Chapter 2: Social Science Research and Text Analysis*

Justin Grimmer† Margaret E. Roberts ‡ Brandon M. Stewart §

April 1, 2019

Texts are increasingly used to make social science inference, because language is the medium of social interaction. While the availability of texts is a huge boon to social science, the sheer volume of text presents a substantial challenge. Thankfully an explosion of new technology has provided social scientists with the ability to utilize much larger data sets. Yet, many of these techniques are developed in other fields or are often developed with other purposes in mind.

In this chapter we provide a framework and a general set of principles for utilizing text as data methods in social science research. Our framework applies generally across ways scholars might utilize text in their research. As we will explain throughout the book, it covers methods that are traditionally known as "machine learning" methods, models that come from statistical approaches, and even qualitative approaches to texts that rely on human coders. Our framework remains general across these different context because it focuses on the goals social scientists pursue when using texts, rather than focusing on the tools that are deployed to accomplish those goals.

We explain how text as data methods can contribute to inference at three stages of the research process: discovery, measurement, and causal inference or prediction. At the *discovery* — or hypothesis generation — phase, text methods can help illuminate new concepts

*Incomplete and preliminary draft from forthcoming book manuscript. Please do not cite or distribute.

†Associate Professor, Department of Political Science, University of Chicago
‡Assistant Professor, Department of Political Science, University of California at San Diego
§Assistant Professor, Department of Sociology, Princeton University

and suggest insights that deserve further investigation. Text as data methods provide researchers with new ways to organize their texts — including identifying clusters in the data, an underlying spectrum, or words that characterize a particular group of people. A new organization can prompt social scientists to read the texts differently and draw connections that they otherwise would have missed.

With concepts in hand — either from a discovery stage, from a theoretical model, or other intuition — scholars often want to *measure* or describe the prevalence of particular concepts in their data or to characterize where individuals and texts align on a spectrum. For example, we might be interested in learning the amount of legislation that falls within a broad range of policy agendas and the comparative manifesto project measures the prevalence of topics in party manifestos across the world. Numerous other papers, books, and dissertations read documents and assign them to categories by hand. The prevalence of text data and the preponderance of methods for approaching measurement has led to an explosion of interest in measuring quantities from text in increasingly diverse ways and from collections of increasing size. Novel and larger sources of text mean that the measures are often granular, providing insights into behavior otherwise difficult to detect. Measurement is the core ingredient of description and is useful as a goal in itself. Description can provide valuable summaries of the data that may inform theories, provide the measures necessary for causal inferences, and characterize the state of the world. To accomplish these goals, researchers have to demonstrate that their method of measurement does indeed describe the concept or behavior they would like to measure –that is, they have to validate.

Once a concept is discovered and the measures are validated, researchers are able to use those measures from text data to make a *causal inference* about the effect of some intervention, or make a *prediction* about what will happen in the future, given the past. For example, researchers might assess the causal effect of a president's speech on the salience of news coverage about a particular topic or they may be interested in how political content

affects users' engagement in online forums. Or, researchers may wonder the extent to which information in text can predict events in the world, like stock market movement or conflict. The methods for discovery, measurement, causal inference and prediction using text data vary not only because they are focused on different stages of the research process, but also vary in how stringently they rely upon specific modeling assumptions.

Discovery, measurement, causal inference and prediction are separate inferential goals. In spite of the differences in inferential goals at each stage of the research process, we advocate for five key principles when using the methods for social science research. Table 1 introduces our five principles, which we develop in depth below.

Table 1: Key Principles About Text as Data and Social Science Research

1) Social Science Inferences are Necessarily Sequential
2) Text as Data Methods Do Not Replace Humans — They Augment Them
3) The codebook function, g, is central: Text as data methods are about compression
4) There is no general theory of language, nor globally best method.
5) Validate,Validate,Validate

Our first principle is that social science inferences are necessarily sequential. We all learn about the world through an iterative process that involves exploration, measurement, and testing. By acknowledging and accepting the sequential approach to social science, we can design methods that explicitly account for and promote replicability and extension.

Our second principle is that text as data methods do not replace humans — they augment them. We will see that the methods we introduce in this book make humans more effective readers, lower the cost of analysis, and in many ways change the way we read. But applying text methods is still a fundamentally qualitative research activity — and as a result, there is no computer-only approach to text-as-data. Humans will need to read, interpret, and explain the output of the methods.

Our third principle is that the codebook function, which we call g, is central. Text

as data methods are about compression and regardless of the method we use, g codifies that compression. When applying text as data methods, the general goal is to reduce the complexity of the text to find some interesting features, rules for classification, or some new view of the data. Given this goal, our efforts center around developing procedures for compressing information and evaluating the content of the compression.

Our fourth principle is that there is no general theory of language and, as a result, no globally best method for analyzing text methods. Researchers have a range of different social scientific goals and text-based methods themselves are developed with different statistical objectives. The result is that there is no one method that will accomplish all tasks and there is no one method that strictly dominates all other methods as a text analytic tool. Different tasks require different techniques.

Our fifth and final principle is Validate, Validate, Validate. The methods we describe in this book all make strong assumptions about how texts are created and then deployed. For example, in Chapter 3 we describe how we discard word order to create text for analyzing quantitatively. In Chapter 4 we describe methods that make unrealistic assumptions about how texts are produced. And across the many methods that we present in this book (and methods that are left out) there are no theorems that show that the methods are guaranteed to capture the most important features of text. Rather, experience has shown that the methods tend to be useful in many settings; however, to assess the utility in any one setting requires extensive validation.

In this chapter we describe our approach to social science inference and our principles for applying text methods. To provide more context for each stage in the research process we first offer a description of one research program: an analysis of how King, Pan and Roberts (2013) discovered, measured, and then tested their theory of how the Chinese government censors social media.

1 Discovery, Measurement, and Causal Inference: How the Chinese Government Censors Posts

Ideas and research programs rarely emerge as a straightforward product, even though they are often reported that way in papers. While research papers often portray thoughtful questions and careful hypotheses as emerging after careful non-data driven contemplation and observation about the world, the reality is that those questions and hypotheses often emerge from an initial inspection of data. This inspection is usually guided by a substantial background knowledge and preexisting theories of how the world works which informs the discovery of a theoretical tension or empirical puzzle. This is perhaps clearest when working with textual data. Often times, research projects begin with a particular goal and then shift to explain an interesting aspect of the data. Other times, research projects discover some previously unknown feature of the data — and then the researchers change focus and seek to explain that new and interesting feature.

The non-traditional approach is evident in King, Pan and Roberts (2013), who develop a measure of censorship in China (KPR hereafter). KPR initially sought to use blog posts as a measure of public opinion in China — an important objective because of the difficulty of running surveys in an authoritarian environment. To this end, KPR downloaded millions of posts from over 1,400 social media websites. After making notes on a few potentially controversial posts and logging the URLs of the posts, the researchers went back to the original posts to better understand the post's context. When they returned to the posts, however, they noticed something surprising: many of the posts were now missing. Rather than the original social media posts, KPR now found pages that proclaimed that the content had been removed — an indication of government censorship.

KPR had accidentally stumbled upon a research design that would enable them to directly measure the rate of Chinese censorship — in doing so they had used the exploration of their

data to *discover their question of interest*. They had also discovered one conceptualization or way to organize the texts: they could view the social media posts as censored or not. The identification of this particular empirical phenomenon — the removal of documents from the web — derives its significance by the researchers embedding it into the broader social science context of censorship in authoritarian regimes. It should also be clear that after the fact it is obvious that some social media posts are censored in China. But before hand, it was far from obvious that collecting social media posts and then revisiting them could provide a valid design for studying Chinese censorship.

This initial exploration of their data led them to generate a hypothesis about why certain texts would be censored. At this point they were ultimately interested in answering a causal inference question, but did not yet have a viable potential intervention that could explain the censorship decision. To identify this potential intervention, a computer-assisted manual examination of the text gave them the impression that censorship rates tended to be much higher when the posts were about potential or existing collective action events in China — when groups of people would potentially come together, which could lead to a protest of the Chinese government. This suggested a new conceptualization that suggested a key quantity to measure: to what extent does a post make a reference to a collective action event?

Critically, this exploration stage gave them reason to believe that other conceptualizations would be less useful for answering the causal question. For example, sentiment, whether the post was critical or supportive of the government, did not seem to explain censorship decisions. Thus, using their texts, KPR were able to *discover* two more conceptualizations of the texts: the events the posts related to and whether the posts were critical of the government, neutral, or supportive of the government. If they were right, then the decision to censor a text would be unrelated to whether the post was critical about the government or not, but would focus on posts that described protest events.

Rather than a process detached from their texts, KPR used both a subset of their data

and statistical methods to discover their question of interest, generate a hypothesis, and to formulate the implied conceptualization. This is what we mean by *discovery* and we will describe a general process to facilitate discovery in Chapter 4. Given this information, they then set out to *measure* the texts according to their conceptualizations. To measure the topics of the posts, they used a supervised learning method, based on the keywords of the posts. KPR identified an initial set of keywords that they hypothesized would allow them to identify whether the posts were about a particular collective action event. They then iteratively refined that list to ensure the keywords captured only collective action events and were neither too broad nor too narrow. Measuring whether a post was censored was straightforward, based on their own records of the post. It was less straightforward, however, to measure the post's sentiment toward the government. As often happens with social science questions, their goal was not to characterize each individual document's sentiment, it was to characterize the proportion of document's within each category across censored and uncensored posts. For this task, they used a specific supervised learning algorithm — ReadMe, which we cover in Chapter 5 — to measure whether censored posts were supportive or critical of the government. Specifically, KPR sampled posts and hand labeled them as critical or supportive of the regime. They then used ReadMe (Hopkins and King, 2010) to extrapolate from the hand-labeled documents to the entire collection of posts and measure the proportion of documents that were supportive or not of the government.

Using an initial subset of posts, KPR were able to discover a theoretically interesting research question. They then moved to a larger data set in order to measure the key quantities of interest. With the measurement in hand, KPR estimated a social science causal effect of interest: the average effect of a post being about collective action on the probability that a post is censored. They also attempt to estimate the average effect of sentiment on the probability of censorship. While the measures necessary to answer those questions are now in hand, the measures alone are not sufficient for answering the causal inference questions. A

design is also necessary to eliminate confounding: other factors that are correlated with either collective action potential or sentiment and the outcome, but are distinct. For example, if collective action social media messages tend to be issued from people who are censored at a high rate we will confuse the message for the person.

To test the hypothesis, KPR utilize exogenous events that occur in China, which allows them to examine the government response. They find that posts around collective action events are censored at an extremely high rate — strong evidence that the government is censoring posts about events that might cause the public to come together. They also find that the government's decision to censor posts or not is essentially unrelated to whether the post supports the government. Their conclusions provide evidence that the government's censorship rules are not designed to merely suppress dissent against the regime. It also provides an example of how text can be used to make a causal inference. In Chapter 6 we describe a general approach for causal inferences with texts, including when text is the *intervention*.

In their first article, KPR make a compelling case based on observational data that censorship focuses on three things: collective action, pornography and criticism of censors. Their data was close to the real world but did not allow for the possibility of manipulating the subject of the post directly. The ability to randomly assign the intervention of interest (in this case the topic of the post) is a key component of the most rigorous causal inference designs. Having established their hypothesis in King, Pan and Roberts (2013), KPR produced a second study, King, Pan and Roberts (2014), that used a randomized field experiment to provide further evidence for their theory of censorship. They created accounts on one hundred social media sites across China and submitted text to these sites, randomly assigning the text to discuss protest events or events that were not related to protests and randomly varying whether the post was critical or supportive of the government. These experimental results produced the same conclusion as the observational results, providing

further verification that protest-related topics were causing censorship.

KPR's ground breaking insights into the Chinese government did not begin with a well-stated question of social science causal inference. But this does not mean that their research design is over-fit, atheoretical, or devoid of usefulness to social science. The initial exploration of the data was based on a subset of the data. This ensured that KPR would have the opportunity to demonstrate that they were wrong in new data where a different pattern might maintain. Theory also has an important role throughout KPR's design. They began with baseline theoretical accounts of how the Chinese government censors posts and using that theory lead them to pose new questions from their data. Their initial work inspired a second study which sought to reinforce claims made through observational data with data collected in a more controlled laboratory-like environment.

While not explicitly designed to test a formal model, the results of KPR's study has also inspired new formal theoretic models that seek to explain and contextualize KPR's findings (Chen, Pan and Xu, 2016; Lorentzen, 2014). In this sense, we see the specific instantiation of our more general point that science is sequential and collaborative.

Using KPR's experience as a reference, in this chapter we describe the research process when using text as data. While we focus on decisions that must be made when working with text as data, our argument is more general and illuminates one model of how empirical social science can proceed. It is also worth noting that scholars must enter into the research process at various stages. In the conclusion to this chapter, we explain how our framework can be used if scholars already have well defined questions or have key measures of quantities of interest in hand.

2 Social Science Inferences Are Necessarily Sequential

The goal of social science is to build robust theoretical explanations for social phenomena. But theory building is rarely simply about accumulating evidence from theoretical deductions and then revising the theory accordingly. This perception is a product of a time when acquiring data, running surveys, conducting experiments, and interviewing subjects was expensive and thinking was comparatively cheap. The cost of thinking has stayed relatively constant over time, but the cost of data collection has plummeted. The result is that it is practical for social scientists — even those on the smallest of budgets — to include new data acquisition as an explicit part of the research process. This means that initial samples of data can be used as an exploration to refine the research question — and to potentially stumble upon new and useful questions. New ways to organize the world suggest new measurements and eventually questions. This leads us to pursue new measures and research designs, which we then use to estimate causal inferences. Or, we use the new measures to make predictions about what will happen in the future.

Rather than the sequential approach to social science being atheoretical, each stage of the research process is both informed by theory and can help us build new theoretical explanations and then test those explanations. For example, developing a conceptualization requires knowing how prior theoretical work organized the world and the extent to which an organization is actually new. Measures are interesting in so far as they correspond to some theoretical quantity of interest (or several quantities of interest under different theories). The estimated causal effects can help us to build new theories and to test existing theories. Causal effects might help us to better understand and determine the extent to which observable implications are found in a data set and which observable implications are not present.[1]

[1]Importantly, the theory building can be with formal models, which are an integral feature of social science but one that is largely outside the scope of this book. Formal models can be used within the research

Learning from data, updating theories, and then testing those revised theories is a stronger move away from how social science is typically *taught* than it is how social science is typically *practiced*. We see sequential social science as a more honest account of how social science research tends to be done. The increased transparency makes clear the value of different methods at particular phases of the research process and enables us to design new methods to support individual research tasks. For example, we describe discovery methods in Chapter 4 of this book, which tend to get little attention in the social science. We think this is due, in part, to the tendency to pretend that conceptualizations are already known before the real research begins.

As we will emphasize throughout the book, the sequential nature of research leads us to regularly recommend that analysts split their sample into a training set and test set. At nearly every stage of the research process — whether discovery, measurement, causal inference, or prediction — we encourage sample splitting. In discovery, this is essential to learn whether a particular organization of texts is only prominent in one subsample, or if the prevalence is found in other data sets. In measurement the use of a split sample is more traditional and ensures that our functions that are used to classify objects are not overfit and the fresh data ensures that we can accurately evaluate the performance of our classifier on new data. In Chapter 6 we show that the training/test split is essential for using machine learning methods in causal inference. The training stage enables us to tune our method, discovering conceptualizations that are likely to give rise to interventions that exert a causal effect on some outcome of interest. The test stage enables us to credibly evaluate the size of the causal effect, while also avoiding some of the common problems with fishing in causal inference. Sample splitting is perhaps most common in applications of prediction, where it

process in order to derive hypotheses and then to test. Formal models can also be built on the output of models, helping researchers to interpret the conclusions of research. But crucially, formal models can also help us to interpret the findings from a study. By thinking carefully about actors' incentives when responding to intervention, we can better insights into what an effect tells us about the behavior of individuals and how it reveals patterns of strategic interaction (Bueno de Mesquita, 2016).

plays a vital role in assessing accuracy.

As we will explain below, sample splitting also solves problems that other mechanisms have been developed to do; for example, the pre-analysis plan. The benefit of having a train/test split is that it enables the researcher to have an explicit discovery phase, which is often ruled out in pre-analysis plans — which implicitly assume that discovery is already complete (Humphreys, Sanchez de la Sierra and Van der Windt, 2013). Further, pre-analysis plans are often only effective when others are present to regulate their content and notice that deviations have occurred. Researchers often lack the time to check pre-analysis plans and current resources for pre-registering studies often makes it difficult to connect a final paper to an initial pre-analysis plan. Pre-registration sources exacerbate this problem by allowing researchers to amend their pre-analysis plan.

The recommendation to split samples is usually met with two different kinds of objections. The first objection is that a particular experiment might be very costly and the decrease in power to split the sample is not justified. As we explain in more detail in Chapter 6, this objection is intuitive, but it is a case where our intuition leads us astray. Especially in instances where stakes are high, interventions are expensive, and the consequences for public policy are clear, we want to have the most confidence in the effects we estimate. A split sample provides robust detection of overfitting, while also ensuring our data cleaning rules take into account the realities of a particular data set. The second objection is that some historical data may occur only once and therefore splitting a sample can happen only once. In many ways this is true, some events only occur once and therefore provide us with only one data set (and therefore one sample split). But as Fowler and Montagnes (2015) suggest, there are often analogous interventions that could be studied. For example, they examine the compelling finding in Healy, Malhotra and Mo (2010) that college football scores affect Congressional elections, finding no indication that National Football League (NFL) games affect the outcomes. This related study is useful, because if the logic in Healy, Malhotra and

Mo (2010) is correct, we would expect to see a similar effect in NFL games. For almost all historical interventions, there are analogous interventions in different times or in different places.

3 Text as Data Methods Do Not Replace Humans — They Augment Them

Text is already pervasive in the social sciences. For centuries scholars have analyzed books, laws, documents, and interview transcripts to learn about the world. Qualitative methodologists have developed methods to improve upon this reading, to ensure that the information in the texts is reported reliably, and to increase the transparency of the research. At the core of these methods is a researcher who not only carefully consumes the content in a book, but also adds analysis and insights to reach a conclusion. Consider, for example, McQueen (2018), who examines how apocalyptic movements contemporary to great political thinkers influences their political philosophy. To reach this conclusion, McQueen (2018) situates her intellectual history in a rich context, analyzing thousands of documents to better understand the context in which Machiavelli, Hobbes, and Morgenthau wrote.

It would be a mistake to suppose that text as data techniques could replace this long research tradition, eliminate the need for careful and close readings of texts, or otherwise obviate the need for analysis. Rather than replace reading or supplant the close reading of text, text as data methods *augment* our reading ability. Text as data methods help us read better, not avoid reading at all. This amplification of human effort improves the analysts ability to discover interesting organizations, measure key quantities of interest, estimate causal effects and make predictions.

Text as data methods are a way to address shortcomings of human researchers, while still preserving researchers' strengths. To better understand why text as data methods augment

rather than replace human readers, we use the "hay" analogy from Hopkins and King (2010). Suppose that our texts are like straws of hay and we see a farm field full of hay. One task that we might have is to analyze a particular straw of hay. We might want to understand its beauty, describe its particular features, or even characterize its color. We might approach text similarly. We might want to analyze a particular poem and describe its linguistic beauty, we might engage with a particular line of text to appreciate its deep meaning, or we could examine the words that are not written to better understand an author's more complete message (Strauss, 1952; Melzer, 2014). While humans are often very adept at this sort of deep reading, text as data methods are rarely designed to be helpful with this specific task. The methods we describe typically fail to identify features of very small and isolated documents.

Now consider a separate task. Suppose that we want to identify features of the straws of hay and sort them into piles of similar hay straws. Humans, as it turns out, are ill-equipped for this sort of task. This is because we have tiny active memories, we tend to get distracted when performing the same task for awhile, and we might make errors because we fail to interpret the instructions correctly. In contrast, algorithms work incredibly well at sorting, provided there is a well-defined objective function or human guidance that a computer replicates. When applied to text, this implies that computers work well when discovering new organizations and compressing texts into lower-dimensions.

Text as data methods, then, are best suited to accomplish tasks that humans find difficult. They are complements, not competitors, to close reading of texts. Text as data methods can guide and assist close reading of texts, conditional on computer discovered organizations, measurements, or causal inferences. For example, Blaydes, Grimmer and McQueen (2018) analyzes a large collection of governmental advice texts, written in Muslim and Christian countries from 1000-1600 CE. To analyze the large of collection of texts, Blaydes, Grimmer and McQueen (2018) introduce a method that discovers coarse and granular text topics and

applies them to the books. Using the organization of texts as a guide, they characterize the divergent evolution of themes in the Christian and Muslim texts. To establish this divergence, the authors rely on both the organization from the method and their own close reading of the texts.

Text as data methods make some activities cheaper. This is particularly true for classification methods. A large class of methods develops rules to assign documents into predetermined set of categories. These methods are usually used as a subsidy to the researcher. Rather than engage in expensive and time consuming hand coding, supervised methods cut down the amount of human intervention that is necessary in order to do the classification. Other classification methods eliminate human coding altogether and instead use proxies to classify documents. Supervised scaling methods, such as WordScores (Laver, Benoit and Garry, 2003), similarly embeds documents into a low-dimensional space with researcher determined anchors for the space.

Machine learning algorithms applied to text clearly have a speed advantage. But it would be a mistake to only use text as data methods when you have a large collection of text because even small text collections can lead to high-dimensional problems. To see why, suppose we are interested in creating a set of categories and then classifying one-hundred open-ended survey responses. This is a relatively small data set. Data scientist Hanna Wallach calls corpora this size "artisanal data," a reference to a data set's smaller size and the careful curation of text collections. Suppose that we have a relatively simple goal: to identify the most interesting partition — an organization that assigns every text to one and only category — of the documents. This relatively simple-sounding goal requires a search over a massive domain. We can, characterize the number of potential partitions from a collection with the *Bell Number*. For example, if we have two documents $\{A, B\}$ then the Bell number is 2 — AB assigned to the same partition and A, B assigned to different partitions. With three documents the Bell number is 5 — $ABC; A, BC; AB, C; AC, B; A, B, C$. With 100

documents the Bell number is $10^{115.68}$ — an incomprehensibly large number. By comparison, scientists estimate that there are approximately 10^{80} atoms in the known universe. Clearly, then, computational methods can help us explore a massive space.

Taken together, text as data methods work directly with qualitative research methods. In fact, text as data methods are used to accomplish a fundamentally qualitative task: to extract meaning and analysis from collections of text. The methods and research designs that we introduce in this book bridge the quantitative and qualitative divide. Text methods work well because their statistical and algorithm foundations enable them to complement human reading of documents. And then conditional on this organization or a set of measurements, we can read our text and glean insights we otherwise might now have encountered.

Text as data methods, then, provide ways to catalyze discovery, make measurement more efficient, or facilitate estimating causal effects. But to do this the methods neither replace human readers, nor supplant qualitative methods. Text as data methods help connect quantitative and qualitative research traditions and help scholars improve their ability to use text to make inferences — a fundamentally qualitative task.

4 The Codebook Function, g, is Central: Text as Data Methods are about Compression

Text is inherently high-dimensional. Consider one of the greatest speeches in American political history. On the eve of his assassination, in the context of a large number of death threats, Martin Luther King, Jr. delivered a speech entitled "I've Been to the Mountaintop" in Memphis, Tennessee. In the rousing speech where King confronts the threats on his life directly, King closes with a prophetic declaration:

> Like anybody, I would like to live a long life. Longevity has its place. But I'm

> not concerned about that now. I just want to do God's will. And He's allowed me to go up to the mountain. And I've looked over. And I've seen the Promised Land. I may not get there with you. But I want you to know tonight, that we, as a people, will get to the promised land!

In many ways, "I've Been to the Mountaintop" is high-dimensional, like other pieces of text. It is high-dimensional in the obvious way — language depends on the order that words are written. In this sense, the speech is unique, only taking on the particular meaning because of the order of the words. This requires the exact sequence of words to convey the exact same meaning. Beyond the order dependence of words, however, interpretation of the speech depends on context, adding further dimensionality. Part of the speech's power comes from when it was delivered and who delivered it: on the eve of his assassination, an iconic civil rights leader discusses his vision of the future. The speech is also powerful because it makes a vivid allusion to biblical stories, in particular the book of Exodus. King is implicitly comparing himself to Moses, the black audience members to the enslaved people escaping Egypt, and the quest for equal rights and justice to the escaped slaves finding a land of their own.

The goal of text as data methods is to develop a *codebook* function that reduces the high-dimensional information in the text to a much lower-dimensional representation that is then used for social science research. We will often refer to this function as g. There are many reasons to reduce information and compress text. First, reducing information can make the text interpretable in a way that would be otherwise difficult. The reduction in dimensionality is a major reason that text as data methods are useful for discovery. Reducing the complexity of text in discovery often provides an organization of the documents that then informs the way we read those documents. It can also cause us to rethink how we view a particular phenomenon, leading to new questions, ideas, organizations, and research projects. When performing a measurement, the reduction of information, enables us to understand what we

are measuring and why it might be relevant as a quantity of interest. For example, reducing political text and roll call voting decisions to a single dimension can be incredibly useful for summarizing political decisions. The hope when doing reduction for measurement is that we preserve the useful and interesting facets of the texts.

A second reason for reducing the dimensionality of text is for statistical properties. When we are attempting to use texts to make predictions, infer how texts affected an individual, or trying to infer how texts are similar, we are unable to work with the entire text because there is simply too much information. This curse of dimensionality can manifest in several statistical applications of text as data methods. Even with the biggest data sets we are simply unable to learn about how all the complicated features and dimensions of a text apply to a particular problem. If we fail to reduce the dimensionality when attempting to make predictions we will overfit and our forecasts will perform poorly. If we do not reduce the dimensionality when attempting to infer the causal effects of a speech we will lack the basic information necessary to reliably estimate causal effects of interest. The only way to proceed, then, is to reduce complexity.

A third reason to focus on reducing the dimensionality of text is that most social science theories are about relatively low-dimensional and/or a small number of categories. For example, political economy theories of politics often suppose that conflict happens along a low-dimensional ideological spectrum. In order to test the observable implications from these theories, then, we need measures of individuals along the dimension. Or, we might be interested in describing the nature of campaign advertisements. This literature tends to create a dichotomy between negative and non-negative advertisements, with more refinement of both categories sometimes used.

Given its importance, we will focus our efforts in this book on understanding the particular g function we are estimating, what properties it might have, and how we know if we have done a good or bad job. The function will take on several forms. The most familiar

version for many social scientists will be using research assistants to hand classify documents into a set of categories, constituting a manual g function. We might use samples of manual codings to statistically learn a g function that places previously unseen documents and assigns them to known categories. Or, we might want a function that will take a collection of documents, partition them into a set of categories and then assign documents to categories. Finally, we might use a g function to discover the treatments that are present in a collection of documents.

As we explain below, the g function's overall goal is to retain the most important content of the documents, while discarding the features of the document that are irrelevant for the analysis at hand. This is why we refer to the g function as a distillation, or summary, of the text. We are not attempting to provide a comprehensive account of what is in the text and all facets of meaning. Rather, we want to retain the most relevant content. This objective is intentionally vague. Further refinement of the goal requires that we know more about what we are trying to accomplish and where we are in the research process. We discuss this in the next section.

5 There is No General Theory of Language Nor Globally Best Method For Text Analysis

While we have an overarching goal of distilling texts into the most useful content, we will not introduce a general theory of language to achieve this summary. Nor will we introduce a single method for reducing the information in the texts.

We do not introduce a general theory of language because we are unaware of one that would be useful for the ways social scientists apply text data in their research. Across the field of computational linguistics there is no comprehensive model of language that is generally applied to comprehend text. The lack of a model of language is due, in large part,

to the complexity of language in the world. As we discussed above, the meaning of language is found not just in the words — but also in the order they are spoken, the characteristic of the speaker, and the social context of the reader. This means that any attempt to comprehensively distill the work down to a general model of language generation, or even a general model for a specific language, will be fraught with difficulties. And will likely be impossible, or so difficult as to not be useful.

The demands on such a general theory for social science would be complicated because the features of interest in the language vary so much based on what we want to learn from the text. For example, the content that we would want to extract from a text if we are interested in learning the topic of discussion is qualitatively different than the content we would extract if we are interested in learning the sentiment from a text. We are interested in fundamentally different things when we attempt to learn the ideology of a speaker as opposed to trying to identify who wrote a text (a task from the field of stylometry). The types of quantities that social scientists hope to extract from texts are diverse and constantly growing.

In short, we do not have a comprehensive model of language both because it is difficult to develop and because the relevant features of text are task-dependent. The lack of a general theory also implies that there is no single best method for analyzing text. The multitude of demands that social scientists place on their methods ensure that there is no one right way to do text analysis. Discovering an organization of texts is fundamentally different than classifying documents into preexisting categories, which is different than inferring the causal effect of a text. In this book we focus on presenting different options and describing the tasks where we might expect they will perform best.

The methods that are applied to text as data also tend to lack the statistical theory found in many other areas of statistics. There are important and deep statistical theoretical properties for the estimators that we will discuss in this book which help us evaluate methods.

Yet, even the most theoretically well developed methods have few theorems that relate the performance of the method back to natural language *as it is spoken* or tie performance to particular social science tasks. The properties are proven conditioning on a (generally very simplified) representation of the text and are relevant to tasks that may not quite align with our uses in the social sciences.

Thus in general, our confidence in the methods that we apply to our own data is not based on their deep theoretical properties. Our confidence tends to be based on empirical evidence: the methods that get adopted tend to be adopted because they have performed well across a range of problems. As we explain below when discussing validation, that means that *every* application of text as data methods requires extensive evaluation, because we cannot be sure that the method will perform well with the specific task and the data we are using. Validation builds confidence with our readers and more importantly, allows us the opportunity to detect cases where a given method fails to perform well for our particular task.

Even though there is no deep theoretical model of language that will be useful for this text, there are still many surprising connections across models developed in completely different contexts for completely different purposes. In this book we will adopt a consistent notation that reflects this connection to hopefully help the reader see the commonalities among approaches. In particular, we adopt the following conventions:

[Summary of material to be added: As notation is finalized, we will give a more in-depth explanation including a breakdown of why notation is formulated as it is.]

- N : number of documents
- i : index of documents,
- M: number of tokens in a document
- m: index of tokens in document
- J : total number of features
- j : index of features
- K: total number of components

- k: index of components
- D_i : an abstract representation of document i including its text, formatting, or any additional information of interest.
- W : $N \times J$ document term matrix
- W_i : document i row in the document term matrix
- W_j: column j of document term matrix
- W_{ij}: entry i, j
- Z: word level assignment to features
- π : $N \times K$ matrix describing weights/loadings on latent variables
- π_{ik} : Document i's weight/loading on feature k
- μ: estimated center, factor, or other measure of tendency
- y_i: the target label/dimensions/category
- X_i: a vector of general covariates for document i of dimension P
- $\hat{y}_i$: same support as y
- β: Regression coefficients from subsequent analysis
- λ: regularization parameter or rate parameter depending on context.
- σ^2: variance
- $p(\cdot)$: a generic unspecified probability distribution.
- ϕ: all other nuisance parameters
- I: index for documents in the training set, or in sample
- O: index for documents in the test set, or out of sample

By returning to this list as new models are introduced, readers will be able to see the similarities in structure across many of the models we will apply at each stage of the research process.

6 Validate, Validate, Validate

There is no universal model of language nor is there a task-independent optimal method for text analysis. And as we explained above, a consequence of this is that there are few (if any) theorems that justify text analysis methods as applied to natural language. Perhaps unsurprisingly, then, every application of text analysis methods requires that we validate the properties of our measures that our compression, g, produces. Validations are necessary because the primary justification for using text as data methods is empirical: we use methods that have worked well in the past when applied to similar problems. Of course, we may

suspect that a data set we are working with is strikingly similar to prior data sets or (we suspect) fits within the general features of other text collections where the method has performed well in the past. This could be true, but it is hard for the researcher to fully understand all the ways a text collection might be different and to understand how those differences could affect the performance of a method.

Given that most justifications for methods in the text as data literature are empirical, validation is critical to establish that a method performs well in a particular data set. But a method performing the literal task that we set out is neither necessary nor sufficient for a method to accomplish our ultimate goal: making a valid social science inference. A method might perform well on the specific task, but when we aggregate up the low-level measures to a unit of interest, bias could accumulate resulting in an inaccurate measure of some aggregate, unit-level behavior. And, even if our method is able to produce accurate unit-level measures, we still have to validate that we are measuring our concept of interest.

The first validation assesses whether our method performs well the specific task it is designed to perform. This sort of validation is most obvious with supervised classification methods, where our task is to replicate human hand coding decisions. Similarly when we are using texts to make predictions, our task is to accurately forecast some variable in the future. To assess the performance on both tasks we use the same basic insight: we can replicate the task on data where we know the answer. In Chapter 5 we describe how validation data sets and cross validation can assess how well our method classifies documents into categories. In Chapter 7 we discuss the conditions under which held out data can help us assess the accuracy of our forecasts.

Validating that other text as data methods perform the specific task at hand is less straightforward, but a variety of tools have also been developed for unsupervised models that we use for discovery (Chapter 4), measurement (Chapter 5), and to facilitate causal inferences (Chapter 6). Some validations are straightforward and involve the researcher

manually reading texts assigned to a particular category and deciding if the organizations from g are coherent. Other validations involve examining the features that are indicative of a category or latent dimension. Using these underlying features the goal is to evaluate if these indicative features are coherent and correspond to some identifiable quantity of interest for the researcher. Another approach is to use quantitative measures of model fit that assess how well a particular text as data method fits the data. While seemingly an attractive option, previous work has shown that quantitative methods tend to do a poor job of measuring the usefulness of models for social science work. In response, there are a series of validation methods that place "humans in the loop," explicitly accounting for and including human judgment in a partially automated process. To do this, these validation techniques exploit two key insights. First, for many tasks we want to assess models based on how humans use the information the model provides. Second, to include human information we need to carefully design experiments to ensure that researchers avoid arbitrary criteria when choosing the model. With these insights in mind, we describe how to use experimental evidence to assess the quality of topics and clusters from a wide array of text as data methods.

Rarely are social scientific tasks — like discovery, measurement, causal inference, or prediction — solely reliant upon a method performing a specific task well. Usually, we are interested in some aggregation of the measures of documents (Hopkins and King, 2010). For example, Catalinac (2016) uses output from a topic model to measure how Japanese political candidates move away from particularistic concerns of their district towards international issues. In order to validate this step, we need to show that combining the individual level decisions avoids introducing systematic bias that makes our measures less than useful. Further, it is often also useful to establish that the measures have *face validity*. That is, we should expect that, at a minimum, our measures pass the inspection of a subject expert. Moving beyond face validity is difficult. But it is possible to more formally demonstrate that a data set adheres to an expected set of patterns. This *hypothesis* validity provides a formal

mechanism for assessing the quality of our measures. Of course, there might be a blurred line between a measure failing a hypothesis validation and a research finding. There are two approaches to combating this. First, it is essential to select hypotheses that are obvious and difficult to explain non-findings. For example, Grimmer (2013) shows that committee chairpersons are more focused on issues under their committee's jurisdiction. Second, several studies have suggested pre-registering hypothesis validations. If pre-registering ensures that the results are presented, regardless of the outcome of the validation.

A final validation provides an assessment that the label we have attached to our compression of the text corresponds to what the measure is actually capturing from the text. This assessment of label *fidelity* is essential to establishing the viability of a particular text compression for a research task. When hand coding or using a supervised method, the key to establishing fidelity is providing transparency about how categories are defined in the codebook and demonstrating that the coders adhered to the coding rules. Demonstrations of fidelity for unsupervised methods can be more challenging. Some techniques are straightforward — providing sample text to the reader, for example. And it is always an option to validate the discovered categories using hand coding and supervised methods. That is, we can take the measures that were originally discovered using a method without predetermined categories and then demonstrate that we can recover those categories with a researcher defined codebook.

A proximate, though distinct, task from validation is model selection — that is choosing a particular g to use for our social scientific tasks. Just like there is no universal tool for text analysis and no application independent method for validation, there is no general approach for model selection. Rather, we will argue that how we select a model will depend upon the particular task at hand. If we are attempting to *discover* some underlying organization or conceptualization then the primary evaluation will be what the g function suggests and whether this provides us with useful insights. Though, we will also want to evaluate the g

function to ensure that we are correctly interpreting the function. When used to *measure* according to some conceptualization, we can ask about the g function's bias and its precision. When used to make a *causal inference* we will also look for evidence that the g function conforms with assumptions that are necessary to avoid bias in our estimates. And when used to make predictions we will look for a g that helps us to make our best prediction of the future.

7 Visual Overview

[Summary of material to be added: This section will ultimately contain a visualization of the core principles including pointers to different parts of the book where they are explored. The goal is to show how the principles weave through the rest of the text.]

8 The Role of Intuition

Having advanced our five principles and previewed our more specific guidance for each research task, the reader may be left wondering why we are silent on other issues. For example, we say little about a preference for model or algorithmic approaches, the best way to represent texts for quantitative analysis, particular tools that we think are best suited for research tasks, or the optimal algorithms to use to obtain valid social science inference. We are intentionally silent on these issues for two reasons. First, the differences are often overstated. Throughout the book we will show that there are surprising connections and close relationships across different methods. And these similarities often undermine the major highlighted differences in the literature. Second, many of the answers to these questions do not lend themselves to principles that always apply. Rather, they are contextual and depend upon the setting. Evaluations depend on the task at hand. The best model depends upon the

task. And how one should set up their workflow to achieve the best work depends on the analysts' own proclivities as much as it depends on advice we might offer.

We are also silent on whether text as data, and machine learning methods in general, should be intuitive. Our silence here is for a different reason. The issue of intuition in machine learning methods is a profound and deep issue. Resolving the proper role of intuition in models — and their application to solve problems — requires resolving long-standing philosophical issues of fairness and justice. Obviously, we will be unable to tackle that here.

Instead, we offer a simple and, we think, relatively uncontroversial, insight into the intuition debate. While analysts may optimize for intuition in a variety of ways, we think analysts should work to offer some insight into why the algorithm works and how it makes the decisions that it does. Even in the least intuitive and most complicated models, analysts can help the reader (or person subjected to the algorithm) to understand why and with what information the decision was made. Or at the very least, understand better than merely a prediction. This is particularly necessary when scholars are optimizing methods to provide intuition. In those instances it is essential that the intuition is conveyed to the reader so they can understand the inner workings of the method.

9 Conclusion: Text as Data and Social Science

[Summary of material to be added: A brief conclusion previewing the rest of the structure of the book and articulating the way that social science repurposes computer science and the way computer science repurposes social science.]

References

Blaydes, Lisa, Justin Grimmer and Alison McQueen. 2018. "Mirrors for princes and sultans: advice on the art of governance in the medieval Christian and Islamic worlds." *Journal of Politics* (4). Forthcoming.

Bueno de Mesquita, Ethan. 2016. *Political Economy for Public Policy.* Princeton University Press.

Catalinac, Amy. 2016. "From Pork to Policy: The Rise of Programmatic Campaigning in Japanese Elections." *The Journal of Politics* 78(1):1–18.

Chen, Jidong, Jennifer Pan and Yiqing Xu. 2016. "Sources of authoritarian responsiveness: A field experiment in China." *American Journal of Political Science* 60(2):383–400.

Fowler, Anthony and B Pablo Montagnes. 2015. "College football, elections, and false-positive results in observational research." *Proceedings of the National Academy of Sciences* 112(45):13800–13804.

Grimmer, Justin. 2013. *Representational Style in Congress: What Legislators Say and Why It Matters*. Cambridge University Press.

Healy, Andrew J, Neil Malhotra and Cecilia Hyunjung Mo. 2010. "Irrelevant events affect voters' evaluations of government performance." *Proceedings of the National Academy of Sciences* 107(29):12804–12809.

Hopkins, Daniel J and Gary King. 2010. "A method of automated nonparametric content analysis for social science." *American Journal of Political Science* 54(1):229–247.

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.

King, Gary, Jennifer Pan and Margaret E. Roberts. 2013. "How Censorship in China Allows Government Criticism but Silences Collective Expression." *American Political Science Review* 107:1–18. http://j.mp/LdVXqN.

King, Gary, Jennifer Pan and Margaret E Roberts. 2014. "Reverse-engineering censorship in China: Randomized experimentation and participant observation." *Science* 345(6199):1251722–1251722.

Laver, Michael, Kenneth Benoit and John Garry. 2003. "Extracting Policy Positions from Political Texts Using Words as Data." *American Political Science Review* 97(2):311–331.

Lorentzen, Peter. 2014. "China's Strategic Censorship." *American Journal of Political Science* 58(2):402–414.

McQueen, Alison. 2018. *Political Realism in Apocalyptic Times*. Cambridge University Press.

Melzer, Arthur M. 2014. *Philosophy Between the Lines*. University of Chicago Press.

Strauss, Leo. 1952. *Persecution and the Art of Writing*. Free Press.