beachy SH, Olson S, Berger AC. 2015. *Genomics-Enabled Learning Health Care Systems: Gathering and Using Genomic Information to Improve Patient Care and Research: Workshop Summary*. Washington, DC: Natl. Acad. Press

Bennett WL, Segerberg A. 2012. The logic of connective action. *Inf. Commun. Soc.* 15(5):739–68

Berk RA. 2008. *Statistical Learning from a Regression Perspective*. New York: Springer

Berman F, Brady H. 2005. *Workshop on cyberinfrastructure for the social and behavioral sciences: final report*. Rep., Natl. Sci. Found., Alexandria, VA. **https://www.sdsc.edu/assets/docs/SBE-CISE-FINAL.pdf**. Accessed Dec. 2, 2018

Bishop CM. 2011. *Pattern Recognition and Machine Learning*. New York: Springer

Bohn R, Short J. 2012. Measuring consumer information. *Int. J. Commun.* 6:980–1000

Bond RM, Fariss CJ, Jones JJ, Kramer AD, Marlow C, et al. 2012. A 61-milllion-person experiment in social influence and political mobilization. *Nature* 489(7415):295–98

Bond R, Messing S. 2015. Quantifying social media's political space: estimating ideology from publicly revealed preferences on Facebook. *Am. Political Sci. Rev.* 109(1):62–78

Bonica A. 2013. Ideology and interests in the political marketplace. *Am. J. Political Sci.* 57(2):294–311

Bonica A. 2016. A data-driven voter guide for U.S. elections: adapting quantitative measures of the preferences and priorities of political elites to help votes learn about candidates. *RSF Russell Sage Found. J. Soc. Sci.* 2(7):11–32

Bonica A, Chilton A, Sen M. 2016. The political ideologies of American lawyers. *J. Legal Analysis* 8(2):277–335

Bonica A, Rosenthal H, Rothman DJ. 2014. The political polarization of physicians in the United States: an analysis of campaign contributions to federal elections, 1991 through 2012. *JAMA Intern. Med.* 174(8):1308–17

Boullier D. 2015. The social sciences and traces of big data: society, opinion, or vibrations? *Rev. Française Sci. Politique* 65(5–6):71–93

boyd D, Crawford K. 2012. Critical questions for big data: provocations for a cultural, technological, and scholarly phenomenon. *Inf. Commun. Soc.* 15(5):662–79

Brady HE. 2009. Causation and explanation in political science. In *The Oxford Handbook of Political Science*, ed. R Goodin, pp. 217–70. Oxford, UK: Oxford Univ. Press

Brady HE, Grand SA, Powell MA, Schink W. 2001. Access and confidentiality issues with administrative data. In *Studies of Welfare Populations: Data Collection and Research Issues*, ed. Natl. Res. Counc., pp. 220–74. Washington, DC: Natl. Acad. Press

Brady HE, McNulty JE. 2011. Turning out to vote: the costs of finding and getting to the polling place. *Am. Political Sci. Rev.* 105(1):115–34

Brady HE, Schlozman KL, Verba S. 1999. Prospecting for participants: rational expectations and the recruitment of political activists. *Am. Political Sci. Rev.* 93(1):153–68

Breiman L. 2001. Statistical modeling: the two cultures. *Stat. Sci.* 16(3):199–231

Chen H, Chiang RHL, Storey VC. 2012. Business intelligence and analytics: from big data to big impact. *MIS Q.* 36(4):1165–88

Christiano LJ. 2012. Christopher A. Sims and vector autoregressions. *Scand. J. Econ.* 114(4):1082–104

Clark WR, Golder M. 2015. Big data, causal inference, and formal theory: contradictory trends in political science. *PS Political Sci. Politics* 48(1):65–70

Clarke RA, Knake R. 2011. *Cyber War: The Next Threat to National Security and What to Do About It*. New York: HarperCollins

Cleveland WS. 2001. Data science: an action plan for expanding the technical areas of the field of statistics. *Int. Stat. Rev.* 69(1):21–26

Conway D. 2013. The data science Venn diagram. **http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram**

Corbett-Davies S, Pierson E, Feller A, Goel S, Huq A. 2017. Algorithmic decision making and the cost of fairness. In *Proceedings of 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Canada*. New York: ACM. **https://arxiv.org/abs/1701.08230**

Cukier K, Mayer-Schoenberger V. 2013. The rise of big data: how it's changing the way we think about the world. *Foreign Aff.* 92(3):28–40

Deutsch KW. 1963. *The Nerves of Government: Models of Political Communication and Control*. New York: Free Press

Donoho D. 2017. 50 years of data science. *J. Comput. Graphical Stat.* 26(4):745–66

Dunlap CJ. 2014. The hyper-personalization of war: cyber, big data, and the changing face of conflict. *Georgetown J. Int. Aff.* 15:108–18

Dustdar S, Nastić S, Šćekić O. 2017. *Smart Cities: The Internet of Things, People, and Systems*. New York: Springer Int. Publ.

Dzau VJ, Ginsburg GS. 2016. Realizing the full potential of precision medicine in health and health care. *JAMA* 316(16):1659–60

Enos RD. 2016. What the demolition of public housing teaches us about the impact of racial threat on political behavior. *Am. J. Political Sci.* 60(1):123–42

Evans P. 2018. Harnessing big data: a tsunami of transformation. In *Opening Government*, pp. 137–44. Acton, ACT, Aust.: ANU Press

Farrell H. 2012. The consequences of the internet for politics. *Annu. Rev. Political Sci.* 15:35–52

Glaeser EL, Cominers SD, Luca M, Naik N. 2018. Big data and big cities: the promises and limitations of improved measures of urban life. *Econ. Inq.* 56(1):114–37

Goff PA, Lloyd T, Geller A. 2016. *The science of justice: race, arrests, and police use of force*. Rep. Cent. Policing Equity, New York, NY

Gomez-Rodriguez M, Leskovec J, Krause A. 2012. Inferring networks of diffusion and influence. *ACM Trans. Knowledge Discov. Data* 5(4):21

Granato J, Scioli F. 2004. Puzzles, proverbs, and omega matrices: the scientific and social significance of Empirical Implications of Theoretical Models (EITM). *Perspect. Politics* 2(2):313–23

Gray J. 2009. Jim Gray on eScience: a transformed scientific method. In *The Fourth Paradigm: Data-Intensive Scientific Discovery*, ed. T Hey, S Tansley, K Tolle, pp. xvii–xxxi. Redmond, WA: Microsoft Res.

Grimmer J, Messing S, Westwood SJ. 2012. How words and money cultivate a personal vote: the effect of legislator credit claiming on constituent credit allocation. *Am. Political Sci. Rev.* 106(4):703–19

Grimmer J, Stewart BM. 2013. Text as data: the promise and pitfalls of automatic content analysis methods for political texts. *Political Anal.* 21(3):267–97

Hanauer DA, Rhodes DR, Chinnaiyan AM. 2009. Exploring clinical associations using '-omics' based enrichment analyses. *PLOS ONE* 4(4):e5203

Harcourt BE. 2007. *Against Prediction: Profiling, Policing, and Punishing in an Actuarial Age*. Chicago: Univ. Chicago Press

Hashem IAT, Chang V, Anuar NB, Adewole K, Yaqoob I, et al. 2016. The role of big data in Smart City. *Int. J. Inf. Manag.* 36:748–58

Hastie T, Tibshirani R, Friedman J. 2016. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Stanford, CA: Stanford Univ. Press. 2nd ed.

Hersh ED. 2013. Long-term effect of September 11 on the political behavior of victims' families and neighbors. *PNAS* 110(52):20959–63

Hilbert M, López P. 2011. The world's technological capacity to store, communicate, and compute information. *Science* 332:60–65

Hochschild J, Sen M. 2015. Genetic determinism, technology, optimism, and race: views of the American public. *Ann. AAPSS* 661:160–80

Hopkins D, King G. 2010. A method of automated nonparametric content analysis for social science. *Am. J. Political Sci.* 54(1):229–47

Hsiang SM, Burke M, Miguel E. 2013. Quantifying the influence of climate on human conflict. *Science* 341:1235367

Hsiang SM, Meng KC, Cane MA. 2011. Civil conflicts are associated with the global climate. *Nature* 476:438–41

Jamieson K. 2018. *Cyber-War: How Russian Hackers and Trolls Helped Elect a President*. New York: Oxford Univ. Press

Jordan M. 2018. Artificial intelligence—the revolution hasn't happened yet. *Medium*. **https://medium.com/@mijordan3/artificial-intelligence-the-revolution-hasnt-happened-yet-5e1d5812e1e7**

Kalil T. 2012. Big data is a big deal. Press release, The White House, Mar. 29. **https://obamawhitehouse.archives.gov/blog/2012/03/29/big-data-big-deal**

Kandel S, Paepeke A, Hellerstein Heer J. 2011. *Wrangler: interactive visual specification of data transformation scripts*. Paper presented at CHI Conference on Human Factors in Computing Systems, May 7–12, Vancouver, BC

Kandel S, Paepeke A, Hellerstein Heer J. 2012. Enterprise data analysis and visualization: an interview study. *IEEE Trans. Vis. Comput. Graph.* 18(12):2917–26

Kaplan F. 2017. *Dark Territory: The Secret History of Cyber War*. New York: Simon & Schuster

Kim IS. 2017. Political cleavages within industry: firm-level lobbying for trade liberalization. *Am. Political Sci. Rev.* 111(1):1–20

Kim IS, Kunisky D. 2018. *Mapping political communities: a statistical analysis of lobbying networks in legislative politics*. Work. Pap., Mass. Inst. Technol., **http://web.mit.edu/insong/www/pdf/network.pdf**. Accessed Dec. 2, 2018

King G, Pan J, Roberts ME. 2013. How censorship in China allows government criticism but silences collective expression. *Am. Political Sci. Rev.* 107(2):326–43

Kitchin R. 2014. The real-time city? Big data and smart urbanism. *GeoJournal* 79(1):1–14

Kitzes J, Turek D, Deniz F. 2017. *The Practice of Reproducible Research: Case Studies and Lessons from the Data-Intensive Sciences*. Oakland: Univ. Calif. Press

Kleinberg J, Ludwig J, Mullainathan S, Obermeyer Z. 2015. Prediction policy problems. *Am. Econ. Rev. Pap. Proc.* 105(5):491–95

Knight W. 2017. The dark secret at the heart of AI. *MIT Technol. Rev.*, May/June. **https://www.technologyreview.com/s/604087/the-dark-secret-at-the-heart-of-ai/**

Laney D. 2001. *3D data management: controlling data volume, velocity, and variety*. Application Delivery Strategies File 949, Feb. 6, META Group. **https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf**

Lasswell HD. 1951. The policy orientation. In *The Policy Sciences: Recent Developments in Scope and Method*, ed. D Lerner, H Lasswell, pp. 3–15. Stanford, CA: Stanford Univ. Press

Laver M, Benoit K, Garry J. 2003. Extracting policy positions from political texts using words as data. *Am. Political Sci. Rev.* 97(2):311–31

Lazer D, Kennedy R, King G, Vespignani A. 2014. The parable of Google flu: traps in big data analysis. *Science* 343(6176):1203–4

LeCun Y, Bengio Y, Hinton G. 2015. Deep learning. *Nature* 521:436–44

Leskovec J, Backstrom L, Kleinberg J. 2009. *Meme-tracking and the dynamics of the news cycle*. Paper presented at 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, June 28–July 1, Paris, France

Libicki MC. 2014. Why cyber war will not and should not have its grand strategist. *Strateg. Stud. Q.* 8(1):23–39

Lin H, Tegmark M, Rolnick D. 2017. Why does deep and cheap learning work so well? *J. Stat. Phys.* 168(6):1223–47

Lugmayr A, Stockleben B, Scheib C. 2016. A comprehensive survey on big-data research and its implications—What is really 'new' in big data?—It's cognitive big data! In *PACIS 2016 Proceedings*, Abstr. 248. **https://aisel.aisnet.org/pacis2016/248**

Luks S, Brady HE. 2003. Defining welfare spells. Coping with problems of survey responses and administrative data. *Eval. Rev.* 27(4):395–420

Lyman P, Varian HR. 2003. *How much information? Executive summary*. Rep. School Inf. Manag. Syst., Univ. Calif., Berkeley, CA. **http://groups.ischool.berkeley.edu/archive/how-much-info-2003/execsum.htm**

Maimon O, Roach L. 2005. *The Data Mining and Knowledge Discovery Handbook*. New York: Springer

Manjoo F. 2016. A plan in case robots take the jobs: give everyone a paycheck. *New York Times*, Mar. 2. **https://www.nytimes.com/2016/03/03/technology/plan-to-fight-robot-invasion-at-work-give-everyone-a-paycheck.html**

Mayer-Schönberger V, Cukier K. 2014. *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. Boston: Houghton Mifflin Harcourt

Mbadiwe T. 2018. Algorithmic injustice. *New Atlantis* 54:3–28
Mergel I. 2016. Big data in public affairs education. *J. Public Aff. Educ.* 22(2):231–48
Miller K. 2012. Big data analytics in biomedical research. *Biomed. Comput. Rev.* Winter 2011/2012:14–21. **http://biomedicalcomputationreview.org/content/big-data-analytics-biomedical-research**
Mosco V. 2014. *To the Cloud: Big Data in a Turbulent World*. New York: Paradigm
Mullainathan S, Spiess J. 2017. Machine learning: an applied econometric approach. *J. Econ. Perspect.* 31(2):87–106
Nagler J, Tucker JA. 2015. Drawing inferences and testing theories with big data. *PS Political Sci. Politics* 48(1):84–88
National Research Council. 2011. *Toward Precision Medicine: Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease*. Washington, DC: Natl. Acad. Press
National Research Council. 2013. *Frontiers in Massive Data Analysis*. Washington, DC: Natl. Acad. Press
Neumann R. 2016. *The Digital Difference: Media Technology and the Theory of Communication Effects*. Cambridge, MA: Harvard Univ. Press
Nickerson DW, Rogers T. 2014. Political campaigns and big data. *J. Econ. Perspect.* 28(2):51–73
NIST (Natl. Inst. Standards Technol.). 2015. *Big data interoperability framework: Volume 1, definitions*. NIST Spec. Publ. 1500-1. **https://bigdatawg.nist.gov/_uploadfiles/NIST.SP.1500-1.pdf**
NITRD (Netw. Inf. Technol. Res. Dev.). 2016. *The federal big data research and development strategic plan*. Rep. Big Data Senior Steering Group, Subcomm. NITRD, Washington, DC. **https://www.nitrd.gov/PUBS/bigdatardstrategicplan.pdf**
Noble S. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York: New York Univ. Press
Oussous A, Benjelloun FZ, Lahcen AA, Belfkih S. 2018. Big data technologies: a survey. *J. King Saud Univ.—Comput. Inf. Sci.* 30(4):431–48
Picon A. 2015. *Smart Cities: A Spatialised Intelligence*. New York: Wiley
Pierson E, Simoiu C, Overgoor J, Overgoor J, Corbett-Davies S, et al. 2017. A large-scale analysis of racial disparities in police stops across the United States. arXiv:1706.05678 [stat.AP]
Pool IS. 1983. Tracking the flow of information. *Science* 221(4611):609–13
Porche IR, Wilson B, Johnson EE, Tierney S, Saltzman E. 2014. Barrier to benefiting from big data. In *Data Flood: Helping the Navy Address the Rising Tide of Sensor Information*, pp. 13–21. Santa Monica, CA: RAND Corp.
Powell J. 2017. Identification and asymptotic approximations: three examples of progress in econometric theory. *J. Econ. Perspect.* 31(2):107–24
Pratt GA. 2015. Is a Cambrian explosion coming for robotics? *J. Econ. Perspect.* 29:51–60
Prior M. 2013. Media and political polarization. *Annu. Rev. Political Sci.* 16:101–27
Rid T. 2012. Cyber war will not take place. *J. Strateg. Stud.* 35(1):5–32
Ripley BD. 1995. *Pattern Recognition and Neural Networks*. New York: Cambridge Univ. Press
Roberts M, Stewart B, Tingley D, Lucas C, Leder-Luis J, et al. 2014. Structural topic models for open-ended survey responses. *Am. J. Political Sci.* 58(4):1064–82
Rogers R. 2013. *Digital Methods*. Cambridge, MA: MIT Press
Russell S, Norvig P. 2009. *Artificial Intelligence: A Modern Approach*. New York: Pearson. 3rd ed.
Salganik MJ. 2017. *Bit by Bit: Social Research in the Digital Age*. Princeton, NJ: Princeton Univ. Press
Samuel A. 1962. Artificial intelligence: a frontier of automation. *Ann. Am. Acad. Political Social Sci.* 340:10–20
Sanger DE. 2018. *The Perfect Weapon: War, Sabotage, and Fear in the Cyber Age*. New York: Crown
Sarle W. 1994. Neural networks and statistical models. In *Proceedings of the Nineteenth Annual SAS Users Group International Conference, Dallas, Texas, Aprl 10–13*. Cary, NC: SAS Inst. **http://www.sascommunity.org/sugi/SUGI94/Sugi-94-255%20Sarle.pdf**
Schmidhuber J. 2015. Deep learning in neural networks: an overview. *Neural Netw.* 61:85–117
Schroeder R. 2018. *Social Theory after the Internet: Media, Technology, and Globalization*. London: UCL Press
Schudson M. 2002. The news media as political institutions. *Annu. Rev. Political Sci.* 5:249–69
Scott JC. 1999. *Seeing Like a State*. London: Yale Univ. Press
Shmueli G. 2010. To explain or to predict. *Stat. Sci.* 25(3):289–310
Sims CA. 1980. Macroeconomics and reality. *Econometrics* 48(1):1–48

Smith G. 2018. *The AI Delusion*. New York: Oxford Univ. Press
Statistical Science. 2003. Tribute to John W. Tukey. *Stat. Sci.* 18(3)
Stephens-Davidowitz S. 2014. The cost of racial animus on a black candidate: evidence using Google search data. *J. Public Econ.* 118:26–40
Tankersley J. 2018. Democrats' next big thing: government-guaranteed jobs. *New York Times*, May 22. **https://www.nytimes.com/2018/05/22/us/politics/democrats-guaranteed-jobs.html**
Taylor GR. 1951. *The Transportation Revolution 1815–1860*. New York: Rinehart
Thagard P. 1992. *Conceptual Revolutions*. Princeton, NJ: Princeton Univ. Press
Theodoridis AG, Nelson AJ. 2012. Of BOLD claims and excessive fears: a call for caution *and patience* regarding political neuroscience. *Political Psychol.* 33(1):27–28
Tinati R, Halford S, Carr L, et al. 2014. Big data: methodological challenges and approaches for sociological analysis. *Sociology* 48(4):663–81
Titiunik R. 2015. Can big data solve the fundamental problem of causal inference? *PS Political Sci. Politics* 48(1):75–79
Townsend AM. 2013. *Smart Cities: Big Data, Civic Hackers, and the Quest for a New Utopia*. New York/London: W.W. Norton
Tukey J. 1962. The future of data analysis. *Ann. Math. Stat.* 33(1):1–67
Turnbull N. 2008. Harold Lasswell's "problem orientation" for the policy sciences. *Crit. Policy Anal.* 2(2):72–91
Varian HR. 2014. Big data: new tricks for econometrics. *J. Econ. Perspect.* 28(2):3–27
Voigt R, Camp NP, Prabhakaran V, et al. 2017. Language from policy body camera footage shows racial disparities in officer respect. *PNAS* 114(25):6521–26
Ward JS, Barker A. 2013. Undefined by data: a survey of big data definitions. arXiv:1309.5821 [cs.DB]
Warner B, Misra M. 1996. Understanding neural networks as statistical tools. *Am. Statistician* 50(40):284–93
Weil F. 2012. The sinews of society are changing. *Huffington Post*, Apr. 17. **https://www.huffingtonpost.com/frank-a-weil/the-sinews-of-society-are_b_1277241.html**
White H. 1992. *Artificial Neural Networks: Approximation and Learning Theory*. Cambridge, MA: Blackwell
Wickham H. 2014. Tidy data. *J. Stat. Softw.* 59(10):1–24
Wiedemann G. 2013. Opening up to big data: computer-assisted analysis of textual data in social sciences. *Forum Qual. Soc. Res.* 14(2):13. **http://www.qualitative-research.net/index.php/fqs/article/view/1949**
Wigner E. 1960. The unreasonable effectiveness of mathematics in the natural sciences. *Commun. Pure Appl. Math.* 13(1):1–14
Wilkerson J, Casas A. 2017. Large-scale computerized text analysis in political science: opportunities and challenges. *Annu. Rev. Political Sci.* 20:529–44
Williams BA, Brooks CF, Shmargad Y. 2018. How algorithms discriminate based on data they lack: challenges, solutions, and policy implications. *J. Inf. Policy* 8:78–115
Yarkoni T, Westfall J. 2017. Choosing prediction over explanation in psychology: lessons from machine learning. *Perspect. Psychol. Sci.* 12(6):1100–22