Chapter 5: Measurement[*]

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

March 19, 2018

1 Introduction

For social scientists, a compelling part of the text as data toolkit is the ability to systematically measure concepts that are not easily quantifiable. This opportunity for measurement has allowed large bodies of text to contribute to the rapidly expanding evidence base of social science in the digital age (King, Schlozman and Nie, 2009; Lazer et al., 2009; Salganik, 2017). Stylistically, these measures sit somewhere between quantitative and qualitative tradition. They provide us with quantifiable summaries of empirical phenomena, but they require interpretation and considerable care in application.

The goal in measurement is to instantiate some concept within our hypothesis or theory in order to facilitate quantification.[1] Thus from our point of view, no measurement can be completely divorced from an account of the nature of the thing being observed. To measure something is to invoke a series of testable and untestable assumptions about how a observable

[*] Preliminary and incomplete draft, please do not circulate.

[†] 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

[1] A natural baseline is provided by Stevens (1946) which defines measurement as the assignment of numeric values to objects according to some set of rules. Duncan (1984) memorably describes the Stevens definition as incomplete in the way that 'playing the piano is striking the keys of the instrument according to some pattern' (Duncan, 1984, p.126). We follow him in expanding the definition to include the goal of quantification. We do however take a slightly broader definition in including the idea of classification.

data is translated to a theoretical concept. Measurement is the vital link that connects our broader conceptual argument to the data we have available as evidence.

Consistent with our key principles outlined in Chapter 2, there is no single approach to measurement which is going to perform best across all settings. However, we think of most measurement as involving the following steps for the researcher,

1. *Define* the conceptualization we want to measure.
2. *Locate* a source of data which contains implications of the identified concept.
3. *Generate* a way to translate data into a latent representation. We call this the g function.
4. *Label* the representation and connect to the identified concept.
5. *Validate* the resulting measure.

In Chapter 4, we described methods for discovery which helped us define the conceptualization we want to measure by sparking a creative idea for new research questions or theory. In Chapter 3, we covered how to locate sources of text data. In this chapter, we turn to steps 3–5 and the development of measures that will test the observable implications of that theory. Regardless of whether our goal is description, causal inference or prediction, we will need some means of systematically measuring phenomena from a complex world.

Measurement is fundamentally about compression — in measuring something we are choosing to throw away the majority of the information about a *specific* situation, in order to focus on a *generalizable* property. This process of compression is formalized with the codebook function, g, which maps between the text and our latent representation that we will use for measurement. This chapter will cover both the methods we use to arrive at a g and the way in which we integrate the latent representation into our broader research process.

In Section 2 we discuss what is involved in moving from a concept to a measurement of that concept. We dispel some common misconceptions, describe the desirable properties of a measurement and detail different types of compression. We then turn to different approaches to learning g: repurposing of unsupervised methods (Section 3), extrapolating from human coding (Section 4), supervised learning without training samples (Section 5), and additional methods beyond classifying or scaling subject matter (Section 6). Section 7 details how to choose a measurement technique and integrate it into a broader research design. We conclude in Section 8 with advice on how to validate the measurements for a particular research project.

This chapter focuses on measuring some property of the text itself. A separate, but related, task is the use of documents to predict some completely separate measurement. This is a process called "nowcasting" (Choi and Varian, 2012) — predicting the present state of the world in a way that accords with another measurement that is, or will be, collected. This includes using psychological language on Twitter to measure heart disease mortality (Eichstaedt et al., 2015), using search terms in Google to predict flu rates reported by the CDC (Ginsberg et al., 2009; Lazer et al., 2014) and using Twitter to predict opinion polls (O'Connor et al., 2010; Beauchamp, 2013). This allows for reporting of an existing measure in a way that is more timely or fine-grained. Many of these approaches use some of the same machine learning techniques we outline below, but for a fundamentally distinct social science task. We take up these approaches in Chapter 7 on prediction.

2 From Concept to Measurement

In Chapter 4, we discussed the task of *discovery* and how we discover our question of interest. There our goal was to develop a good research question and spark the concepts that would form the basis of our measurement strategy. In this chapter we take up the task of quantifying the concept in our actual documents. In practice, these two tasks are not always cleanly separated, often because the discovery portion is not talked about at all. By separating out these two distinct stages, we can approach each with appropriate methods and design new methods that are optimized for the particular task. Even though we see discovery and measurement as separate, we imagine that in most broad research programs there will be iteration back and forth between these two stages as the research question evolves.

There is no universally best method for measuring a concept and so we cannot offer a 'one size fits all' approach. Instead, we describe a wide variety of approaches, each of which leverages information from the analyst in a different way.

2.1 Three Misconceptions

Before delving further into how to approach measurement with text, we want to address three common misconceptions about measurement using quantitative text data. We have found these misconceptions to be sticky because they are at least partially true. Yet, they have pernicious effects which are worth dispelling before proceeding further.

Each of these misconceptions have a common theme that issues around quantitative text analysis are somehow distinct or unusual within social science. However, essentially everything we raise in this chapter is not unique to text: it is common to measurement in social science writ large. While text analysis doesn't raise fundamentally new concerns, it does throw old ones into relief.

Misconception 1: Quantitative text analysis automates measurement. As we pointed out in Chapter 2, text as data methods do not replace humans, they augment them. The same is true for their application in measurement. What text analysis methods do offer — and what gives this misconception the longevity of a partial truth — is the possibility to decrease the marginal human cost of producing a measurement on a new document to almost zero. However, there is still a requirement of a substantial investment of human time; that is, there is no fully automatic procedure for measurement.

While the methods described below often radically reduce the need for human effort in some part of the measurement process, there is generally a tradeoff where we require increased human attention to some other area. For example, a (non-expert) human may no longer need to personally read each document, but we now require someone with more programming knowledge and an expert human to validate results. Often these trades of one type of human effort for another are well worth it, but we shouldn't confuse this with complete automation. There is no free lunch. If any method was completely automatic, this chapter would be a lot shorter.

Misconception 2: Unsupervised methods require fewer assumptions than supervised methods. The quantitative text analysis literature is full of binary heuristics for organizing models: generative vs. discriminative, statistical vs. algorithmic, and most prominently, supervised vs. unsupervised. These binary heuristics are useful but are often taken too seriously. In textbooks and articles, unsupervised methods are described as a human provides only a document feature matrix and the algorithm produces a series of insights which are then labeled entirely after the fact. Supervised methods are presented as measurement black boxes which merely replicate human effort with nothing left to discover. There are some elements of truth in these simple sketches, but there is considerably more to most methods than this simplistic view and over-reliance on the supervised/unsupervised divide can cloud our thinking. The important distinction is *what information the model uses* to learn about the data.

A consequence of this caricature of supervised and unsupervised models is the persistent idea that unsupervised models have fewer assumptions. Unsupervised methods have *different* assumptions, but they are not in any meaningful sense fewer. We should always be cautious about methods that purport to require nothing from the user because the *information has to come from somewhere.* If it isn't coming from labeled data, it is usually coming from strong assumptions about how the data was generated. These assumptions don't need to be exactly right for the method to be *useful*, but we will take special care as we discuss the methods below to point out what information the method draws on to produce g.

What then does it mean for a method to be useful? Unsupervised methods are often presented as tools for discovery and supervised methods are presented as tools for measurement. This is too narrow. The tools themselves are separate from the task we put them to: essentially all methods can be used for either discovery or measurement. In this chapter, we focus on what it means to look at any of these approaches as a tool for measurement but in the process we will introduce specific techniques that might also be useful for other tasks.

We aren't going to try to overturn decades of use of the terms supervised and unsupervised, but we will delineate methods based on what information the model has available to learn from the data. We have covered unsupervised methods which use information about the data generating process or a distance metric and we will cover classical supervised methods that extrapolate from known gold-standard coding. We will also cover a lot of the space between, such as models that make use of labeled data and a generative model. When we speak of unsupervised methods we generally mean those primarily driven by a data generating process or measure of distance and by supervised methods we generally mean those that leverage gold-standard data. But, it is important to not get too hung up on the labels.

Misconception 3: There is a task-agnostic degree of compression In the presentation of clustering and topic models in the chapter on discovery, we discussed setting the degree of compression in the model in terms of choosing number of components, K. This choice of dimensionality is often made manually in clustering and mixed membership models with practitioners seeking guidance about the way to choose the "correct" value. The choice of K is determined by a combination of the data set, the task and the model; any two are insufficient.

All measurement is about compression: what information we keep and what we can discard. Thus any time we measure something we make a choice about complexity. There is a choice of K when we code documents into categories, but there is also a choice of dimensionality when a quantitative scholar uses Gross Domestic Product (GDP) as a proxy for the economy. There is even an analog of K in qualitative description — the amount of detail. This should make it clear that there is no true underlying K that describes an event, only choices that serve the particular research question, because any measurement is a substantial reduction of reality. As King, Keohane and Verba (1994) write

> *the difference between the amount of complexity in the world and that in the thickest of descriptions is still vastly larger than the different between this thickest of descriptions and the most abstract quantitative or formal analysis.* No description, not matter how thick, and no explanation, no matter how many explanatory factors go into it, comes close to capturing the full "blooming and buzzing" reality of the world. There is no choice but to simplify. (43)

On the one hand, this means that choosing K is not some novel challenge that comes with the adoption of machine learning — it is an essential part of every tradition of social science. On the other hand, we should not expect to have some procedure that given only a dataset will choose K for us.

We contend that the choice of K requires a combination of a dataset, a task and the model. The task encompasses the argument we want to make and the goals we have for the representation. This is required to select K because different tasks require different complexities of representation. If someone hands you a set of documents and says 'What is K?', the question is fundamentally ill-posed. If someone gives you a procedure for choosing K which is data, task or model independent, it may be useful in some circumstances, but it cannot be universally optimal.

2.2 What Makes a Good Measurement

The role of a good measurement scheme is to achieve a simplification of real world phenomena that allow us to accurately describe a process or test the observable implications of our theory. Because no measurement can reflect the world in all its complexity, the properties of a good measurement is intrinsically linked to the particular problem that the analyst is trying to solve. As analysts we have to be clear about the ways in which our measures fail and the positive properties that they have.

While computer scientists have primarily focused their text analysis efforts on applications in information retrieval and prediction, social scientists have almost exclusively focused on texts as a mechanism for measurement. Texts have been used to measure things as diverse as culture (Bail, 2014), political agendas (Grimmer, 2010), support for violent jihaad (Nielsen, 2012) and international events (Schrodt, 2012). It is not surprising that social scientists have focused on applications in measurement as there is a long history in the social sciences of both manual content analysis (Krippendorff, 2004; Neuendorf, 2002) and measurement more broadly (Stevens, 1946; Lazarsfeld and Barton, 1951; Thurstone, 1959).[2] This difference of focus is important. A model which is able to, for instance, identify the pieces of spam in your email inbox will not necessarily do a good job of characterizing the proportion of your email that are spam. Measurement also places heavy demands on the analyst to define what the measurement is and justify its role in the research process.

It will be helpful to begin with a statement of the properties we aspire to in our measures. These five characteristics constitute a demanding set of criteria which may not be fully satisfied in any one paper or project.

[2] The first computer assisted content analysis was Sebeok and Zeps (1958) which analyzed folktales. Hays (1960) was the first with political documents followed by Stone et al. (1962); Stone and Hunt (1963); Stone, Dunphy and Smith (1966) with the General Inquirer.

Principle 1: Measures should have clear goals. A measurement is a reduction of the original source material. There is no universally right or wrong piece of information to preserve about the original source and as such a measure should have a clearly stated goal about what it is trying to capture about the world. This implies that the measure has clear *scope and purpose* which helps ensure that it is used properly in the initial analysis and hopefully ensures it is not inadvertently misused by future scholars. Where possible this statement should be sufficiently simple that it is clearly understandable by non-experts in the field.

Principle 2: Source material should always be identified and ideally made public. Texts have varying meaning which is sensitive to attributes of the context in which it was produced. In order to clearly communicate the meaning of the measurement the source of the original material should always be clearly stated. Although it may not always be possible, making the source texts publicly available both facilitates future and allows for independent validation of the measure.

Principle 3: The coding process should be explainable and reproducible. In order to be used as a scientific measurement we need to understand how the measure is constructed. While this may take many different forms, in principle the reader should have available sufficient instructions that given the raw input documents the resulting measure can be reconstructed. In the case of human coders this generally involves releasing a comprehensive codebook which defines clear procedures for workers implementing the coding system. For computer assisted tasks this involves a complete description of the model or software used to produce the coding. The "explainable" characteristic requires that these materials ought to be understood by the broadest possible audience. In our view, a measurement cannot properly be a part of science if we don't know how it was generated.

Principle 4: The measure should be validated. Our measures should be chosen such that the labels we give them are understandable to a broad audience. There is always a conceptual gap between the way a phenomenon is actually measured and the (generally aspirational) name that we give it. However, we need to strive to keep this gap as small as possible. Validitation is the process of establishing for our readers and ourselves that the value produced by the g function from the text, maps well to the theoretical concept that it purports to measure. We discuss the many forms of validity and reliability in Section 8 below.

Principle 5: Limitations should be explored, documented and communicated to the audience. Some degree of measurement error is inevitable — but the consequences of measurement error are closely tied to how we intend to use the measure in our research. Thus, it is crucial that analysts identify and disclose when the measure performs well and when it doesn't. This allows readers and users of our measures to calibrate their expectations appropriately. For example some measures discriminate well along the entire range of their possible values, while others perform best only at the extremes or in the middle. Defining the scope conditions of our measure helps to ensure that it is used and interpreted correctly.

By expanding the possible sources of information to all written documents, text as data methods allow us to encode measurements which are closer to our theoretical quantities of interest. Traditionally social scientists primarily make use of data that were originally collected for a purpose distinct from the researcher's current work (Salganik, 2017). The use of existing data is powerful because it lowers the cost of research by allowing scholars to piggyback off pre-existing work but at the cost of introducing a gap between our theoretical concept and our quantifiable measure. Gross Domestic Product (GDP) measurements aren't collected for a specific academic study but they are often use as a surrogate of economic strength. Crucially, GDP is not necessarily an appropriate measure for every argument which needs a measure of economic health.

The text as data methods described in this chapter decrease the cost of measuring new quantities from text, making it easier for scholars to design their own data. This allows us to increase the breadth of our questions and more carefully tailor our measures to our argument. Regardless of the measurement strategy we adopt, it will be necessary to think carefully about how our realized measurement matches the theoretical ideal we are attempting to approximate. No amount of automated assistance can obviate the need for human judgment in matching measures to our theorized concepts.

There is a considerable literature on the theory of measurement in both the natural and social sciences. By contrast to many of these fields, Text as Data is in its infancy. While in this chapter we draw on insights from philosophy of science, psychometrics and other traditions of measurement, none of these other substantial bodies of literature applies directly to the challenges of text data. The great insights of these earlier traditions were not generated in a vacuum; but were instead the rationalization and codification of the best procedures generated through trial and error in the early development of the field (Duncan, 1984). A full theory of measurement is beyond the scope of this work, but we offer the best guidance we have from the results of our own process of trial and error.

Once we have a measurement that matches our theory we can do either descriptive, causal or predictive inference. Because text as data often gives us access to the ability measure new phenomena systematically, it is often sufficiently interesting to simply count things. Descriptive inference has a poor reputation in the social sciences, but in early investigations of new phenomena simply being able to quantify the frequency of something can be powerful first step in understanding it. Some of the most profound social scientific contributions have been to measurement of a phenomenon. In this chapter we focus on the task of descriptive inference, but the measures developed here can also be used in a causal and predictive framework as we explore in Chapters 6 and 7 respectively.

3 Repurposing Discovery Methods

The previous chapter outlined a number of techniques for discovery: single-membership clustering models, mixed-membership topic models and low-dimensional embedding strategies. These techniques can also be re-purposed as measurement strategies although the details of their use will differ somewhat. Although discovery and measurement are substantively different tasks, the methods we used for discovery were all finding g functions and thus can also be applied to measurement.

In the initial years of use within the social sciences, the role of discovery and measurement has frequently been blurred as the different fields applying these approaches began to make sense of what text analysis methods were capable of. A clear way to separate the two is to conduct them on different sets of data. For example, Grimmer and King (2011) discover their conceptualization of how legislators communicate in press-releases using a small corpus from a single senator. They then use a different measurement strategy on a different corpus to confirm their finding. In practice this clear separation is rare, but it does have advantages as we will describe below in Section 8.

In this section we rethink the methods we used for discovery and talk about some considerations that become more salient in the context of measurement. We also describe amendments to the models that are motivated by measurement concerns.

3.1 Rethinking Unsupervised Methods

What we typically think of as unsupervised models: clustering, topic models and embeddings, all work on a principle of compression. What makes the latent variables learned by these models interpretable to humans is that they are trying to squeeze high-dimensional information through a low-dimensional channel. For example, a standard topic model needs to represent all J individual words that appear in a document using only K topics. Because

J is generally much larger than K, the model puts words together that frequently co-occur because there are only a small number of channels to work with.

The need to explain all the observed words in a document using the K topics tends to favor words that are frequent. Language in general tends to be quite noun-heavy with generally well over 50% of words used being nouns. This tends to weight topic models towards picking up subject matter rather than, for example, sentiment. This is where topic models get their name. We haven't made any explicit decision in the model that forces it to pick up subject matter, it is the interaction of the way language is used combined with a modeling approach that weights more heavily frequent words that produces this result.

This all means that unsupervised methods are more likely to work well 'off-the-shelf' if the goal of measurement is to pick up subject rather than style, tone, or some other more subtle property of the text. However, the methods will always be sensitive to what is prominent in the text themselves meaning that by changing the contents of the document-feature matrix we can also change the type of latent feature that will be discovered. For example, using part-of-speech tagging to discard everything except the adjectives and adverbs in a set of texts will tend to push the model more towards sentiment.

What makes these models ideal for discovery is that they require relatively little information from the analyst at the beginning of the analysis; instead, the role of the analyst is primarily concentrated in the process of interpretation and validation of the measure. When deployed for measurement, these methods tend to work best in settings where we want our measurement to explain all of the content of our pre-processed text and/or we don't have strong prior expectations about the nature of what the text contains. If an analyst has a specific quantity that she wants to extract, particularly if that quantity is not contributing to a large portion of the pre-processed text, the use of unsupervised methods can be an incredibly frustrating experience because they might not pick this quantity up. Below we highlight some of the commonly used methods that fall under the rubric of discovery methods applied to measurement and highlight the circumstances where we might expect them to excel.

3.1.1 Simple Latent Variable Models of Word Counts

A common paradigm for unsupervised methods is to specify a model that tells a simplified story about how the observed data arises from a low-dimensional latent variable. In our notation, this involves specifying $p(W, \pi, \mu) = p(W|\pi, \mu)p(\mu, \pi)$ where W is the observed documents (typically represented as a document-feature matrix), π_i is the latent measurement for document i and μ is the set of parameters mapping from the latent measurement to the observed data. A particularly common form of this model is a latent linear multinomial model where

$$W_i \sim \text{Multinomial}(\underbrace{\sum_j W_{i,j}}_{\text{\# tokens in doc}}, \sum_k \overbrace{\pi_{i,k}}^{\text{document loadings}} \underbrace{\mu_k}_{\text{topics}}), \tag{3.1}$$

where N_i is the observed number of words in the document, μ_k has typical element $\mu_{k,j}$ which gives the probability of observing feature j under topic k. This general formulation encompasses multinomial single-membership models, mixed-membership topic models and with small modification multidimensional scaling models. Each different option implies a different form for the latent measurement π.

Commonalities Across Models Latent linear models of word counts are designed to compress information about the presence or absence of all J features in a document into a K-dimensional variable π_i. Because J is generally much larger than K, this involves efficiently compressing information about the words. It is this compression that is going to cause these models to produce semantically interesting results.

Each of these models compresses information in approximately the same way. Words which commonly co-occur are grouped together in a single latent dimension. This means that the latent dimensions encode collections of words that frequently co-occur in the corpus. We refer to these latent dimensions as topics below because they frequently convey subject matter.

Single Membership Clustering When the latent variable, π_i, is a vector with one 1 and the rest 0, we get a single membership clustering model. Intuitively, the model learns a set of K archetypes (as described by μ_k) which partition the documents. The document is represented by the archetype which it is closest to based on the distance measure implied by the data generating process. Because every document is explained by one and only one topic, the topics learned are typically quite diffuse, covering many different words in the vocabulary, which in turn can make them somewhat more challenging to interpret. However, because the representation at the document level is incredibly simple, it is very straightforward to summarize results of a corpus since we can simply report how many documents fall into category k by counting up the values of the latent variable ($\sum_i \pi_{i,k}$).

Mixed Membership Topic Models We obtain a mixed membership model when π_i is a vector of proportions (i.e. a value on the $K-1$ simplex, a set of K non-negative numbers which sum to one). Whereas the single membership clustering groups documents into topics, the mixed membership model represents each document as a combination of topics. We can think of this as decomposing each document into achetypical representations. This adds additional modeling flexibility beyond the single membership clustering. For example, imagine that we have a book which is equally about computer science, social science and statistics. We need never have seen such a book before in order to represent it well with the model, we only need to have a concept of each category separately. The limit of this representation is that we cannot represent a document that is more extreme along a given dimension than the topic's archetypical representation μ_k.

The topics learned under a mixed membership model are often easier to interpret than those under a single membership model. Because the topic only needs to represent a part of the document, the topics often put the majority of their weight on only a small handful of words which lends them a highly focused quality that makes simple word lists more informative. The downside is that this makes the document measurements π_i correspondingly more hard to work with. The document measure $\pi_{i,k}$ represents the proportion of words in document i that arise from topic k. When aggregating over collections of documents this can lead to subtleties of interpretation. In the single membership case $\sum_{i \in \mathcal{I}} \pi_{i,k}$ yields the proportion of documents assigned category k in the set $\mathcal{I}$. In the mixed membership case, the same quantity is the expected proportion of words that come from topic k in a given document.

Low-Dimensional Scaling Models When π_i is continuous, a closely related model yields the IRT scaling model known as Wordfish (Slapin and Proksch, 2008). In our notation the model is

$$W_{ij} \sim \text{Poisson}\left(\exp(\alpha_i + \psi_j + \mu_j \pi_i)\right) \tag{3.2}$$

where α_i is a document-level offset capturing the length of the document, ψ_j is a word-level offset capturing the frequency of the word and μ_j and π_i are the word-specific and document-specific latent variables respectively.

This is closely related to the multinomial models below due to connections between the Poisson and Multinomial model (see Lowe 2016 for an account of the unifying similarities between Wordfish and other approaches to scaling). While in theory these models could be estimated with any number of dimensions, they are in practice estimated with only 1 as in the case of Wordfish, or a very small number (2-3) in the case of alternative models. Continuous latent variable models can be difficult to interpret outside the context of the spatial models they are motivated with. Wordfish, for example, is designed to measure the spatial location of actors within a political space. As pointed out in the discovery chapter, this relies on an assumption of *ideological dominance*, that is that the ideological dimension we want to discover is the primary dimension of variation in the text.

3.2 Incorporating Side Information

Most of the discovery models we presented in the last chapter use relatively few inputs: the content of the documents, the choice of the model and some hyperparameters such as the number of latent components. This is what makes these approaches particularly appealing for discovery — they require relatively little information from the analyst up front. However, social scientists often have considerably more information about their documents which they encode as metadata, document-level variables which describe additional information about the document such as its author, date of publication, place of publication etc.

A common pattern in social science research is to look at the way the topical content of documents varies with the metadata. For example Catalinac (2016) examines how the content of Japanese party manifestos change over time and DiMaggio, Nag and Blei (2013) examine how different newspapers over time frame government funding for the arts. These approaches are an extension of the haystack metaphor we discussed in Chapter 2. Instead of only carrying about studying a single strand of hay or characterizing the entire hay stack, many scholars want to characterize individual haystacks separated over time, by organization or by author.

These descriptive goals have lead to the development of models for measurement that explicitly bring in this metadata information. In order to study the way that U.S. senate offices allocate press attention across issues, Grimmer (2010) developed a custom model that explicitly incorporates the role of the author. Quinn et al. (2010) and Blei and Lafferty (2006) develop models particularly designed to track change over time. While these models are built for particular kinds of corpus structure, we can think generally about topic models that are able to incorporate covariates which capture this structure. In fact, in the last chapter we already encountered a general example of a covariate-based topic model, the Structural Topic Model. In this section, we place this model in a broader context of different ways to make sure of side information.

Incorporating covariate information into topic models can serve multiple distinct purposes and help tailor the model to the task at hand. In the context of mixed-membership topic models Mimno and McCallum (2012) introduce a useful distinction between *upstream* covariate models and *downstream* covariate models. Each has a different goal with *upstream* covariates serving to model a particular structure in order to improve our topics interpretability and estimate an association with topics. *Downstream* covariate models are specifically designed to generate topics which are predictive of the covariate. We explain the two approaches below and add a third kind of covariate information based on topical content. Finally we consider some of the ways to incorporate non-covariate side information into topics.

We give nearly all of the examples below in terms of single membership and mixed membership topic models because these are the settings where such models have been developed in practice. The core insights are applicable to other settings as well such as scaling models (e.g. Kim, Londregan and Ratkovic Forthcoming).

3.2.1 Upstream Covariate Models

Recall the canonical Latent Dirichlet Allocation model described in Chapter 4 and shown in graphical model notation in Figure 3.2.1.

In an upstream covariate topic model, we replace the global prior (α) for the document-topic proportions (π_i) with a document-specific prior formed as a function of document-level covariates (X_i). The "upstream" name comes from the graphical model, where the covariates

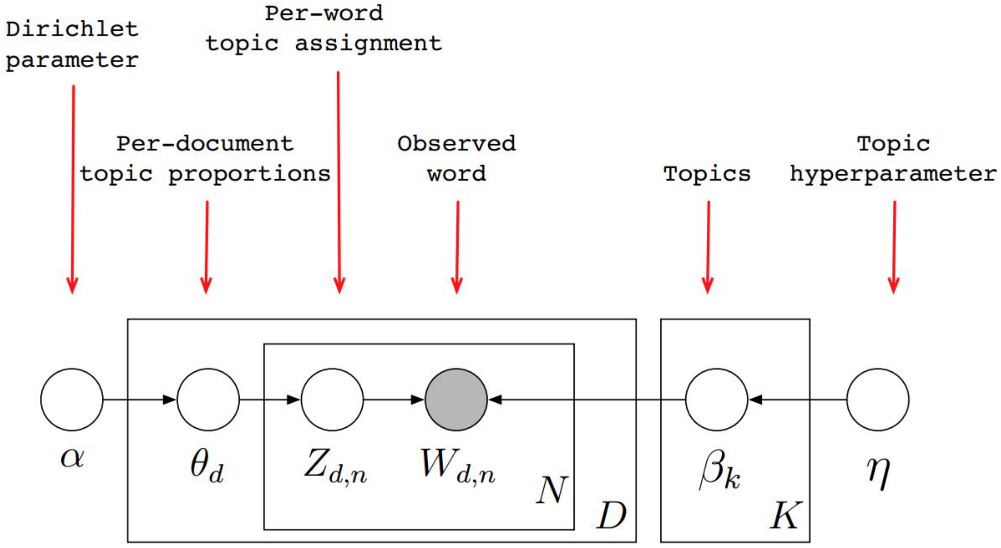


Figure 1: A placeholder figure depicting the LDA model. Will update to include new notation and include the location of the upstream covariate.

are in an area of the graph such that following the flow of the arrows downstream leads to the document-topic proportions. In practice this means that upstream covariate models allow for information to be shared, but unlike the downstream covariate models we will talk about below, the topics are not trying to explain the covariates.

Some of the earliest extensions to the original LDA model were upstream covariate models including organizing documents by author (Rosen-Zvi et al., 2004) or time (Blei and Lafferty, 2006). Further extensions have generalized these specific types of covariates to more general covariate infrastructures (Mimno and McCallum, 2012; Roberts et al., 2014). The approach has been applied to single membership models as well (Quinn et al., 2010).

Upstream covariate models have been popular in the social sciences as a way to refine measurements. Expanding on the pioneering qualitative work of Fenno (1978), Grimmer (2010) explores the question of how legislators portray themselves to their constituents and cultivate a representational style. To get at this style, Grimmer (2010) collects 24,000 press releases put out by U.S. senate offices in 2007. Importantly though, Grimmer's interest is in the allocation of attention by a given senator across topics, rather than the topic of any individual press release. To get at this phenomenon, Grimmer (2010) introduces the Expressed Agenda model which is an upstream covariate single-membership model where each press release is about a single topic and each senator is characterized by a distribution over topics. Using the upstream covariate structure allows the model to share information across different press-releases from the same office and more directly estimate the specific quantity of interest — the expressed agenda of each senate office.

Because the upstream model allows for sharing information through the prior, the upstream covariates won't matter for incredibly long documents. As document lengths tend towards infinity, the data overwhelms the prior and the estimate of the topics for a given documents is driven only by the words it contains. Of course, as documents become extremely long, it becomes more appealing to break them up into smaller chunks so that individual co-occurences become more meaningful. Topic models that induce a hierarchical structure like upstream covariate models can then be used to aggregate back to the true quantity of interest.

3.2.2 Downstream Covariate Models

In a downstream covariate model (see Figure 3.2.2), the covariate is generated from the individual token latent variable placing it downstream of the topic-proportions. This forces the topics to explain both the words *and* the covariate value for a given document. These models thus serve two distinct purposes which are fundamentally different than upstream covariate models: maximizing predictive performance and projecting into a common space.

To motivate the predictive performance case, imagine that we want to predict the number of stars that a movie reviewer gives to a movie (Y_i) on the basis of the text of the review (W_i). The hope, for example in the original LDA paper (Blei, Ng and Jordan, 2003), was that topics from the LDA model would be good features for use in a classifier. In practice,

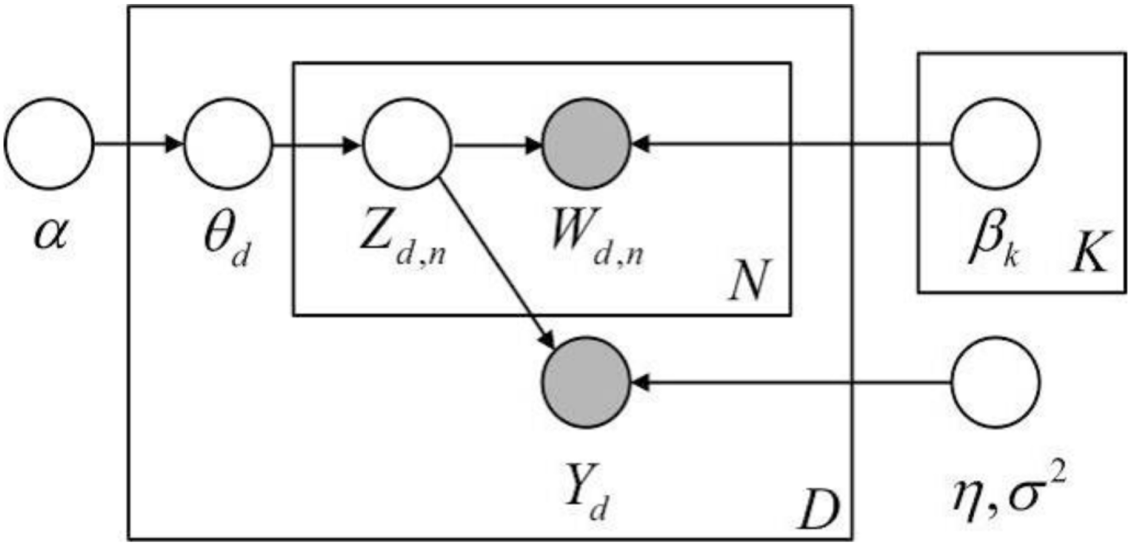


Figure 2: Placeholder figure for downstream covariate models.

they often were not. While the topics learned by LDA (π_i) were good predictors of the words actually used, they weren't necessarily good predictors of our outcome of interest Y_i. This is the problem Mcauliffe and Blei (2008) addressed by introducing Supervised LDA (sLDA). Their insight was that by placing the variable we want to predict, Y_i downstream of the topics, the topics are forced to explain both W_i and Y_i. For a new document we can use the observed words to infer π_i and then use π_i to predict the outcome Y_i.[3] Thus including the downstream covariate helps ensure that the topics will be good predictors, an idea that will recur in later chapters on causal inference and prediction.

In the case of sLDA we typically have a low-dimensional variable Y that we ultimately want to predict and the high-dimensional words W. We can also have a high-dimensional Y that we want to project into a common space with the words. A common use case is where Y are votes and W is text. For example, we might consider bills before congress and their votes (Gerrish and Blei, 2012; Kim, Londregan and Ratkovic, Forthcoming) or supreme court opinions and votes (Lauderdale and Clark, 2014). These models project both the votes

[3] In practice we are actually using a function of the token level variables Z because in practice it performs better. When π_i is predicting Y_i it can become too disconnected from the words themselves (Mcauliffe and Blei, 2008).

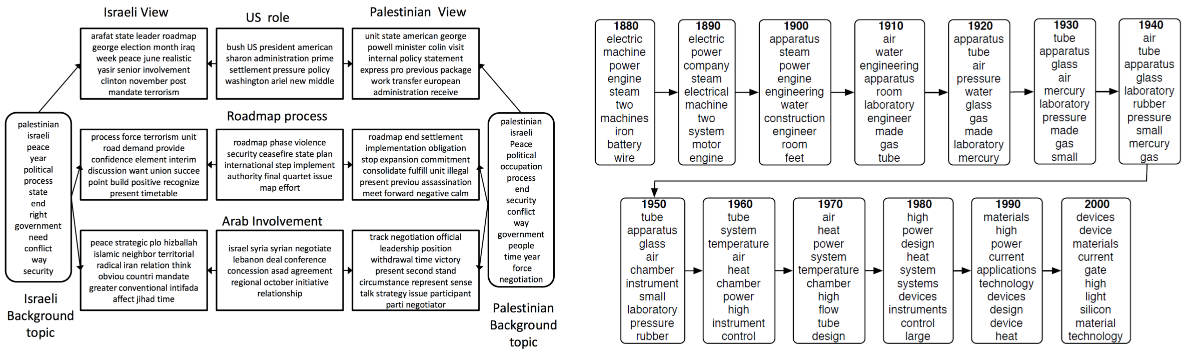

Israeli View
US role
Palestinian View
arafat state leader roadmap george election month iraq week peace june realistic yasir senior involvement clinton november post mandate terrorism
bush US president american sharon administration prime settlement pressure policy washington ariel new middle
unit state american george powell minister colin visit internal policy statement express pro previous package work transfer european administration receive
Roadmap process
process force terrorism unit road demand provide confidence element interim discussion want union succee point build positive recognize present timetable
roadmap phase violence security ceasefire state plan international step implement authority final quartet issue map effort
roadmap end settlement implementation obligation stop expansion commitment consolidate fulfill unit illegal present previou assassination meet forward negative calm
Arab Involvement
peace strategic plo hizballah islamic neighbor territorial radical iran relation think obviou countri mandate greater conventional intifada affect jihad time
israel syria syrian negotiate lebanon deal conference concession asad agreement regional october initiative relationship
track negotiation official leadership position withdrawal time victory present second stand circumstance represent sense talk strategy issue participant parti negotiator
palestinian israeli peace year political process state end right government need conflict way security
Israeli Background topic
palestinian israeli Peace political occupation process end security conflict way government people time year force negotiation
Palestinian Background topic
1880 electric machine power engine steam two machines iron battery wire
1890 electric power company steam electrical machine two system motor engine
1900 apparatus steam power engine engineering water construction engineer room feet
1910 air water engineering apparatus room laboratory engineer made gas tube
1920 apparatus tube air pressure water glass gas made laboratory mercury
1930 tube apparatus glass air mercury laboratory pressure made gas small
1940 air tube apparatus glass laboratory rubber pressure small mercury gas
1950 tube apparatus glass air chamber instrument small laboratory pressure rubber
1960 tube system temperature air heat chamber power high instrument control
1970 air heat power system temperature chamber high flow tube design
1980 high power design heat system systems devices instruments control large
1990 materials high power current applications technology devices design device heat
2000 devices device materials current gate high light silicon material technology

Figure 3: Two different views of models where topical content varies by an observed covariate. (Left) This image comes from Ahmed and Xing (2010) and models data from BitterLemons, a website designed to provoke dialog between Israelis and Palestinians on the Israeli-Palestinian conflict. The image shows three topics (US role, Roadmap process, Arab involvement) across two different levels of a content covariate indicating the background of the author (Israeli, Palestinian). (Right) This comes from Blei and Lafferty (2006) and shows the evolution of a single topic over time learned from a corpus of articles from the journal *Science*. Here time is the content covariate.

and the text into a common space which can be an effective way to make use of data from multiple sources, both of which we believe inform the underlying measure.

3.2.3 Topical Content Models

A third kind of covariate-based model allows for variation in the language used within a given topic on the basis of covariate. These models allow the topic-specific distribution of words to vary by time (Blei and Lafferty, 2006), geography (Eisenstein et al., 2010), ideological affiliation (Ahmed and Xing, 2010) or an arbitrary categorical covariate (Eisenstein et al., 2010; Roberts et al., 2014). Using a content covariate model essentially creates multiple versions of each topic, one for each level of the content covariate. These are constrained either by sparsity in the deviations between the different versions of the topic (Eisenstein et al., 2010; Roberts et al., 2014) or through a smoothness assumption such as continuity through time (Blei and Lafferty, 2006).

Topical content models can be powerful when we want to understand the differences in language use between different sub-populations who are talking about the same thing but using slightly different words. For example, Republicans might use the phrase ‘death tax’ while Democrats use the phrase ‘estate tax,’ even though they are referencing the same set of issues. Even when the differences in language use are not the primary quantity of interest, these models can be helpful for marginalizing out a difference that is not the main quantity of interest. For example, Lucas et al. (2015) show that content covariates can be useful when working with multilingual text collections. Their idea is to machine translate to a common language and then use a content covariate indicating the language to marginalize out idiosyncracies in translation.

3.2.4 Non-covariate side information

We have focused above on document-level covariate information but there are otherwise to incorporate substantive knowledge into discovery models. When information is available about the topics such that they are sparse (Wang and Blei, 2009; Williamson et al., 2010), have a hierarchical structure (Paisley et al., 2015), or contain certain preferred or banned connections between words (Andrzejewski, Zhu and Craven, 2009), we can integrate this into our model. We can also use interactive human feedback as a source of information when it is available (Vikram and Dasgupta, 2016). Hu et al. (2014) introduce an approach to interactive topic model where an analyst is presented with an organization of the text and can provide feedback with the model being quickly refit on the fly.

3.3 Concerns in Repurposing Unsupervised Methods for Measurement

All the methods described above could be useful in performing discovery. While discovery is mostly concerned with how to spark a good idea, measurement requires making a specific claim about what the g function is capturing and how it approximates some theoretical concept. When these methods are turned to the measurement task, we reasonably have additional concerns. Below we address a number of concerns expressed about using unsupervised methods for measurement and give a general reply below. Thankfully, all the concerns have a common response: the value of the measurement is justified not by the model but by our validation of the result as a useful social scientific measure of some quantity. These are tools which can lead to useful measurements in practice, but also may not. To help keep the ideas tractable, we discuss the concerns in the case of mixed-membership topic models; however, the concerns are substantially more general.

Concern 1: The method always returns a result. This concern is often phrased as the objection: 'if this model always yields clusters/topics/scales (even when the data is random), how can this be a valid measurement strategy?' As with many other statistical models, we first assume the model and then find the parameters which best fit the data. In the same way that a linear regression will always return a line, the latent variable models will always return some latent variables and thus some g function. This is why validation is such an important part of the research process. It is not the results of the model that let us know that we have measured something, it is the process of labeling and validating the g function.

The objection does correctly identify that latent variable models tend to be poor tools for asserting the existence of a particular underlying structure — even in purely synthetic data where we know such structure to exist. Running a cluster model with $K = 5$ provides only minimal evidence that there are five clusters in the data. Rather, what the method produces is the best 5-cluster approximation to whatever structure does exist. This neither guarantees that there is an underlying cluster structure, nor does it guarantee anything about the number of clusters. In general, we don't even think there is a 'true' underlying structure in real world data, merely things that are more and less useful. It is useful to know, that even if we accepted the idea of a true structure, this wouldn't be the way to go about estimating it.

Concern 2: Opaque differences in estimation strategies. For any one model there are several different approaches to estimating the best fitting parameters. Although the details of estimation are largely beyond the scope of this book, we want to emphasize that these differences can be particularly salient in certain applications.

To make this concrete, consider the vanilla LDA model. This is a Bayesian model where the goal in estimation is to approximate the posterior distribution. There are three broad approaches to inference for this model: collapsed Gibbs sampling (Griffiths and Steyvers, 2004; Yao, Mimno and McCallum, 2009), variational inference (Blei, Ng and Jordan, 2003) and spectral methods (Arora et al., 2013).[4] These methods all have benefits and drawbacks which are hard to enumerate in full without a technical background. Unfortunately these differences can be important for certain problems and quantities of interest (Asuncion et al., 2009; Boyd-Graber, Mimno and Newman, 2014). In practice, we recommend that authors always indicate the approach to estimation in addition to the particular software implementation they used.

Concern 3: Sensitivity to un-intuitive hyperparameters. We have already discussed the selection of K, probably the most obvious parameter that one manually sets in using topic models. However, there are a number of other hyperparameters which we set explicitly or implicitly that can substantially affect performance (Wallach, Mimno and McCallum, 2009; Wallach et al., 2010). Ultimately these choices are difficult because practioners may have few intuitions about how to set them optimally. We find the best strategy when forced to set them manually is to consider many different options and choose based on substantive

[4] Technically the spectral methods don't approximate the posterior distribution, they provide a point estimate of some parameters of the model.

fit.

Concern 4: Instability in results. The final concern is also at the heart of the other concerns: the instability of latent variable models, especially relative to other common statistical models in the social sciences like linear regression. Essentially all the methods described above lead to non-convex (and thus not easily solvable) optimization problems. In practice, results are sensitive to the initial starting values and each time the model is run, we can get a different result. The different results arise because the model is finding locally optimal solutions instead of a globally optimal solution.

This is understandably unsettling to practitioners who are concerned about whether their results are accurate. A number of articles have explored the problem in some depth. For example. Roberts, Stewart and Tingley (2016) explain the issue, provide strategies for exploring different solutions, and offer different initialization strategies (see also Chuang et al. 2015 which provides additional visualization approaches).

The instability issue isn't a simple problem of reproducibility (being able to generate the same results from the same data consistently). The analyst can set the random number generator seed in order to ensure that results are consistent across computers.[5] Roberts, Stewart and Tingley (2016) go further and propose an initialization strategy that relies on a deterministic and (under certain conditions) globally consistent estimator developed by Arora et al. (2013).[6] This approach does not guarantee we will find the globally best solution in any one case, but it does perform well empirically and offers asymptotic guarantees. Even when we can find stable initialization strategies, the models are often relatively sensitive to seemingly minor changes such as preprocessing or changings in the underlying data (Denny and Spirling, Forthcoming; Wilkerson and Casas, 2017). This instability in turn makes

[5] In practice, even this can be difficult due to different versions of the software and different levels of numerical precision on different machines. However, these are engineering problems that are quickly being worked out.

[6] See also the careful numerical analysis and initialization strategy of Lancichinetti et al. (2015).

the distinction between globally and locally optimal solutions less relevant. Because small perturbations of the data can cause a local optimum to potentially become global, the distinction of one that is higher than the others is arguably less relevant- they simply measure different things which may at different times provide the best fit to the data you have. We still believe that have a locally optimal solution is important though because it lets us know that given what we are measuring and the data we have, this measurement is the best fit to the data.

Denny and Spirling (Forthcoming) eloquently articulate the core concern in their paper on the impact of text pre-processing steps for unsupervised methods,

> To underline our philosophical point here, note that the issue is not simply that dishonest researchers might cynically pick a specification they like and run with it, to the detriment of scientific inquiry. The more subtle problem is that well-meaning scholars would have *no idea of the truth value* of their findings. A particular feature of unsupervised models of text is that *there are typically many possible specifications, and many plausible stories about politics that can be fit to them, and validated, after estimation.* Fundamentally then, a lack of attention to preprocessing produces a potentially virulent set of “forking paths” (in the sense of Gelman and Loken, 2014) along which researchers interpret their results and then suggest further cuts, tests and validation checks *without realizing that they would have updated had they preprocessed their documents differently* (emphasis ours)

This concern is held more broadly by a number of social scientists who are skeptical of the ability of discovery methods to be used for measurement. The authors ultimately suggest that in the absence of strong theory, we would like to see that our results are robust to different preprocessing decisions and ideally we would average over different alternatives.

Wilkerson and Casas (2017) make a similar appeal for robustness across different parameter values more broadly.

Rethinking stability Unsupervised methods don't have direct information about the quantity that the analyst wants to measure. When we have an objective, easily evaluatable criterion for performance, these discussions are more straightforward: we simply search over the space of possible solutions and find the one that performs best on that criterion. What is remarkable about discovery methods is that they offer a workable *surrogate* criterion to an unspecified measurement objective. Sometimes this works to measure the concept of interest and sometimes it doesn't.

The key to the critiques of Denny and Spirling (Forthcoming) and Wilkerson and Casas (2017) is a presumption of a stable, underlying target that these methods are trying to measure. Under this worldview, every different model solution is a different approximation to some underlying truth in the document. Under this view, it absolutely makes sense to pursue a result that is robust to different variations of the model.

We argue instead that every re-specification of the model is capturing something different about the documents. We shouldn't confuse instability in the *measurement of a fixed concept* with differences in the *choice of concept we are measuring*. While there are differences in what we can measure from a set of documents, a properly validated measurement is constrained by the actual content of the documents.

If this seems counter-intuitive we offer the simple assertion that most scholars would find it absurd if someone claimed there was only one right way to read a text. But this is what a stable, underlying target measurement implies in this context. This stability seems natural when compared to many other quantitative activities, like estimating a well-defined estimand using an experiment or regression analysis. But that presumption of stability is not there in the more comparable qualitative manual content analysis tradition. Consider two different teams of qualitative scholars approaching the same set of documents — they would come up with different coding schemes. We wouldn't dismiss the findings of either team because they chose to employ different methods and answer potentially different questions. We should judge the content analysis enterprise not by whether multiple methods come to the same solution, but by what the resulting measure tells us about the world.

Repurposing discovery methods for measurement ultimately places the burden on the analyst to carefully validate the resulting measures. As we will see, the supervised methods we describe below often rely on strong assumptions and require validation. Thus, from our perspective, the need for validation was inevitable.

4 Methods for Known Categories from Human Coding

The methods in the previous section focused on the case where the human provided at best indirect information about the content of the categories that the model is intended to define. This ensures that the categories are well-separated by whatever metric the model uses to measure distance, but it certainly does not guarantee that they are the categories the analyst intended or wanted to measure. In fact, when the analyst has a very specific idea about what she wants to measure in the data, she may have a very bad time trying to get unsupervised methods to find those specific things.

In this section, we approach the problem from a different paradigm, one that is usually associated with manual coding and supervised learning. In this setting the analysts understand the categories well enough to be able to do the categorization themselves (i.e. sort any document into its appropriate category), although perhaps not at scale. Although many of the methods described are more general, we will focus on the specific case where our goal is for each document to be placed into a single mutually exclusive and exhaustive category. This is the most common use cases for known categorization schemes because other types of latent representations we have considered above (mixed membership, multiple binary features and scales) can be harder for humans to code.

Categorization tasks of this sort have long been a component of artificial intelligence (AI), but the general approach to them has changed over time. In the early days of AI, coders tried to implement the rules and logic of how to place things into categories. In our vocabulary, they tried to program the g function directly by trying to abstract the rules used by human coders. That is, they tried to get the machine to solve the problem in the human way. Then the machine learning revolution happened and the default approach started to change. AI experts realized with samples of the input D (in our case the text) and the desired output Y (the category label), basic statistical algorithms could *learn* the g function from the data. Surprisingly, even with a relatively small number of samples this procedure, which forms the basis of modern machine learning, can dramatically beat attempts to manually elaborate the logic and rules implicit in g. This means that the algorithm makes decisions differently than a human would, but ultimately it is trained to replicate human judgment as closely as possible. As more and more data have become available, we have been able to tackle more and more complex tasks with higher and higher accuracy.

The key to these automated approaches is in the feature representation we create for D. Cases where we have better features require less complicated models and are easier to gain high performance. One thing that makes text problems relatively straightforward is that the bag-of-words representation is a pretty good feature set for problems that are about broad topical content. Compare, for example, to trying to build a classifier for images — including every pixel in order is constitutive of the entire image, but each pixel by itself doesn't communicate a lot of information about what the image of the picture is of. By contrast, for assessing the broad topic of a document in particular, a single word has a decent amount of meaning on its own. This has lead to considerable human effort in engineering feature transformations in images and text that perform well in a variety of settings. However, as our document collections get larger we can rely more and more on the statistical model to take a fairly unruly feature set (like individual pixels) and adaptively find the right feature representation to make a classifier.

Below we cover a range of ways to approach the document classification problem when the analyst is able to classify documents into categories at the outset. We cover both cases where we trying to encode the logic of the classification decisions in various ways and others where we rely on samples of coded documents.

We start with manual content analysis and how a team of human coders can be trained using a codebook to place documents into a set of categories. This is a natural place to start both because it shows how the qualitative content analysis tradition fits into our view of text analysis and because manual coding is an essential first step for many (but not all) of the processes of learning a classifier. We then discuss approaches to crowd-sourcing, which use much higher numbers of coders who each receive much less training. We then turn to approaches to encoding simple logic of how to do a classification directly with a particular focus on popular dictionary-based methods. Finally, we cover how to learn the g function from random and non-random samples of human-generated labels. In Section 5 we will discuss approaches to learning the g function from the data that do not rely on human coding of individual documents.

4.1 Human Coding

The oldest tradition of content analysis is human coding. In short, a codebook is written by a principal investigator and then documents are placed into usually mutually exclusive and exhaustive categories. Coding of this sort predated modern approaches to statistical text analysis and are still extremely useful for small datasets and coding of complex concepts.

Even though we have human coders there is still a g function- it is the combination of the codebook, training given to the coders and their internal thought processes when assigning codes. An advantage here is that humans can fill in gaps in coding schemes with context in a way that machines are unlikely to be able to match in the near future. The disadvantage is that the process is much slower and less replicable. Whether these tradeoffs are worth it is going to be a determination that is context specific.

Ambiguities in language, limited attention of coders, and nuanced concepts make the reliable classification of documents difficult–even for expert coders. Unfortunately teams of expert coders are few and far between, the modal manual content analysis project in the social sciences is, as Schrodt (2006) colorfully put it describing the collection of events data, " legions of bored students flipping through copies of The New York Times" (2). Complications arise because of the deeply contextual nature of language that makes it very difficult to specify an entire codebook *ex ante*. For this reason, we recommend an explicit exploratory/discovery phase in which initially, a concise codebook is written to guide coders, who then apply the codebook to an initial set of documents. When using the codebook–particularly at first–coders are likely to identify ambiguities in the coding scheme or overlooked categories. This subsequently leads to a revision of the codebook. Only after coders apply the coding scheme to documents without noticing ambiguities is a "final" scheme ready to be applied to a separate set of documents used for analysis. This ensures that ambiguities have been sufficiently addressed without risk of overfitting.

Manual content analysis has particular advantages and disadvantages when it comes to content 'drift.' When a content stream is evolving over time (say a project like Policy Agendas which tracks bill text over years), there can be changes that create new ambiguities or cause problem for the codebook. Maybe new categories evolve or collapse and merge with others. Because coders can actually come talk to you as the analyst, it is often easier to pick up that this is happening. They can raise ambiguities and give feedback when algorithmic coding will just continue to spit out codes which (when performance is un-monitored) can be become arbitrarily bad. The downside is that when the need for a change is detected, automated methods can quickly revisit old documents whereas human coding cannot. This creates a bit of a paradox where in human coding schemes detecting the problem is easier, but solving it is substantially harder. Over long periods of time coders will often change which can introduce a inconsistency in the human coding even if there is not underlying drift in the material.

Thankfully, manual content analysis is a richly developed field and there are whole books on how to write a good codebook and train coders. We particularly recommend Krippendorff (2012) and Neuendorf (2016). Although beyond the scope of this book, learning to write an excellent codebook and train coders is important even in the context of automated methods as manual coding often forms the basis of training and validation sets for automated methods.

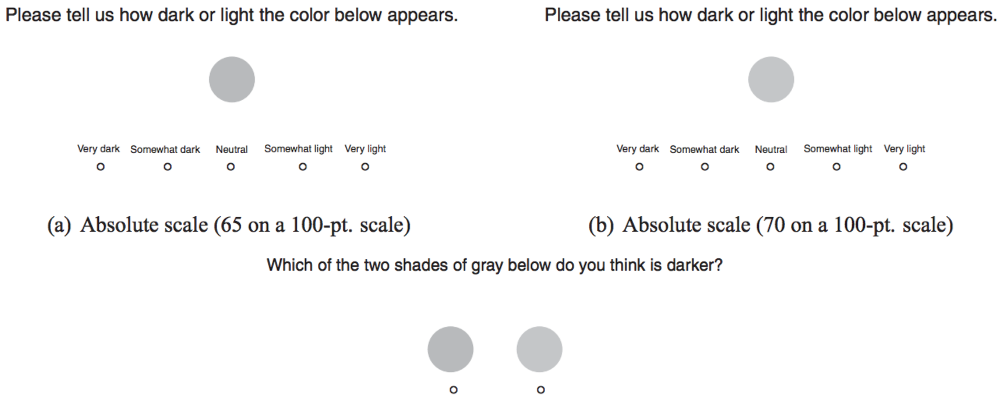
4.2 Crowd Sourcing

In the last ten years, the introduction of online labor markets such as Mechanical Turk has radically altered the landscape of recruitment for annotators, survey respondents and participants in experiments (Snow et al., 2008; Kittur, Chi and Suh, 2008; Buhrmester, Kwang and Gosling, 2011; Berinsky, Huber and Lenz, 2012; Budak, Goel and Rao, 2016). Labor markets provide access to untrained or lightly trained workers at scales that would be unfathomable in a typical university setting. For tasks that can be easily explained, are relatively straightforward and can be quickly completed, online labor markets can be a valuable way to collect document annotations.

It is tempting to think about these online labor markets as infinitely large pools of workers but they are actually much smaller than it might otherwise seem. For example, Amazon's Mechnical Turk is the most popular online labor market for academic research. A recent study estimates that there are less than two hundred thousand unique workers, with about 2–5 thousand active at any given time (Difallah, Filatova and Ipeirotis, 2018). These workers tend to have skewed demographic and political characteristics relative to the population as a whole, although certainly not as skewed as the population of undergraduates (Huff and Tingley, 2015).

Differences between crowd-sourcing and human coders There are three big differences between working with a small team of human coders and a huge crowd. First, in the crowd setting fixed costs have to be lower, so intensive training is less feasible. In practical terms this limits the kind of work that can be done by crowd-workers to comparatively simple tasks that don't require expert background. The advantages of low-fixed costs lead to the second big difference — it is more feasible to quickly scale-up crowd workers to enable high through-put coding. In a university or industry setting, hiring new workers takes time, but thousands of annotations can be collected for a simple task in just a few hours. The third major difference, is that we have to fully embrace the inevitability of error in annotations. With a small coding team we can conceivably train them until error rates are tolerably low, but in crowd-sourcing we have to come up with ways to reconcile conflicting labels. Thankfully this is an old problem which has seen renewed interest in recent years as a result of online labor markets (Dawid and Skene, 1979; Sheng, Provost and Ipeirotis, 2008; Zhang et al., 2014; Benoit et al., 2016). We return to these methods in Section 8 on validation.

Some would add a fourth major difference, that crowd-sourcing is cheaper. It is certainly true that crowd-sourced annotations *can* be obtained cheaper on Mechanical Turk than from undergraduate annotators, but it is questionable whether or not they should be. Because the workers are independent contractors, wages paid to them for tasks fall outside the scope of minimum wage laws. We note that while for the purposes of statistical text analysis it can be helpful to think of the crowd as a certain type of 'algorithm', the workers are very much humans and so ethical considerations about appropriate compensation apply (Fort, Adda and Cohen, 2011; Mason and Suri, 2012; Shank, 2016).



Please tell us how dark or light the color below appears.

(a) Absolute scale (65 on a 100-pt. scale)

Please tell us how dark or light the color below appears.

(b) Absolute scale (70 on a 100-pt. scale)

Which of the two shades of gray below do you think is darker?

(c) Comparison (65 and 70 on a 100-pt. scale)

Figure 4: A figure from Carlson and Montgomery (2017) explaining how pairwise comparison tasks can be more tractable than absolute scale tasks. Task C is going to be much more consistent than either tasks A or B.

Changing the Task In practice, the key to using crowd sourcing has generally been to find the optimal way to divide the coding task into small ‘bite-sized’ chunks. Now that online labor markets have been around for over a decade, researchers have started to creatively develop new tasks that can more easily be discretized. Carlson and Montgomery (2017) introduce a system called SentimentIt which provides a framework for generating scaled representations by collecting thousands of responses to quick pairwise comparison text.

The core insight is beautifully illustrated by an example from Carlson and Montgomery (2017) which is depicted in Figure 4. For many tasks, such as assessing how dark or light a color is, it is quite difficult to assign an absolute scale without some kind of reference point. However, it is very tractable to make a reliable assessment of a comparison between two values. These pairwise comparisons can in turn be used to recover an underlying latent dimension — an idea that has been around in psychometrics since at least Thurstone (1927) and arguably, traces back to the 1800’s.

The SentimentIt platform provides a way to use Mechnical Turk to collect many such pairwise judgments which are then used to model a single underlying latent dimension of interest. Carlson and Montgomery (2017) shows that the system is highly reliable and compares favorably to other approaches. Remarkably, coders seem to need considerably less training to make a pairwise judgment than to accurately assign an absolute scale. The need to make pairwise comparisons causes the number of evaluations needed to scale with the number of texts squared, so this general approach works best with a moderate number of documents (their largest case is less than 2000 documents).

The pairwise comparison approach relies heavily on the idea that the underlying concept being compared has only a single dimension. If there is more than one dimension, comparisons could be based on different dimensions for different respondents. In some cases, we may want to assess an underlying dimension that summarizes a more complex and varied set of considerations. In a relevant non-text setting, Kaufman, King and Komisarchik (2017) set out to measure the 'compactness' of legislative districtness. In the U.S. laws surrounding gerrymandering, compactness is an important, if somewhat ill-defined concept. Intuitively, the idea is that a compact district is one where the boundaries are not unevenly close to the geographic center of the district and the shape is 'regular.' There have been hundreds of proposals for measures of compactness, but none seem to get at the essence of human judgment.

Kaufman, King and Komisarchik (2017) set out to measure compactness from the image of a legislative district (which we can think of as analogous to the text). They find that pairwise comparisons of the sort that underlying the SentimentIt system fail, and thus introduce a system based on ranking a larger set of 100 items. They find that judgments are much more consistent in the ranking task than the comparison task. Presumably this is because respondents are forced to choose a single dimension under which to rank all the items in the ranking setting, but can effectively switch dimensions on every pairwise comparison. Crucially the ranking method still recovers one dimension by construction and implicitly assumes that there is a single salient dimension among which all items are ranked. Still, it seems to perform better in settings where that dimension is multifaceted and complex.

The ranking approach would be difficult to apply to the text setting. Ranking 100 images is a much more tractable task than ranking 100 documents, even relatively short documents. However, ranking more than 2 items at a time might be a useful direction to explore in the text analysis setting.

From Online Labor to Mass Collaboration In his book *Bit by Bit*, Matt Salganik lays out a vision of three different forms of what he terms *mass collaboration* (Salganik, 2017). This involves an important philosophical pivot from seeing the crowd as a source of online labor to a collection of potential scientific collaborators. The first of his forms is Human Computation is functionally equivalent to what we call crowd-sourcing and has been widely used in text analysis. The third form, Distributed Data Collection, involves using large, distributed groups of people to collect data. This has also been used to a limited extent in the text setting, for example in the Malawi Journal Project which collects conversational journals on conversations about AIDS/HIV in Malawi (Watkins and Swidler, 2009). Chakrabarti and Frye (2017) show that these conversations can be fruitfully analyzed with a mix of traditional qualitative and quantitative text analysis methods.

His second form is the Open Call, a setting where a researcher poses a clearly defined problem and solicits solutions from the world. These problems can be large and open-ended, for example, the Netflix challenge which offered a one million dollar prize to the team that could improve Netflix's internal recommendation engine by more than 10%. They can also be more limited in scope, such as Foldit which turned a complex optimization problem, protein folding, and turned it into a simple game that people could play online (Cooper et al., 2010). The Open Call is similar to the Shared Task system in computer science, where many teams of researchers work to find the best performing solutions to a single problem. Yet to this point, these kinds of shared tasks have essentially been non-existent in the social sciences. Salganik's recent work on the Fragile Families Challenge, a competition to predict outcomes for children in the Fragile Families and Child Well-Being Study (Reichman et al., 2001) is perhaps the closest example. We think however that these kinds of coordinated efforts, potentially reaching out to non-academics, could be an important and underutilized resource in the analysis of text.

4.3 Learning with Human-Generated Rules

When moving beyond human coding of texts with known categories, we can directly write down the g-function using a set of human-generated rules or we can learn the g function inductively from human-labeled examples. Arguably the most fundamental of the human-generated rules are keyword or dictionary based methods.

Keyword Counting The simplest version of the dictionary approach is counting instances of a particular word or phrase. Imagine that we have a collection of newspaper articles and we want to measure interest in U.S. presidential politics. We could count the number of articles that use the word 'president' at least once. As a measure this has the advantage that it is clear and easy to communicate to our audience. However, it isn't clear that this measure is able to pick up on discourse about presidential politics; companies, other countries and organizations of all sorts have presidents. We could instead opt for a slightly more complicated option and count instances of the phrase 'White House'. This has the advantage of producing fewer false positives (although there are, of course, other houses which are white), but many articles about presidential politics may not contain the phrase 'White House'. While counting words is straightforward methodologically it is difficult to accurately represent complex latent concepts.

The most compelling cases for keyword counting are those where the particular word choice is highly meaningful and the evidence is supplemented by careful reading. For example, in his book, *Ghetto: The Invention of a Place, the History of an Idea*, ethnographer

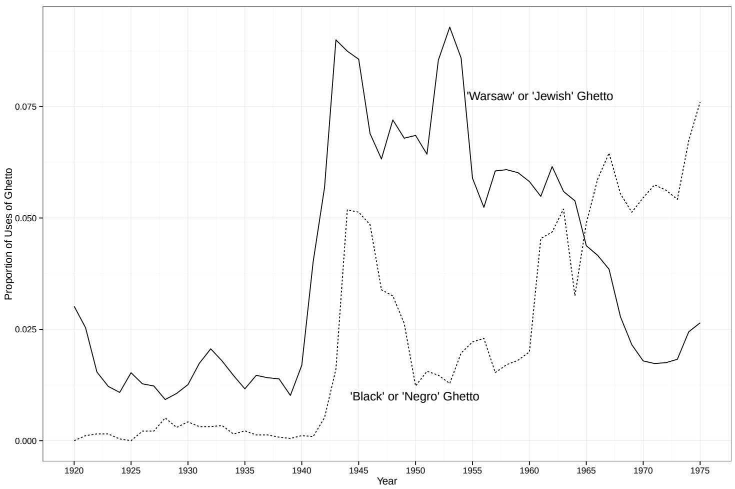


Figure 5: Figure from Duneier (2016) tracking alternate uses of ghetto as a proportion of uses of the word 'ghetto' through time in the Google Books corpus of English language books (Michel et al., 2011).

Mitchell Duneier studies the evolution of the idea of the ghetto, tracing its changing meaning from the 16th century origin of the term through the present (Duneier, 2016). Using a primarily qualitative-historical evidence base, Duneier (2016) argues that the notion of the black/urban ghetto supplanted the original idea of the Jewish ghetto in the early 1960's. This qualitative evidence is supplemented with a simple keyword analysis, tracking usages of the terms 'Warsaw ghetto' and 'Jewish ghetto' and comparing them over time with uses of 'Black ghetto' and 'Negro ghetto' in the Google Books corpus of English language books. The result, shown in Figure 5, shows a clear pattern.

In isolation, the quantitative evidence might be problematic. The total uses of the word 'ghetto' far outnumber the specific uses with the clarifiers Warsaw/Jewish/Black/Negro. There are also serious limits to using the Google Ngrams corpus to make inferences about the popularity of a phrase over time as pointed out by Pechenick, Danforth and Dodds (2015). But in the context of a richer qualitative analysis (including in depth reading of many of the books that are part of this corpus and newspaper articles from the time), a simple graph like Figure 5 provides an additional piece of evidence that helps to make the broader point of a cross-over in the way the concept of the ghetto is understood.

Dictionary Methods Dictionaries use the rate key words appear in a text to classify documents into categories or to measure the extent to which documents belong to particular categories. For example, suppose the goal is to measure the *tone* in newspaper articles (for example, Eshbaugh-Soha 2010): whether articles convey information positively or negatively. Dictionary methods use a list of words with attached tone scores and the relative rate words occur to measure a document's tone. A *dictionary* to measure tone is a list of words that are either dichotomously classified as positive or negative or contain more continuous measures of their content. Formally, each word j will have associated score μ_j. For the simplest measures, $\mu_j = -1$ if the word is associated with a negative tone and $\mu_j = 1$ if associated with a positive tone. If $M_i = \sum_{j=1}^{J} W_{ij}$ words are used in document i, then dictionary methods can measure the tone, π_i, for any document as,

$$\pi_i = \sum_{j=1}^{J} \frac{\mu_j W_{ij}}{M_i}.$$

Scholars often use π_i as an approximately continuous measure of document tone, but it also can be used to classify documents into tone categories if a decision rule is assumed along with the dictionary method. Perhaps the simplest coding rule would assign all documents with $\pi_i > 0$ to a positive tone category and $\pi_i < 0$ to a negative tone.

Tone is just one type of analysis a dictionary method can perform. The general idea of dictionaries make them relatively easy and cheap to apply across a variety of problems: identify words that separate categories and measure how often those words occur in texts (for some recent examples that use dictionaries to measure a variety of concepts, see Kellstedt 2000; Burden and Sanberg 2003; Laver and Garry 2000; Young and Soroka 2011). Finding the separating words is also relatively easy. There are a variety of widely used off the shelf dictionaries that provide keywords for a variety of categories (for example, Bradley and Lang 1999; Hart 2000; Pennebaker, Francis and Booth 2001; Turney and Littman 2003). And if scholars have documents already coded into categories, dictionaries can be produced using existing methods. Monroe, Colaresi and Quinn (2008) describe a variety of methods that measure how well words separate already identified categories of interest, (see also Taddy (2010) and Diermeier et al. (2011)). Any one of these methods could be used to produce dictionary-like scores of words, which could then be applied in other contexts to classify documents.

For dictionary methods to work well, the scores attached to words must closely align with how the words are used in a particular context. If a dictionary is developed for a specific application, then this assumption should be easy to justify. But when dictionaries are created in one substantive area and then applied to another, serious errors can occur. Perhaps the clearest example of this is shown in Loughran and McDonald (2011). Loughran and McDonald (2011) critique the increasingly common use of off the shelf dictionaries to measure the tone of statutorily required corporate earning reports in the accounting literature. They point out that many words that have a negative connotation in other contexts, like `tax`, `cost`, `crude` (oil) or `cancer`, may have a positive connotation in earning reports. For example, a health care company may mention `cancer` often and oil companies are likely to discuss `crude` extensively. And words that are not identified as negative in off the shelf dictionaries may have quite negative connotation in earning reports (`unanticipated`, for example).

Dictionaries, therefore, should be used with substantial caution, or at least coupled with explicit validation. When applying dictionaries, scholars should directly establish that word lists created in other contexts are applicable to a particular domain, or create a problem specific dictionary. In either instance, scholars must validate their results. But measures from dictionaries are rarely validated. Rather, standard practice in using dictionaries is to assume the measures created from a dictionary are correct and then apply them to the problem. This is due, in part, to the exceptional difficulties in validating dictionaries. Dictionaries are commonly used to establish granular scales of a particular kind of sentiment, such as tone. While this is useful for applications, humans are unable to produce the same granular measures reliably (Krosnick, 1999). The result is that it is essentially impossible to derive gold standard evaluations of dictionaries based on human coding of documents.

The consequence of domain specificity and lack of validation is that most analyses based on dictionaries are built on shaky foundations. What is difficult is that it is easy to assign a weight to a word and then produce a document level score by averaging over those weights. The problem is that this needn't produce a measure with any particular properties. Even granted that we have selected the correct features to weight, the fundamental question is whether variations produced in the weights constitute underlying differences in the quantity being measured or if they are just introducing noise. As with all methods, we recommend a case-specific validation.

4.4 Supervised Learning with Random Samples

We have considered two important scenarios for defining a g function so far: the case where the analyst writes a codebook and allows the g function to be interpreted in the mind of the coder and a second where the g function is explicitly defined by a set of rules or a dictionary. A third strategy is to learn the g function from the data by training an algorithm to predict the category of interest from the text using a sample of coded documents. We define a model such that our prediction $\hat{Y}_i$ for document i is generated by a function $f_\mu(W_i)$ where W_i is the document representation and μ are the parameters of the model. We then learn the parameters of the model by observing random samples of the input W and the output Y (called training samples). This essentially turns the measurement problem into a *prediction* *task*, leveraging statistical models to predict the label that the analyst would have assigned herself. This approach is quite general and has the advantage that once the parameters of the model are learned, classifying additional documents is essentially costless.

In this approach, the human is responsible for providing three sources of information: labels for the training documents, the representation of the input, and the class of model to learn. Having the labels typically implies that the analyst has already developed a coding scheme and a g function that lives in her head. We assume for now that the analyst selects samples at random and assigns the labels Y with perfect accuracy to the selected documents.

The second step is the feature representation of the input. This is an important and underappreciated part of the process. The representation chosen should match the type of task that the analyst is engaged in. A recurring theme throughout this section is that the features we use to predict the label Y are substantially more important than the class of models. This point often gets lost in the literature where authors are trying to provide advice which is generic to any type of g function that the analyst might provide. However, *we will choose better features over better models every time.* A good baseline set of features is the vector of term counts W_i which we use here.

The third and final step is choosing a class of models to learn. There are a dizzying array of choices; but all the different options provide different ways of sharing information. If we make no assumptions about how features map to the output, we can only classify documents that are exactly like a sample we observed in the training set. Different classes of models define how the learned g function maps from the input representation to the label. We can choose relatively simple models that are often easy to interpret and require relatively little data to train, but may not be flexible enough to predict human labels. More complex models are hungrier for data and may be relatively opaque to the analyst, but can often do quite complex prediction problems.

We will start with a careful walk through of a baseline model, Naive Bayes (Maron and

Kuhns, 1960), and some of its extensions. The model is important both because it performs reasonably well on its own and it forms the foundation of other more complicated models. We then turn to a broader class of machine learning methods which can be used for classification and prediction.

4.4.1 Naive Bayes

Suppose we wanted to place to build a model which could place each document D_i into a mutually exclusive and exhaustive categories using the document-feature matrix as input. Let π_i denote a k length vector where element $\pi_{i,k} = 1$ if document i is in category k and $\pi_{i,k} = 0$ otherwise. For the documents in our training set $\pi_{i,k} = 1$ when $Y_i = k$ and 0 otherwise. We are interested in the probability that an unseen document is in category k given that we observe document D_i. Using Bayes rule, we can write this probability as:

$$\begin{aligned} p(\pi_{i,k} = 1|D_i) &= \frac{p(\pi_{i,k} = 1, D_i)}{p(D_i)} && (4.1) \\ &= \frac{p(\pi_{i,k} = 1)p(D_i|\pi_{i,k} = 1)}{p(D_i)} && (4.2) \end{aligned}$$

The denominator is going to be constant across all categories, so we can simply drop it and normalize at the end of the process. This leads to a formula whose components each have a clear meaning,

$$\underbrace{p(\pi_{i,k} = 1|D_i)}_{\text{prediction}} \propto \underbrace{p(\pi_{i,k} = 1)}_{\text{prevalence of class } k} \overbrace{p(D_i|\pi_{i,k} = 1)}^{\text{category-specific language model}} \qquad (4.3)$$

We will denote the $p(\pi_{i,k} = 1)$ with the parameter α. It is straightforward to estimate α from the data. Because we have a random sample of documents, we can simply calculate the proportion that fall into each class.

$$\hat{\alpha} = \frac{\sum_i^N I(Y_i = k)}{N} \tag{4.4}$$

To this point, we have been agnostic about the form of the category-specific language model. Unfortunately, the document D_i is complicated to model without any assumptions. To handle this we will switch to the document-feature matrix representation W_I and make the eponymous ”naive” assumption. We will assume that each token in the document is drawn independently conditional on the class. This allows us to write

$$W_i|\pi_{i,k} = 1 \sim \text{Multinomial}(\overbrace{\sum_j W_{i,j}}^{\text{Number of tokens}}, \underbrace{\mu_k}_{\text{word probabilities for class } k}) \tag{4.5}$$

$$p(W_i|\pi_{i,k} = 1) \propto \prod_{j=1}^{J} \mu_{k,j}^{W_{i,j}} \tag{4.6}$$

where μ_k is the vector of probabilities that indicate the probability of drawing each word in class k. The Naive Bayes idea is quite general and if we had features of a different sort (for example continuous), we could have factorized in a different way.

This naive assumption of independence conditional on the document class is almost certainly wrong but it gets us to a set of parameters that we can easily estimate the necessary parameters.

$$\widehat{\mu_{k,j}} = \frac{\overbrace{\sum_i^N \pi_{i,k} W_{i,j}}^{\text{Count of feature } j \text{ in documents of class } k}}{\underbrace{\sum_i \sum_j \pi_{i,k} W_{i,j}}_{\text{Count of features in documents of class } k}} \tag{4.7}$$

We can think of estimating the parameters for class k by collecting all of the training documents assigned to class k together and calculating the frequency with which each word is used.

This does create a problem with words that are sufficiently rare that they never appear under a given class. This causes our estimate of the probability of their occurrence to be zero which means that the given class is completely ruled out if it occurs. We want to avoid this problem and so we add a small constant c for each word

$$\hat{\mu}_{k,j} = \frac{c + \sum_i^N \pi_{i,k} W_{i,j}}{Jc + \sum_i \sum_j \pi_{i,k} W_{i,j}} \tag{4.8}$$

When $c = 1$ this is like adding a new document for each class that contains one of every word in the vocabulary.

Putting this all together, our algorithm can be given as:

$$p(\pi_{i,k} = 1 | W_i) \propto \alpha_k \prod_{j=1}^{J} \mu_{k,j}^{W_{i,j}} \tag{4.9}$$

If we want to make the best prediction for a given class, we often rewrite it using the `arg max` notation to indicate that we want the most probable category,

$$\hat{y}_i = \texttt{arg max}_k \left[\hat{\alpha}_k \prod_{j=1}^{J} \hat{\mu}_{k,j}^{W_{i,j}} \right] \tag{4.10}$$

$$= \texttt{arg max}_k \left[\log(\hat{\alpha}_k) + \sum_j^J W_{i,j} \log(\hat{\mu}_{k,j}) \right] \tag{4.11}$$

where taking the log improves numerical stability but preserves the maximum.

Naive Bayes uses the training documents to form a prototypical document for each class. When choosing a class, the algorithm weighs the similarity of the new document to each prototype correcting for the rarity of the class. The Naive Bayes classifier is an attractive model to study because it is comparatively simple, performs well and contains several of the elements that will exist in additional methods we will study. Before moving on to a broader list of machine learning classifiers, we make a few remarks about the Naive Bayes method.

The assumptions in Naive Bayes are almost certainly wrong. Our derivation of Naive Bayes started with Bayes Rule and in order to simplify $p(D_i|\pi_{i,k} = 1)$ we made the strong assumption that it could be represented as a class-conditional set of independent multinomial draws. There is a whole series of Naive Bayes methods with different conditional distributions which all assume some form of independence. In general, we believe that all of these assumptions are wrong. Importantly though, the assumption being wrong does not necessarily keep the method from being useful or accurate. A particularly dangerous trap in automated content analysis is dismissing methods out of hand, or spending time to change them, because the assumptions are wrong without assessing the effect of those assumptions.

Why then do we make these assumptions and what effect do they have on our classifier? The advantage of the assumption is that it reduces the number of parameters we need to fit the model down to just $(J-1)K + (K-1)$. This means that the model is more limited in its ability to fit the data — for example, it is unable to capture interactions between words — but, it is also takes relatively little data to fit the model well. The independence assumption does tend to make the Naive Bayes model overconfident in its predictions. What the independence assumption tells the model is that each additional word count is a new piece of information separate from all the previous pieces. In fact, we typically believe that words are "bursty" once a particular word has been used in a text, it is more likely to be used in again. Consequently it is common to see the predictions of a Naive Bayes model be generally correct but where the predicted probabilities are not well calibrated. By this we mean that for a series of documents which are all predicted to be in class k 90% of the time, substantially fewer than 90% are.

One consequence of the independence assumption is that it tends to make Naive Bayes models overconfident. Words that appear together in a document are generally positively correlated and so the Naive Bayes model believes that it has $\sum_j W_{i,j}$ pieces of information about a document, when in fact it has somewhat less. This leads to a situation where the model is often chooses the correct category, but predicted probabilities derived from $\hat{\pi}_i$ are poorly calibrated such that averaging over a group of predictions provides a poor estimate of the total proportion in a given class. Thus when we want to estimate the proportion of documents in class k, these estimates might be somewhat poor.

Naive Bayes is a generative model. In estimating the Naive Bayes classifier our goal was to estimate the probability a document was in a particular class given the contents of the document, $p(\pi_{i,k} = 1|D_i)$. Yet, in the process, we introduced and maximized a joint distribution describing how the model was generated $p(\pi, D_i) = p(D_i|\pi_{i,k} = 1)p(\pi_{i,k} = 1)$. In fact, we can think of this as a mixture model of the type we discussed in Section 3 where some of the components are labeled in advance by the analyst.

One attractive part of generative model is that they provide us a way to build, adapt and interpret algorithms. For example, when introducing the algorithm above we discussed adding a small constant c to deal with the problem of estimating zero probability of seeing some words. We can interpret this process as adding a Bayesian prior where the pseudo-document arises as a consequence of the prior distribution. Similarly, we can address concerns about burstiness in the words by adopting a more complicated counterpart to the multinomial, the Dirichlet compound multinomial (Elkan, 2006). Generative models make great building blocks for more complex models as we already saw with the mixture of multinomials giving rise to latent Dirichlet Allocation and eventually the Structural Topic Model.

However, for classification tasks, there is something a bit odd about generative models. If our interest is in learning $p(Y|D)$ why should we be trying to learn $p(D|Y)$ at all? Can't we just directly maximize $p(Y|D)$ to learn a function mapping from the input to the outcome label without making assumptions about the inputs? In fact, we can and this gives rise to discriminative models, many of which we talk about in the next section. In these models we only condition on the features we use to represent the documents and so we don't need to generate a model for them. This will be familiar to many social and data scientists from regression, where we assume no generative model for the features X we use to predict Y.

The discriminative analog of the Naive Bayes model is multinomial logistic regression (sometimes called logistic regression or the softmax classifier in computer science). The model can be given as:

$$p(\pi_{i,k} = 1|X_i) \propto \exp(X_i\beta) \tag{4.12}$$

where we have switched to X_i to emphasize that they are generic features that need not be word counts. Ng and Jordan (2002) show theoretically and empirically that while logistic regression has higher asymptotic accuracy, Naive Bayes approaches its asymptotic accuracy faster. In other words, for small training samples we should expect Naive Bayes to provide better performance than logistic regression. This makes intuitive sense. The additional assumptions we make to arrive at Naive Bayes provide us with some leverage on the problem, but if we have enough data we can eventually learn all the parameters we need to.

Discriminative models are also useful members of our toolkit because we don't require separate implementations for different types of covariates. In order to add additional information to Naive Bayes, for example the number of upvotes on a social media post, we would need to alter the model in order to specify a generative process for the upvotes. In logistic regression, we can simply add the variable in along with the word counts. While there is a price of convenience, writing down a model of the input variables does confer other benefits, such as making it easier to deal with missing data in the inputs.

Generative and discriminative models don't always exist in a clear dichotomy. A model might specify a generative process for some variables but not for others. Part of Naive Bayes success on small datasets is almost certainly attributable to the generative structure, but it also explains why one seldom sees Naive Bayes in production level systems containing millions of training examples.

Naive Bayes is a Linear Classifier A common thread in the models that we explore in this section is that they can be expressed as models that are linear in some set of features. This means that the classification decision can be expressed as:

$$\hat{y}_i = f\left(\beta_0 + \sum_{j=1}^{J} \beta_j x_{i,j}\right) \tag{4.13}$$

where the function $f()$ converts the linear combination of the features to the output space of the prediction.

By examining the logged form of the Naive Bayes classifier we can see this form:

$$\begin{aligned} \hat{y}_i &= \underbrace{\texttt{arg max}_k}_{f()} \left[\underbrace{\log(\hat{\alpha}_k)}_{\beta_0} + \sum_{j}^{J} \underbrace{W_{i,j}}_{x_{i,j}} \underbrace{\log(\hat{\mu}_{k,j})}_{\beta_j} \right] && (4.14) \\ &= f\left(\beta_0 + \sum_{j=1}^{J} \beta_j x_{i,j}\right) && (4.15) \end{aligned}$$

Linear classifier are quite fast to train and fit and much less prone to overfitting. However, they cannot capture meaningful interactions between the features. This works well in settings where the features that we provide to the algorithm are meaningful and useful on their own. In the next section we will consider a broader set of options that can build better features at the cost of demanding more data.

4.4.2 Machine Learning

Naive Bayes is one of the simplest machine learning models, but there are a host of other options. Here we consider a number of ways to model the conditional density. These methods are general to all kinds of prediction problems with arbitrary inputs and essentially arbitrary outputs.

Essentially all machine learning classifiers can be seen as an example of the following optimization problem (inspired by Mullainathan and Spiess 2017).

$$\texttt{arg min}_\phi \sum_{i=1}^{N} \overbrace{L(\underbrace{f(x_i,\phi)}_{\text{function of } x}, y_i)}^{\text{training loss}} + \underbrace{R(\phi,\lambda)}_{\text{regularization}} \tag{4.16}$$

In words, we minimize the error in our training set (as determined by the loss function $L(\cdot)$) by choosing ϕ which defines a function $f(\cdot)$ subject to a complexity restriction $R(\cdot)$ whose strength is determined by λ. This might seem overwhelming at first, but let's start with simple ordinary least squares model. Here the loss function is the sum of squared errors, the function is a linear predictor and there is no regularizer, such that,

$$L(f(x_i,\phi), y_i) = (y_i - f(x_i,\phi))^2 \tag{4.17}$$

$$f(x_i,\phi) = x_i\phi \tag{4.18}$$

$$R(\phi,\lambda) = 0 \tag{4.19}$$

$$\texttt{arg min}_\phi \sum_{i=1}^{N} (y_i - x_i\phi)^2 \tag{4.20}$$

This model will be familiar to many social scientists. An attractive part of the model is the relatively simplicity. We can think of the model prediction $x_i\phi = \sum_j x_{i,j}\phi_j$ as summing over the features for document i with a different weight, ϕ_j, for each feature $x_{i,j}$. Although we will consider substantially more complicated classes of function $f(\cdot)$ in the models below, we will find that they can often be represented as a linear function on some transformed feature set. In fact, one way to frame these more complex models is that they are searching for the right transformation of the features to make the linear model perform well. We will consider three broad classes of models: linear models with fixed basis functions, kernel methods and adaptive basis function models. These types of models overlap and influence each other quite a bit, but these three categories (along with the ensemble methods considered in the next section) cover the core insight of most of the models in the literature. We focus below on how the model makes predictions and generally gloss over how to estimate the parameters. We refer readers to Murphy (2012) for more on this point.

Fixed Basis Functions Fixed basis function models are the closest to the basic ordinary least squares model that we will consider here. The linear predictor $f(x_i, \phi) = x_i\phi$ implies that the prediction is linear and additive in the predictor variables x_i. However, the model can represent non-linear and interactive functions by including non-linear transformations of the original covariates that we will denote $h(x_i)$. We now have the linear predictor $f(x_i, \phi) = h(x_i)\phi$ which is linear in the parameters, but non-linear in the original predictors!

Many social scientists will likely be familiar with the idea of adding polynomials or interactions as a type of fixed basis function. For example, we could define the non-linear transformation on each predictor as $h(x_{i,j}) = x_{i,j}, x_{i,j}^2, x_{i,j}^3 \dots x_{i,j}^a$ to add up to the a-th polynomial. We could also include interactions such that $h(x_i) = x_{i,j}x_{i,j'}$. This transformation is a basis function expansion of the original predictors. As it turns out, any continuous function can be represented as a linear function on some set of basis functions. This implies that if we only have the right basis functions, we can use a simple linear model to capture any continuous function. The trick is specifying the "right basis functions."

Fixed basis function models are those that commit to a set of basis functions prior to estimation and optimize the weights ϕ. We can think of two types of basis functions: global and local. The coefficient of a global basis function, such as the polynomials discussed above, is affected by data over the entire range of x. An alternative is a local basis function, which models the data within a particular local region of X. Different basis functions are appropriate for modeling certain types of functions. Splines are adept at modeling smooth functions while wavelets can handle sharp discontinuities. Even relatively simple representations of text such as document-feature matrices are high-dimensional. When we start adding basis functions, the dimensionality can explode quickly and we can reach a point where the number of features J exceeds the number of documents N. In this setting, we will no longer be able to estimate the parameters ϕ without some additional kind of constraint. We can add a regularizer $R(\phi, \lambda)$ which expresses a preference for simpler functions.

Different kinds of regularizers yield different kinds of results. The 'ridge' regularizer, $|\phi|_2$ shrinks the parameters smoothly towards zero. By contrast the 'lasso' regularizer $|\phi|_1$ induces sparsity which is set many of the parameters in ϕ exactly to zero. The latter is often helpful in settings where we believe that most features are essentially uninformative about the quantity we are trying to estimate but we don't know which. For example, in predicting the sentiment of a product review, the vast majority of the words might be describing the product itself but carry relatively little information on whether or not the view of the product is positive or negative.

If the parameters are truly sparse, then using a sparsity promoting penalty in the high-dimensional case is uncontroversial. However, in social science and language more broadly we might believe that the world is full of small but generally non-zero effects; that is, we believe the parameter vector is *dense*, not sparse. Hastie, Tibshirani and Friedman (2013) introduce the 'bet on sparsity' principle which is the advice that we "use a procedure that does well in sparse problems, since no procedure does well in dense problems." Regardless of one's philosophical position, sparsity is an important part of the regularization toolkit.

Fixed basis function models with regularization covers an enormous range of machine learning methods and could have a book's worth of material all on their own. The models give us great control over the complexity of the function we learn. These models will function best when the features carry a lot of independent information about the outcome, they will also tend to perform better than alternatives when data is limited. We consider alternatives when we don't have good features below.

2) Kernel Methods [Summary of material to be added: A description of kernel methods as a more flexible alternative to fixed basis functions that must be identified in the past. Explain both the gaussian kernel and string kernel. Clarify that the interpretation here is that we are essentially measuring similarity to the training data and predicting off of that, making kernel methods a kind of template matching. Explain kernel ridge regression and SVM/hinge loss. What is lost in these methods and the ones that follow is simple interpretation of what is doing the work in the prediction.]

3) Adaptive Basis Functions [Summary of material to be added: The third option is to inductively learn the basis expansion $h(x_i)$. This moves the feature engineering choices from the analyst to the algorithm. This typically requires substantially more data and is less efficient than a correctly specified fixed basis function (no free lunch). We present trees as a basic form of an adaptive basis function which induce a particular type of partition in the feature space that lends itself to simple interactions and dichotomization. Present neural networks and their deep variants as a linear classifier with adaptive basis function by demonstrating how the final layer is multinomial logistic regression and everything to that point is simply creating the feature space for it. Include a brief discussion of Long Short Term Memory networks and how they are able to more easily incorporate elements of the feature space like word order.]

A Note About Features [Summary of material to be added: The more we have to learn the features from the data, the more data we are going to need. Easy to get lost in the weeds though- if we have great features be sure to find a way to use them. One of the biggest trends of late aughts early teens is that big enough datasets came online that in many areas these approaches that engineer features were able to outstrip hand-engineered features. Still doing so doesn't always preserve interpretability.]

4.4.3 Ensembles

[Summary of material to be added: Ensembles and how they work in general. Bagging → Random Forests, Boosting → Gradient Boosted Machines. A brief discussion of the Superlearner and its properties.]

4.5 Supervised Learning with Non-Random Samples

[Summary of material to be added: The section above implicitly assumed random samples, but random samples can be incredibly to get in practice. If we are in the streaming setting where things are measured over time, random samples are hard. Even when we can take a random sample, rare events pose a substantial problem where we may not be able to get enough instances of our class of interest. Thus we outline a few methods for dealing with non-random examples.]

4.5.1 Active Learning

[Summary of material to be added: A short discussion of active learning and its application to supervised learning.]

4.5.2 Stratified Sampling Methods

[Summary of material to be added: A short discussion of stratified sampling methods including the factor optimal design of Taddy (2013).]

4.5.3 ReadMe

[Summary of material to be added: A short discussion of the Hopkins and King (2010) method with a particular focus on the assumptions that you have randomly sampled within class. Include the recent update by Jerzak, King and Strezhnev (2018) and other contributions in the literature on proportion estimation.]

5 Methods for Known Categories without Human Coding

Thus far, we've introduced two general ways to provide "supervision" when obtaining a g function, directly write-down the g function as with dictionary based methods or learn g from a sample of labeled training data. In this section, we explore other approaches to providing supervision. This is a very active area of research because generating training data can be extremely expensive. Thus, the holy grail would be a method which could generate accurate classifications from a pre-determined system without needing human-coded documents. Of course, the information to do the classification has to come from somewhere. The methods in this section explore some clever sources of information that an analyst might have access to.

In Richard Nielsen's book *Deadly Clerics: Blocked Ambition and the Paths to Jihad* he explores the questions of why some Islamic clerics start to express support for violent Jihad (Nielsen, 2017). He argues that a major driver of a turn towards an ideology that supports violence is blocked ambition — circumstances where the cleric is shut out of prominent state jobs effectively thwarting their goals. To build evidence on this question, he collected the first census of islamic clerics of representation on the web. Statistical text analysis methods are used throughout Nielsen (2017) in a variety of insightful and clever ways, but one of the central applications is assessing whether clerics support violent Jihad.

There is some valuable, but non-systematic, sources of information about cleric ideology including biographies, endorsements — for example by Ayman al-Zawahiri, the head of al-Qaeda, and prior academic texts which provide lists of known jihadist clerics. However, most of these sources capture only the most prominent examples and don't provide information on the majority of individuals. Instead Nielsen (2017) uses the writings that clerics place online including their articles, books, sermons and fatwas (fatwas are legal rulings in Islam). These texts capture their *expressed ideology* (as opposed to privately held beliefs) which are the object of focus here.

Nielsen (2017) aggregates all the texts for a particular cleric together and estimates a 'Jihad Score' which is a continuous measure of the support for violent Jihad. The underlying model is closely connected to the Naive Bayes (NB) model we explained in depth above and is similar in style to Wordscores (Laver, Benoit and Garry, 2003). Under the NB model assumptions, the score for document i is calculated by as the log-odds of being in the Jihad class assuming equal prior odds for Jihad and not (Beauchamp, 2011; Perry and Benoit, 2017). The score for document i is

$$\text{Jihad Score}_i = \frac{1}{\sum_j w_{i,j}} \sum_j w_{i,j} \text{log} \left(\frac{\mu_{j,1}}{\mu_{j,0}} \right) \tag{5.1}$$

where $\mu_{j,\cdot}$ indicates the probability of seeing word j given the document is either from a jihadist ($\mu_{j,1}$) or not ($\mu_{j,0}$). When the probability of seeing the observed word is higher in a

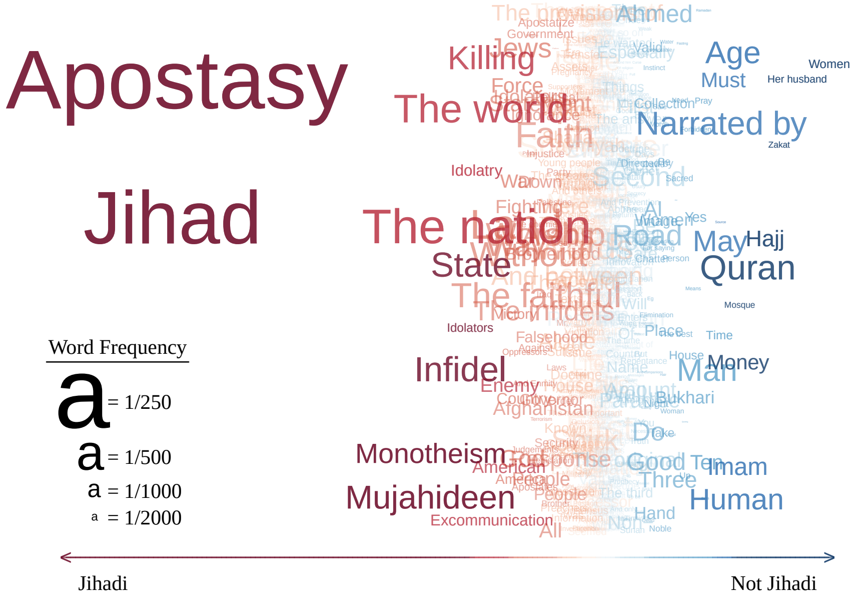


Figure 6: Plot of weights on words with english translations. Figure 5.2 from Nielsen (2017).

jihadist document the score contribution for that word will be positive, when the probability of seeing the observed word is higher in a non-jihadist document the score contribution will be lower. Thus the log-ratio of probabilities acts like a weight for each word which is then simply summed up. These weights can be examined such as in Figure 6.

To estimate the probabilities of seeing each word under jihadist texts, Nielsen (2017) uses the ‘The Jihadist’s Bookbag’ a collection of documents designed to provide an introduction to jihadist ideology. To get the probabilities of seeing each word under a non-jihadist text, Nielsen used a curated set of documents from known non-jihadist clerics. With these key parameters estimated, he forms the scores for each cleric in his dataset and validates them in a number of different ways. With this validate measure of expressed ideology in hand, he provides support for his main hypothesis that blocked ambition is a significant cause of clerics’ turning to support violence.

5.1 Supervision of Convenience

The average cleric in Nielsen's sample has 523 documents. This makes human coding even a small random sample of clerics for jihadist ideology prohibitively difficult. This difficulty renders the use of the methods described in Section 4, which are all based on human coding in some respect, impractical. Nielsen instead turns to a ready sample of documents chosen by Jihadists themselves as representative of Jihad and a curated sample from known non-jihadist clerics.

This approach is an example of what we call the 'supervision of convenience' which involves using found data in order to pin down the categories. This can be a low-cost strategy to producing a classifier that often works surprisingly well in practice. The complication is that this procedure can bias your measures in unpredictable directions for at least two reasons. First, the labels haven't been assigned by the analyst and thus come from an unknown process. The analyst needs to ensure that the contents of those documents are reflective of the categories they wish to code. Even granted that the codes for the found data are as the analyst would have assigned them, the second concern is that the documents were non-randomly selected and thus the estimated probabilities of words given latent category are unlikely to be accurate. This is particularly true in practice, as supervision of convenience often arises by choosing extreme cases in one direction or the other where language patterns might be quite distinct.

Nielsen (2017) is aware of these issues and provides nine-distinct validations designed to establish the validity of his scores. Much like other assumptions such as the bag-of-words or generative models of text are wrong, treated supervision of convenience like it is a random sample is an incorrect assumption which compels us towards more careful validation. However, like many other cases, it may work well in practice.

5.2 Predictability as Measurement

A second analytic move in Nielsen (2017) worthy of consideration is the use of predictability as measurement. This is a common approach in scaling methods where humans would have a difficult time creating a gold standard judgment of the exact continuous value to assign to some underlying trait. This is closely related to the fictitious prediction problem which we outlined in Chapter 4, where we use a *prediction* task for which prediction is never of interest to measure the *intensity* of some dimension.

Of course in the case of Nielsen (2017), the prediction isn't fictitious at all: it would actually be useful to classify clerics into jihadist and not. While Nielsen (2017) is generally careful to treat the jihad scores as an indicator of how likely it is that the individual supports violent jihad, he does at time seem to suggest that this also measures intensity of support. Others are more direct in making the claim that the prediction of a class provides a measurement of intensity along a dimension such as ideology (Beauchamp, 2011) and polarization (Gentzkow, Shapiro and Taddy, 2016). In these tasks the predictability of the outcome is a surrogate function for the underlying dimension of interest.

Class Affinity Model Aware of this tension, Perry and Benoit (2017) identify an approach to address what they call "a fundamental disconnect between the classification philosophy and the goals of scaling". They write that

> Instead of predicting class membership, our objective in such problems is to *scale a continuous characteristic*, through measuring the fit of a text to a set of known classes based on its degree of similarity to typical texts from these classes. (emphasis theirs)

To address these concerns they introduce the Class Affinity Model. The affinity model is motivated by the mixed-membership generative model

$$\begin{aligned} U_{i,m} &\sim \text{Categorical}(\pi_i) \\ \pi_i &= E\left[\frac{1}{M}\sum_{m=1}^{M} Z_m\right] \\ w_i &\sim \text{Multinomial}(M, \sum_{k=1}^{K} \pi_{i,k}\mu_k) \end{aligned}$$

where $U_{i,m}$ is the categorical affiliation for each token, π_i is the document affinity represented as proportional allocation to each category, and the observed word vector is drawn based on the category conditional distribution over words μ_k. The category specific distributions, μ_k, are estimated using 'clearly polar examples of each reference class' (Perry and Benoit, 2017). Perry and Benoit (2017) also offers a series of diagnostics, comparisons to existing methods, methods to obtain uncertainty and regularized estimation approaches.

This model is closely related to two models we have already seen. It is the LDA model without the Bayesian priors, except that μ_k are estimated with supervision. It is also the Naive Bayes model where the document can have mixed membership.

The class affinity model provides a theoretical foundation for models which both rely on the supervision of convenience and predictability as measurement. The model still uses extreme texts as in supervision of convenience, but gives a clearer interpretation of the result the model is providing — the mixture of the extreme class distributions which best approximates the document. Perry and Benoit (2017) also acknowledges the limitations of prediction as a measurement. We caution however that while the philosophical foundations are clearer, the affinity measure on dimension k is still about the best convex combination of extreme class distributions which approximates the given document. Perry and Benoit (2017) have done a great service by being clear about what their dimension is and analysts need to be equally clear about what they want the measure to be so that they can assess whether the method is appropriate for their case.

Accuracy as the Quantity of Interest Peterson and Spirling (2018) also tackle the question of predictability as a measure. In contrast to Perry and Benoit (2017), they embrace the idea and give a defense of measuring political polarization by the accuracy of classifying people into parties. Thus the argue that when House of Commons speeches cannot be used to accurately predict the party of the speaker, polarization is low. This provides a measure of polarization in settings where voting data is not available or informative and allows it to be traced over time. This suggests an alternate way of thinking about predictability, not as prediction of a particular class as a continuous measure of some latent dimension, but the accuracy aggregated over a period as a measure of separation between two sets of speech. Importantly this implies that any changes of accuracy are a result of polarization, suggesting the assumption that conditional on polarization, the method requires constant predictability of text over time. This will work in some settings more than others, but Peterson and Spirling (2018) provides clear guidance in their paper and online appendix.

Supervision of convenience and prediction problems as measurement can both be useful tools. However both require careful validation and deep consideration of the measure that the analyst is trying to construct.

5.3 Feature-Based Supervision

In the examples of the last two subsections, we are still soliciting information from the analyst through documents that are labeled in some way. A recently developed strategy is for the analyst to provide expert knowledge about the decision-making process (like the rule/dictionary based methods) but still learn the parameters of a classifier. These approaches are quite new at the time of writing, but are an important direction for where these methods might go in the future.

5.3.1 Anchor Methods

We already saw previously with Naive Bayes and Multinomial Inverse Regression that generative models can increase efficiency when their assumptions hold. It turns out that the same conditional independence assumption can be combined with certain assumptions about the document features to produce a classifier without any training data at all. These methods rely on strong assumptions and have not yet been widely applied to text, but we cover them here because they suggest a possible way forward for analysts to specify categories with an information source that is different than randomly coded documents.

Electronic medical records have become an important source of information for targeting medical interventions to patients and providing clinical decision support. Electronic medical records are similar to documents in that they can be formulated as set of discrete features, albeit with somewhat lower dimensionality than text. The goal is to infer certain patient phenotypes which are similar to the latent categories that we try to classify documents into. Thus just as in the text analysis case we are trying to use a relatively high-dimensional discrete feature set to predict some latent class for each observation. The problem is that gold-standard data is difficult to come by and systems that require separate training by institution are not feasible in practice.

[Summary of material to be added: A short description of the anchor and learn framework of Halpern et al. (2014) and further developed in (Halpern, Horng and Sontag, 2015; Halpern et al., 2016; Joshi et al., 2016). Based on previous work on positive only labels by Elkan and Noto (2008). The core idea is that an anchor provides an unambiguous signal of the latent class when it is present and is thus like a noisy label. This can be combined with a conditional independence assumption to learn the entire classifier. Discovered anchors have been used in the text as a key component of one spectral approach to learning topic models

(Arora et al., 2013) and have been further applied to supervised and unsupervised settings by: Reing et al. (2016) and Nguyen et al. (2015). Lund et al. (2017) extends to the case of multiword anchors.]

5.3.2 Distant and Weak Supervision

[Summary of material to be added: A description of distant supervision, an old idea for generating training samples based on weak rules. Connect to recent work by Ratner et al. (2017) out of Chris Ré's lab on models to denoise training data generated in this manner. These approaches still create training sets but essentially via rules that are specified by the analyst making them somewhat more like anchor-methods than supervised learning with non-random sampling.]

6 Beyond Subject Matter

[Summary of material to be added: Many of the methods above are implicitly geared towards broad properties in text such as topic matter. Here we outline a few more subtle things we can measure in text and their applications in measurement. A natural piece we omit is stylometry which we treat as a prediction problem and put in Chapter 7.]

6.1 Complexity

[Summary of material to be added: A discussion of measuring complexity from text. A short explanation of typical readability scores and their limits. A discussion of recent work by Benoit, Munger and Spirling (2017). Highlight applications in Hengel (2017) and Spirling (2016).]

6.2 Sentiment Analysis

[Summary of material to be added: A separate, explicit discussion of sentiment analysis. The tools in the previous sections can be used for sentiment analysis, but sentiment can be complex and it is often helpful to have methods which can capture more nuance like negation and contextual meaning. Outline the problem and its context (Pang and Lee, 2008). Discuss the work by Socher et al. (2013) on deep neural networks and the way the collection of the sentiment treebank enabled that work.]

6.3 Text Reuse

[Summary of material to be added: Discuss applications of plagiarism detection. Explain the core intuition for how it works. Because the relevance of this may not be intuitive to many social scientists give applications include Bail (2012), Grimmer (2013*b*) and Wilkerson, Smith and Stramp (2015).]

6.4 Network Approaches to Text

As we described above, bag-of-words latent variable models implicitly use word co-occurence patterns to discover a sense of textual meaning. We can represent these co-occurences explicitly as a weighted undirected network (where each word is a node and each edge represents their co-occurence within a unit of text). This word network can be analyzed using the insights of the rich statistical literature on social network analysis (Carley, 1990, 1997*b*,*a*). For example, Rule, Cointet and Bearman (2015) leverages a community detection algorithm and numerous approaches to network visualization in order to study the evolution of discourse in U.S. State of the Union addresses from 1790 to 2014. Community detection algorithms are closely related to the clustering algorithms described in the chapter on discovery, but here they are applied to the vocabulary creating groupings of terms rather than groupings of documents.

In the examples we have given thus far, the quantity of interest has been the document or a collection of documents rather than the vocabulary terms themselves. This is also compatible with a network-based approach to text analysis. Bail (2016) studies how advocacy organizations capture public attention. Drawing on the network analysis literature, Bail (2016) hypothesizes that organizations which can create cultural bridges by connecting typically disparate themes can more easily connect with new audiences. Using a dataset of advocacy organizations for issues around autism spectrum disorder, he creates a network where each advocacy organization is a node and the weighted edges between them indicate overlapping use of nouns and noun phrases. He then measures cultural bridges by betweenness centrality. An organization with high betweenness centrality has a large number of shortest paths between organizations going through it. He finds that organizations which engage in substantial cultural bridging get two and a half times as many comments from new social media users. It is not only the particular method, but the broader analytic approach that is defined by the network perspective.

Network analysis approaches adopt a relational approach to meaning. However, while not as explicit in the modeling, the relational idea of language is also central to the understanding of how LDA and other topic models function (DiMaggio, Nag and Blei, 2013). All of these models rely on the correlational structure in the words in order to capture semantic meaning. While the network approaches rely only on the co-occurences between words, LDA implicitly uses word triples and other high order tuples as well. This can lead to richer models when substantial amounts of data are available but also more difficult estimation problems (as we discuss in more detail below).

The network analysis perspective also brings a distinct theoretical background to text analysis. Social network analysis has a well-developed theoretical and methodological toolkit, some of which can be ported to understand text. Still, just as with all the methods we have discussed, the theoretical background does not guarantee that the methods will be successful in any particular case, resulting in the need to validate.

6.5 Multiplicative Latent Variable Models

The latent variable models we described above are relatively straightforward generative models of text. For example, the mixed membership topic models have an additive form where the distribution for a given document is a convex combination (i.e. a weighted average with non-negative weights that sum to one) of the topic specific distribution. A limitation of this kind of model is that the model is unable to predict or adequately explain a document which is more extreme than the individual topics.[7] In practice this means that while a document can be represented by multiple topic, the topics cannot interact to produce a more precise prediction given by their overlap. If we have a model of scholarly disciplines and we represent a document as a mixture of the 'humanities' and 'computer science', we cannot get a higher probability of the phrase 'digital humanities' than either of the original topics would give alone.

We can get around this limitation by using models which forego the latent additive model infrastructure and instead multiply topic features together and renormalize. This can provide a much more flexible model.[8] Since the terms are multiplied together if the loading of a topic on a given word is 1, it will be as though the topic wasn't there. If the values is much larg Hinton and Salakhutdinov (2009) introduce the Replicated Softmax model which

[7] Any convex combination lies on the interior of the convex hull. This means the document representation in a model like LDA cannot predict a word distribution which is sharper than the individual topics (Hinton and Salakhutdinov, 2009).

[8] The flexibility comes because each topic can enforce a constraint on a particular word (or not). Imagine a multiplication over a set of topic terms for some feature j such that the probability of a word in a document up to normalization over the features is $p(w_{i,j}) \propto \prod_{k=1}^{K} \mu_{k,j}^{\pi_{i,k}}$ where $\pi_{i,k}$ is a one if the topic is present in that document and zero otherwise. Since the terms $\mu_{k,j}$ are multiplied together, if the loading for a given topic is approximately 1, it will be as though the topic wasn't there (up to the normalization over features). Values far from one can enforce a constraint that feature j is always there ($\mu_{k,j} \gg 1$) or never there ($\mu_{k,j} \ll 1$). See Murphy (2012) Chapter 27.

represents each document with some number of binary features which indicates which topics it contains.[9] The document fit is then estimated by multiplying the topic-word features together for those documents and then renormalizing. This allows meaningful interactions between the topics such that a document formed from the combination of the ‘computer science’ and ‘humanities’ features can look substantially different than simply combining two separate documents, one from ‘computer science’ and one from ‘humanities’, together. Many newer flexible approaches to modeling text have this feature including those based on deep neural network infrastructures (Gan et al., 2015; Ranganath et al., 2015), the core ideas of which we explain in more detail in the section on supervised learning.

Multiplicative latent variable models are much more flexible models than the basic LDA topic model. However, this flexibility also makes them correspondingly more difficult to interpret. To continue our example above, consider examining the ‘computer science’ topic from an LDA model using the words which are most probable under that topic. Under the model, these words will always have high probability for any document with a large ‘computer science’ component. Under the multiplicative latent variable models, the high probability words are contingent on the interaction between all the topics present in a document. This is a fantastic amount of expressive power, but also makes it very difficult to contemplate how one would easily evaluate all those interactions.

Some models have made use of multiplicative latent variable structures by limiting complexity in some way to make intepretation more tractable. Paul and Dredze (2012) limit the number of factors interacting through a sparsity constraint. Eisenstein, Ahmed and Xing (2011) limit the types of interactions. Multidimensional extensions of the scaling models like Wordfish that we described above are also multiplicative in this way, but simplify by having only a very small number of dimensions.

[9] This is a form of a Restricted Boltzmann machine designed particularly for categorical data. The model has been extended in a number of directions including to make use of word order (e.g. Larochelle and Lauly 2012)

6.6 Information Extraction

[Summary of material to be added: A very brief review of a very broad area including entity extraction, role labeling, template filling and events data.]

7 Integrating Measurement into the Research Process

Having reviewed a number of approaches to generating g functions we provide an overview of a few related concepts including choosing a particular method, sample splitting, labeling concepts and performing analysis with the learned measures.

7.1 Choosing a Method

No single method of measurement is going to be right for all settings— the choice will always be specific to the context. In this section, we try to offer some broad, informal guidance about when certain methods will be the most useful. The first high-level choice is whether you want to repurpose a discovery method or use one of the methods for known categories or loosely, a choice between a more unsupervised or a more supervised approach. Asking yourself the following questions can help push you in the right direction.

Do you have a sharply defined idea of the dimensions you want to measure? If you do, you likely would prefer to use on e of the methods for known categories because you can more precisely shape the measure to approximate your vision. It can be unpleasant to try to coerce a discovery method to produce the concepts you want. If you have a broad idea of the dimensions you are interested in but lack specifics, you may want to start with the discovery methods, at least initially.

Are you interested in the entire contents of a document or just a part of it? By construction, most of the discovery methods have to create a latent representation which explains all the observed features.[10] This means that discovery methods can be ill-suited to explaining one aspect of a document such as the sentiment of a movie review, because some part of that learned representation has to explain other features such as the type of movie and the summary of the plot. This can sometimes be alleviated by careful selection of the feature set (for example removing all words from the movie review that do not convey sentiment) but this requires a strong *a priori* set of knowledge. By contrast, many of the methods for known categories can use only the components of the features that algorithm needs to explain the labeled documents making them easier to use in these settings.

Do you have a large number of dimensions/categories? When the number of dimensions under consideration is very large it can be difficult for a human to enumerate all the possibilities and provide enough training examples to use a classical supervised technique with high accuracy. The exact boundaries of "a large number" are application specific, but supervised learning is often effective with 2-7 dimensions while 100 or more can be extremely difficult to handle. However, unsupervised methods often work extremely effectively at 100 dimensions or more and can produce conceptually-precise, meaningful categories at this scale.

These questions and our presentation of methods has been organized around grouping methods which have a similar mechanism for incorporating information from the analyst into the measurement process. When reading the broader literature, there is another difference in choosing a method that might seem important, but we argue is not: whether the model is motivated from a stochastic data generating process (e.g. Latent Dirichlet Allocation) or an algorithmic perspective (like k-means). These differences are very salient when trying to move between different literature, but the differences in language often mask the similarities

[10] Analogously they assess distance between two documents based on all observed features.

in how the methods operate.

Breiman et al. (2001) describes these two cultures for modeling data in the following way,

1. The Stochastic Data Model Culture:

   The model arises from a generative process which we use inference to reverse. New methods here are typically developed and motivated by writing down a probabilistic model and finding parameters that best fit the data given that model.

2. The Algorithmic Model Culture:

   In this culture we treat the data as generated from a completely unknown process and we simply want to find a predictive function that minimizes some loss function.

Breiman et al. (2001) wonderfully details some of the subtle differences between these two cultures for a statistical audience that has primarily been dominated by the stochastic data culture. While some approaches are more amenable to one view or another, there is often a direct corollary between the two approaches. Take for example the most striking difference the objective function. Algorithmic methods are often designed to minimize a certain distance/loss function. For example, Ordinary Least Squares regression (OLS) minimizes the sum of the squared residuals which is closely related to the Euclidean distance between the observed data and our prediction. Stochastic methods tend to maximize the fit to a prespecified probability distribution. Yet these two things are very closely linked. In fact, there is a bijection (i.e. a one-to-one correspondence) between exponential family distributions which includes most probability distributions people use, and the class of regular Bregman divergences which includes most distance metrics people use (Banerjee et al., 2005).[11] Thus, for many methods, there is both an algorithmic and stochastic cultural interpretation and

[11] A divergence is a weaker concept than a distance as it need not be symmetric. Some Bregman divergences include: squared loss, logistic loss, squared euclidean distance, Mahalanobis distance, KL-divergence. Exponential family distributions include: Gaussian, Poisson, Bernoulli, Binomial, Exponential, von Mises-Fisher distribution.

presentation. Given the similarities we think the more important distinction to focus on is what information the algorithm uses to identify the measure.

7.2 Sample Splits

It is difficult to do the work of articulating the conceptualization and measuring it at the same time. Humans are human after all, and we might worry about allowing our biases or theoretical expectations to creep into our measurement or interpretation. Concerns about this kind of subjective measurement error are why interviewers will sometimes be sent in blind to key expectations about a subject or coders will not see the outcomes when doing the coding. If we want to create a measure of some property in the text, we want to minimize the degree to which our measurement is affected by something outside the text.

In describing methods to avoid this kind of measurement error, King, Keohane and Verba (1994) counsel,

> Our advice in these circumstances is, first, to try to use judgments made for entirely different purposes by *other researchers*. This element of arbitrariness in qualitative or quantitative measurement guarantees that the measure will not be influenced by your hypotheses, which presumably were not formed until later . . . If you are the first person to use a set of variables, it is helpful to let *other informed people* code your variables without knowing your theory of the relationship you wish to evaluate. (156)

Text analysis lends itself to measures that are not available in other settings, so it might be useful to separate the coding from other outcomes or covariates. One way to address this problem is by a train-test split.

By developing the measure in the training set and then applying the measure in the test set, we can protect ourselves from certain kinds of overfitting. As we will show in the text chapter, sample splits are important for causal inference, but we think they can also be helpful for descriptive inference. When using a sample split it is important to validate in the training set while developing the label but also in the test set. This allows us to ensure that measure is still capturing what we intend it to capture.

7.3 Performing Analysis and Uncertainty

Generally the goal of measuring some quantity is to make either a descriptive, causal or predictive inference. All these techniques generally rely on some notion of means or conditional means. However, it should not be taken for granted that just because we can measure something means that the average across measurements is meaningful. This is a point on which much work, in text and elsewhere, has been fairly cavalier. For example, it is potentially problematic to report the mean of a Likert scaled variable because the categories are not equal width apart. This is an issue with which the broader field may need to grapple at some point.

[Summary of material to be added: A short explanation of properties necessary for analysis with measurements — drawing from the psychometric literature. For application consider the debate on manifesto positions and the psychophysical foundations for certain representations used by Lowe et al. (2011).]

[Summary of material to be added: A description of the problems with measuring and incorporating uncertainty in analysis. The core problem is that it is hard enough to get accurate uncertainty estimates of the latent quantity, it is even harder to incorporate into analysis. Discuss the general concept of uncertainty in the Bayesian models via approximations to the posterior (both variational and gibbs) including cases where the underlying quantity of interest is directly represented in the model as in Grimmer (2010). Overview bootstrap approaches to uncertainty (Benoit, Laver and Mikhaylov, 2009; Lowe et al., 2011; Lowe and Benoit, 2013). Unfortunately it isn't super clear what is to be done with this uncertainty. Review approaches in Stewart and Zhukov (2009); Roberts et al. (2014); Fong and Tyler (2017); Grimmer, King and Superti (2018) and their limitations.]

8 Validation

[Summary of material to be added: The argumentation below is fairly complete, but we intend to include more worked examples of specific validations throughout this section.]

Validation is a concept to which we continually return and in this section we outline some of the ways to validate your findings. All measurement is about compression and the process of validation is ensuring that we have correctly understood what we compressed and what information we discarded. There are myriad validation techniques and no single clear organizing framework. However there is one underlying principle that unites validation techniques: vulnerability to being proven wrong. By repeatedly exposing ourselves to being proven wrong we start to build a base evidence that can be convincing both to ourselves and our readers.

No amount of validation can completely assure as we are correct and different settings will call for different levels of certainty and thus, different levels of validation. In all circumstances though, we emphasize trying to build in an evaluation loop- a regular process of validating the measure at regular intervals throughout the research process and over time (for ongoing measurements). This provides us a a regular, external check on our model. What is hopefully clear, is that we do not believe any text model is self-justifying and thus we should avoid the blind use of any method without a validation step.

What does it mean to validate? Essentially it means to show that we are right about what a measure is capturing about the world. In measurement we are implicitly invoking an assumption of concept homogeneity (Goertz, 2008; Przeworski, Teune et al., 1970; Gerring and Thomas, 2005) which states that two observations assigned the same value of Y are equivalent with respect to the conceptual dimension the measure is seeking to recover. We are attempting to evaluate both that the observations assigned a particular value are capturing what we theoretically claim (label fidelity) and that the observations are conceptually equivalent in the above sense.

8.1 Validation with Gold-standard Data

We start by setting out the infrastructure for the simplest case. This case is likely to never exist in practice and thus for most projects, some validations from the next section will also be required. However, this simple setup allows us to establish a useful baseline. We assume that we are in the canonical setting for supervised learning which contains four components: a set of categories which are mutually exclusive and exhaustive, a set of documents with gold-standard labels for those categories which are randomly sampled from the population of interest, a set of unlabeled documents we wish to measure and a way of obtaining a g function which extrapolates from the gold-standard labels to unlabeled documents. By gold-standard we mean that the labels are correct and of the form that we want. We can evaluate the accuracy of the system relative to the gold-standard labels.

What accuracy means is relative to the analysts specific quantity of interest. Ideally there is some specific loss function for the given application. For example, consider the spam filter for your email — in this case it only matters that documents are correctly sorted and it is more costly to put an important message in the spam folder than for a piece of spam to filter into the inbox. By contrast, many tasks in the social sciences are trying to 'characterize the haystack' and measure the proportion of documents in a given category for a given stratum (such as a time period or author). These two applications have different *loss functions* and performance that might be desirable in a spam filter might provide systematically biased estimate of the proportion of spam in a person's inbox.

We can't choose your loss function for you. Below we provide some ways of summarizing performance, but the best summary of classifier performance is going to be application specific.

The Idealized Case Imagine an idealized case where the gold-standard labeled dataset gets asymptotically large. We can divide our gold-standard documents into a train and test set. The training documents can be used to obtain g and we can evaluate accuracy on the test set. The performance of the model can be directly compared to the distribution of gold-standard documents in the test set. We make this comparison using a *confusion matrix*: a $K \times K$ cross-tabulation where the rows describe predictions and the columns characterize the gold-standard codes, where each cell describes the number of documents that received that classification from each. In the simplified setting where the model produces a predicted class for a new document and no other information (e.g. a predicted probability), the confusion matrix a straightforward summary of the typically most relevant performance information.

Table 1: Example of Confusion Matrix

	C_1	C_2	C_3	C_4	Machine
C_1	$\mathrm{N}_{1,1}$	$\mathrm{N}_{1,2}$	$\mathrm{N}_{1,3}$	$\mathrm{N}_{1,4}$	$\mathrm{N}_{1,}$
C_2	$\mathrm{N}_{2,1}$	$\mathrm{N}_{2,2}$	$\mathrm{N}_{2,3}$	$\mathrm{N}_{2,4}$	$\mathrm{N}_{2,}$
C_3	$\mathrm{N}_{3,1}$	$\mathrm{N}_{3,2}$	$\mathrm{N}_{3,3}$	$\mathrm{N}_{3,4}$	$\mathrm{N}_{3,}$
C_4	$\mathrm{N}_{4,1}$	$\mathrm{N}_{4,2}$	$\mathrm{N}_{4,3}$	$\mathrm{N}_{4,4}$	$\mathrm{N}_{4,}$
Human	$\mathrm{N}_{,1}$	$\mathrm{N}_{,2}$	$\mathrm{N}_{,3}$	$\mathrm{N}_{,4}$	N

For example, $\mathrm{N}_{1,2}$ describes the number of documents an automated method coded as a 1 that the gold-standard labeled a 2. In general, if the on-diagonal elements of Table 1 are large, then the method is performing well.

Directly using confusion matrices to evaluate model performance can be difficult–so it is standard to summarize the matrices. The three most prominent summaries used are: (1) *Accuracy*–the proportion of correctly classified documents, (2) *Precision* for category k–the number of documents correctly classified into category k, divided by the total number of documents classified as category k, and (3) *Recall* for category k–the number of correctly classified category k documents divided by the number of gold-standard documents in category k.

$$\text{Accuracy} = \frac{N_{1,1} + N_{2,2} + N_{3,3} + N_{4,4}}{N} \; ; \; \text{Precision}_k \quad = \quad \frac{N_{k,k}}{N_{k,}} \; ; \; \text{Recall}_k = \frac{N_{k,k}}{N_{,k}}$$

Instead of a confusion matrix and the associated criteria, we can use the test set to estimate any loss function of interest.

[Summary of material to be added: Include some alternative loss functions in specific projects.]

Cross-validation This approach to validation is difficult to apply in most settings because our supply of gold-standard labeled documents is not effectively infinite. Even in settings where we have a large test set, we may want to make decisions about the appropriate model in the training set and thus need to divide it further. In these settings, *cross-validation* can be used to replicate the ideal procedure (Efron and Gong, 1983; Hastie, Tibshirani and Friedman, 2001). In V-fold cross validation, the training set is randomly partitioned into V ($v = 1 \ldots, V$) groups. The model's performance is assessed on each of the groups, ensuring all predictions are made on data out of sample. For each group v, the model is trained on the $V - 1$ other groups, then applied to the V^{th} group to assess performance. Cross-validation is extremely powerful–it avoids overfitting by focusing on out of sample prediction and selects the best model for the underlying data from a set of candidate models (this is known as the Oracle property) (van der Vaart, Dudoit and van der Laan, 2006).[12]

While cross-validation is an excellent method of approximating out-of-sample fit, this

[12] These properties only apply when all steps (including selection of relevant predictors) is handled within the training set of the cross-validation and not within the full labelled data. See Section 7.10.2 of Hastie, Tibshirani and Friedman 2001 for more information on the proper application of cross-validation.

only provides an accurate estimate of the prediction error of interest if the gold standard documents are a random sample from the population of interest. If for example, the goal is to categorize internet news posts today from a training set of news posts from 1993 we are unlikely to accurately estimate our error. While this example may seem far-fetched, many articles attempt to build systems which will classify documents that have not yet been written. In cases where there is conceptual drift or some other change in the data generating process, we will no longer have an accurate estimate of the error.

The importance of gold-standard data The above examples assume that we have gold-standard data which is perfectly accurate but this is often incorrect. Human coding is often thought about as the gold-standard, but on difficult tasks humans can disagree all the time. Standard practice in these settings is to iteratively train coders to have high-levels of intercoder reliability. Unfortunately, this is not sufficient to establish the accuracy of the resulting coding.

[Summary of material to be added: A short description of the key findings of Grimmer, King and Superti (2018) on bias corrections and bounds for disagreements among human coders.]

Ongoing Evaluations In many large projects, data is designed to be continually collected over a long period of time. In these settings we strongly recommend building in habitual evaluation of classifier performance into maintenance of the system. Changes over even relatively short periods of time can dramatically change performance. Consider a classifier designed to assess sentiment on Twitter. Not only is language rapidly evolving but so is the population of people using the platforms. These kinds of changes are not accounted for in the original accuracy estimate.

8.2 Validation without Gold-standard Data

When gold-standard data does not exist prior to the start of measurement, the process is somewhat harder. We might lack gold-standard data for any number of reasons — the most obvious setting being the use of unsupervised methods. However, many methods we think of as supervised don't come with gold-standard data either: e.g. all kinds of scaling, dictionary methods, and complexity measures. This is okay — many of these unsupervised and supervised methods shift the burden of human effort from the initial stages of labeling documents to the later stages of labeling and validation. We also emphasize that even when gold-standard data is available, the evaluations below can still be useful for evaluating the behavior of the classifier.

We divide evaluations into four types: surrogate labels, non-expert human evaluation, model-based assessments and correspondence to external information. These categories are neither mutually exclusive, nor exhaustive, and they are all designed to supplement *expert human evaluation*, although such evaluation is quite difficult to talk about in the abstract. Each of these four supplementary approaches is situational and generally the best validations are tailored to the application.

8.2.1 Surrogate Labels

While we may not have gold-standard data, we still may be able to collect surrogate labels which are a close enough approximation to let us know that we are on the right track. In the psychometric literature this often falls under the broader category of *convergent validity*, establishing that our measure accords with other relevant measures.

Approximate Labels In many circumstances we may not have labels of the type we need for a proper evaluation set because they are approximate in some way. For example, in approaches that produce a non-categorical representation (such as scaling, topic models and dictionaries), it may be more tractable to collect labels that are coarsened versions of the gold-standard label. For example, in a topic model of newspaper articles, an analyst may be able to assess whether or not an article is substantially about the 'environment,' but not that the environment accounts for .18 of the tokens in the document. Thus the data isn't really gold-standard because it is not in the form of the latent representation we want, but it can still be incredibly useful for assessing accuracy.

These labels can be systematically collected after the fact, but they can also be types of found data. In analyzing Senate press releases, Grimmer and King (2011) use coarse topics assigned by press secretaries. As we describe in more detail below, Nielsen (2017) uses lists of jihadists and non-jihadists as validations of his jihad scores. Often in looking for initial measures, analysts may run across approximate and possibly non-systematic measures which can serve as validation exercises for the newly created measure.

Approximate labels can be an important part of the validation process but they aren't gold-standard data. We highlight two ways in which this could potentially lead us astray. This isn't a reason to not use these methods, but rather to be cautious about what they can and can't tell us.

First, approximate labels have a tendency to tell us only about the most extreme cases. This is part of a broader point that we should be cautious of measures which are difficult or impossible to obtain an exact gold-standard. For example, in the context of scaling or topic models, the algorithm often produces a level precision in the score that we can't evaluate with our course labels. This can lead to circumstances where we fail to notice that our measures lack a discriminating power that they would appear to have on their face. Goertz (2008) identifies a concern in measurement more generally that he calls the "the gray zone," a range of intermediate points along a scale that are not really distinguished from each other. He gives an example using the Polity scale for democracy which ranges from -10 to +10. He notes that if you take only the most extreme values (-10 and +10) which account for 23%

of the data and replace all other values with random uniform draws, you get a measure that correlates at .5 with the measure of interest. The concern here is that we can easily trick ourselves into believing that we have a measure that discriminates well along an entire scale but really just picks up a coarse distinction between two extremes.

The second concern is that surrogate labels are often non-randomly selected and thus not representative of the population of interest. This might be because the surrogate labels are found data, or because we are interested in evaluating the accuracy of a fairly rare category which would require an inordinately large number of randomly sampled documents to efficiently evaluate. Thankfully evaluation designs developed for the events data literature can be helpful in this latter case (King and Lowe, 2003).

Smaller Gold-Standard Data Sometimes a gold-standard dataset can be collected but simply not at the scale or with the accuracy necessary to perform standard supervised learning. For example, after running a clustering model to partition a set of documents, an analyst could write out a codebook and code a small set of heldout documents, use the model to estimate the categories and compare the results using the techniques shown in the previous section.

Even in the case of scaling we can get absolute measures which are approximately gold-standard. Lowe and Benoit (2013) lay out a procedure for validating scaled manifestos using both pairwise comparisons and a 0-100 scaling. These measures might be too noisy to directly train a regression-based scaling method, but they can be valuable for validation.

Partial Category Replication When validating unsupervised topic or clustering models, one of the challenges can be that the number of dimensions is too high to feasibly train human coders or do standard supervised learning. However, even in settings with 100 or more categories it is often the case that only a subset of categories (or an aggregation of categories) is relevant to the quantity of interest. In these settings, it is possible to write a codebook for only the subset of categories of interest (grouping the rest into an other category) and perform standard supervised learning. Grimmer (2013*a*) uses this approach to establish that the supervised method ReadMe is able to capture aggregate categories of interest about legislative credit claiming.

Full Replication Ideally, with infinite time and resources, we could conduct a full replication of the finding using gold-standard data. This essentially moves us back to the previous section and the methods for assessment described there. This is going to be impractical in many settings, but is an important possibility to consider.

8.2.2 Non-Expert Human Evaluation

It is often helpful to have non-expert human evaluations. The trick to these designs is finding a way to solicit relevant information from the evaluators. in particular, non-expert humans can also help evaluate *semantic validity*, the claim that the learned dimension is internally homogenous and distinctive. We provided overviews of some of the relevant experimental designs in discovery chapter including the cluster quality evaluations of (Grimmer and King, 2011) and the intrudor detection tasks of (Chang et al., 2009). Pairwise comparisons and ranking tasks are another useful way of soliciting evaluations (Lowe and Benoit, 2013; Carlson and Montgomery, 2017; Kaufman, King and Komisarchik, 2017).

8.2.3 Model-Based Assessments

One of the most straight-forward assessments, but perhaps the least revealing, is model-based assessments. Here we assess the fit of the model based either on the criterion it was originally designed to optimize or some related external criterion. Because we are interested in the models only instrumentally for what they can tell us about the world and not as ends unto themselves, the model-based assessments are not as revealing as the other approaches.

Still, they are often the easiest validations to compute and thus can be useful supplements, if certainly not replacements, for other modes of validation.

For generative models of text, the most straightforward model-based evaluation is heldout log-likelihood which assesses the fit of the model in a sample not used in training the model. This provides an indication of the ability of the compressed latent variable to reproduce the observed data that was used in the generative model (e.g. the word counts).

Unfortunately due to the nature of latent variable models, even calculating the heldout log-likelihood exactly can be intractable (Wallach et al., 2009). For a given test document i we want to calculate $p(w_i|w_{\text{train}}) = \int_{\pi_i,\mu} p(w_i, \pi_i, \mu)$ but the integral is intractable. For the topic model case, Wallach et al. (2009) offers an overview and comparison of existing approximation techniques. These methods are only really interpretable comparatively. Foulds and Smyth (2014) show an approach for directly estimating relative performance between models using annealing paths. Optimizing held-out likelihood maximizes the performance of the model on a task that we are not actively engaged in: predicting term count vectors for documents. It protects us from a certain kind of overfitting, but does not ensure that we have captured anything about the underlying meaning of the documents.

[Summary of material to be added: These methods are often use to set K. A short description here will explain why we don't find that very helpful.]

[Summary of material to be added: A description of further model-based assessments that are slightly external to the model such as posterior predictive checks (Gelman, Meng and Stern, 1996; Mimno and Blei, 2011; Mimno, Blei and Engelhardt, 2015) and surrogate criteria like cohesiveness and exclusivity (Mimno et al., 2011; Roberts et al., 2014).]

8.2.4 Correspondence to External Information

If measures are interesting they should have interesting relationships to external information. When these relationships be anticipated and specified beforehand, we can evaluate *hypothesis* *validity.* For example, we might expect swings in congressional speech on the house floor should be predictable based on major news events (Quinn et al., 2010) or that committee chairs in the senate talk more about the issues related to their committee (Grimmer, 2013*b*).

The key to hypothesis validity is carefully choosing the hypotheses so that a failure to see the observed relationship is extremely unlikely except when the measure is performing poorly. This often creates a bit of a paradox where researchers show that their new measures has some expected relationships and then in their key finding shows an unexpected relationship.

8.3 Example: Validating Jihad Scores

Validation strategies are often idiosyncratic to the particular case in which they are used because they focus on the particular areas where the analyst wants to allow the opportunity to prove themselves wrong. When designing checks for you own work, be sure to clarify what it what mean for the validation to fail and how the claim being validated fits into the broader architecture the argument. Designing custom validation checks can be a great way of clarifying what properties are important to you as analyst about your measure.

To give a sense of what these checks might look like, we describe the validation process for the development of the jihad scale in Nielsen (2017) the context for which we described in Section 5. Nielsen (2017, Ch. 5) uses a nine-part validation to assess different assumptions and performance qualities of his measure. This is a particularly difficult case as part of the motivation for his scaling exercise was the extreme difficulty in obtaining gold standard data.

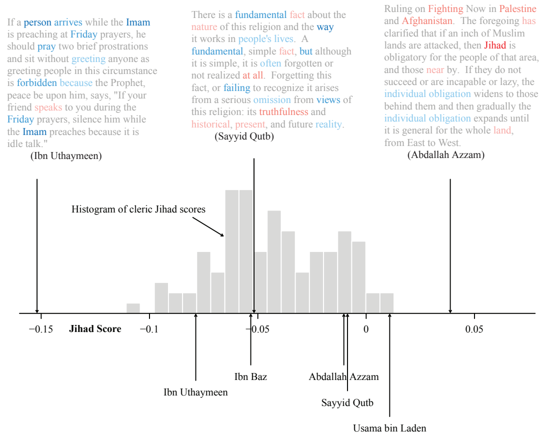
Validation 1: Representativeness of the labeled documents Recall that Nielsen (2017) uses a 'supervision of convenience' strategy in which the scale is pinned down by two extreme: a set of documents from known jihadists called the Jihadist's Bookbag. These documents are a form of 'found data', they were not annotated for this purpose and they aren't a set of documents we are trying to classify. As a result, it is important to think about whether or not they are representative of the kinds of *his conception* of jihad. There is theoretical justification for this set of texts, the jihadists themselves are essentially defining what jihad is through the construction of this set, but he needed to verify that the contents comport with his description of jihad.

To get a feel for the documents he ran an LDA topic model with $K = 5$ clusters. In the book he presents the top five words for each topic (using FREX scores) and titles of three documents per topic that have the highest proportions of the given topic. He then further groups the authors of these documents into single membership clusters based on their use of different topics to show patterns of particular Jihadist authors. He concludes through these assessments and reading of the material that these documents are accurate reflections of his conception of jihadist ideology.

Validation 2: Sensitivity to alternate dataset To ensure that his findings were not driven by the particular set of labeled documents he chose, Nielsen (2017) generated a new scaling using documents in the online jihadist library "Pulpit of Monotheism and Jihad." He originally did not choose this dataset because it was too ideologically broad. Still he shows it produces qualitatively similar findings to his scores.

Validation 3: Assessing the word weights Scaling models of this sort have the distinct advantage that the way a document is scored is fairly transparent. Nielsen (2017) examined the word weights to see that they seemed on their face reasonable. He also visualized the translated results (Figure 6). This helps to assure us that the classifier is doing something reasonable while also communicating more information to the reader about what is indicative of Jihad.

Validation 4: Comparison of the Resulting Scores for Known Individuals Nielsen (2017) reports on some baseline checks on the nature of the scores. He shows the distribution of scores and labels some key figures as well as providing translated samples of their writing (Figure 7). He couples this with descriptions of why these clerics are reasonably placed along the scale. He also investigates all the clerics at the most extreme end of the scale and confirms that they are in fact extreme jihadists.



If a person arrives while the Imam is preaching at Friday prayers, he should pray two brief prostrations and sit without greeting anyone as greeting people in this circumstance is forbidden because the Prophet, peace be upon him, says, "If your friend speaks to you during the Friday prayers, silence him while the Imam preaches because it is idle talk."
(Ibn Uthaymeen)

There is a fundamental fact about the nature of this religion and the way it works in people's lives. A fundamental, simple fact, but although it is simple, it is often forgotten or not realized at all. Forgetting this fact, or failing to recognize it arises from a serious omission from views of this religion: its truthfulness and historical, present, and future reality.
(Sayyid Qutb)

Ruling on Fighting Now in Palestine and Afghanistan. The foregoing has clarified that if an inch of Muslim lands are attacked, then Jihad is obligatory for the people of that area, and those near by. If they do not succeed or are incapable or lazy, the individual obligation widens to those behind them and then gradually the individual obligation expands until it is general for the whole land, from East to West.
(Abdallah Azzam)

Figure 7: Figure 5.3 from Nielsen (2017) showing validation of individual clerics and translated excerpts from their writing.

Validation 5–8: Checks against External Sources If the jihad scores are accurate, we might expect that they align with other sources of expert knowledge. Nielsen (2017) assesses four sources of such knowledge: (5) biographies of clerics (which may identify them as jihadists), (6) a list of supporters of al-Qaeda generated by Ayman al-Zawahiri, the current leader of al-Qaeda, (7) a set of three jihadist websites listing both fellow jihadists and clerics that they explicitly reject, and (8) assessments by counter-terrorism specialists appearing in (McCants, Brachman and Felter, 2006). In each case he demonstrates that his jihad scores are predictive of the classifications in these lists and in the case of errors, carefully explains the nature of the error and why it is unlikely to problematic for the validity of the final scores.

Validation 9: Careful Reading of the Texts Finally, the most important validation, Nielsen carefully read many of the texts.

These validations are all quite specific to the particular case but they run the gamut of establishing that the scores are predictive of the things they should be predictive of and satisfy assumptions that are important to the model. Of course, there are always more validations that one can do, but this assessment is quite extensive. The validation section takes up 20 pages in the book (pg 114-130, appendix pg 204-209), includes 6 figures and 2 tables, an allocation of space that is unlikely to fit in a typical journal article. However, the scores are core to the book's argument and consequently careful validation is warranted.

9 Conclusion

In this treatise on social measurement, sociologist Otis Dudley Duncan wrote,

> Measurement is one of many human achievements and practices that grew up and came to be taken for granted before anyone thought to ask how and why they work.
>
> (Duncan, 1984, p.119)

Measurement is a process that we often take for granted. Existing measures aren't always carefully examined and the coverage of measurement in graduate training is uneven. An advantage of text-as-data methods being relatively new is that they have brought a renewed attention to the assumptions that go into measurement and the importance of careful validation. That said, the field is still new and we are far away from a rigorous foundational theory of measurement and validation.

[Summary of material to be added: More concluding elements and transition to causal inference. As Will Lowe memorably wrote in a conclusion for a draft paper: ‘as above, but shorter’]

References

Ahmed, Amr and Eric P Xing. 2010. Staying informed: supervised and semi-supervised multi-view topical analysis of ideological perspective. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing.* Association for Computational Linguistics pp. 1140–1150.

Andrzejewski, David, Xiaojin Zhu and Mark Craven. 2009. Incorporating domain knowledge into topic modeling via Dirichlet forest priors. In *Proceedings of the 26th Annual International Conference on Machine Learning.* ACM pp. 25–32.

Arora, Sanjeev, Rong Ge, Yonatan Halpern, David Mimno, Ankur Moitra, David Sontag, Yichen Wu and Michael Zhu. 2013. A practical algorithm for topic modeling with provable guarantees. In *International Conference on Machine Learning.* pp. 280–288.

Asuncion, Arthur, Max Welling, Padhraic Smyth and Yee Whye Teh. 2009. On smoothing and inference for topic models. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence.* AUAI Press pp. 27–34.

Bail, Christopher A. 2012. "The fringe effect: Civil society organizations and the evolution of media discourse about Islam since the September 11th attacks." *American Sociological Review* 77(6):855–879.

Bail, Christopher A. 2014. "The cultural environment: measuring culture with big data." *Theory and Society* 43(3):465–482.
URL: *http://dx.doi.org/10.1007/s11186-014-9216-5*

Bail, Christopher Andrew. 2016. "Combining natural language processing and network analysis to examine how advocacy organizations stimulate conversation on social media." *Proceedings of the National Academy of Sciences* 113(42):11823–11828.

Banerjee, Arindam, Srujana Merugu, Inderjit S Dhillon and Joydeep Ghosh. 2005. "Clustering with Bregman divergences." *Journal of machine learning research* 6(Oct):1705–1749.

Beauchamp, Nick. 2011. "Using Text to Scale Legislatures with Uninformative Voting." New York University Mimeo.

Beauchamp, Nick. 2013. Predicting and interpolating state-level polling using Twitter textual data. In *New directions in analyzing text as data workshop.*

Benoit, Kenneth, Drew Conway, Benjamin E Lauderdale, Michael Laver and Slava Mikhaylov. 2016. "Crowd-sourced text analysis: Reproducible and agile production of political data." *American Political Science Review* 110(2):278–295.

Benoit, Kenneth, Kevin Munger and Arthur Spirling. 2017. "Measuring and Explaining Political Sophistication Through Textual Complexity.".

Benoit, Kenneth, Michael Laver and Slava Mikhaylov. 2009. "Treating words as data with error: Uncertainty in text statements of policy positions." *American Journal of Political Science* 53(2):495–513.

Berinsky, Adam J, Gregory A Huber and Gabriel S Lenz. 2012. "Evaluating online labor markets for experimental research: Amazon. com's Mechanical Turk." *Political Analysis* 20(3):351–368.

Blei, David M, Andrew Y Ng and Michael I Jordan. 2003. "Latent dirichlet allocation." *Journal of machine Learning research* 3(Jan):993–1022.

Blei, David M and John D Lafferty. 2006. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*. ACM pp. 113–120.

Boyd-Graber, Jordan, David Mimno and David Newman. 2014. "Care and feeding of topic models: Problems, diagnostics, and improvements." *Handbook of mixed membership models and their applications* 225255.

Bradley, MM and PJ Lang. 1999. "Affective Norms for English Words (ANEW): Stimuli, Instruction, Manual and Affective Ratings." University of Florida Mimeo.

Breiman, Leo et al. 2001. "Statistical modeling: The two cultures (with comments and a rejoinder by the author)." *Statistical Science* 16(3):199–231.

Budak, Ceren, Sharad Goel and Justin M Rao. 2016. "Fair and balanced? quantifying media bias through crowdsourced content analysis." *Public Opinion Quarterly* 80(S1):250–271.

Buhrmester, Michael, Tracy Kwang and Samuel D Gosling. 2011. "Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data?" *Perspectives on psychological science* 6(1):3–5.

Burden, Barry and Joseph Sanberg. 2003. "Budget Rhetoric in Presidential Campaigns from 1952 to 2000." *Political Behavior* 25(2):97–118.

Carley, Kathleen. 1990. *Content analysis*. The encyclopedia of language and linguistics. Edinburgh: Pergamon Press.

Carley, Kathleen M. 1997*a*. "Extracting team mental models through textual analysis." *Journal of Organizational Behavior* pp. 533–558.

Carley, Kathleen M. 1997*b*. "Network text analysis: The network position of concepts." *Text analysis for the social sciences: Methods for drawing statistical inferences from texts and transcripts* 4:79–100.

Carlson, David and Jacob M Montgomery. 2017. "A pairwise comparison framework for fast, flexible, and reliable human coding of political texts." *American Political Science Review* 111(4):835–843.

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

Chakrabarti, Parijat and Margaret Frye. 2017. "A mixed-methods framework for analyzing text data: Integrating computational techniques with qualitative methods in demography." *Demographic Research* 37:1351–1382.

Chang, Jonathan, Sean Gerrish, Chong Wang, Jordan L Boyd-Graber and David M Blei. 2009. Reading tea leaves: How humans interpret topic models. In *Advances in neural information processing systems.* pp. 288–296.

Choi, Hyunyoung and Hal Varian. 2012. "Predicting the present with Google Trends." *Economic Record* 88(s1):2–9.

Chuang, Jason, Margaret E Roberts, Brandon M Stewart, Rebecca Weiss, Dustin Tingley, Justin Grimmer and Jeffrey Heer. 2015. TopicCheck: Interactive Alignment for Assessing Topic Model Stability. In *HLT-NAACL.* pp. 175–184.

Cooper, Seth, Firas Khatib, Adrien Treuille, Janos Barbero, Jeehyung Lee, Michael Beenen, Andrew Leaver-Fay, David Baker, Zoran Popović et al. 2010. "Predicting protein structures with a multiplayer online game." *Nature* 466(7307):756.

Dawid, Alexander Philip and Allan M Skene. 1979. "Maximum likelihood estimation of observer error-rates using the EM algorithm." *Applied statistics* pp. 20–28.

Denny, Matthew J and Arthur Spirling. Forthcoming. "Text preprocessing for unsupervised learning: why it matters, when it misleads, and what to do about it." *Political Analysis* .

Diermeier, Daniel, Jean-Francois Godbout, Bei Yu and Stefan Kaufmann. 2011. "Language and Ideology in Congress." *British Journal of Political Science* . Forthcoming.

Difallah, Djellel, Elena Filatova and Panos Ipeirotis. 2018. "Demographics and Dynamics of Mechanical Turk Workers.".

DiMaggio, Paul, Manish Nag and David Blei. 2013. "Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of US government arts funding." *Poetics* 41(6):570–606.

Duncan, Otis Dudley. 1984. *Notes on social measurement: Historical and critical.* Russell Sage Foundation.

Duneier, Mitchell. 2016. *Ghetto: The invention of a place, the history of an idea.* Macmillan.

Efron, Bradley and Gail Gong. 1983. "A Leisurely Look at the Bootstrap, the Jackknife, and Cross-Validation." *American Statistician* pp. 36–48.

Eichstaedt, Johannes C, Hansen Andrew Schwartz, Margaret L Kern, Gregory Park, Darwin R Labarthe, Raina M Merchant, Sneha Jha, Megha Agrawal, Lukasz A Dziurzynski, Maarten Sap et al. 2015. "Psychological language on Twitter predicts county-level heart disease mortality." *Psychological science* 26(2):159–169.

Eisenstein, Jacob, Amr Ahmed and Eric P Xing. 2011. "Sparse additive generative models of text.".

Eisenstein, Jacob, Brendan O'Connor, Noah A Smith and Eric P Xing. 2010. A latent variable model for geographic lexical variation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing.* Association for Computational Linguistics pp. 1277–1287.

Elkan, Charles. 2006. Clustering documents with an exponential-family approximation of the Dirichlet compound multinomial distribution. In *Proceedings of the 23rd international conference on Machine learning.* ACM pp. 289–296.

Elkan, Charles and Keith Noto. 2008. Learning classifiers from only positive and unlabeled data. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining.* ACM pp. 213–220.

Eshbaugh-Soha, Matthew. 2010. "The Tone of Local Presidential News Coverage." *Political Communication* 27(2):121–140.

Fenno, R.F. 1978. *Home style: House Members in their districts.* Little, Brown.
URL: *https://books.google.com/books?id=RUckAQAAIAAJ*

Fong, Christian and Matthew Tyler. 2017. "Regression with Classier-Generated Covariates.".

Fort, Karën, Gilles Adda and K Bretonnel Cohen. 2011. "Amazon mechanical turk: Gold mine or coal mine?" *Computational Linguistics* 37(2):413–420.

Foulds, James R and Padhraic Smyth. 2014. Annealing Paths for the Evaluation of Topic Models. In *UAI.* pp. 220–229.

Gan, Zhe, Changyou Chen, Ricardo Henao, David Carlson and Lawrence Carin. 2015. Scalable deep Poisson factor analysis for topic modeling. In *International Conference on Machine Learning.* pp. 1823–1832.

Gelman, Andrew, Xiao-Li Meng and Hal Stern. 1996. "Posterior predictive assessment of model fitness via realized discrepancies." *Statistica sinica* pp. 733–760.

Gentzkow, Matthew, Jesse M Shapiro and Matt Taddy. 2016. Measuring polarization in high-dimensional data: Method and application to congressional speech. Technical report National Bureau of Economic Research.

Gerring, John and Craig W Thomas. 2005. "Comparability: A key issue in research design." *Committee on Concepts and Methods. Working Paper Series* 4:1–20.

Gerrish, Sean and David M Blei. 2012. How they vote: Issue-adjusted models of legislative behavior. In *Advances in Neural Information Processing Systems.* pp. 2753–2761.

Ginsberg, Jeremy, Matthew H Mohebbi, Rajan S Patel, Lynnette Brammer, Mark S Smolinski and Larry Brilliant. 2009. "Detecting influenza epidemics using search engine query data." *Nature* 457(7232):1012–1014.

Goertz, Gary. 2008. Concepts, Theories, and Numbers: A Checklist for Constructing, Evaluating, and Using Concepts or Quantitative Measures. In *The Oxford Handbook of Political Methodology*, ed. J.M. Box-Steffensmeier, H.E. Brady and D. Collier. Oxford University Press.
URL: *https://books.google.com/books?id=BpsLCx0SHtwC*

Griffiths, Thomas L and Mark Steyvers. 2004. "Finding scientific topics." *Proceedings of the National academy of Sciences* 101(suppl 1):5228–5235.

Grimmer, Justin. 2010. "A Bayesian hierarchical topic model for political texts: Measuring expressed agendas in Senate press releases." *Political Analysis* 18(1):1–35.

Grimmer, Justin. 2013*a*. "Appropriators not position takers: The distorting effects of electoral incentives on congressional representation." *American Journal of Political Science* 57(3):624–642.

Grimmer, Justin. 2013*b*. *Representational style in Congress: What legislators say and why it matters.* Cambridge University Press.

Grimmer, Justin and Gary King. 2011. "General purpose computer-assisted clustering and conceptualization." *Proceedings of the National Academy of Sciences* 108(7):2643–2650.

Grimmer, Justin, Gary King and Chiara Superti. 2018. "The Unreliability of Measures of Intercoder Reliability and What to do About it.".

Halpern, Yoni, Steven Horng and David Sontag. 2015. "Anchored Discrete Factor Analysis." *arXiv preprint arXiv:1511.03299* .

Halpern, Yoni, Steven Horng, Youngduck Choi and David Sontag. 2016. "Electronic medical record phenotyping using the anchor and learn framework." *Journal of the American Medical Informatics Association* p. ocw011.

Halpern, Yoni, Youngduck Choi, Steven Horng and David Sontag. 2014. Using anchors to estimate clinical state without labeled data. In *AMIA Annual Symposium Proceedings.* Vol. 2014 American Medical Informatics Association p. 606.

Hart, RP. 2000. *Diction 5.0: The Text Analysis Program.* Thousand Oaks, CA: Sage-Scolari.

Hastie, Trevor, Robert Tibshirani and Jerome Friedman. 2001. *The Elements of Statistical Learning.* Springer.

Hastie, Trevor, Robert Tibshirani and Jerome Friedman. 2013. *The elements of statistical learning.* Springer.

Hays, David C. 1960. "Automatic content analysis." *Rand Corporation. Santa Monica* .

Hengel, Erin. 2017. Publishing while female. Technical report Technical report, Technical report, University of Liverpool.

Hinton, Geoffrey E and Ruslan R Salakhutdinov. 2009. Replicated softmax: an undirected topic model. In *Advances in neural information processing systems.* pp. 1607–1614.

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.

Hu, Yuening, Jordan Boyd-Graber, Brianna Satinoff and Alison Smith. 2014. "Interactive topic modeling." *Machine learning* 95(3):423–469.

Huff, Connor and Dustin Tingley. 2015. "Who are these people? Evaluating the demographic characteristics and political preferences of MTurk survey respondents." *Research & Politics* 2(3):2053168015604648.

Jerzak, Connor T, Gary King and Anton Strezhnev. 2018. "An Improved Method of Automated Nonparametric Content Analysis for Social Science.".

Joshi, Shalmali, Suriya Gunasekar, David Sontag and Joydeep Ghosh. 2016. "Identifiable Phenotyping using Constrained Non-Negative Matrix Factorization." *arXiv preprint arXiv:1608.00704* .

Kaufman, Aaron, Gary King and Mayya Komisarchik. 2017. "How to Measure Legislative District Compactness If You Only Know it When You See it.".

Kellstedt, Paul. 2000. "Media Framing and the Dynamics of Racial Policy Preferences." *American Journal of Political Science* 44(2):245–260.

Kim, In Song, John Londregan and Marc Ratkovic. Forthcoming. "Estimating Ideal Points from Votes and Text." *Political Analysis* .

King, G., R.O. Keohane and S. Verba. 1994. *Designing Social Inquiry: Scientific Inference in Qualitative Research.* Princeton paperbacks Princeton University Press.
URL: *https://books.google.com/books?id=A7VFF-JR3b8C*

King, Gary, Kay Schlozman and Norman Nie. 2009. "The changing evidence base of social science research." *The Future of Political Science: 100 Perspectives* pp. 91–93.

King, Gary and Will Lowe. 2003. "An automated information extraction tool for international conflict data with performance as good as human coders: A rare events evaluation design." *International Organization* 57(3):617–642.

Kittur, Aniket, Ed H Chi and Bongwon Suh. 2008. Crowdsourcing user studies with Mechanical Turk. In *Proceedings of the SIGCHI conference on human factors in computing systems.* ACM pp. 453–456.

Krippendorff, Klaus. 2004. *Content analysis: An introduction to its methodology.* Sage.

Krippendorff, Klaus. 2012. *Content analysis: An introduction to its methodology.* Sage.

Krosnick, Jon. 1999. "Survey Research." *Annual Review of Psychology* 50(1):537–567.

Lancichinetti, Andrea, M Irmak Sirer, Jane X Wang, Daniel Acuna, Konrad Körding and Luís A Nunes Amaral. 2015. "High-reproducibility and high-accuracy method for automated topic classification." *Physical Review X* 5(1):011007.

Larochelle, Hugo and Stanislas Lauly. 2012. A neural autoregressive topic model. In *Advances in Neural Information Processing Systems.* pp. 2708–2716.

Lauderdale, Benjamin E and Tom S Clark. 2014. "Scaling politically meaningful dimensions using texts and votes." *American Journal of Political Science* 58(3):754–771.

Laver, Michael and John Garry. 2000. "Estimating Policy Positions from Political Texts." *American Journal of Political Science* 44(3):619–634.

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.

Lazarsfeld, Paul F and Allen H Barton. 1951. "Qualitative measurement in the social sciences: Classification, typologies, and indices." *The policy sciences* pp. 155–192.

Lazer, David, Alex Sandy Pentland, Lada Adamic, Sinan Aral, Albert Laszlo Barabasi, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann et al. 2009. "Life in the network: the coming age of computational social science." *Science (New York, NY)* 323(5915):721.

Lazer, David, Ryan Kennedy, Gary King and Alessandro Vespignani. 2014. "The parable of Google flu: traps in big data analysis." *Science* 343(6176):1203–1205.

Loughran, Tim and Bill McDonald. 2011. "When is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks." *Journal of Finance* 66(1):35–65.

Lowe, Will. 2016. "Scaling things we can count." *Online verfügbar unter http://dl. conjugateprior. org/preprints/scaling-things-we-can-count. pdf, zuletzt geprüft am* 16(2016):99–132.

Lowe, Will and Kenneth Benoit. 2013. "Validating estimates of latent traits from textual data using human judgment as a benchmark." *Political Analysis* 21(3):298–313.

Lowe, Will, Kenneth Benoit, Slava Mikhaylov and Michael Laver. 2011. "Scaling policy preferences from coded political texts." *Legislative studies quarterly* 36(1):123–155.

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.

Lund, Jeffrey, Connor Cook, Kevin Seppi and Jordan Boyd-Graber. 2017. Tandem anchoring: A multiword anchor approach for interactive topic modeling. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).* Vol. 1 pp. 896–905.

Maron, Melvin Earl and John L Kuhns. 1960. "On relevance, probabilistic indexing and information retrieval." *Journal of the ACM (JACM)* 7(3):216–244.

Mason, Winter and Siddharth Suri. 2012. "Conducting behavioral research on Amazons Mechanical Turk." *Behavior research methods* 44(1):1–23.

Mcauliffe, Jon D and David M Blei. 2008. Supervised topic models. In *Advances in neural information processing systems.* pp. 121–128.

McCants, William, Jarret Brachman and Joseph Felter. 2006. Militant Ideology Atlas: Research Compendium. Technical report MILITARY ACADEMY WEST POINT NY COMBATING TERRORISM CENTER.

Michel, Jean-Baptiste, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K Gray, Joseph P Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant et al. 2011. "Quantitative analysis of culture using millions of digitized books." *science* 331(6014):176–182.

Mimno, David and Andrew McCallum. 2012. "Topic models conditioned on arbitrary features with dirichlet-multinomial regression." *arXiv preprint arXiv:1206.3278* .

Mimno, David and David Blei. 2011. Bayesian checking for topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing.* Association for Computational Linguistics pp. 227–237.

Mimno, David, David M Blei and Barbara E Engelhardt. 2015. "Posterior predictive checks to quantify lack-of-fit in admixture models of latent population structure." *Proceedings of the National Academy of Sciences* 112(26):E3441–E3450.

Mimno, David, Hanna M Wallach, Edmund Talley, Miriam Leenders and Andrew McCallum. 2011. Optimizing semantic coherence in topic models. In *Proceedings of the conference on empirical methods in natural language processing.* Association for Computational Linguistics pp. 262–272.

Monroe, Burt, Michael Colaresi and Kevin Quinn. 2008. "Fightin' Words: Lexical Feature Selection and Evaluation for Identifying the Content of Political Conflict." *Political Analysis* 16(4):372.

Mullainathan, Sendhil and Jann Spiess. 2017. "Machine learning: an applied econometric approach." *Journal of Economic Perspectives* 31(2):87–106.

Murphy, Kevin P. 2012. *Machine learning: a probabilistic perspective.* MIT press.

Neuendorf, Kimberly A. 2002. *The content analysis guidebook.* Sage.

Neuendorf, Kimberly A. 2016. *The content analysis guidebook.* Sage.

Ng, Andrew Y and Michael I Jordan. 2002. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In *Advances in neural information processing systems.* pp. 841–848.

Nguyen, Thang, Jordan Boyd-Graber, Jeffrey Lund, Kevin Seppi and Eric Ringger. 2015. Is your anchor going up or down? Fast and accurate supervised topic models. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.* pp. 746–755.

Nielsen, Richard. 2012. "Jihadi radicalization of muslim clerics." *unpublished paper, Cambridge MA: Harvard University* .

Nielsen, Richard A. 2017. *Deadly Clerics: Blocked Ambition and the Paths to Jihad.* Cambridge University Press.

O'Connor, Brendan, Ramnath Balasubramanyan, Bryan R Routledge and Noah A Smith. 2010. "From tweets to polls: Linking text sentiment to public opinion time series." *ICWSM* 11(122-129):1–2.

Paisley, John, Chong Wang, David M Blei and Michael I Jordan. 2015. "Nested hierarchical Dirichlet processes." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37(2):256–270.

Pang, B. and L. Lee. 2008. "Opinion mining and sentiment analysis." *Foundations and Trends in Information Retrieval* 2(1-2):1–135.

Paul, Michael and Mark Dredze. 2012. Factorial LDA: Sparse multi-dimensional text models. In *Advances in Neural Information Processing Systems*. pp. 2582–2590.

Pechenick, Eitan Adam, Christopher M Danforth and Peter Sheridan Dodds. 2015. "Characterizing the Google Books corpus: Strong limits to inferences of socio-cultural and linguistic evolution." *PloS one* 10(10):e0137041.

Pennebaker, James, Martha Francis and Roger Booth. 2001. *Linguistic Inquiry and Word Count: LIWC 2001*. Mahway, NJ: Erlbaum Publishers.

Perry, Patrick O and Kenneth Benoit. 2017. "Scaling Text with the Class Affinity Model." *arXiv preprint arXiv:1710.08963* .

Peterson, Andrew and Arthur Spirling. 2018. "Classification Accuracy as a Substantive Quantity of Interest: Measuring Polarization in Westminster Systems." *Political Analysis* 26(1):120–128.

Przeworski, Adam, Henry Teune et al. 1970. The logic of comparative social inquiry. Technical report.

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.

Ranganath, Rajesh, Linpeng Tang, Laurent Charlin and David Blei. 2015. Deep exponential families. In *Artificial Intelligence and Statistics*. pp. 762–771.

Ratner, Alexander, Stephen H Bach, Henry Ehrenberg, Jason Fries, Sen Wu and Christopher Ré. 2017. "Snorkel: Rapid training data creation with weak supervision." *arXiv preprint arXiv:1711.10160* .

Reichman, Nancy E, Julien O Teitler, Irwin Garfinkel and Sara S McLanahan. 2001. "Fragile families: Sample and design." *Children and Youth Services Review* 23(4-5):303–326.

Reing, Kyle, David C Kale, Greg Ver Steeg and Aram Galstyan. 2016. "Toward interpretable topic discovery via anchored correlation explanation." *arXiv preprint arXiv:1606.07043* .

Roberts, Margaret E., Brandon M. Stewart and Dustin Tingley. 2016. Navigating the Local Modes of Big Data: The Case of Topic Models. In *Computational Social Science: Discovery and Prediction*, ed. R. Michael Alvarez. New York: Cambridge University Press chapter 2, pp. 51–97.

Roberts, Margaret E, Brandon M. Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson and David G Rand. 2014. "Structural Topic Models for Open-Ended Survey Responses." *American Journal of Political Science* 58(4):1064–1082.

Rosen-Zvi, Michal, Thomas Griffiths, Mark Steyvers and Padhraic Smyth. 2004. The author-topic model for authors and documents. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence.* AUAI Press pp. 487–494.

Rule, Alix, Jean-Philippe Cointet and Peter S Bearman. 2015. "Lexical shifts, substantive changes, and continuity in State of the Union discourse, 1790–2014." *Proceedings of the National Academy of Sciences* 112(35):10837–10844.

Salganik, Matthew. 2017. *Bit by Bit: Social Research in the Digital Age.* Princeton University Press.

Schrodt, Philip A. 2006. "Twenty years of the Kansas event data system project." *The political methodologist* 14(1):2–8.

Schrodt, Philip A. 2012. "Precedents, progress, and prospects in political event data." *International Interactions* 38(4):546–569.

Sebeok, Thomas A and Valdis J Zeps. 1958. "An analysis of structured content, with application of electronic computer research, in psycholinguistics." *Language and Speech* 1(3):181–193.

Shank, Daniel B. 2016. "Using crowdsourcing websites for sociological research: The case of Amazon Mechanical Turk." *The American Sociologist* 47(1):47–55.

Sheng, Victor S, Foster Provost and Panagiotis G Ipeirotis. 2008. Get another label? improving data quality and data mining using multiple, noisy labelers. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining.* ACM pp. 614–622.

Slapin, Jonathan B and Sven-Oliver Proksch. 2008. "A scaling model for estimating time-series party positions from texts." *American Journal of Political Science* 52(3):705–722.

Snow, Rion, Brendan O'Connor, Daniel Jurafsky and Andrew Y Ng. 2008. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the conference on empirical methods in natural language processing.* Association for Computational Linguistics pp. 254–263.

Socher, Richard, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Ng and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing.* pp. 1631–1642.

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.

Stevens, S. S. 1946. "On the Theory of Scales of Measurement." *Science* 103(2684):677–680.
URL: *http://science.sciencemag.org/content/103/2684/677*

Stewart, Brandon M. and Yuri M. Zhukov. 2009. "Use of force and civil–military relations in Russia: an automated content analysis." *Small Wars & Insurgencies* 20:319–343.

Stone, Philip J, Dexter C Dunphy and Marshall S Smith. 1966. “The General Inquirer: A Computer Approach to Content Analysis.”.

Stone, Philip J and Earl B Hunt. 1963. A computer approach to content analysis: studies using the general inquirer system. In *Proceedings of the May 21-23, 1963, spring joint computer conference.* ACM pp. 241–256.

Stone, Philip J, Robert F Bales, J Zvi Namenwirth and Daniel M Ogilvie. 1962. “The general inquirer: A computer system for content analysis and retrieval based on the sentence as a unit of information.” *Behavioral Science* 7(4):484–498.

Taddy, Matt. 2013. “Measuring political sentiment on twitter: factor optimal design for multinomial inverse regression.” *Technometrics* 55(4):415–425.

Taddy, Matthew A. 2010. “Inverse Regression for Analysis of Sentiment in Text.” *Arxiv preprint arXiv:1012.2098* .

Thurstone, Louis L. 1927. “A law of comparative judgment.” *Psychological review* 34(4):273.

Thurstone, Louis L. 1959. “The measurement of values.”.

Turney, P and ML Littman. 2003. “Measuring Praise and Criticism: Inference of Semantic Orientation from Association.” *ACM Transactions on Information Systems (TOIS)* 21(4).

van der Vaart, AW, S Dudoit and MJ van der Laan. 2006. “Oracle Inequalities for Multifold Cross Validation.” *Statistics and Decisions* 24(3).

Vikram, Sharad and Sanjoy Dasgupta. 2016. Interactive bayesian hierarchical clustering. In *International Conference on Machine Learning.* pp. 2081–2090.

Wallach, Hanna M, David M Mimno and Andrew McCallum. 2009. Rethinking LDA: Why priors matter. In *Advances in neural information processing systems.* pp. 1973–1981.

Wallach, Hanna M, Iain Murray, Ruslan Salakhutdinov and David Mimno. 2009. Evaluation methods for topic models. In *Proceedings of the 26th annual international conference on machine learning.* ACM pp. 1105–1112.

Wallach, Hanna, Shane Jensen, Lee Dicker and Katherine Heller. 2010. An alternative prior process for nonparametric Bayesian clustering. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics.* pp. 892–899.

Wang, Chong and David M Blei. 2009. Decoupling sparsity and smoothness in the discrete hierarchical Dirichlet process. In *Advances in neural information processing systems.* pp. 1982–1989.

Watkins, Susan Cotts and Ann Swidler. 2009. “Hearsay ethnography: Conversational journals as a method for studying culture in action.” *Poetics* 37(2):162–184.

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.
URL: *https://doi.org/10.1146/annurev-polisci-052615-025542*

Wilkerson, John, David Smith and Nicholas Stramp. 2015. "Tracing the flow of policy ideas in legislatures: A text reuse approach." *American Journal of Political Science* 59(4):943–956.

Williamson, Sinead, Chong Wang, Katherine Heller and David Blei. 2010. "The IBP compound Dirichlet process and its application to focused topic modeling.".

Yao, Limin, David Mimno and Andrew McCallum. 2009. Efficient methods for topic model inference on streaming document collections. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM pp. 937–946.

Young, Lori and Stuart Soroka. 2011. "Affective News: The Automated Coding of Sentiment in Political Texts." *Political Communication* . Forthcoming.

Zhang, Yuchen, Xi Chen, Denny Zhou and Michael I Jordan. 2014. Spectral methods meet EM: A provably optimal algorithm for crowdsourcing. In *Advances in neural information processing systems*. pp. 1260–1268.