Chapter 1

The Joys of Big Data for Historians

In this chapter, we look at the emergence of the idea of "big data" for historians, examining some case studies in the broader field of the digital humanities. We discuss the limits of big data in terms of historical practice and make an argument for why all of us — whether we employ the methods discussed in this book or not — need to be aware of why this matters.

A macroscope is a bit like a microscope or a telescope, but instead of allowing you to see things that are small or far away, the macroscope makes it easier to grasp the incredibly large. It does so through a process of compression, by selectively reducing complexity until once-obscure patterns and relationships become clear. Often, macroscopes produce textual abstractions or data visualizations *in lieu* of direct images.[11]

Pointed at human history, the macroscope offers a stark contrast to what has become standard historical practice. Rather than through compression, good historians, like good detectives, test their merit through expansion: the ability to extract complex knowledge from the smallest crumbs of evidence that history has left behind. By tracing the trail of these breadcrumbs, a historian might weave together a narrative of the past. A historian's macroscope offers a complementary,

[11] On visualizations and abstractions, Lorraine Daston has written "these techniques aim at more than making the invisible visible. They aspire to all-at-once-ness, the condensation of laborious, step-by-step procedures into an immediate *coup d'oeil*, …. What was a painstaking process of calculation and correlation — for example, in the construction of a table of variables — becomes a flash of intuition." See Lorraine Daston, "On Scientific Observation," *Isis* 99.1, March 2008, 97–110. This is the goal of the macroscope: to highlight immediately what often requires careful thought and calculation, sometimes more than is possible for a single person.

but very different, path to knowledge. It allows you to begin with the complex and winnow it down until a narrative emerges from the cacophony of evidence.[12]

In this, an important distinction needs to be made between our understanding of the micro- and macroscope versus micro- and macro-history; the two do not dovetail perfectly together. Microhistory involves the rigorous and in-depth study of a single story or moment in history, whereas macrohistory susses out long-term trends and eddies, such as Fernand Braudel's longue durée. A macroscope, by studying large quantities of data, could fit into microhistory; imagine the parsing of hundreds of thousands of tweets around a single American presidential debate, for example. Yet it also helps us understand macrohistory: tracing fluctuating word and topic frequency over decades, or even centuries. Macroscopes are not bound by time, but rather quantity. They also draw upon a rich heritage of computational history, stretching back to early cutting-edge work done with censuses in the 1960s.[13]

As history becomes digitized in ever-increasing scales, historians without the ability to research both micro- and macroscopically may be in danger of becoming mired in evidence or lost in the noise. This book is aimed at those historians who aspire to turn the macroscope on their own research, an increasingly important skill in our historical moment.[14] It is neither exhaustive in scope nor terribly deep in any one methodology; instead, we have filled this book with the first

[12] We are neither the first to coin the term macroscope, nor the first to point its gaze toward human history. It was originally used by Joël de Rosnay (1979), *The Macroscope: A New World Scientific System* (Harper & Row, New York, NY) to discuss complex societies and has most recently been popularized by Kay Börner (March 1, 2011) "Plug-and-Play Macroscopes." *Communications of the ACM*, 54(3), 60, in the context of tools that allow one to see human activities at a distance. Macroscopes have been brought up in the humanities as well, such as by Tim Tangherlini (see his 'Tracking Trolls: New Challenges from the Folklore Macroscope, eHg Annual Lecture, Royal Netherlands Academy of Arts and Sciences, December 12, 2013, https://www.ehumanities.nl/ehg-annual-lecture-tim-tangherlini-ucla/) to similar effect. In literary criticism and history, similar concepts have been called "distant reading," as articulated by Franco Moretti (2005), *Graphs, Maps, Trees: Abstract Models for Literary History* (London, New York, NY: Verso,) or "macroanalysis" in Matthew L. Jockers (2013), *Macroanalysis: Digital Methods & Literary History* (Urbana-Champaign, IL: University of Illinois Press). In *Uncharted: Big Data as a Lens on Human Culture* (2013, New York, NY: Riverhead,) on culturomics, Erez Aiden and Jean-Baptiste Michel wrote of their intent "to build a scope to study human history."

[13] Chad Gaffield (2020), 'Clio and Computers in Canada and Beyond: Contested Past, Promising Present, Uncertain Future' *The Canadian Historical Review* 101(4) pp. 559–584, https://doi.org/10.3138/chr-2020-0020.

[14] Jo Guldi and David Armitage effectively argue the importance of macroscopic thinking in their monograph *The History Manifesto* (2014, Cambridge: Cambridge University Press).

stepping-stones for many different roads. We hope interested readers will follow those paths into areas yet unexplored.

Big Data

"How big is big?" we rhetorically ask: big data for literature scholars might mean a hundred novels ("the great unread"),[15] for historians it might mean an entire array of 19th century shipping rosters,[16] and for archaeologists it might mean every bit of data generated by several seasons of field survey and several seasons of excavation and study — the materials that *don't* go into the GIS. For computer scientists, they are often focused not just on materials of a scope that can't be read but on volumes of information that elude processing by conventional computer systems, such as large collections of multimedia websites or the shocking amount of information generated by experiments such as CERN's Large Hadron Collider.

For us, as humanists, *big is in the eye of the beholder*. If there are more data than you could conceivably read yourself in a reasonable amount of time, or that require computational intervention to make new sense of them, it's big enough! These are all valid answers. Indeed, there is some valid hesitancy around the use of the term "data" itself, as it has a faint whiff of quantifying and reducing the meaningful life experiences of the past to numbers. We believe that this book outlines various methods to computationally explore historical data in a way that would previously seem resistant to qualification. With that proviso in mind, of course, we do take "big data" to be a central concept in this book.

As scholars, we have all used varying degrees of datasets and considered them "big." On one extreme lie the datasets that stretch the constraints of individual researchers and personal computing, including large web-scale datasets, such as the 80-terabyte sweep of much of the publicly accessible World Wide Web in 2011, made available by the Internet Archive.[17] Such projects may require high performance computing and specialized software. But others work with more manageable sets of data: popular music lyrics, proposals or papers

[15] See Matthew L. Jockers (2013), *Macroanalysis: Digital Methods & Literary History* (Urbana-Champaign, IL: University of Illinois Press), and Margaret Cohen (1999), *The Sentimental Education of the Novel*, Princeton, NJ: Princeton University Press.

[16] See Trading Consequences (2014), "Trading Consequences | Exploring the Trading of Commodities in the 19th Century," http://tradingconsequences.blogs.edina.ac.uk/.

[17] Internet Archive (October 26, 2012), "80 Terabytes of Archived Web Crawl Data Available for Research," *Internet Archive Blog*, http://blog.archive.org/2012/10/26/80-terabytes-of-archived-web-crawl-data-available-for-research/.

submitted to conferences, databases of dissertations, historiographical inquiries, correspondence, oral history interviews, and beyond. While much early work focused on textual data, the "visual turn" within the digital humanities increasingly leverages machine learning and other techniques for working with visual data at scale. In any case, for us, big data is simply more data that you could conceivably read yourself in a reasonable amount of time or, even more inclusively, information that requires, or can be read with, computational intervention to make new sense of it.

Big Data analysis skills are on the verge of no longer being a "nice to have" for historians but nearly a necessity. Historians must be open to the digital turn, thanks to the astounding growth of digital sources and an increasing technical ability to process them on a mass scale.[18] Both trends are discussed in this book. Historians are collectively witnessing a profound transformation in how they research, write, disseminate, and interact with their work. As datasets expand into the realm of the big, computational analysis ceases to be "nice to have" and becomes a requirement. While not all historians will have to become fluent with data (just as not all historians are oral historians, or use GIS, or work within communities), digital historians will become part of the disciplinary mosaic. Computational skills may be increasingly viewed as akin to language requirements: in some cases, a nice-to-have and in other programs, a serious requirement to graduate.

In this chapter, we introduce you to our definition of big data, what opportunities it affords, where it came from, and the broader implications of this "era of big data." New and emerging research tools are driving cutting-edge humanities research, often funded by transnational funding networks. Historians are asking new questions of old datasets with new tools, as well as finding new avenues on previously inaccessible terrain. After this survey of the current state of affairs, we then turn our eyes to the historical context of this current scholarly moment. Here we see the joys of abundance but also the dangers of information overload.[19] The contours of this challenge and opportunity are fascinating and help anchor the discussion that follows. While there will certainly be bumps on the road ahead to come, we generally see a promising future for an era of big data.

[18] The foundational text in this field is Daniel J. Cohen and Roy Rosenzweig (2005). *Digital History: A Guide to Gathering, Preserving, and Presenting the Past on the Web*. This work primarily focuses on putting information on the web, whereas we explore various ways to improve your research with that information.

[19] Roy Rosenzweig (2003), "Scarcity or Abundance? Preserving the Past in a Digital Era," *American Historical Review*, 108(3), 735–762, available online at http://chnm.gmu.edu/digitalhistory/links/pdf/introduction/0.6b.pdf.

Putting Big Data to Good Use: Historical Case Studies

For 239 years, men, women, and children (initially Londoners, subsequently all Britons accused of major trials) stood for justice before judges at the Old Bailey. Thanks to a popular fascination with crime amongst Londoners, the events were recorded in the *Proceedings*: initially as pamphlets, later as bound books and journals replete with advertisement, and subsequently becoming official publications. These sources provide an unparalleled look into the administration of justice and the lives of ordinary people in 17th, 18th, and 19th century England (their publication ceased in 1913).[20] While there are gaps, this is an exhaustively large source for historians using customary methods who sought to answer questions of social and legal history in the 20th and 21st centuries.

The Old Bailey records are, by the standards of historians and most humanists, *big data*. They comprise 127 million words, cover 197,000 trials, and have been transcribed to a high standard by two typists working simultaneously to reduce error rates. Computers can read this amount of information quickly, but it would take years for a single scholar to read this (and by then they probably would have forgotten half of what they had read). As the research team put it: "it is one of the largest bodies of accurately transcribed historical text currently available online. It is also the most comprehensive thin slice view of eighteenth and nineteenth-century London available online."[21] Tackling a dataset of this size, however, requires specialized tools. Once digitized, it was made available to the public through keyword searches. Big data methodologies, however, offered new opportunities to make sense of this very old historical material.[22]

The *Data Mining with Criminal Intent* project sought to do that.[23] A multinational project team, including scholars from the United States, Canada, and the

[20] "The Proceedings — Publishing History of the Proceedings — Central Criminal Court," April 2013, http://www.oldbaileyonline.org/static/Publishinghistory.jsp.

[21] Dan Cohen *et al*. (August 31, 2011), *Data Mining with Criminal Intent: Final White Paper*, originally at http://criminalintent.org/wp-content/uploads/2011/09/Data-Mining-with-Criminal-Intent-Final1.pdf; now at https://pdfs.semanticscholar.org/80d1/bfb2d186097ca83353d81d254c1e3fa2151d.pdf. We provide a short url: https://tinyurl.com/criminal-intent-whitepaper. Sharon Howard maintains a set of digital history projects that focus on the 're-use, enhancement and exploration' of data from the Old Bailey Online and the London Lives 1690-1800 projects, at https://london.sharonhoward.org. These, and the related projects at https://sharonhoward.org/digital-history-projects.html are strongly recommended.

[22] The best explanation of this is probably the Piled Higher and Deeper (PhD Comics) video explaining British historian Adam Crymble's work. See "Big Data + Old History," YouTube video, 6 September 2013, http://www.youtube.com/watch?v=tp4y-_VoXdA.

[23] For an incredible overview of this, see Tim Hitchcock (9 December 2013), "Big Data for Dead People: Digital Readings and the Conundrums of Postivism," *Historyonics Blog*, http://historyonics.blogspot.ca/2013/12/big-data-for-dead-people-digital.html.

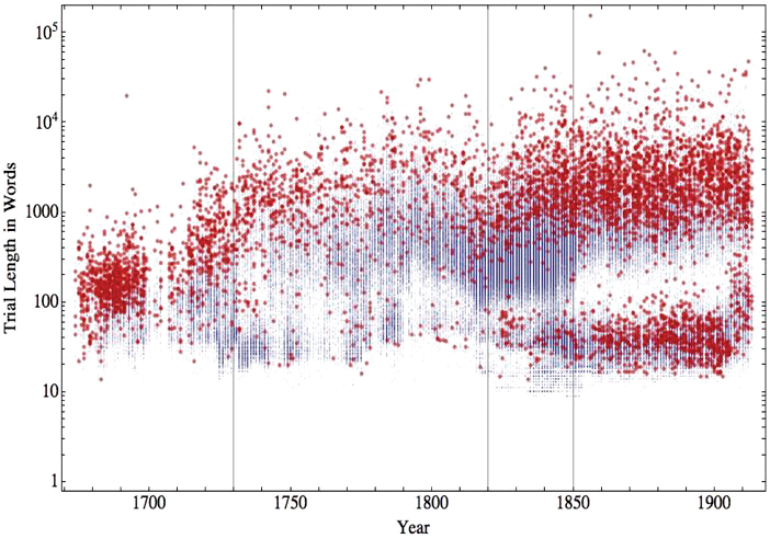

Fig. 1.1 The length of trials, in words, in Old Bailey Proceedings. Copyright Tim Hitchcock and William J. Turkel, 2011. Used with permission.

United Kingdom, sought to create "a seamlessly linked digital research environment" for working with these court records. Pulling their gaze back from individual trials and records, the team made new and relevant discoveries. Using a plugin (a little program or component that adds something to a software program) for the open source reference and research management software *Zotero* (http://www.zotero.org/), Fred Gibbs at George Mason University developed a means to look at specific cases (e.g. those pertaining to "poison") and look for commonalities. For example, "drank" was a common word; closer to that, the word "coffee"; conversely, the word "ate" was conspicuously absent. In another development, closely related trials could be brought up; through comparing differences in documents (using Normalized Compression Distance, or the standard tools that compress files on your computer) one can get the database to suggest trials that are structurally similar to the one a user is currently viewing. And, most importantly, taking this material and visualizing it allowed the research team to discover new findings, such as a "significant rise in women taking on other spouses when their husbands had left them," or the rise of plea bargaining around 1825, based on the figure below (Fig. 1.1) showing a dramatic shift in the length of trials in that time period.[24]

[24] Dan Cohen *et al*. (August 31, 2011), *Data Mining with Criminal Intent: Final White Paper*, https://pdfs.semanticscholar.org/80d1/bfb2d186097ca83353d81d254c1e3fa2151d.pdf; short url: https://tinyurl.com/criminal-intent-whitepaper.

All of these tools were put online, and researchers can now access the *Old Bailey* both through the traditional portal as before and through a dedicated Application Programming Interface or API. APIs are a set of rules or specifications that let different software programs talk to each other. In the case of the *Old Bailey*, the API allows researchers to use programs like Zotero, or their programming language of choice, or visualization tools such as the freely accessible Voyant Tools, to access the database. The project thus leaves a legacy to future researchers, to enable them to point their macroscope toward the trials, to make sense of that exhaustive dataset of 127 million words.

One eye-opening study that emerged out of this massive undertaking — "The Civilizing Process in London's Old Bailey" by Sara Klingenstein, Tim Hitchcock, and Simon DeDeo — made a provocative argument that could only have been made via computational intervention. The researchers used an innovative method of creating "bags of words" out of related words from *Roget's Thesaurus* and were able to convincingly demonstrate a major shift that occurred in the early 19th century: violent crimes began to be both discussed and treated differently. Before that time, while the crime itself was mentioned, the violence within it was not a discrete category. The work helps confirm the hypothesis of others, which sees a "civilizing process" as a:

> Deep-rooted and multivalent phenomenon that accompanied the growing monopoly of violence by the state, and the decreasing acceptability of interpersonal violence as part of normal social relations. Our work [in this article] is able to track an essential correlate of this long-term process, most visibly in the separation of assault and violent theft from nonviolent crimes involving theft and deception.[25]

As the *New York Times* explained in their coverage of the study, the routine nature of violence in the 1700s gave way by the 1820s to an emphasis on "containing violence, a development reflected not just in language but also in the professionalization of the justice system."[26] Distant reading enabled historians to pull their gaze back from the individual trials and to consider the 197,000 trials as a whole. The "civilizing process" study will probably, in our view, emerge as

[25] Sara Klingenstein, Tim Hitchcock, and Simon DeDeo (2014), "The Civilizing Process in London's Old Bailey," *Proceedings of the National Academy of Sciences U. S. A.*, **111**(26), 9419–9424. Available online at http://www.pnas.org/content/111/26/9419.full.

[26] Sandra Blakeslee (June 16, 2014), "Computing Crime and Punishment," *The New York Times*, June 16, 2014, http://www.nytimes.com/2014/06/17/science/computing-crime-and-punishment.html.

one of the defining stories of how big data can reveal new things that were not possible without it.

The high quality of the *Old Bailey* dataset is, however, an outlier: most researchers will not have ready access to nearly perfectly transcribed databases like these criminal records. Even digitized newspapers, which on the face of it would seem to be excellent resources, are not without serious issues at the level of the optical character recognition (OCR).[27] There is still joy to be found, however, in the less-than-perfect records. The *Trading Consequences* project is one example of this. Bringing together leading experts in natural language processing (NLP) and innovative historians in Canada and the United Kingdom, this project confirmed and enhanced existing understandings of trade by tracing the relationships between commodities and global locations. They adapted the Edinburgh Geoparser[28] to find and locate place names in text files, and then identified commodities that were mentioned in relationship with these place names. One useful example the project gave us was that of cinchona bark, a raw material used to make quinine (an anti-malaria drug) — it was a useful example because it allowed them to confirm the secondary literature on research questions around the trade of cinchona. They confirmed via computational analysis that until the 1850s it was closely related to Peruvian forests as well as markets and factories in London, England; by the 1860s, it had shifted towards locations in India, Sri Lanka, Indonesia, and Jamaica. By drawing on thousands of documents, changes over time can be quickly spotted. Even more promisingly, it can be put online in database fashion, enabling historians to test out their own hypotheses and findings. There are several ways to access their database, including searches on specific commodities and locations, and it can all be found at http://tradingconsequences.blogs.edina.ac.uk/access-the-data/. For historians interested in this form of work, their project white paper is of considerable utility.[29]

Moving beyond anecdotes and case studies, the team drew on over six million British parliamentary paper pages, almost four million documents from Early Canadiana Online,[30] and smaller series of letters, British documents ("only"

[27] Ian Milligan (2013), "Illusionary Order: Online Databases, Optical Character Recognition, and Canadian History, 1997–2010," *Canadian Historical Review*, 94(4), 540–569.

[28] Beatrice Alex, Kate Byrne, Clare Grover, Ricahrd Tobin (2016) *The Edinburgh Geoparser* https://www.ltg.ed.ac.uk/software/geoparser/ . Beatrice Alex has also published a tutorial for using the geoparser at *The Programming Historian* https://programminghistorian.org/en/lessons/geoparsing-text-with-edinburgh.

[29] Ewan Klein *et al.* (March 2014), *Trading Consequences: Final White Paper*, http://tradingconsequences.blogs.edina.ac.uk/files/2014/03/DiggingintoDataWhitePaper-final.pdf.

[30] Now *Canadiana Online* http://www.canadiana.ca/

140,010 images, for example), and other correspondences. This sort of work presages the growing importance of linked open data, which points towards the increasing trend in putting information online in a format that other computer programs can read. In this way, Early Canadiana Online can speak to files held in another repository, allowing us to harness the combined potential of many numerous silos of knowledge.

An automated process would take a document, turn the image into accessible text, and then "mark it up:" London, for example, would be marked as a "location," grain would be marked as a "commodity," and so forth. The interplay between trained programmers, linguists, and historians was especially fruitful for this: for example, the database kept picking up the location "Markham" (a suburb north of Toronto, Canada) and historians were able to point out that the entries actually referred to a British official, culpable in smuggling, Clements Robert Markham. As historians develop technical skills and computer scientists develop humanistic skills, fruitful collaborative undertakings can develop. Soon, historians of this period will be able to contextualize their studies with an interactive, visualized database of global trade in the 19th century. Yet, while collaborative outcomes are laudable when dealing with *big* questions, we still believe that there is a role for the sole or small group undertaking. We personally find that exploratory research can be fruitfully carried out by a sole researcher, or in our case, four of us working on a problem together. There's no one size of a team that will fit all questions.[31]

Beyond court and trading records, census recordings have long been a staple of computational inquiry. In Canada, for example, there are two major and ambitious projects underway that use computers to read large arrays of census information. The Canadian Century Research Infrastructure project, funded by federal government infrastructure funding, draws on five censuses of the entire country in an attempt to provide a "new foundation for the study of social, economic, cultural, and political change."[32] Simultaneously, researchers at the Université de Montréal are reconstructing the European population of Quebec in the 17th and 18th centuries, drawing heavily on parish registers.[33] This form of history harkens

[31] The topic of equitable and effective project management in digital history should be considered at the outset. We suggest beginning by reading the discussion and considering the artefacts collected by Lynne Siemens (nd) 'Project Management' *Digital Pedagogy in the Humanities: Concepts, Models, and Experiments* https://digitalpedagogy.mla.hcommons.org/keywords/project-management/

[32] For more information, see the Canadian Century Research Infrastructure Project website at http://www.ccri.uottawa.ca/CCRI/Home.html.

[33] See the Programme de recherche en démographie historique at http://www.genealogy.umontreal.ca/fr/LePrdh.

back to the first wave of computational research, discussed later in this chapter, but shows some of the potential available to historians computationally querying large datasets.

Any look at textual digital history would be incomplete without a reference to the Culturomics Project and Google Ngrams.[34] Originally co-released as an article and an online tool, a team from Harvard University (the composition of which is discussed shortly) collaborated to develop a process for analyzing the millions of books that Google has scanned and applied OCR to as part of its Google Books project. This project indexed word and phrase frequency across over five million books, enabling researchers to trace the rise and fall of cultural ideas and phenomena through targeted keyword and phrase searches and their frequency over time.[35] The result is an uncomplicated but powerful look at a few hundred years of book history. One often unspoken tenet of digital history is that very simple methods can produce incredibly compelling results, and the Google Ngrams tool exemplifies this idea.[36] In terms of sheer data, this is the most ambitious and certainly the most widely accessible (and publicized) big history project in existence. Ben Zimmer used this free online tool to show when the United States began being discussed as a singular entity rather than as a plurality of many states by charting when people stopped saying "The United States are" in favor of "The United States is" (Fig. 1.2):[37]

This is a powerful finding, both confirmatory of some research and suggestive of future paths that could be pursued. There are limitations, of course, with such a relatively simple methodology: words change in meaning and typographical characteristics over time (compare the frequency of *beft* with the frequency of *best* to see the impact of the "medial s"), there are OCR errors, and searching only on words or search phrases can occlude the surrounding context of a word. Some of the hubris around culturomics rankled some historians,[38] but taken on its own merits, the Culturomics Project and the Ngram viewer have done wonders for

[34] Michel *et al*. (2011), "Quantitative Analysis of Culture Using Millions of Digitized Books," *Science*, **331**, 6014, and http://books.google.com/ngrams.

[35] Remember, to do a phrase search, use quotation marks around your phrase. Otherwise, the search will return results where the words are in close proximity, which can skew your results. See https://books.google.com/ngrams/info.

[36] Aptly put in Ted Underwood (20 February 2013), "Wordcounts are Amazing," *The Stone and the Shell Research Blog*, http://tedunderwood.com/2013/02/20/wordcounts-are-amazing/.

[37] Ben Zimmer (18 October 2012), "Bigger, Better Google Ngrams: Brace Yourself for the Power of Grammar," *TheAtlantic.com,* http://www.theatlantic.com/technology/archive/2012/10/bigger-better-google-ngrams-brace-yourself-for-the-power-of-grammar/263487/.

[38] Historians' responses largely played out across social media.

Fig. 1.2 "The United States are" versus "The United States is".

popularizing this model of big digital history and continue to be recurrent features in the popular press, academic presentations, and lectures.[39]

Culturomics also presents historians with some professional cautionary notes, as Ian Milligan brought up in a *Journal of the Canadian Historical Association* article.[40] The authorship list of the paper and the tool was extensive: 13 individuals and the Google Books team. There were mathematicians, computer scientists, scholars from English literature, and psychologists. However, there were no historians present on the list.[41] This is suggestive of the degree to which historians had not then yet fully embraced digital methodologies, an important issue given the growing significance of digital repositories, archives, and tools.[42] Given the significance of culturomics and its historical claims, this

[39] See, for example, C. Tumbe (2019), "Corpus linguistics, newspaper archives and historical research methods", *Journal of Management History*, Vol. 25 No. 4, pp. 533–549, or Victor Zakharov and Andrei Masevich, "Diachronic Corpora as Research Tool in Humanities." *Slavonic Natural Language Processing in the 21st Century* (2019): 226.

[40] Ian Milligan (2012), "Mining the 'Internet Graveyard': Rethinking the Historians' Toolkit," *Journal of the Canadian Historical Association*, 23(2), 21–64.

[41] A situation which continues – in a similar vein, a 2019 study that attracted international attention claims to evaluate 'subjective wellbeing' over two hundred years, drawing on Google Books and comparing word use to a *modern* 'list of words and rate them on their valence, indicating how good or bad individual words make them feel', was published and none of the authors were historians. Hills, T.T., Proto, E., Sgroi, D. *et al.* Historical analysis of national subjective wellbeing using millions of digitized books. *Nat Hum Behav* (2019) doi:10.1038/s41562-019-0750-z.

[42] One of the issues raised by the 2016 piece in the LA Review of Books by Allington, Golumbia and Brouillette was the way DH work, in their formulation, relied too much on tools built by tech

omission did not go unnoticed. Writing in the American Historical Association's professional newspaper, *Perspectives*, then-AHA President Anthony Grafton tackled this issue. Where were the historians, he asked in his column, when this was a historical project conducted by a team of doctoral holders from across numerous disciplines?[43] To this, project leaders Erez Leiberman Aiden and Jean-Baptiste Michel responded in a comment, noting that while they had approached historians and used some in an advisory role, no historians met the "bar" for meriting inclusion in the author list: every one of the project participants had "directly contributed to either the creation or the collection of written texts (the 'corpus'), or to the design and execution of the specific analyses we performed." As for why, they were clear:

> The historians who came to the [project] meeting were intelligent, kind, and encouraging. But they didn't seem to have a good sense of how to yield quantitative data to answer questions, didn't have relevant computational skills, and didn't seem to have the time to dedicate to a big multi-author collaboration. It's not their fault*: these things don't appear to be taught or encouraged in history departments right now*.[44] (emphasis added).

To some degree it is an overstatement, as the previous examples in this chapter illustrate. Historians are doing amazing work with data. Yet the teaching of digital history perspectives and method are not yet in the mainstream of the profession, something which was true both in 2015 with the first edition of this book and in 2021 with the second one. This is certainly changing, albeit slowly: witness American Historical Association panels on both teaching and doing digital history, as well as the emergence of exciting and well-received books such as *Technology and the Historian*.[45] The *Historian's Macroscope* is very much a part of this conversation.

---

companies like Google or in concert with non-humanities scholars. Amardeep Singh, in a blog post response, suggested that it is in the *outcomes* of such projects that we should judge their historical or humanistic value, not their origin. See Amardeep Singh (2016) 'In Defense of Digital Tools (by a Non-Tool)' http://www.electrostani.com/2016/05/in-defense-of-digital-tools-by-non-tool.html.

[43] Anthony Grafton (March 2011), "Loneliness and Freedom," *AHA Perspectives*, http://www.historians.org/perspectives/issues/2011/1103/1103pre1.cfm.

[44] Comment by Jean-Baptiste Michel and Erez Lieberman Aiden on Anthony Grafton, "Loneliness and Freedom."

[45] Adam Crymble (2021) *Technology and the Historian: Transformations in the Digital Age.* University of Illinois Press, Urbana. An earlier work well worth your time is Toni Weller (ed) (2013),

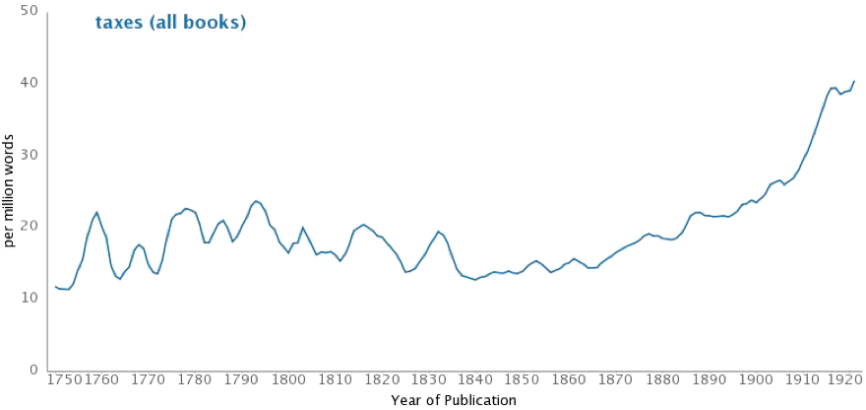

Fig. 1.3 Bookworm search of "taxes", all books.

While the Ngram database is a good way to begin to understand and think about textual analysis, other important resources exist for humanities scholars; tools like Bookworm, a textual visualization platform developed by Ben Schmidt, a historian at Northeastern University, and Erez Lieberman Aiden. To use their platform, visit http://benschmidt.org/OL/. In *Bookworm*, we see that size is not everything. Its "Open Library" search functionality "only" searches a little over a million books (the openlibrary.org collection), but it draws those books from the pre-1923 corpus; this means that it has functionality not available to a tool that incorporates copyrighted works. Consider the following search depicted in Fig. 1.3.

The frequency of taxes is increasing, with initial valleys and peaks from 1750 through 1830, and then a steady ascendency up to and continuing into the end of the database in 1923. In the former Ngram database, this would be all we could see. But, due to the granularity of the newer database, we can see where discussion of taxes has been increasing. If we focus our search so it simply considers books published in the United States, we see the graph depicted in Fig. 1.4.

The peak here occurs much earlier — the American Revolution — and only begins to trend upwards into much later in the 19th century. We are seeing more refined results, and the overall contours become clearer when we look at taxes in the United Kingdom (Fig. 1.5).

History in the Digital Age, Routledge, Abingdon, UK. The *Open Syllabus* project has collated over 7 million course syllabi from universities and colleges across the world. Searching for 'digital history' finds over 150 separate works: https://opensyllabus.org/results-list/titles?findWorks="digital%20history"

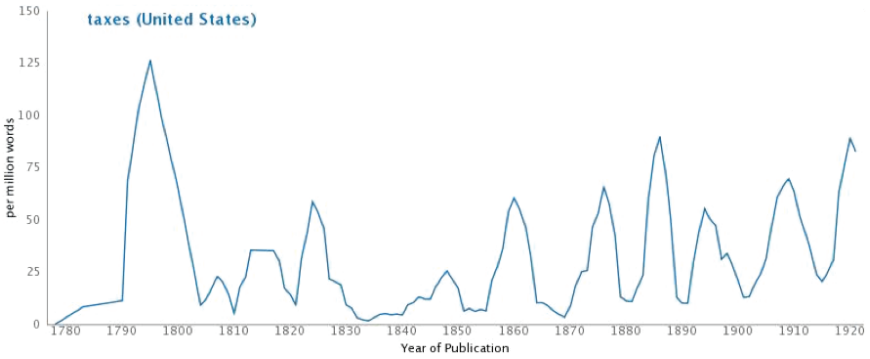

Fig. 1.4 Bookworm search of "taxes" for books published in the United States.

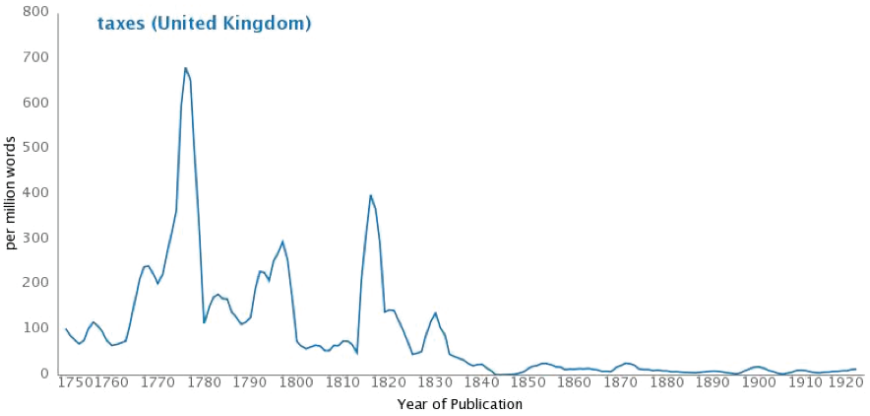

Fig. 1.5 Bookworm search of "taxes" for books published in the United Kingdom.

Taxes assumed far more significance in the late 17th century than they did thereafter, unlike the United States where they became more relatively significant.

While this occludes context at a glance, because included books are not under copyright we can move from the distant reading level of relative word frequency to individual books by clicking on a given year. We discover that the peak year is 1776, the American Revolution, and that it was the discussion of taxes as a central role. By clicking on each year, readers can be taken to the texts themselves: in this case, discussions of the *Wealth of Nations*, political pamphlets published in England concerning taxation, speeches published, and government reports. There are other options: users can discover the percentage of books in

which a word appears, for example, or in the American context go down to the level of the author's state. It thus serves the distant reading purpose of looking at overall trends, but also the close reading perspective of finding individual works. *Bookworm* is even more versatile than we have the space to do it justice here: it has been adapted by the Hathi Trust Research Centre,[46] to explore and contextualize State of the Union addresses in the United States, and beyond.

Maps can also play a significant role in uncovering new historical dimensions, as a project like *ORBIS: The Stanford Geospatial Network Model of the Roman World* vividly demonstrates. *ORBIS*, developed by Walter Scheidel and Elijah Meeks at Stanford University, allows users to explore Roman understandings of space as temporal distances and linkages, dependent on stitching together roads, vehicles, animals, rivers, and the sea.[47] Taking the Empire, mostly circa 200 AD, *ORBIS* allows visitors to understand that world as a geographic totality: it is a comprehensive model. As Scheidel and Meeks explained in a white paper, the model consists of 751 sites, most of them urban settlements but also including important promontories and mountain passes and covers close to 10 million square kilometers (~4 million square miles) of terrestrial and maritime space. There are 268 sites that serve as seaports. The road network encompasses 84,631 kilometers (52,587 miles) of road or desert tracks, complemented by 28,272 kilometers (17,567 miles) of navigable rivers and canals.[48] Wind and weather are calculated, and the variability and uncertainty of certain travel routes and options are well provided. 363,000 "discrete cost outcomes" are made available. A comprehensive, "top down" vision of the Roman Empire is provided (given how we are used to understanding space), drawing upon conventional historical research. Through an interactive graphical user interface, a user — be they a historian, a lay person, or student — can decide their start, destination, month of travel, whether they want their journey to be the fastest, cheapest, or shortest, how they want to travel, and then specifics around whether they want to travel on rivers or by road (from foot, to rapid march, to horseback, to ox cart, and beyond). For academic researchers, they are now able to begin substantially more finely-grained research into the economic history of antiquity.

[46] https://bookworm.htrc.illinois.edu/

[47] Walter Scheidel, Elijah Meeks, and Jonathan Weiland (2012), "ORBIS: The Stanford Geospatial Network Model of the Roman World," http://orbis.stanford.edu/#.

[48] Walter Scheidel, Elijah Meeks, and Jonathan Weiland (May 2012), "ORBIS: The Stanford Geospatial Network Model of the Roman World," http://orbis.stanford.edu/orbis2012/ORBIS_v1paper_20120501.pdf

The *ORBIS* project has moved beyond the scholarly monograph into an accessible and comprehensive study of antiquity. Enthusiastic reactions from media outlets as diverse as the *Atlantic*, the *Economist*, the technology blog *Ars Technica*, and the ABC television network all demonstrate the potential for this sort of scholarship to reach new audiences.[49] *ORBIS* takes big data, in this case an entire repository of information about how long it would take to travel from point A to point B (remember, 363,000 cost outcomes), and turns it into a modern-day *Google Maps* for antiquity.

Another project that uses mapping to great *historical* effect is Vincent Brown's *Slave Revolt in Jamaica, 1760–1761*.[50] Using the open source mapping tool *Leaflet.js*, Brown's animated map depicts an unfolding narrative about the rebellion of the enslaved Black men and women on Jamaica and makes a cartographic argument about the intersection of landscape, rebellion, and counterinsurgency. Reading the textual sources of the revolt, Brown argues, necessarily means that we do not know what the rebels were thinking; but

> we learn something else by plotting the combatants' movements in space. Tracing their locations over time, it is possible to discern some of their strategic aims and to observe the tactical dynamics of slave insurrection and counter-revolt.

Another approach to working with data is to take the approach of 'literate programming'; that is, when confronted by some kind of historical information that can be quantified (whether that is census data or simple word counts from a corpus of pamphlets or diaries), we weave our reflection around the data, and our computer code, to create a *reproducible* document that not only communicates our research but shows every step in our analysis. By 'reproducible' we mean that the document can be opened on one's own machine and all of the computations reproduced; such documents are often called 'notebooks' and they can mix text with a variety of programming languages, most commonly Python or R. The

[49] "Travel Across the Roman Empire in Real Time with ORBIS," *Ars Technica*, accessed June 25, 2013, http://arstechnica.com/business/2012/05/how-across-the-roman-empire-in-real-time-with-orbis/; "London to Rome, on Horseback," *The Economist*, accessed June 25, 2013, http://www.economist.com/blogs/gulliver/2012/05/business-travel-romans; Rebecca J. Rosen (May 23, 2012), "Plan a Trip Through History With ORBIS, a Google Maps for Ancient Rome," *The Atlantic*, http://www.theatlantic.com/technology/archive/2012/05/plan-a-trip-through-history-with-orbis-a-google-maps-for-ancient-rome/257554/.

[50] Brown, Vincent. 2012. Slave Revolt in Jamaica, 1760–1761: A Cartographic Narrative. https://revolt.axismaps.com.

National Library of Scotland's 'Data Foundary' takes this approach from a teaching and outreach perspective. The NLS has many collections already digitized, but by using the Jupyter Notebook platform, its staff are able to walk the reader through *how* to load the data up, and how to ask historical questions of the data. In their 'Exploring A Medical History of British India' notebook[51] Digital Library Research Intern Lucy Havens shows the reader not just the nuts and bolts of loading the data up, but also builds into the notebook potential research questions like, 'What is the rhetoric around vaccinations and public health?' or 'How does the language around mental hospitals change over time?' We can imagine more works of digital history taking this approach, equipping the reader not only with the data and tools to assess the argument as it is made, but also the foundation for taking the research further.

From textual analysis explorations in the Old Bailey, to the global network of 19th century commodities, to large collections of Victorian novels or even millions of books within the Google Books database, or to the travel networks that made up the Roman Empire, to cartographic narratives, to reproducible research, thoughtful and careful employment of big data has much to benefit historians today. Each of these projects, representative samples of a much larger body of digital humanities work, demonstrates the potential that new computational methods can offer to scholars and the public.[52] None of them would be possible, of course, if information professionals had not done such a fantastic job of collecting all this material — in a theme that we will return to, librarians and archivists often do the heavy lifting in curating, collecting, and preserving these traces of the past in aggregate form for us.

Early emergences: Humanities computing, and the emergence of the digital humanities

To know where we are today, and indeed where we are going, we need to understand where we as a discipline came from. In this section, we provide a brief overview of the evolution of the digital humanities and digital history: the intellectual tradition that has led to the projects discussed earlier in this chapter. This is not an exhaustive history, but provides a basic sense of where our discipline has emerged

[51] Havens, Lucy, 2020. Exploring A Medical History of British India. https://data.nls.uk/tools/jupyter-notebooks/exploring-a-medical-history-of-british-india/

[52] The interested reader should follow http://www.digitalhumanitiesnow.org and @dhnow on Twitter to be kept abreast of current developments and projects in digital history. The resource is a machine–human collaboration meant to surface the best recent digital humanities work, which is itself a leveraging of big data that has only recently become possible. See also https://www.zotero.org/groups/2725721/a_collection_of_digital_history_projects/

in part; it does not capture the many diverse national and linguistic traditions that underlie the very diverse field of the broader digital humanities today.[53] Writing a history of such a diverse field is difficult (and is not our goal). Such a history would draw from many different origin stories, from corpus linguistics, computer science, English literature, historical studies, archaeology, and so forth — but we anchor our brief overview within the twin fields of digital humanities and quantitative history.

Our first edition centred this discussion around Roberto Busa's work in the 1950s and 60s developing (with the help of many unsung computer operators) a computational index to the writings of Thomas Aquinas. And it is no doubt important, but it was not the only strand that eventually comes together to form this braided thing 'digital history'. As wider access to digital computers grew on university campuses in the 1960s, as Susan Hockey recounts in her history of humanities computing, this period saw the rise of scholars interested in the digital opportunities offered by large-scale concordances — automatic searching and cross-referencing. Scholars began with collections of texts and subsequently moved into areas such as authorship attribution and quantitative approaches to literary style; notably, in 1964, statisticians Frederick Mosteller and David Wallace used computers to attempt to identify the authors of a dozen disputed *Federalist Papers*; an attempt generally deemed successful.[54] Conferences and journals emerged, such as the *Computers and the Humanities*, accompanied by the establishment of research centers.

Historians would seem initially well positioned, by the 1950s and 1960s, to become involved with large-scale computational inquiries. The *Annales* school aimed to dramatically expand the scope of historical inquiry. In particular, the work of Fernand Braudel is worth briefly discussing. Braudel, amongst the most instrumental historians of the 20th century, pioneered a distant approach to history: his inquiries spanned large expanses of time and space, that of civilizations, the Mediterranean world, as well as smaller (only by comparison) regional histories of Italy and France. His approach to history did not necessitate disengagement with the human actors on the ground, seeing instead that beyond the

[53] The annual digital humanities conference is testament to the diversity of the field; the DH 2014 Conference Call for Proposals was translated into more than 20 languages alone. As historians, we acknowledge that "digital history" can be viewed as having a rather different foundational narrative, emerging out of work in oral and public history (though our own personal trajectories are more from the digital humanities side of things than from those subfields of history). See for instance the post by Stephen Robertson, where he draws out the differences between "digital humanities" and "digital history." http://drstephenrobertson.com/2014/05/23/the-differences-between-digital-history-and-digital-humanities/

[54] Susan Hockey (2004), "The History of Humanities Computing," in *A Companion to Digital Humanities*, Susan Schreibman, Ray Siemens and John Unsworth (eds), Oxford: Blackwell.

narrow focus of individual events lay the constant ebb and flow of poverty and other endemic structural features.[55] While his methodology does not apply itself well to rapidly changing societies (his focus was on the long-term slow change), his points around distant reading and the need for collaborative interdisciplinary research are a useful antecedent for today's researchers.

Quantitative and computational history was on the rise by the late 1960s. Articles and special issues on the subject littered several contemporary journals, including *The American Historical Review*, *The Journal of American History*, *The Journal of Contemporary History*, and *History and Theory*. In 1965, 35 historians attended a three-week seminar on computing in history at the University of Michigan; by 1967, over 800 scholars were receiving a newsletter aimed at computing for history.[56] Two AHA conferences on quantification in history were held in 1967, and many of the earliest issues of *Computers and the Humanities* featured the work of historians.

For historians, however, computational history became associated with demographic, population, and economic histories. For a time in the 1970s, it looked like history might move wholesale into quantification, with the widespread application of math and statistics to the understanding of the past. By 1972, at least half a dozen new journals and magazines were devoted to some aspect of computing and history.[57] Literature scholars pursued textual analysis; historians, to generalize a bit, preferred to count. A book like Michael Katz's *The People of Hamilton, Canada West*, which traced economic mobility over decades using manuscript censuses, was a North American emblem of this form of work. These were fruitful undertakings, providing invaluable context to the more focused social history studies that fleshed out periods under study.[58] A potential downside, however, was that computational history became associated with

[55] The *longue durée* stands opposite the history of events, of instances, covering instead a very long time span in an interdisciplinary social science framework. In a long essay, Braudel noted that one can then see both structural crises of economies, and structures that constrain human society and development. For more, see Fernand Braudel, "History and the Social Sciences: The Longue Durée" in Fernand Braudel, *On History*, trans. Sarah Matthews (Chicago: University of Chicago Press, 1980), 27. The essay was originally printed in the *Annales E.S.C.*, no. 4 (October–December 1958).

[56] Charles M. Dollar (1969), "Innovation in Historical Research: A Computer Approach," *Computers and the Humanities*, **3**(3).

[57] Joel H. Silbey (1972), "Clio and Computers: Moving into Phase II, 1970–1972," *Computers and the Humanities*, **7**(2).

[58] Michael Katz (1975), *The People of Hamilton, Canada West: Family and Class in a Mid-Nineteenth-Century City*, Cambridge, MA: Harvard University Press. See also A. Gordon Darroch and Michael

quantitative studies. This was not aided by some of the hyperbole that saw computational history as making more substantial "truth" claims, or the invocation of a "scientific method" of history.[59] As mainstream historians increasingly questioned this strand of "objectivism," itself a trend dating back to the 1930s, cliometrics became estranged from the mainstream of the profession.[60] This stigma would persist early into the 21st century, even while literary scholars pursued increasingly sophisticated forms of textual analysis, social networking, and online exploratory portals. This 1960s and 1970s explosion of digital work represented the first "wave" of computational history.

Yet, in part due to hubris, debates over seminal works such as *Time on the Cross*, and a more general move towards social history within some elements of the historical profession, computational history retreated in the 1970s. The first wave had come to an end. Yet it would re-emerge in the 1990s with the advent of personal computing, easy-to-use graphical user interfaces, and improvements in overall accessibility. This represented a second "wave" of computational history. Painstaking punchcards and opaque input syntax gave way to relatively easy to use databases, GIS, and even early online networks such as H-Net and USENET. Early conferences, like Hypertext and the Future of the Humanities, held at Yale in 1994, included some of the founders of modern digital history.[61] The *Journal of the Association for History and Computing* (*JAHC*) was published between 1998 and 2010, growing out of annual meetings of the American Association for History and Computing, itself founded in Cincinnati in January 1996. Seeing digital methods as transforming *both* the creation and dissemination of history, *JAHC* published forward-thinking articles on hypertext, digital

D. Ornstein (1980), "Ethnicity and Occupational Structure in Canada in 1871: The Vertical Mosaic in Historical Perspective," *Canadian Historical Review*, **61**(3), 305–333.

[59] Computational historians in 1967 were already arguing against the common notion that quantitative history needed to be positivist history. Vern L. Bullough (1967), "The Computer and the Historian–Some Tentative Beginnings," *Computers and the Humanities*, **1**(3).

[60] Witness the debate over Robert William Fogel and Stanley L. Engerman (1974), *Time on the Cross: The Economics of American Negro Slavery*, New York, NY: W.W. Norton and Company, which was condemned for reducing the human condition of slavery to numbers. On the flip side, it also provided context in which to situate individual stories. The debate continues in countless historical methods classes today. For the general estrangement between the historical profession and cliometrics, see Ian Anderson (2008), "History and Computing," *Making History: The Changing Face of the Profession in Britain*, http://www.history.ac.uk/makinghistory/resources/articles/history_and_computing.html, (accessed 20 December 2012).

[61] Attendees included Gregory Crane and Roy Rosenzweig. Program at https://web.archive.org/web/20011129042909/http://www.cis.yale.edu/htxt-conf/program.html.

teaching methods, barriers to adoption, and new means of representing the past (whether through sound, graphics, maps, or the Web). The back issues represent a snapshot of the cutting-edge work going on.[62] Crunching away at census manuscripts, or attempting to identify authorship, or counting words, the broad interdisciplinary scholarly field known as humanities computing emerged, bringing together historians, philosophers, literary scholars, and others under a broad computational tent.[63]

Projects like 'The Map of Early Modern London', led by Janelle Jenstad, emerged first (in 1997) by taping a historical map to the classroom wall, then working with a digital version and marking up hyper-textual versions of historical and literary texts as both research and pedagogy. That is, it involves digital tool use for humanistic purposes, but also explores what this humanistic exploration implies for the tools.[64] Is this 'digital history'? Is it 'digital humanities'? One could spend pages defining humanities computing, and its eventual successor the digital humanities. Indeed, several other authors have. In a provocative essay, "What is Humanities Computing and What is it Not," John Unsworth defined the field as "a practice of representation, a form of modeling or [...] mimicry. It is [...] a way of reasoning and a set of ontological commitments, and its representational practice is shaped by the need for efficient computation on the one hand, and for human communication on the other."[65] Simply using word processors, e-mail, or communicating by list-servs did not a humanist computationist make. To be "DH," required a new method of thinking. The shift towards the digital humanities was not simply a shift in nomenclature, although there are elements of that as well. Patrik Svensson has traced the shift from humanities computing to the digital humanities, showing how the nomenclature change staked a new definition that was even more inclusive, and broad: inclusive of questions of design, the born-digital, new media studies, and more emphasis on tools with less emphasis on more straightforward methodological discussions.[66]

---

[62] Available at http://quod.lib.umich.edu/j/jahc/browse.html.

[63] Willard McCarty (2005), *Humanities Computing*, Basingstoke, England; New York, NY: Palgrave Macmillan.

[64] Janelle Jenstad, ed. Map of Early Modern London. History of MoEML. https://mapoflondon.uvic.ca/history.htm.

[65] John Unsworth (November 8, 2002), "Unsworth: What Is Humanities Computing and What Is Not?," https://web.archive.org/web/20190424171153/http://computerphilologie.uni-muenchen.de/jg02/unsworth.html.

[66] Patrik Svensson (2009), "Humanities Computing as Digital Humanities," *Digital Humanities Quarterly* **3**(3), http://www.digitalhumanities.org/dhq/vol/3/3/000065/000065.html.

As recounted in Matthew Kirschenbaum's take on the history of the digital humanities, the term also places different emphasis on different parts of the phrase. As the National Endowment for the Humanities Chief Information Officer put it to him, he "appreciated the fact that it seemed to cast a wider net than 'humanities computing' which seemed to imply a form of computing, whereas 'digital humanities' implied a form of human-ism."[67]

This expanded definition seemed to extend to digital history as well. As previously discussed, digital history has an uneasy relationship to digital humanities; some see the former as a subset of the latter, while others see them as overlapping but not hierarchically connected communities. Digital history, for one, sits closer to the public humanities than many of its counterparts. The group is also much less well represented at digital humanities conferences than those coming from literature and modern language backgrounds,[68] leading some popular accounts of the digital humanities to ignore digital history entirely.[69] We do not plan on resolving those differences here, but instead will draw from every corner of the big tent of digital humanities in order to facilitate training new digital historians.

In short, the term "digital humanities" is difficult to define. A fun way to open a digital humanities or history course is to visit Jason Heppler's fun website, "What is Digital Humanities?"[70] Participants in the annual Day of DH, (which began with the University of Alberta, began to rotate between institutions, and is now somewhat defunct), were annually asked to provide their own definitions. Between 2009 and 2014, Heppler compiled 817 entries, all made available in raw data format. A visitor to his website can refresh it for a new definition: ranging from the short and whimsical ("[a]s a social construct," "[t]aking people to bits"), to the long and comprehensive ("[d]igital Humanities is the critical study of how the technologies and techniques associated with the digital medium intersect with and alter humanities scholarship and scholarly communication") to more specific definitions focused on making or digital preservation. It's still one of the best compilations.

---

[67] Matthew G. Kirschenbaum (2010), "What Is Digital Humanities and What's It Doing in English Departments?," *ADE Bulletin*, no. 150, 55–61.

[68] See Weingart's blog series on digital humanities conferences, http://www.scottbot.net/HIAL/?tag=dhconf.

[69] Stephen Richardson (23 May 2014), "The Differences between Digital History and Digital Humanities," *drstephenrobertsom.com*, http://drstephenrobertson.com/2014/05/23/the-differences-between-digital-history-and-digital-humanities/. See the chapter in Adam Crymble's (2021) monograph discussing the 'origin myths' of computing in historical research.

[70] Jason Heppler (2013), "What Is Digital Humanities?," http://whatisdigitalhumanities.com/.

Amongst such crowded and thoughtful conversation, we hesitate to add our own definition. A definition, however, is in order for the purposes of this book. We believe that the digital humanities are partly about understanding what digital tools have to offer, but also — and perhaps more importantly — an understanding of what the digital does and has done to our understanding of the past and ourselves.

That definition raises an obvious question: 'who do we imagine as 'ourselves'?' We failed to ask this question, in the first edition. Roopika Risam, on the other hand, does: 'What forms of the "human" are sanctioned when artificial intelligence can reproduce human processes?'[71] We can just as easily substitute 'Big Data' in for 'artificial intelligence' there. Risam shows how the use of digital tools which emerge principally in the Global North can re-enact colonial violence, can marginalize, or diminish ways of being that do not 'fit' paradigms assumed by the (often white, western, male) creators of these tools. If digital history and the digital humanities take these tools on without thinking through their origins, their training data, and what 'counts'[72] then we will not create truer histories, or truer inquiries. Thus, we need to be careful about how we imagine our historical big data, and our own particular versions of digital history / digital humanities.

The historical big data that we currently work with is largely data that was collected by individuals and governments working in pre-digital days; the data is a reflection of structures of power and relationships from those eras. But at the same time, more and more data that we *will* work with in the future will be data that was collected by, and from, digital services (both commercial and governmental); consider how much metadata lies behind a single Instagram

[71] Roopika Risam, New Digital Worlds: Postcolonial Digital Humanities in Theory, Praxis and Pedagogy. Evanston, Illinois: Norhwestern University Press, 2018, 127.

[72] On which in particular, see the entire chapter, 'What Gets Counted Counts' in Catherine D'Ignazio and Lauren F. Klein (2020). Data Feminism, Cambridge MA: MIT Press. pp. 97–124. Retrieved from https://data-feminism.mitpress.mit.edu/pub/h1w0nbqp. Digital tools emerge from larger social contexts. Thus the field as a whole, if the Alliance for Digital Humanities' Organizations can be taken as representative, has recently taken steps aimed at ameliorating this — "Our goal is to discern and weed out structures based upon white privilege, heterosexist bias, ableist presumptions, sexism, and other oppressions, including but not limited to those exacerbated by the use of technology. These are issues which confront us all, and they require all of us to challenge and change the systems that promote, perpetuate, or tolerate injustice and inequalities, including our own.". ADHO Statement on Black Lives Matter, Structural Racism, and Establishment Violence. July 18 2020. https://adho.org/blm-structuralracism-establishmentviolence.

tweet, or a Facebook post. We should therefore think about our definition within D'Ignazio and Klein's '7 Principles' for working with data. As they say, "data are part of the problem, to be sure. But they are also part of the solution." Their principles[73]:

1. Examine power — we need to be aware how our tools/data embody the operations of power (to create this data, to extract it, to analyze it)
2. Challenge power — as historians working with big data, we will miss important parts of the story if we don't consider point 1 without reflection on how these structures continue
3. Elevate emotion and embodiment by 'valu[ing] multiple forms of knowledge'
4. Rethink binaries and hierarchies — systems of classification, especially in computation, have to knock off the edges to make people fit into boxes.[74]
5. Embrace pluralism — who are we ignoring when we use data from a particular source?
6. Consider context — 'data are not neutral or objective. They are products of unequal social relations'.
7. Make labor visible — especially in digital history, the labour of working with data falls unequally. Data do not simply emerge ready-to-be-explored. What are the conditions that enable you to work with this information?

In this book, with this in mind, we peel back the layers of a particular approach to big data using specific tools such as topic modeling and network analysis. We try to live up to these principles, but we also should recognize that 'failure' is also a core value of digital history and be prepared to discuss why and how.[75] As we say to our students, 'you don't have to be 'techy' to do digital history. But you do have to be willing to be wrong and to own why'. When we fail, that is the beginning of a conversation towards a better iteration.[76]

[73] D'Ignazio and Klein, 17–18.

[74] On which note, see Allison Parrish reminds us that 'programming is forgetting' . Allison Parrish, 2016. 'Programming is Forgetting: Toward a New Hacker Ethic'. *Open Transcripts*. http://opentranscripts.org/transcript/programming-forgetting-new-hacker-ethic/#

[75] Quinn Dombrowski (2019) *Towards a Taxonomy of Failure.* http://quinndombrowski.com/blog/2019/01/30/towards-taxonomy-failure. See also Croxall and Warnick, 'Failure' Digital Pedagogy in the Humanities https://digitalpedagogy.hcommons.org/keyword/Failure and Graham, 2019, *Failing Gloriously and Other Essays* Grand Forks, ND: The Digital Press at the Univeristy of North Dakota Press.

[76] See Martin's 2020 discussion of the failures of the teaching of digital history in Canada, especially the first edition of this volume. Kim Martin (2020) 'Clio, Rewired: Propositions for the Future of Digital History Pedagogy in Canada' *The Canadian Historical Review*, **101**(4), pp. 622–640.

Why this all matters now: The third wave of computational history

We have entered an era of big data. Big data emerges from the context of the constant, always-on, notifications, check-ins, National Security Agency (NSA)-metadata-gleaned, pervasive computing world. Even our thermostats and doorbells are watching what we do to generate insights from our big data lives.[77] People purchase (overt) home surveillance devices like Amazon's 'Ring' doorbell; they use (covert) home surveillance devices like Amazon Alexa who listens in on every conversation. Universities require students to install invasive software to proctor exams, cameras on, observing the student's room, face, expressions, and computer use.

There's a lot of data out there.

This makes decisive movement towards digital methods, with explicit ethical considerations over *using* that data, over the next ten or 20 years, imperative for the profession. Yet while big data is often explicitly framed as a problem for the future, it has already presented fruitful opportunities for the past. The most obvious place where this is true is archived copies of the publicly accessible Internet, as co-author Ian Milligan has argued in his 2019 book *History in the Age of Abundance? How the Web is Transforming Historical Research*.[78] The advent of the World Wide Web in 1991 has had revolutionary effects on human communication and organization, and its archiving presents a tremendous body of non-commercialized public speech. There is a *lot* of it, however, and large methodologies will be needed to explore it. It is this problem that we believe makes the adoption of digital methodologies for history especially important.

As we saw above, historians have been through these questions before. In the 1960s, large censuses met punch-card computing to result in significant scholarly contributions that continue to contextualize more focused studies today. Now, as the digital humanities flourish, we can see the current historical interest in them as owing its roots to that initial flurry of interest. Putting this into historical context, if the first wave of computational history emerged out of humanities computing, and the second wave developed around textual analysis (and H-Net, Usenet, and GIS), we believe that we are now in a third revolution in computational history. There are three main factors that make this instrumental: decreasing storage costs, with particular implications for historians; the power of the Internet and

[77] A great overview of big data can be found in Viktor Mayer-Schönberger and Kenneth Cukier (2013), *Big Data: A Revolution that Will Transform How We Live, Work, and Think*, Boston, MA: Eamon Dolan Book; an overview of the perils of all this are explored in Cathy O'Neil (2016) Weapons of Math Destruction. Crown Random House: New York.

[78] Ian Milligan, *History in the Age of Abundance? How the Web is Transforming Historical Research* (Montreal & Kingston: McGill-Queen's University Press, 2019).

cloud computing; and the rise of open source tools.[79] We are asking similar questions, in many cases, to the 1960s pioneers: just with more powerful and even more accessible tools (for all the frustrations that the tools discussed in this book can occasionally produce, they luckily are not punch-card based).

Significant technological advances in how much information can be stored herald a new era of historical research and computing that we need to prepare for *now*. Historical methods need to develop in order to keep up with where our profession might go in the next ten or 20 years. In short, we can retain more of the information produced every day, and the ability to retain information has been keeping up with the growing amount of generated data. As author James Gleick argued:

> The information produced and consumed by humankind used to vanish — that was the norm, the default. The sights, the sounds, the songs, the spoken word just melted away. Marks on stone, parchment, and paper were the special case. It did not occur to Sophocles' audiences that it would be sad for his plays to be lost; they enjoyed the show. Now expectations have inverted. Everything may be recorded and preserved, at least potentially.[80]

This has been made possible by the corollary to Moore's Law (which held that the number of transistors on a microchip would double every two years), Kryder's Law.[81] He argues, based on past practice, that storage density will double approximately every 11 months. While this law may be more descriptive than predictive, the fact remains that storage has been getting cheaper over the last ten years and has enabled the storage and hopeful long-term digital preservation of invaluable historical resources.

We store more than we ever did before, and increasingly have an eye on this digital material to make sure that future generations will be able to fruitfully explore it. The creation or generation of data and information do not, obviously, in and of themselves guarantee that they will be kept — for that we have the field of digital preservation. In 2011, humanity created 1.8 zettabytes of information.

---

[79] Presaged in the aforementioned Roy Rosenzweig (2003), "Scarcity or Abundance? Preserving the Past in a Digital Era," *American Historical Review*, **108**(3), 735–762, available online at http://chnm.gmu.edu/digitalhistory/links/pdf/introduction/0.6b.pdf.

[80] James Gleick (2011), *The Information: A History, A Theory, A Flood*, New York: Pantheon.

[81] Chip Walter (25 July 2005), "Kryder's Law," *Scientific American*, http://www.scientificamerican.com/article.cfm?id=kryders-law.

This is not an outlier: from 2006 until 2011, the amount of data expanded by a factor of nine.[82] All such statistics are hard to calculate, but on some estimates, 2020 would have seen the creation of 44 zettabytes of data (a zettabyte equals one trillion gigabytes, by the way; that's a 1 with 21 zeros).[83]

These data take a variety of forms, some accessible and some inaccessible to historians. In the latter camp, we have walled gardens and proprietary networks such as Facebook, corporate databases, server logs, security data, and so forth. Save a leak or forward-thinking individuals, historians may never be able to access those data. Yet in the former camp, even if smaller than the latter one, we have a lot of information: Facebook sees 4 petabytes of *new* data every day, including 350 million photos[84]; the hundreds of millions of tweets sent every day over Twitter; the blogs, ruminations, comments, and thoughts that make up the publicly-facing and potentially archivable World Wide Web.[85] Beyond the potentialities of the future, however, we are already in an era of archives that dwarf previously conceivable troves of material. While this book is not about accessing these archives in particular, it does concern itself with the methods necessary to access these sorts of datasets. Historians need to begin to think computationally *now* so that our profession is ready to access these data in the next generation.

The shift towards widespread digital storage, preserving information longer and conceivably storing the records of everyday people on an ever more frequent basis, represents a challenge to accepted standards of inquiry, ethics, and the role of archivists.[86] How should historians respond to the transitory nature of historical sources, be it the hastily deleted personal blogs held by MySpace, or the destroyed websites of GeoCities? How can we even use large repositories such as the over two million messages sent over USENET in the 1980s alone? Do we have ethical responsibilities to website creators who may have had an expectation of privacy, or in the least had no sense that they were formally publishing their webpage in 1996? These are all questions that we, as professionals, need to tackle. They are, in a word, disruptive.

---

[82] John Gantz and David Reinsel (June 2011), "Extracting Value from Chaos," IDC iView, https://web.archive.org/web/20190404101630/https://www.emc.com/collateral/analyst-reports/idc-extracting-value-from-chaos-ar.pdf.

[83] World Economic Forum. 2019 'How Much Data is Generated Each Day?' https://www.weforum.org/agenda/2019/04/how-much-data-is-generated-each-day-cf4bddf29f/

[84] World Economic Forum. 2019.

[85] Twitter (January 2015), "Getting all tweets for research purposes | Twitter Developers," Twitter.com, https://twittercommunity.com/t/getting-all-tweets-for-research-purposes/30501 ; YouTube (May, 29, 2013), "Statistics — YouTube," YouTube, http://www.youtube.com/yt/press/statistics.html.

[86] Milligan, *History in the Age of Abundance?*

It is important to pause briefly, however, and situate this claim of a revolutionary shift due to ever-bigger datasets into its own historic context. Humanists have long grappled with medium shifts and earlier iterations of this big data moment, which we can perhaps stretch back to the objections of Socrates to the written word itself.[87] As the printing press and bound books replaced earlier forms of scribed scholarly transmission, a similar medium shift threatened existing standards of communication. Martin Luther, the German priest and pivotal figure in the Protestant Reformation, argued that "the multitude of books [were] a great evil;" this 16th century sentiment was echoed as well by Edgar Allen Poe in the 19th century and Lewis Mumford as recently as 1970.[88] Bigger is certainly not better, at least not inherently, but it should equally not be dismissed out of hand.

Accessing the third wave today

This big data, however, is only as useful as the tools that we have to interpret it. Luckily, two interrelated trends make interpretation possible: more powerful personal computers, and more significantly, accessible open-source software to make sense of all these data. Prices continue to fall, and computers continue to get more powerful: significantly for research involving large datasets, the amount of Random-Access Memory, or RAM, that computers have continues to increase. Information loaded into RAM can be manipulated and analyzed very quickly. Even humanities researchers with limited research budgets can now use computers that would have been prohibitively expensive only a few years ago. Online computing environments that leverage high performance computing are similarily becoming available; Google's 'Colaboratory' for instance is a hosted service for Python language programming, that provides access to powerful graphical processing units (like the more familiar CPU, or central processing unit, but designed for high-end mathematics, emerging out of the need for ever better graphics in video games).[89] 'Notebooks', or files where the code and the observations *about* the code are integrated (and so can be re-run in the 'colaboratoy'), can be shared easily.

It is, however, the ethos and successes of the open-source movement that have given digital historians and the broader field of the digital humanities wind in their sails. Open-source software is a transformative concept that moves beyond simply

[87] See the *Phaedrus* dialogue by Plato, available in English translation at http://classics.mit.edu/Plato/phaedrus.html.

[88] James Gleick (2012), *The Information: A History, A Theory, A Flood*, New York: Pantheon.

[89] Google. 'Colaboratory — Frequently Asked Questions' https://research.google.com/colaboratory/faq.html.

"free" software: an open source license means that the code that drives the software is freely accessible, and users are welcome to delve through it, make changes that they see fit, and distribute the original or their altered version as they see fit. Open-source software is often mandated by funding agencies, and in any case, open-source tools and infrastructure run the computing world today. Notable open-source projects include Mozilla Firefox, Linux, the Zotero reference-management software system and the Tropy research photo image-management software developed by George Mason University's Centre for History and New Media (CHNM), the WordPress and Drupal website Content Management System (CMS) platforms, and even entire Microsoft Office-like competitors such as LibreOffice.

For humanists, then, carrying out large-scale data analysis no longer requires a generous salary or expense account. Increasingly, it does not even require potentially expensive training. Take, by way of introduction, the *Programming Historian*, an open-source textbook dedicated to introducing computational methods to humanities researchers — itself written and developed both using open-source publishing software on the popular collaborative code repository website GitHub.com.[90] The *Programming Historian* is a useful introduction, as well, to the potential offered by these programs, which include several that we saw in our preface, as well as some other go-to programs:

- **Wget:** A program, run on the command line, that lets you download entire repositories of information. Instead of right-clicking on link after link, wget can quickly download an entire directory or website to your own computer.
- **MALLET:** The MAchine Learning for LanguagE Toolkit provides an entire package of open-source tools — notably topic modeling, which takes large quantities of information and finds the 'topics' that appear in them.
- **Python** and **R:** All-purpose, open, programming languages with a wide ecosystem of pre-built packages for all manner of tasks from general statistics to topic modeling; see for instance Melanie Walsh's 'Introduction to Cultural Analytics' combined course and workbook at https://melaniewalsh.github.io/Intro-Cultural-Analytics/welcome.html for Python, and Silge and Robinson's 'Text Mining with R' open-access textbook for R at https://www.tidytextmining.com/. The RStudio environment (https://rstudio.com/) can make working with R much more effective. An online version of RStudio is available at https://rstudio.cloud and can be a good place to start.

[90] https://programminghistorian.org/

- **Jupyter notebooks** and services like the previously mentioned Google **Colab** or **Mybinder.org** which serve bundled code and narrative text (that is, the notebooks) in 'containers' through the browser — a computer in the cloud configured to demonstrate a particular piece of analysis or visualization. Tim Sherratt, a historian and hacker in Australia, has used these to great effect for his 'GLAM Workbench' (Galleries, Libraries, Archives, and Museums), a suite of notebooks served through mybinder, https://glam-workbench.github.io/.

These tools are just the tip of the iceberg and represent significant change. Free tools, with open-source documentation written for and by humanists, allow us to unlock the potential inherent in big data. Many custom built tools for digital history or digital humanities work can be found on code-sharing websites like Github.com or Gitlab.com.

Big data represents a key component of this third wave of computational history. By this point, you should have an understanding of what we mean by big data and some of the technical opportunities we have to explore it. The question remains, however: what can this do for humanities researchers? What challenges and opportunities does it present, beyond the short examples provided at the beginning of the chapter?

The limits of big data, or big data and the practice of the history

A quick proviso is in order here. Will big data have a revolutionary impact on the epistemological foundation of history? As historians work with ever-increasing arrays of information, we need to consider the intersection of this trend with broader discussions around the nature of history and the past itself. At first glance, larger amounts of data would seem to offer the potential of empirical advances in 'knowing' history: the utopian promises of a record of all digital communications (turned into a dystopian reality by the NSA, unfortunately), for example, or the ability to process the entirety of a national census over several years. In some ways, it is evocative of the modernist excitement around "scientific" history.[91] Big data has substantial implications for historians, as we again move from studies based on positive examples to the findings of overall trends from extensive computational databases. This is similar to debates from the 1970s and 1980s, where quantitative historians faced criticisms around the reproducibility of their results that were derived from early computer databases and often presented only in

[91] Robert William Fogel (1983), "'Scientific History' and Traditional History," in Robert William Fogel and G.R. Elton (eds), *Which Road to the Past? Two Views of History*, New Haven, NJ and London: Yale University Press.

tables. Yet we believe that, for all the importance of big data, it does not offer any change to the fundamental questions of historical knowledge facing historians.

For all the excitement around the potential offered by digital methods, it is important to keep in mind that it does not herald a transformation in the epistemological foundation of history. We are still working with traces of the past. The past did happen, of course. Events happened during the time before now: lives lived, political dynasties rose and fell, and working people made do with the limited resources that they had before them. Most of the data about these experiences and events, however, have disappeared: the digital revolution may make it possible to consider more than before, perhaps even on an order of magnitude.[92] Yet even with terabytes upon terabytes of archival documents, we are still only seeing traces of the past today. While there is a larger debate to be had about the degree of historical knowledge possible, we believe that, amidst the excitement of big data, this is a point that needs to be considered.

History, then, as a professional practice, involves the crafting of this available information and transforming it into scholarly narratives.[93] Even with massive arrays of data, historians do not simply cut and paste findings from computer databases; such an approach would be evocative of the "scissors and paste" model noted by historian R.G. Collingwood.[94] Having more data is not a bad thing. With more data, in practical terms, there is arguably a higher likelihood that historical narratives will be closer in accordance with past events, as we have more traces to base them on. But this is not a certainty. History is not merely a reconstructive exercise, but also a practice of narrative writing and creation.[95]

Throughout the methodological chapters that follow, the "subjective" appears throughout: decisions about how many topics to search for as we generate

[92] See, for a brief introduction, Keith Jenkins (1991), *Rethinking History*, London: Routledge and Alun Munslow (2010), *The Future of History*, New York, NY: Palgrave Macmillan.

[93] In the years ahead, it is possible the digital turn will render the word "narrative" too confining for describing what historians produce. We continue to use the word in this book, however projects like *SCALAR* and *ORBIS* are making the term increasingly inaccurate. A more encompassing term may be "historiographies."

[94] R.G. Collingwood (1965), *The Idea of History*, London: Oxford University Press. 'Scissors and Paste' also happens to be the name of a digital history project by Melodee Beals, exploring 1750–1850 newspaper reprint culture in the English-speaking world; the website is at http://scissorsandpaste.net/ and the entire project repository can be found at OSF from the Center for Open Science https://osf.io/nm2rq/; the repository includes not only the code, but also all of the scholarly works that Beals derived from the corpus.

[95] A point made by a good number of historians, but see Keith Jenkins (1991), *Rethinking History*, London: Routledge and Alun Munslow (2010), *The Future of History*, New York, NY: Palgrave Macmillan for concise introductions to this line of reasoning.

a topic model, for example; what words to exclude from computations; what words to look for; or what categories of analysis to assume. Underlying this too are fundamental assumptions made by computational linguists, such as the statistical models employed or even premises around the nature of language itself.[96] At a more obvious and apparent level, much of what we do also involves transforming data, or altering these traces of the past into new and in many cases — for the purpose of our narrative construction — more fruitful forms. Many of the techniques discussed in this book involve text, yet many sources are more than just texts: they have images, texture, smell, or are located in specific areas. Even if attuned to issues of historical context, these can and often will be lost.

On a philosophical level, however, the digital transformation of sources does not represent a significant challenge to historical practices. Historians are always changing their sources and are always engaged in choices and decisions. What notes to take? What digital photograph to snap? Who to interview? What sources will lead to a better publication? What will my dean/manager/partner think? What will help build my career? To this litany of issues, digital techniques add new questions that we will discuss in this book: how to break up texts (should you separate "tokens" on word breaks, punctuation, or so forth, as discussed in Chapter 2), what topic modeling algorithm to use, whether to ignore non-textual information, or whether or not a network is a useful visualization of one's results.

If we do not believe that big data is here to correct what other historians are doing in their subdisciplines, however, this also means that we do not believe that the issues briefly outlined above are fundamental shortcomings in our methodologies. Yes, mediums will be transformed, and decisions will be made, but this is within the normal bounds of the historical tradition. Digital history does not offer direct truths, but only new ways of interpreting and understanding traces of the past. More traces, yes, but still traces: brief shadows of things that were.

But to what end? Trevor Owens draws attention to the purpose behind one's use of computational power — generative discovery versus justification of a hypothesis. For Owens, if we are using computational power to deform our texts, we are trying to see things in a new light, new juxtapositions, to spark new insight.[97] Stephen Ramsay talks about this too in *Reading Machines* where he discusses the work of Lisa Samuels and Jerome McGann: "Reading a poem backward is like viewing the face of a watch sideways — a way of unleashing the

[96] A point discussed at the Stanford Digital Humanities Reading Group, as recounted by Mike Widner, "Debating the Methods in Matt Jockers's Macroanalysis," *Stanford Digital Humanities Blog*, https://digitalhumanities.stanford.edu/debating-methods-matt-jockerss-macroanalysis, accessed 6 September 2013.

[97] Trevor Owens (February 3, 2012), "Deforming Reality with Word Lens," *Trevor Owens*, http://www.trevorowens.org/2012/02/deforming-reality-with-word-lens/.

potentialities that altered perspectives may reveal."[98] Owen's purpose in highlighting "justification" against "discovery" is not to condemn one approach over another, but rather to draw attention to the fact that:

> When we separate out the context of discovery and exploration from the context of justification we end up clarifying the terms of our conversation. There is a huge difference between "here is an interesting way of thinking about this" and "This evidence supports this claim."[99]

This then is the nub of big data for digital history. Digital approaches to the past use computational power to force us to look at the materials differently, to think about them playfully, and to explore what these sometimes jarring deformations could mean.

Conclusion

This chapter has provided a basic introduction to the joys and pitfalls of abundance in this new era of big data. Certainly, the authors of this book are all hopeful about the potential of digital history and big datasets. This potentially has already been fruitfully realized in several successful projects: from the criminal trials of the *Old Bailey Online*, the Roman travel patterns of *ORBIS*, the global commodities of the *Trading Consequences* project. We, however, do believe that this all needs to be both contextualized in terms of the ethical implications and the field's historical development, and nuanced with respect to the knowledge claims that can be made. Looking at more recent history, the *Torn Apart / Separados* project is an excellent example of a project that foregrounds its ethical uses of big data to witness the activities of the United States' Immigration and Customs Enforecement agency.[100]

---

[98] Stephen Ramsay (2011), *Reading Machines: Toward an Algorithmic Criticism*, Urbana, IL: University of Illinois Press, p. 33.; Lisa Samuels and Jerome McGann (1999) 'Deformance and Interpretation' *New Literary History,* 30.1, pp. 25–56.

[99] Trevor Owens (February 3, 2012), "Deforming Reality with Word Lens," *Trevor Owens*, http://www.trevorowens.org/2012/02/deforming-reality-with-word-lens/.

[100] Manan Ahmed, Alex Gil, Moacir P. de Sá Pereira, Roopika Risam, Maira E. Álvarez, Sylvia A. Fernández, 2018. Torn Apart/Separados http://xpmethod.columbia.edu/torn-apart/volume/2/index For the full collaborator list please see http://xpmethod.columbia.edu/torn-apart/credits.html; This project is notable not only for its subject matter and for the way it mobilizes digital and humanities scholars and professionals to produce 'nimbly…curated and applied knowledge', but also for the way it envisions the work as an on-going process of continual updates, a kind of scholarly journal platform.

As the next chapter notes, we believe that we are in a transitory moment within both the historical field and the digital humanities more broadly: the "DH" moment is in full flow. Considered engagement with this field, however, requires an understanding of where we have come from and where we are at this present moment in time. Much of what you will encounter and read over the next five chapters will seem new, for it certainly feels like we are straddling a line between revolution and continuity; resolving this tension is going to be a central part of historians' tasks over the coming years. In the chapters that follow, we draw on elements of both perspectives, fully understanding that this book itself is as much a historical artefact of this particular moment as are the tools themselves that we study.

The point of this chapter, however, is to stress that, for all the novelty, there are still points of continuity. Debates over historical knowing have continued into the digital era, but still remain much as they have been over the last hundreds of years: digital history is no more objective, nor no more subjective, than what has come before. The fundamental questions remain the same around humanistic inquiry: What can we know of the past? Whose voices can we try to remediate? In what ways can we use historical knowledge in the present day, from informing policy decisions, to inspiring marginalized communities, to simply telling entertaining stories?

As the next chapter shows, a new moment is upon us and a familiarity with macroscopic knowledge is becoming ever-important. As our world is profoundly reshaped by the digital revolution, as historians increasingly engage with digitized sources, and as we begin to reflect on how to study the 1990s, digital methods become even more pressing. Yet another DH moment is upon us, one made all the more critical by the sheer amount of information that we now have available to us.