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

hbrady@berkeley.edu

August 15, 2018

The Challenge of Big Data and Data Science

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 in a 2015 "framework" report on "Big Data" (NIST, 2015, 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" (OSTP 2012) that would build 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, 6)." In just the first six months of 2018, the *New York Times* published articles telling us "AI and Big Data Could Power a New War on Poverty" (January 2, 2018), "Big Data Comes to Dieting," (January 25, 2018), "How Democracy can Survive Big Data" (March 22, 2018), and "How Big Data is 'Automating' Inequality" (May 4, 2018). "Big data" appeared about 560 times per year in JSTOR from 2014-2017 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 seventeen 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 their utility (e.g., boyd & Crawford 2012; Donoho 2017).

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" have implications for 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, non-profits, and people have combined with the enormous increases in computing power and improvements in data science methods to change society in fundamental ways, 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 on line in 2022 will generate 90 terabtyes every night; genomic data is doubling every 9 months, and it is currently being produced at approximately 10 terabytes per day; the large Hadron collider at CERN generates 140 terabytes per day. The 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, and prediction 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 "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 that parallel the striking changes in transportation in the 19th century. Historians report that with the invention of the steam engine, the time for and cost of travel dropped dramatically in the 1800's creating new trading networks, new opportunities for migration, new kinds of cities with commuter suburbs, and new understandings of the world. In 1816, using horse-driven stagecoaches, mule-driven canal-boats, or sailing packets that averaged two to eight miles per hour, a trip between Philadelphia and Quebec (560 miles) took 103 hours – over four days at five to six miles per hour. By 1860 with the advent of railroads and steam-boats that went fifteen to thirty miles per hour, the time and cost for travel dropped by over two-thirds, and the same trip took just 31 hours – just over one day (estimated from Taylor 1951, Chapter VII, 141). With the regular use of jet planes in the 1960s, people could fly from New York to London in less than seven hours compared to the two-weeks it took by ship in the mid-nineteenth century—a fifty-fold improvement over 100 years. These transportation changes—along with electrification at the end of the 19th century—revolutionized society in ways that had enormous implications for politics, economics, and society.

Changes every twenty years in information technologies punctuated the history of the late 19th, 20th and early 21st century: telephones (1870-1890s), phonographs (1870-1890s), cinema (1890-1920s), radio (1900-1920s), television (1940-1950s), mainframe computers (1940-1950s), personal computers (1970-1980s), 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 *extensive digital datatification* in which myriad events are now digitally recorded, *widespread connectedness* in which events and people are identified so that they can be linked up with one another, *pervasive networking* where people are embedded in a community of users who interact with one another and become nodes in larger networks, and *ubiquitous computerized 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 first looked at the production of "words" in the American mass media (e.g., radio, television, records, movies, newspapers, books, etc.) 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 nine per cent per year. He also found that 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 percent per year" so that "each item of information produced faces a more competitive market and a smaller audience on average (Pool 1983, 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 also said that "Computer networking is for the first time bringing the costs of a point-to-point medium, data communication, down to the range of costs characteristics of mass media (611)."

Subsequent studies by political scientists and others (Lyman & Varian 2003; Bohn & Short 2012; IJC, 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 (subject matters of books). They also measured information in digital bytes instead of words so that the measures reflect the proliferation of images such as video which communicate many more bytes per second than do words through text or speech (Bohn & Short 2012, 986). Hilbert & Lopez (2011, 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 ten to fifteen years. Finally, they computed a new quantity – the growth in the world's computational power in Millions of Instructions Per Second or MIPS, and they found that the world's humanity guided general-purpose computation grew at an impressive compound annual growth rate of 58% per capita between 1986 and 2007. Embedded applications-specific computation grew even faster, at 83%.

This research identifies several notable trends that have produced the "Big Data" revolution. First, there is the tsunami of data about societal events, and digital communications are overtaking analog. This *extensive digital datafication* (Cukier & Mayer-Schoenberger 2013, 29) creates data in a format that can be readily stored and processed by computers. "Recording" might be used instead of the ugly neologism "datification," but it seems too passive for processes that are transmogrifying human interactions into data. Even though some of these data are relatively unstructured as 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.

Third, one feature of the changing information environment 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—*networking*—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 in politics through the creation of networks that are closed with respect to dissenting opinions (Neumann 2016).

Finally, whereas the communication of information traditionally involved sending messages from one place to another 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* by programs and algorithms that transform inputs into quite different outputs. Computers use programs to produce new products 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. For the first time in history, aside from naturally produced information from the environment, there is non-human production of new information. Nature and humans no longer have a monopoly on authoring. We now live in an era where 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, 2010) 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 seventy year tradition continues of building computers and robots with "Artificial Intelligence (AI)" (Russell & Norwig 2009) that can perform human-like tasks such as playing games of chess or driving cars. Some of the methods developed by AI 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 21st century the need for better ways to process and use data, especially in the sciences, were discussed under the rubric of "cyberinfrastructure" (Atkins et al. 2003; Berman & Brady 2005), but more recently the terms "Big Data" and "Data Science" have become popular.

"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, especially given Moore's "law" which successfully predicts the doubling of the number of transistors per square inch every eighteen months, 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 number of bytes, the variety in terms of text, images, audio, video, sensor, social media, and other forms of data, 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 terms of variety adds the difficulties of translating data from one form to another, and the growth in terms of velocity leads to the need to edit data “on-the-run” and to choose what is important. More recently a fourth concern, checking on the veracity of the data, adds another layer of complexity on top of volume, variety, and velocity.

Size, complexity, and technological challenges provide one definition of big data (Ward and Barker, 2013; National Research Council 2013), but they do not seem like a sufficient basis for heralding a sea-change in our data environment since the race between dataset 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, 5).” But the statistician David Donoho 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 (Donoho 2017, 747).” 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, e-mail, 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 and automobile trips, tax payments, photographs of social gatherings, weather and environmental measurements, digital images from microscopes and telescopes, and much more. When combined with the fact 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— a picture emerges less of “big data” than of “immersive data” that surrounds us and affects our lives on a daily basis. 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 (Neumann 2016; Lugmayr et al 2017; Schroeder 2018) that requires new perspectives to deal with datification, connectedness, networking, and computer authoring. These phenomena stem from the invention of new technologies including innovative methods in data science.

“Data Science” – The companion idea of “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” (Cleveland 2001, 21) by providing more resources for “computing with data” (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 (Gray 2007 in Hey et al. 2009), computer scientist Jim Gray advocated for “data-driven science” as a new scientific paradigm that uses large collections of data to make scientific discoveries. In the first part of his talk Gray proposed that there was a "need for tools to help scientists capture their data, curate it, and then visualize it" (Gray 2007, xxv), and in the second part he proposed 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" (xxv).

Starting from these ideas about data science, NIST describes data science as "the extraction of actionable knowledge directly from data through a process of discovery, or hypothesis formulation and hypothesis testing (NIST 2015, 7)." One well-known Venn diagram (Conway 2013) places data science at the intersection of three areas: computer programming skills, mathematics and statistics, and substantive expertise in a field of research. The diagram includes "machine learning (ML)" as an important aspect of data science because ML 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 for us 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 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, 2). Tukey's impact on statistics has been immense (Statistical Science 2017), and his concept of data analysis covers much of the same ground as data science. In a 2017 article, the statistician David Donoho argued 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, 748).

Donoho proposed that data science should encompass six activities to which I've added one more (5.5 below) as well as adding my gloss in brackets:

> "1. Data Gathering, Preparation, and Exploration" [e.g., 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];
>
> "2. Data Representation and Transformation" [e.g., relational databases; networks; other mathematical structures for data];
>
> "3. Computing with Data" [e.g., R and Python; programming packages, text manipulation languages; cluster and cloud computing; reproducible workflows];
>
> "4. Data Modeling" [e.g., determining or hypothesizing data generating probability functions, structural and predictive modeling];
>
> "5. Data Visualization and Presentation" [e.g., types of visualizations and graphs; rules for labeling and presenting data; psychological impacts of various displays];

> 5.5 Data archiving, indexing, and search and data governance such as standards for open data and reproducibility; determining rules for access and privacy protection where necessary.
>
> "6. Science about Data Science" [e.g., ways that people do data science and the impacts of data science and big data on society]." (Donoho 2017, 755).

Judging from this list, data science borrows methods and techniques that go beyond the traditional core of statistics which is largely encompassed in "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 computing with data. 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 the exact way that the applied data science being done in these disciplines can be incorporated into these programs. For example, in addition to benefitting 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 listed above means that there must be some way for universities to bring together the people working on them to learn from one another and to be able to teach the next generation of students and scholars what they need to know 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 where 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 and some statistics packages such as BMDP-SPSS-SAS-STATA-GAUSS, 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, however, is hard to find. Just as the 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 of rails, roads, and rivers, and even the development of new social norms such as standard time zones—the information revolution is more than just computers or any other single thing. It also involves sensors, data bases, programming languages, artificial intelligence, telecommunications, machine learning, social media, the Internet, and many other inventions. Neither "big data" nor "data science" nor any other words or phrases encompass all these innovations. The term cyberinfrastructure might have been a useful one, but it has not caught on. "Artificial Intelligence" is too limited. One leading data science scholar (Jordan 2018) argues for the use of the term "Intelligent infrastructure," 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 that are being affected by "big data" (Chen at al. 2012; Cukier & Mayer-Schoenberger 2013, 2014; Evans 2018; Mosco 2014). We cannot provide an exhaustive review of the possible societal impacts of big data and data science, but it is worth listing a few prominent examples to show how they deserve more scrutiny by political scientists. I have chosen cyberwarfare and homeland security, smart cities, medicine, and the media.

Several recent books have proposed that "cyber warfare" is here and a threat to international security (e.g., Clarke & Kanke 2011; Kaplan 2017), but skeptics (Rid 2012, Libicki 2014) have argued that while cyber disruptions may be a problem, they do not constitute classical warfare like the Japanese attack on Pearl Harbor on December 7, 1941 which involved a purposeful and publicly claimed act of violence for political advantage. Some of the leading examples of "Cyber War" such as the Stuxnet virus's introduction into Iranian centrifuges (an essential part of Iran's nuclear fuels enrichment program) that led to their destruction 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, 5) as recent events involving Russia and the 2016 American election make clear (Sanger 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 Spatialized 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 datasets on people and their relationship to schools, social welfare agencies (Brady et al. 2001), medical care, and police, and there are similar datasets on physical structures and their relationship to streets, services, land-use and zoning. Second, the reduced costs of sensors, wireless networks, video cameras and the ability to connect them with an "Internet of Things" makes it possible to monitor and sometimes remotely control air pollution, traffic, electricity usage, utilities, parking, safety, water usage, police and fire 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 geocoding 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 RFID tags which makes it possible to track movements throughout the city (Hashemet al. 2016)

Using these data, the city and its operations can be described, managed, and evaluated. Traffic, air pollution, or poverty maps can provide useful *descriptions* for those trying to understand where to live, where to travel, or what to do. Improved conditions on a real-time basis can be *managed* 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, and 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).

"Toward Precision Medicine," a 2011 report of the National Research Council of the National Academy of Sciences defined precision medicine as "the tailoring of medical treatment to the individual characteristics of each patient" (125). To practice precision medicine, information about the individual must be combined with medical knowledge about how people vary in their response to illnesses and treatments (Dzau et al. 2016). Individual information would come from electronic medical records and genomic data. The 2011 report suggested creating a new taxonomy of human disease based upon molecular biology that would serve as the starting place 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, Olson, Berger 2015), and these data would be used to explore clinical associations (Hanauer et al. 2009; Miller 2011-12). 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 the first examination of the mass media in the *Annual Review of Political Science* in 2002, Michael Schudson 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." (Schudson 2002, 249). Yet he does not even mention the Internet or World Wide Web, and he focuses on the relative merits of state versus commercial controlled media, journalism as "the story of the interaction of reporters and government officials" (255), and the cultural norms that shape coverage of topics such as homosexuality and crime. Schudson concludes that "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" (263). Not even a hint comes through about the possible anarchy of uncontrolled news "sources" and direct leader-follower communications now bedeviling a world with Facebook, Google, and Twitter.

By 2012, Farrell's *Annual Review* 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 intense people to find one another. Although Prior still concludes in his 2013 *Annual Review* article on "Media and Political Polarization" that "Internet use shows few signs of ideological segregation (Prior 2013, 122)," he takes the Internet seriously. And communications theorists such as Bennett and 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,* James C. Scott (1999) has chronicled how states have misused census and other information. What will it mean when societies, businesses, and governments have access to large datasets 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, cell-phone 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 kinds of problems are created when information can be hacked and digital systems are vulnerable to viruses? When medical diagnoses or city operations depend upon algorithms that sometime fail? What kinds of 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 and the cloud, and as they become embodied in autonomous machines with their own light-weight power sources, to what degree do they become an organism that has 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 (Tankersly 2018) as ways to deal with the possibility of job loss due to robots. What kinds of political problems does this raise, or is a 1962 article right when it concludes that "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, 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 entire universe: 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 3-D microscopy of cells at many different wavelengths to understand their operations, scanning animal and human brains and bodies using 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 datasets could be useful to political scientists such as fMRI data for those studying political psychology (Theodoridis 2012) or those studying the impacts of climate change on politics (Hsiang et al. 2013).

Social scientists have benefitted from many new data sources as well. As of roughly 1980, political scientists had available a limited number of datasets, 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 and 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 thirty years, the volume and variety of data have increased enormously beyond these areas especially in terms of 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 of millions of registered voters in Los Angeles and their precinct location 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 towards the Republican party. Using voter registration files for the city of Chicago, Enos (2016) examines the impact of racial threat on voter turnout by using a natural experiment in which public housing buildings with over 25,000 African Americans living in them were demolished. He categorizes each voter's race using a Bayesian classifier based upon 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 which presumably reduced the sense of threat. Ansolabehere and 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 socio-economic 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 to 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 (PACS) 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." (298) The technique provides estimates for first-time candidates who have no roll-call records available to estimate their political positions, and Bonica shows that using his ideological estimates for candidates only provides "a negligible reduction in predictive power of legislative voting behavior" (308) compared to roll-call votes. In other papers he connects these data with contributions by lawyers (Bonica et al. 2015) and doctors (Bonica et al. 2014) by linking the contributions dataset to listings of these professionals. In a more recent paper he (Bonica 2016) 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" (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 develops a three staged process "for measuring preferences and expressed priorities across issue dimensions that combines topic modeling, ideal point estimation, and machine learning methods." (18) 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, In Song 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 Congressional bills that had been lobbied, and information about tariffs and trade (Kim 2017, 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, 13)

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 datasets on voting, lobbying, campaign contribution, trade, tax, welfare, police reports, 311 calls, and many areas often provide the (legally) definitive data on these activities, but the datasets can contain errors (Luks & Brady 2003). Moreover, in order to get a dataset that represents different areas and that has enough cases for analysis, they often require, as in the police report studies, *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 is added as in the work by Bonica and Kim. Brady et al (2001, 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 socio-economic 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., income data are reliable for welfare programs because they are part of the application process but education data are not). Intensive linking to other datasets can often provide a tremendous expansion in their utility, 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 rather than the complete universe of those who might have done things. For example, data on police traffic stops tells us who was stopped but not who should have been stopped. Campaign contribution data tell us who gave money, but they provide no way to figure out the rate of giving because we only know 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 as in the police stops example where we might 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 and 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., that for Barack Obama and Mitt Romney) necessitating steps to adjust for different base frequencies for liking candidates. For those candidates for whom there is an independent measure of ideology from Congressional roll-call data, the correlation between the two measures of ideology was .47 and .42 for Democrats and Republicans respectively (Bond & Messing 2015, 68). Similarly, using Twitter users from six countries, Barbera (2015) identify Twitter "followers" of three or more political actors and use 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 and his collaborators (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 though retweets and that a small number of people are key players. Gomez and his collaborators (Gomez 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 be 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, Backstrom and Kleinberg (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 sources imitate one another producing persistence of memes and recency in which new phrases are preferred producing extinction of older memes.

In a Facebook study (Bond et al. 2012), researchers 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 mid-term election. The "social message group" of sixty million people were also shown up to six faces of their friends who had reported that they voted on Facebook that day. The "informational message group" of over six hundred thousand people received just 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 2 percentage points more likely to say that they had voted than those in just the informational message group, and other significant effects were found.

King and his colleagues studied the motivation of Chinese Internet censorship by following the fate of blog posts over time (King et al. 2013). 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” (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 four percent 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, web data are typically highly selective in terms of socio-economic characteristics (especially by having more young people, although older people are catching up) and they often depend upon 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 2014). Missing data is also a problem as in the studies estimating ideology using Facebook and Twitter data. The compensating merits are that they 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 efficiencies, 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 U.S. 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 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.

A recent *Annual Review* article (Wilkerson & Casas 2017) and one somewhat older article (Grimmer & Steward 2013) provide excellent overviews of the profusion of content analysis methods developed in the last fifteen 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 range from the search for particular words or phrases as in the Stephens-Davidovitz (2014) or Leskovec, Backstrom and Kleinberg (2009) articles described earlier; the determination of what fractions of text fit into pre-determined categories as in King et al (2013) and Grimmer et al. (2012) described earlier; the classification of each text into pre-determined categories using supervised learning; the classification of text into unknown categories using unsupervised clustering methods; or the ideological scaling of political texts such as party platforms (Laver, Benoit, & Garry 2003).

These methods require careful use because as Grimmer and Steward note "All Quantitative Models of Language are Wrong—But Some are Useful" and "Quantitative Methods Augment Humans, Not Replace Them" so "Validate, Validate, Validate" (Grimmer & Steward 269, 270, 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 so that one cannot guarantee a globally correct solution, lasso or ridge regression that strive for simplicity that might under-fit the data, and models with many parameters that often try to estimate values for each document with what amounts to small amounts of data. To perform these tasks they often use estimation methods such as the EM algorithm or Bayesian MCMC that take a long time to converge and can be tricky to use. For a discussion of the issues involved see, for example, 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 has progressed to a point where quantifying large bodies of text is possible and arguably an improvement over human coding because of its effectiveness 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 – Sol Hsiang and his colleagues (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 Interstate Conflict" dataset to study the impact of weather on civil conflicts. They use the El Nino/Southern Oscillation (ENSO) in weather to identify their model, and they find that the probability of new civil conflicts doubles during El Nino 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 (Voigt et al. 2017) use body camera data from stops by Oakland police officers 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 towards people they have stopped. They note that "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, 5)

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 or the modifiable areal unit problem (MAUP) in geography that 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 scientists 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 it means 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 underway. These courses include programming in R or Python, an emphasis on resampling approaches to understanding statistics, an overview of the data sources described earlier, 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 Boullier (2015), Salganik (2017), Mayer-Schoenberger & Cukier (2014), Rogers (2013), and Mosco (2015), provide some food for thought about what will happen when we make “the world self-aware and self-describing” (Evans 2018, 141).

New Data 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 (Oussous et al. 2017; Varian 2014) 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, Paepcke, Hellerstein, & Heer 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 upon 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. 2018, 13).” In Kitzes et al. (2018) reproducibility is exemplified 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 work-flow. 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 towards requiring it for their 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 upon 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 upon body mass, age, serum insulin, etc.) whereas statisticians talk about predicting the value of a *dependent variable* based upon *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, where the former refers to problems where there is information on the relevant classes (e.g., specimens already classified into separate species) and the later refers to problems without this information. Supervised learning uses methods with a dependent variable like discriminant analysis or logistic regression whereas unsupervised learning uses clustering, factor analysis, or multi-dimensional scaling. Once the newcomer to the field of data science recognizes these differences in nomenclature, books on "pattern recognition" (Ripley 1995), artificial intelligence (Bishop & Norvig 2009), "machine learning" (Bishop 2011) and "statistical learning" (Hastie et al. 2016) seem less arcane and much more approachable. Newcomers can also benefit from articles that bridge the gaps including (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 these models can be estimated effectively and efficiently using Markov Chain Monte Carlo (MCMC) and other methods. Second, "smoothing" or "regularizing" approaches that require the estimation of non-linear 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 like the bootstrap, bagging, boosting, Bayesian model averaging, and random forests have become common-place 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., E-M and genetic algorithms, MCMC methods, back-propagation) to estimate models with complicated density mixtures, large numbers of parameters, multiple local maxima, and knotty non-linearities 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, bioinformatics and other areas (LeCun, Bengio, & Hinton 2015). Deep learning is a variant of the canonical feed-forward neural network which involves multi-layer 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" or 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 a dog and not a cat so that 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 upon the number of hidden layers and the number of hidden variables or neurons.

We have known for over twenty-five years that systems with at least one hidden layer were "universal approximators" (White 1992) that could, with relatively arbitrary activation functions, approximate to any degree non-linear 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. 2018), and this evokes a famous paper by the physicist Eugene Wigner on “The Unreasonable Effectiveness of Mathematics in the Natural Sciences” (Wigner 1960). 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: (1) Forming concepts and providing measures of them; (2) Providing reliable descriptive inferences; (3) Making causal inferences from past experience; and (4) 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. 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, and one of the most informative studies of concept formation (Thagard 1992) used AI 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 2011) that allow for the evaluation of uncertainty in typologies, but concepts such as an “atom,” “species,” “democracy” or “topic” are still very deep ideas that are based upon 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 because 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 it, and to display it effectively. Big Data helps us with descriptive inferences because it often provides a complete list of arrests, registered voters, or food stamps recipients, but the problem of defining the proper universe remains since we may care about crimes, potential voters, or those eligible for food stamps. 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 it represents 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 lots of data (either many cases or 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 Titiunik 2015, Lazer et al. 2014). The second inclination is 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, 39, 32) 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 "We can learn from a large body of information things we could not comprehend when we used only smaller amounts." The third and more defensible inclination is the notion that making reliable causal inferences is so hard that we should focus on prediction. This idea led to vector auto-regression methods in macroeconomics (Sims 1980; Christiano 2012) forty 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 tradeoffs.

There are certainly practical and technical problems for which achieving a good prediction using machine or statistical learning is a satisfactory, and perhaps optimal, answer to the problem. 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 whether fMRI data can be used to infer whether people's memories about faces are accurate. 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 upon 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 where 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, 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 upon 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 auto-regressions with concerns about causal mechanisms and macro-economic models (Christiano 2012). Athey (2018) discusses some of the 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 upon 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 increases (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, especially in the field of criminal justice. A number of writers (Mbadiwe 2018; Harcourt 2007; Williams, et al. 2018) have worried that algorithms used to assign bail, decide upon sentences, or place prisoners in various levels of detention have “baked” into them predictions that are not causal, that reproduce pre-existing 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, Pierson, Feller, Goel, & Huq 2017, 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 voters who are under-represented may become even more under-represented because it costs more to mobilize them (Brady, et al. 1999).

Athey (2018) notes that predictive algorithms can not only be unfair, they 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 their work on political campaigns, on the impacts of the media, on the operation of cities, on terrorism and cyber-warfare, 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, this would be a useful turn back towards the "policy sciences" advocated by Harold Lasswell (1951; Turnbull 2008), but political scientists will undoubtedly find themselves taking on new roles that will require debate and discussion within the profession.

REFERENCES

Ahlquist JA, Breunig C. 2012. Model-based Clustering and Typologies in the Social Sciences. *Political Analysis.* 20(1): 92-112

Albus JS. 1984. Robots and the Economy. *The Futurist*. 18(6): 38-44.

Alvarez R. 2016. *Computational Social Science: Discovery and Prediction (Analytical Methods for Social Research).* Cambridge, Cambridge University Press.

Ansolabehere S, Hersh E. 2012. Validation: What Big Data Reveal About Survey Misreporting and the Real Electorate. *Society for Political Methodology*. 20(4): 437-459

Athey S. 2018. The Impact of Machine Learning on Economics. http://www.nber.org/chapters/c14009.pdf

Atkins DE, Droegemeier KK, Feldman SI, et al. 2003. Revolutionizing Science and Engineering Through Cyberinfrastructure: Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure

Bail CA, 2014. The cultural environment: measuring culture with big data. *Theory and Society.* 43(3/4): 465-482

Barberá P. 2015. Birds of the Same Feather Tweet Together: Bayesian Ideal Point Estimation Using Twitter Data. *Political Analysis.* 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. Institute of Medicine. The National Academy Press.

Bennett WL, Segerberg A. 2012. The Logic of Connective Action. *Information, Communication & Society.* 15(5): 739-768

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

Berman F, Brady H. 2005. Final Report. NSF SBE-CISE Workshop on Cyberinfrastructure in the Social Sciences, http://www.sdsc.edu/sbe/

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

Bohn R, Short J. 2012. Measuring Consumer Information*, International Journal of Communication*

Bond RM, Fariss CJ, Jones JJ, et al. 2012. A 61-milllion-person experiment in social influence and political mobilization. *Nature.* 489: 295-298

Bond R, Messing S. 2015. Quantifying social Media's Political Space: Estimating Ideaology from Publicly Revealed Preferences on Facebook. *American Political Science Review.* 109(1): 62-78

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: The Russell Sage Foundation Journal of the Social Sciences.* 2(7): 11-32

Bonica A. 2013. Ideology and Interests in the Political Marketplace. *American Journal of Political Science.* 57(2): 294-311

Bonica A, Rosenthal H, Rothman DJ. 2014. The Political Polarization of Physicians in the United States: An Analysis of Campaign Conributions to Federal Elections, 1991 Through 2012. JAMA Intern Med. 174(8):1308-1317

Boullier D. 2015. *The Social Sciences and traces of big data: Society, opinion, or vibrations?* Sciences Po University Press. 65(5-6): 71-93

boyd D, Crawford K. 2012. Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon. *Information, Communication, and Society*. 15(5): 662-79.

Brady HE. 2009. Causation and Explanation in Political Science. IN Robert Goodin (editor). *The Oxford Handbook of Political Science*. Oxford: Oxford University Press.

Brady HE, Grand SA, Powell MA, Schink W. National Research Council. 2001. Access and Confidentiality Issues with Administrative Data. In *Studies of Welfare Populations: Data Collection and Research Issues.* pp. 220-274. Washington DC, National Academies Press.

Brady HE, Schlozman KL, Verba S. 1999. Prospecting for Participants: Rational Expectations and the Recruitment of Political Activists. *The American Political Science Review.* 93(1): 153-168

Brady HE, Schlozman KL, Verba S. 2015. Political Mobility and Political Reproduction from Generation to Generation. *The ANNALS of the American Academy of Political and Social Science*. 657(1): 149-173

Brady HE, McNulty JE. 2011. Turning Out To Vote: The Costs of Finding and Getting to the Polling Place. *The American Political Science Review.* 105(1): 115-134

Breiman L. 2001. Stastistical Modeling: The Two Cultures. *Statistical Science.* 16(3): 199-231

Chen H, Chiang RHL, Storey VC. 2012. Business Intelligence and Analytics: From Big Data to Big Impact. *MIS Quarterly*. 36(4): 1165-1188

Christiano LJ. 2012. Christopher A. Sims and Vector Autoregressions. *The Scandinavian Journal of Economics*. 114(4): 1082-1104. Oxford, UK.

Clark WR, Golder M. 2015. Big Data, Causal Inference, and Formal Theory: Contradictory Trends in Political Science. *PS: Political Science and Politics.* 48(1): 65-70

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

Cleveland WS. 2001. Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics. 69(1): 21-26

Conway D. 2013. The data science venn diagram.

Corbett-Davies S, Pierson E, Feller A, et al. 2017. *Algorithmic Decision Making and the Cost of Fairness*. Proceedings of 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Canada

Cukier K, Mayer-Schoenberger V. 2013. The Rise of Big Data: How It's Changing the Way We Think About the World. *Foreign Affairs.* 92(3): 28-40

Deutsch KW. 1963. The Nerves of Government: Models of political communication and control. *The Free Press.* New York.

Donoho D. 2017. 50 Years of Data Science. *Journal of Computational and Graphical Statistics.* 26:4, 745-766

Dustdar S, Nastić *S, Šćekić O. 2017. Smart Cities: The Internet of Things, People, and Systems. Switzerland, Springer International Publishing*

Dunlap CJ, The Hyper-Personalization of War: Cyber, Big Data, and the Changing Face of Conflict. *Georgetown Journal of International Affairs.* 15:108-118

Dzau VJ, Ginsburg GS. 2016. Realizing the Full Potential of Precision Medicine in Health and Health Care. *JAMA*. 316(16): 1659-1660

Enos RD. 2016. What the Demolition of Public Housing Teaches Us abou the Impact of Racial Threat on Political Behavior. *American Journal of Political Science.* 60(1): 123-142

Evans P. 2018. Harnessing big data: A tsunami of transformation. In *Opening Government, ANU Press.* pp.137-144. Australia/New Zealand

Glaeser, EL, Cominers SD, Luca M, and Naik N. 2018. Big Data and Big Cities: the Promises and Limitations of Improved Measures of Urban Life. *Economic Inquiry*. 56(1):114-137.

Goff PA, Lloyd T, Geller A. 2016. The Science of Justice: Race, Arrests, And Police Use Of Force. Center for Policing Equity. Rep. New York, NY

Gomez-Rodriguez M, Leskovec J, Krause A. 2012. Inferring Networks of Diffusion and Influence. *ACM Transactions on Knowledge Discovery 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). *Perspectives on Politics.* 2(2):313-23.

Gray J. 2007. Jim Gray on eScience: A Transformed Scientific Method. In *The Fourth Paradigm.* Hay et al. xvii-xxxi

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. *American Political Science Review.* 106(4): 703-719

Grimmer J, Stewart BM. 2013. Text as Data: The Promise and Pitfalls of Automatics Content analysis Methods for Political Texts. *Political Analysis.* 21(3): 267-297

Hanauer DA, Rhodes DR, Chinnaiyan AM. 2009. Exploring Clinical Associations Using '-Omics' Based Enrichment Analyses. *PLoS One.* 4(4): 1-7

Harcourt BE. *Against Prediction: Profiling, Policing, and Punishing in an Actuarial Age.* Chicago. The University of Chicago Press.

Hashem IAT, Chang V, Anuar NB, et al. 2016. The Role of Big Data in Smart City. *International Journal of Information Management,*

Hastie T, Tibshirani R, Friedman J, et al. 2008. *The Elements of Statistical Learning – Data Mining, Interference, and Prediction.* Stanford, 2nd ed.

Hastie T., Friedman J., Tibshirani R. 2001. Neural Networks. In: *The Elements of Statistical Learning*. Springer Series in Statistics. Springer, New York, NY

Hersh ED. 2013. *Long-term effect of September 11 on the political behavior of victims' families and neighbors*. Proceedings of the National Academy of Sciences, 110(52): 20959-20963

Hey T, Tansley S, Tolle K. 2009. *The Fourth Paradigm Data-Intensive Scientific Discovery.* Redmond, Washington

Hilbert M, L*ópez 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. ANNALS, *AAPSS.* 661: 160-180

Hsiang SM, Meng KC, Cane MA. 2011. Civil conflicts are associated with the global climate. *Nature.* 476: 438-441

Hsiang SM, Burke M, Miguel E. 2013. Quantifying the Influence of Climate on Human Conflict. *Science*. 341: 1235367

Jordan M. 2018. Artificial Intelligence – The Revolution Hasn't Happened Yet. *Medium*

Kandel S, Paepeke A, Hellerstein, Heer J. 2011. *Wrangler: Interactive Visual Specification of Data Transformation Scripts.* Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. Vancouver, Canada

Kandel S, Paepeke A, Hellerstein, Heer J. 2012. Enterprise Data Analysis and Visualization: An Interview Study. *IEEE Trans. Vis. Comput. Graph.* 18(12): 2917-2926

Kaplan F. 2016. *Dark Territory: The Secret History of Cyber War.* NY, Simon & Schuster

Kim IS, Kunisky D. 2017. Mapping Political Communities: A Statistical Analysis of Lobbying Networks in Legislative Politics. (Abstr.)

*Kim IS. 2017. Political Cleavages within Industry: Firm-level Lobbying for Trade Liberalization. *American Political Science Review.* 111 (1): 1-20

*King G, Pan J, Roberts ME. 2013. How Censorship in China Allows Government Criticism but Silences Collective Expression. *The American Political Science Review.* 107 (2): 326-343

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, CA: University of California Press.

Kleinberg J, Ludwig J, Mullainathan S, Obermeyer Z. 2015. Prediction Policy Problems. *America Economic Review: Papers & Proceedings*. 105(5): 491-495

Knight W. 2017. The Dark Secret at the Heart of AI. *MIT Technology Review.* Cambridge, MA

Laney D. 2001. 3D Data Management: Controlling Data Volume, Velocity, and Variety. META Group

Lasswell HD, 1951. The Policy Orientation. In D. Lerner and H. Lasswell, *The Policy Sciences*: Recent developments in scope and method, Stanford, Stanford University Press, 3-15

Laver M, Benoit K, Garry J. 2003. Extracting Policy Positions from Political Texts Using Words as Data. *The American Political Science Review.* 97(2): 311-331

Lazer D, Kennedy R, King G, Vespignani, 2014. The Parable of Google Flu: Traps in Big Data Analysis. Science. 343 (6176): 1203-1204.

LeCun Y, Bengio Y, Hinton G. 2015. Deep learning. *Nature.* 521: 436-444

Leskovec J, Backstrom L, Kleinberg J. 2009. *Meme-tracking and the Dynamics of the New Cycle*. Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, France

Libicki MC. 2014. Why Cyber War Will Not and Should Not Have Its Grand Strategist. *Strategic Studies Quarterly.* 8(1): 23-39

Lugmayr A, et al. 2016. *A Comprehensive survey on Big-Data Research and Its Implications – What is Really 'New' in Big-Data? – It's Cognitive Big Data!* Proceedings of PACIS. Taiwan.

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

Lyman P, Varian HR. 2003. How Much Information? Executive Summary. http://groups.ischool.berkeley.edu/archive/how-much-info-2003/execsum.htm

Maimon O, Roach L. 2005. The Data Mining and Knowledge Discovery Handbook. Springer

Manjoo F. 2016. A Plan in Case Robots Take the Jobs: Give Everyone a Paycheck. *The New York Times*. March 2

Mayer-Schönberger V, Cukier K. 2014. *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. UK, John Murray

Mbadiwe T. 2018. Algorithmic Injustice. *The New Atlantis*. (54):3-28

Mergel I. 2016. Big Data in Public Affairs Education. *Journal of Public Affairs Education*. 22(2) 231-248

Miller K. 2012. Big Data Analytics in Biomedical Research. *Biomedical Computation Review.* http://biomedicalcomputationreview.org/content/big-data-analytics-biomedical-research

Mosco V. 2014. To the Cloud: Big Data in a Turbulent World. New York.

Mullainathan S, Spiess J. 2015. Machine Learning: An Applied Econometric Approach. *Journal of Economic Perspectives*. 31(2): 87-106

Nagler J, Tucker JA. 2015. Drawing Inferences and Testing Theories with Big Data. *PS: Political Science and Politics.* 48(1): 84-88

National Research Council. 2013. Frontiers in Massive Data Analysis*. The National Academies Press.* Washington DC.

Neumann R. 2016 *The Digital Difference: Media Technology and the Theory of Communication Effects.* Cambridge, MA. Harvard University Press

Nickerson DW, Rogers T. 2014. Political Campaigns and Big Data*. The Journal of Economic Perspectives*. 28(2): 51-73

NIST. 2015. Big Data Interoperability Framework: Volume 1, Definitions. National Institute of Standards and Technology

NITRD. 2016. The Federal Big Data Research and Development Strategic Plan

Oussous A, Benjelloun FZ, Lahcen AA, Belfkih S. 2017. Big Data technologies: A survey. *Journal of King Saud University – Computer and Information Sciences*.

Picon A. 2015. *Smart Cities: A Spatialised Intelligence.* UK, Wiley

Pierson E, Simoiu C, Overgoor J, et al. 2017. A large-scale analysis of racial disparities in police stops across the United States. eprint arXiv:1706.05678 (Abstr.)

Porche IR, Wilson B, Johnson EE, et al. 2014. Barrier to Benefiting from Big Data. In *Data Flood.* US, RAND Corporation

Powell J. 2017. Identification and Asymptotic Approximations: Three Examples of Progress in Econometric Theory. *Journal of Economic Perspectives*. 31(2): 107-24.

Pratt GA. 2015. Is a Cambrian Explosion Coming for Robotics? *The Journal of Economic Perspectives*. 29(#): 51-60

Rid T. 2012. Cyber War Will Not Take Place. *Journal of Strategic Studies.* 35(1): 5-32

Ripley BD. 1995. *Pattern Recognition and Neural Networks.* Cambridge, NY. Cambridge University Press.

Rogers R. 2013. Digital Methods. *The MIT Press.* Cambridge, MA

Russell SJ, Norvig P. 1995. Artificial Intelligence: A Modern Approach

Salganik MJ. 2017. *Bit by Bit: Social Research in the Digital Age.* Princeton, NJ. Princeton University Press

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.* Proceedings of the Nineteenth Annual SAS Users Group International Conference, Cary, NC

Schmidhuber J. 2015. Deep Learning in Neural Networks: An Overview. *Neural Networks.* 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. *Annual Review of Political Science.* 5:249-69

Shmargad Y. 2018. Structural diversity and tie strength in the purchase of a social networking app. *Journal of the Association for Information Science and Technology.* 69(5): 660-674

Shmueli G. 2010. To Explain or to Predict. *Statistical Science.* 25(3): 289-310

Sims CA. 1980. Macroeconomics and Reality. *Econometrics.* 48(1): 1-48

Scott JC. 1999. *Seeing Like a State.* London, Yale University Press

Statistical Science. Tribute to John W. Tukey. 2003. Institute of Mathematical Statistics. 18(3)

Stephens-Davidowitz S. 2014. The cost of racial animus on a black candidate: Evidence using Google search data. *Journal of Public Economics.* 118: 26-40

Tankersley J. 2018. Democrats' Next Big Thing: Government-Guaranteed Jobs. *New York Times,* May 22

Taylor GR. 1951. The Transportation Revolution 1815-1860. New York: Rinehart

Thagard P. 1992. *Conceptual Revolutions.* Princeton, New Jersey. Princeton University Press.

Theodoridis AG, Nelson AJ. 2012. Of BOLD Claims and Excessive Fears: A Call for Caution *and Patience* Regarding Political Neuroscience. *Political Psychology.* 33(1): 27-28

Tinati R, Halford S, Carr L, et al. 2014. Big Data: Methodoligical Challenges and Approaches for Sociological Analysis. *Sociology.* 48(4): 663-681

Titiunik R. 2015. Can Big Data Solve the Fundamental Problem of Causal Inference? *PS: Political Science & Politics.* 48(1): 75-79

Townsend AM. 2013. *Smart Cities: Big Data, Civic Hackers, and the Quest for a New Utopia.* NY/London, W.W. Norton & Co

Turnbull N. 2008. Harold lasswell's "problem orientation" for the policy sciences. *Critical Policy Analysis.* 2(2): 72-91

Varian HR. 2014. Big Data: New Tricks for Econometrics. *The Journal of Economic Perspectives.* 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 Early Edition.*

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. *The American Statistician.* 50(40): 284-293

Weil F. 2012. The Sinews of Society Are Changing. *The Huffington Post.* 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 Publishers

Wickham H. 2014. Tidy Data. *Journal of Statistical Software*. 59(10): 1-24

Wiedemann G, 2013. Opening up to Big Data: Computer-Assisted Analysis of Textual Data in Social Sciences Forum Qualitative Sozialforschung / Forum: Qualitative Social Research, 14(2), Art. 13, http://www.qualitative-research.net/index.php/fqs/article/view/1949

Wigner E. 1960. The Unreasonable Effectiveness of Mathematics in the Natural Sciences. *Communications in Pure and Applied Mathematics.* 13(1)

Wilkerson J, Casas A. 2017. Large-Scale Computerized Text Analysis in Political Science: Opportunities and Challenges. *Annual Review of Political Science.* 20: 529-44

Williams BA, Brooks CF, Shmargad Y. 2018. How Algorithms Discriminate Based on Data they Lack: Challenges, Solutions, and Policy Implications. *Journal of Information Policy.* 8: 78-115

Yarkoni T, Wastfall J. 2017. Choosing Prediction Over Explanation in Psychology: Lessons From Machine Learning. *Perspectives on Psychological Science.* pp.1-23

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