Turning History into Data:
Data Collection, Measurement, and Inference in HPE

Alexandra Cirone[*] Arthur Spirling[†][‡]

Abstract

There are a number of challenges that arise when working with historical data. On one hand, scholars often find themselves with too much archival data to read, code, or compile into large-N datasets; on the other hand, scholars often find themselves dealing with too little information and problems of missing data. Selection bias, time decay, confirmation bias, and lack of contextual knowledge can also be potential obstacles. This essay serves to identify common threats to inference when performing historical data collection, and provide a number of best practices that can guide potential scholars of historical political economy. We also discuss new advances in data digitization, text-as-data, and text analysis that allow for the quantitative exploration of historical material.

Keywords: missing data, selection bias, digitization, OCR, text-as-data, text analysis

[*] Department of Government, Cornell University, White Hall, Ithaca, NY 14850; aec287@cornell.edu.

[†] Arthur Spirling, Department of Politics, Center for Data Science, New York University, 60 5th Ave, New York, NY 10011; arthur.spirling@nyu.edu

[‡] The authors would like to thank Jeff Jenkins, Broadstreet editorial team, and anonymous reviewers for their particularly useful comments and feedback.

Introduction

It is a truth universally acknowledged, that working with historical data can be challenging. While poring over dusty tomes with elaborate script in a wood-paneled archive seems idyllic, more often than not researchers are grappling with incomplete inventories, damaged records, and indecipherable text. In short, they are attempting a puzzle, knowing that at least a few pieces are missing. Two seemingly opposite problems emerge. On the one hand, scholars find themselves with too much data: the digitization of national libraries and archives yields thousands of records which are extremely time-consuming to read, code, or compile into organized large-N data sets. On the other hand, analysts find themselves with too little information: they may be unclear as to what is missing and why, or what was never there in the first place. This is an intimidating state of affairs.

Yet it is possible to identify common threats to inference when performing historical data collection, and there are a number of best practices that can guide potential scholars of historical political economy (HPE). And thanks to technological advances in data digitization, text-as-data, and methods of text analysis, one can quantitatively explore large amounts of text. This essay discusses the role of data in historical political economy, and specific challenges scholars face when working with that data.

Part I discusses the finding of historical data. More specifically it reviews a number of common challenges a researcher may face, including selection bias, missing data, and time decay. It grounds these concerns in terms of their potential threats to inference, and connects the specifics of HPE to broader statistical problems (and solutions).

Part II discusses how to convert this found history into data. It reviews advances in the collection, digitization, and conversion of historical records. This section highlights the role of OCR (optical character recognition) in digitization, and how historical records make standard OCR challenging. It nonetheless gives reasons for optimism, given recent advances (especially) in digital humanities. We also emphasize the possibilities of text-as-data as an academic subfield for HPE, with new ideas on how to treat historical sources as data to be collected, processed, and analyzed.

Finally, Part III discusses an application of these methods that is relevant for HPE, in the form of text as data in legislative studies. Historical data such as roll call votes or legislative speeches have long been used to study the polarization and behavior of elites, in the both the US and other countries, and we discuss both the use of such methods and recent advances. These examples hopefully provide inspiration for the uses and range of text analysis in HPE.

1 Data Collection, Selection Bias, and Inference

All data collection has the potential to induce bias in subsequent estimates. But for reasons we explain below, this is especially likely for historical work. Records are subject to selection bias, in that the data when *collected* is not representative of the population of interest. The very existence of a historical document, and the contents within them, are endogenous to preferences and norms of the time. Authors choose to record specific information, governments choose to publish certain types of information, and elites choose to preserve only a subset of records. But even if collected comprehensively in some sense, the data that is *preserved* for the analyst is likely a non-random sample (Lustick 1996; Collier and Mahoney 1996; Inwood and Maxwell-Stewart 2020). Beyond these issues, researchers should also be aware of other problems like confirmation bias and time decay. Of course, the threat of the "drunkards search"—looking only at data that is easiest to obtain—is a long-standing concern in all observational social science work (Kaplan 1964). The issue in historical political economy is that the searchers don't know they are drunk. We now discuss a number of common data challenges, their potential threats to inference, and corresponding examples from work in historical political economy.

Missing Data and Sample Selection Bias

The practice of historical research often leaves something to be desired; at it simplest, the researcher needs certain data to test her hypotheses, yet she does not have it. What is worse, she may not know the extent to which what she does have is useful. The central issue here is that, relative to the present, it is difficult to create new data about the past: researchers are choosing from the records available. And the reason data is missing may be directly related to the outcome of interest. Thus understanding *why* data is missing is vital for understanding its implications for research.

First, and most optimistically: data could be missing because of an accident, or events fundamentally unrelated to the data generating process. The existence of historical data is never a given, and despite the best efforts of archivists over the centuries, data gets lost. Bookworms, water damage, or brittle materials can hasten the demise of pre-modern records; and repositories can fall prey to natural damage or societal conflict. For example, the 1890 US Census was burned in a fire; the Libraries of Alexandria, Constantinople, and the US Library of Congress were destroyed numerous times; the Public Record Office of Ireland was destroyed in 1922; in the 21st century, archives and libraries in Egypt, Iran, Iraq, Bosnia and Herzegovina were looted or destroyed.

In the case where the subset of data that is destroyed is “missing completely at random", it should not induce bias in empirical work. It does, however, potentially limit the types of questions researchers can explore. In any case, the plausibility of such random damage is not a given. A careful researcher should provide a detailed discussion of this, and the larger data generating process.

Second, the data could be missing as a function of context or culture. Perhaps customs were simply not recorded—potentially because they were so commonplace, there was thought to be no need. Often then, culture and belief systems must be inferred from a patchwork of different historical sources. This is particularly true when it relates to the practices of ‘ordinary’ citizens; the exploits of kings were recorded, while the daily life of peasants generally were not.

An interesting example of this can be found in marriage practices. Historians have sought to examine the lives of Russian peasants, using the rates of marriage in serf estates in the 18th and 19th centuries. In particular, and counter to existing literature, historian Bushnell

(2017) provides evidence of female resistance to marriage and high celibacy rates in a select number of Russian provinces. Bushnell does meticulous archival research and triangulates a variety of different archival sources (household inventories, documents of landlords, petitions from serfs demanding female compliance, etc) to document this phenomenon; however, the reasons as to why female peasants embraced this practice, or the cultural or contextual beliefs that made this practice possible, are lost to the historical record (Dennison 2018).

Data could also be missing because some groups were systematically and deliberately left out of history. These "archival silences" are common, almost by definition, for marginalized individuals (Thomas 2017). An example of this phenomenon is found in the work of Doniger (2010), who notes that historical records on Hinduism tend to present the perspectives of dominant groups at the expense of women and lower castes. Similarly, we have shockingly little information on the lives of those captured and transported in the Atlantic Slave Trade. Recent data coordination efforts are attempting to mitigate this lack of data, however (see "Enslaved: Peoples of the Historical Slave Trade," at `www.enslaved.org`).

Third, data could be missing due to poor state capacity, lack of resources, lack of institutions, lack of proximity to record keepers, or other government weakness. Clearly, stable countries, states, or towns with ample resources might be better at keeping and storing records than remote areas, poor areas, or areas facing civil war, economic collapse, or famine. Particularly if there is variation within a country, the lack of data can tell a story by itself, by indicating levels of state capacity or infrastructure. For example, Garfias and Sellars (2020) focus on outbreaks of epidemics in Mexico in the 18th century, to study rent seeking by politicians and the value of local office. To do so, they digitized archival data on office sales in 102 districts in central Mexico from 1702 to 1750, as well as data on smallpox, measles, and 'matlazahuatl' disease. When compiling this information, the author note that data from districts in Mexico City and larger districts are disproportionately available, compared to smaller and more rural districts. The reason for this is not known—it could be that elites did not bother to record data in less notable, provincial areas; or it could be that officials simply had no information about these places. Of course, an alternative explanation could be they were remote and had fewer outbreaks (Sellars 2020).

Fourth, some historical records or data that are missing might be omitted for strategic reasons. This could be sensitive or politically damaging information classified or left out (Connelly et al. 2020), or certain records preserved over others for political purposes. For example, Suryanarayan and White (2020)'s recent study of social-status and inequality in post-Civil War American South notes such challenges. In 1900, the US census decided to purposefully stop reporting county level tax collections. This means that researchers can construct measures of state capacity and hollowing out before that point, but not after. The authors chose the Reconstruction period as a test case for bureaucratic manipulation, so it is perhaps no surprise to see strategic reporting of administrative data; however, the government's reporting decision is challenging for researchers trying to find evidence of changing state capacity.[1]

Finally, in trying to assess the extent of a missing data problem, it is important to recognize the difference between missing data and data that never existed. One looks like the other, but absence of evidence is not necessarily evidence of absence. Determining which is which is especially difficult in cases in developing regions, especially those with few administrative records. For example, a county in the 19th Century records a district with no school data. Did that region literally have zero schools, or did it have schools but no data was collected? Upon further research, you might discover that district is mountainous with (then) low population, and thus little demand; the administrative district exists, but there were indeed no schools built.

A good example of this issue relates to the historical study of political dynasties (Smith 2018; Berlinksi, Dewan and van Coppenolle 2014), or to what extent politicians or candidates have had a blood relative in power. Collecting data on dynastic links for candidates from pre-19th century elections is often difficult. Typically, a combination of name matching and

[1] In this case, the authors had to be creative in conceptualizing alternative measures; they looked at county age heaping, bureaucratic employment at county level, and census collector level age heaping analysis. They found that hollowing out of the state (whose data was omitted from the census) continued into the 20th century.

biographic research from historical records is used to code a dynastic link, but if familial data is missing, it is not clear whether a politician can be coded as "non-dynastic" or should instead be coded as missing. This choice matters, because it will change the number of dynasties recorded. If the researcher could show case-specific evidence that family connections were glorified in politics, and frequently mentioned in newspapers, this might assuage fears about missing dynastic links that are actually present. Alternatively, one could subset the sample to cases where dynastic links are more reliable (Smith 2018), or try to estimate a proxy based on common surnames (Querubin 2016; Geys and Smith 2017). But this potentially introduces measurement error (Smith 2018), and interferes with scholars' descriptive inferences.

How can we mitigate these challenges? One best practice is to incorporate challenges with the historical data directly into the research design. This first and foremost means grounding the research design using theory (Gordon and Simpson 2020), and letting such choices guide data collection. More specifically, the incorporation of Directed Acyclical Graphs (DAGs), a tool of causal inference, can help clarify approaches (Pearl 2009; Cunningham 2020). DAGs are diagrams that map the causal relationship between the main independent variable of interest (the "treatment"), and the dependent variable (the "outcome"), and serve to illustrate the causal relationship between two. However, in showing these relationships, it is also essential to model all mediating, colliding, and confounding paths. As Schneider (2020) demonstrates, using a number of examples from economic history, the use of DAGs can help researchers explicitly model sample selection bias.

HPE researchers should be explicit about what materials will be included, and how they were chosen. In traditional survey research, one would define a population of interest and thus a *sampling frame* (Lewis-Beck, Bryman and Futing Liao 2004): literally, the total set of items from which one is drawing a subset to study. Obviously, in many historical projects, elucidating the population itself may be prohibitively difficult: the analyst simply does not know what the 'complete' record looks like, and is thus unclear about what their sample might represent. In part to mitigate such concerns, Lee (2017) recommends a series of steps. These include an exploratory trip to the archives, the cataloguing of available data, and a selection criteria that will guide (and justify) which files are to be consulted and digitized. Further, Lee (2017) provides an application such a sampling frame method, using material from the National Archives of India on the Indian Emergency (1975-1977). This criteria should be made as part of the research design, and then all material reviewed within it; additional material may be sampled later, with the creation of additional sampling frames. This type of systematic data planning helps avoid the bias introduced by "going down the rabbit hole" of *ad hoc* collecting of historical records.

For estimating causal effects, incomplete historical data can make it difficult to appropriately account for both observed and unobserved confounding variables (Gordon and Simpson 2020; Dunning 2012), and even small amounts of missing data can bias results (Broderick, Giordano and Meager 2020). Empirically, there are a number of methods to deal with missing data, biased data, or endogenous sample selection. These include imputation, weighting, or subsetting the sample by geographic area or time period where data is more complete (Gelman and Hill 2006; Lall 2017). Sample selection models can also be used to deal with nonrandom data (Vance and Ritter 2014). Finally, one could adopt design-based inference—for example, utilizing natural experiments to control for endogeneity (Dunning 2012).

Time Decay

There are also threats to inference relating to the passage of time itself. This not only affects what data is available, but how data can be compared or aggregated. In general, we expect to have fewer records the further back we go. Partly this is due to technological and cultural shifts: we can imagine Norman bureaucrats kept less comprehensive and less numerous records than Edwardian ones. But it is also due to the way that data decays: assuming a *fixed* rate of loss (e.g. floods or fires) per period, we will have less to work with from 1000 years ago than 100 years ago. And there is no particular reason to assume a fixed rate—it may well be increasing. In any case, the historical record will always be incomplete.

This causes some subtle problems that may not be immediately apparent to social science researchers. For one thing, it is difficult to know precisely when a given class of event began or ended. As an example, suppose one is interested in the history of a given idea—such as “equal treatment" as a justification for progressive tax regimes (e.g. Scheve and Stasavage 2016). Finding the “first" use of a particular term or concept is vital to such a claim, but this cannot be known with certainty—there could always be an earlier instance that was missed (or is missing). Similarly, the “last" time a given idea is invoked cannot be definitively established. This “Signor-Lipps" (Signor and Lipps 1982) effect is well known in paleontology, but it has implications for historical political economy too: some care is required in our claims about the rise and fall of particular practices.

In historical research, we often try to collect data that tracks individuals across time periods. We want to “link" the (often census) records over time. New automated methods can assist with this creation of longitudinal datasets by matching people acroess sources (e.g. Feigenbaum et al. N.d.). However, the linkage success rate will be higher for individuals who stayed in one place and kept their name; it will be much lower for individuals who moved geographic location, occupation, or changed their name (Inwood and Maxwell-Stewart 2020). Thus the universe of of successfully linked records may not be representative, and is likely to be systematically different, from the population of interest.

A related issue is that information collected for the *same* units can change in nature over time, either due to new data collection processes or variable definitions, or other contextual dynamics. Nix, Dahis and Qian (2020) provide an excellent example of this tendency, by looking at the documenting of racial identity in the US census, from 1880-1940 using longitudinal (matched census) data. Among other demographics, census collectors gathered data on race, and the authors are able to show that a large number of African American men experienced a change in racial identity over time (being recorded as black in one wave of the census, and then later being recorded as white). In the time period of the study, the United States had a culture that was increasingly discriminatory, so “passing for white" became a strategic response (for those who could). The census is already an interesting example of potential inconsistency in data collection; enumerators varied within and across censuses, and in this example race was judged by the enumerator. But this also demonstrates that a supposedly “objective" demographic indicator was not fixed, and changed as a function of context.

Similarly, when collecting data over time, researchers must be sensitive to the fact that while country, state, district, village borders seem stable now, their creation and maintenance is highly fluid and endogenous. A town in one historical period might not geographically correspond to its modern equivalent, as place names change over time. The creation of boundaries has been studied by a series of papers that look at the historical foundations of the modern territorial state. For example, Abramson (2017) uses data on European states between 1100 and 1790 to show that territorial boundaries were born of patterns of economic development and fragmented political authority, such that the emergence of towns and cities then caused the formation of small and independent states. Acharya and Lee (2018) note that the mutual recognition of state boundaries among rulers is a contemporary phenomenon. They use a formal model and the illustrative case the English-Scottish border disputes since the 17th Century to show that the territorial state system was a solution to manage competition in the market for governance. Thus it is important to pay careful attention to the definition of units of analysis over time.

Knowing the Case

In depth knowledge of the historical case is exceptionally important. It is vital for understanding the data generating process and avoiding (too much) measurement error.[2] While this is intuitive, advances in the availability of digital collections has also made it easier to access and even directly download data, which could very easily reduce engagement with the case and its particulars (and might make it easy to ignore the selection bias that comes as a result of digitization choices made by someone other than the researcher). Putnam (2016) writes that increased access to information is “radically diminishing the role of place-specific

[2] While we focus primarily on data collection, it is also important for data analysis; natural experiments rely on correct interpretation of historical facts in order to justify identifying assumptions (Kocher and Monteiro 2016).

prior expertise as a prerequisite to discovery." Putnam was speaking of historians, but this problem is shared across all fields in historical political economy; it is easy to begin work with data out of context. Gaps in contextual knowledge, culture and norms, or details of the historical time period in question can lead to inaccurate collection or interpretation of data.

Sometimes historical records contain inaccurate data, from their creation. Scholars of historical cartography are likely familiar with the fictitious Mountains of Kong, a large mountain range that crossed much of West Africa in the maps created between 1798 and 1890. Yet these mountains were inserted as a result of a single explorer's account and subsequent speculation. They were not real; they "never existed except in the imagination of explorers, mapmakers, and merchants" (Bassett and Porter 1991). In an interesting case of cartographic path dependence, where maps that followed were reproduced with this original mistake, these mountains were present in map-making for almost a century. Bassett and Porter (1991) argue this demonstrates that maps are social constructs rather than objective “scientific" facts, drawn for specific audiences in their historical context.

Sometimes historical records involve strategic fabrication. While the modern world grapples with new forms of “fake news", just because a historical document exists doesn't mean it is real. A researcher without contextual knowledge might assume it is. One notable example is that of Benjamin Franklin, the esteemed 18th century US diplomat, statesman, and Founding Father. While serving as diplomat to Paris in 1782, he was trying to achieve reparations from Great Britain and wanted to engineer sympathy. He fabricated a supplement to the Boston Independent Chronicle, detailing fictitious atrocities committed by Native American populations on American citizens, to circulate in Europe with the hopes of having this sensational 'news' reprinted.[3] It was typographically convincing and professionally executed (Franklin was a printer, after all), and even included fake advertisements discussing land for sale and a missing horse. At the time, while some elites recognized it as

[3] To see the document and its transcription, consult the US National Archives at https://founders.archives.gov/documents/Franklin/01-37-02-0132.

sensational and therefore not real, it was picked up by European news outlets. If it could trick a contemporary, it could very well fool a modern day researcher.

Case specific knowledge is also important in recognizing how the organizational units of society, and therefore our data, are formed; this is something we often take for granted (or presume are locally exogenous). This could be data on country, state, or municipality borders; or this could be the classification of economic or civil society organizations; or this could be thresholds for voting, welfare assistance, career progression. This data is political. For example, Slez (2020) uses historical evidence to analyze the determinants of electoral populism in the United States, and in doing so highlights the case of the strategic division of the Dakota Territory in 1889 into the two separate states we know today. The ultimate division of territory was a byproduct of political compromise and strategic motivations of local actors, and was only one of many proposals put forward. But these states's borders are highly endogenous, and were formed in direct response to changes in political geography and a specific set of historical events. Any data collected on administrative units should take these possibilities into consideration.

Researchers could fall victim to incorrect interpretation of historical records. This was the case for Dr. Naomi Wolf, and her yet-to-be-released book "Outrages: Sex, Censorship and the Criminalisation of Love," which looked at the persecution of homosexuality in 19th century England. Focusing on the British Obscene Publications Act of 1857, she collected archival data from Old Bailey court records, coding the phrase "death recorded" as evidence an execution had occurred. Except in the Victorian legal language of the time, this phrase actually meant the opposite – that the presiding judge had abstained from issuing a death sentence (Ward 2021). This was pointed out by a number of journalists, and the book has been retracted for revisions. In fact, the historical case of sentencing at the Old Bailey is an example of another contextual challenge. Historians have documented that even individuals convicted often did not receive the punishments corresponding to the officially recorded sentence (Shoemaker 2019). As a result, recorded outcomes might not reflect realized outcomes. A lack of contextual knowledge can have serious implications for data collection.

Related, a researcher might fail to "read history forward" in the process of data collection (Capoccia and Ziblatt 2010; Ahmed 2010; Møller 2020), and thus attribute modern day biases and interpretation to historical texts. This not only could mean that the selection of historical evidence is biased, but that a researcher might infer ex-ante causes from ex-post consequences. This could manifest in a number of ways. For example, our modern day conceptions of data sources such as newspapers, government records, parliamentary archives, etc might not match the reality of the time period in which we study. Today, the *New York Times* is a dominant, respected, national mainstream outlet in the United States. While it has existed since 1851, in its initial form it was one of many papers on the brink of bankruptcy, struggling for readership in a Gilded Age of yellow journalism where party interests often dictated the news. As Porwancher (2011) documents, its reputation for objectivity, and standards of ethical and fair reporting, only developed as a result of its purchase by Adolph S. Ochs in 1896 (and even this was a strategic rebranding that did not always reflect its reporting). Researchers unfamiliar with the era and looking for a reliable source of data might first turn to the pre-1896 New York Times, and overestimate its influence or mischaracterize its partisan bias.

These data collection concerns can not be fully avoided. But researchers should make every effort to be aware not only of the data generating process, but also how the historical material was recorded and then preserved. Historical political economy is not a field where one "drops in, gets the data, and gets out." This often means consulting with, and even vetting historical analysis, with historians (Møller 2020). Not only are historians experts in the case in question, but HPE scholars can benefit from their open vantage point. As Kreuzer (N.d.), writes: *"Unlike social scientists, historians are far less hemmed in by epistemological priors, theoretical presuppositions, or methodological exigencies and consequently are capable of providing accounts that are nuanced, comprehensive, and, above all, attuned to changes across time."* A rigorous treatment of a historical case by a trained historian or other expert, unconstrained by a specific research design, provides a buffer against reading history forward.

Confirmation Bias and Other Problems

It is important for a researcher to be aware of their own individual biases, and how this affects their selection of historical evidence (Lustick 1996; Lee 2017). In particular, a researcher might introduce bias by "cherry picking" data: seeking out sources and evidence most consistent with their own argument, while suppressing evidence against (Moravcsik 2019). This practice is often unconscious or unwitting, in that a researcher might not realize how confirmation bias is dictating their selection of historical evidence.

It could also manifest in the extent to which a single historical account can dictate the subsequent thinking and the analysis of a research project. Lustick (1996) demonstrates this by discussing Barrington Moore, whose canonical work helped define the comparative study of democratization across Europe. Moore (1966)'s analysis of the English case relied on a specific interpretation of the English Civil war as a bourgeois revolution (which fit his theory very well); however, this is only one of many accounts historians had put forth and had even fallen out of favor at the time. Today, algorithmic advances in search engines (from libraries to digital archives to Google) reinforce confirmation bias, by suggesting best matches or similar sources to the one searched. An enthusiastic scholar awash in sources might forget that this method is less than systematic; as Underwood (2014) writes "full- text search can confirm almost any thesis you bring to it."

The inclusion of secondary sources on sensitive topics adds another potential source of bias; a researcher must account for the information within the secondary source, as well as the motivations and biases of the historical text's original author. Some historical records are more objective than others, and secondary sources often aggregate primary source data in ways that are not immediately transparent. One useful example of this comes from Dell (2012*b*), who posits a relationship between severe drought and an increased likelihood of insurgent activity during the Mexican revolution (1910-1918). This required collecting data on municipality-level rainfall, land ownership, and incidences of violence against the government (as well as modern day development outcomes). Data on insurgency was collected from multiple regional histories and from the Encyclopedia of Mexican Municipalities,[4] a secondary source that is a compilation of various administrative records and accounts from historians, to detail major events of the local area. This project provides a nice demonstration of the challenges of secondary data. First, different sources might be more or less inclined to record violent events, for a variety of reasons, and second, there may be consistency issues if each case is being sourced from different secondary sources. Alternative coding criteria or challenges to the coding of individual cases could change results (Maurer 2013), though notably Dell provides a detailed Data Appendix (Dell 2012*a*) in which she lists each municipality and the explicit historical record wording that drove the coding (a best practice for this type of challenge).

The best way for a researcher to guard against confirmation bias is to be aware of its existence. To this end, they should make the selection and interpretation of historical records more transparent. A number of recent scholarly papers and initiatives have provided resources to encourage transparency in qualitative research (Moravcsik 2019). In political science, this led to the creation of the Qualitative Transparency Deliberations (QTD)[5], an initiative of the Qualitative and Multi-Method Research (QMMR) section of the American Political Science Association (APSA). The QTD initiative formed 13 working groups, focusing on a number of topics on transparency in qualitative research, and issued a series of reports covering everything from research ethics and transparency, to analytic approaches and forms of evidence, to research in challenging contexts (Jacobs et al. 2021). Similarly, the Annotation for Transparent Inquiry (ATI) Initiative[6] is a platform that allows scholars to generate and share digital annotations, which provide additional information about data sources or interpretation. Procedures outlined in such initiatives can help historical political economy researchers be upfront about their data collection, and their choices of historical cases.

[4] See http://www.inafed.gob.mx/work/enciclopedia/EMM15mexico/index.html

[5] See www.qualtd.net/

[6] See https://qdr.syr.edu/ati/ati-initiative.

Historical political economy can also take cues from experimental work. Transparency in quantitative research is relatively straightforward, and typically involves the sharing of data, code, and other files for replication. Experimental work has incorporated preregistration in to the research process, by having scholars submit a "pre-analysis plan" (PAP) that consists of a document of how a researcher will collect and analyze data in their experiment. This is submitted to a public repository before a project begins (Olken 2015; Humphreys, Sanchez de la Sierra and van der Windt 2013).[7] For both qualitative and historical researchers, pre-analysis plans are more difficult to construct—thanks to missing data and incomplete cases, exploratory research, or theory building. And, of course, historical data is typical observational, meaning it has been collected or viewed before the project begins. Yet recent advances have adapted PAPs to qualitative research (Haven and Grootel 2019; Pineiro and Rosenblatt 2016; Kern and Gleditsch 2017), with modifications such as establishing data collection protocols before visiting the archives, and clearly delineating the inductive and deductive aspects of the research.

Now that we have a growing set of best practices of how to be transparent about the selection of qualitative evidence, how the evidence is analyzed, and how the ultimate results are framed, we can better address the various biases that affect historical data collection.

2 Turning History into Data

Research in historical political economy involves turning history into data, through the large scale collection and digitization of primary source materials. On the supply side, the last decade has seen a sharp rise in the digital preservation of history. National libraries are increasingly scanning documents in their collections and providing them online for interested readers; notable examples include *Gallica* (BnF) in France, the *British Library*, the US *Library of Congress*, the *National Library of Norway*, and Russia's *Presidential Library* (among others). The Covid-19 pandemic increased the demand for digital resources, and

[7] Or see DeclareDesign at https://declaredesign.org/pap

libraries have attempted to meet this shift. For example, in 2020 Mexico's General Archive of the Nation established an initiative called "Archives from Home (#archivosdesdecasa)" to encourage citizens to access its digital collections. There are also multi-country, coordinated initiatives to improve access to historical text and preserve cultural heritage; two notable ones were funded by the European Union, including the 2008-2011 EU IMPACT project,[8] and Europeana,[9] an online digital repository for European newspapers from 20 countries, dating from 1618 to 1996. While much of early digitization efforts focused on English, more are being added; thus, scholars of China can consult the The China Biographical Database Project[10] or the forthcoming digitization of 1.5 million archival files by the Chinese government as part of the Qing History Project (Mao and Ma 2012). While these projects are encouraging, it is possible that the earlier problem of scarcity gives rise to a new one of over-abundance. That is, there are simply too many texts. Without systematic metadata or ability to search, researchers may find themselves overwhelmed (see, e.g., Connelly et al. 2020). Certainly, researchers will need new methods and skills to harvest this data.

For repositories of data, already digitized and in machine readable format, it is relatively easy to scrape and analyze this data, using a number of open-source tools such as R or Python (Munzert et al. 2015). But historical records are often not available in an electronic format and must be digitized by the researcher, and some historical texts prove more challenging than others. It is important to recognize the unique challenges that come with collecting, processing, and converting of historical data, and here we discuss new technologies that make data conversion possible, namely advances in Optical Character Recognition (OCR) and text analysis in the form of "text-as-data."

[8] See https://cordis.europa.eu/project/id/215064.

[9] See https://www.europeana.eu/en/about-us.

[10] A joint initiative from Harvard University, Academia Sinica, and Peking University, at http://projects.iq.harvard.edu/cbdb

OCR for Historical Data

There are a number of tools currently available to help extract information from historical text. The digitization of historical sources usually involve the creation of page images, produced by scanning or taking a digital photograph of a document page. While this displays the format and context of the original, an image's text is neither searchable nor manipulable (and space intensive to store). This is where "optical character recognition" (OCR) may be used: this process moves images to machine-readable text.

In principle, using OCR is straightforward. The types of images with the best conversion rates are typically in color or grey-scale, high resolution (300 DPO or more), and taken with documents in a flat or resting position (so text lines are straight as possible). Then, using the selected software, the images are pre-processed and the layout, character and language settings, and dictionaries are chosen. There are a number of OCR software programs, though most common are ABBYY Fine Reader, Adobe Acrobat Pro, or Tesseract. When using OCR software, we are typically concerned with accuracy, or percentage of characters recognized correctly, out of the total volume of characters converted. The accuracy rates of OCR are very high for modern data; as high as 99% of characters are correctly recognized.

Yet historical documents present a challenge for OCR. In particular, printed documents from the Gutenberg age up until and including the 19th century have a number of features that frustrate standard software.[11] They have atypical characters and non-standard typography; in particular, ligatures such as 'ct', 'ff', 'ffl', and 'fi' lead to high error rates. These texts also have complex layouts or ink that is fading or has bled through the pages.

Historical languages also feature wide variation in usage across time and space; in a single historical document by a single author, the same word can be spelled in a number of different

[11] To further complicate matters, sometimes scholars have prior scans of historical documents from years ago, that cannot be updated due to fragile materials; for example, a collection of George Washington's letters was digitized from scanned microfilms, which distorts the original image. See https://userpages.umbc.edu/~skane/pubs/MM-36.pdf.

ways.[12] As a result, preexisting modern training data may be of little use for historical sources. Instead, individual OCR models must be trained on the specific typography of a historical source, which involves creating transcriptions from a portion of real data, from scratch. This cannot generalize to other documents, and can be particularly time and labor intensive.

Handwritten manuscripts also provide a particular challenge for OCR; the conversation of handwriting to machine-readable text is called "handwriting recognition" (HR). Signatures are unique enough to be used for identify verification purposes, implying there's almost infinite variation in the style, shape, and format of characters used. The joining of letters as a result of cursive handwriting, or inconsistent spacing also pose difficulties; not to mention the fact that handwritten documents are subject to smudges, cross outs, stray marks, and general wear and tear. For now, handwriting recognition has been limited, and mainly used to recognize common keywords in historical texts to allow them to be indexed and searchable, later on.

Accuracy rates for historical OCR vary, but are substantially lower—HPE scholars should know that even with the clearest images and high-tech OCR software, there still will be mistakes. As Hill and Hengchen (2019) write in reference to OCR, it s a stylized fact that "researchers spend 80% of their time pre-processing data, and only 20% analysing it." Data cleaning will be necessary, and ideally by someone with knowledge of the source context. The error rate here is highly variable, but this demonstrates the importance of running a digitization test, or pilot, on a managable subsample of texts. Scholars should choose a small but representative sample of every type of material, and run it through OCR software to gauge time and error rates. If it takes longer to debug the OCR output than to manually

[12] Take an example from a 15th century book of manners for children, a digitized tome held in the British Library's collections (Unknown 15th century manuscript): the phrases 'Pyke notte thyne errys nothyr thy nostrellys' and 'And chesse cum by fore the, be not to redy' simultaneously feature in the book, but note the atypical language and the inconsistent spelling of the simple word "not". These translate to 'Don't pick your ears or nose', and 'Don't be greedy when they bring out the cheese'.

enter the data, then it might be more efficient to send the project out for data entry (there are firms that can provide transcription services), or invest in specialized equipment and trained professionals.

On the bright side, OCR for historical data is a separate field with a number of useful resources (Cordell and Smith 2019; Piotrowski 2012; Tanner 2006). Software is improving by adding a multitude of fonts, character sets, languages, and page attributes. There have been a number of advances in character recognition specifically for historical work, constructing generalized models trained on a range of pre-19th century fonts that could be applied to a wider range of historical sources, and whose document and word accuracy rates reach over 90% (Springmann and Lüdeling 2017). Further, scholars are developing clean corpora and lexicons of historical texts on a case-by-case basis; for example, for 18th and 19th century Slovenian books and newspapers (Ines Jerele 2012), 19th century British newspapers Simon Tanner (2009), American newspapers before the Civil War Cordell (2015), and a number of other countries (19th century newspapers from Finland, Germany, Mexico, the Netherlands, the United Kingdom, and the United States (Cordell et al. forthcoming) among others. Documents written in languages other than English, or even multilingual text, are less of a problem that they once were (Lucas et al. 2015), and history buffs can even explore the realm of historical cryptography using cypher databases (Megyesi 2017; Megyesi et al. 2020).

Advances in OCR and digitization are also being driven by the rapidly growing field of digital humanities. Those fields "combine traditional qualitative methods with quantitative, computer-based methods and tools, such as information retrieval, text analytics, data mining, visualization, and geographic information systems (GIS)" (Piotrowski 2012). They are also concerned with the availability of digital text, and have contributed countless methodological innovations in natural language text processing for historical records. Thus Digital Humanities are at the very least adjacent to, if not sharing the same sandbox, as historical political economy (Warwick, Terras and Nyhan 2018).

Text-as-Data with Historical Data

Political economists engage history in a number of ways, and HPE scholars are increasingly turning to new methods of "text-as-data," which treat historical sources as data to be collected, processed, and analyzed. This follows a larger pattern in political science (Grimmer and Stewart 2013; Wilkerson and Casas 2017), economics (Gentzkow, Kelly and Taddy 2019), sociology (Evans and Aceves 2016), and policy (Gilardi and Wüest 2020). The study of history has always included literary discourse, and qualitative analysis of the meanings, rhetoric, and topics found in historical texts. But new advances in text analysis are now incorporating statistics and machine learning (supervised or unsupervised) methods to extract and analyze information in the form of electronic text.

What does this look like? First, historical sources are compiled, and then sample of the available texts is chosen (guided by research principles). Once digitized, this raw text is then converted into a structured form of data, depending on the method; most common in political science is to extract features (selected words) into a "document-feature matrix." Converting text into a numeric matrix allows for a wide range of statistical analyses. Importantly, this type of text analysis can provide information about the manifest characteristics of the data, or information about was communicated in the text, as well as latent characteristics, or information about the author of the text (Benoit 2020).

As we have described above in other contexts, being clear about the research question, the unit of analysis and the population of interest is vital. In some cases, it is possible to collect the complete corpus of relevant texts (i.e. the entire set of the annual State of the Union speeches in the United States). In other cases, facing constraints on time or resources, historical scholars may only be able to obtain a subset of documents.

Once a scholar has assembled a corpus of digitized and processed historical text, the door opens to a variety of text analysis methods. There is no globally best method for automated text analysis (Grimmer and Stewart 2013), and different research questions require different models. Text-as-data and text analysis are research methods with their own set of rules, procedures, and choices to be made by the researcher. A full treatment is outside the scope of this article; interested readers can begin with Benoit (2020), or other resources such as Wilkerson and Casas (2017), Slapin and Proksch (2014), Grimmer and Stewart (2013), and Denny and Spirling (2018). Causal inference is possible with text analysis, though considerable care is required (Roberts, Stewart and Nielsen 2020). For HPE scholars specifically, at least two issues stand out: language changes, and metadata decay.

First, the meaning of words varies over time: the word 'gay' as used in Congress or in newspapers will connote something very different in 1920 vs 2020. And indeed, scholars have studied such variation in some detail (Rodman 2020). But there are less obvious examples, and it is incumbent upon scholars to not simply count all instances of a term as if they are identical. More subtly, very different words can connote the same thing, and this needs to be accounted for in any measurement procedure: 'wireless' and 'radio' being an example. Beyond meaning, frequency changes too. It is too easy to take models designed for the present day—with a particular understanding of how common a given word is, or the education level required to comprehend it—and apply them without thought to times past. Benoit, Munger and Spirling (2019), for example, note that early US presidential addresses may appear artificially rarefied by today's standards, simply because words (like 'husbandry') that were then commonplace have become much rarer.

Second, scholars might encounter issues with metadata decay. In many projects, the metadata—who gave the speech, from which party, at what time—is as important as the text itself. Above, we made the point that all data decays, and by extension this goes for information that we need about documents. Thus tasks such as disambiguation (e.g. who *exactly* the speaker is, and to whom they are referring) become vital. But this is time-consuming and expensive, and in many cases may never be fully resolved beyond a 'best guess'. Researchers should plan for this cost and frustration (!)

3 Application: Text as Data in Legislative Studies

Legislative proceedings provide a rich source of data. In historical political economy, voting and parliamentary debates have been used to study a wide range of outcomes from party development to democratization, using 19th-20th century data from Britain, Canada, Germany, Denmark, Sweden, and the US (among other countries; for a review, see Cirone (2020)). In the study of American politics, there has been particular interest in understanding polarization of elites, with voting and text data used to measure the ideology of legislators, and to map the dimensions of political conflict over time. Here we explore how these approaches have spread from the US context to other places, and the evolution of that work, including recent innovations from scholars of HPE.

While certainly not the first to study roll call votes *per se* (see, e.g. Lowell 1902), it is hard to overstate the contribution of Poole and Rosenthal (1997) for understanding the US Congress. Their key innovation was to see the votes in Congress as an "item response" similar to the way that term is used in education. In that literature, students get questions on a test 'correct' or 'incorrect'. From large numbers of these responses and a simple statistical model, we can place test-takers on a continuum from lower to higher ability. In the case of voting in a parliament, essentially the same model can be used on the 'yes' or 'no' votes of parliamentarians to build up an ideological spectrum on which they all sit. This, in turn, can be used to study policy disagreement, gridlock, and polarization (measured as the distance between party ideal points). Poole and Rosenthal created multiple iterations of these scaling procedures, with the core intellectual product originally known as "NOMINATE".[13] These ideas led to a number of applications and the growth of a large literature, studying topics like central bank 'hawkishness' (Hix, Høyland and Vivyan 2010), or the dynamics of Supreme Court Judge ideologies over time (Martin and Quinn 2002).

Roll call votes are useful historical data, but both within and outside the US, may not be informative in some contexts. For example, voting behavior could be strategic, or a function of party control (Roberts 2007; Clinton and Meirowitz 2004). There also may be complications when studying political conflict over different historical periods. Scaling allows for the comparison of polarization and within-legislator position change across time;

[13] The database and code corresponding to NOMINATE scores, as well as background information on Congress members, since 1789 is also publicly available via the VoteView website (Lewis and Sonnet 2021). See `https://www.voteview.com/data`

however, changes in policy, legislative agendas, or party positions can induce difficulties in interpretation of the estimates (see, e.g. Spirling and McLean 2007). HPE scholars may be more likely to encounter this issue. For example, Bateman, Clinton and Lapinski (2017) use roll call voting to study how elite conflict over African American civil rights evolved in the US Congress since 1877. They demonstrate that the content of policies can change over time in ways not picked up by DW-NOMINATE, and that such estimates are also difficult to interpret if the party's position on policy changes over time. Solutions to this problem could include incorporating more information in the statistical analysis to reflect policy change over time (Bailey 2013; Clinton and Meirowitz 2004; Clinton, Jackman and Rivers 2004), or imputing votes for legislators based on historical context (Bateman, Clinton and Lapinski 2017).

Given these concerns, scholars have turned to new sources of legislative data to augment or replace roll calls. Most directly, automated text analysis can be used to extract word frequencies, specific phrases, or other text data, These may then directly inform the estimation of ideal points in legislative speeches (e.g. Laver, Benoit and Garry 2003; Slapin and Proksch 2014; Lauderdale and Herzog 2016). Alternatively, scholars have used speeches to explore the substantive meaning of ideological dimensions. For example, Diermeier et al. (2012) analyze the content of all speeches made in the US Senate from 1989-2004, with supervised learning. The central concern is to find words and phrases most associated with Republican and Democratic senators, and thus understand the themes and issues that separate modern parties. The methods used in this and related studies have been of considerable interest themselves (e.g Monroe, Colaresi and Quinn 2017).

While the recent past has been of primary interest to political scientists, scholars of HPE have used recently digitized records to estimate dynamics over a much longer window. For example, Jensen et al. (2012) extend the spirit of the studies above to include not only Congress but the public political discourse: they make use of the *Congressional Record* and the *Google Ngrams* corpus since 1873. The authors identify political phrases spoken by legislators on the floor of the House of Representatives, to impute both partisanship and polarization over time. They map these partisan phrases to phrases used in the 2 million+ database of Google books from the 19th century, effectively creating a time series measure of political ideology in the literary discourse. Their analysis suggests that polarization in public discourse is a good predictor of legislative inefficiency, and that today's polarization is not the worst of all time.

Despite—or perhaps because of—the attention on text to improve our understanding of ideology, scholars have turned to this data to do other things in HPE. So-called "topic models," in which the focus is on estimating the themes in legislative discussions over time have proved popular (Quinn et al. 2010). Text analysis can also be used to measure the "sophistication" of political communication of both spoken and written speeches over long periods, both in the US (Benoit, Munger and Spirling 2017) and the UK Spirling (2016) contexts. In that literature, the goal is to understand how elected politicians respond—in terms of how they express themselves—to changing norms of democracy and technology.

Of course, when HPE scholars take text methods to 'new' periods, they need to adjust for the estimands of interest. For example, Spirling and Eggers (2016) use parliamentary speeches in the 19th and 20th century of the UK to study the emergence of the Shadow Cabinet, a legislative institution of the opposition that for most of its history, did not officially record its members or activities. (Since it served as an informal institution, albeit an important one, scholars have very little data to work with.) By using a new way to measure the 'burstiness' of individual contributions to debate, the authors establish that the Second Reform Act (1867) was associated with a small set of very active individuals on the opposition benches (consistent with the existence of a set of MPs in leadership roles). These active individuals were more likely to be promoted to Cabinet office in later sessions, providing further evidence of an operational Shadow Cabinet. The main methodological point here is that the authors use text to estimate something specific to Westminster systems, rather than 'off-the-shelf' from say, the US context. In more recent times, scholars have moved beyond such 'mere' description, to more causal claims. Thus, Blumenau (2019) combines a design-based inference research design with data from UK parliamentary debates to demonstrate how the presence of female ministers significantly affects the participation of other female MPs. He exploits dynamic interactions recorded in 460,000 speeches across a decade and quantifies 'influence' (an otherwise difficult phenomenon to measure).

Most recently, scholars of HPE have come full circle, and used recent advances in computer science to return to the study of non-US parliament ideology. Here, Rheault and Cochrane (2020), for example, consider a 'word embedding' approach to modeling parties in the UK, US and Canada. The key idea here is to more explicitly take the *context* of words and phrases into account when modeling their authors and speakers. Writ large, this quantitative approach with appropriate sensitivity to local circumstances is exactly what HPE should aim to do.

4 Conclusion

It is clear that there have been enormous strides in the collection, digitization, and analysis of historical records. Up until recently, the costs of analyzing and coding archival records have been prohibitive, and have limited the scope of historical research and the types of research questions scholars can pursue. But advances in OCR and text conversion, combined with automated text analysis that can scrape, link, code, and analyze data, opens up centuries of possibility for historical political economy scholars.

There are challenges, to be sure. Missing data will always be a problem, and problems with selection bias abound; in this, research transparency and research design must play a leading role in HPE research. Cleaning and coding historical data takes time and effort; though interdisciplinary initiatives are sharing advice and tools across fields, to make this process easier. Text-as-data and text analysis methods have incredible potential, but must be used with care; as with any sophisticated qualitative method, a scholar must invest in learning (and then keeping up with) this fast-paced and technical field.

Historical research will always be more challenging than modern research; but therein lies the adventure. As Walter Prescott Webb wrote, in his 1959 presidential address to the American Historical Association, *"... adventure into the wilderness of the past, that* *wild country wherein one can be lost for days or weeks or months... and the lovely feature about this delirious experience is that the historical explorer moves among the dangers and hardships with complete immunity until finally he comes out in print, in point-blank range of the critic."* (Webb 1959). No one said historical political economy was for the academic faint of heart.

References

Abramson, Scott F. 2017. "The Economic Origins of the Territorial State." *International Organization* 71(1):97–130.

Acharya, Avidit and Alexander Lee. 2018. "Economic Foundations of the Territorial State System." *American Journal of Political Science* 62(4):954–966.
URL: *https://onlinelibrary.wiley.com/doi/abs/10.1111/ajps.12379*

Ahmed, Amel. 2010. "Reading History Forward: The Origins of Electoral Systems in European Democracies." *Comparative Political Studies* 43(8-9):1059–1088.
URL: *https://doi.org/10.1177/0010414010370436*

Bailey, Michael A. 2013. "Is Today's Court the Most Conservative in Sixty Years? Challenges and Opportunities in Measuring Judicial Preferences." *The Journal of Politics* 75(3):821–834.
URL: *https://doi.org/10.1017/S0022381613000443*

Bassett, Thomas J. and Philip W. Porter. 1991. "'From the Best Authorities': The Mountains of Kong in the Cartography of West Africa." *The Journal of African History* 32(3):367–413.
URL: *http://www.jstor.org/stable/182661*

Bateman, David A., Joshua D. Clinton and John S. Lapinski. 2017. "A House Divided? Roll Calls, Polarization, and Policy Differences in the U.S. House, 1877–2011." *American Journal of Political Science* 61(3):698–714.
URL: *http://www.jstor.org/stable/26379519*

Benoit, Ken. 2020. *Text As Data.* SAGE Handbook of Research Methods in Political Science and International Relations.

Benoit, Kenneth, Kevin Munger and Arthur Spirling. 2017. "Measuring and Explaining Political Sophistication through Textual Complexity." *American Journal of Political Science* 63(2):491–508.

Benoit, Kenneth, Kevin Munger and Arthur Spirling. 2019. "Measuring and explaining political sophistication through textual complexity." *American Journal of Political Science* 63(2):491–508.

Berlinksi, Samuel, Torun Dewan and Brenda van Coppenolle. 2014. "Franchise Extension and the British Aristocracy." *Legislative Studies Quarterly* 39(4):531–558.

Blumenau, Jack. 2019. "The Effects of Female Leadership on Women's Voice in Political Debate." *British Journal of Political Science* p. 1–22.

Broderick, Tamara, Ryan Giordano and Rachael Meager. 2020. "An Automatic Finite-Sample Robustness Metric: Can Dropping a Little Data Change Conclusions?" Working Paper.
URL: *https://arxiv.org/abs/2011.14999*

Bushnell, John. 2017. *Russian Peasant Women Who Refused to Marry: Spasovite Old Believers in the 18th–19th Centuries*. Bloomington: Indiana University Press.

Capoccia, Giovanni and Daniel Ziblatt. 2010. "The Historical Turn in Democratization Studies: A New Research Agenda for Europe and Beyond." *Comparative Political Studies* 43(8-9):931–968.

Cirone, Alexandra. 2020. The extension of factor analysis to three-dimensional matrices. United Kingdom of Great Britain and Northern Ireland: Edward Elgar Publishing pp. 353–367.

Clinton, Joshua D. and Adam Meirowitz. 2004. "Testing Explanations of Strategic Voting in Legislatures: A Reexamination of the Compromise of 1790." *American Journal of Political*

Science 48(4):675–689.
URL: *https://onlinelibrary.wiley.com/doi/abs/10.1111/j.0092-5853.2004..x*

Clinton, Joshua, Simon Jackman and Douglas Rivers. 2004. "The Statistical Analysis of Roll Call Data." *American Political Science Review* 98(2):355–370.

Collier, David and James Mahoney. 1996. "Insights and Pitfalls: Selection Bias in Qualitative Research." *World Politics* 49(1):56–91.
URL: *http://www.jstor.org/stable/25053989*

Connelly, Matthew J, Raymond Hicks, Robert Jervis, Arthur Spirling and Clara H Suong. 2020. "Diplomatic documents data for international relations: the Freedom of Information Archive Database." *Conflict Management and Peace Science* .

Cordell, Ryan. 2015. "Reprinting, Circulation, and the Network Author in Antebellum Newspapers." *American Literary History* 27.

Cordell, Ryan, David A. Smith, Abby Mullen, Jonathan Fitzgerald and Thanasis Kinias. forthcoming. *Going the Rounds: Virality in Nineteenth Century American Newspapers.* Forthcoming, Minnesota University Press.

Cordell, Ryan and David Smith. 2019. "A Research Agenda for Historical and Multilingual OCR." Report for the Andrew W. Mellon Society.
URL: *https://repository.library.northeastern.edu/files/neu:f1881m409*

Cunningham, Scott. 2020. *Causal Inference: The Mixtape.* Yale University Press.

Dell, Mellissa. 2012*a*. "Data Construction Appendix." *Appendix to Working Paper: Path dependence in development: Evidence from the Mexican Revolution* .

Dell, Mellissa. 2012*b*. "Path dependence in development: Evidence from the Mexican Revolution." *Working Paper* .

Dennison, Tracy. 2018. "John Bushnell. Russian Peasant Women Who Refused to Marry: Spasovite Old Believers in the 18th–19th Centuries." *American Historical Review* 123(5):1790 – 1791.
URL: *http://search.ebscohost.com.proxy.library.cornell.edu/login.aspx?direct=truedb=aphAN=1334627*
live

Denny, Matthew J. and Arthur Spirling. 2018. "Text Preprocessing For Unsupervised Learning: Why It Matters, When It Misleads, And What To Do About It." *Political Analysis* 26(2):168–189.

Diermeier, Daniel, Jean-François Godbout, Bei Yu and Stefan Kaufmann. 2012. "Language and Ideology in Congress." *British Journal of Political Science* 42(1):31–55.

Doniger, Wendy. 2010. *The Hindus: an alternative history.* Oxford University Press, USA.

Dunning, Thad. 2012. *Natural Experiments in the Social Sciences: A Design-Based Approach.* Cambridge University Press.

Evans, James A. and Pedro Aceves. 2016. "Machine Translation: Mining Text for Social Theory." *Annual Review of Sociology* 42(1):21–50.

Feigenbaum, James, Leah Boustan Ran Abramitzky, Katherine Eriksson and Santiago Perez. N.d. "Automated Linking of Historical Data." *Working Paper*. Forthcoming.

Garfias, Francisco and Emily A. Sellars. 2020. "Epidemics, Rent Extraction, and the Value of Holding Office." *Journal of Political Institutions and Political Economy* 1(4):559–583.

Gelman, Andrew and Jennifer Hill. 2006. *Data Analysis Using Regression and Multilevel/Hierarchical Models.* Cambridge University Press.

Gentzkow, Matthew, Bryan Kelly and Matt Taddy. 2019. "Text as Data." *Journal of Economic Literature* 57(3):535–74.
URL: *https://www.aeaweb.org/articles?id=10.1257/jel.20181020*

Geys, Benny and Daniel M. Smith. 2017. "Political Dynasties in Democracies: Causes, Consequences and Remaining Puzzles." *The Economic Journal* 127(605):F446–F454.
URL: *https://onlinelibrary.wiley.com/doi/abs/10.1111/ecoj.12442*

Gilardi, Fabrizio and Bruno Wüest. 2020. "Text-as-Data Methods for Comparative Policy Analysis." Working Paper.
URL: *https://www.fabriziogilardi.org/resources/papers/Gilardi-Wueest-TextAsData-Policy-Analysis.pdf*

Gordon, Sanford C. and Hannah K. Simpson. 2020. "Causes, theories, and the past in political science." *Public Choice* 185(3):315–333.
URL: *https://ideas.repec.org/a/kap/pubcho/v185y2020i3d10.1007*$_s 11127 - 019 - 00703 - 6.html$

Grimmer, Justin and Brandon M. Stewart. 2013. "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts." *Political Analysis* 21(3).

Haven, Tamarinde L. and Dr. Leonie Van Grootel. 2019. "Preregistering qualitative research." *Accountability in Research* 26(3):229–244.
URL: *https://doi.org/10.1080/08989621.2019.158014*

Hill, Mark J. and Simon Hengchen. 2019. "Quantifying the impact of dirty OCR on historical text analysis: Eighteenth Century Collections Online as a case study." *Digital Scholarship in the Humanities* 34.

Hix, Simon, Bjørn Høyland and Nick Vivyan. 2010. "From doves to hawks: A spatial analysis of voting in the Monetary Policy Committee of the Bank of England." *European Journal of Political Research* 49(6):731–758.

Humphreys, Macartan, Raul Sanchez de la Sierra and Peter van der Windt. 2013. "Fishing, Commitment, and Communication: A Proposal for Comprehensive Nonbinding Research Registration." *Political Analysis* 21(1):1–20.

Ines Jerele, Tomaž Erjavec, Daša Pokorn Alenka Kavčič-Čolić. 2012. "OPTICAL CHARACTER RECOGNITION OF HISTORICAL TEXTS: END-USER FOCUSED RESEARCH FOR SLOVENIAN BOOKS AND NEWSPAPERS FROM THE 18TH AND 19TH CENTURY." 21.
URL: *http://elib.mi.sanu.ac.rs/files/journals/ncd/21/ncd21117.pdf*

Inwood, Kris and Hamish Maxwell-Stewart. 2020. "Selection Bias and Social Science History." *Social Science History* 44(3):411–416.

Jacobs, Alan M., Tim Büthe, Ana Arjona, Leonardo R. Arriola, Eva Bellin, Andrew Bennett, Lisa Björkman, Erik Bleich, Zachary Elkins and Tasha Fairfield. 2021. "The Qualitative Transparency Deliberations: Insights and Implications." *Perspectives on Politics* p. 1–38.

Jensen, Jacob, SURESH NAIDU, ETHAN KAPLAN, LAURENCE WILSE-SAMSON, DAVID GERGEN, MICHAEL ZUCKERMAN and ARTHUR SPIRLING. 2012. "Political Polarization and the Dynamics of Political Language: Evidence from 130 Years of Partisan Speech [with Comments and Discussion]." *Brookings Papers on Economic Activity* pp. 1–81.
URL: *http://www.jstor.org/stable/41825364*

Kaplan, Abraham. 1964. *The conduct of inquiry: Methodology for Behavioral Science.* San Francisco: Chandler.

Kern, Florian G. and Kristian Skrede Gleditsch. 2017. "Exploring Pre-registration and Pre-analysis Plans for Qualitative Inference." *Working Paper* .
URL: *https://www.researchgate.net/profile/Florian*$_{K}ern4/publication/319141144_{E}xploring_{P}re-registration_{a}nd_{P}re-analysis_{P}lans_{f}or_{Q}ualitative_{I}nference/links/599455d60f7e9b98953af045/Exploring-Pre-registration-and-Pre-analysis-Plans-for-Qualitative-Inference.pdf?$

Kocher, Matthew A. and Nuno P. Monteiro. 2016. "Lines of Demarcation: Causation, Design-Based Inference, and Historical Research." *Perspectives on Politics* 14(4):952–975.

Kreuzer, Marcus. N.d. *Analyzing Time and History: Surveying the Tools of Comparative Historical Analysis.* Oxford Handbook on Politics in Time.

Lall, Ranjit. 2017. “How Multiple Imputation Makes a Difference.” *Political Analysis* 24(4).

Lauderdale, Benjamin E and Alexander Herzog. 2016. “Measuring political positions from legislative speech.” *Political Analysis* pp. 374–394.

Laver, Michael, Kenneth Benoit and John Garry. 2003. “Extracting Policy Positions from Political Texts Using Words as Data.” *The American Political Science Review* 97(2):311–331.
URL: *http://www.jstor.org/stable/3118211*

Lee, Alex. 2017. “The Library of Babel Problem: Hypothesis Testing With Archival Sources.” *Working Paper* .
URL: *http://www.rochester.edu/college/faculty/alexander*$_l$*ee/wp* − *content/uploads/*2017/11/*archives*3.*pdf*

Lewis-Beck, Michael, Alan Bryman and Tim Futing Liao. 2004. *The SAGE Encyclopedia of Social Science Research Methods.* Sage Publications: Thousand Oaks, California.
URL: *https://methods.sagepub.com/reference/the-sage-encyclopedia-of-social-science-research-methods*

Lewis, Jeffrey B., Keith Poole Howard Rosenthal Adam Boche Aaron Rudkin and Luke Sonnet. 2021. “Voteview: Congressional Roll-Call Votes Database.”.

Lowell, Abbott Lawrence. 1902. “The Influence of Party on Legislation in England and America.”.

Lucas, Christopher, Richard A. Nielsen, Margaret E. Roberts, Brandon M. Stewart, Alex Storer and Dustin Tingley. 2015. “Computer-Assisted Text Analysis for Comparative Politics.” *Political Analysis* 23(2):254–277.

Lustick, Ian S. 1996. "History, Historiography, and Political Science: Multiple Historical Records and the Problem of Selection Bias." *The American Political Science Review* 90(3):605–618.

Mao, Liping and Zhao Ma. 2012. ""Writing History in the Digital Age": The New Qing History Project and the Digitization of Qing Archives." *History Compass* 10(5):367–374.
URL: *https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1478-0542.2012.00841.x*

Martin, Andrew D and Kevin M Quinn. 2002. "Dynamic ideal point estimation via Markov chain Monte Carlo for the US Supreme Court, 1953–1999." *Political analysis* 10(2):134–153.

Maurer, Noel. 2013. "Did drought abet the start of the Syrian civil war?".
URL: *https://noelmaurer.typepad.com/aab/2013/09/did-drought-abet-the-start-of-the-syrian-civil-war.html*

Megyesi, B., Blomqvist N. Pettersson. 2017. "The DECODE Database: Collection of Historical Ciphers and Keys." *Proceedings of the 2nd International Conference on Historical Cryptology: HistoCryp* .
URL: *http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-385920*

Megyesi, Beáta, Bernhard Esslinger, Alicia Fornés, Nils Kopal, Benedek Láng, George Lasry, Karl de Leeuw, Eva Pettersson, Arno Wacker and Michelle Waldispühl. 2020. "Decryption of historical manuscripts: the DECRYPT project." *Cryptologia* 44(6):545–559.

Monroe, Burt L., Michael P. Colaresi and Kevin M. Quinn. 2017. "Fightin' Words: Lexical Feature Selection and Evaluation for Identifying the Content of Political Conflict." *Political Analysis* 16(4):372–403.

Moore, Barrington. 1966. *Social Origins of Dictatorship and Democracy: Lord and Peasant in the Making of the Modern World.* Beacon Press.

Moravcsik, A. 2019. "Transparency in Qualitative Research." *SAGE Research Methods Foundations* .

Munzert, Simon, Christian Rubba, Peter Meißner and Dominic Nyhuis. 2015. *Automated Data Collection with R: A Practical Guide to Web Scraping and Text Mining.* Wiley and Sons.

Møller, Jørgen. 2020. "Reading History Forward." *PS: Political Science amp; Politics* .

Nix, Emily, Ricardo Dahis and Nancy Qian. 2020. "Choosing Racial Identity in the United States, 1880-1940." *NBER Working paper* .
URL: *https://www.nber.org/papers/w26465*

Olken, Benjamin A. 2015. "Promises and Perils of Pre-analysis Plans." *Journal of Economic Perspectives* 29(3):61–80.
URL: *https://www.aeaweb.org/articles?id=10.1257/jep.29.3.61*

Pearl, Judea. 2009. *Causality: Models, reasoning, and inference.* Cambridge University Press, Cambridge.

Pineiro, Rafael and Fernando Rosenblatt. 2016. "Pre-Analysis Plans for Qualitative Research." *Revista de Ciencia Politica* 36(3):785–796.

Piotrowski, Michael. 2012. *Natural Language Processing for Historical Texts.* Morgan and Claypool Publishers.

Poole, Keith and Howard Rosenthal. 1997. *Congress: A Political-Economic History of Roll Call Voting.* New York: Oxford University Press.

Porwancher, Andrew. 2011. "Objectivity's Prophet." *Journalism History* 36(4):186–195.

Putnam, Lara. 2016. "The Transnational and the Text-Searchable: Digitized Sources and the Shadows They Cast." *The American Historical Review* 121(2):377–402.
URL: *https://doi.org/10.1093/ahr/121.2.377*

Querubin, Pablo. 2016. "Family and Politics: Dynastic Persistence in the Philippines." *Quarterly Journal of Political Science* 119(2):151–181.

Quinn, Kevin M, Burt L Monroe, Michael Colaresi, Michael H Crespin and Dragomir R Radev. 2010. "How to analyze political attention with minimal assumptions and costs." *American Journal of Political Science* 54(1):209–228.

Rheault, Ludovic and Christopher Cochrane. 2020. "Word Embeddings for the Analysis of Ideological Placement in Parliamentary Corpora." *Political Analysis* 28(1):112–133.

Roberts, Jason M. 2007. "The Statistical Analysis Of Roll-Call Data: A Cautionary Tale." *Legislative Studies Quarterly* 32(3):341–360.
URL: *https://onlinelibrary.wiley.com/doi/abs/10.3162/036298007781699636*

Roberts, Margaret E, Brandon M Stewart and Richard A Nielsen. 2020. "Adjusting for confounding with text matching." *American Journal of Political Science* 64(4):887–903.

Rodman, Emma. 2020. "A timely intervention: Tracking the changing meanings of political concepts with word vectors." *Political Analysis* 28(1):87–111.

Scheve, Kenneth and David Stasavage. 2016. *Taxing the rich: A history of fiscal fairness in the United States and Europe*. Princeton University Press.

Schneider, Eric B. 2020. "Collider bias in economic history research." *Explorations in Economic History* 78:101356.
URL: *http://www.sciencedirect.com/science/article/pii/S0014498320300516*

Sellars, Emily A. 2020. "Archival Silences and Historical Political Economy.".
URL: *https://broadstreet.blog/2020/10/21/archival-silences-and-historical-political-economy/*

Shoemaker, Bob. 2019. "Why Naomi Wolf Misinterpreted Evidence From The Old Bailey Online.".
URL: *http://www.historymatters.group.shef.ac.uk/naomi-wolf-misinterpreted-evidence-bailey-online*

Signor, Philip W., III and Jere H. Lipps. 1982. Sampling bias, gradual extinction patterns and catastrophes in the fossil record. In *Geological Implications of Impacts of Large Asteroids and Comets on the Earth*, ed. Leon Silver and Peter Schultz. Geological Society of America.

Simon Tanner, Trevor Muñoz, Pich Hemy Ros. 2009. "Measuring Mass Text Digitization Quality and Usefulness: Lessons Learned from Assessing the OCR Accuracy of the British Library's 19th Century Online Newspaper Archive." *D-Lib Magazine* 15.
URL: *http://www.dlib.org/dlib/july09/munoz/07munoz.html*

Slapin, Jonathan B. and Sven-Oliver Proksch. 2014. *Words as Data: Content Analysis in Legislative Studies*. The Oxford Handbook of Legislative Studies: Oxford University Press.

Slez, Adam. 2020. *The Making of the Populist Movement State, Market, and Party on the Western Frontier*. Oxford University Press.

Smith, Daniel M. 2018. *Dynasties and Democracy The Inherited Incumbency Advantage in Japan*. Stanford University Press.

Spirling, Arthur. 2016. "Democratization and Linguistic Complexity: The Effect of Franchise Extension on Parliamentary Discourse, 1832–1915." *The Journal of Politics* 78(1):120–136.

Spirling, Arthur and Andy Eggers. 2016. "The shadow cabinet in Westminster systems: modeling opposition agenda setting in the House of Commons, 1832–1915." *British Journal of Political Science* 48(2).

Spirling, Arthur and Iain McLean. 2007. "UK OC OK? Interpreting optimal classification scores for the UK House of Commons." *Political Analysis* pp. 85–96.

Springmann, Uwe and Anke Lüdeling. 2017. "OCR of historical printings with an application to building diachronic corpora: A case study using the RIDGES herbal corpus." *Digital Humanities* 11.
URL: *http://www.digitalhumanities.org/dhq/vol/11/2/000288/000288.html*

Suryanarayan, Pavithra and Steven White. 2020. "Slavery, reconstruction, and bureaucratic capacity in the American south." Working Paper.

Tanner, S. G. 2006. *A Guide to Capturing Text from Historical Documents: Report commissioned by the Oxford University Digital Library.* King's College London.

Thomas, David, Simon Fowler Valerie Johnson. 2017. *The Silence of the Archive.* ALA Neal-Schuman.

Underwood, Ted. 2014. "Theorizing Research Practices We Forgot to Theorize Twenty Years Ago." *Representations* 127(1):64–72.

Unknown. 15th century manuscript. "The boke of curtesy,' beginning, 'Litylle chyldrynne here may y e lere." *Egerton MS 1995* f. 58.
URL: *https://www.bl.uk/collection-items/the-lytille-childrenes-lytil-bok*

Vance, Colin and Nolan Ritter. 2014. "Is peace a missing value or a zero? On selection models in political science." *Journal of Peace Research* 51(4):528–540.

Ward, Richard. 2021. "Sentencing.".
URL: *https://www.digitalpanopticon.org/Sentencing*

Warwick, Claire, Melissa Terras and Julianne Nyhan. 2018. *Digital Humanities in Practice.* Cambridge University Press.

Webb, Walter Prescott. 1959. "History as High Adventure." *The American Historical Review* 64(2):265–281.
URL: *http://www.jstor.org/stable/1845443*

Wilkerson, John and Andreu Casas. 2017. "Large-Scale Computerized Text Analysis in Political Science: Opportunities and Challenges." *Annual Review of Political Science* 20(1):529–544.