The History of Science and the Science of History: Computational Methods, Algorithms, and the Future of the Field

Abraham Gibson, *Arizona State University*
Cindy Ermus, *University of Texas at San Antonio*

Abstract: Many researchers insist that computational methods will transform the historical profession, while an equally large number reject these claims as unwarranted hype. This study attempts to place the debate in historical and social context. The essay is divided into three parts. The first part offers a brief review of computational history. It asks whether the "computational turn" bears any resemblance to quantitative history and how it fits within the digital humanities. The second part describes the authors' recent attempts to apply computational methods to an existing project in the history of science using a standardized workflow. It demonstrates that each step of the workflow adds another layer of subjectivity. The third part reflects on what computational methods mean for the historical profession. It systematically reviews the positive aspects of computational history (open access, interdisciplinary collaboration, and new perspectives) as well as the negative aspects (inequality, fragility, and the threat of automation) and offers prescriptions based on the authors' experiences.

The digital revolution often evokes images of a technocratic future, but its impact on the past is equally profound. Computers have transformed not only the historical record but the historical profession as well. Historians have eagerly embraced digital tools and methods, and we now rely on computers for every single aspect of the trade, from collection and curation to analysis and interpretation. We can now teach courses, write books, and build careers without

Abraham Gibson is a National Science Foundation Postdoctoral Fellow in the Center for Biology and Society at Arizona State University. His research examines the history of biology, the ethics of engineering, and the social impact of data science. He recently published *Feral Animals in the American South: An Evolutionary History* (Cambridge, 2016), and he is now writing a book about scientific holism during the interwar years. Center for Biology and Society, Arizona State University, 427 East Tyler Mall, Tempe, Arizona 85287-4501, USA; abraham.gibson@asu.edu.

Cindy Ermus is Assistant Professor of History at the University of Texas at San Antonio and a Visiting Scholar in the Center for Biology and Society at Arizona State University. She specializes in the history of science, medicine, and the environment in France and the Atlantic World from the eighteenth century to the present. She is the editor of a volume entitled *Environmental Disaster in the Gulf South: Two Centuries of Catastrophe, Risk, and Resilience* (LSU, 2018) and is currently completing a monograph on the 1720 Plague of Provence. She is also cofounder and executive editor at AgeofRevolutions.com. Department of History, University of Texas at San Antonio, College of Liberal and Fine Arts, 1 UTSA Circle, San Antonio, Texas 27249, USA; cindy.ermus@utsa.edu.

Isis, volume 110, number 3. © 2019 by The History of Science Society.
All rights reserved. 0021-1753/2019/0110-0009$10.00.

ever reading a physical journal or, increasingly, without ever leaving the house. To paraphrase Lincoln Mullen, we are all digital historians now.[1]

And yet, despite our critical dependence on computers, historians are generally ignorant about even the most basic concepts in computer science. Our two disciplines do not share methods, tools, or theories. We publish in different journals, and we are trained on different sides of campus. It is true that computational research often requires heavy-duty processing power, which is a nonstarter in most cash-strapped history departments, but communication is the much larger obstacle. Historians and computer scientists speak very different languages. Basic ideas like "Jupyter notebooks" and "natural language processing" are unintelligible to most historians, whereas "historiography" and "presentism" mean nothing to most computer scientists.

Historians of science have a unique opportunity to help bridge the gap. As practitioners of one of the only disciplines whose members frequent every library on campus, we are conversant in both the humanities and the STEM fields.[2] We are inherently sensitive to social, cultural, and technological entanglements and can thus serve as mediators between these two worlds. In theory, anyway. In practice, the fabled learning curve discourages most historians of science from even trying. This essay describes one modest effort to change that. We wanted to know whether digital neophytes with no previous training can learn and apply computational methods to the history of science and whether the fruits of computational history and philosophy of science (HPS) are worth all the labor. Our findings are instructive.

The essay is divided into three parts. The first part reviews the contested place of numbers in the history of the profession. We briefly explain how computational history is related to both quantitative history and the digital humanities in general. The second part describes our recent attempts to learn and apply computational methods to a project in the history and philosophy of science using a standardized workflow. Each step of the workflow adds another layer of subjectivity, and each engages the existing literature in its own unique way. The third and final part reflects on what computational methods mean for HPS. We contemplate what the profession will look like in the future, and we offer prescriptions based on our experiences.

I

Historians have used numbers to help them navigate "big data" since antiquity, but quantitative studies of the past accelerated following the end of World War II. Some cite Father Roberto Busa and his punch-card concordances as the origin of quantitative history, though one must also cite the female programmers who designed and executed many of his projects. Some point to the rise of cliometrics, the *Annales* school, and the new social history as the origin of the field. These perspectives emphasized the multitudes (what Jesse Lemisch called "history from the bottom up"), which accorded well with computational thirst for data.[3]

Even so, quantitative history faced significant challenges. In 1962 the president of the American Historical Association (AHA), Carl Bridenbaugh, emphatically condemned "that Bitch-goddess, QUANTIFICATION," in his presidential address to the organization. Bernard Bailyn

[1] Lincoln Mullen, "Digital Humanities Is a Spectrum; or, We're All Digital Humanists Now," in *Defining Digital Humanities: A Reader*, ed. Melissa Terras, Edward Vanhoutte, and Julianne Nyhan (Farnham, Surrey: Ashgate, 2013), pp. 237–238.

[2] Stephen Weldon, "Recent Trends in Research in the Field of the History of Science," Webinar Presentation for the Collections Forum on Science, Technology, and Engineering, Center for Research Libraries, Annual Council Meeting, 19 Apr. 2013.

[3] Markus Krajewski, "Tell Data from Meta: Tracing the Origins of Big Data, Bibliometrics, and the OPAC," *Osiris*, 2017, N.S., 32:224–240; Melissa Terras and Julianne Nyhan, "Father Busa's Female Punch Card Operatives," in *Debates in the Digital Humanities*, ed. Matthew K. Gold and Lauren F. Klein (Minneapolis: Univ. Minnesota Press, 2016) (hereafter cited as ***Debates in the Digital Humanities*** **[2016]**), pp. 60–65; and Jesse Lemisch, "The American Revolution Seen from the Bottom Up," in *Towards a New Past: Dissenting Essays in American History*, ed. Barton J. Bernstein (New York: Pantheon, 1968), pp. 3–45.

lamented the "soul-entrapping difficulties" of quantitative history in his own presidential address to the AHA nineteen years later.[4] Following the controversial claims about American slavery in *Time on the Cross*, many historians forswore quantitative methods altogether. By the 1980s, quantitative history was anathema. Historians who once praised quantification now recanted with equal zeal. Those who retained faith in numbers faced "outright hostility."[5]

Even as historians abandoned quantitative methods, they embraced digital tools. In the span of a generation, word processors replaced typewriters. Hard drives replaced file cabinets. The biggest change was the explosion of the internet. Card catalogues and academic journals were posted online, providing historians with unprecedented access to the profession's innumerable historiographies. The early 2000s witnessed the rise of academic blogging, the launch of widespread digitization initiatives, and the creation of MOOCs, among other new features. Many of these self-identifying digital historians sought to distance themselves from their still-discredited quantitative forebears.[6]

By the early 2010s, the field of digital humanities was cohering around the "big tent" metaphor that first emerged during the Alliance of Digital Humanities Organizations annual conference in 2011. It later served as a central organizing framework for the first volume of *Debates in the Digital Humanities* in 2012. More recently, researchers have acknowledged that while the big tent metaphor invites scholars of every stripe to take up digital methods, it has failed to articulate what, precisely, the digital humanities entail.[7] In fact, the lack of clarification is something of a running joke in the field. Perhaps you have encountered the Twitter bot (@DHDefined) that tweets random definitions of the digital humanities every four hours.

Within the big tent of digital humanities, the field of computational history has grown exponentially. Computational history is the result of two major developments: the rise of data science and the development of increasingly powerful computers. There are many ways to interrogate data, but the goal is generally the same: identify signals within the noise. Historians of science are at the forefront of this movement. Last year, *Historical Studies in the Natural Sciences* published a special issue on "Histories of Data and the Database," while *Technology and Culture* published a special issue entitled "Shift CTRL: New Directions in the History of Computing." Two years ago, the annual issue of *Osiris* focused on "Data Histories." The journal you are reading, *Isis*, has published numerous articles about computational history in recent years, and several discipline-specific journals have featured a growing number of articles based on computational studies.[8]

[4] Carl Bridenbaugh, "The Great Mutation," *American Historical Review*, 1963, 68:315–331, on p. 326; and Bernard Bailyn, "The Challenge of Modern Historiography," *ibid.*, 1982, 87:1–24, on p. 6.

[5] Robert William Fogel and Stanley L. Engerman, *Time on the Cross: The Economics of American Negro Slavery* (Boston: Little, Brown, 1974); "Passages from Quantitative History to Digital Humanities," panel at AHA Annual Meeting, Chicago, 4 Jan. 2019; and Adam Tooze, "Trouble with Numbers: Statistics, Politics, and History in the Construction of Weimar's Trade Balance, 1918–1924," *Amer. Hist. Rev.*, 2008, *113*:678–700, on p. 679 (quotation).

[6] Cameron Blevins, "Digital History's Perpetual Future Tense," in *Debates in the Digital Humanities* (2016), pp. 308–324.

[7] Lauren F. Klein and Matthew K. Gold, "Digital Humanities: The Expanded Field," in *Debates in the Digital Humanities* (2016), pp. ix–xvi.

[8] Nate Silver, *The Signal and the Noise: Why So Many Predictions Fail—But Some Don't* (New York: Penguin, 2015); "Histories of Data and the Database," *Historical Studies in the Natural Sciences*, 2018, 48(5); "Shift CTRL: New Directions in the History of Computing," *Technology and Culture*, 2018, 59(4 [suppl.]); Elena Aronova, Christine von Oertzen, and David Sepkoski, eds., "Data Histories," *Osiris*, 2017, N.S., 32; Anna-Luna Post and Andreas Weber, "Notes on the Reviewing of Learned Websites, Digital Resources, and Tools," *Isis*, 2018, *109*:796–800; Leonid Zhmud and Alexei Kouprianov, "Ancient Greek *Mathēmata* from a Sociological Perspective: A Quantitative Analysis," *ibid.*, pp. 445–472; Brian Ogilvie, "Scientific Archives in the Age of Digitization," *ibid.*, 2017, *107*:77–85; Colin Allen and the InPhO Group, "Cross-Cutting Categorization Schemes in the Digital Humanities," *ibid.*, 2013, *104*:573–583; and Stephen P. Weldon, "Ordering the Discipline: Classification in the History of Science," *ibid.*, pp. 537–539.

Given the purported value of computational methods, how should those of us with limited computer skills respond? Researchers offer a variety of opinions. Some insist that every historian should be trained in computer science and that every historian should learn how to read and write code. As early as 1968, Emmanuel Le Roy Ladurie predicted that "the historian of tomorrow will be a programmer or will not be at all." Others insist that the opposite is true and that reconciliation is impossible. Jill Lepore recently caused a stir when she remarked in an interview with the *Chronicle of Higher Education* that digital humanities are not the *real* humanities. The stakes are high. "If we do not wake up soon to the new realities of big data," Roberto Franzosi warns, "computer scientists will leave us behind, biting the dust in this road to knowledge."[9]

II

Our experiences using computational history reflect several broader trends in the field. For example, this essay was funded with "soft money." This is an increasingly common practice in digital humanities, and it is consistent with the larger shift toward the "adjunct-ification," or "Uberization," of work in the early twenty-first century.[10] In this case, one of us (AG) secured a National Science Foundation–funded postdoctoral fellowship that provided immersive training in computational methods at the Laubichler Lab for Computational HPS. Located in the Center for Biology and Society at Arizona State University, this lab uses high-powered computers to identify networks, patterns, and trends within the swarm of big data that is the history of science. The second author (CE) joined the lab in a formal capacity, and this article as coauthor, relatively soon thereafter.

While they prefer to start integrated training at the undergraduate level, lab members agreed to host our project for several reasons. First, much of their research has dealt with data that was born digital, including a spate of publications on recent topics like the Anthropocene, the microbiome, and evolutionary medicine. They were intrigued that our research dealt with digitized (rather than born-digital) sources, and they wanted to show that computational methods could be applied to older topics. Second, they were keen to show that computational methods apply to both large-scale *and* small-scale topics. David Armitage and Jo Guldi had recently published *The History Manifesto*, which celebrated computational methods because they enabled *longue durée* analyses.[11] Our research would help the lab demonstrate that computational methods likewise enable increasingly detailed microhistories.

The Laubichler lab has created a generalized workflow, which we used in this project. The workflow offers a six-step guide to starting and successfully executing any project in computational HPS. What follows is a brief rundown of the methods we employed, the data we analyzed, and the

[9] Emmanuel Le Roy Ladurie, "La fin des érudits," *Nouvel Observateur*, 8 May 1968; Evan Goldstein, "The Academy Is Largely Itself Responsible for Its Own Peril," *Chronicle of Higher Education*, 13 Nov. 2018; and Roberto Franzosi, "A Third Road to the Past? Historical Scholarship in the Age of Big Data," *Historical Methods*, 2017, 50:227–244, on p. 241. See also Shawn Graham, Ian Milligan, and Scott Weingart, *Exploring Big Historical Data: The Historian's Macroscope* (London: Imperial College Press, 2015).

[10] Miriam Posner, "Here and There: Creating DH Community," in *Debates in the Digital Humanities* (2016), pp. 265–273; and Julia Flanders, "Time, Labor, and 'Alternate Careers' in Digital Humanities Knowledge Work," in *Debates in the Digital Humanities*, ed. Matthew K. Gold (Minneapolis: Univ. Minnesota Press, 2012) (hereafter cited as ***Debates in the Digital Humanities*** **[2012]**), pp. 292–308. For broader context see Alex Rosenblat, *Uberland: How Algorithms Are Rewriting the Rules of Work* (Berkeley: Univ. California Press, 2018).

[11] David Armitage and Jo Guldi, *The History Manifesto* (New York: Cambridge Univ. Press, 2014). Regarding preferred training practices at the Laubichler lab see B. R. Erick Peirson, Julia Damerow, and Manfred Laubichler, "Software Development and Transdisciplinary Training at the Interface of Digital Humanities and Computer Science," *Digital Studies* (2016), http://doi.org/10.16995/dscn.17.

conclusions we drew. Despite dreams of machine-like exactitude, every step in the process drips with value-laden judgment calls.

Step 1: Planning

Countless books and articles promise that computational methods allow one to ask entirely new types of historical questions (see Manfred D. Laubichler, Jane Maienschein, and Jürgen Renn, "Computational History of Knowledge: Challenges and Opportunities," in this Focus section). Perhaps; but most of us would first like to know whether they can help answer existing questions. We were especially eager to apply computational methods to a very specific topic: biological ideas about cooperation during the interwar years. The topic was not chosen at random. Over the past few years, there have been murmurings in the field about an organismic synthesis that dominated biological thinking throughout the early twentieth century.[12] There is also precedent. Others have used computational methods to look at similar ideas, like systems biology, complexity, and emergence.[13] So, it is a bit like the discovery of Neptune. We are not just pointing in the dark. The data suggests that *something* is looming, and we orient our instruments accordingly.

The types of questions that one asks determine the types of methods that one employs. This is not intuitive for historians, who seldom discuss methodology in print and who have even been accused of hiding methods. By comparison, digital historians are obsessed with methods. They write and teach about methods more than any other subfield of history, and their articles are as likely to describe the steps a researcher used as the conclusions he or she reached. Some have suggested that historians take a cue from the sciences and include a clearly demarcated methods section in all their publications. Others have suggested that scholars publish two distinct papers, one on methods and one on interpretations. Still others suggest that we integrate methods and theory into a single "braided narrative."[14]

Step 2: Collecting

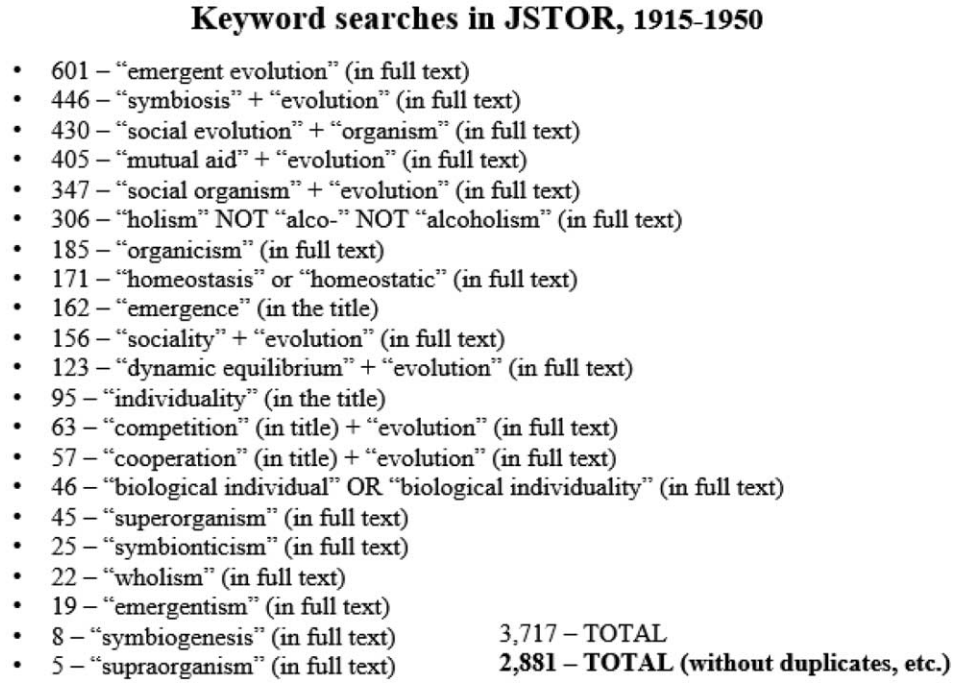
Most studies in computational HPS rely on preexisting, largely self-contained datasets. Recent examples include comprehensive analyses of the *Journal of the History of Biology* and the *Stanford Encyclopedia of Philosophy*, among others.[15] There is only one *JHB*, but all historians have a collection of carefully curated PDFs on our respective computers. In an attempt to emulate noncomputational research practices (and to keep this project manageable), we ran more than twenty keyword searches in JSTOR (see Figure 1), often combining queries to help us winnow the data. Limiting our search to the years between 1915 and 1950, we ended up with more than

[12] Jane Baedke, "O Organism, Where Art Thou? Old and New Challenges for Organism-Centered Biology," *Journal of the History of Biology*, 2019, 52:293–324; Emily Herring, "'Great Is Darwin and Bergson His Poet': Julian Huxley's Other Evolutionary Synthesis," *Annals of Science*, 2018, 75:40–54; Erik L. Peterson, *The Life Organic: The Theoretical Biology Club and the Roots of Epigenetics* (Pittsburgh: Univ. Pittsburgh Press, 2017); Daniel J. Nicholson and Richard Gawne, "Neither Logical Empiricism nor Vitalism, but Organicism: What the Philosophy of Biology Was," *History and Philosophy of the Life Sciences*, 2015, 37:345–381; Maurizio Esposito, *Romantic Biology, 1890–1945* (London: Pickering & Chatto, 2013); and Abraham H. Gibson, Christina L. Kwapich, and Martha Lang, "The Roots of Multilevel Selection: Concepts of Biological Individuality in the Early Twentieth Century," *Hist. Phil. Life Sci.*, 2013, 35:505–532.

[13] Jürgen Jost, Nils Bertschinger, and Eckehard Olbrich, "Emergence," *New Ideas in Psychology*, 2010, 28:265–273; and Yawen Zou and Manfred D. Laubichler, "From Systems to Biology: A Computational Analysis of the Research Articles on Systems Biology from 1992 to 2013," *PLOS ONE*, 25 July 2018, pp. 1–16.

[14] Lincoln A. Mullen, "A Braided Narrative for Digital History," in *Debates in the Digital Humanities*, ed. Matthew K. Gold and Lauren F. Klein (Minneapolis: Univ. Minnesota Press, 2019), pp. 382–388.

[15] B. R. Erick Peirson *et al.*, "Quantitative Perspectives on Fifty Years of the *Journal of the History of Biology*," *J. Hist. Biol.*, 2017, 50:695–751; and Allen and the InPhO Group, "Cross-Cutting Categorization Schemes in the Digital Humanities" (cit. n. 8).



Keyword searches in JSTOR, 1915-1950

- 601 – "emergent evolution" (in full text)
- 446 – "symbiosis" + "evolution" (in full text)
- 430 – "social evolution" + "organism" (in full text)
- 405 – "mutual aid" + "evolution" (in full text)
- 347 – "social organism" + "evolution" (in full text)
- 306 – "holism" NOT "alco-" NOT "alcoholism" (in full text)
- 185 – "organicism" (in full text)
- 171 – "homeostasis" or "homeostatic" (in full text)
- 162 – "emergence" (in the title)
- 156 – "sociality" + "evolution" (in full text)
- 123 – "dynamic equilibrium" + "evolution" (in full text)
- 95 – "individuality" (in the title)
- 63 – "competition" (in title) + "evolution" (in full text)
- 57 – "cooperation" (in title) + "evolution" (in full text)
- 46 – "biological individual" OR "biological individuality" (in full text)
- 45 – "superorganism" (in full text)
- 25 – "symbionticism" (in full text)
- 22 – "wholism" (in full text)
- 19 – "emergentism" (in full text)
- 8 – "symbiogenesis" (in full text)
- 5 – "supraorganism" (in full text)

3,717 – TOTAL

2,881 – TOTAL (without duplicates, etc.)

Figure 1. Keyword searches in JSTOR, 1915–1950.

3,700 documents, a mixed bag of articles, essays, and letters that we stored as PDFs in Zotero. Like all datasets, this corpus was a thoroughly subjective artifact from the start.[16]

Historians have yet to acknowledge fully the profound changes that accompany our wholesale adoption of keyword searches.[17] Instead of reading whole books from cover to cover, we increasingly cherry-pick PDFs using "control + F." In doing so, we risk exaggerating the historical significance of any given person, event, or idea—and yet the practice is ubiquitous. Also, many databases are designed around the search function, which means that users are necessarily experiencing the data through recycled frameworks already on hand. We know that the meaning of words changes over time, and that meaning varies depending on context, yet searches fail to capture this dynamism. Some have suggested that we cite our searches, but searches can be relativized according to location, browsing history, and even device.

Step 3: Preprocessing

Deleting duplicates helped us prune the corpus down to approximately 2,900 documents. None of these items was born digital. Instead, they were digitized and placed online, an increasingly common practice that has revolutionized the study of history. Trillions of words have already been digitized, and trillions more will be soon. As the historian Tim Hitchcock has observed, "We are witnessing the creation of the Western print archive, second edition." Moreover, many

[16] Andrew Stauffer, "Introduction: Searching Engines, Reading Machines," *Victorian Studies*, 2011, *54*:63–68, esp. p. 65.

[17] Hieke Huistra and Bram Mellink, "Phrasing History: Selecting Sources in Digital Repositories," *Hist. Meth.*, 2016, *49*:220–229; Lara Putnam, "The Transnational and the Text-Searchable: Digitized Sources and the Shadows They Cast," *Amer. Hist. Rev.*, 2016, *121*:377–402; and Jeffrey M. Binder, "Alien Reading: Text Mining, Language Standardization, and the Humanities," in *Debates in the Digital Humanities* (2016), pp. 201–217.

of these files now boast optical character recognition (OCR), which means that machines can read the documents. This has led to what Lara Putnam calls the "data-fication" of the historical record. Above all, it means that words are now "available" for analysis.[18]

The words may be available, but they still need to be harvested (see Julia Damerow and Dirk Wintergrün, "The Hitchhiker's Guide to Data in the History of Science," in this Focus section). If we hope to play with the data, they have to be in "computable form." Toward that end, researchers in the Laubichler lab developed Amphora, a tool that collects data from PDFs and copies them into more readable and manipulatable TXT files. Users can upload documents, collect metadata, and analyze extracted text.[19] Even then, the data needs to be prepared in a way that makes it easier to process. For example, we wanted to examine change over time, so we manually divided the files by year. As ever, the dataset retains the personality of its creators.

Step 4: Processing

Computers offer many different ways of interrogating big data, from "distant reading" to "topic modeling" to "network analyses." Our aims were relatively modest. Within the corpus, we sought to identify which topics were most popular at a given time, how various topics were connected to one another, and how the conversation changed over time. We had 2,881 different TXT files with millions of words, which we aggregated into a single "bag of words" using Wordsmith. Then we compared our corpus against the Baker-Brown index, which accounts for the overabundance of filler words like "the," "and," and the like. Creating a "stop list" helped filter out overrepresented words like "volume" and "page," but it added another layer of curation.

Next, we generated network statistics using ORA, a dynamic metanetwork assessment and analysis tool developed by CASOS at Carnegie Mellon, which contains hundreds of metrics for comparing networks, groups, and individuals across a dynamic metanetwork perspective. Notably, this step required advanced knowledge of coding. Students with computational training helped execute the most challenging aspects, which underscores the field's reverence for collaboration but also its frustrating opacity. Even if we learn the basics of code, historians are not trained to fact-check algorithms. Unless we expect every historian to learn advanced coding, using computational methods will mean outsourcing historical analysis to hand-crafted algorithmic "black boxes" we had no part in building.[20]

Step 5: Visualizing

We used VOSviewer, a tool for constructing bibliometric networks that was designed at Leiden University, to help us visualize our data (see Figure 2). One can spin the visualization on three axes, zooming in and out to view connections at various scales. The size of a word and its proximity to other words signal its frequency and its connectedness. In some cases, the insights affirm what we already know. Words like "stimuli" and "reaction" group closely every year. So too do "physics" and

[18] Nanna Bonde Thylstrup, *The Politics of Mass Digitization* (Cambridge, Mass.: MIT Press, 2019); Tim Hitchcock, "Confronting the Digital: Or How Academic History Writing Lost the Plot," *Cultural and Social History*, 2013, *10*:9–23, on pp. 9–10; and Putnam, "Transnational and the Text-Searchable," p. 400.

[19] Manfred D. Laubichler, Jane Maienschein, and Jürgen Renn, "Computational Perspectives in the History of Science: To the Memory of Peter Damerow," *Isis*, 2013, *104*:119–130, on pp. 121–122; https://github.com/diging/amphora; and Julia Damerow, B. R. Erick Peirson, and Laubichler, "The Giles Ecosystem—Storage, Text Extraction, and OCR of Documents," *Journal of Open Research Software*, 2017, 5:1–5.

[20] Kathleen Carley and Jeff Reminga, "ORA: Organization Risk Analyzer," CASOS Technical Report CMU-ISRI-04-106, Carnegie Mellon University (2004), pp. 1–45; Benjamin M. Schmidt, "Do Digital Humanists Need to Understand Algorithms?" in *Debates in the Digital Humanities* (2016), pp. 546–555; and Frank Pasquale, *The Black Box Society: The Secret Algorithms That Control Money and Information* (Cambridge, Mass.: Harvard Univ. Press, 2015).

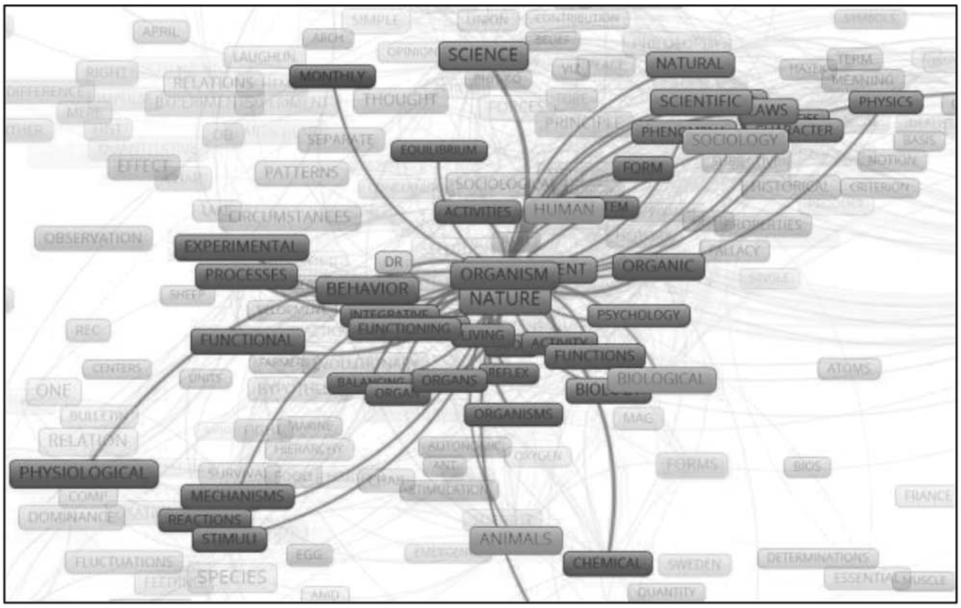


Figure 2. Visualization based on JSTOR search results for 1945.

"laws." In other cases, the visualizations reveal unanticipated connections. We had expected "organism" to feature prominently every year, especially since some variation of the word appeared in five of our JSTOR searches, but we were surprised by the breadth of its connections. Words like "Hitler" and "Nazis" start to cluster around "holism" in the second half of the corpus, partly because fascists called the nation-state an "organism." Additional research reveals that they were not alone. World leaders from across the political spectrum invoked organismic metaphors between the wars. The same ideas that were cited in support of fascism were also cited in support of communism, capitalism, and cosmopolitanism. They were mobilized in support of nationalism *and* globalism. Others have looked at these connections in various national contexts, but our interpretation of the data suggests that a global history of interwar holism is also within reach.[21]

Visualizations can prompt one to rethink an entire project, but that is not their only selling point. Charts and graphs connote scientism, especially if they are in color and aesthetically pleasing. Miriam Posner notes that these visualizations can sometimes look "terrifyingly authoritative" to undergraduate students, but they have a similar effect on professional historians. In light of these facts, it is important to remember that visualizations are constructed artifacts. Toward that end, several scholars have shown that our current toolkit of charts, graphs, and visualizations is not inevitable but the result of many subjective decisions.[22]

[21] Peterson, *Life Organic* (cit. n. 12); Anne Harrington, *Reenchanted Science: Holism in German Culture from Wilhelm II to Hitler* (Princeton, N.J.: Princeton Univ. Press, 1996); Gregg Mitman, *The State of Nature: Ecology, Community, and American Social Thought, 1900–1950* (Chicago: Univ. Chicago Press, 1992); and Daniel P. Todes, *Darwin without Malthus: The Struggle for Existence in Russian Evolutionary Thought* (New York: Oxford Univ. Press, 1989).

[22] Miriam Posner, "What's Next: The Radical, Unrealized Potential of Digital Humanities," in *Debates in the Digital Humanities* (2016), pp. 32–41; Lauren Klein *et al.*, "The Shape of History: Elizabeth Palmer Peabody's Feminist Visualization," *Feminist Media Histories*, 2017, 3:149–153; Alberto Cairo, *How Charts Lie: Getting Smarter about Visual Information* (New York: Norton, 2019); and Kevin Kee and Timothy Compeau, eds., *Seeing the Past with Computers: Experiments with Augmented Reality and Computer Vision* (Ann Arbor: Univ. Michigan Press, 2019).

Step 6: Reflecting

If we choose to expand this project, we could parse the same data in many different ways. We could determine whether certain branches of science or certain national styles were more prone to holistic fervor. We could chart citation networks within the corpus, and we could ask why practitioners in radically different fields developed such strikingly similar ideas. We could ask whether the inclusion of political organisms changes how we understand "organism choice."[23] We could expand beyond JSTOR, adding thousands of OCR-enabled books and articles from the Internet Archive, HathiTrust, and the Biodiversity Heritage Library. We could build similar corpora in different languages by clicking on "View Full Text" within these resources and then letting Google Chrome's auto-translate widget do the rest. For the time being, these potential projects would require one to input metadata for thousands of articles manually, which means that it is unlikely to happen without additional soft money.

The final step of the workflow invites one to reflect not only on the project at hand but also on the profession at large. Accordingly, we spend the remainder of this essay reflecting on the significance of computational methods for HPS more generally.

III

There is an increasingly popular trope that refers to the "light side" and the "dark side" of the digital humanities.[24] The positive aspects of digital humanities—the light side—are obvious to most people. For example, nearly everyone in the field promotes open access. The most popular textbooks in the digital humanities are written and published entirely online.[25] Primary and secondary sources are routinely scanned and shared on the internet for free. Open and equal access to data has a democratizing effect, which expands the canon and diversifies the guild.

Computational methods also place historians in league with strange company, but most of us agree that this is a light-side item. United by a shared methodology rather than shared interests, historians who use computational methods are as likely to cite physicists, geneticists, and librarians as fellow historians. We are all part of a larger "computational turn" that is affecting every branch of knowledge. These methods have proven especially popular among literary scholars, although this wholesale embrace is not without controversy. Meanwhile, scientists are also turning computational tools on themselves, analyzing the scientific enterprise from its broadest contours to its finest details, seeking a "science of science."[26] In each case, these fields rely on the same two ingredients: voluminous data and computational power.

[23] B. R. Erick Peirson *et al.*, "The Diversity of Experimental Organisms in Biomedical Research May Be Influenced by Biomedical Funding," *Bioessays*, 2017, 39:1–9; Nathan Crowe *et al.*, "The Diversification of Developmental Biology," *Studies in History and Philosophy of Biological and Biomedical Sciences*, 2015, 53:1–15; and Michael R. Dietrich, Rachel A. Ankeny, and Patrick M. Chen, "Publication Trends in Model Organism Research," *Genetics*, 2014, 98:787–794.

[24] Brian Greenspan, "Are Digital Humanists Utopian?" in *Debates in the Digital Humanities* (2016), pp. 393–409; Wendy Hui Kyong Chun and Lisa Marie Rhody, "Working the Digital Humanities: Uncovering Shadows between the Dark and the Light," *Differences: A Journal of Feminist Cultural Studies*, 2014, 25:1–26; and Chun *et al.*, "The Dark Side of the Digital Humanities," in *Debates in the Digital Humanities* (2016), pp. 493–509.

[25] For work promoting open access see Jane Maienschein *et al.*, "Data Management and Data Sharing in Science and Technology Studies," *Science, Technology, and Human Values*, 2019, *44*:143–160; Mark D. Wilkinson *et al.*, "The FAIR Guiding Principles for Scientific Data Management and Stewardship," *Scientific Data*, 2016, 3, article 160018; Jeffrey M. Perkel, "A Toolkit for Data Transparency," *Nature*, 2018, *560*:513–515; and Morgan Taschuk and Greg Wilson, "Ten Simple Rules for Making Research Software More Robust," *PLOS Computational Biology*, 2017, *13*:1–10. Among the online textbooks in the field see Gold and Klein, eds., *Debates in the Digital Humanities* (2016); Graham *et al.*, *Exploring Big Historical Data* (cit. n. 9); Gold, ed., *Debates in the Digital Humanities* (2012); Daniel J. Cohen and Roy Rosenzweig, *Digital History: A Guide to Gathering, Preserving, and Presenting the Past on the Web* (Philadelphia: Univ. Pennsylvania Press, 2005); and Gold and Klein, eds., *Debates in the Digital Humanities* (cit. n. 14).

[26] Dennis Tenen, "Blunt Instrumentalism: On Tools and Methods," in *Debates in the Digital Humanities* (2016), pp. 83–91; and Santo Fortunato *et al.*, "Science of Science," *Science*, 2018, 359(6379):eaao0185. On the popularity of computational methods

If we are having a forthright discussion about the light side of computational history, then we should be equally frank about its dark side. Some worry about the fate of our narrative craft when historical documents are divorced from their context. Others have expressed concern about overabundance.[27] Given the glut of data, it is important to remember that the world has not been digitized evenly or equally and that the types of sources we digitize influence the types of stories we tell. Historians are going to have to work harder to make sure less visible histories do not get swamped. And yet, despite this excess, data is more transient than ever. Digital artifacts are surprisingly fragile, subject to corruption, malfunction, and the "dangerous potency" of the delete key. Broken links and deepfakes threaten to eat our evidentiary base from under us. Threats of disasters and cyberwarfare loom in the background. Given this transience, deciding what to save becomes a big decision. Meanwhile, many university libraries destroy physical records as a matter of course. Some have not bought a physical book in years.[28]

Sometimes it is not clear if a given development belongs on the light side, the dark side, or somewhere in between. Consider the case of "cliodynamics," which purports to treat history like a science. Established by Peter Turchin, the field uses mathematical modeling to identify and examine large-scale patterns of complexity that emerge in different places and at different times and thus helps elucidate the "first laws of history."[29] Last year, Turchin and his colleagues published an article in the *Proceedings of the National Academy of Sciences* in which they claimed to uncover a single dimension of complexity structuring global variation in human social organizations throughout history.[30] Though tantalizing, these claims should raise two red flags. First, these laws of history would apply to the past *and* the future. This might allow for prognostication, but it could also serve as a pretext for social control. Second, but no less important, the field of cliodynamics readily cedes analysis of the historical record to computers. This raises important questions about whether the historical profession can be automated and whether there is any place for history in the future.

in literary studies see Ted Underwood, "Dear Humanists: Fear Not the Digital Revolution," *Chron. Higher Educ.*, 27 Mar. 2019; Underwood, *Distant Horizons: Digital Evidence and Literary Change* (Chicago: Univ. Chicago Press, 2019); Joseph P. Dexter *et al.*, "Quantitative Criticism of Literary Relationships," *Proceedings of the National Academy of Sciences*, 2017, *114*:e3195–e3204; Stephen Robertson, "The Differences between Digital Humanities and Digital History," in *Debates in the Digital Humanities* (2016), pp. 289–307; Stephen Ramsay, *Reading Machines: Toward an Algorithmic Criticism* (Urbana: Univ. Illinois Press, 2011); Jean-Baptiste Michel *et al.*, "Quantitative Analysis of Culture Using Millions of Digitized Books," *Science*, 2011, *331*:176–182; and Franco Moretti, *Graphs, Maps, Trees: Abstract Models for Literary History* (London: Verso, 2005). For works that criticize computational literary studies see Nan Z. Da, "The Digital Humanities Debacle," *Chron. Higher Educ.*, 27 Mar. 2019; and Da, "The Computational Case against Computational Literary Studies," *Critical Inquiry*, 2019, *45*:601–639.

[27] For worries about narrative craft see Laura K. Nelson, "Computational Grounded Theory: A Methodological Framework," *Sociological Methods and Research*, Nov. 2017, https://journals.sagepub.com/doi/full/10.1177/0049124117729703; Toni Weller, "Introduction," in *History in the Digital Age*, ed. Weller (London: Routledge, 2012), pp. 1–20, esp. p. 8; and Rebecca Onion, "Snapshots of History," slate.com, 5 Feb. 2014. Regarding concerns about overabundance see Ian Milligan, *History in the Age of Abundance? How the Web Is Transforming Historical Research* (Montreal: McGill-Queens Univ. Press, 2019); and Roy Rosenzweig, "Scarcity or Abundance? Preserving the Past in a Digital Era," *Amer. Hist. Rev.*, 2003, *108*:735–762.

[28] Putnam, "Transnational and the Text-Searchable" (cit. n. 17), pp. 389–390 (uneven digitization); Weller, "Introduction," pp. 4, 11 (quotation); and Andrew Stauffer, "My Old Sweethearts: On Digitization and the Future of the Print Record," in *Debates in the Digital Humanities* (2016), pp. 218–229 (transience of physical records).

[29] Peter Turchin, *Historical Dynamics: Why States Rise and* Fall (Princeton, N.J.: Princeton Univ. Press, 2003); and "How the New Science of Computational History Is Changing the Study of the Past," *MIT Technology Review*, 23 June 2016.

[30] Peter Turchin *et al.*, "Quantitative Historical Analysis Uncovers a Single Dimension of Complexity That Structures Global Variation in Human Social Organization," *Proc. Nat. Acad. Sci.*, 2018, *115*:e144–e151. See also Nick Tosh, John Ferguson, and Cathal Seoighe, "History by the Numbers?" *ibid.*, p. e5840; and Thomas E. Currie, "Quantitative Analyses of Cultural Evolution Require Engagement with Historical and Archaeological Research," *ibid.*, pp. e5841–e5842.

Researchers are already using artificial intelligence (AI) to identify connections among millions of different articles, and they are already using AI to peer-review article manuscripts.[31] Some of these developments will apply to all disciplines, and some will be tailored to history. For example, one recent study showed that all the major elements of storytelling, the 5 W's + H (who, what, when, where, why, and how), can now be automated. Some experts predict that AI will be able to write bestselling books within the next thirty years, but, given the uncanny (and somewhat unnerving) quality of OpenAI's recently released text generator, that estimate may well prove conservative.[32] The ultimate goal, according to Father Busa, has always been the "automation of every possible analysis of human expression."[33] Some object that ceding our birthright to algorithms will spell the end of history, but that is inaccurate. It would not mean the end of history, only the end of historians.

These are sobering prospects, but at least we are not going it alone. Computers are reshaping every profession on Earth. Experts predict that we will be capable of automating half of the jobs in the United States within the next ten to twenty years. Beyond that, no job is safe from AI. President Obama acknowledged as much in his farewell address to the nation. "The next wave of economic dislocation won't come from overseas," he warned. "It will come from the relentless pace of automation that makes many good, middle-class jobs obsolete." He cited the need for a "new social compact" that would help us navigate, and survive, rough waters ahead.[34] Researchers are already drafting first principles.[35]

As historians of science and technology, we are uniquely qualified to help contextualize these developments and to help our fellow citizens make sense of it all.[36] Our blind faith in algorithms certainly warrants closer inspection. As numerous scholars have shown, algorithms are sociotechnical systems. They are cultural artifacts with "knobs and dials" that can be tweaked to achieve subtly different results.[37] Our best hope as a profession—our best hope as a society—lies in our ability to acknowledge and overcome these algorithmic biases. It is noteworthy that many (most?) of the leading voices in the closely related fields of "algorithmic accountability," "algorithmic justice," and "critical code studies" belong to women and minority scholars—groups that have been underrepresented in software development.[38] These scholars have demonstrated that structures of

[31] Sampo Pyysalo, "LION-LBD: A Literature-Based Discovery System for Cancer Biology," *Bioinformatics*, 2018, 35:1–9; and Douglas Heaven, "The Age of AI Peer Reviews," *Nature*, 2018, 563:609–610.

[32] Franzosi, "Third Road to the Past?" (cit. n. 9); Katja Grace *et al.*, "When Will AI Exceed Human Performance? Evidence from AI Experts," arXiv:1705.08807v3 [cs.AI], 3 May 2018; and Alec Radford *et al.*, "Language Models Are Unsupervised Multitask Learners," OpenAI, 14 Feb. 2019.

[33] Steven E. Jones, "The Emergence of the Digital Humanities (as the Network Is Everting)," in *Debates in the Digital Humanities* (2016), pp. 3–15 (quoting Busa).

[34] Kai-Fu Lee, *AI Superpowers: China, Silicon Valley, and the New World Order* (New York: Houghton, Mifflin & Court, 2018); David Kaufman, "Watch Out Workers, Algorithms Are Coming to Replace You—Maybe," *New York Times*, 18 Oct. 2018; and Barack Obama, "Farewell Address to the Nation," 10 Jan. 2017.

[35] Meredith Whittaker *et al.*, *AI Now Report 2018* (New York: New York Univ., 2018).

[36] Ivan Flis, Evina Steinová, and Paul Wouters, "Digital Humanities Are a Two-Way Street," *Isis*, 2016, *107*:346–348, esp. p. 348.

[37] Perkel, "Toolkit for Data Transparency" (cit. n. 25), p. 513. See also Brent Daniel Mittelstadt *et al.*, "The Ethics of Algorithms: Mapping the Debate," *Big Data and Society*, July–Dec. 2016, 6:1–21; Nick Seaver, "Algorithms as Culture: Some Tactics for the Ethnography of Algorithmic Systems," *ibid.*, July–Dec. 2017, 20:1–12; Seaver, "What Should an Anthropology of Algorithms Do?" *Cultural Anthropology*, 2018, 33:375–385, esp. p. 385; and Hallam Stevens, "A Feeling for the Algorithm: Working Knowledge and Big Data in Biology," *Osiris*, 2017, 32:151–174, esp. p. 173.

[38] Safiya Umoja Noble, *Algorithms of Oppression: How Search Engines Reinforce Racism* (New York: NYU Press, 2018); Virginia Eubanks, *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor* (New York: St. Martin's, 2018); Roopika Risam, *New Digital Worlds: Postcolonial Digital Humanities in Theory, Praxis, and Pedagogy* (Chicago: Northwestern Univ. Press, 2018); Marie Hicks, *Programmed Inequality: How Britain Discarded Women Technologists and Lost Its Edge in Computing* (Cambridge, Mass.: MIT Press, 2017); Cathy O'Neill, *Weapons of Math Destruction: How Big Data Increases Inequality*

power and privilege are easily reproduced in digital humanities. They remind us that *we* are big data, that our web searches, financial transactions, and interpersonal relationships are subject to quantification and manipulation. Lara Putnam was referring to the historical record when she used the word "data-fication," but the exact same process applies to each of us.[39]

On the basis of our experiences with computational HPS and digital humanities, we can offer the following suggestions to other historians. First, we should learn to embrace collaboration. Many historians still believe in the "monastic ideal of scholarship," which emphasizes years of solitary contemplation, but digital history is a group effort.[40] It is not always going to be easy. For example, every word of this brief article was hotly debated by its two coauthors. Being of two minds, however, we embrace the holistic ideal that is afforded by numerous authors with diverse perspectives. Second, we should reject our Luddite impulses. Innumerable studies from every corner of the labor force suggest that the highest results are achieved when skilled workers utilize advanced tools like AI and machine learning. Historians of science are no different.[41] Finally, we should embrace the field's revolutionary potential, but we should not surrender what makes us unique. It is obvious that computers can analyze historical datasets, but they cannot yet interpret them. Therefore, we must demand a place for humans in the digital humanities.[42]

and Threatens Democracy (New York: Broadway, 2016); Joy Buolamwini, "When the Robot Doesn't See Dark Skin," *New York Times*, 21 June 2018; Bethany Nowviskie, "Resistance in the Materials," in *Debates in the Digital Humanities* (2016), pp. 176–183; Meredith Broussard, *Artificial Unintelligence: How Computers Misunderstand the World* (Cambridge, Mass.: MIT Press, 2018); danah boyd and Kate Crawford, "Critical Questions for Big Data," *Information, Communication, and Society*, 2012, *15*:662–679; and Joy Rankin, *A People's History of Computing in the United States* (Cambridge, Mass.: Harvard Univ. Press, 2018).

[39] Putnam, "Transnational and the Text-Searchable" (cit. n. 17), p. 400. See also Jacqueline Wernimont, *Numbered Lives: Life and Death in Quantum Media* (Cambridge, Mass.: MIT Press, 2019); Elizabeth Losh and Wernimont, eds., *Bodies of Information: Feminist Debates in Digital Humanities* (Minneapolis: Univ. Minnesota Press, 2019); Casey Fiesler and Blake Hallinan, "'We Are the Product': Public Reactions to Online Data Sharing and Privacy Controversies in the Media," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (New York: Association for Computing Machinery, 2018), Paper 53; and Gina Neff and Dawn Nafus, *Self-Tracking* (Cambridge, Mass.: MIT Press, 2016).

[40] Vincent Brown, "Narrative Interface for New Media History: *Slave Revolt in Jamaica, 1760–1761*," *Amer. Hist. Rev.*, 2016, *121*:176–186, on p. 184.

[41] Kellen Funk and Lincoln A. Mullen, "The Spine of American Law: Digital Text Analysis and U.S. Legal Practice," *Amer. Hist. Rev.*, 2018, *123*:132–164, esp. p. 162.

[42] Elizabeth Losh *et al.*, "Putting the Human Back into the Digital Humanities: Feminism, Generosity, and Mess," in *Debates in the Digital Humanities* (2016), pp. 92–103.