Diplomatic History After the Big Bang

Using Computational Methods to Explore the Infinite Archive

David Allen and Matthew Connelly*

When the first edition of *Explaining the History of American Foreign Relations* was published in 1991, it would have been hard to explain to readers why historians would soon find computers indispensible for doing their research.[1] The World Wide Web was only just being born, mainly to facilitate communication among scientists. Apple's first laptop, the "PowerBook 100," went on sale that year with a measly 20MB hard drive, and cost over $4,000 in today's dollars. There were no digital cameras for the consumer market, on-site library catalog terminals were hard to use, and scholarly article databases still lay far in the future.

Nowadays, it is hard to understand how anyone ever wrote a book without word processing software, or why anyone today would not use bibliographic and note-taking tools. Whereas the first contributors to this collection a quarter century ago would have had to spend many hours at the library just to find book reviews and check citations, these tasks can today be done with a few mouse clicks. Historians are not just collecting gigabytes of digitized documents. They are organizing them as text-searchable pdfs in annotated databases, a practice that would have filled previous generations of scholars with wonder.

---

* As this essay will explain, the authors have been part of a multidisciplinary team that has been exploring applications for text processing and machine learning in diplomatic history. Much of what we have learned has come from collaboration, especially with David Madigan, Rex Douglass, Daniel Krasner, Ian Langmore, Sasha Rush, and Shawn Simpson. Their individual contributions are also noted in the discussion of specific research methodologies below. We would also like to acknowledge the generous support of the Brown Institute for Media Innovation and the John D. and Catherine T. MacArthur Foundation.

[1] Michael J. Hogan and Thomas G. Paterson (eds.), *Explaining the History of American Foreign Relations* (New York, 1991).

Remarkable though they are, these are still just technical improvements on time-honored historical tradecraft. The keyboard and screen take the place of the typewriter, databases substitute for filing cabinets and card catalogs, and track changes replace sticky notes. But the advent of digital media and recent advances in information technology portend much more dramatic changes in the very nature of our field, as is already happening in law, journalism, and literary studies. If diplomatic historians do not adapt, we may one day find ourselves buried beneath an avalanche of electronic records, far too many to cope with using traditional methods.

This essay describes how new research in computational techniques could determine how scholars work after what State Department historian William McAllister has called the "Big Bang" in historical source materials. This explosion began with the sudden release of a quarter of a million cables through Wikileaks, which in turn is now dwarfed by the 1.4 million declassified cables from the State Department's Central Foreign Policy Files, spanning the years 1973-1977. Even this pales in comparison with what is being created by newer forms of communication, including some forty million e-mails generated by Bill Clinton's administration, and two *billion* e-mails produced while Hillary Clinton was Secretary of State. How will archivists and historians possibly cope with it all?[2]

In fact, while the digital archive is potentially infinite, as William Turkel notes, archivists already have to conduct "triage" with the relatively small collections from the 1970s.[3] They lack the resources to cope with the much larger challenges to come in preservation, declassification, and curation, raising the question of whether archival integrity and proper finding aids will become a thing of the past. The search is on for technology that might accelerate the processing

---

[2] William McAllister, "The Documentary Big Bang, the Digital Records Revolution, and the Future of the Historical Profession," *Passport* 41:2 (September 2010): 12-17.

[3] "Interchange: The Promise of Digital History," *Journal of American History* 95 (September 2008): 455.

of large collections without resorting to crude sampling methods that would decimate our archive.

The risks in this new era of "big data" are great, but so too are the opportunities. Applying computational methods to massive corpora of electronic records is not only essential to preserve meaningful access to the archival record. It may become possible to provide precise answers to heretofore intractable questions, such as the rank order of different issues and different areas for policymakers, and how, specifically, official secrecy distorts the historical record. It will also prompt entirely new kinds of questions, now that we can begin to visualize the flow of information through the foreign policy bureaucracy, in Washington and across the wider world of American diplomacy. All this may require fundamental changes in the practice of history, such as learning programming languages in addition to foreign languages, sharing our "data" with other researchers, and forming multi-disciplinary teams to conduct large-scale experiments.

Diplomatic history is unusually rich in documents, most of them in the public domain and many already available in digital form. Historians of American foreign relations therefore have a unique opportunity to lead the way, developing methods and standards that other scholars will learn from when Facebook and Twitter become the archive of contemporary social and cultural history. But reviewing previous attempts to apply computational methods to history reveals many pitfalls, whether neglecting older but still important questions and sources, raising false expectations of scientific objectivity, or drawing younger scholars into what may turn out to be methodological dead-ends. Even so, perhaps the greatest risk of all would be to not rethink how we approach archival research when the archive itself is about to explode.

Computational Methods 1.0 and 2.0: From Cliometrics to the Digital Humanities

Great claims about the potential of statistics and computers in historical research are not new. As long ago as Frederick Jackson Turner there were calls for a quantitative political history, focusing on roll call voting and election returns, which came of age in the mid-twentieth-century in the work of William Aydelotte and others.[4] In the 1950s, social science historians began to use punch cards and sorters to study urban, demographic, and African-American history. By the 1960s, a growing interest in quantitative and statistical history was spread across a wide variety of fields.[5] Economic historians started to array data and analyze them with novel modeling techniques, dubbing themselves econometricians or cliometricians.[6] Across the Atlantic, the historians of the *Annales* used perforated tape to create data series for commodity prices over the *longue durée*, in what they called "serial history."[7] Articles proliferated in the *Journal of American History* and the *American Historical Review* announcing the need to train graduate students in statistical methods, even computer programming, and lauding what computers could achieve.[8] Books released in the field described armies of research assistants and thousands of hours of data input, and came with supplementary volumes full of tables and mathematical formulas. The most zealous members of

---

[4] William O. Aydelotte, "Quantification in History," *American Historical Review* 72 (1966): 803-825.

[5] Robert P. Swierenga, "Computers and American History: The Impact of the "New" Generation," *Journal of American History* 60 (1974): 1045-1070.

[6] Robert W. Fogel, "The New Economic History," *Economic History Review* 19 (1966): 642-656; Robert W. Fogel, *Railroads and American Economic Growth: Essays in Econometric History* (Baltimore, 1964); Avner Greif, "Cliometrics After 40 Years," *American Economic Review* 87 (1997): 400-403; Douglass C. North, "Cliometrics–40 Years Later," *American Economic Review* 87 (1997): 412-414.

[7] François Furet, "Quantitative History," *Daedalus* 100 (1971): 151-167; Emmanuel Le Roy Ladurie, *The Territory of the Historian* (Chicago, 1979).

[8] Swierenga, "Computers and American History"; Jerome M. Clubb and Howard Allen, "Computers and Historical Studies," *Journal of American History* 54 (1967): 599-607; William G. Thomas, II, "Computing and the Historical Imagination," in Susan Schreibman, Ray Siemens, and John Unsworth (eds.), *A Companion to Digital Humanities* (Oxford, 2004), digitalhumanities.org/companion/.

the movement argued that quantitative methods would eventually take over history, turn it into a science, and rid the profession of ideological cant.[9]

Historians committed to traditional methods reacted in predictable fashion. American Historical Association President Carl Bridenbaugh railed against those who worshipped "at the shrine of that bitch-goddess, QUANTIFICATION."[10] Arthur Schlesinger, Jr., told the American Sociological Association that, although he did not deny the value of quantitative methods, nonetheless "almost all important questions are important precisely because they are *not* susceptible to quantitative answers."[11] The debate erupted into public view in 1974, with the publication of *Time on the Cross*, a computational attempt to reinterpret slavery.[12] It was not the first time that the "new" economic history had considered slavery, but Robert Fogel and Stanley Engerman promised a computational revolution not just in historical methods, but also in public attitudes towards slavery itself, and therefore in the contemporary politics of race.[13] While the book garnered coverage in *Time* and *Newsweek*, scholars took issue with cherry-picked and unrepresentative data, misapplication of formulas and theory, and an ideological agenda that belied its pretensions to scientific certainty.[14] In the aftermath, Fogel began talking of the limits rather than the promise of quantification, the messianic fervor dissipated, and adherents fell away into other fields.[15]

---

[9] Robert P. Swierenga (ed.), *Quantification in History: Theory and Research* (New York, 1970).

[10] Carl Bridenbaugh, "The Great Mutation," *American Historical Review* 68 (1963): 326.

[11] Arthur Schlesinger, Jr., "The Humanist Looks at Empirical Social Research," *American Sociological Review* 27 (1962): 770.

[12] Robert William Fogel and Stanley L. Engerman, *Time on the Cross: The Economics of American Negro Slavery* (Boston, 1974).

[13] Alfred H. Conrad and John R. Meyer, "The Economics of Slavery in the Ante Bellum South," *Journal of Political Economy* 66 (1958): 95-130.

[14] Herbert G. Gutman, "The World Two Cliometricians Made: A Review-Essay of F + E = T/C," *Journal of Negro History* 60 (1975): 53-57; C. Vann Woodward, "The Jolly Institution," *New York Review of Books* 21 (2 May 1974); Eric Foner, "Redefining the Past," *Labor History* 16 (1975): 127-138.

[15] Robert William Fogel, "The Limits of Quantitative Methods in History," *American Historical Review* 80 (1975): 329:350; Charlotte Erickson, "Quantitative History," *American Historical Review* 80 (1975): 351-365.

As quantitative methods became dominant in economics and political science, equations disappeared from mainstream historiography. The cliometricians had engendered greater reflexiveness about method, and for many this required a new focus on epistemology. As François Furet wrote at the time, the laborious creation of data series made the historian, still male, "aware that he has constructed his own facts," something that Furet called "a revolution in the historiographical consciousness."[16] Looking back a quarter of a century later, Joyce Appleby observed that while quantitative history had made it impossible to deny the structural inequities in American history, the statistics had not spoken for themselves. Historical analysis required revealing the power relations that produced the numbers, and that work would generally be qualitative in nature.[17]

While social and cultural history took off, cliometrics all but died with *Time on the Cross*. A zeal for information technology was not reborn in the historical discipline until the coming of digital humanities. Its current form is traceable to the mid-1990s, when the internet appeared to promise a solution to what was already seen as a crisis in the humanities. So the most important primers on the digital humanities declare that it not only revolutionizes the humanities, it "upends academic life as we know it."[18] For one of digital history's early trailblazers, Roy Rosenzweig, interactive CD-ROMs and hyperlinked projects were steps on the "road to Xanadu," because they promised to bring history to a wider audience.[19] For another advocate,

---

[16] Furet, "Quantitative History," 160.

[17] Joyce Appleby, "The Power of History," *American Historical Review* 103 (1998): 5-6.

[18] Matthew K. Gold, "Introduction: 'The Digital Humanities Moment'," in Matthew K. Gold (ed.), *Debates in the Digital Humanities* (Minneapolis, 2012), ix. See also Anne Burdick, Johanna Drucker, Peter Lunenfeld, Todd Presner, Jeffrey Schnapp, *Digital_Humanities* (Cambridge, MA, 2012), and the collection edited by one of history's few representatives in the digital humanities, Daniel J. Cohen: Daniel J. Cohen and Tom Scheinfeldt (eds.), *Hacking the Academy: New Approaches to Scholarship and Teaching from Digital Humanities* (Ann Arbor, 2013).

[19] Roy Rosenzweig, "The Road to Xanadu: Public and Private Pathways on the History Web," *Journal of American History* 88 (2001): 548-579. Rosenzweig was the key figure in earlier moves towards digital history. See Roy Rosenzweig, ""So, What's Next for Clio?" CD-ROM and Historians," *Journal of American History* 81 (1995): 1621-1640); Roy Rosenzweig and Michael O'Malley, "Brave New World or Blind Alley? American History on the World

digital history is "a revolution in the history profession that will change the way history is done at every level of scholarship and teaching and throughout the libraries and databases historians use in their everyday work."[20]

The field of digital humanities has given rise to countless web-based projects, internet forums, and "unconferences." But one of the central preoccupations for humanists in the field is whether, and how, all of the blogs, tools, and visualizations add up to a new kind of scholarship that can and should pass peer-review. Digital historians' proudest achievements have not been new discoveries or grand narratives – conspicuous by their absence – but easy-to-use interactive tools and free public platforms like Zotero. So does digital history have to meet the same standard as any other field of history, that is, to demonstrate that it has created new and important knowledge about the past? Or is its main role precisely to expose self-important academic pretensions and "shake things up," as Michael Frisch argues?

An even more fundamental debate concerns how we define and delimit the field. William G. Thomas suggests that it "is about the medium, not the method," so almost any history on the web is digital history. Many practitioners celebrate collaboration, and point out that not every member of a team needs to have the same skillset. But for Daniel Cohen and William Turkel, only programming historians can do truly advanced research in digital history. All this creates a first mover problem, as Kirsten Sword asks: "Is it wise and fair to launch graduate students into their own, largely unsupported, digital projects when the 'best' work appears in large scale,

---

Wide Web," *Journal of American History* 84 (1997): 132-155; Roy Rosenzweig, "Scarcity or Abundance? Preserving the Past in a Digital Era," *American Historical Review* 108 (2003): 735-762; Roy Rosenzweig, "Can History Be Open Source? Wikipedia and the Future of the Past," *Journal of American History* 93 (2006): 117-146; Daniel J. Cohen and Roy Rosenzweig, *Digital History: A Guide to Gathering, Preserving, and Presenting the Past on the Web* (Philadelphia, 2006); Roy Rosenzweig, *Clio Wired: The Future of the Past in the Digital Age* (New York, 2010).

[20] Orville Vernon Burton, "American Digital History," *Social Science Computer Review* 23 (2005): 206-220.

collaborative ventures, and when scholarly articles and monographs remain our common professional currency?"[21]

One thing seems clear. There is not yet any agreement on how to define "digital humanities" or the subfield of "digital history." Debate will likely only be settled when digital historians produce "field-defining" work, the kind of work that commands the respect of the rest of the academy. But discussions about digital history are mainly happening among digital historians, rather than historians more broadly. In literary studies, an *avant-garde* led by Franco Moretti and Matthew Jockers has been bold and successful enough to spark a backlash in the core of the discipline.[22] Yet historians generally evidence little more than mild curiosity about digital techniques, which they are as likely to read about in the arts section of their newspaper as in a scholarly journal.[23] By promising a revolution that has not yet come, digital history aims to challenge every scholar, but does not challenge anyone in particular. For outsiders, it is all too easy to see digital history as something to teach with, like world history, or even as a harmless form of public outreach.

This is unfortunate, and not just for the digital historians themselves. For literary scholars, the development by Google and HathiTrust of vast corpora of digitized books has made

---

[21] "Interchange: The Promise of Digital History," 488, 442-451, 461, 464.

[22] Franco Moretti, *Graphs, Maps, Trees: Abstract Models for Literary History* (London, 2007); Franco Moretti, *Distant Reading* (London, 2013); Matthew L. Jockers, *Macroanalysis: Digital Methods and Literary History* (Urbana, 2013); Stephen Ramsay, *Reading Machines: Toward An Algorithmic Criticism* (Urbana, 2011). Moretti and Jockers work with the Stanford Literary Lab, the publications of which can be found at litlab.stanford.edu/?page_id=255. For the backlash, see Stephen Marche, "Literature is Not Data: Against Digital Humanities," *Los Angeles Review of Books*, October 28, 2012, lareviewofbooks.org/essay/literature-is-not-data-against-digital-humanities; Scott Selisker and Holger S. Syme, "In Defense of Data: Responses to Stephen Marche's "Literature is Not Data"," *Los Angeles Review of Books*, November 5, 2012, lareviewofbooks.org/essay/in-defense-of-data-responses-to-stephen-marches-literature-is-not-data.

[23] See the "Humanities 2.0" series in the *New York Times*, especially Patricia Cohen, "Digital Keys for Unlocking the Humanities' Riches," *New York Times*, November 16, 2010, nytimes.com/2010/11/17/arts/17digital.html, and Patricia Cohen, "Digital Maps Are Giving Scholars the Historical Lay of the Land," *New York Times*, July 26, 2011, nytimes.com/2011/07/27/arts/geographic-information-systems-help-scholars-see-history.html.

computational analysis possible, but not imperative.[24] The advent of electronic records, on the other hand, is already bringing profound changes to the very nature of our archives, whether scholars of American foreign relations realize it or not. While the growth of digitized and born-digital primary source collections is usually seen as an unmitigated good, the effects are complex, and in some ways quite worrying.

The Archival Explosion

Historians of the contemporary national security state have long been coping with the problem of "big data." What counts as "big," after all, is always relative to what has come before. As long ago as 1961, the advisory committee to the State Department's *Foreign Relations* series warned of "the fantastic expansion of materials in the archives," which it called "a crisis of major proportions."[25] Twenty years later, Gerald K. Haines and J. Samuel Walker argued that one of the principal problems faced by historians was "an almost overwhelming task of sifting and winnowing an enormous amount of documentation."[26] The foreign policy bureaucracy vastly expanded over the course of the Cold War with the growth of the State Department and the Pentagon and the addition of new players like the National Security Council, USAID, and a host of intelligence agencies. The horizons of diplomacy widened too, as decolonization increased not just the number of states but also the burdens of global management, ranging from the eradication of disease to the control of world population growth. As McAllister puts it, our work

---

[24] Google Books, books.google.com; "Google Books History," google.com/googlebooks/about/history.html; HathiTrust Digital Library, hathitrust.org.

[25] "Public Report of the Advisory Committee," 1961, John F. Kennedy Presidential Library, Arthur M. Schlesinger, Jr., Personal Papers, Box WH-12.

[26] Gerald K. Haines and J. Samuel Walker, "Some Sources and Problems for Diplomatic Historians in the Next Two Decades," in Gerald K. Haines and J. Samuel Walker (eds.), *American Foreign Relations: A Historiographical Review* (Westport, 1981), 335-336.

over the past decades has embraced "more actors, more topics, more interaction, more documents, and more historiographical approaches," adding up "to a vastly larger universe of study."[27]

What has changed is that more and more of the sources that make up this universe are either being digitized or were "born digital," in the sense that they were originally created as electronic records. This is bringing about a qualitative transformation even in the study of the more distant past, since the individual researcher can now sift, search, and sort enormous collections with truly unprecedented storage and processing capabilities. The Federalist Papers have long been a test bed for statisticians who seek to perfect techniques in authorship attribution of anonymous documents. Now there is the prospect of applying this and other statistical techniques to much larger corpora, such as the fifteen thousand letters Benjamin Franklin wrote or received over the course of his life, which the Mapping the Republic of Letters project has begun to quantify and visualize to better understand his cosmopolitan connections even before he became one of America's first diplomats.[28] For later periods, over 450 volumes of the *Foreign Relations of the United States* (FRUS) series dating to 1861 are available digitally from the State Department Office of the Historian and the University of Wisconsin's Digital Collections, and it would be a relatively simple matter to determine, among other things, changes in the relative frequency of references to this or that country or individual across the whole 150-year corpus.[29]

These kinds of analysis *might* be simple, except that there are not yet any web applications for the historian to identify anonymous authors, extract and map locations from documents, or

---

[27] McAllister, "The Documentary Big Bang," 16.

[28] Mapping the Republic of Letters, "Visualizing Benjamin Franklin's Correspondence Network," republicofletters.stanford.edu/casestudies/franklin.html.

[29] *Foreign Relations of the United States*, State Department Office of the Historian, history.state.gov/historicaldocuments; *Foreign Relations of the United States*, University of Wisconsin Digital Collections, uwdc.library.wisc.edu/collections/FRUS.

conduct frequency analysis even for document collections that have already been digitized, like *FRUS*. The only tool that is typically offered users of online archives is the search engine. And when it comes to the electronic reading rooms maintained by every federal department and agency to store documents released under the Freedom of Information Act (FOIA), these search engines can be quite primitive or even non-functional.[30] Even so, the State Department site alone offers access to 85,000 records, which someone with modest coding skills can "scrape" (or copy) to create their own database.[31] More challenging to the would-be digital historian are the Remote Archives Capture (RAC) terminals, only accessible at presidential libraries. In an attempt better to manage declassification, the CIA and National Archives digitized millions of pages of documents, and those digital copies are now starting to be released at libraries from Truman onwards.[32] The CIA has declassified some 11 million pages of its own records, but only makes them available at Archives II in College Park, through the CREST system.

These are public sources, and in theory anyone could print out and re-digitize the RAC and CREST materials. Much of our public record is already being scanned and sold for profit, as Roy Rosenzweig pointed out many years ago.[33] ProQuest maintains the Digital National Security Archive, home to over 700,000 pages of FOIAed documents.[34] It also owns the History Vault, an agglomeration of documents that includes the National Security Files of the Kennedy, Johnson, and Nixon White Houses, plus the archives of all the major American newspapers.[35] Gale/Cengage maintains the Declassified Documents Reference System (DDRS), composed of around 500,000 pages of documents, mainly released through the Mandatory Declassification

---

[30] See, for instance, Central Intelligence Agency, "Freedom of Information Act Reading Room," foia.cia.gov.
[31] Department of State, "Freedom of Information Act," foia.state.gov.
[32] National Archives and Records Administration, "The Remote Archives Capture Program (RAC)," archives.gov/presidential-libraries/declassification/rac.html.
[33] Rosenzweig, "The Road to Xanadu," esp. 564-577.
[34] Digital National Security Archive, nsarchive.chadwyck.com/home.do.
[35] ProQuest History Vault, hv.conquestsystems.com/historyvault/hv.jsp?pageid=home.

Review (MDR) requests that scholars have filed at Presidential libraries over the last forty years. DDRS can therefore tell us a great deal not just about history, but also historiographical fashion and declassification policy.[36] But until recently, the only way to explore it and all the aforementioned databases has been by intuiting what terms will yield interesting results from the omnipresent search engine.

Having access to more and more digitized documents has already made it cheaper and easier to conduct primary source research compared to traditional archival expeditions. But the advent of "born digital" electronic record collections will likely bring much more profound changes to the nature of our research, beginning with reducing our dependence on scanning and Optical Character Recognition (OCR). OCR, after all, almost always produces some garbled text depending on image quality and the quirks of the software – one reason why search engines do not always produce even documents that contain the specified search terms. And scanning does not, by itself, yield metadata, or "data about data," like the author, recipient, date, and subject of a text. Names and locations embedded in clean text can be extracted through what data scientists call "named-entity recognition." It is an error-prone process, since computers cannot tell the difference between, say, Paris Hilton the celebrity and the seat of French government.

Electronic records, on the other hand, usually come complete with native metadata, which allow for many more – and more rigorous – forms of analysis. Consider, for instance, the State Department Central Foreign Policy File (CFPF), which has been one of the core collections for the study of American foreign relations since 1907. In the late 1960s, the State Department began to experiment with automatically sorting airgrams and began to convert all cables to machine-readable microfilm in the middle of 1973. One reason they did it was precisely to

[36] Declassified Documents Reference System, galenet.galegroup.com/servlet/DDRS.

generate data about diplomacy, such as through the "Traffic Analysis by Geography and Subject" (TAGS) system. These TAGS are one of 68 different fields of metadata that are now available online together with full-text cables through the National Archives and Records Administration's (NARA) "Access to Archival Databases" system (AAD).[37] "P-reel," or paper documents that were sent by diplomatic pouch, are currently only obtainable at College Park, but the metadata for each is also available at the AAD site. Along with "subjects," "concepts," and other information, the metadata provides a history of how each document was declassified. Each field adds another layer for potential analysis, and because humans filled each in at the time of record creation and archival preservation, their inconsistencies are revealing and interpretable. There are also hundreds of thousands of withdrawal cards, albeit with more limited metadata. But knowing the sender, recipient, date, and subject makes it possible for the first time to conduct systematic analysis of the overall agenda and volume of American diplomacy. With each new installment, the CFPF will become ever more central in the study of diplomatic history since 1973. So too will the Internet Archive, which has been copying documents from the .gov domain since 1995. The collection now totals a staggering 37 million unique PDFs, which the Internet Archive has begun to make available to research teams together with a high-performance computing cluster.[38]

In assessing what we have and what is yet to come in digital form, it is important to realize what we have lost. The CFPF at first glance appears overwhelming both in its size and its seeming completeness. But over one hundred thousand cables were corrupted in the transition to electronic records. For certain periods, most or all of the documentary record has simply been lost. This includes most of the cables from the first half of December 1975, for instance, and 92%

[37] National Archives and Records Administration, Record Group 59, Central Foreign Policy Files, 1973-1977, Access to Archival Databases (AAD), aad.archives.gov/aad/series-description.jsp?s=4073&cat=WR43&bc=,sl.
[38] The Wayback Machine offers access to a subset of this data, but it is already indispensible. See archive.org/web/.

of the telegrams from June 1976. Gone are records pertinent to the Indonesian invasion of East Timor, and the American response to the Soweto Uprising. Moreover, an increasing proportion of what survived intact is still unreleased. Whereas for the 1973 cables, 13% were withheld, for 1976, it was 24%. Withholding often occurs because certain collections are more likely to have national security or personally sensitive information, and NARA has no easy way to prioritize documents that require closer scrutiny. All of these records were simply printed out and reviewed page-by-page. And the years in which most of these cables were reviewed coincided with a dramatic decline in appropriations for declassification, from $232 million in 2001 to $48 million in 2004. Spending on declassification has not recovered, such that the inflation-adjusted budget is just 15% of what it was in the late 1990s.[39]

Consequently, archivists have felt compelled to delete millions of other documents – the exact number is impossible to determine – because they lacked the staff to screen all of them and prioritized those that appeared to have more enduring historical significance. Materials now lost forever include whole classes of cables concerning government-sponsored research, cultural diplomacy, passports, and visas.[40] Moreover, what remains was scanned or inputted into the State Archiving System in the order in which it was submitted to records managers. Documents created in 1975 might not have been put away until 1980, and are therefore currently inaccessible without lengthy FOIA delays. Using a keyword search to identify a P-reel document that is available yields, after consultation with a container list, a whole box of random documents that happened to be scanned at the same time. This loss of archival integrity makes it impossible even to produce a thematic or institutional finding aid, a problem that will become all the more

[39] Information Security Oversight Office (ISOO), "Annual Report to the President" (2012), archives.gov/isoo/reports/2012-annual-cost-report.pdf, 26.

[40] David Langbart, William Fischer, and Lisa Roberson, "Appraisal of records covered by N1-59-07-3-P," June 4, 2007.

acute when archivists with deep knowledge of these collections retire and their institutional knowledge is lost with them. It is already a disaster for the historian, virtually eliminating any chance of making a serendipitous discovery in neighboring files, or gaining any greater understanding of the context in which these documents were produced.

The rapid expansion of the virtual archive of American foreign relations can thus distract us from the pitfalls and dark corners awaiting the unwary researcher. So far, historians who wish to explore this archive have only been able to use a search engine. It is not unlike a flashlight, which we shine into the archive if only because we cannot think of what else to use. But computer science is beginning to produce a whole array of new techniques to explore virtual archives, the equivalent of infrared lenses and autonomous drones. It would be foolhardy for historians not at least to try to use them before we stumble much further into the darkness, and before millions more historical records are lost forever.

Computational Methods 3.0

So how should we grapple with these digital repositories, once we realize that they are disjunctive and disorganized, and that a large (but unknown) part of the original documentary record is unavailable because of corrupt files, deletions, and withholding? Imagine beginning a book on Henry Kissinger's stint as Secretary of State, and approaching the source base in the traditional manner taught to graduate students for decades: read everything, and then read around. It would take well over a lifetime to read the cables for those few years, and, nearly three years to read just the ones that were sent from Moscow.[41] One would need to add transcripts of Kissinger's

[41] Calculations based on a reader working an eight-hour day with no days off.

telephone conversations, all 15,000 of them, plus the records of his meetings, the State Department papers that were not stored electronically, and Defense, Treasury, and intelligence records. Then combine all of that with the personal and non-governmental archives that have been mined so profitably over the past decades by historians of American foreign relations. Finally, research in the archives of other countries and international organizations in order to correct for the intrinsic bias in a single government's records is vital to truly international history. Just thinking about the scale of what this will eventually entail makes it immediately clear that new approaches are needed.

What if we made a virtue of necessity, and approached the archive in entirely new ways? Normally, when we go to a large physical archive, we enter with some idea of the key topics, consult the finding aids, learn the scope and content of the collections, and start ordering everything that seems relevant. We then look at the documents one by one, and try to glean insights. Sometimes they give us leads that we follow into other files, until we begin to think that we have some sense of how everything is connected. But we never have a very clear picture of the larger whole, since we never see more than a fraction of the full collection. This is the virtue and vice of "close reading."

Now it is possible to "read" an *entire* archive and analyze *every* available document and withdrawal card at the same time. We can use this power to perform a "first cut," determine the thematic topics that are statistically most prevalent, reveal what kinds of documents are particularly likely to be withheld or redacted, and rank all available documents according to their relevance to our research interests. All this can be done based on the features within the documents themselves, without presuming that we already know what topics are important or sensitive, or what terms might yield documents relevant to our research. We can then alternate from this kind of "distant reading" to close reading of the usual kind, only now with more

confidence that – if we cannot *actually* see everything – we at least do not have the tunnel vision that results from only reading the results of search queries.[42] So the old and the new are not mutually exclusive, indeed quite the opposite.

For collections like the CFPF that have irretrievably lost their archival integrity, computational methods may offer the only hope of creating order from the chaos and producing anything like a proper finding aid. Lawyers have already discovered this when faced with huge corpora of documents produced through legal discovery. There is now a multi-billion dollar industry devoted to "e-discovery," albeit one that closely protects its intellectual property. Journalists, who write the first draft of history, were the first to create free public platforms based on machine learning and natural language processing (NLP). These systems automatically cluster documents and organize them in virtual files and folders in ways that resemble textual archives, and thus help the individual researcher determine where and how to begin reading.[43]

Historians can make excellent use of these systems, but we should also be helping computer scientists to develop new ones. This requires collaborative research, as the digital humanists argue, but our experience suggests that historians can contribute even if they lack coding skills. In fact, advanced work using NLP and machine learning requires much more knowledge of mathematics and computer science than all but a handful of historians are likely to possess. As part of a strong, multidisciplinary team, historians play a critical role in defining worthwhile questions to investigate, advising on the tradeoffs of various research protocols, and determining whether the results are valid and interesting or are merely an artifact of a flawed

---

[42] This distinction between "close" and "distant" reading is borrowed from Franco Moretti. See Franco Moretti, "Conjectures on World Literature," *New Left Review* 1 (January/February 2000): 54-68; Moretti, *Distant Reading* (London, 2013). Matthew Jockers prefers to think in terms of "macro" and "microanalysis," arguing that algorithms are not actually *reading* documents (a human activity if ever there was one), but analyzing them. See Jockers, *Macroanalysis*, 22-31.

[43] See, for instance, "DocumentCloud," documentcloud.org/home; "The Overview Project," overview.ap.org.

research design. If historians do not start working together with data scientists to create reliable tools for our research, we will not have any say – nor perhaps any understanding – of the methods, compromises, and tradeoffs in putting them together.

Historians and data scientists will also have to work with the professionals who will largely determine what kinds of research we will be able to do in the future. This begins with records managers, who decide whether digital collections will be more (or less) "future-proof." Archivists have been talking about these challenges for much longer than historians, and we should enter their discussions with all due humility.[44] We also have a role to play in helping them determine what records have permanent historical significance. Ironically, what seem like mundane documents on communications procedures and records management may be the most important of all, since they are an indispensible means to reconstruct how a collection came together. Destroying them is the equivalent of throwing out the owner's manual. If instead archivists keep in mind the potential for data-mining, computer scientists can more easily develop tools to help them process text collections. In the meantime, deleting electronic records has to be a last resort.

In addition to records managers and archivists, librarians play a critical role in helping historians and data scientists negotiate access to collections owned by private vendors. Librarians decide what digital collections to acquire, and as customers they are in the best position to communicate with vendors about the needs of researchers. The greatest need is usually to have the raw data, ideally through an Application Programming Interface (API). Vendors are usually open to this idea since they understand that new analytical tools can greatly enhance the value of their collections.

---

[44] Joshua Sternfeld, "Archival Theory and Digital Historiography: Selection, Search, and Metadata as Archival Processes for Assessing Historical Contextualization," *The American Archivist* 74 (2011): 544-575; Alexandra Chassanoff, "Historians and the Use of Primary Sources in the Digital Age," *The American Archivist* 76 (2013): 458-480.

If historians can join forces with data scientists, archivists, and librarians, there is the prospect of creating a vibrant new field of research. Text processing has long depended on a relatively small number of datasets, that are large, machine-readable, public, and rich with metadata. The State Department cables meet all these criteria, are more voluminous, and bear on matters of great and enduring historical significance. As more and more government communications are released, slowly but still decades sooner than most private or corporate e-mails, there will be many more such datasets. And because there is a sizable scholarly community devoted to their study both now and stretching into the future (unlike, for instance, the Enron e-mails that currently provide one dataset), it will be easier to develop and pursue a joint research agenda likely to result in original and important discoveries.[45]

It is still very early days for digital history. We barely know what kinds of questions could be asked of our documents in future, let alone how to answer them. There are, however, several fields of research that should prove particularly useful for historians of American foreign relations. All rely on some form of "distant reading" of thousands, even millions of documents. Each technique has the potential to combine the idea of a cold start in a new archive with the sense of serendipitous discovery familiar to all historians, by finding patterns across time and anomalies that depart from those patterns.

[45] Jessica Leber, "The Immortal Life of the Enron E-mails," *MIT Technology Review*, 2 July 2013, technologyreview.com/news/515801/the-immortal-life-of-the-enron-e-mails/.

Fields of Research

Counting

Counting is the simplest kind of computation, but it can help to answer some fundamental questions. In the absence of tools to turn words into data, historians have resorted to using search engines to tabulate the number of references to this or that historical term.[46] Aside from the problem of corrupted text due to imperfect OCR, this method does not take into account how a corpus changes over time. If newspapers grow in size, for instance, the frequency of most terms will also increase. The Google Ngram Viewer would appear to solve this problem, since it displays word frequency relative to other words published in a given year in the Google Books corpus.[47] But it does not allow users to see what part of the corpus is being quantified and graphed. This is a fatal flaw for historians who want to understand the nature of their sources before building arguments on top of them.[48]

---

[46] See, for instance, James Belich, *Replenishing the Earth: The Settler Revolution and the Rise of the Anglo-World, 1783–1939* (New York, 2009), 151-2.

[47] "Google Ngram Viewer," https://books.google.com/ngrams and https://books.google.com/ngrams/info; Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden, "Quantitative Analysis of Culture Using Millions of Digitized Books," *Science* 331 (14 January 2011): 176-182, sciencemag.org/content/331/6014/176.full.html.

[48] Some historians have used Ngram graphs with no apparent awareness of this problem. See, for example, Daniel Lord Smail and Andrew Shryock, "History and the 'Pre'," *American Historical Review* 118 (June 2013): 710-712.

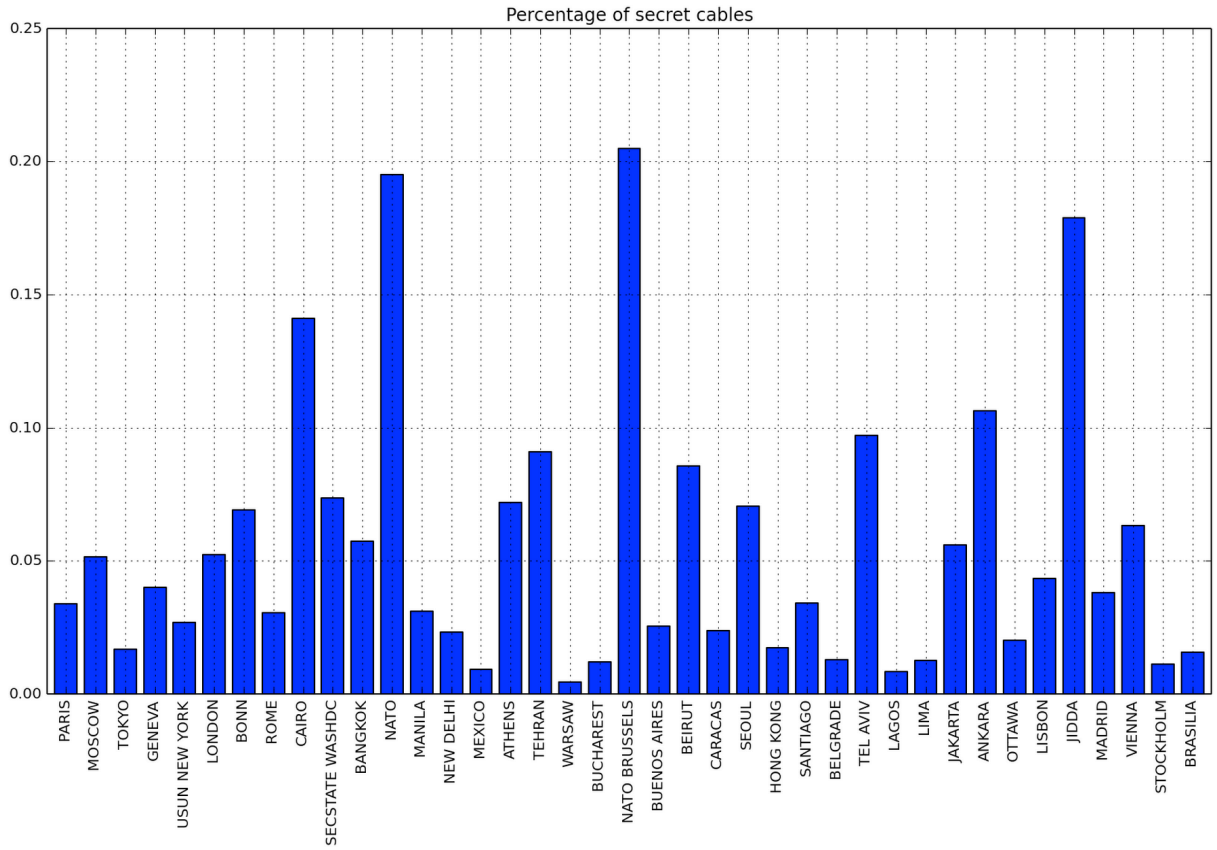

Figure 1. *The percentage of cables, sent by American embassies and offices around the world, originally classified secret, 1973-1976. The embassies, on the x-axis, are arranged in order of total volume, from left to right. Source: cleaned version of State Department Cables, Access to Archival Databases, National Archives and Records Administration. We thank Daniel Krasner for the production of this graph.*

These problems can be overcome. When they are, international historians will want to know how the volume of diplomatic activity changed over time, and where it tended to focus. They will likely debate whether the number of telephone conversations or cables transmitted provides a way to measure interest rather than just activity. With metadata, there will be new layers of analysis, and new questions. For instance, which embassies deal with the most information that is classified secret, and why (*Figure 1*)? Does that reflect the sensitivity of those communications, or is it an indicator of which embassies tended to overclassify?

Once it becomes easier to turn words into data, new debates will begin about how to interpret that data. The relative frequency with which diplomats use the term "human rights" in

confidential communications, for instance, may or may not indicate whether human rights were a priority in foreign policy, much less whether that was predictive of how they would treat a friendly dictator. It does show whether, when, and to what extent, they believed human rights were worthy of discussion, something human rights scholars fiercely debate.[49] Now that this data is becoming more readily available, it will become hard to ignore, and standards of evidence will begin to shift.

Traffic Analysis

Counting is useful, but it can only go so far. As William Sewell has written, we have always needed better ways to understand "lumpy, uneven" time, to describe how the pace of history appears to speed up or slow down.[50] Measuring the rate of communications is a promising way of measuring the speed of diplomatic activity. But when we plot graphs to reveal these "lumps," do we segment them by year, month, week, day, or hour? No one choice is more objective than another, but too small an increment will reveal many blips, rather than real bursts of activity. Conversely, if the increment is too long, these bursts will become invisible because they average out over time.

---

[49] See, for instance, the periodization debate present in Samuel Moyn, *The Last Utopia: Human Rights in History* (Cambridge, MA, 2010); Sarah B. Synder, *Human Rights Activism and the End of the Cold War: A Transnational History of the Helsinki Network* (New York, 2011); and Elizabeth Borgwardt, *A New Deal for the World: America's Vision for Human Rights* (Cambridge, MA, 2005).

[50] William H. Sewell, Jr., *Logics of History: Social Theory and Social Transformation* (Chicago, 2005), 9, passim.

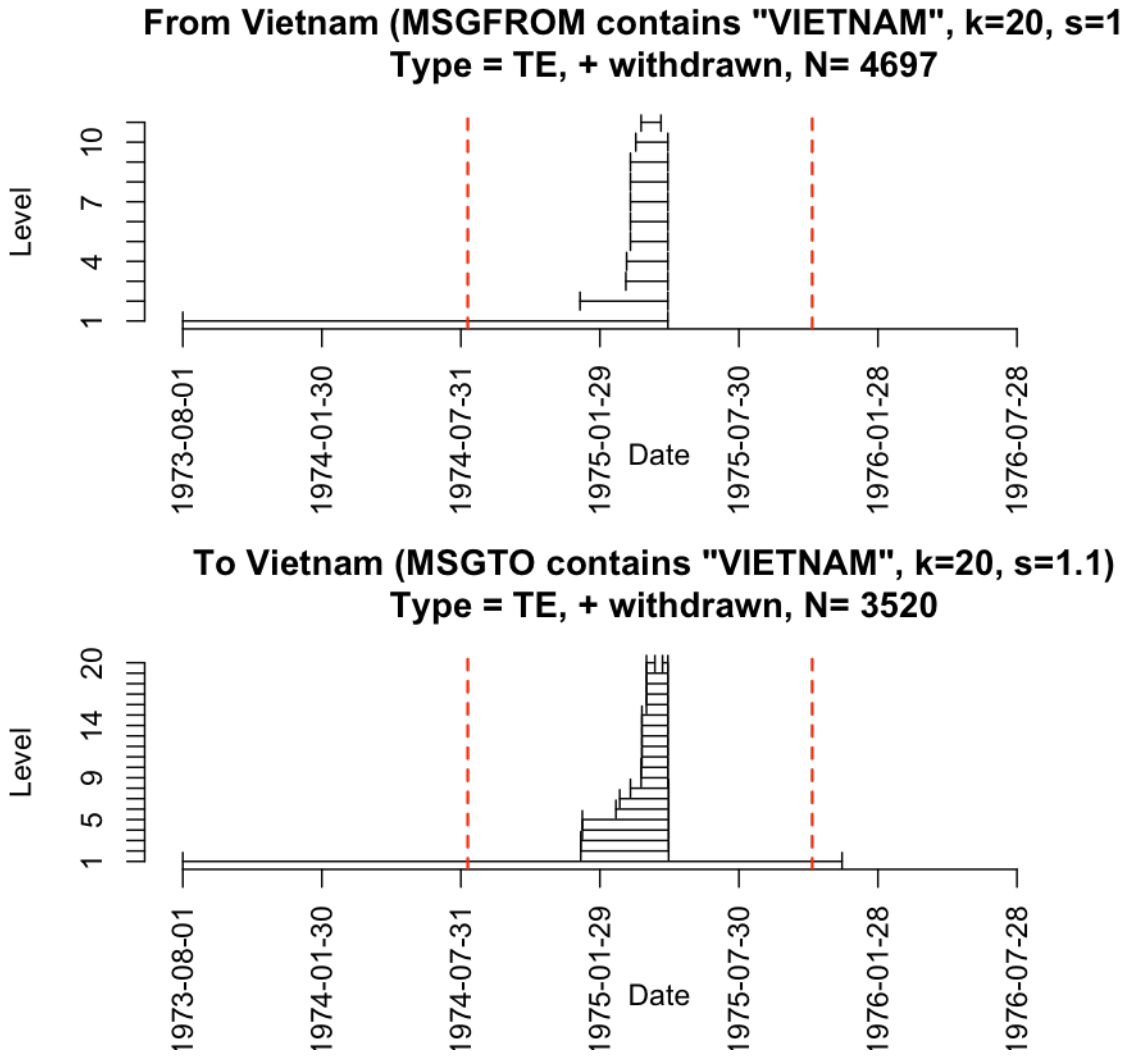

Figure 2. *"Burstiness" graphs for the fall of Saigon, 1973-1976. These two graphs are produced by statistical work that charts traffic analysis. In both cases, a higher "burst" level represents heightened speed of communications, in and out of Vietnam. The upper limit of a "burst" is represented by "k," while "s" is a measure of graphical smoothness. N is calculated by adding the total number of cables sent "to" or "from" Vietnam over the period to the declassified metadata of cables that remain withheld. Dotted lines give dates for the resignation of President Richard Nixon, and the firing of Henry Kissinger as Secretary of State. Increased "burstiness" tracks important military events as North Vietnamese troops moved south, as well as resulting refugee crises. The precipitous stop occurs on April 30, 1975, with the evacuation of the American embassy in Saigon. Source: cleaned AAD. We thank Shawn Simpson for the statistical work that produced these graphs.*

One approach that helps to resolve this conundrum is precisely to focus on this quality of "burstiness" in streams of communications.[51] The idea is to segment time into levels of activity as measured by the observed time between cables. We demarcate the beginning and end of episodes

[51] For the statistical basis of "burstiness", see Jon Kleinberg, "Bursty and Hierarchical Structure in Streams," *Data Mining and Knowledge Discovery* 7 (2003): 373-397.

by the escalation or de-escalation of that activity. With this model, we can then identify "bursty" time spans across the entirety of a collection, or between two embassies, and see how they relate to events. This allows us to precisely map the duration of a crisis, and to compare what we find to public assertions of what was going on. For instance, plotting the "burstiness" of communications between Vietnam and Washington between 1974 and 1976 shows how "bursts" track the dates of significant military events and how the final crisis had started long before the Ford administration admitted as much. Only a deeper dive into the documents shows that, in the end, embassy communications were mainly about refugees (*Figure 2*).

This is an experimental approach, but it already shows that "bursts" of activity can be a function of how the communications flow changes when a Secretary of State moves through the network. When he or she goes to a foreign capital, some communications that would ordinarily have been internal to Foggy Bottom become external, leading to heightened cable traffic. It will take more experimentation before this method can reveal unstudied events. We might, for instance, be able to determine whether there are particular types of cables or language within cables that are predictive of bursts of activity, or how particular kinds of metadata interact with the text. We know that intelligence agencies model and measure "chatter" to predict terrorist attacks. We ought to be able to develop our own models for diplomatic communications to predict and model other kinds of events and non-events.

Topic Modeling

Searches can be a powerful way to interact with documents, but they are time-consuming, frustrating, and imprecise. We might instead want to find documents related to a topic without presuming that we already know the main topics in a collection. In probabilistic topic modeling,

computer scientists attempt to find the hidden thematic structure in large sets of documents. There are various kinds of models, the most common of which is Latent Dirichlet Allocation (LDA).[52] Topic modeling finds words that are likely to relate to each other statistically, and turns them into strings of probabilities that make up a theme. To model a corpus, it assumes that a certain number of topics *generated* the documents, which reverses our usual intuition, and that each of the documents in the corpus therefore represents those topics to some extent. It is an "unsupervised" technique, which means that, once the parameters are specified, the algorithm is autonomous, automatically generating the combination of topics that provide the best solution.

Computers cannot actually recognize the difference between a meaningful topic and one that merely reflects strings of words that are often used, like conjunctions and modifiers. These and other common words, or "stopwords," must be excised with discretion and a broad historical imagination for what the underlying themes of the cable collection might be, before the model is run again. Once that is done, we have strings of words that are statistically representative of themes in the text. When we recognize the interrelationship of terms our inclination is to label them, something humans can do much better than machines.[53] This "science" is probabilistic and interpretive all the way down.

---

[52] For summaries of topic modeling, especially Latent Dirichlet Allocation (LDA), see Jockers, *Macroanalysis*, 118-153; David M. Blei, "Topic Modeling and Digital Humanities"; David M. Blei, "Probabilistic Topic Models," *Communications of the ACM* 55 (April 2012): 77-84. For the original statistical explanations, see David M. Blei, Andrew Y. Ng, Michael I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research* 3 (2003): 993-1022; Thomas L. Griffiths and Mark Steyvers, "Finding scientific topics," *Proceedings of the National Academy of Science* 101 (April 6, 2004): 5228-5235. For an early historical use of topic modeling, see David J. Newman and Sharon Block, "Probabilistic Topic Decomposition of an Eighteenth-Century American Newspaper," *Journal of the American Society for Information Science and Technology* 57 (2006): 753-767.

[53] Jonathan Chang, Jordan Boyd-Graber, Sean Gerrish, Chong Wang, David M. Blei, "Reading Tea Leaves: How Humans Interpret Topic Models," *Neural Information Processing Systems* (2009), cs.princeton.edu/~blei/papers/ChangBoyd-GraberWangGerrishBlei2009a.pdf. For more on the strings of words that topic modeling generates, and the coherence of the strings, see Jey Han Lau, Karl Grieser, David Newman, and Timothy Baldwin, "Automatic Labelling of Topic Models," *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics* (June 19-24, 2011): 1536-1545; Jey Han Lau, David Newman, Sarvnaz Karimi, and Timothy Baldwin, "Best Topic Word Selection for Topic Labelling," *Proceedings of the 23rd International Conference on Computational Linguistics Posters* (August 23-27, 2010): 605-613; David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin,

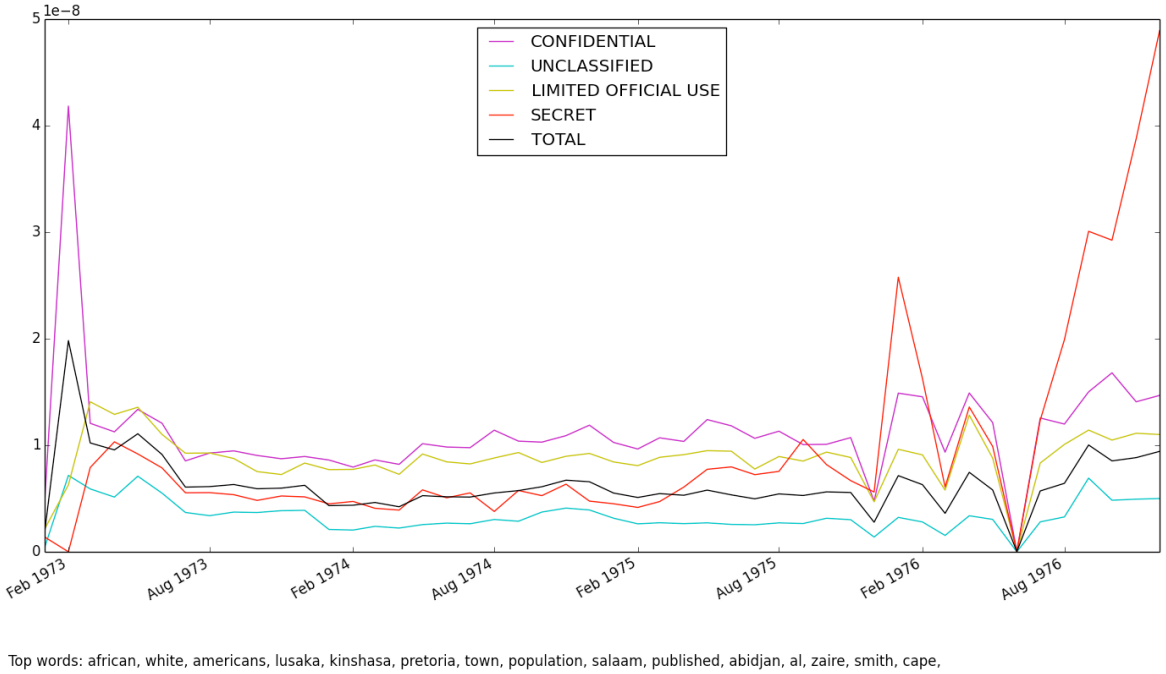

Figure 3. *Secrecy levels of cables in an "Africa policy" topic. Topic modeling produces strings of words that are statistically related to one another in a corpus: in this case, words pertaining to policy in Africa. On the graph, note the rise in volume and secrecy as Henry Kissinger becomes personally involved in the end of white rule in Rhodesia. Source: cleaned AAD. We thank Ian Langmore, Daniel Krasner, and Maja Rudolph for the production of this graph.*

Topic modeling is a fairly new and quickly developing field in natural language processing, and the computer scientists doing such research are keen to work with historians to develop better ways of interacting with text. Models have been developed that take into account historical change, networks of documents, syntax, "burstiness," and even, most interestingly, the influence of one document on another over time.[54] When used appropriately, topic modeling

---

"Automatic Evaluation of Topic Coherence," *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL* (June 2010): 100-108.

[54] David M. Blei and John D. Lafferty, "Dynamic Topic Models," *Proceedings of the 23rd International Conference on Machine Learning* (2006): 113-120; Jonathan Chang and David M. Blei, "Relational Topic Models for Document Networks," *Proceedings of the 12th International Conference on Artificial Intelligence and Statistics* (2009): 81-88; Jordan Boyd-Graber and David M. Blei, "Syntactic Topic Models," *Neural Information Processing Systems* (2008), cs.princeton.edu/~blei/papers/Boyd-GraberBlei2009.pdf; Thomas L. Griffiths, Mark Steyvers, David M. Blei, and Joshua B. Tenenbaum, "Integrating Topics and Syntax," *Neural Information Processing Systems* 17 (2005), psiexp.ss.uci.edu/research/papers/composite.pdf; Gabriel Doyle and Charles Elkan, "Accounting for Burstiness in Topic Models," *Proceedings of the 26th International Conference on Machine Learning* (2009), cseweb.ucsd.edu/~elkan/TopicBurstiness.pdf; Sean M. Gerrish and David M. Blei, "A Language-based Approach to Measuring Scholarly Impact," *Proceedings of the 26th International Conference on Machine Learning* (2010), cs.princeton.edu/~blei/papers/GerrishBlei2010.pdf.

finds the hidden intellectual structures in our documents. Imagine that we wanted to write a book about how the term "national security" has changed meaning over time. Simple searching for the term in various databases will be helpful, but frustrating. Topic modeling has the potential not only to identify documents in the archive that are thematically related to national security, whether or not the term itself is actually used, but to show how the words related to the concept have changed over time. Other applications might include showing how the language of public diplomacy differs from that of private diplomacy, and how certain topics tend to be more highly classified, or take longer to declassify.

Going "Off-Topic"

This kind of "distant reading" can also be used to find anomalies. Anomalies are one way that we might restore the classic serendipitous finding in paper archives for the digital era. So, if we take the CFPF from the Kissinger years and apply topic modeling to all the telegrams, we find that individual embassies have specific signatures (*Figure 4*). The Moscow embassy talks a lot about the USSR, and very rarely about anything else. The London embassy, however, serves as a clearinghouse for multiple issues (Europe, the Commonwealth, trade policy, and so on), so its signature is much more diverse. If we know that particular embassies have particular signatures when they are "on" topic, we can see what happens when they go "off" topic.[55]

[55] Ian Langmore was the one who came up with this idea as part of the Declassification Engine project.

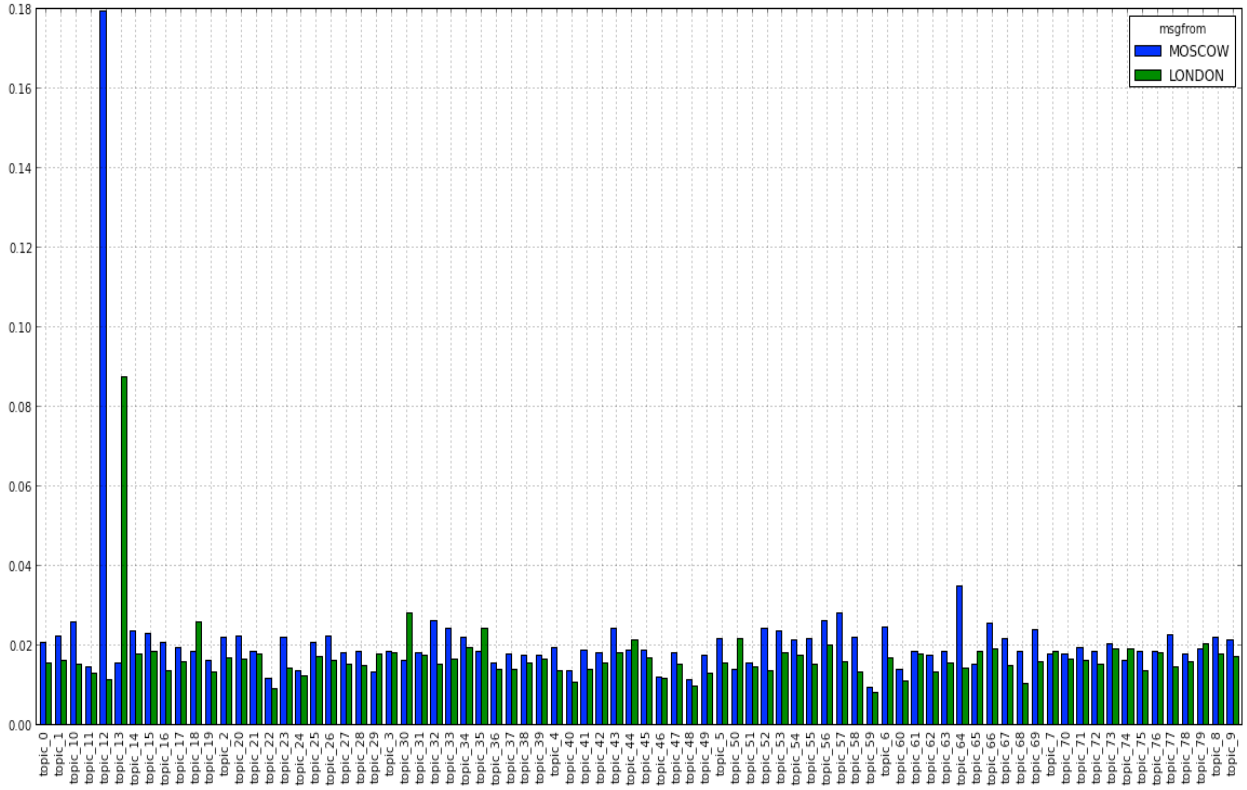

Figure 4. *Topic modeling the State Department Cables as a corpus reveals that individual embassies have specific topic signatures, so that eighteen percent of Moscow cables concern topic 12, "Soviet, Moscow, October…" and nine percent of London cables concern topic 13, "London, Bonn, Rome…" These signatures can be used to predict where a given cables "should" have originated from. If this does not match the actual sender, a cable is said to be "off" topic. We thank Ian Langmore for the production of this graph.*

In a first run of this method, for instance, we found an unusual backchannel communication between the Soviet and American political counselors in Paris, discussing the Arab-Israeli war of 1973.[56] We also found examples of cables sent from unexpected embassies, like a report on a Kissinger meeting with Willy Brandt sent once the Secretary had landed in Moscow, and a piece of Kremlinology pertaining to Leonid Brezhnev's power in the Politburo, relayed through the Finnish representative to the talks, held in Geneva, that led to the Helsinki

[56] Paris to SecState, "Middle East Situation: Soviet Embassy Approach," October 7, 1973, AAD, 1973PARIS26220.

Accords.[57] This method could apply not just to space, but also to time. If we have enough metadata, we might even be able to estimate when a document was written, thereby helping to find documents that represent certain themes in foreign policy before or after they become especially common. By predicting where a document came from, or who wrote it, we can find the unpredictable.

Authorship Attribution

If topic modeling is a very new field, authorship attribution is very old. It dates back to the medieval scholastics, who tried to find ancient authorities for their documents. More recently, common statistical problems have involved verifying Shakespeare plays and the writers of individual Federalist papers. In their famous 1963 article on the latter, Frederick Mosteller and David L. Wallace used the frequency of words like "to," "from," and "upon" to assign contested articles to Alexander Hamilton or James Madison.[58] Since then, computers have made authorship attribution more rapid, powerful, and accurate.[59]

This method tries to solve various problems. In the easiest cases, there might be an anonymous document that could have been written by a small pool of authors, like a presidential speech or an embassy telegram sent under an ambassador's name. Or we might have a much

---

[57] Moscow to State, "SecState's Meetings with Brandt and Scheel," March 25, 1974, AAD, document number 1974MOSCOW04262; Geneva to State and Bonn, "Soviet Leadership Matters Bonn for Secretary's Party," February 15, 1975, AAD, 1975GENEVA01022.

[58] Frederick Mosteller and David L. Wallace, "Inference in an Authorship Problem," *Journal of American Statistical Association* 58 (1963): 275-309.

[59] For summaries of the various methods from literary scholars, see Jack Grieve, "Quantitative Authorship: And Evaluation of Techniques," *Literary and Linguistic Computing* 22 (2007): 251-270; Matthew L. Jockers and Daniela M. Witten, "A comparative study of machine learning methods for authorship attribution," *Literary and Linguistic Computing* 25 (2010): 215-223; For technical summaries, see Patrick Juola, "Authorship Attribution," *Foundations and Trends in Information Retrieval* 1 (2006): 233-334; Moshe Koppel, Jonathan Schler, and Shlomo Argamon, "Computational Methods in Authorship Attribution," *Journal of the American Society for Information Science and Technology* 60 (2009): 9-26.

larger number of candidates, such as an anonymous memorandum that might have been written by one of thousands of different foreign service officers. Or, especially in the case of documents like National Security Council or Policy Planning Staff memoranda, we might want to determine who wrote specific sections of a given document. In each case, we would need authenticated examples of the writing of all of the candidates. But even if we have no idea who wrote a given text, authorship attribution techniques are reasonably accurate in determining their age and gender. All of these questions are ongoing fields of research in statistics that can help to solve longstanding problems with our sources.

Network Analysis

Another well-developed field in statistics and computer science is social network analysis. The idea here is to analyze large collections of texts through the social network that is evoked in them, in ways that builds on economic and sociological theory.[60] A network is a collection of people ("nodes") connected by links ("edges"). The most pertinent examples for historians of American foreign relations are the bureaucrats in government departments, and the networks of informants that embassies use to gain intelligence about the country on which they report. Network analysis provides us with ways to view structures like this across a corpus, and to see how the actions of one node have consequences for the rest of the system. Once we have extracted the network in historically representative ways, we can then begin to model it in a way that helps us to read the documents anew.

[60] For a good summary, see David Easley and Jon Kleinberg, *Networks, Crowds, and Markets: Reasoning about a Highly Connected World* (New York, 2010), cs.cornell.edu/home/kleinber/networks-book/.

Franco Moretti has already demonstrated the power of network analysis in analyzing individual plays, using network models to question traditional concepts in the study of literature. What happens, he asks, when you take Hamlet or Claudius out of *Hamlet*?[61] How might literary scholars define the "centrality" of a character?[62] Once historians of American foreign relations have large corpora of documents, we can imagine offering new answers to classic questions of diplomatic history and international relations theory. How, for instance, does power shift across the bureaucracy, and what are the policy implications? We might study how networks developed in ways that would not be expected by the power structure of the State Department or the federal government, charting the informal networks that drove policies. We could see how the network structures of the State Department and the National Security Council promoted different policies, and why one triumphed over the other. We might look at how specific bureaucratic networks formed in relation to specific geographic or issue questions. We could investigate how networks of local informants to a diplomatic mission changed before and after a crisis, or with new local governments. More broadly for international history, we might graph alliance structures on a host of different questions, viewing what happens if we remove certain countries from the equation. Modeling is not the end of the story, of course. Any models we make need to be tied back to documents and to traditional techniques and standards of proof.

Mapping

[61] Moretti, "Network Theory, Plot Analysis," *New Left Review* 68 (March-April 2011), newleftreview.org/II/68/franco-moretti-network-theory-plot-analysis.

[62] Moretti, "'Operationalizing': Or, the Function of Measurement in Literary Theory," *New Left Review* 84 (November-December 2013), newleftreview.org/II/84/franco-moretti-operationalizing.

Of all the digital techniques outlined here, the most established in history is that of mapping. Geographical Information Systems (GIS) have been a key part of historical geography for many years.[63] Its application has been part of a "spatial turn" in certain strands of historiography, especially environmental history.[64] For instance, GIS was used in Geoff Cunfer's innovative history of the Great Plains, tools like the *Digital Atlas of Roman and Medieval Civilizations* and *A Vision of Britain through Time*, and a pioneering digital history project led by William G. Thomas, III, and Edward L. Ayers, "The Differences Slavery Made".[65] Now, with the pioneering work of Stanford University's Spatial History Lab, what Richard White calls "spatial history" seems ripe for takeoff.[66] Already Caroline Winterer has used maps in her investigation of the republic of letters, and more work is in progress.[67] The historiography of American foreign relations has not quite taken a "spatial turn" in the sense of adopting, as White and others have, the ideas of Henri Lefebvre and other geographers.[68] But international historians have always been interested in space in one way or another. After all, the discipline is fundamentally about how territory is conceived *geo*politically.

GIS allows for the mapping of any data point that has a corresponding location, as so many of our sources do. Once we have that kind of data, all kinds of maps can be placed on top of one another, adjusted to show change over time, and so on. Mapping brings the prospect of

---

[63] Ian N. Gregory and Paul S. Ell, *Historical GIS: Technologies, Methodologies and Scholarship* (New York, 2007).

[64] Anne Kelly Knowles (ed.), *Placing History: How Maps, Spatial Data, and GIS Are Changing Historical Scholarship* (Redlands, 2008); David J. Bodenhamer, John Corrigan, and Trevor M. Harris (eds.), *Spatial Humanities: GIS and the Future of Humanities Scholarship* (Bloomington, 2010).

[65] Geoff Cunfer, *On the Great Plains: Agriculture and Environment* (College Station, 2005); "Digital Atlas of Roman and Medieval Civilizations," darmc.harvard.edu/icb/icb.do?keyword=k40248&pageid=icb.page188865; "A Vision of Britain through Time," visionofbritain.org.uk; William G. Thomas, III, and Edward L. Ayers, "An Overview: The Differences Slavery Made: A Close Analysis of Two American Communities," *American Historical Review* 108 (2003): 1299-1307; "The Differences Slavery Made," valley.lib.virginia.edu.

[66] Richard White, "What is Spatial History?" (1 February 2010), stanford.edu/group/spatialhistory/cgi-bin/site/pub.php?id=29.

[67] Caroline Winterer, "Where is America in the Republic of Letters," *Modern Intellectual History* 9 (2012): 597-623; stanford.edu/group/spatialhistory/cgi-bin/site/projects.php.

[68] Henri Lefebvre, *The Production of Space* (Chicago, 1991).

visualizing some of the most basic aspects of international history. One of our first efforts, as part of our "Declassification Engine" project at Columbia University, has been to begin mapping cable traffic month by month.[69] From there, imagine overlaying material capabilities with military potential and alliance structure, for instance, or comparing those capabilities with the amount of bureaucratic attention paid to the countries in question. Classic questions of international history will suddenly become, quite literally, visible.

Conclusion

Historians might understandably be nervous about the idea of having to learn an entirely new set of methods to explore contemporary diplomatic archives. In fact, at least some of the aforementioned approaches can and will produce tools that anyone will be able to use. But it could take some time, and meanwhile there are tremendous opportunities for new discoveries to be made by historians who are willing to work collaboratively. To improve certain methods, such as network extraction and topic modeling, computer scientists need to consult with historians and others with deep knowledge of the documents. This research tends to be an iterative process, in which methods are continually refined to produce results that are both valid and significant. While computer scientists necessarily focus on making new discoveries in their own field – and are not usually eager merely to apply known techniques – many look to other disciplines to demonstrate that these discoveries really do help us better understand real world phenomena. As the Princeton computer scientist David Blei writes, even if new statistical models "are meant to

[69] Declassification Engine, "The Sphere of Influence," declassification-engine.org/index.py?section=sphere#". We thank Kalev Leetaru and Dainis Kiusals for putting together this project.

help interpret and understand texts," it is still the historian's job to do the interpreting. "Using humanist texts to do humanist scholarship," he concludes, "is the job of a humanist."[70]

We as historians can do a better job once we realize, as even anti-cliometricians did four decades ago, that each time we write "most" or "likely" we are making quantitative or probabilistic judgments. At least some of those judgments could now be made more precisely, and in a way that can either be validated or disproven. Unlike a lot of the quantitative data used in social science research, ours will not come from coding by research assistants or self-reporting in surveys. We do not, in other words, have to join the international relations researchers who struggle to rate wars on a scale of one to five, or fall into the fallacy sociologists commit when they equate attitudes recorded in polls with actual behavior. We have the immense advantage of using primary sources, and can now use them in a whole new way. Even if we can never use the whole corpus, we have enough of it to mitigate the selection bias and out-of-sample issues that bedevil other disciplines. We need not become obsessed with running regressions or pursuing statistical significance as an end in itself. Instead we can combine computational methods with our traditional strength in closely reading our sources and attending to their context.

Twenty-five years from now, or whenever the next edition of *Explaining the History of American Foreign Relations* is published, much of this essay will likely seem dated. That, at least, is our hope. Our hope is that more and more digitized and born-digital documents will become available, more even than we envisage here. This could come about if historians begin to crowd-source the problem by pooling our archival discoveries in a virtual archive. We hope too that advances in data science continue at an even more rapid pace, too rapid for us to imagine all the implications. Above all, we hope that historians will take up some of these discoveries, determine

---

[70] David M. Blei, "Topic Modelling and Digital Humanities," *Journal of Digital Humanities* 2 (Winter 2012), journalofdigitalhumanities.org/2-1/topic-modeling-and-digital-humanities-by-david-m-blei/.

which have practical utility, and help to develop new ones. Some will be found wanting, but we cannot be afraid to fail. If we do not at least try to come up with new means to cope with the infinite archive, we will not even realize what we have missed. If at least some of us start to work together and work across disciplines, we can begin an exciting new period of experimentation, and perhaps even lead the way in the reinvention of history as a data science.