Historical Research and Digital Methods

In Conversation with Jo Guldi

Just want to see them, only live a few hours.

Anaclet Pons

Professor of Contemporary History
University of Valencia – Department of Contemporary History
anaclet.pons@uv.es

Jo Guldi

Professor of Quantitative Methods
Emory University
https://www.joguldi.com

Abstract

Jo Guldi, Full Professor of Quantitative Methods at Emory University, where she also recently joined the Department of Computer Science, is one of the most competent researchers in the field of Digital History. The interview we offer is an overview of her vast historiographical output, but also a warning to those who might be tempted to remain technologically and methodologically outdated. Without denigrating traditional research, Jo Guldi shows, in theory and in practice, how research in text mining – and elsewhere – stimulates interdisciplinary innovation in all humanities disciplines, including history. And this, in turn, is a call to change the training we offer, to ask whether the historians of today and tomorrow will be able to write the history of our time if they have not been trained in the analysis of texts as data.

Keywords: Algorithms, Digital History, Digital Humanities, Digital space, History, Methodology, Quantitative methods, Technology, Text Mining, United States

Combining attention to algorithms and databases with care for the concerns of historical theory and social theory can produce a really robust practice

We all live in an era of technological change

It can be said that Jo Guldi is a traditional historian and that she is not a traditional historian at all. She is traditional because her subjects are, when she studies topics such as the history of British ideas about property rights or the history of the landscape, the land and the water. But she is anything but traditional because she is a scholar who uses machine learning and other big data methods to approach traditional humanities concerns. She is a historian of her time, someone who argues that a world awash in text requires new interpretative tools, that can reconcile the quantitative approaches of data science with the nuanced approach of traditional history, an "hybrid knowledge."

Jo Guldi in her office.

Anaclet Pons – *We would normally ask how you first came to enter to the "digital" field. But it seems that your case is not the usual one, because we could say that has been in contact with this field from your earliest childhood. Really?*

Jo Guldi – Really! I grew up in the company town of Texas Instruments, in a school district where children were offered robust courses in math and coding taught by particularly talented instructors who had the knack of making these problems fun. History, by contrast, was relatively neglected, and I had to find my way there gradually, after degrees in Literature and an emphasis on ancient languages. The benefit of learning a little math and code at a young age was that learning more later came with ease.

I believe the inverse is true as well; students who learn a little of writing, history, literature, and languages can always learn more throughout their life. The best approach to secondary education is a balanced one, because the world demands citizens who have the skills of interacting with information of diverse kinds.

Anaclet Pons – *When did you start using this first skill to help you work, and why?*

Jo Guldi – Through my years of undergraduate and graduate study, I focused on learning about the past and reading cultural theory. If I used a computer, it was to search a library catalog, to write, or perhaps to blog. But because I kept in touch with friends whose interest was more technical than my own, I was in conversations in the 2000s where people – outside the history profession – were asking me for detailed answers about how technology and data were changing how historians did their research. At first, these were conversations, and later they were blog entries, increasingly informed by reading what other historians were saying in answer, and sometimes informed by my own early experiments on Google Books. Eventually, I hired data scientists with more skills than my own to run further experiments, and much later, I re-learned how to code in order to unpack the black box of the algorithm for myself.

So it has been one of the surprises of my career that I became a technologist without intending to be. What was essentially a hobby for me, at a time when I was leaning into the history of political economy, turned out to be one of the most relevant parts of my practice. I gave in to questions about the implications of the digital not because of any individual proclivity to evangelize technology – far from it. I am an ancient languages scholar who fell in love with questions of political economy in the eighteenth and nineteenth century. I surrendered to the

conversation about data and algorithms because at every university I have visited, questions about the implications of data and technology have remained relevant.

In service to the profession, I've done my best to consider those conversations with a serious attitude. When questions were raised about what was inside the black box of the algorithm or whether text mining methods might lead to discovering new events, at first I assumed that other scholars would leap into this space of investigation, and that my own work was done. But again and again, I found myself invited to address other historians about these question, and as I read I realized that most of the pressing questions remained unanswered. There were so few of us who were working on the problem of digital methods for the historical profession. It was out of a sense that the questions needed to be answered, and answered correctly, that I began to lean in. Today, the problem of data and algorithm is fully half of my scholarly output, with another half remaining centered on global problems of political economy, especially around the history of displacement and policies for stopping or repairing displacement since the nineteenth century.

Anaclet Pons – *You were awarded the first "Digital History" position in the United States, at the University of Chicago. Can you explain what that meant?*

Jo Guldi – Colleagues at the University of Chicago applied to the Mellon Foundation for a two-year postdoctoral fellowship in "Digital History." To my surprise, those colleagues endorsed my application, which consisted of a very traditional, archivally-researched book about the history of roads in eighteenth-century Britain and a collection of blog entries about the implications of data for the historical profession, and offered me a job.

As a Digital History, I was asked to teach Chicago's graduate students about new methods. I cobbled together a course reviewing contemporary discussions about the practicalities of network analysis and mapping, the ethics and bias of tools, experimental work in History, and the potential of crowdsourcing for the university more generally. It was a terrific challenge to construct a program of study out of whole cloth, but I really enjoyed it. I have been teaching some version of that course – with increasing specificity on advances in methods in a rapidly-changing field – since 2008. Today, I'm developing a multi-semester track to teach graduate students about ethics and bias in data, the array of

available tools for different questions, and advanced methods of text mining that require a hands-on approach to code.

Taking the position in Digital History had complicated implications for the rest of my career. From my point of view, my interest in the implications of data for the History profession had been endorsed by a top journal and some of the top scholars in the field, and I was being invited more frequently to talk about my publications in Digital History than my publications about political economy. When David Armitage and his colleagues kept inviting me to address the History Department at Harvard about the future of the profession, I began to believe that the field needed a work of theory to pull together the major questions that I had seen emerge in the classroom – for instance, whether a data-intensive practice of History would also be a form of History that asked questions over decades and centuries rather than months and years.

Choosing to enter the conversation in Digital History was in many respects a gamble, and it also put me in a difficult position in certain respects. In North America, a majority of positions in the History field are still structured by the shape of nineteenth-century nation-states; thus one is hired for a permanent position as a professor of Britain or British Empire or of China or Africa, but not as a "digital historian" or "professor of historical methods."

Because most jobs that are available are those that were available in the past, I took tenure-track positions in two departments of history, and in both cases I was hired as a British historian. In both of these institutions, I had some colleagues who were convinced that my publications and teaching should center on the question of how Britain changed over time, and only secondarily about question of methods. In the tenure and promotion process, I found myself required to work very hard to explain that I was talking to two communities – the British studies community and the community of historians and social scientists as a whole, for whom the question of methods was the most relevant question. Certain senior colleagues felt firmly that my packet should be reviewed by "traditional" historians rather than the colleagues whom I was engaging in more recent work on digital methods, or that my teaching was letting down the department by focusing on questions like the development of capitalism or the promise of new methods. Those battles were probably particular to individuals and specific departments at a moment at time, but they required a great deal of energy. In that sense, choosing the "Digital History" track came with extra burdens that simply would not

have existed had I exclusively pursued the road laid out for me – that of a traditional historian of British political economy.

Meanwhile, I felt strongly that I was responding to a call that came not merely from the University of Chicago and Harvard, but also from the many journal editors and conference organizers and dozens of history departments around the world that had summoned me to talk about the implications of data for historical practice. One definition of "service to the profession" in our trade is answering the questions that colleagues believe should be taken seriously. For me, answering the call to think critically about the implications of a data-driven society for the university and the profession was a form of service to my colleagues and students.

The Research (Why I began to code)

Anaclet Pons – *Let's take a look at your work. Since the publication of your first book,* Roads to Power[1]*, you have been involved in the research and development of new forms of infrastructure to support the work in the digital humanities. Could tell us about Democracy lab?*

Jo Guldi – *Roads to Power* was a history of infrastructure-building in Britain during the century when the British state began to invest in improvements to roads, ports, and lighthouses. It employed studies in political economy, the history of technology, and social history to understand the roads, who built them, how money was diverted at a national level rather than a local level, and how working-class travelers used the roads to imagine their own communities. My questions about infrastructure thus began as a historical set of questions about the origins and implications of modernity.

When I began teaching and writing about the implications of data for historians and for the modern university more generally, there was a natural continuity between my inquiries about the history of eighteenth-century infrastructure and infrastructure in the present day. I found myself teaching a history of capitalism course that traced the uses of technology to shape economic systems and public space from the draining of Northern Europe in the fourteenth century through the twentieth century. I also found myself wrestling with questions about how nineteenth-century socialism resulted in a proliferation of new

1. Jo Guldi, *Roads to Power: Britain Invents the Infrastructure State*, Cambridge and London, Harvard University Press, 2012.

forms of infrastructure – from public parks and public housing to sewers, sidewalks, and urban planning departments. I wanted to know at what point "participatory" projects began to raise questions about whether expert engineers were needed to guide technology.

I was asking questions about the history of infrastructure, participation, and expert rule in part to give my students a historical handle on how to understand contemporary debates about social media – where utopian technologists claimed that a world of exchanges on the internet portended the expansion of democracy through the Middle East, while journalists revealed unpleasant truths about how private capital and authoritarian leaders were employing the same technology for the purposes of social control.

I became convinced that a responsible response was not merely to study infrastructure but also to carefully choose the shape of the scholarly infrastructure in which I myself invested. I was motivated by watching projects like Matthew Desmond's EvictionLab, where sociologists mapped eviction in contemporary America by census tract, thus giving important tools to activists hoping to make a difference in eviction policy in the contemporary United States. How might historian support self-knowledge and critical reasoning about the past and the present in contemporary politics?

What I knew was that historians of North America, Latin America, Europe, Hong Kong, and Africa have access to digitalized versions of the parliamentary and congressional debates of many nations. I reasoned that as my lab worked on refining white-box tools for understanding historical change, we should make our work available to activists, citizens, and teachers who wanted to conduct their own inquiries.

Democracy Lab put one emphasis on robust tools for examining historical change that would allow researchers to see aggregate trends as well as the voices of the dispossessed. But because of our questions about infrastructure, we put an equal emphasis in grant-writing and building on creating public-facing infrastructure to allow researchers, activists, teachers, or other citizens to ask questions of their own. With our tools, it's possible to ask how Nancy Pelosi's speech changed in Congress over a decade, or how she differs from another speaker in Congress. The interface is intended to mirror the approach to text mining for historical analysis laid out in my new book, *The Dangerous Art of Text Mining*, where I pair historical theories of event, periodization, and memory with algorithms from data science, within the context of an iterative,

critical approach that takes for granted that multiple, complementary interpretations of the past are possible.

Democracy Lab is still in the process of looking for funding, but we have prototype versions of the app available for public use.[2]

Anaclet Pons – *One of the first projects you worked on, with Cora Johnson Roberson*[3]*, was Paper Machines, a free toolkit for historians. Can you explain what it was and why it is not running today?*

Jo Guldi – *Paper Machines* was my first experiment in designing infrastructure that mirrored scholars' concerns with synthesis and transparency. A plug-in to the open-access app Zotero – which has been described as "iTunes for scholarly citations" – *Paper Machines* allowed a researcher to visualize the common features of a set of text-based documents, for example the most frequent words over time.

The problem with scholarly infrastructure is that it requires maintenance and ongoing development. The scholars who have continued to develop robust tools for scholarly infrastructure – for instance the authors of Zotero, EvictionLab, the Atlantic Slave Trade database, and the Old Bailey – have grown research labs around them, typically with the support of their institutions. As an assistant professor who was the lone methodological specialist on the tenure track at institutions that concentrated on the "traditional" history of nation states, I had to set aside some of my scholarly ambitions. I wrote grants to support research on the history of political economy rather than grants to aid the development of infrastructure. As a recently promoted full professor heading to an institution with a long trajectory of library support for scholarly infrastructure projects, I'm optimistic about my future ability to support meaningful infrastructure for the public.

The lesson for other scholars here is that infrastructure-building is not a one-scholar game; it depends on being at one of the few institutions that have made a meaningful commitment to research infrastructure. The lesson for the profession is that the infrastructure that supports the scholarly research of the future is a public good – and as such, it deserves public support – and public debate – from the profession in ways that have yet to be debated.

2. The version for the British House of Commons and House of Lords is here: github.com/stephbuon/hansard-shiny. The Congress interface is here: github.com/stephbuon/congress-shiny

3. corajr.com/

Exposed to Big Data

Anaclet Pons – *You say that you designed Paper Machines to help with your next monograph,* The Long Land War[4]*, a history of land reform movements in the twentieth century. How did it help you? In other words, you were immediately exposed to big data. How has that changed your fellowship?*

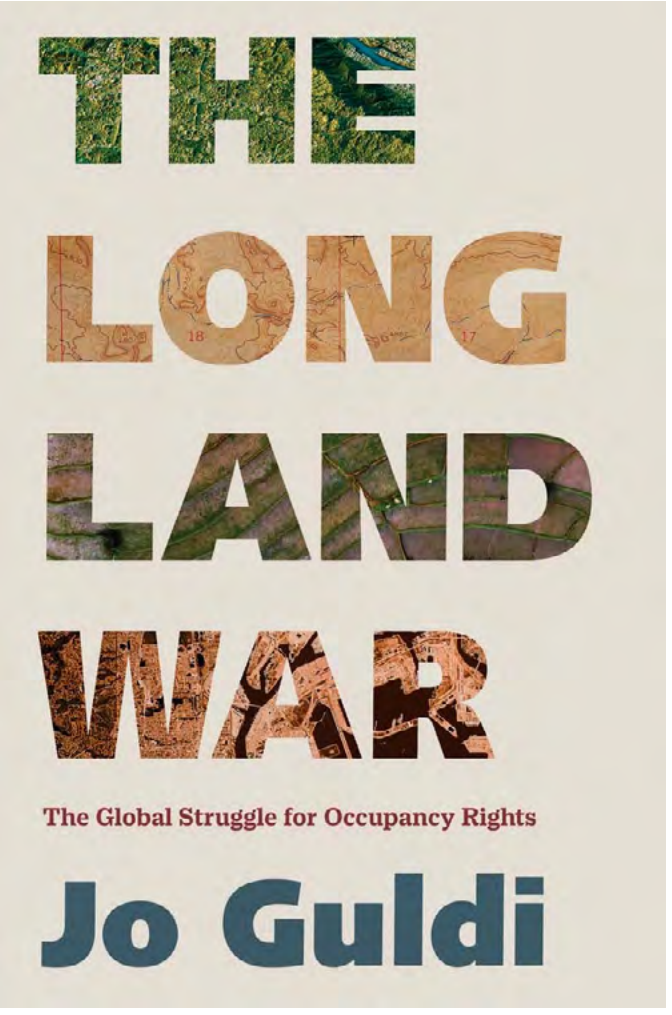
Jo Guldi – I originally planned to use digital technologies to investigate the history of political economy over two centuries. Through the process of engaging with archives and refining my sense of how methods should be applied, that one project split into four separate projects – *The Long Land War*[5], *The Dangerous Art of Text Mining*[6], *How Not to Kill Your Landlord*[7], and a manuscript called *A Distant Reading of Property*. In the first of the books out – *The Long Land War* – I leaned on a global collection of twentieth-century archives and ultimately chose not to publish my early experimentations with text mining methods.

Anaclet Pons – *It is inevitable to mention* The History Manifesto[8]*, co-authored with David Armitage, on the role of history and the humanities in a digital age. What is your view today, with the benefit of hindsight, of the criticism this book has generated?*

Jo Guldi – The published record of debate[9] contains two piles of reactions – a large set of positive reactions, and a smaller, but extremely heated, set of counteractions that denounced the *Manifesto,* typically making some version of the following complaint: that the *Manifesto* didn't describe all of the practices of historians; that digital tools were untried and that we hadn't provided any case studies in our short pamphlet, and that micro histories of gender, race, and class needed to be defended in a more digital age. The last headings of reactions is really the only ones that matter, and they have been spelled out more robustly

4. Jo Guldi, *The Long Land War: The Global Struggle for Occupancy Rights*, New Haven, Yale University Press, 2022.
5. Jo Guldi, *The Long Land War: The Global Struggle for Occupancy Rights*, New Haven, Yale University Press, 2022.
6. Jo Guldi, *The Dangerous Art of Text Mining: A Methodology for Digital History*, Cambridge, Cambridge University Press, 2023.
7. Jo Guldi, *How Not to Kill Your Landlord*, New York, Dutton, forthcoming.
8. Jo Guldi and David Armitage, *The History Manifesto*, Cambridge, Cambridge University Press, 2014.
9. See: "AHR Exchange – On *The History Manifesto*," *American Historical Review*, vol. 120, n° 2, 2015. DOI: doi.org/10.1093/ahr/120.2.527; "La longue durée en débat - Histoire des sciences," *Annales. Histoire, Sciences Sociales*, vol. 70, n° 2, 2015.

in a separate debate in recent years through the work of interlocutors like Jessica Marie Johnson[10], as well as STS critiques of the bias of digitalized archives from writers such as Cathy O'Neil[11].



Jo Guldi, *The Long Land War: The Global Struggle for Occupancy Rights*, New Haven, Yale University Press, 2022.

10. Jessica Marie Johnson, "Black Beyond Data," *Arts & Sciences Magazine*, vol. 19, n° 2, 2022.

11. Cathy O'Neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*, New York, Crown, 2016.

Historians who practice text mining have learned a great deal from these debates. Colleagues like Lauren Klein[12] and Richard Jean So[13] have dedicated their career to showing that text mining can help us to study the history of exclusions of race and gender. In *The Dangerous Art of Text Mining*, I spend the first half of the book reviewing these debates, underscoring the use of critical history to reveal unpleasant truths about society and institutions, and reviewing how text mining can be an ally in this struggle.

Anaclet Pons – *There are two central themes in the book. On the one hand, the crisis in the humanities, linked to the fact that historians, in particular, have retreated from the longue-durée, generally privileging short-termism. On the other hand, there is an opportunity to reverse this trend by taking advantage of the availability of digital data, the tools for analyzing it, and the methods of communication. That is, to use of distant reading, data visualization and other digital tools specifically designed to answer broader historical questions. In this way, you say, historians could be critical arbiters of the flood of data that surrounds us. Are Digital Humanities, partially, an answer to this identity crisis? What do you think of the AI chatbot and how do you think it could impact disciplines such as ours?*

Jo Guldi – In the new book – *The Dangerous Art of Text Mining* – I return to this argument in a new form: the world of GPT is a world flooded with automatically generated text that bears a specious relationship to historical fact. I spell out a theory and offer case studies to support a practice of historically-robust text mining that can be used to extract factual perspectives on the past. I work within a framework that I call "critical search," where I renounce from the beginning the possibility of a single metanarrative on the past, but instead embrace the possibility of many true perspectives, equally built on fact, each of which refracts the bias of the archive, theory, and algorithm use to create a perspective on History. In other words, it is possible to use one algorithm to create a *longue-durée* history of working-class perspectives that appear in the parliamentary debates of Great Britain, and another algorithm to create a *longue-durée* history of how parliamentarians in aggregate spoke about working-class persons. These two perspectives complement each other and give a more complete view of the past. But both of them are based, ultimately, on counting words that were recorded

12. lklein.com/
13. mcgill.ca/english/staff/richard-jean-so

at moments in time, and which can be corroborated with the archival record for further support.

I remain a serious believer that our age of truthiness requires an embrace of enlightenment and critical perspectives from the humanities and social sciences. We see in GPT that irrational exuberance from Silicon Valley is capable of creating an abundance of tools which are not well designed to complement the need for historical truth. Journalists are already writing about what happens when a lawyer plans a case based on the half-truths or actual falsehoods manufactured when GPT attempts to synthesize textual documents. The university students of today will in the near future enter professions like the law, policy, social work and engineering where they will be responsible for actual facts. They need tools that are refined enough to accurately extract information from text. Reviewing the truth-value of facts about the past is one of the skills taught in History Departments and other social science and humanities departments. There's a real need for individuals with the skills to examine the origin and distortion of fact to engage the mechanisms by which fake facts are being manufactured in our own age.

Competently engaging facts today requires the skills of critical thinking about archives, where facts come from, as well as representation, and how easily facts can be distorted – but more than that is required. True competence in auditing today's GPT technology requires a detailed knowledge of how the algorithms work, what biases they contain, and what biases are structured into the archives they use as their sources. So tomorrow's citizen needs an apprehension of both data science and critical thinking from the liberal arts.

A handful of university programs today have recognized this urgent need. At Pennsylvania State University and Emory University, undergraduates can major in programs like Quantitative Methods, where data science and machine learning are taught alongside the skills of critical thinking and social theory from the humanities and social sciences. When I was offered a position in Emory's QTM department this past year, I leapt at the opportunity. Emory's 600 QTM majors take courses that elsewhere are taught in a Data Science program by PhD's in engineering. At Emory, however, those 600 QTM majors are being taught by PhD's in political science, history, literature, economics, and statistics. The students learn statistics, but also critical theory. They think about algorithms in detail, but they also reason about gender and race bias. They will learn from me how to understand the bias of historical archives implicit in the reckoning of algorithms from Silicon Valley, and

therefore how to productively engage their skepticism, to review facts, and to build algorithms that serve society better.

I see critical thinking about data as an inevitable frontier for the university of a data-driven society, and I believe that Emory is pioneering a model that other universities would do well to follow. This kind of labor is not for every individual with a History PhD, but it does have certain implications for how members of History Departments talk to their deans and provosts about what History has to offer, and why new hires and graduate students in History Departments might be needed in the future. It remains vital for the preservation of a culture of fact that Data Science be taught from within a Liberal Arts tradition, where the skills of critical thinking remain part of the perspective. As the traditional site of the examination of archives for historical fact, History departments have a great deal to offer any university that takes such a turn.

I'm less certain what the future of History Departments is with respect to the use of data. The numbers tell us that History majors are down, while demand for Data Science degrees remains high. In any department suffering from a fall in enrollments, I would ask colleagues to consider whether a modern History Department should not offer a course on the "history of the modern fact" as a requirement for an undergraduate degree – or at least for any degree in History. Hires who are capable of teaching a deeper engagement with historical truth via the algorithm have even more to offer such a shifting curriculum – although there are very few individuals today who gain training in working with algorithms as part of their History PhD.

Anaclet Pons – *In the book* – The Dangerous Art …– *you talk about a revolution in the use of macroscopic data, of counting things that can be measured, as a guide to doing history, to looking at the past. Does big data improve our ability to handle historical information? Is it a return to the quantitative social science history based on statistical analysis? Is Digital History the exclusive domain of big data? Does it necessarily privilege quantitative over qualitative? Is it rather a different approach to data and its analysis?*

Jo Guldi – The data-driven strategy that I use and recommend for historians today is text mining for historical analysis, an approach that begins with counting words over time. Because the material of study is text, questions about bias and representation are still at the forefront of the issue. The cultural history issues of how women are represented in the text are exactly what we study – for instance, when we use linguists' algorithms to extract all the adjectives paired with words for women and girls over the nineteenth century. Issues of intellectual and political

conceptualization are similarly relevant, as are the issues associated with the "linguistic turn" in social history, where historians have asked how working-class people and ethnic groups described themselves and their ambitions for political order.

These approaches are a far cry from the methods of the "Quantitative Turn" in History during the 1960s and 1970s, when large-scale databases of wages and nutrition were used to make sweeping arguments about the history of slavery which may historians viewed as reductive in the extreme. While demographic history still exists in some places, those questions are routinely handled in departments of Economics in North America. In Britain, the new practitioners of Digital History often use maps and geographical data sets to inquire into the effects of the railroads on the age of marriage and birth rates.

Meanwhile, a majority of historians in History departments in North America today remain concerned with issues of representation. To these historians, any given tool from beneath the enormous umbrella of the digital humanities may or may not be useful; they may or may not believe that archiving historical objects and images with the Omeka software is useful to themselves and their students. A network analysis may or may not reveal information about the characters in their story.

But I believe that text mining is exceptional in this regard. The tools of text mining can be used on virtually any project about which there some large-scale digitalized repository of text is relevant – whether that repository is the German novel, the Dutch newspaper, the political debates of Congress, or judicial rulings of some court.

Other ways of understanding the past?

Anaclet Pons – *What does text mining offer in terms of innovative ways of understanding the past?*

Jo Guldi – Text mining for historical analysis allows researchers to make visible hitherto invisible patterns of change over time, pinpointing events and periods that are distinctive with respect to a style of representation, a concept, or a material concern. In *The Dangerous Art of Text Mining*, I offer case studies which include looking for all the ways that speakers in parliament talked about the future, decade by decade, or all of the historical events referenced in parliament.

I believe that success in text mining is a matter not merely of "innovative" algorithms. Some of the best work in Digital History engages old algorithms from the 1950s and 60s, putting vintage math into dialogue

with modern questions about gender, race, and class. Asking those questions with concern not just for the most numerous events – but also for the events that reflect concerns about colonies or the working class – means being able to drill down from the aggregate overview to a perspectival view where word count helps us to think critically about how the institution spent its time.

Anaclet Pons – *Does digital scholarship allow you to do things that you wouldn't be able to do in a traditional archive?*

Jo Guldi – No researcher would be able to read a century of parliamentary debates to annotate every sentence where a speaker refers to events "in the future" or "years from now." With an algorithm, we can automate that search. Similarly, we can look for all of the sentences where women or girls are the subject of the sentence, and we can ask what the verbs are – what is it that speakers in parliament imagining women and girls doing? And we can ask the inverse question: what are all the verbs where women or girls are the object of the verb – what is imagined as being done to women and girls?

Questions of this kind allow us to aggregate information about ideas and representation in political texts. We can look for the "average sentence" of a year, a decade, or a particular speaker. We can ask highly abstracted questions about how different institutions imagined Britain, its colonies, or the rest of the world.

Anaclet Pons – *The "digital turn" has changed the way almost all of us access and search for sources, analyze historical content and present our research. How are the new tools, methods and perspectives changing the way in which people think about and do scholarship? Will some of the focus of the previous scholarly record be displaced by the new goals of Digital History?*

Jo Guldi – During the Pandemic, digitalized newspapers offered a life-line for many researchers and students who no longer had access to physical archives. Some very good books were written, and sometimes those books took a new angle that was incredibly illuminating – I think of Niall Whelehan's *Changing Land*[14], which used newspapers to reconstruct an international Irish diaspora on three continents, tracing a *longue-durée* story about actors and ideas that would have been impossible with traditional archives. The Pandemic showed many scholars the wealth of resources available in digitalized form. But for

14. Niall Whelehan, *Changing land: diaspora activism and the Irish Land War*, New York, NYU Press, 2021.

every such book, there were also a dozen historians who could not wait to get back to their archives. As a group, historians are highly sensitive to what is left out of any given archive. Few historians are naïve about the limits of digitalized newspapers or the Hathi Trust or other online repositories as a resource, and I think that's for the good.

What we did not see during the Pandemic was a wholesale movement towards text mining, or the application of algorithms to digitalized archives in order to reveal trends that are difficult to detect by simply reading. The field as a whole is a long way from embracing text mining for historical analysis, which holds the greatest precision for asking nuanced questions about representation and how it has changed.

Anaclet Pons – *You say that digital tools should be conceived as a new form of philology and point out that digital analysis, using the tools available today, is a technique that should be handled with the same scrupulous caution as any other historical method. Can you explain it a little more? Any common mistakes to avoid?*

Jo Guldi – In *The Dangerous Art of Text Mining* I recount a story from the classroom. Given a count of the most frequent adjectives applied to women by speakers in Britain's parliament, the students noticed that a frequent phrase was "ignorant women." They charted the count of the phrase over time and noticed that it peaked in the late nineteenth century. Asked to interpret the phrase, the students ignored my advice of going back to the text to ask how speakers in parliament were using the phrase. They reasoned from the graph itself and wrote a response paper arguing that Britain had suffered in the nineteenth century a plague of ignorant women.

What the students were missing, in this case, was a sense of the bias of the archive and the importance of interpretation given outside context. Most history majors who have some experience of careful reading would understand that a rise in the count of "ignorant women" is a signal of a moment of prejudice among the speakers in parliament, and it is unlikely to correspond to a demographic trend of ignorance in the population. The historian might grow curious about the context in which this phrase was used, and might move from word count to close reading, and discover that "ignorant women" was a phrase commonly invoked around the time of the Contagious Disease Acts, when Britain began regulating prostitution. The prominence of the phrase is an indication of how members of parliament interpreted the spread of syphilis; they regarded it as the fault of "ignorant women," largely

exonerating the soldiers who visited prostitutes from any responsibility for the spread of the disease.

I think there's an important lesson here for historians about how to talk about the skills of history, and why the skills of historical interpretation remain vital even in a data-driven age. One might say that the risk of universities with popular Data Science programs risk turning out legions of students who cannot interpret a chart of the count of the phrase "ignorant women." Taken to an extreme, the thought experiment is unpleasant. A world of data analysts who lack a liberal arts education risks reducing contemporary political discourse to the level of the nineteenth century – where prejudice is unquestioned because social facts remain invisible to the uncritical imagination.

Just because people count things doesn't mean that data workers don't need historical methods, social theory, or critical thinking. On the contrary, I believe that those of us who count things – especially those of us who count words -- need those skills all the more.

Combining historical research and digital methods

Anaclet Pons – *Would you say that Digital History requires more than just a diverse set of tools, but that it also requires a different way of approaching it? What is the biggest challenge in combining historical research and digital methods?*

Jo Guldi – In my chapters on "critical search," I lay out a strategy for engaging secondary sources, data-driven analysis of text, and traditional reading of primary sources. I believe that opportunities for critical thinking enter into the process whenever one shifts from one more to the next – and that the more opportunities, the better. In practice, this means that it's nearly impossible to adequately interpret the phrase "ignorant women" by just calling up a graph of word count over time; a researcher needs, at minimum, to read the passages of text where this phrase is used, and better, to read some informed secondary sources that treat the problem of historical prejudice against women in the time of the Contagious Disease Acts. For students with a background in Computer

Science or Data Science, this task is very difficult, but it comes easily to students of History or other social science and humanities fields.

Anaclet Pons – *You say that you realized that you had a lot to say about the nature of texts as objects, and that these questions led you to the code. Do you think that this is understood by all, or most, academics?*

Jo Guldi – There's a real gap between the appreciation of how to treat text in the liberal arts and in the information science fields. Many of the publications that purport to research history published by researchers from Computer Science, Data Science or Informatics look very little like research – or historical fact-finding – to our eyes. That's because there's been so little dialogue across those disciplines about the value of the historical method, about when finding a signal in the past constitutes a historical discovery.

The problem of understanding society and how it is changing is too valuable to be left in the hands of people who simply do not work with questions of archive, sources, bias and truth.

Anaclet Pons – *What does it mean to be a digital scholar? Because there is a lot of heterogeneity in the field.*

Jo Guldi – I think Ian Milligan[15] is right when he says that all scholars in the humanities are digital these days. Don't almost all of us use Microsoft Word, search library catalogs, and keyword search digital repositories? Milligan make a good case for teaching students to be critical about how they engage this vast array of technology.

While almost every historian can integrate some of Milligan's ideas into their syllabus, there is a smaller set of historians who have actively engaged with refining data-driven methods in dialogue with theories of history and historical interpretation. This is where the real work is being done, because the methods of history are so specific to a literature not widely consulted beyond our field. Combining attention to algorithms and databases with care for the concerns of historical theory and social theory can produce a really robust practice with data capable of supporting work that reads institutional archives against the grain, to borrow a phrase from postcolonial history. At the moment there are

15. ianmilligan.ca/. See: Ian Milligan, *The Transformation of Historical Research in the Digital Age* (Elements in Historical Theory and Practice), Cambridge, Cambridge University Press, 2022. DOI: doi:10.1017/9781009026055

a dozen practitioners actively contributing to this set of methods, and the majority of the practitioners are in Europe.

Anaclet Pons – *Ian Milligan argues that the profession is marked by a mood of resistance to quantitative learning, an attitude that has structured the recent history of the profession and shaped another generation of scholars who are ill equipped to deal with the archives he has surveyed. Do you agree? Do you think the problem is that history is a system of knowledge developed in the 19th century with problems to remain valid in our digital world?*

Jo Guldi – Luke Blaxill[16] has also made a similar argument. Meanwhile, the Macarthur Prize just went to a demographic historian, Steven Ruggles[17], whose scholarship suggests that history journals in aggregate are publishing more quantitative data than they were ten or twenty years ago.

There is no question in my mind that the historical method is on the ascendency, and must remain on the ascendency in a world that values fact. The historical method implies studying change over time by comparing sources and by tracing every fact to the sources that support its fact. It is implicit today in the practice of law and journalism, and it's also absorbed in fact-finding in most of the social science and humanities field.

What Blaxill, Ruggles, and Mulligan have argued is that in most nearby fields – including law, but also Art History and Political Science – technology has been slowly but surely been adopted over the past several decades. History has been resistant, particularly in North America, where to my knowledge there are no leading programs of History which teach digital methods in a robust way, perhaps with the exception of Columbia, where Matthew Connelly's lab[18] uses text mining to investigate declassified documents and where Ira Katznelson[19] has been leaning into how Machine Learning approaches can create rigorous historical findings, or Princeton, where Matthew Jones[20] has long been teaching a history of statistics class that requires students to run sample data sets in R.

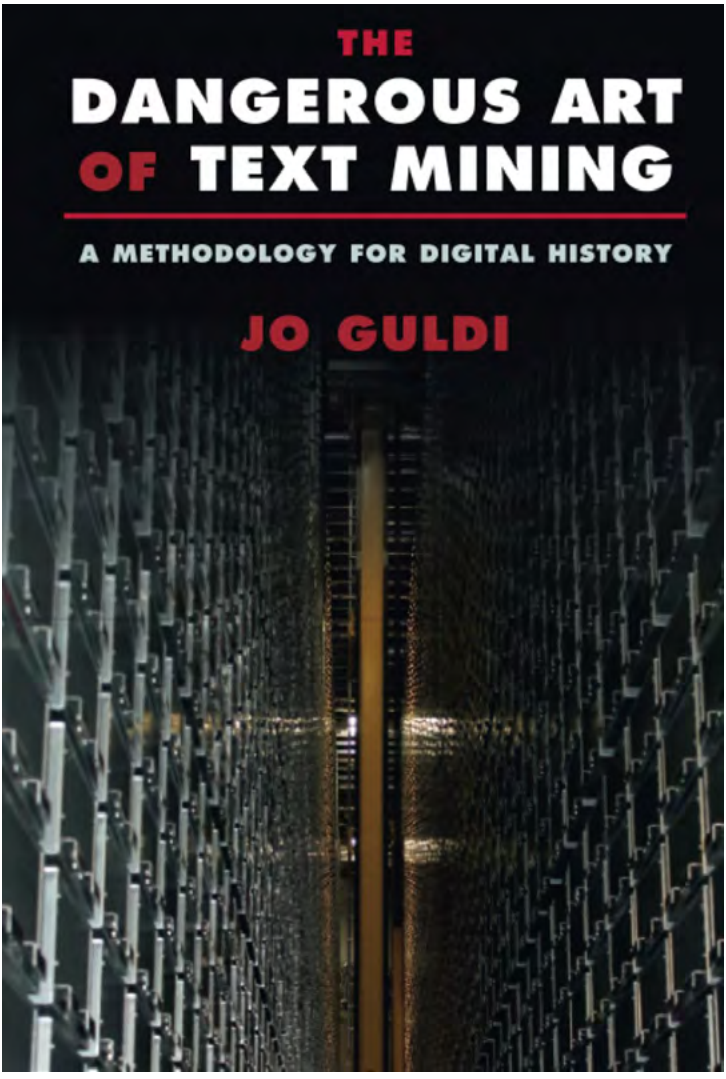
16. lukeblaxill.com/
17. macfound.org/fellows/class-of-2022/steven-ruggles#searchresults
18. matthewconnelly.net/
19. polisci.columbia.edu/content/ira-i-katznelson
20. nescioquid.org/contacts.html

Imagine: the graduates of that program will have at their fingertips skills unparalleled by any other PhD's of their generation.

The History Manifesto attempted to make an intervention in this structure, but I would be quick to point out my own ignorance of why the apparent divide exists. In the *Manifesto*, we suggested that the "two cultures" described by Snow structure this resistance. On anecdotal evidence, senior faculty who themselves feel uncomfortable with a spreadsheet are unlikely to require their own PhD students to learn new methods. Yet we're a diverse lot, and I know senior faculty at many programs who have professional backgrounds in cybersecurity or coding that required them to have many advanced digital skills. It's worth considering that the direction taken by History Departments as a whole is informed by a collective imagination richer than any individual psychology of hostility or enthusiasm. For indeed, we are an enlightenment field.

History is sometimes a conservative discipline because its loyalty is above all to truth; historians will not accept a method on the basis reputation. They must ultimately feel the it has been vetted and that they can vouch for the facts produced. A black box is insufficient. There has been an *avant-garde* of historians who have been pressing the methods in the search to create a historical method for ascertaining the facts of change over time. It has taken a decade or more of research to produce methods that are sufficiently robust to support historical questions of the multitude and nuance that characterize today's historical practice. But now that they exist – and now that they have been vetted in the highest journals in the field and embraced by some of the best practitioners in every field – I believe they will be emulated.

I cannot predict exactly the shape that expertise will take in the future. Columbia is not alone. A handful of History Programs like Clemson have committed to a multiple-semester track in digital methods. But whether the instruction in new methods will take place only at a minority of elite institutions like Columbia, which can support a diverse program of methodological training, or whether those skills will be taken up across the diversity of departments, each competing to show off the new methods, is an open question for now.



Jo Guldi, *The Dangerous Art of Text Mining: A Methodology for Digital History*, Cambridge, Cambridge University Press, 2023.

The distinctive features of Digital History

Anaclet Pons – *Stephen Robertson*[21] *has said we need to understand the distinctive features of Digital History, that distinguish the discipline from digital literary studies in particular, arguing that historians are different in the way they use the Web and in the computational tools they favor. Do you share this view?*

Jo Guldi – Stephen is absolutely right, and he's one of the historians who helped me to understand this. I spent the first part of my explorations of the digital trying to catch up with progress in digital literature and social science. It was through the good influence of folks like Stephen Robertson, Lara Putnam[22], Lauren Klein, Lincoln Mullen[23], Melvin Wevers[24], and Tim Hitchcock[25] – including a lot of informal conversations about the future of the field – that I eventually re-centered my investigations on the problem of change over time. I learned a great deal from the theory of history people as well – especially conversations with Ethan Kleinberg[26] and Stephen Tanaka[27]. The editors at the *American Historical Review* also really pressed me in this direction.

My newest book, *The Dangerous Art of Text Mining*, directly grows out of those conversations. It forwards a series of approaches to text mining that directly respond to specifically historical questions like periodization and memory. There is more to be done, but my survey of existing methods pulls together a variety of strategies for understanding why a month, year, or decade differed from those before it. Between the approaches, it is possible to capture underrepresented voices even while using other tools to characterize the bias of institutions over time.

I call my approach "text mining for historical analysis" rather than "Digital History." The interventions described in the book are tailored specifically to historical questions, drawing on the theory of history.

---

21. Stephen Robertson, "The Differences between Digital Humanities and Digital History," *in* Matthew Gold and Lauren Klein (eds.), *Debates in the Digital Humanities 2016*, Minneapolis, University of Minnesota Press, 2016, pp. 289-307. DOI: doi.org/10.5749/9781452963761; and Stephen Robertson, "The Properties of Digital History," *History and Theory*, vol. 61, 2022, pp. 86-106. DOI: doi.org/10.1111/hith.12286
22. history.pitt.edu/people/lara-putnam
23. lincolnmullen.com/
24. melvinwevers.nl/
25. sussex.ac.uk/profiles/336034
26. wesleyan.edu/academics/faculty/ekleinberg/profile.html
27. Stephen Tanaka, "The Old and New of Digital History," *History and Theory*, vol. 61, n° 4, 2022, pp. 3-18. DOI: doi.org/10.1111/hith.12284

Because these strategies directly reflect the historical method, they will be immediately more relevant to most historians. I would imagine that an approach so centered on history will feel much more relevant to many colleagues.

If we contrast text mining for historical analysis to the broader umbrella of digital humanities approaches – for instance, collecting and annotating digital images, or marking up text to annotate rhetorical strategies – those approaches sometimes feel peripheral to the central task of characterizing change over time. Perhaps they feel way like a couture skillset, relevant to certain scholars, but overall tangential to the practice of history.

In contrast, practically every historian deals with making sense of change over time through reckoning with large collections of text at some point in every project. So text mining for historical analysis is highly relevant to practically every historical project. For that reason, my hope would be to see these approaches being taught in leading programs in history. Text mining for historical analysis will directly feed into historical research for any researchers who have a robust, digitalized corpus of text where they need perspectives on turning points and other changes in representation and conceptualization.

While my book emphasizes the historical method, there are also a variety of approaches from outside of history, informed by other questions in the social sciences and humanities, which remain relevant to certain historians. I believe the tools that were pioneered by our cousins in Literature and Sociology have covered a great deal of ground, and they may be of terrific importance in certain cases. For instance, the work by Richard Jean So, Austin Kozlowski[28] and James Evans[29] on understanding differential representations of race[30] has many applications in our field. In literature, Andrew Piper[31] and Ted Underwood[32] have been leaning into *longue-durée* studies of change over decades, pioneering

28. austinkozlowski.com/about-me/
29. sociology.uchicago.edu/directory/james-evans
30. Austin C. Kozlowski, Matt Taddy, and James A. Evans, "The Geometry of Culture: Analyzing the Meanings of Class through Word Embeddings," *American Sociological Review*, vol. 84, n° 5, 2019, pp. 905-949. DOI: doi.org/10.1177/0003122419877135
31. piperlab.mcgill.ca/about.html
32. ischool.illinois.edu/people/ted-underwood

tactics that many historians will want to follow. But our field is vast, and the diversity of practitioners will probably borrow a diversity of tactics.

Anaclet Pons – *You say that technology, like nature, has always been with us. Indeed, there is something elemental in the modern transformation of scholarship. New work with algorithms potentially participates in such a tide, and what is thrilling about it is the sense that any scholar, anywhere in the world, might contribute to its movement. So, Digital History is not so much a field or sub-field in this rich and varied landscape, as a universal approach to history*[33]*. What do you think will be the biggest challenges in the field in the next years?*

Jo Guldi – One version of how things go is that most history departments in North America never hire a specialist in digital methods, content to hire researchers and students who are passive consumers of online tools for inspecting digitalized archives that have been designed by for-profit companies like Elsevier.

In the best-case scenario, Elsevier's content designers would integrate the new methodological approaches developed by people like me, and all researchers are able to use them. Reading a few methodological articles or books during a historiography class would be sufficient to enhancing the training of most students so that they understand the capacity of the new methods and can integrate them with their work. Such a transition would probably feel relatively seamless. A few elite universities might decide to concentrate on hiring specialists in methods, developing programs for the training of methodologically-astute students who might go on to enhance the new methods, their innovations embraced by a generation of students via the search interface developed by private companies.

In the worst-case scenario, however, a company like Elsevier might hire data scientists without consulting with the community of digital historians. Or they simply might not upgrade their infrastructure. Then, the vast majority of historian researchers and students would be stuck using essentially inherently biased tools that at best limit what kinds of signals they can search for and at worst actually distort their impression of history, channeling historical research in a biased direction, that is, the world where the most quantitatively numerous signal matters the most. Careful readers of social history from the past and teachers of

33. See: Daniel J. Story, Jo Guldi, Tim Hitchcock, and Michelle Moravec, "History's Future in the Age of the Internet," *The American Historical Review*, vol. 125, n° 4, 2020, pp. 1337-1346. DOI: doi.org/10.1093/ahr/rhaa477.

social theory would caution that numbers aren't anything, but biased tools tend to reproduce biased analysis. In such a world, a few specialists in digital methods like me might continue to do our work, but it would be regarded as peripheral to the everyday practice of historical research. Meanwhile, in ignorance, students will execute keyword searches through antiquated and biased frameworks, neither students nor faculty cognizant of the fact that the algorithm has distorted their findings all along, nor that another more robust set of findings lies just out of view.

There are other possibilities, to be sure. The terms of engagement with research need not be set by the profit models associated with private companies. Just as archivists and scholars collaborated to establish the HathiTrust as a nonprofit serving scholars, archivists and historians could collaborate in designing research infrastructure. Or a generation of scholars might become intrigued by the possibilities offered by text mining for historical analysis, and, a few scholars at a time, individuals might decide to engage the new methods, motivated purely by a sense of competition for the most robust argument and the most innovative perspective on the past. It is clear that the new methods can produce discoveries that could not be produced in any other way.

What I think is least likely of all is the scenario in which text mining, network analysis, and map analysis become so popular that they displace "traditional history" altogether. While I have no crystal ball, I see little evidence of this happening. The pace of change is slow, and History remains far behind other fields. The question is whether mainstream departments of History will embrace any opportunities to teach the new methods in the formal and structured way that any deep body of knowledge requires – that is, through engagement with historiographical issues, theoretical issues from across the disciplines, and minute, supervised work with detailed hands-on case studies. Short summer courses and practice cannot provide the kind of education that a PhD student needs in order to become a rigorous user of the new methods, let alone to enter into a dialogue with the international community of practitioners of text mining for historical analysis. At the moment, perhaps the only History Department to provide such a rigorous program of training in text mining is Clemson University in South Carolina.

Anaclet Pons – *I notice we have forgotten a question about your prior book. Please include one explaining what it is about.*

Jo Guldi – *The Long Land War* (2022) is a history of the land reform and rent control movements of the twentieth century – the first such history to be published in perhaps half a century. On the one hand, it

represents the fruit of a long engagement with questions of land law and occupancy rights in Britain and its empire, pursued through digital archives and traditional paper archives over three continents and a decade of research. On the other hand, the book continues my investigations into the history of infrastructure through the theorization of what I call "information infrastructure," or the vast system of maps, bibliographies, publications, and other information systems contrived by nations, nonprofits, and international bodies as tools for shaping human behavior. *The Long Land War* pivots around the fate of maps and bibliographies used by the United Nations to support small farmers in the developing world. When certain UN offices had their budgets cut in the 1970s, the map and bibliography system fell apart. In its place emerged a grassroots system of mapping village property lines, cases of pollution, and plans for local development, which was called "participatory mapping." Much of the book weighs the prejudices, strengths, and challenges of both centralized information systems (like those at the UN) against decentralized systems (like those of the participatory map).