Political Science and Big Data: Structured Data, Unstructured Data, and How to Use Them

JONATHAN GROSSMAN
AMI PEDAHZUR

BIG DATA ARE A SALIENT FEATURE of the information tsunami that characterized the end of the twentieth and the beginning of the twenty-first century. As a result of the incessant rise in computational power, communication velocity, and storage capacity, new knowledge is accumulating at an exponential rate. Between 2006 and 2011, the amount of data in the world increased almost ninefold. Today, it is expected to double every two years.[1]

Computer and data scientists have been studying big databases for a while. Some of them see the analysis of such data as a panacea for all scientific questions—an omnipotent power that renders theory and within-case analysis redundant and can predict future trends and generate policy recommendations with the magic of numbers. Conversely, social scientists in general and political scientists in particular have only recently turned their attention to big data analysis. In contrast to most

[1]Min Chen, Shiwen Mao, and Yunhao Liu, "Big Data: A Survey," *Mobile Networks and Applications* 19 (April 2014): 171–209, at 171.

JONATHAN GROSSMAN is a postdoctoral fellow at the Leonard Davis Instititue for International Relations, the Hebrew University of Jerusalem. His main areas of research are social science research methodology, diaspora politics, and diplomatic history. **AMI PEDAHZUR** is a professor of government at the University of Texas at Austin. His most recent research explores the evolution of warfare since the industrial revolution.

POLITICAL SCIENCE QUARTERLY | Volume 135 Number 2 2020 | www.psqonline.org

DOI: 10.1002/polq.13032

data analysts, political scientists are far more suspicious of big data, and rightly so.

Our objective in this article is to solidify the conceptual foundations for the use of big data in political science and policy research. We discuss the main benefits of big data in the discipline and, more importantly, their potential to advance scholarship in the field by drawing on both structured and unstructured data. We contend that political scientists have not expressed much excitement about the rise of big data, for good reason. The analysis of increasingly large troves of data has been a feature of political research for several decades now. As investigators were unveiling the benefits of such large data depots, they became increasingly aware of their limitations: big databases have offered insufficient information about context, processes over time, and interactions among variables and actors. Without the ability to capture these nuances, generalizations and statistical inferences based on big data have often been flawed or superficial.

These problems stem, to a large degree, from the structured nature of the big data sets that researchers commonly analyze. Discussions of big data tend to ignore the fact that such structured databases constitute only a small part of what qualifies as big data, whereas most data in the world are unstructured. The recent surge in the literature on evidence-based research methods that focus on in-depth analyses of unstructured sources is thus timely. While data scientists often refer to unstructured data of any size as a hurdle, we maintain that in light of the advancements in both technology and political science methodology, unstructured big data are now easier to collect and organize and can prove useful in solving some of the most pertinent issues that preoccupy the discipline. When used with caution and rigor, they can help us test theories and establish causal relationships without losing meaning or context.

To substantiate our claims, we begin by clarifying the terminology for data and big data. Second, we contextualize this discussion by reviewing the recent, and quite limited, use of the term "big data" in political science and policy publications; the actual analysis of big data in such studies, which is considerably more prevalent; and the challenges of using structured big data in social science research. Third, we briefly discuss, through examples, the merits of unstructured big data and their potential to contribute to burgeoning methodological evidence-based approaches in political science. Fourth, we argue that the research tradition of historical institutionalism and the research method of process tracing can particularly benefit from incorporating unstructured big data sources. In the final section, we discuss the reasons why most historical institutionalists and process tracers are still unenthusiastic

about big data, and we offer a pathway to overcome these obstacles by utilizing unstructured big data in an organized and systematic manner.

WHAT MAKES DATA "BIG"?

Big data are a product of the information revolution—breakthroughs in information technology in the late twentieth century, both in hardware and software, that have transformed human society, economy, and culture in most parts of the world.[2] From a historical perspective, the information revolution is still in its infancy. While it is too early to fully assess its magnitude, it would not be presumptuous to assert that this revolution has had a profound impact on political science and policy. Among other changes, it reshaped how the scholarly community in general and political scientists in particular approached scientific endeavors—the ways they collected, organized, and analyzed data. To explain this shift, we first have to delineate the differences between structured and unstructured data.

For most people, the word "data" means an organized set of values, such as an Excel spreadsheet. These are *structured data*—data that can fit squarely into a table or a relational database, where every row is an observation, every column a variable, and the cells at the intersection of rows and columns contain values. Consequently, the entire database or parts of it can be subject to quantitative analysis. Election results, census records, financial transactions, temperature measurements, and GPS coordinates are all examples of structured data; they transform observable phenomena into measurable and legible forms. Data of this kind, however, are but a fraction of the total amount of data in the world. According to different estimates,[3] 80 to 95 percent of existing data are *unstructured data,* that is, data that cannot fit snugly into rows and columns. Unstructured data may take the form of text, audio, video, or any other observable manifestation. The content of a political speech, the video recording of that speech, the blog post commenting on the video, and the academic article analyzing the post are all unstructured.[4]

---

[2]See Manuel Castells, *The Rise of the Network Society,* 2nd ed., vol. 1, *The Information Age: Economy, Society, and Culture* (Malden, MA: Wiley-Blackwell, 2010), 28–33.

[3]"Data, Data Everywhere," *The Economist,* 27 February 2010, accessed at https://www.economist.com/special-report/2010/02/25/data-data-everywhere, 9 April 2020; Drew Robb, "Semi-Structured Data," *Datamation,* 3 July 2017, accessed at https://www.datamation.com/big-data/semi-structured-data.html, 9 April 2020; and Christie Schneider, "The Biggest Data Challenges That You Might Not Even Know You Have," *IBM Watson Blog,* 25 May 2016, accessed at https://www.ibm.com/blogs/watson/2016/05/biggest-data-challenges-might-not-even-know/, 9 April 2020.

[4]On the differences between structured and unstructured data, see Rob Kitchin, *The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences* (Los Angeles: Sage, 2014), 5–6; and

Because of the messy and eclectic nature of unstructured data, attempts to investigate them with conventional statistical methods would often be futile.

To analyze such unruly data in a quantitative way, one needs to impose a structure upon them by coding selected data points as observations and variables—that is, rows and columns. As one data analyst commented, "In analytics there is no such thing as unstructured data, just data that structure has not yet been applied to."[5] For example, researchers interested in the data surrounding events review media reports and other unstructured sources to identify discrete political incidents such as protests, terrorist attacks, or violent clashes. Based on their codebook—a document that specifies and standardizes the project's rules for collecting and reviewing raw data and transforming them into structured data—they have to decide whether an incident meets the inclusion criteria, in which case they would code it as an observation in a structured database. In accordance with the codebook, they would assign values to the attributes of each event—for instance, the date, location, and type of a protest or terrorist attack or the number of participants or casualties.[6] The reverse procedure of turning a structured database into unstructured data is hardly feasible, as data structuring is, by nature, a reductive process that inevitably entails the loss of details and context. Such preparation of data for quantitative investigation is not a novel approach—the quantification of unstructured data has been in existence since the invention of writing.[7] However, the technological developments of the information age enabled the automated processing of massive amounts of raw data rapidly and at a low cost.[8]

Another approach to structuring data is through text-as-data methods, in which scholars use computer software to turn unstructured data into quantitative data, based on such parameters as the occurrence frequency of words in documents or the type of sentiments expressed in them. In text-as-data projects, investigators usually break the text into specific

---

Amir Gandomi and Murtaza Haider, "Beyond the Hype: Big Data Concepts, Methods, and Analytics," *International Journal of Information Management* 35 (April 2015): 137–144, at 138.

[5]Hjalmar Gislason, "There Is No 'Unstructured Data' in Analytics," Medium, 15 July 2017, accessed at https://medium.com/@hjalli/there-is-no-unstructured-data-in-analytics-8c5d06944b23, 9 April 2020.

[6]Philip A. Schrodt, "Event Data in Foreign Policy Analysis," in Laura Neack, Jeanne A.K. Hey, and Patrick Jude Haney, eds., *Foreign Policy Analysis: Continuity and Change in Its Second Generation* (Englewood Cliffs, NJ: Prentice Hall, 1995), 145–166.

[7]Viktor Mayer-Schönberger and Kenneth Cukier, *Big Data: A Revolution That Will Transform How We Live, Work, and Think* (Boston: Houghton Mifflin Harcourt, 2013), chap. 5.

[8]Wei Wang, Ryan Kennedy, David Lazer, and Naren Ramakrishnan, "Growing Pains for Global Monitoring of Societal Events," *Science 353* (September 2016): 1502–1503.

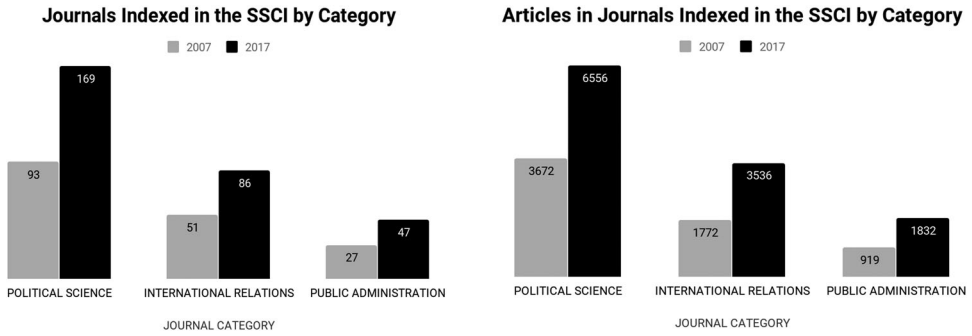
units of analysis, such as articles, chapters, paragraphs, sentences, or single words. Next, they use computer software to quantify these units and analyze the resulting data statistically.[9] For example, in their political theory article, Lisa Blaydes and colleagues examined nearly 10,000 sections from 46 medieval political advice books; they identified recurring terms in the text, combined synonyms into single terms, and nested individual themes (such as a ruler's moral character) under broader ones (in this case, "the art of rulership"). Finally, they tracked the occurrence and weight of these themes in medieval books over time. The statistical analysis shows that while the emphasis on religious issues in political advice texts written by Christians declined over time, Muslim writers continued using religious discourse throughout the period in question. Moreover, the analysis identifies the mass migration of various peoples to the Middle East during the High Middle Ages as a possible influence on Muslim political thought. In particular, the invasion of Turks and Mongols into the Islamic world between the eleventh and thirteenth centuries correlates with a strong emphasis in political advice texts on the desired qualities of the ideal ruler.[10] Such methods enable large-scale conclusions that were previously inaccessible or only accessible through more time-consuming and labor-intensive research. On a smaller scale, we model the text-as-data approach in the next section by counting the occurrence of the term "big data" in political science articles.

The amount of unstructured data in the world not only is massive, but keeps growing perpetually. Consider the case of academic scholarship—the one type of unstructured data that pertains to every political scientist, including purely quantitative ones. We compared the number of political science journals and articles indexed by the Web of Science Social Sciences Citation Index at two points in time—2007 and 2017. We sampled the three most relevant subfields for the purpose of this article: political science, international relations, and public administration. As Figure 1 shows, in one decade, all subfields experienced a dramatic rise. Most notably, political science journals jumped by 82 percent. As for the articles, the increase is even more staggering. Political science and public administration experienced 100 percent and 99 percent jumps, respectively. Simply put, to launch

[9]Justin Grimmer and Brandon M. Stewart, "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts," *Political Analysis* 21 (Summer 2013): 267–297; and John Wilkerson and Andreu Casas, "Large-Scale Computerized Text Analysis in Political Science: Opportunities and Challenges," *Annual Review of Political Science* 20 (2017): 529–544.

[10]Lisa Blaydes, Justin Grimmer, and Alison McQueen, "Mirrors for Princes and Sultans: Advice on the Art of Governance in the Medieval Christian and Islamic Worlds," *Journal of Politics* 80 (October 2018): 1150–1167.

FIGURE 1
Increase in the Number of Political Science Journals and Articles, 2007–2017



Source: Charts created with data from a Web of Science search conducted on 16 February 2019. Left panel: Journals indexed in the Social Sciences Citation Index in 2007 and 2017 by Web of Science subject category. Right panel: Articles in journals indexed in the Social Sciences Citation Index in 2007 and 2017 by Web of Science subject category.

a new study today, political scientists must first review twice as much or more literature than they did a decade ago.

These numbers show that even the most basic scholarly engagement with unstructured data—that is, the process of producing a literature review—already involves struggling with large quantities of material that amass at an inexorable rate. In fact, as one calculation shows, even hypothetical scholars dedicating all their time to reading core publications in their field (and assuming that they are high-speed English readers) would miserably fail to keep up with the literature.[11] The information revolution has turned the search for relevant literature into a desperate race against time and, consequently, a constant source of anxiety for many researchers.[12]

While this information overload is not a recent phenomenon,[13] the term "big data" is relatively new. It first emerged in the information technology industry in the mid-1990s[14] and made its academic debut in a

[11]Michael Billig, *Learn to Write Badly: How to Succeed in the Social Sciences* (Cambridge: Cambridge University Press, 2013), 27–28.
[12]Kenneth Einar Himma, "The Concept of Information Overload: A Preliminary Step in Understanding the Nature of a Harmful Information-Related Condition," *Ethics and Information Technology* 9 (December 2007): 259–272.
[13]David Bawden and Lyn Robinson, "The Dark Side of Information: Overload, Anxiety and Other Paradoxes and Pathologies," *Journal of Information Science* 35 (April 2009): 180–191; and Ann Blair, *Too Much to Know: Managing Scholarly Information before the Modern Age* (New Haven, CT: Yale University Press, 2010), 55.
[14]Steve Lohr, "The Origins of 'Big Data': An Etymological Detective Story," Bits Blog (*New York Times*), 1 February 2013, accessed at https://bits.blogs.nytimes.com/2013/02/01/the-origins-of-big-data-an-etymological-detective-story, 9 April 2020.

1998 computer science paper.[15] In the two decades that followed, it gained popularity rapidly.[16] This proliferation notwithstanding, there is little agreement, both inside academia and outside it, as to what exactly constitutes big data. The most common definitions are based on the "three Vs" framework that Doug Laney presented in an unpublished 2001 report (which, curiously, did not include the words "big data" at all).[17] According to these definitions, data are big if they are high in *volume* (the sheer size of the data set is large), *velocity* (data are produced in or almost in real time), and *variety* (data come in different types and formats and may be structured or unstructured).[18] Over the years, authors have come up with additional Vs, such as *veracity, variability,* and *value,*[19] but volume, velocity, and variety remain the core attributes. Many big data definitions also underline the technological innovations and capabilities and the sophisticated methods required to gather, store, and analyze such data.[20]

These attributes of big data, despite their prevalence in the literature, are more of a general guideline than a definition. Since the introduction of the term, specialists have been attaching to it different and often contrasting meanings. By way of illustration, consider a survey that the University of California, Berkeley, School of Information conducted in 2014. Attempting to clarify ambiguity around the term, the author invited 43 experts in various sectors and industries—from tech and education to food and fashion—to define big data.[21] A majority of respondents (29) highlighted the novel technological and methodological capabilities required to collect, store, organize, and analyze such data.

---

[15]Francis X. Diebold, "On the Origin(s) and Development of the Term 'Big Data,'" (Social Science Research Network, 26 September 2012), accessed at https://papers.ssrn.com/abstract=2152421, 9 April 2020.

[16]A Google Scholar search (21 September 2019) yielded more than 600,000 papers (excluding patents) in which the expression "big data" occurred, while a Web of Science core collection query of papers containing "big data" in any search field produced more than 73,000 results.

[17]Doug Laney, "3D Data Management: Controlling Data Volume, Velocity and Variety" (research note, META Group, Stamford, CT, 6 February 2001).

[18]Erik W. Kuiler, "From Big Data to Knowledge: An Ontological Approach to Big Data Analytics: From Big Data to Knowledge," *Review of Policy Research* 31 (July 2014): 311–318, at 311; and Andrej Zwitter, "Big Data and International Relations," *Ethics & International Affairs* 29 (Winter 2015): 377–389, at 378–379.

[19]Gandomi and Haider, "Beyond the Hype," 139.

[20]Andrea De Mauro, Marco Greco, and Michele Grimaldi, "A Formal Definition of Big Data Based on Its Essential Features," *Library Review* 65 (2016): 122–135, at 125–127; and David Lazer and Jason Radford, "Data Ex Machina: Introduction to Big Data," *Annual Review of Sociology* 43 (2017): 19–39, at 20–21.

[21]Jennifer Dutcher, "What Is Big Data?," Berkeley School of Information Blog, 3 September 2014, accessed at https://web.archive.org/web/20180121162550/, https://datascience.berkeley.edu/what-is-big-data, 9 April 2020.

Many respondents referred to the volume of big data (22) and their cost and potential value—commercial, scientific, or intellectual (17). Fewer participants mentioned data variety, complexity, or messiness (18) and data velocity (11).[22] Alongside these more standard conceptualizations, many responses failed to offer a clear definition of big data and, instead, relied on vague language—"storytelling," "challenges and opportunities," "a cultural shift," "a rhetorical device," and even "anything related to data analytics or visualization" were among the answers given. As these answers suggest, many definitions of big data are ambiguous, and therefore lack the qualities that make good concepts in the social sciences.[23]

Further, the way data analysts in the business sector or the STEM disciplines (science, technology, engineering, and mathematics) define big data is not entirely compatible with the way social scientists should view it. For a start, the big data sets that social scientists analyze are large in terms of their disciplines, but they are often considerably smaller in volume and less resource-intensive than what many STEM scientists or data analysts refer to as big data.[24] For these reasons, social science researchers need their own definition and use of big data appropriate to their field.

The two common features of big data that are most relevant to political science are the variety of data and the technological means required to extract, organize, and analyze them. For political scientists, then, big data should be defined as structured and unstructured data of different provenances, which they can access, review, and process with the help of digital technology. Of utmost importance here is the variety of big data: today, more than 98 percent of stored data in the world are digital (the rest being stored on paper, film, vinyl, and other analog media).[25] As most of these data are unstructured, many political questions would greatly benefit from the wealth of knowledge that the information revolution has made available to us in the form of organizational and personal, official and unofficial, digitized or digitally born documents, photos, video and audio recordings, websites, art, video games, and so on.[26] Yet scholars in the discipline have

[22]Replication materials for our content analysis of the University of California, Berkeley, survey can be accessed at https://github.com/jonathan-grossman/Political-Science-and-Big-Data.

[23]See John Gerring, *Social Science Methodology: A Unified Framework* (Cambridge: Cambridge University Press, 2012), 116–131.

[24]Lev Manovich, "Trending: The Promises and the Challenges of Big Social Data," in Matthew K. Gold, ed., *Debates in the Digital Humanities* (Minneapolis: University of Minnesota Press, 2012), 461; and Alex Street, Thomas A. Murray, John Blitzer, and Rajan S. Patel, "Estimating Voter Registration Deadline Effects with Web Search Data," *Political Analysis* 23 (Spring 2015): 225–241, at 238.

[25]Mayer-Schönberger and Cukier, *Big Data,* 9.

[26]See Sharan B. Merriam and Elizabeth J. Tisdell, *Qualitative Research: A Guide to Design and Implementation,* 4th ed. (San Francisco: Jossey-Bass, 2016), chap. 7.

largely ignored unstructured big data so far (although, as will be shown later, this situation is gradually changing). As we discuss in the next section, while political scientists have been reluctant to join the conversation about big data, many of them have actually been using structured big data without referring to them as such.

BIG DATA AND POLITICAL SCIENCE

Political scientists witnessed the transformation of big data into a "buzzword,"[27] a "catchall term,"[28] and even a "meme."[29] However, a content analysis of the 133 political science articles about big data that were indexed in the Web of Science Social Sciences Citation Index[30] (as of 1 January 2019) suggests that the discussion of big data in the discipline is still in its infancy.[31] The earliest article on big data in a political science journal was published in 2012.[32] As Figure 2 shows, since 2014, there has been a marked increase, although not a dramatic or steady one, in the number of such articles. To a large degree, this discussion is currently confined to specialized journals (such as those focusing on intelligence and technology, for example *Intelligence and National Security* or *Policy & Internet*), methodological contributions, and special issues and symposia—30 out of the 133 articles were published in the same issue with at least two other articles from the database; 41 articles share the same volume with at least two other articles. For example, eight articles were published in the same issue and volume of *Review of Policy Research* in 2014, and four articles were published in one volume of *Politics and Governance* in 2018.

Not all articles whose topic was categorized as "big data" by the Web of Science discuss big data in any meaningful way. At least in some of them, "big data" serves as a buzzword (or, perhaps, as clickbait) to catch the attention of readers or journal editors without much actual discussion of

[27]Christopher Eldridge, Christopher Hobbs, and Matthew Moran, "Fusing Algorithms and Analysts: Open-Source Intelligence in the Age of 'Big Data,'" *Intelligence and National Security* 33 (2018): 391–406, at 393; and Damien Van Puyvelde, Stephen Coulthart, and M. Shahriar Hossain, "Beyond the Buzzword: Big Data and National Security Decision-Making," *International Affairs* 93 (November 2017): 1397–1416.

[28]Connie L. McNeely and Jong-on Hahm, "The Big (Data) Bang: Policy, Prospects, and Challenges," *Review of Policy Research* 31 (July 2014): 304–310, at 304.

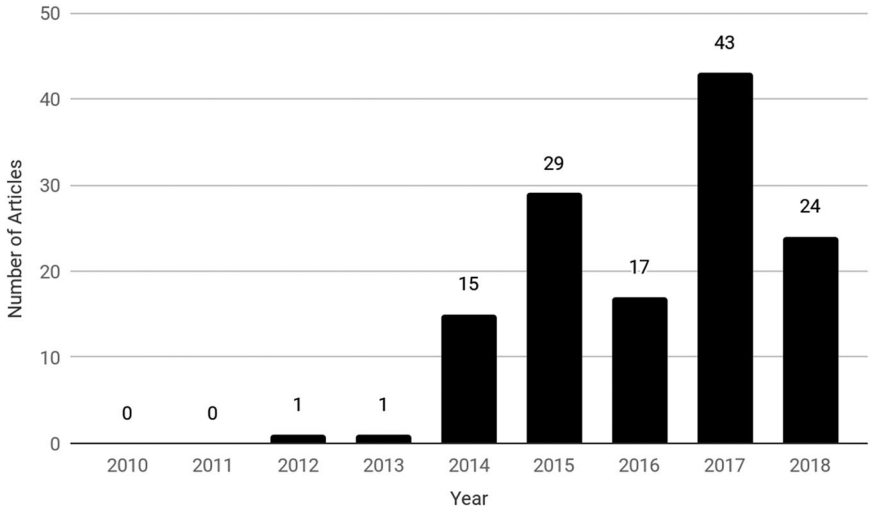
[29]Jan Youtie, Alan L. Porter, and Ying Huang, "Early Social Science Research about Big Data," *Science and Public Policy* 44 (February 2016): 65–74, at 65.

[30]The Social Sciences Citation Index is part of Clarivate Analytics' Web of Science academic index (https://webofknowledge.com).

[31]See the online appendix for the methodology and results of the content analysis, at https://github.com/jonathan-grossman/Political-Science-and-Big-Data/blob/master/Online-Appendix.pdf. Replication materials can be accessed at https://github.com/jonathan-grossman/Political-Science-and-Big-Data.

[32]The full list of articles and all the variables in this analysis can be accessed at https://github.com/jonathan-grossman/Political-Science-and-Big-Data.

FIGURE 2
Political Science Articles about Big Data, by Year



Source: Chart created with data from a Web of Science search conducted on 1 January 2019 with the following specifications: DOCUMENT TYPE = "Articles"; WEB OF SCIENCE CATEGORIES = "Political Science," "International Relations," and "Public Administration."

this concept. Thus, in 16 articles, the expression "big data" only occurs in the article title, abstract, list of references, or list of keywords, but not in the content of the article. Fifteen articles mention big data only once in the narrative. In 35 articles, the term occurs three times or fewer. Only 52 articles—less than 40 percent of the total articles—offer a definition of big data that is workable to some extent. Further, the discussion of big data in political science is highly fragmented—less than half of the articles in the database cite each other.[33]

As these content and citation analyses indicate, political scientists have only recently begun to integrate the term big data into their scholarship. Discussions of big data have yet to occupy a niche in mainstream political science publications. Nonetheless, the tables of contents of leading quantitatively oriented political science journals confirm that political scientists have been using digital tools to structure and analyze large data repositories for some time now, albeit without explicitly referring to their data as "big" or taking part in recent debates about big data. Established research fields such as electoral studies, public opinion research, event data analysis, and comparative policy analysis are only a few examples of

[33]For a citation network analysis of the 133 articles, see the online appendix at https://github.com/jonathan-grossman/Political-Science-and-Big-Data/blob/master/Online-Appendix.pdf.

the prevalence of big data sources in political science. The proliferation of big databases in these fields has been a gradual process rather than a revolution—the size and complexity of data that political scientists analyze have been increasing for decades, since before the information age, in keeping with technological advancements in storage and processing capabilities.[34]

And yet, political scientists have been reluctant to identify their sources as big data when there was no analytical justification for using this term. For instance, Gary King and colleagues' 2013 censorship study on China features an automated collection, coding, and analysis of 11,382,221 social media posts from 1,382 Chinese websites. The researchers sampled each post several times to track changes to it. As they found out, the Chinese government was likely to censor posts mentioning collective action even when the posts were supportive of the government—and less likely to delete posts that did not mention collective action even when these posts were highly critical of the government.[35] Adopting a similar approach to write about civil conflicts, Nils Metternich and his colleagues created "a dataset of over two million machine-coded daily events" extracted from more than 75 news sources. By applying network analysis and game theory methods to this database, they created a model to predict antigovernment conflicts in Thailand.[36] Although the term "big data" does not occur even once in these and similar papers, they would qualify as studies of structured big data according to common definitions. When studying the use of big data in political science, then, we should consider that while papers rarely mention the term "big data," many of them use this kind of data in practice.

Political scientists who engage in such research endeavors are predominantly working within the quantitative tradition. They detect trends and patterns in structured databases and employ statistical methods to differentiate signal from noise, identify correlations, and generate predictions and policy recommendations. In many instances, this use of big data for research is necessary. No one would expect demographers to conduct their own national census or students of voter turnout rates to be

[34]Michael D. Ward, Nils W. Metternich, Cassy L. Dorff, Max Gallop, Florian M. Hollenbach, Anna Schultz, and Simon Weschle, "Learning from the Past and Stepping into the Future: Toward a New Generation of Conflict Prediction," *International Studies Review* 15 (December 2013): 473–490.

[35]Gary King, Jennifer Pan, and Margaret E. Roberts, "How Censorship in China Allows Government Criticism but Silences Collective Expression," *American Political Science Review* 107 (May 2013): 326–343.

[36]Nils W. Metternich, Cassy Dorff, Max Gallop, Simon Weschle, and Michael D. Ward, "Antigovernment Networks in Civil Conflicts: How Network Structures Affect Conflictual Behavior," *American Journal of Political Science* 57 (October 2013): 892–911.

present at voting centers throughout Election Day to record the attendance of each and every voter. Structured data are better equipped than unstructured data to address some problems, just as quantitative methods can answer certain questions better than qualitative ones.[37] When data are objective and fully represent the phenomenon in question, such analyses can be highly effective.

Unfortunately, this is not always the case. At the outset, big databases seemed to offer a solution to the challenges associated with quantitative data analysis: why use a sample when we can study the entire population? However, this optimism was short-lived. An ever-growing body of literature indicates that big data are not as representative or free of bias, manipulation, and interpretation as we would like to believe.[38] For instance, while Facebook has been the most popular online social platform across generations, most social scientists work with Twitter data because the latter are considerably easier to collect and structure than Facebook posts.[39] In addition, 10 million tweets may represent 10 million people—but only those people who have an internet connection and Twitter account and actively tweet. As a case in point, one study of the 2009 German parliamentary election claimed that the proportion of tweets mentioning a political party before the election could predict the number of votes that this party would gain.[40] However, as a later study revealed, that analysis was correct only because the researchers had arbitrarily removed from the results the small but much-discussed Pirate Party, which was mentioned in about a third of the tweets but won only 2.1 percent of the votes.[41]

In a similar vein, an ever-updating stream of news media articles may capture hundreds or thousands of events each day, but it will fail to include unreported events or events reported by news outlets that

[37]See Gary Goertz and James Mahoney, *A Tale of Two Cultures: Qualitative and Quantitative Research in the Social Sciences* (Princeton, NJ: Princeton University Press, 2012), 2–3.

[38]danah boyd and Kate Crawford, "Critical Questions for Big Data," *Information, Communication & Society* 15 (June 2012): 662–679, at 666–71; Lazer and Radford, "Data Ex Machina," 28–31; and Matthew J. Salganik, *Bit by Bit: Social Research in the Digital Age* (Princeton, NJ: Princeton University Press, 2018), 17–41.

[39]Ines Mergel, "Building Holistic Evidence for Social Media Impact," *Public Administration Review* 77 (July–August 2017): 489–495, at 490.

[40]Andranik Tumasjan, Timm O. Sprenger, Philipp G. Sandner, and Isabell M. Welpe, "Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment," Fourth International AAAI Conference on Weblogs and Social Media, May 2010, accessed at https://www.aaai.org/ocs/index.php/ICWSM/ICWSM10/paper/view/1441, 9 April 2020.

[41]Andreas Jungherr, Pascal Jürgens, and Harald Schoen. "Why the Pirate Party Won the German Election of 2009 or The Trouble with Predictions: A Response to Tumasjan, A., Sprenger, T. O., Sander, P. G., & Welpe, I. M. 'Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment,'" *Social Science Computer Review* 30 (May 2012): 229–234.

are not represented in the specific index that the researchers consulted. Nils Weidmann compared coverage of insurgent attacks in Afghanistan in a military database with coverage of such events in a media-based data set. Weidmann concludes that an event was more likely to be reported in the media when the number of casualties was relatively high, the event took place relatively close to a city or a town, and the area had cellphone coverage.[42] However, official data are not necessarily more reliable than media reports. Governments, corporations, and other entities may restrict access to some or all of their data, manipulate them, or fail to collect data efficiently. According to a recent study of officer-involved fatal shootings in California and Texas, open-source records are in fact more accurate and comprehensive than official databases; the former report 25 to 50 percent more such cases than state or federal databases, which suffer from underreporting and classification errors.[43]

Another problem arises when researchers forgo asking questions altogether. In the age of big data, some analysts have challenged the scientific method of raising questions, formulating hypotheses, and using theory. Instead, they claim that since we have so much data at our disposal, these data should be allowed to "speak for themselves," revealing patterns and correlations that would tell us what is important without knowing in advance what we were hoping to find.[44] While this inductive approach is mainly characteristic of analysts in the business sector, such as ones searching for promising investment opportunities, a growing number of data and STEM scientists now purport to solve social problems with the power of big data. Ascribing little importance to unquantifiable details, these pundits often lack not only training in social science methods but also substantial knowledge of the unique social, cultural, and political circumstances of the issue under study. Instead, they confidently interpret the available structured data according to patterns and correlations that they find in them,[45] while paying little attention, if at all, to social theory. As one review of the literature on Twitter and political action shows, even researchers who

[42]Nils B. Weidmann, "A Closer Look at Reporting Bias in Conflict Event Data," *American Journal of Political Science* 60 (January 2016): 206–218.
[43]Howard E. Williams, Scott W. Bowman, and Jordan Taylor Jung, "The Limitations of Government Databases for Analyzing Fatal Officer-Involved Shootings in the United States," *Criminal Justice Policy Review* 30 (March 2019): 201–222, at 216.
[44]Rob Kitchin, "Big Data, New Epistemologies and Paradigm Shifts," *Big Data & Society* 1 (April–June 2014): 1–12, at 3–4; and Mayer-Schönberger and Cukier, *Big Data,* chap. 4.
[45]Kitchin, "Big Data," 3–5.

include theoretical discussion in their papers do not always ground the results of their big data analysis in any theory.[46]

Not surprisingly, such data-centered works tend to be superficial, if not erroneous. Their authors, devoid of regional or historical context, often infer causation from spurious correlations.[47] Without a theoretical, conceptual, and contextual framework and clearly defined hypotheses to guide data collection and analysis, any datum in the vast ocean of big data may appear relevant and important.[48] The website Spurious Correlations records such instances of statistically significant yet absurd findings. For example, there is a 95.86 percent correlation between the number of civil engineering doctorates that are awarded each year and the consumption of mozzarella cheese in the United States.[49] Researchers who downplay the importance of context (and, sometimes, of common sense) and do not know their subject matter well may conclude that the fact that two variables are correlated necessarily means that one caused the other.[50]

Some scholars maintained that big data were likely to encompass not only explanatory powers but also predictive ones. However, as it soon turned out, our ability to accurately forecast outcomes such as terrorist attacks and election results has yet to improve significantly.[51] For instance, the vast majority of predictive models for the 2016 U.S. presidential election, which forecast that Hillary Clinton would defeat Donald Trump, proved to be wrong.[52] Political events are often highly complex. They involve a large number of agents who interact with one another in unexpected and often irrational ways and conceal their motives and strategies. Neither the laws of nature nor a finite set of actors, which render large volumes of structured data highly effective in forecasting the weather or determining the best move in a game of

[46]Peter Cihon and Taha Yasseri, "A Biased Review of Biases in Twitter Studies on Political Collective Action," *Frontiers in Physics* 4 (August 2016), https://doi.org/10.3389/fphy.2016.00034.
[47]Gary Smith, "The Exaggerated Promise of So-Called Unbiased Data Mining," *Wired,* 11 January 2019, accessed at https://www.wired.com/story/the-exaggerated-promise-of-data-mining, 9 April 2020.
[48]Gary King, Robert O. Keohane, and Sidney Verba, *Designing Social Inquiry: Scientific Inference in Qualitative Research* (Princeton, NJ: Princeton University Press, 1994), 44.
[49]Spurious Correlations, accessed at http://tylervigen.com/spurious-correlations, 9 April 2020.
[50]Nate Silver, *The Signal and the Noise: Why So Many Predictions Fail—but Some Don't* (New York: Penguin, 2015), 253.
[51]Lars-Erik Cederman and Nils B. Weidmann, "Predicting Armed Conflict: Time to Adjust Our Expectations?," *Science* 355 (February 2017): 474–476.
[52]James Markarian, "What the Election Taught Us about Predictive Analytics," *Forbes,* 8 February 2017, accessed at https://www.forbes.com/sites/forbestechcouncil/2017/02/08/what-the-election-taught-us-about-predictive-analytics, 9 April 2020.

chess, exist in most political interactions.[53] Further, the prediction of political events is often based on flawed data, especially when computer programs automatically extract and code qualitative sources such as news reports without proper quality control by experts. Consequently, biases and errors in unstructured sources might be integrated into structured databases and distort the results of statistical analyses.[54]

While political science is not immune to these practices, the discipline as a whole is actually moving toward a more theory-guided and context-informed analysis of big data. As part of this trend, researchers have been paying increasing attention to the content of unstructured big data sources. Most of these studies employ text-as-data approaches that turn the unstructured sources into highly structured data sets ready for statistical analysis at the cost of historical detail. One recent example is Azusa Katagiri and Eric Min's article about the effectiveness of private and public diplomatic signals, which uses the 1958–1963 Berlin Crisis as a case study. To test their theoretical hypotheses that actions speak louder than words and that private signals are both less noisy and more effective than public ones, Katagiri and Min processed more than 18,000 U.S. diplomatic documents. By building and training statistical models adapted to the unique circumstances of the Berlin Crisis, the authors identified documents discussing Soviet threats to use force and translated their narrative into quantitative data. Based on the results of their statistical analysis, Katagiri and Min argue that Soviet material actions were considerably more effective than private signals in influencing how U.S. policymaking elites perceived Soviet threats. However, private signals of Soviet resolve were still more compelling than public ones in convincing U.S. officials that the Soviet Union was willing to use force to achieve its objectives in Berlin. Thus, the quantitative analysis supports the researchers' hypotheses.[55]

Quantitative studies of this type capitalize on the potential of unstructured big data, but the insight derived from them only applies to general trends and therefore should not replace a contextualized in-depth reading of historical documents, as the authors of one such work frankly admit.[56]

---

[53]Cederman and Weidmann, "Predicting Armed Conflict"; Keith Dowding, "So Much to Say: Response to Commentators," *Political Studies Review* 15 (May 2017): 217–230, at 226; and Silver, *The Signal and the Noise*, 2.

[54]Cederman and Weidmann, "Predicting Armed Conflict," 475; Idean Salehyan, "Best Practices in the Collection of Conflict Data," *Journal of Peace Research* 52 (January 2015): 105–109, at 108; and Wang et al., "Growing Pains."

[55]Azusa Katagiri and Eric Min, "The Credibility of Public and Private Signals: A Document-Based Approach," *American Political Science Review* 113 (February 2019): 156–172.

[56]Blaydes, Grimmer, and McQueen, "Mirrors for Princes and Sultans," 1153–1154.

Many other scholars supplement their quantitative analysis of structured big data with a thorough analysis of traditional sources such as oral interviews.[57] For instance, in their study of the influence of election campaigns on voter turnout rates, Ryan Enos and Anthony Fowler quantitatively investigate a big data repository of voter participation but reinforce their statistical findings with insight from oral and email conversations with campaign managers and strategists.[58] Still other researchers analyze a random sample of unstructured big data. Greg Distelhorst and Diana Fu, in their article about citizenship in authoritarian regimes, employ this method. After compiling a database of more than 8,000 online appeals by Chinese citizens to local authorities, the authors randomly selected a subsample of 500 documents for close reading and interpretive analysis.[59] Such applications of unstructured "small" data have proven useful for contextualizing the statistical investigation of structured big data sets and for enhancing sampling validity. At the same time, they do not tap into the full potential of unstructured big data.

THE PROMISE OF UNSTRUCTURED BIG DATA

Political research that is primarily based on the scrutiny of mostly unstructured texts has become more rigorous, systematic, and transparent than ever. Taking advantage of the rich detail found in such traditional sources as archival documents, oral interviews, and memoirs, as well as of the investigators' theoretical, historical, and regional expertise,[60] scholars have been able to tackle meaningful questions about political behavior in innovative ways and to publish their findings in mainstream political science journals. This direct contact between political scientists and historical data is a significant break from past practices—not long ago, secondary literature written by historians was the main source of information for political scientists working on historical case studies, who rarely consulted primary data.[61] Today, many of them collect their own historical evidence.

---

[57]Tobias Bornakke and Brian L Due, "Big–Thick Blending: A Method for Mixing Analytical Insights from Big and Thick Data Sources," *Big Data & Society* 5 (January–June 2018), https://doi.org/10.1177/2053951718765026.

[58]Ryan D. Enos and Anthony Fowler, "Aggregate Effects of Large-Scale Campaigns on Voter Turnout," *Political Science Research and Methods* 6 (October 2018): 733–751.

[59]Greg Distelhorst and Diana Fu, "Performing Authoritarian Citizenship: Public Transcripts in China," *Perspectives on Politics* 17 (March 2019): 106–121, at 110.

[60]Marcus Kreuzer, "The Structure of Description: Evaluating Descriptive Inferences and Conceptualizations," *Perspectives on Politics* 17 (March 2019): 122–139, at 125.

[61]Ian S. Lustick, "History, Historiography, and Political Science: Multiple Historical Records and the Problem of Selection Bias," *American Political Science Review* 90 (September 1996): 605–618; and Theda Skocpol, "Emerging Agendas and Recurrent Strategies in Historical Sociology," in Theda Skocpol, ed., *Vision and Method in Historical Sociology* (Cambridge: Cambridge University Press, 1984), 356–385, at 382.

Two recent articles in this journal illustrate both the potential and challenges of the dramatic increase in unstructured data for advancing political science research. In his 2018 article, Christopher Lamb shows how careful organization and examination of a large number of unstructured sources enable political scientists to correct common misconceptions in the literature and, as a result, develop better theories. Lamb makes this point with regard to the *Mayaguez* incident, a highly complex and contested historical event in which U.S. and Cambodian forces clashed over the seizure of a U.S. cargo ship by the Khmer Rouge in May 1975. Lamb demonstrates that even though a staggering number of sources on the incident have become available to researchers, there is still no consensus as to what really happened there and how the United States managed the crisis.[62] Moreover, Lamb argues that the rise in the number of sources on the incident did not contribute to our understanding of it: scholars used the newly declassified information to describe the crisis—the qualitative equivalent of letting the data "speak for themselves"—rather than explaining it through the testing of alternative hypotheses.[63]

Lamb uses the available evidence on the *Mayaguez* incident, including records that have only recently been declassified, to build a painstakingly detailed chronology of the events in Cambodia and Washington on 12–15 May 1975. He then draws on this timeline to criticize key arguments in the extant literature on the crisis. For instance, he cites a "widely overlooked interview" with *Veteran* magazine, in which former Secretary of Defense James R. Schlesinger admitted that he had purposely delayed a White House order to sink the Cambodian fishing boat carrying the U.S. crew of the *Mayaguez*. As Schlesinger retroactively explained, "it seemed to me it would destroy our own purposes to sink a ship which would have killed the Americans we were trying to save."[64] Other scholars failed to notice this crucial piece of evidence that did not exist in conventional sources such as government records. This oversight is indicative of a typical problem in the information era—unstructured data are not only growing, they are also scattered across many places and are often hard to locate. Finding a needle in a haystack is only possible if we can find the haystack.

Another example of how the growing availability of unstructured data, coupled with a systematic approach to analyzing these data, can result in more accurate political science research is Stefano Recchia's 2015 article.

---

[62]Christopher J. Lamb, "The *Mayaguez* Crisis: Correcting 30 Years of Scholarship," *Political Science Quarterly* 133 (Spring 2018): 35–76.
[63]Lamb, "The *Mayaguez* Crisis," 74–75.
[64]Lamb, "The *Mayaguez* Crisis," 53.

This article offers a new reading of France's March 2003 thwarting of a U.S.-British plan to introduce a resolution draft at the United Nations Security Council, which would have authorized the use of force against Iraq. Relying on declassified documents and oral interviews, Recchia constructs a comprehensive timeline of the events that led to President Jacques Chirac's public threat to veto the proposed resolution. Recchia uses these rich data to show, by means of counterfactual analysis, that France did not categorically oppose the use of force as many specialists argue. In fact, President Chirac was likely to support or at least abstain on a similar resolution draft that would have allowed Iraq more time to comply with the Security Council's demands regarding the control and surrendering of certain weapons before resorting to armed force. The George W. Bush administration, which was already preparing the invasion of Iraq, was unwilling to postpone it. Instead, Washington and its allies decided to carry out the attack without United Nations support.[65]

Lamb's and Recchia's works are examples of how the meticulous collection and sequencing of unstructured data—and particularly the construction of detailed timelines consisting of distinct interactions between actors—can revolutionize research. In the age of big data, as more and more information on such political events as the 1975 *Mayaguez* incident and the 2003 invasion of Iraq becomes available, it is increasingly challenging to keep track of the details of these events for inferential purposes. A growing number of political scientists have attempted to address these challenges by integrating methodological and conceptual contributions from such disciplines as economics, psychology, history, and law into their research. As we propose in the next section, the political science literature on historical institutionalism and process tracing offers a productive methodological framework for working with unstructured big data.

HISTORICAL INSTITUTIONALISM, PROCESS TRACING, AND THE PROMISE OF UNSTRUCTURED BIG DATA

Historical institutionalism adds dynamism to political science research by introducing time as a focal element, while process tracing allows researchers to systematically arbitrate between rival hypotheses to find causal mechanisms. Both historical institutionalism and process tracing emphasize evidence. Unstructured big data can provide such evidence—for example, in the form of news articles from numerous origins, the full text of or excerpts from multiple books and accounts, or diplomatic correspondence from the

[65]Stefano Recchia, "Did Chirac Say 'Non'? Revisiting UN Diplomacy on Iraq, 2002–03," *Political Science Quarterly* 130 (Winter 2015): 625–654.

archives of foreign ministries. By harnessing the potential of such data, historical institutionalists and process tracers (who are sometimes the same researchers) can dramatically enhance the accuracy and scope of their work and magnify the impact of their scholarship.

Historical Institutionalism and Big Data

Since the 1990s, scholars working within the research tradition of historical institutionalism have been underscoring the significance of time for political analysis. Historical institutionalism centers on identifying temporal sequences—that is, the timing and order of events—as critical for making causal claims and testing theories regarding change and continuity in the life of institutions (rules, procedures, norms, and organizations). To further understand the influence of temporality on political processes, historical institutionalists have introduced into the discussion of formal and informal institutions such concepts as path dependence, positive and negative feedback, and increasing and diminishing returns.[66] A central quest of many historical institutionalist research designs is to identify critical junctures—short periods of time (relative to the length of the institutional path) during which structural constraints on political actors are relaxed so that "the range of plausible choices open to powerful political actors expands substantially and the consequences of their decisions for the outcome of interest are potentially much more momentous."[67]

Unlike the previous generation of historical institutionalists, who mostly relied on the scholarship of historians,[68] many current researchers

---

[66]Giovanni Capoccia and R. Daniel Kelemen, "The Study of Critical Junctures: Theory, Narrative, and Counterfactuals in Historical Institutionalism," *World Politics* 59 (April 2007): 341–369; Orfeo Fioretos, Tulia Gabriela Falleti, and Adam D. Sheingate, eds., *The Oxford Handbook of Historical Institutionalism* (Oxford: Oxford University Press, 2016); Orfeo Fioretos, ed., *International Politics and Institutions in Time* (New York: Oxford University Press, 2017); B. Guy Peters, *Institutional Theory in Political Science: The New Institutionalism,* 3rd ed. (New York: Continuum, 2012), chap. 4; Paul Pierson, "Increasing Returns, Path Dependence, and the Study of Politics," *American Political Science Review* 94 (June 2000): 251–267; Paul Pierson, *Politics in Time: History, Institutions, and Social Analysis* (Princeton, NJ: Princeton University Press, 2004); Paul Pierson and Theda Skocpol, "Historical Institutionalism in Contemporary Political Science," in Ira Katznelson and Helen V. Milner, eds., *Political Science: State of the Discipline* (New York: W.W. Norton, 2002), 693–721; Thomas Rixen, Lora Anne Viola, and Michael Zürn, eds., *Historical Institutionalism and International Relations: Explaining Institutional Development in World Politics* (Oxford: Oxford University Press, 2016); and Sven Steinmo, Kathleen Ann Thelen, and Frank Longstreth, eds., *Structuring Politics: Historical Institutionalism in Comparative Analysis* (Cambridge: Cambridge University Press, 1992).

[67]Capoccia and Kelemen, "The Study of Critical Junctures," 343.

[68]For example, Ruth Berins Collier and David Collier, *Shaping the Political Arena: Critical Junctures, the Labor Movement, and Regime Dynamics in Latin America* (Notre Dame, IN: University of Notre Dame Press, 2002); and Peter A. Hall, "The Movement from Keynesianism to Monetarism: Institutional Analysis and British Economic Policy in the 1970s," in Sven Steinmo, Kathleen Ann Thelen, and Frank

are working closely with primary sources, which render their work more comprehensive and less prone to selection bias.[69] In both cases, researchers use sources that are, for the most part, unstructured. For example, a 2014 article by Ramazan Kilinç draws on the historical institutionalist literature on critical junctures and self-reinforcing sequences to explain democratic consolidation using a Turkish case study. Kilinç relies on various English and Turkish unstructured sources—research literature, news articles, party programs, political speeches, and legal records—to define the two-year period that followed the 1997 military intervention as a critical juncture for democratic consolidation in the country; during that period, important Islamist groups, whose political participation was curtailed by the military, adopted a liberal democratic stance in a successful attempt to gain access to power. Because these groups chose to pursue a democratic path at that specific point in time, they became path dependent (that is, committed to the democratic discourse that they defended and to the constituencies that supported them because of this democratic position). This commitment ignited a self-reinforcing sequence that resulted in democratic consolidation: first, the Islamist Justice and Development (AK) Party won the 2002 parliamentary elections as the flagbearer of democratization, market economy, social reforms, and EU membership. Subsequently, the AK Party carried out measures that limited the military's ability to intervene again in the political system and broadened the party's own base of support.[70]

A salient characteristic of historical institutionalism is that it seeks to explain real-world cases rather than predict general political behavior.[71] While the use of structured big data to generate predictions has been problematic in many instances, using unstructured big data sources to answer historical riddles is more feasible and—when done systematically, as evidenced in Kilinç's analysis of Turkish politics—less fallible. Real-world cases, after all, need real-world empirical evidence. That is not to say that historical institutionalism is merely descriptive and divorced from theory. Historical institutionalist explanations can be generalized

---

Longstreth, eds., *Structuring Politics: Historical Institutionalism in Comparative Analysis* (Cambridge: Cambridge University Press, 1992), 90–113.

[69]Orfeo Fioretos, "Institutions and Time in International Relations," in Orfeo Fioretos, ed., *International Politics and Institutions in Time* (New York: Oxford University Press, 2017), 25.

[70]Ramazan Kilinç, "Critical Junctures, Catalysts, and Democratic Consolidation in Turkey," *Political Science Quarterly* 129 (Summer 2014): 293–318.

[71]Sven Steinmo, "Historical Institutionalism," in Donatella Della Porta and Michael Keating, eds., *Approaches and Methodologies in the Social Sciences: A Pluralist Perspective* (Cambridge: Cambridge University Press, 2008), 134.

and elaborated into theories that may, in turn, be employed to predict institutional behavior. Thus, drawing on his temporal analysis, Kilinç also makes the theoretical argument that, with respect to democratic consolidation, "the timing of events facilitates or impedes the causal processes that structural conditions lead."[72]

In addition, unlike many other approaches, historical institutionalism views political phenomena as structured in time and space and therefore deeply embedded in historical context.[73] One piece of good advice for historically oriented political scientists who wish to understand the context of the problem they attempt to address is not to limit their purview to analyzing a handful of key primary documents. Rather, as Alexander George and Andrew Bennett suggest, they should survey a large array of media accounts from the period under study to understand not only the facts directly related to their question but also the climate in which decisions and actions were taken and the information that was available to actors, policymakers, and the public at the time.[74] As Deborah Larson remarks, "Journalistic analyses and interpretations of speeches provide a code book by which to decipher the meaning of a document."[75] While viewing a single newspaper over time may result in the incorporation into the study of biases held by specific journalists and editors, using a multitude of sources can lead to a more balanced depiction of the zeitgeist. Browsing unstructured big data repositories such as digital collections of newspapers or, with respect to recent years, online news archives and social media posts can offer this balance.

The latter point illuminates the two features that make unstructured big data particularly appealing for historical institutionalists: they are laden with voluminous and variegated historical narrative, and they are often digitally searchable. Historical institutionalists must identify the point at which an institution started following a particular path as well as when it stopped following it.[76] For this purpose, they must have enough information on the institution and the different actors that

[72]Kilinç, "Critical Junctures, Catalysts, and Democratic Consolidation in Turkey," 295.
[73]Peter A. Hall, "Politics as a Process Structured in Space and Time," in Orfeo Fioretos, Tulia G. Falleti, and Adam Sheingate, eds., *The Oxford Handbook of Historical Institutionalism* (Oxford: Oxford University Press, 2016), 31–50.
[74]Alexander L. George and Andrew Bennett, *Case Studies and Theory Development in the Social Sciences* (Cambridge, MA: MIT Press, 2005), 97.
[75]Deborah Welch Larson, "Sources and Methods in Cold War History: The Need for a New Theory-Based Archival Approach," in Colin Elman and Miriam Fendius Elman, eds., *Bridges and Boundaries: Historians, Political Scientists, and the Study of International Relations* (Cambridge, MA: MIT Press, 2001), 347.
[76]Pierson, *Politics in Time,* esp. 44–46.

influenced and were influenced by it over what is often a very long stretch of time.[77] When formulating a critical juncture argument, they equally need to engage in massive pursuit for historical evidence in order to map the main actors in the juncture and their interactions, build and test counterfactual arguments, and understand the legacy of the juncture.[78] To that end, identifying the dates and locations of events and the participants in them with a maximal degree of certainty is vital. Sources, however, often suffer from description bias, providing inconsistent or inaccurate dates and names. Thus, the more independent sources we have on an item and the more diversified and reliable they are, the better our chances to eliminate errors and discrepancies and establish an accurate time frame.[79]

As the information age matures, such digital primary and secondary sources accumulate at an unprecedented rate. Archives, libraries, publishers, museums, media outlets, government agencies, for-profit and non-profit organizations, businesses and companies, and even private amateurs and aficionados are increasingly digitizing their collections and granting direct access to them, for free or for a fee. New digital archives appear every day, while existing ones are growing bigger and bigger as digitally born material is being published and analog material is being scanned. In its strategic plan for the years 2018–2022, the U.S. National Archives and Records Administration committed to scan and make available to the public half a billion pages by 2024.[80] Other countries are increasingly committed to the digitization of records. The government of New Zealand, for example, launched the digitization project of its national archives in 2017.[81] The United Kingdom archives government information that was published online, including tweets by official government organizations.[82] Private archiving projects have also become

[77]Kathleen Thelen and James Conran, "Institutional Change," in Orfeo Fioretos, Tulia G. Falleti, and Adam Sheingate, eds., *The Oxford Handbook of Historical Institutionalism* (Oxford: Oxford University Press, 2016), 57.

[78]Giovanni Capoccia, "Critical Junctures and Institutional Change," in James Mahoney and Kathleen Thelen, eds., *Advances in Comparative-Historical Analysis* (Cambridge: Cambridge University Press, 2015), 169–173.

[79]Marc Trachtenberg, *The Craft of International History: A Guide to Method* (Princeton, NJ: Princeton University Press, 2006), 147–148.

[80]National Archives and Records Administration, "Strategic Plan 2018–2022," February 2018, accessed at https://www.archives.gov/about/plans-reports/strategic-plan/strategic-plan-2018-2022, 14 August 2019.

[81]Archives New Zealand, "What's Been Digitised," 31 May 2019, accessed at https://archives.govt.nz/search-the-archive/what-we-have/whats-been-digitised, 9 April 2020.

[82]U.K. Government Web Archive, Twitter Archive, accessed at https://webarchive.nationalarchives.gov.uk/twitter/, 9 April 2020.

more common and sophisticated. Today, for example, invaluable information on the operations and members of British commando forces in World War II, including detailed timelines, maps, photos, letters, and booklets, can be found on personal websites founded and maintained by the veterans' families.[83]

No less meaningful for historical institutionalists is the ability to digitally mine, save, organize, search, and retrieve big data sources. Given the gargantuan and constantly growing volume of historical records, researchers can never read all the relevant sources all the way through (although they should peruse enough sources to establish the indispensable historical context). Luckily, the possibility of digitally searching these sources' content allows investigators to quickly and easily find those chunks of text (and, increasingly, of audio, video, and visual material) that pertain to their research questions and units of analysis. As Paul Pierson, one of the forefathers of historical institutionalism in political science, muses, the digital revolution "makes it possible to examine huge quantities of text, increasing researchers' ability to accurately map mass and elite political expression over time."[84] In the online appendix,[85] we suggest several ways of identifying and finding important elements in a large text corpus through regular expressions—common patterns in the text—as well as other methods.[86]

To illustrate both the value of unstructured data and the prospects of unstructured *big* data in such research, consider Kathryn Sikkink's contribution to a recent volume on historical institutionalism in international politics. Sikkink demonstrates that the consolidation of human rights institutions in Latin America in the 1980s harks back to the 1940s. This argument undermines an important study of this topic, according to which the consolidation of such institutions only happened in the 1970s, when social movements in the region began to advocate for human rights. Sikkink explains that the post–World War II support of Latin American governments for human rights set in motion a prolonged critical juncture during which the path of human rights institutions

[83]See, for example, the Combined Operations Project, accessed at https://www.combinedops.com/about_site_background.htm, 9 April 2020; and Commando Veterans Archive, accessed at http://www.commandoveterans.org/, 9 April 2020.

[84]Paul Pierson, "Power in Historical Institutionalism," in Orfeo Fioretos, Tulia G. Falleti, and Adam Sheingate, eds., *The Oxford Handbook of Historical Institutionalism* (Oxford: Oxford University Press, 2016), 138.

[85]Available at https://github.com/jonathan-grossman/Political-Science-and-Big-Data/blob/master/Online-Appendix.pdf.

[86]On regular expressions, see Jeffrey E.F. Friedl, *Mastering Regular Expressions,* 3rd ed. (Sebastapol, CA: O'Reilly, 2006), 4.

slowly and incrementally changed, finally stabilizing and reaching path dependence in the 1980s. Arriving at this conclusion required not only the methodological toolbox of historical institutionalism, but also the careful examination of various historical accounts, legal documents, and organizational reports over a long span of time.[87]

Sikkink's study showcases the crucial role that sensitivity to historical context and detail play when building sound historical institutionalist arguments. However, in the digital age, these arguments can become more robust by drawing on the immense pool of independent historical sources that the internet offers. Unstructured big data such as minutes of discussions in Latin American and other governments and parliaments, op-eds and reports in Latin American and other newspapers, or the texts of political speeches could have reinforced (or, perhaps, modified) Sikkink's conclusions. Navigating such large collections of documents has become easier thanks to the digital searchability of many online sources; by using keywords such as "human rights" (or "derechos humanos"/ "direitos humanos" in Spanish/Portuguese), researchers could follow these institutions more closely and accurately.

Process Tracing and Big Data

When historical institutionalists want to make causal claims and theorize causal mechanisms based on empirical evidence, they must choose an adequate method of analysis.[88] Process tracing—a qualitative method of within-case analysis that is popular among historical institutionalists[89] as well as other political scientists—can particularly benefit from unstructured big data. Drawing on theoretical and methodological contributions from history and psychology, process tracers put rival hypotheses to the test of existing evidence in order to establish causality and identify causal mechanisms.[90] In formal process tracing, researchers use Bayesian reasoning to evaluate the likelihood of a hypothesis given the evidence at hand and their prior knowledge, which they translate into numerical values. Later on, they

[87]Kathryn Sikkink, "Timing and Sequencing in International Politics: Latin America's Contributions to Human Rights," in Orfeo Fioretos, ed., *International Politics and Institutions in Time* (New York: Oxford University Press, 2017), 231–250.
[88]Fioretos, "Institutions and Time in International Relations," 23–24.
[89]Capoccia and Kelemen, "The Study of Critical Junctures," 358.
[90]Derek Beach and Rasmus Brun Pedersen, *Process-Tracing Methods: Foundations and Guidelines* (Ann Arbor: University of Michigan Press, 2013); Andrew Bennett and Jeffrey T. Checkel, eds., *Process Tracing: From Metaphor to Analytic Tool* (Cambridge: Cambridge University Press, 2015); George and Bennett, *Case Studies and Theory Development,* chap. 10; John Gerring, *Case Study Research: Principles and Practices* (Cambridge: Cambridge University Press, 2007), chap. 7; and Ingo Rohlfing, *Case Studies and Causal Inference: An Integrative Framework.* (Basingstoke: Palgrave Macmillan, 2012), chap. 6.

update this probability as new evidence unfolds.[91] Alternatively, the causal inference in process tracing may take a more narrative, although by no means less rigorous, form.[92] In either case, the collection and assessment of evidence are theory-guided, context-sensitive, and explicit (although many other case studies, such as Lamb's article on the *Mayaguez* incident and Recchia's article on France's threat to veto the attack against Iraq, employ an implicit form of process tracing). Like historical institutionalists, process tracers must be highly familiar with their subject matter and its history and they are expected to describe every stage of their inference in a transparent way that allows readers to replicate it.[93]

Jacob Ricks and Amy Liu, in an appendix to their methodological article about process tracing, demonstrate how this method can help researchers test hypotheses. In one of their exemplary case studies, the outcome of interest is the prominence, since 2001, of the Thai Rak Thai (TRT) Party in Thailand—a country characterized by weak political parties and unstable coalitions. Ricks and Liu ask whether the TRT's unprecedented success was attributable to the personal clout of its leader, Thaksin Shinawatra, or to the institutional changes brought about by the 1997 Constitution. After presenting these two hypotheses, the authors establish timelines containing the most important events that happened between the hypothesized causes and the outcome of interest; create causal graphs that visualize possible causal links between these events; identify counterfactual outcomes—what would have happened had Thaksin not founded the TRT Party or had the 1997 constitutional changes not been implemented; and determine which types of evidence are needed to support or refute each hypothesis.[94]

Putting the two hypotheses to the test of evidence, Ricks and Liu survey a rich body of secondary literature on Thai politics. These data support the institutional hypothesis and weaken the personal one: Thaksin already led a political party in the 1996 election (that is, before the 1997 constitutional reforms) and suffered a crushing defeat; therefore, his personality was probably not sufficient to attain electoral victory. Moreover, there is evidence that after the 2001 election, constitutional

---

[91]Tasha Fairfield and Andrew E. Charman, "Explicit Bayesian Analysis for Process Tracing: Guidelines, Opportunities, and Caveats," *Political Analysis* 25 (May 2017): 363–380.

[92]David Collier, "Understanding Process Tracing," *PS: Political Science & Politics* 44 (October 2011): 823–830.

[93]Beach and Pedersen, *Process-Tracing Methods*, 123–126; Christopher Darnton, "Archives and Inference: Documentary Evidence in Case Study Research and the Debate over U.S. Entry into World War II," *International Security* 42 (Winter 2017/18): 84–126; and Kreuzer, "The Structure of Description."

[94]Jacob I. Ricks and Amy H. Liu, "Process-Tracing Research Designs: A Practical Guide," *PS: Political Science & Politics* 51 (October 2018): 1–5, appendix, at 9–18.

restrictions prevented the secession of one of the major factions that composed the TRT and thus ensured the party's integrity. This evidence lends support to the claim that it was the 1997 constitutional reform that allowed the TRT to remain united and form a strong coalition. While Ricks and Liu could not find decisive evidence in the surveyed literature to prove or discredit either hypothesis beyond any doubt, their deep familiarity with the Thai political context allows them to conclude that given the evidence at hand, the institutional hypothesis best explains the TRT's success. At the same time, they do not discount the rival hypothesis and acknowledge Thaksin's individual contribution to this outcome.[95] This example shows researchers developing and employing good research design by relying on process tracing as well as area expertise and knowledge of context to solve a political puzzle.[96]

When inferring causality, *time* is of the essence. Like historical institutionalists, process tracers equally seek to identify exact timeframes and attach events and interactions to specific points in time.[97] They are required to establish a detailed and accurate timeline divided into years, months, days, or even hours and minutes, depending on the level of granularity of inference. Such a timeline constitutes the first layer of research. Once it is identified, documented, and verified, it provides a foundation for multilayered analysis that may include temporal, spatial, and networked elements. As with historical institutionalism, unstructured big data sources such as digitized archival documents, interview transcripts, news articles, book manuscripts, and even secondary literature can provide thick political, historical, social, and cultural description for this purpose.[98]

Thus, to test the hypothesis that Thai constitutional reform accounted for the TRT Party's success, Ricks and Liu had to show that the Constitution predated the TRT's electoral victory. Otherwise, the hypothesis would be invalid.[99] While tracing this particular sequence was not difficult to do—the Constitution was changed in 1997 and the election took place in 2001, as could be easily verified in any reliable chronology of political events in Thailand—other timelines require much greater fragmentation and could thus benefit markedly from drawing on big data.

---

[95]Ricks and Liu, "Process-Tracing Research Designs," appendix, 15–18.
[96]See also James Mahoney, "Process Tracing and Historical Explanation," *Security Studies* 24 (June 2015): 200–218.
[97]Collier, "Understanding Process Tracing," 824; and Ricks and Liu, "Process-Tracing Research Designs," 2.
[98]See Kitchin, "Big Data," 10.
[99]Ricks and Liu, "Process-Tracing Research Designs," appendix, 16.

In the 1962 Cuban missile crisis, for example, a myriad of interactions between many actors occurred within a few days. It is now possible to reconstruct many of these exchanges thanks to a great selection of primary documents from various provenances that are available online; typing the phrase "documents on the Cuban missile crisis" into any web search engine will return an overwhelming number of relevant results. This, however, is not enough. Because of the density of events related to this crisis, process tracers would need to know and record the exact hour—and even the minute—in which these events started and ended to identify causal links between the many distinct interactions.

Even though historical institutionalists and process tracers use data that are, for the most part, unstructured, so far they have paid scant attention to the potential of unstructured big data. By embracing the big data revolution, adherents of these approaches can attain unprecedented levels of granularity and hence of accuracy and reliability, in keeping with the growing standards of transparency and reproducibility in the social sciences.[100] However, unstructured big data also pose new challenges for political scientists. The final section of this article discusses these challenges as well as the best ways to face them.

THE CHALLENGES OF UNSTRUCTURED BIG DATA AND HOW TO OVERCOME THEM

The incredible quantity of big data tests scholars' aims to conduct research with measurable goals and a reasonable time frame. Not only is the volume of data overwhelming, but data can also be misleading and onerous. Finding the evidence that we need in an ocean of irrelevant and false information is a constant challenge. Even if we encounter some details that seem pertinent, we still have to verify their accuracy. In this respect, the big data revolution is a blessing and a curse, as it magnifies and exacerbates reliability problems that have always been part of any inquiry of unstructured sources. While the amount of data in the world keeps growing, most available data are of dubious quality.[101] As Tom Nichols bluntly puts it, "The Internet lets a billion flowers bloom, and most of them stink."[102]

---

[100]R. Michael Alvarez, Ellen M. Key, and Lucas Núñez, "Research Replication: Practical Considerations," *PS: Political Science & Politics* 51 (April 2018): 422–426; and Sean Yom, "Analytic Transparency, Radical Honesty, and Strategic Incentives," *PS: Political Science & Politics* 51 (April 2018): 416–421.
[101]Silver, *The Signal and the Noise*, 250.
[102]Thomas M. Nichols, *The Death of Expertise: The Campaign against Established Knowledge and Why It Matters* (New York: Oxford University Press, 2017), 108.

The sheer amount of data about a certain topic is not necessarily an indicator of accuracy. Making quantitative judgments regarding the quality of data—that is, attempting to determine which pieces of information are correct based solely on the number of sources that support them—would often be precarious. In the age of "fake news," 99 news articles that back up a particular claim might be the product of one disinformation campaign, whereas the single report that substantiates the rival hypothesis might emanate from the only independent journalist bold enough to uncover the truth or committed enough to defend it. Moreover, using evidence volume as a proxy for evidence reliability might result in the exclusion of minorities and women, who are often underrepresented in big data repositories.[103] Thus, the researchers' expertise in the context and history of the period and phenomena they investigate, as well as their ability to evaluate the reliability of their sources, are prerequisites for a good unstructured big data analysis. Data scientists or quantitatively oriented political scientists who were not trained in historical methods, or even a historically oriented investigator who does not specialize in the area or period in question, are likely to encounter difficulties when dealing with such sources without guidance from or collaboration with other experts.[104]

When researching unstructured data, we often strive to find the needles-in-haystacks—those elusive pieces of evidence that provide insight and truth and can confirm a hypothesis, impugn rival explanations, or both.[105] When our data consist of a large number of documents, it is easy to miss a name that occurs only once in the entire corpus. Robert A. Caro, the famous biographer and twice Pulitzer Prize winner, recalls in his memoir the invaluable advice of a trusted editor: "Turn every page. Never assume anything. Turn every goddamned page."[106] However, when there are too many pages, it might be impossible to turn each and every one of them. Caro writes that when he embarked on his research at the Lyndon Baines Johnson Library and Museum in Austin, Texas, upon learning that the LBJ archives contain around 32 million pages, he had to concede that "there would be no turning every page here."[107] His realization highlights the indispensability of digitally searching big data

[103]Brooke Foucault Welles, "On Minorities and Outliers: The Case for Making Big Data Small," *Big Data & Society* 1 (April–June 2014), https://doi.org/10.1177/2053951714540613.
[104]See George and Bennett, *Case Studies and Theory Development,* 96.
[105]Stephen Van Evera, *Guide to Methods for Students of Political Science* (Ithaca, NY: Cornell University Press, 1997), 30–32.
[106]Robert A. Caro, *Working: Researching, Interviewing, Writing* (New York: Alfred A. Knopf, 2019), 11.
[107]Caro, *Working,* 84.

by keywords, strings, and regular expressions, since carefully reviewing and annotating all the relevant sources, as investigators used to do prior to the information era (and, as Caro's anecdote demonstrates, were often unable to accomplish even in the age of small data), is no longer feasible—there are simply too many books to read, video clips to watch, or audio recordings to hear.

Moreover, even when we know how to find and extract the information we need, we must be able to keep track of these details in order to analyze them. Historians who examine primary sources arrange their research notes and present their findings in a logical order for their work to be meaningful.[108] Political scientists studying historical data are no different in this regard. To make sense of unstructured sources, and especially of large digital collections of such sources, we must treat our data according to clear organizing principles.[109] Unlike structuring strategies whose goal is the statistical analysis of data, historical institutionalism and process tracing do not entail the irrevocable quantification or reduction of fine-grained historical narrative. Although scholars engaging in such research endeavors must be able to locate, organize, and retrieve only those facts relevant to their questions and cast aside the noise, they must also be able to consult the raw material whenever they need for the purposes of context, replication, or further inquiry.

One possible way to attain this goal is to create a qualitative codebook in the form of a spreadsheet or a relational database. Such a codebook contains, in a tabular and digitally searchable form, all the germane dates, events, actors, and locations, as well as links to the original sources. On the one hand, the codebook would give researchers a bird's-eye view of all the meaningful items in their study and allow them to array actors and events in order of time, importance, influence over outcome of interest, or any other order. A relational database can additionally allow researchers to create causal links and other connections between actors and events. On the other hand, immediate access to the original sources would permit researchers the close and careful reading of historical sources that is central to historical institutionalism and process tracing, offset the loss of context that is characteristic of structured big data, and ensure the replicability of their study.

---

[108]Anthony Brundage, *Going to the Sources: A Guide to Historical Research and Writing,* 6th ed. (Hoboken, NJ: Wiley, 2017), 114–121.
[109]George and Bennett, *Case Studies and Theory Development,* 90.

To further improve and contextualize the retrieval of relevant details, researchers can allocate columns in the spreadsheet to *descriptive metadata*—data that describe the content of a resource and allow users to look up that resource based on its attributes in a way similar to the category labels in an email inbox or a reference manager.[110] Throughout the research process, the investigators should constantly update the codebook in accordance with the data they find while documenting their steps and decisions to guarantee their project's transparency and replicability. In the online appendix, we provide more elaborate instructions on how to create such a codebook.[111]

The qualitative codebook is just one method for organizing unstructured big data, and it requires some computer knowledge and skills. More traditional methods, such as identifying causal mechanisms by drawing causal graphs[112] or set diagrams,[113] or using an intuitive timeline and mind-mapping software that help writers to keep track of their narrative and characters, may be equally useful in organizing the plethora of details found in unstructured big data. In any event, analyzing such data are most effective when they are machine-readable, fully searchable, and, preferably, downloaded in toto, so that researchers, reviewers, and readers can revisit the data even if the creators or owners of the data change or remove them from their original location. Manually downloading massive amounts of web pages, text documents, or media files can be a time-consuming task. More efficient solutions would be clipping web pages with a designated browser extension that imports the source into some software, such as a note-taking application, a computer-assisted qualitative data analysis (CAQDA) package, or a reference manager; running an automated script that extracts the data from the internet;[114] or using a more user-friendly desktop or web-based scraping service that boasts a graphical user interface. Alternatively, researchers can delegate data mining to a host of human workers through crowdsourcing web services such as Amazon Mechanical Turk.[115]

---

[110]Richard Gartner, *Metadata: Shaping Knowledge from Antiquity to the Semantic Web* (Cham, Switzerland: Springer, 2016), 6–7; and Richard Pearce-Moses, *A Glossary of Archival and Records Terminology,* Archival Fundamentals Series (Chicago: Society of American Archivists, 2005), 113.

[111]Available at https://github.com/jonathan-grossman/Political-Science-and-Big-Data/blob/master/Online-Appendix.pdf.

[112]Ricks and Liu, "Process-Tracing Research Designs," 2–3.

[113]James Mahoney, Khairunnisa Mohamedali, and Christoph Nguyen, "Causality and Time in Historical Institutionalism," in Orfeo Fioretos, Tulia G. Falleti, and Adam Sheingate, eds., *The Oxford Handbook of Historical Institutionalism* (Oxford: Oxford University Press, 2016), 71–88.

[114]Gabe Ignatow and Rada Mihalcea, *An Introduction to Text Mining: Research Design, Data Collection, and Analysis* (Los Angeles: Sage, 2018).

[115]Wilkerson and Casas, "Large-Scale Computerized Text Analysis," 531.

Finally, regardless of the ways data are found, downloaded, and organized, we believe that human researchers rather than machines should be in control of analyzing unstructured big data to ensure that nuanced contextual analyses accompany any quantitative analysis. We do not reject the possibility that throughout the collection and examination of data, technology can help us unveil patterns that we have not yet considered. As it is often practically impossible to read or view all the available evidence, automated methods can help us with this effort. For example, we can use CAQDA software and text-as-data methods to check the frequency of words in a corpus of sources. In this way, we can become familiar with the concepts, actors, locations, and ideas that predominated different moments in time, visualize those findings as a word cloud or a list of words,[116] and narrow down both our digital searches and our analytic focus. Nonetheless, we do not recommend a wholly inductive approach. As this article has indicated multiple times, there are no perfect digital shortcuts to immersing ourselves in the context and history of the political phenomenon that we wish to explain.

CONCLUSION

"Big data" might not be a buzzword in political science yet, but it is nevertheless widely used by political scientists. So far, this use has been almost exclusively in the form of statistical analyses of structured databases. Through this article, we seek to challenge this state of affairs by advocating the analysis of unstructured big data in political research, especially in the context of historical institutionalism and process tracing.

Given that most data in the world are unstructured to some degree, making inferences from exclusively structured data may result in a "drunkard's search" bias in that it would often include only those data that are easiest to obtain and analyze. To deliberately confine our investigation to what constitutes, at best, 20 percent of the available data in the world is to restrict the scope of our research, if not cherry-pick our evidence. In the age of information, the nexus between accessible unstructured data of unprecedented magnitude and novel methods, both quantitative and qualitative, for the analysis of such data has the potential to shed light on formerly hidden political mysteries. It can lead us from descriptive or correlative studies to an accurate depiction of causal political processes and mechanisms and thus enhance our understanding of the world in which we live.

---

[116]Klaus Krippendorff, *Content Analysis: An Introduction to Its Methodology,* 3rd ed. (Los Angeles: Sage, 2013), 189–192.

Of course, as we emphasize throughout the article, investigators should choose their method and data according to their research question—some questions are best answered with the statistical inference of structured data, others with in-depth analyses of unstructured sources, and still others with a combination of both—one common approach, for example, is identifying a causal effect with the former and theorizing the causal mechanism with the latter.[117] In any case, researchers should do their best to avoid the pitfalls of big data while acknowledging that avoiding them altogether is unfeasible.

More research and documented findings are necessary before the majority of political scientists feel at ease studying structured or unstructured big data. Although big data analysis can be done today using common software that are widely available for free or at a low price, it still requires varying degrees of computational skills. Python and R, for example, are two powerful open-source programming languages that many political scientists use to extract, process, and analyze big data sources. On the one hand, these programs are freely available and enjoy a vast universe of users, developers, downloadable packages and modules, online courses, blogs, and guidebooks. On the other hand, they are very unintuitive—mastering them requires a steep learning curve and considerable time resources; moreover, these skills may come at the expense of acquiring others that may be more useful for certain research projects, such as gaining expertise in the language, history, and politics of the studied area or country.[118]

Certainly, a wide digital divide exists between scholars who know how to write algorithms in a programming language, those who do not code but are capable of using relatively sophisticated (and often pricey) digital tools with a graphical user interface (for example, relational databases, text mining applications, or CAQDA software), and those with only basic digital literacy. For the moment, big data are fully accessible only to the first group and, with a lesser degree of functionality, to the second one.[119] Developing software packages that would simplify the search, extraction, organization, and analysis of such data for members of all groups is among the most pressing challenges that social science is facing today.

Making these tools available for free or at an affordable price, so as to increase equal access across different countries, institutions, and

[117]Gerring, *Case Study Research,* 43.
[118]See Thomas B. Pepinsky, "The Return of the Single-Country Study," *Annual Review of Political Science* 22 (2019): 187–203.
[119]See also Henry E. Brady, "The Challenge of Big Data and Data Science," *Annual Review of Political Science* 22 (2019): 297–323.

individuals, is another major challenge. Given the ever-increasing availability of user-friendly digital tools for social science research, on the one hand, and the rapid proliferation of open-source software, on the other hand, there is good reason to believe that these challenges are temporary. Scholars may also increasingly develop the necessary skill-sets to code data or collaborate across disciplines and beyond academia with people that already possess these skills. We look forward to the contributions that these advances will make to the study of big data in political science.*

*We thank the editors and anonymous reviewers of *Political Science Quarterly* and Raelene Camille Wyse for their helpful comments on earlier drafts of this article.