Article

Sociological Methods & Research
2020, Vol. 49(1) 3-42

DOI: 10.1177/0049124117729703
journals.sagepub.com/home/smr

Computational Grounded Theory: A Methodological Framework

Laura K. Nelson[1]

Abstract

This article proposes a three-step methodological framework called computational grounded theory, which combines expert human knowledge and hermeneutic skills with the processing power and pattern recognition of computers, producing a more methodologically rigorous but interpretive approach to content analysis. The first, pattern detection step, involves inductive computational exploration of text, using techniques such as unsupervised machine learning and word scores to help researchers to see novel patterns in their data. The second, pattern refinement step, returns to an interpretive engagement with the data through qualitative deep reading or further exploration of the data. The third, pattern confirmation step, assesses the inductively identified patterns using further computational and natural language processing techniques. The result is an efficient, rigorous, and fully reproducible computational grounded theory. This framework can be applied to any qualitative text as data, including transcribed speeches, interviews, open-ended survey data, or ethnographic field notes, and can address many potential research questions.

[1] 960B Renaissance Park, Department of Sociology and Anthropology, Northeastern University, Boston, MA, USA

Corresponding Author:
Laura K. Nelson, 960B Renaissance Park, Department of Sociology and Anthropology, Northeastern University, Boston, MA 02115, USA.
Email: l.nelson@northeastern.edu

Keywords

computational text analysis, grounded theory, computational grounded theory, inductive analysis, unsupervised machine learning, supervised machine learning, natural language processing, word scores

New computational techniques developed by computer scientists and computational linguists have placed sociological content analysis on the verge of major changes. The increasing popularity of these techniques, along with newly available sources of text as data—both "big" and small—have ignited debates about what methods sociologists should use to extract meaning from text (Bail 2014; Biernacki 2012, 2015; DiMaggio, Nag, and Blei 2013; Lee and Martin 2015; Mohr and Bogdanov 2013; Reed 2015; Spillman 2015). As these debates are carried out in sociology journals, others outside of the social sciences are applying these new techniques to social data, but they are doing so without the specific theoretical groundings important to sociologists.

Other disciplines in the social sciences, including political scientists (Grimmer 2010; King, Pan, and Roberts 2013; Schwartz and Ungar 2015) and psychologists (Tausczik and Pennebaker 2010; Yu and Ho 2014), have already incorporated computational methods into their repertoire, adapting a range of computational techniques to assist them with their analyses. Some of these social science researchers are using these methods to code concrete elements in text, such as counting events (Nardulli, Althaus, and Hayes 2015) or identifying subject–verb–object triads (Franzosi 2010). Sociologists, particularly sociologists of culture, are instead adapting these techniques to questions centered on interpretation and meaning (Bail 2014; DiMaggio et al. 2013). While many sociologists agree on the need for more formal ways to measure culture, there is not yet a consensus on the appropriate role computation should hold in attempts to measure meaning (Biernacki 2015; DiMaggio et al. 2013; Lee and Martin 2015; Mohr 1998; Reed 2015; Spillman 2015).

A general lack of standardized guidelines and training around computer-assisted text analysis in sociology is producing a risky situation for the potential haphazard and undisciplined use of text analysis methods. To better guide the use of these tools to measure meaning, this article formulates best practices for using computer-assisted text analysis to conduct a specific type of sociological research often used to measure meaning in text: inductive grounded theory. I propose a method called computational grounded theory,

which combines expert human knowledge and skills at interpretation with the processing power and pattern recognition brought by computers. I argue that the result is a more methodologically rigorous, reliable, and fully reproducible grounded theory approach to content analysis.

Grounded theory (Glaser and Strauss 1999, 2005) is a method designed to allow categories and themes to emerge inductively from data, culminating in data-driven but abstract theoretical understandings of the underlying social world (Charmaz 2014). It has long been used by sociologists to conduct rigorous theory-producing research. There are challenges with this method, however. The nature of grounded theory necessitates a series of "judgment calls," as researchers code and interpret data, bringing "subjectivities," "predispositions," and "personalities" into the analysis process (quoted in Saldana 2015:8). Because of this, it is difficult to validate and reproduce (Biernacki 2012). Additionally, grounded theory does not scale and thus cannot incorporate recent access to large amounts of unstructured social data (Bail 2014). To update grounded theory for contemporary research, I add computational techniques, providing the ability to incorporate massive amounts of data into theory-generating research in a rigorous and reliable fashion, mitigating the shortcomings of purely qualitative research. By staying grounded in an interpretive relationship with the data, my proposed method also mitigates the shortcomings of purely computational methods, namely, the output from computational methods is often difficult to interpret in meaningful ways. By combining the interpretive and computational approaches, my proposed framework overcomes the shortcomings of each individual method, delivering both quantity *and* quality, breadth *and* depth (Franzosi 2010:146), and allows researchers to leverage both close and distant reading (Moretti 2013) to better measure meaning.

In this article, I detail a three-step framework to carry out computational grounded theory research, providing specific computational techniques and empirical examples for each stage of the process. The goal is to provide researchers with a guide for the application of computational techniques to conduct inductive sociologically based empirical research. The article is structured as follows. I begin with a summary of recent discussions and debates about computational techniques and content analysis, relating them to more traditional content analysis approaches. I then detail a three-step approach that combines inductive grounded theory with deductive quantitative tests, utilizing computational techniques for each. Step 1, the pattern detection step, involves using computational techniques to reduce complicated, messy text into simpler, more interpretable lists or networks of words in order to reveal patterns within the text in an unbiased and reproducible

fashion. Step 2, the pattern refinement step, involves a reengagement with the data through a computationally guided deep reading of the text. Step 3, the pattern confirmation step, applies further computational techniques to assess the validity of the inductively identified patterns. These three steps combine for an efficient, rigorous, and fully reliable and reproducible computational grounded theory.

Content Analysis: An Overview

Content analysis is a cornerstone method in sociology. It is used in conjunction with other methods, such as ethnography, interviews, and comparative historical methods, and bridges many subfields, including but not limited to media studies, cultural sociology, and political sociology (Biernacki 1997; J. H. Evans 2002; Ferree et al. 2002; Franzosi 2004; Griswold 1987; Krippendorff 2013; Neuendorf 2001).

Sociologists have attempted to develop content analysis techniques that meet three requirements for scientific analysis: It should be (1) reliable—the analysis will produce the same results every time; (2) intersubjectively valid—two informed analysts will interpret the results in a similar fashion; and (3) fully reproducible—by providing a detailed description of the data processing steps and analytical strategy, as well as the data itself, any researcher will be able to independently reproduce the full analysis. There are various approaches to formal content analysis in sociology falling into two broad, often interrelated, categories: coding, which involves labeling bits of data according to what they indicate, and the humanities, or interpretive, approach, which instead involves interpreting the text holistically.

The first, coding approach, entails breaking down the text into its constituent parts via coding small parts of the larger text into categories. This typically involves a researcher or team of researchers developing a codebook, or dictionary, with which to categorize text. The process is iterative, with codes being modified as they are applied to better fit the data. The researcher then typically trains research assistants to code text into these predetermined categories (see, e.g., Krippendorff 2013; Neuendorf 2001; Saldana 2015). To increase accuracy, multiple coders will code the same text and their agreement on codes is checked via an intercoder reliability score. When the text is coded with a reasonably high intercoder reliability (or, sometimes, by one expert researcher), the researcher then analyzes the coded text, looking for different types of patterns within and between texts (e.g., J. H. Evans 2002; Ferree et al. 2002; Griswold 1987).

A number of problems plague this approach to content analysis. First, generating categories is subjective by nature; like everyone, researchers are subject to confirmation bias. Categories that are based on assumed knowledge about the text may or may not fit the text, and difficulties in parsing complicated text may prevent researchers from considering other relevant categories. Additionally, because of the difficulty for researchers to describe why texts are categorized a certain way, and because of the number of judgment calls required, readers may find it difficult to judge the adequacy of the categorization.[1] Second, content analysis is not easily reproducible. It is difficult to get the same person to code the same article in the same way twice, let alone train an entirely new team to code a corpus in the same way as a previous team. Third, it is very time-consuming. This type of content analysis can only be done on a small amount of text, or by taking a sample of a larger corpus, which leaves out the majority of available text that can be used as data.

Frustration with reliably and efficiently coding a large body of text has led sociologists to seek alternative methods to conduct content analysis. Two major threads have developed in the attempts to create a more robust way to measure meaning via textual analysis. First, scholars have sought to inductively but quantitatively measure meaning using formal structural methods, such as clustering techniques and block modeling (Bearman and Stovel 2000; Carley 1994; Friedland et al. 2014; Martin 2000; Mische and Pattison 2000; Mohr 1998; Mohr and Duquenne 1997; Pachucki and Breiger 2010; Tilly 1997). Researchers using these methods typically identify a unit of meaning in a text or other cultural artifact, develop a measure of a tie between elements, and finally use structural techniques to uncover latent structures in the cultural landscape, directly, they claim, measuring meaning structures. One of the challenges of measuring culture in this way, however, is deciding what counts as a cultural element and, secondarily, deciding what counts as a tie. Many of the elements in the studies cited above are not formally measured but are identified via traditional coding procedures (e.g., Mische and Pattison 2000; Mohr and Duquenne 1997). These formal structural methods thus reproduce the problem of coding text using human coders and do not solve the problems inherent in traditional content analysis.

A second, humanist, thread is to altogether avoid reducing texts to bits of data, instead treating the text as a whole, irreducible, object. This argument is taken to the extreme by Biernacki (2012), who proposes that there is no possible way to make coding scientific—the process of coding itself obscures rather than reveals what content analysts seek to explain. It is unclear,

however, how to validate or reproduce this humanistic approach, leading scholars who seek to rest content analysis on more scientific grounds back again to the search for more formal ways to measure meaning in text (Lee and Martin 2015).

Others straddle the coding and humanist approaches: They are sympathetic to coding but insist on the role of expert-based hermeneutics in the process. Unlike the structural approach, many of these researchers are still ambivalent about the role computers should play in the process (Reed 2015; Spillman 2015).

I argue that recent developments in language and text analysis offer a way to combine the structural approach with the humanist approach, preserving the superior abilities to interpret text holistically provided by humans but incorporating the formal rigor, reliability, and reproducibility of computer-assisted methods.

Computer-assisted Content Analysis

A remarkably wide range of computer-assisted text analysis techniques is readily available requiring a minimum amount of coding skills, including simple word or phrase frequency counts, more sophisticated supervised and unsupervised machine learning algorithms, and natural language processing techniques that incorporate language structure and relations between words into the calculations.[2] Like all methods, text analysis techniques should be chosen based on the research question and the available data to answer that question. Fortuitously, the wide range of available techniques allows for their use to answer a variety of questions utilizing different types of data.

An array of social scientists, in particular political scientists (Bonilla and Grimmer 2013; Grimmer 2010, 2013; King et al. 2013; Monroe, Colaresi, and Quinn 2008), psychologists (Tausczik and Pennebaker 2010; Yu and Ho 2014), and linguists (P. F. Brown et al. 1993; Hinton et al. 2012), are using computer-assisted text analysis techniques to bolster, and in some cases completely replace, traditional content analysis methods in their disciplines (see Hirschberg and Manning 2015, for a summary of recent developments in this larger field). Researchers in these disciplines, who recognize the potential of using text as data, have developed specific tools and methodological best practices to enable robust computer-assisted text analysis in their respective disciplines. For example, psychologists have spent years developing the Linguistic Inquiry and Word Count dictionary (Tausczik and Pennebaker 2010) to measure how language reveals internal psychological processes; political scientists and organizational scholars have

developed DICTION (Alexa and Zuell 2000) to measure different dimensions of political language; and computer scientists and computational linguists are continually developing mathematical models to more accurately extract relevant information from a collection of texts (see, e.g., Manning, Raghavan, and Schütze 2008).

Sociologists are beginning to adapt many of these tools for sociological research (Bail 2012; Hanna 2013; Mohr et al. 2013), but the field has not yet developed discipline-specific and agreed upon tools or best practices around the use of computer-assisted text analysis for sociological research, in particular, inductive analysis or analyses aimed at measuring meaning structures. Additionally, the multitude of possible computer-assisted techniques, combined with the rapidly changing nature of the field, pose challenges to sociologists. Which of the many tools should be used, and can they be trusted to generate the desired analysis?

To bring together and structure the many tools available into a best practices framework that can be used by sociologists, this article proposes a three-step computational grounded theory approach to measuring meaning through text. The first two steps are the pattern detection and pattern refinement steps. The first step uses computational methods to reduce messy and complicated text to interpretable groups of words, helping researchers cut through the noise inherent to text-based data. The second step returns to a deep reading of the text and incorporates holistic interpretation. These two steps help researchers inductively explore text to uncover data-driven and meaningful patterns. The third step involves using computational methods to more reliably test the validity of the inductively identified patterns in the text.

Collectively, this proposed sequence of techniques allows for the use of expert substantive knowledge to guide questions and hypotheses about text, differentiating it from purely computer-driven approaches, and this sequence also makes the entire content analysis process more transparent, reproducible, and efficient. Additionally, because it is automated, the only limitation on the amount of text that can be analyzed is the availability of computational resources. Thus, this method can be scaled to incorporate "big data" of almost infinite size, enabling researchers to explore questions with more quantities of, and more inclusive, data than previously possible. This approach thus brings inductive content analysis closer to the validity, reliability, reproducibility, and scalability necessary for scientific research. Finally, this approach can be used on a variety of data, including primary sources such as newspapers, diaries, or transcribed speeches as well as interviews, open-ended survey responses, and even ethnographic field notes, making it applicable to many sociological fields.

The sequence of steps presented in this article may not be appropriate for every project—different projects may employ these steps in different orders depending on the nature of the data and the research question—and the suite of methods surveyed below are not the only ones available. Furthermore, these methods, like all methods, are not foolproof. They are only as good as the research question and the data and the match between the method and the question. Like any quantitative method, numbers in and of themselves do not make an analysis. The measures and numbers produced by the methods described below, like the output from regression models, need to be interpreted by the researcher, which takes careful and systematic reading. Texts are particularly convoluted forms of data, however, and these techniques can help researchers cut through the complexity and subterfuge of language to extract different types of grounded meaning from text in a methodologically rigorous fashion.

Computational Grounded Theory: A Methodological Framework

Overview of Computer-assisted Text Analysis Techniques

Computer-assisted text analysis techniques fall into three main categories: lexical-based, text classification, and natural language processing.[3]

- Lexical-based techniques, which are done at the word level, can include simple methods like counting words and phrases. More complicated lexical-based techniques include word scores that aim to identify important or distinctive words (Monroe et al. 2008), and relational semantic network techniques, such as mapping networks of words that occur near one another (Lee and Martin 2015).
- Computer-assisted text classification is typically done via machine learning. Machine learning is a general field in computer science that seeks to develop ways for computers to learn without being explicitly programmed. In text analysis, machine learning is most often used to either categorize text into predetermined categories, known as supervised machine learning, or automatically classify text into computationally derived categories known as unsupervised machine learning. Different unsupervised machine learning algorithms do different things, but in general, some algorithms categorize text into mutually exclusive categories, while others are written with the assumption that one document can be classified into multiple categories.

- Natural language processing techniques incorporate language structure, word context, and word features into the analysis, such as a word's semantic context or a word's part of speech.

The three-step computational grounded theory framework described below combines these three categories into a methodologically rigorous approach to measure meaning in text.

Before getting to the specifics of the framework, a quick note on data and software. There are many out-of-the-box tools to do computer-assisted text analysis (e.g., WordStat, and to a lesser extent, qualitative data analysis packages such as MAXQDA), but all of these tools, which are typically proprietary, restrict you to the specific techniques available in the software, which may or may not be the best for every question and type of data.[4] Additionally, computational methods and tools change and advance at a rapid pace, and out-of-the-box proprietary software typically do not keep up with the changing field. The most flexible, and reproducible, way to do computer-assisted text analysis is to instead learn a scripting language such as Python or R, which, even with the higher learning curve, provide a much more powerful and productive way to carry out computer-assisted methods of any form.[5] Table 1 presents these two languages and the libraries and modules researchers can use for the various techniques outlined in this article.

Python and R can additionally be used to transform texts, often saved in different file formats such as .pdf, .doc, and .txt files, into a dataframe on which text analysis techniques can be used. This data processing step is a necessary step at the start of any computer-assisted text analysis project and must be completed before embarking on step 1 below (see Figure 1 for an example of this process). Once the data are processed into a Python or R dataframe, researchers can begin the three-step text analysis process. Figure 2 provides a graphical depiction of this framework. I describe each step in detail below.

Step 1: Pattern Detection Using Human-centered Computational Exploratory Analysis

One of the principle ways computer-assisted text analysis techniques can help sociologists explore text is by reducing complicated, messy text into simpler, more interpretable lists or networks of words. When compared to one another or when their frequencies are measured across texts, the lists or networks of words can suggest relevant patterns within the text, which can

Table 1. Computer-assisted Text Analysis Software.

Language	Module/Package	Techniques	Websites
Python	NLTK (Natural Language Tool Kit)	Natural language processing, preprocessing, text frequency counts, and interface with resources like WordNet	http://www.nltk.org
	pandas	Data munging, dataframes, and working with metadata	http://pandas.pydata.org
	scikit-learn	Topic modeling and supervised machine learning	http://scikit-learn.org/stable/
	pyLDAvis	Visualizing topic models	https://pypi.python.org/pypi/pyLDAvis
R	tm	Preprocessing text, frequency counts, and TF-IDF weighting	https://cran.r-project.org/web/packages/tm/index.html
	openNLP	Natural language processing	https://cran.r-project.org/web/packages/openNLP/index.html
	lda	Topic modeling	https://cran.r-project.org/web/packages/lda/index.html
	stm	Topic modeling	https://crn.r-project.org/web/packages/stm/index.html
	LDAvis	Visualizing topic models	https://cran.r-project.org/web/packages/LDAvis/index.html

Note: This table is up to date as of December 2016. Because this software changes rapidly, researchers should investigate the most up-to-date options when starting a project. Both Python and R are open source and free and are the best options to date to implement transparent, reproducible, flexible, and up-to-date text analysis techniques. For a quick tutorial on Python, see https://www.codeschool.com/courses/try-python; and for a quick tutorial on R, see https://www.codeschool.com/courses/try-r.

lead to extracting meanings embedded in the text. While the output still must be interpreted by humans, computational exploratory analysis can suggest categories relevant to the text that researchers had not previously considered because of their preconceived notions about, or the complexity of, the text (Grimmer and Stewart 2011) and can help researchers avoid their biases and the natural volatility that comes with reading large bodies

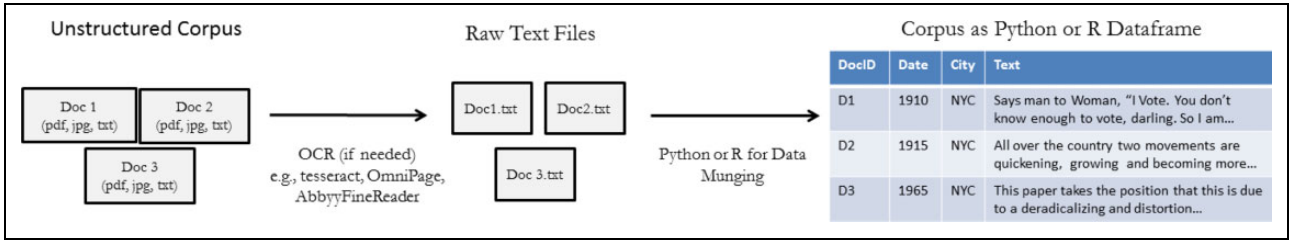

Figure 1. Corpus construction: From text to dataframe. This figure demonstrates a possible path from a collection of texts, saved in separate files, to a digital dataframe suitable for further computer-assisted text analysis techniques. Often historical texts are in the form of pdf or jpg images and thus require an intervening step using optical character recognition software. More contemporary texts are already digitized. Once digitized, the researcher can use Python or R to transform the separate files into one dataframe, with metadata attached to each text (in this example, date of publication and the city in which it was published).

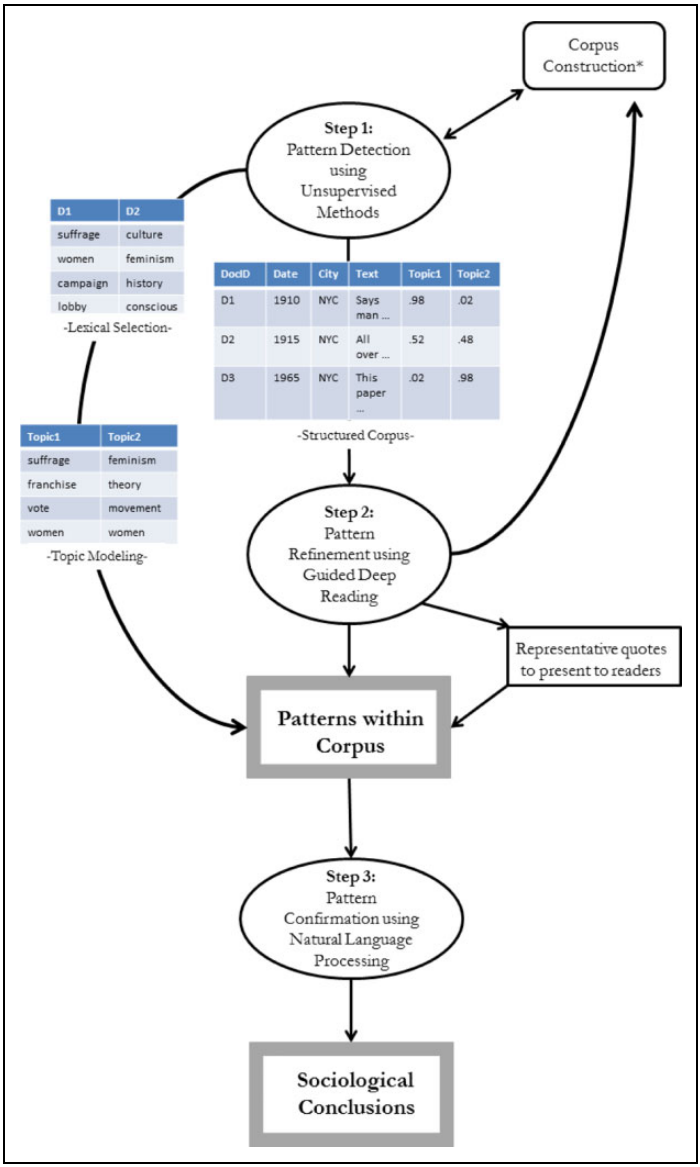

Figure 2. Three-step computational grounded theory framework: From dataframe to conclusion. This figure graphically represents the three-step computational grounded theory process. Step 1 serves two purposes: It outputs interpretable lists of

of text. Similar to traditional coding, the techniques used in this first step also classify text into categories. While this step still involves a number of subjective judgment calls, similar to traditional coding, these decisions are written directly into the process of computer-assisted coding, so, unlike human-coded text, the output of computationally coded text is fully and immediately reproducible.

In sum, in proper grounded theory form, these techniques can help researchers discover new ideas, codes, or concepts while remaining grounded in the data (Glaser and Strauss 1999), but the computational portion brings the field closer to reproducible and scientifically valid grounded theory.

Reducing complex texts to informative groups of words can be done using all three categories of computer-assisted text analysis techniques—lexical selection, classification, and natural language processing—and can additionally be done with guiding input from the researcher. I cover two of these techniques, lexical selection and unsupervised machine learning, here.

Lexical Selection. Lexical selection techniques are the most simple in the automated text analysis tool kit and aim to identify the important content words in a corpus. These techniques can be used to quickly summarize or compare groups of texts without assuming any prior knowledge of the text. One approach to reducing complicated text to informative groups of words, these methods can suggest important differences in the ways in which issues are discussed in different texts. This is typically done by weighting words by their frequency in one document or set of documents compared to the words' frequency in the entire corpus. Words that are frequent in one document and do not occur across many documents are considered defining of that document (or group of documents). Words that occur in all documents will not help distinguish documents.

One calculation that quickly compares two categories is a difference of proportions analysis, which simply calculates the difference in proportional word frequencies between two texts (Monroe et al. 2008). Words with the largest positive and negative differences are distinctive of each text.[6] As a

Figure 2. (continued). words, as shown on the left side of the figure, and produces a structured corpus which can guide the deep reading step (step 2) as shown in the center of the figure. The first two steps, including potentially collecting more data, are iterative and can be done as many times as needed to identify useful patterns in the data. Step 3 is the final confirmation step that formalizes the patterns identified in steps 1 and 2 and allows the researcher to draw reliable conclusions from the data.

concrete example, take two hypothetical documents, D1 and D2, each containing 750 words. Assume the word *about* occurs in D1 100 times with a proportional word frequency of 13.3 percent, and in D2 95 times, with a proportional word frequency of 12.7 percent. The difference of proportions for the word *about* is 0.6. The word *suffrage*, alternatively, occurs 50 times in D1, with a proportional frequency of 6.7 percent, and 5 times in D2, with a proportional frequency of 0.7 percent. The difference of proportion for the word *suffrage* is thus 6.0. Even though the word *about* is used proportionally more in both documents, its low difference of proportion indicates it does not distinguish one document from the other. The word *suffrage,* alternatively, while occurring less frequently in both documents, has a higher difference of proportions score and is thus a distinguishing word for D1. In other words, the word *suffrage* better suggests the content of D1 than the word *about*. If these examples were carried further, we would end up with a list of words that distinguish D1 from D2. Instead of reading each document and summarizing or coding each one, the researcher can now simply interpret these lists of weighted words. Furthermore, unlike coded documents, these lists of words are easy to display in a table and can be included for the reader to view and evaluate. I return to this technique below.

Unsupervised Text Classification. A more sophisticated way to identify patterns across text is by using clustering and topic modeling algorithms, which, instead of simple word frequencies, use the co-occurrence of words in documents to uncover themes within a corpus. These methods, called unsupervised text classification, simultaneously estimate themes or topics within a corpus and classify individual documents, or portions of documents, into those categories. The output from these algorithms is a list of either the most frequent words per cluster or the highest weighted words per topic. These lists of words can suggest the content of a cluster or topic, and these lists as a whole can also quickly summarize or visualize a large corpus.

Because unsupervised text classification shifts the moment of interpretation from creating categories to interpreting estimated categories, like lexical selection techniques, it moves researchers one step away from the data and from their accompanying cultural and historical biases. Furthermore, these algorithms will (usually) classify texts the same way every time, making the classification step fully reproducible.[7]

Topic modeling is one popular way to carry out unsupervised text classification.[8] Briefly, the intuition behind topic modeling is that each document in a corpus is produced or "structured" from a set number of topics. Topic

modeling algorithms analyze the co-occurrence of words within a document over a large number of documents to, in effect, reverse engineer these topics from the larger corpus. More practically, topic modeling algorithms, like lexical selection methods and clustering algorithms, serve to reduce a complicated corpus to simpler, interpretable, groups of words. The output of a topic modeling algorithm is lists of weighted words, where each list is a topic and where higher weighted words in a list are more indicative of that topic, and it represents each document as a distribution over topics, which can be used to detect thematic patterns across documents.

Topic modeling is gaining traction in sociology for three main reasons. First, because it is automated, it can be quickly applied to any sized corpus. Second, its output is reliably interpretable and recognizable to human readers, that is, the topics tend to align with what sociologists think of when they talk about themes. Third, the popular topic modeling algorithms, like Latent Dirichlet Allocation and Structural Topic Models (STM), assume that documents are constructed from multiple topics and that each individual word can be used in a variety of ways. As DiMaggio et al. (2013) explain, these assumptions closely match assumptions that sociologists, and in particular those who do sociology of culture, have about text and discourse. In sum, the efficiency of topic modeling algorithms, the easily interpretable results, and the justifiable assumptions built into the algorithms make topic modeling a natural tool for cultural sociologists using text as data.[9]

To return to a toy example about suffrage and feminism to make this technique more concrete, imagine a corpus of three documents, D1, D2, and D3, where D1 is about suffrage, D2 is about both feminism and suffrage, and D3 is about feminism (note that this toy example would not work in practice, as topic modeling requires hundreds, if not thousands, of documents to be effective). We can use a topic modeling algorithm with two topics to reverse engineer the two themes of feminism and suffrage. If accurate, this topic model will output two weighted lists of words. In one list, T1 (topic 1), the highest weighted words might be *suffrage*, *franchise*, *vote*, and *women*, which would indicate these words co-occur in more documents than you would expect if the words were distributed at random. These words together would suggest the "suffrage" topic. The other list, T2, might have highest weighted words such as *feminism*, *theory*, *movement*, and *women,* which would suggest the "feminism" topic. Topic modeling also outputs each document's distribution over all topics. In this example, D1, which we know is about suffrage, would have a document distribution weighted toward T1; D2, which is about both suffrage and feminism, would have weights more evenly

distributed between the two topics; and D3, which we know is about feminism, would be weighted toward T2. Reading documents that have high weights for their respective topics would give us more insight into the content of each topic.

While sociologists have found topic modeling helpful, these algorithms have come under a lot of criticism outside of sociology for their inaccuracy, exceedingly naive assumptions, and poor predictive performance (Chuang et al. 2012, 2013; Lancichinetti et al. 2014).

One set of problems plaguing topic modeling is the number of qualitative decisions required on the part of the researcher. In the above example, we knew the number of topics before calculating the model. In practice, researchers do not know the number of topics in a corpus. There is much debate and disagreement over how to determine the number of topics to specify in a topic model. While there are some mathematical approaches to determining the number of topics, for social scientists, who are typically motivated by a particular substantive research question, the best way to determine the number of topics is by the usefulness of the output. This is generally done by examining the weighted word lists for a number of models, with a different number of topics prespecified for each (e.g., 20, 30, or 40 topics), to determine which model produces the most semantically coherent and substantively interpretable topics (DiMaggio 2015).[10]

In addition to the problem of choosing the number of topics for each model, topic modeling requires a number of other preprocessing choices on the part of the researcher, and there are not yet clear guidelines on how to choose these parameters. Additional decisions made by the researcher include whether or not to remove stop words,[11] to stem words,[12] to exclude frequent and infrequent words, and more. Each of these decisions will again produce different groupings. Given these multiple choices, there are many ways just one algorithm will group words into topics, and there are multiple algorithms available, producing hundreds of ways to reasonably group a single corpus (Grimmer and Stewart 2011). While there have been some attempts to provide guidelines about these decisions, the general conclusion is that the output should be judged based on how helpful it is to the researcher. It is difficult to argue that any one approach is "the best" way to group the corpus, particularly when researchers are asking different questions of the same corpus. There is at present no objective way to determine the single best model for a text, and the "objective" methods that have been proposed are often not the most substantively helpful (Blei 2012; DiMaggio 2015).

The critical literature in general agrees that if the goal is to use a computer to replicate the hand-coding of documents, or to most accurately place texts

into reliable categories, other supervised machine learning algorithms or more simple clustering algorithms are more accurate than most topic models. If the goal, however, is to perform an initial and fully inductive analysis of thousands of pages of text and visualize the output, topic modeling can be exceptionally helpful. Furthermore, the weaknesses of these tools can be ameliorated by steps 2 and 3 in the computational grounded theory process, where researchers refine the patterns identified by topic modeling with guided deep reading and then confirm them with additional computational techniques.

In sum, unsupervised topic modeling is an excellent research tool for some purposes but not for others and so should always be used with caution. These algorithms offer good tools to quickly summarize the main themes in a corpus so that researchers can make broad comparisons between groups of texts. They can help researchers look at their data in new and perhaps surprising ways, and they can sometimes suggest categories not immediately apparent to human readers. The goal when using unsupervised machine learning should not be to reproduce an existing coding scheme, and it should not be used to identify a particular topic of interest; the goal should rather be to encourage different ways of thinking about and categorizing text. In short, it should most often be used as this first reproducible pattern detection step and should be followed by the two further steps detailed below.

Example: Women's Movements. I use a real-world example from my research on women's movements to illustrate how the two techniques described above, lexical selection and text classification, can be combined to inductively but computationally uncover patterns within a corpus. I return to this example in the next two steps below.

The question motivating this research centered on explaining why a similar debate divided the first and second wave feminist movements in the United States. The existing literature on women's movements has typically claimed the politics of the second wave movement (1964 to the early 1980s) were distinct from the first wave (early 1800s to 1920; Cott 1987; S. Evans 1980; Rosen 2000). I show instead that geographical differences within the first and second waves drove a similar debate within each wave, and these geographical differences persisted over time. In short, geography trumped time in determining the politics of women's organizations. To demonstrate this, I collected the literature produced by women's organizations in two cities that bounded the major debates within each wave—Chicago and New York City—from both the first and second waves. The goal of the content analysis, done on this literature, was to uncover the underlying cognitive frameworks, or political logics (Armstrong 2002), shaping the political

stances of women's movement organizations in these two cities and two time periods.[13] The methodological challenge was to inductively but reliably identify these logics using literature from two time periods that sometimes used remarkably different language and addressed different political issues and to do so in a way that was transparent and reproducible.

In the first step of my analysis, the pattern detection step, I used a combination of difference of proportions (a lexical selection technique) and STM (a structural topic modeling algorithm) on the literature produced by four core women's movement organizations, Hull House in Chicago and Heterodoxy in New York City in the first wave, and the Chicago Women's Liberation Union (CWLU) in Chicago and Redstockings in New York City in the second wave (see Nelson 2015 for an explanation of how I identified the four organizations).[14] I began the analysis with four pairwise difference of proportion calculations to extract the most defining words for each pair of organizations. Table 2 presents the most distinctive words for two pairwise comparisons using the difference of proportion analysis. These words can be analyzed to suggest patterns within the corpus (which I do below).

To further explore the themes addressed in this literature and to categorize the text into those different themes, I followed this analysis with structural topic modeling. To determine which model was best for my corpus, I ran four topic models, with 20, 30, 40, and 50 topics, respectively. Examining the top weighted words for each model, I found some of the topics in the 20-topic model combined two issues into one. For example, one of the topics in the 20-topic model had the top weighted words: *car*, *can*, *women*, *doctor*, *gonorrhea*, and *infect*. I found these words combined the issue of car maintenance and the issue of sexually transmitted infections, meaning the 20-topic model had too few topics. Conversely, I found the 50-topic model produced multiple topics on the same issue. For example, the top three weighted words in one topic in the 50-topic model were *class*, *year*, and *art*, while in another topic they were *hullhouse*, *children*, and *school*, and in yet another they were *school*, *class*, and *boy.* I interpreted these three topics to all be about one issue: different types of classes offered by Hull House. For my purposes, I was looking for more general topics than the specific types of classes Hull House taught, so I determined the 50-topic model was too specific. The 40-topic structural topic model, alternatively, produced topics that were comfortably distinct from one another, yet general enough to be interpretable for my purposes, so I used this model for my analysis (Table 3 summarizes this model). Notably, however, many topics were comparable across all of these four models (e.g., abortion, the Vietnam War, movement history, and legal issues), so I do not anticipate the results to be substantially different if I had

Table 2. Most Distinctive Words, Difference of Proportions.

First Wave[a]		Second Wave[b]	
Hull House (Chicago)	Heterodoxy (New York City)	CWLU (Chicago)	Redstockings (New York City)
hullhouse	woman	chicago	movement
club	man	children	women
miss	women	center	men
school	life	union	radical
given	know	school	feminist
year	world	work	male
members	like	cwlu	political
chicago	sanger	vietnam	history
mr	men	nixon	womens
classes	said	people	feminism
house	home	office	revolution
boys	just	day	love
work	say	health	feminists
years	don't	city	left
social	little	working	power
held	way	vietnamese	oppression
clubs	think	legal	class
mrs	things	war	female
residents	want	care	personal
room	sex	womankind	woman
children	right	government	really
evening	masses	workers	consciousness
neighborhood	make	south	consciousness-raising
italian	things	medical	group
building	good	home	theory
various	business	hospital	groups
plays	law	rape	action
summer	case	abortion	new
city	control	help	oppressed
association	birth	pay	supremacy

Note: This table presents a list of the most distinctive words in the Hull-House texts compared to the heterodoxy texts, and the CWLU texts compared to the Redstockings texts, using a difference of proportions calculation. CWLU = Chicago Women's Liberation Union.

[a]Words with highest and lowest difference of proportions, Hull-House literature - heterodoxy literature. Words with the highest difference of proportion are distinct to Hull House, while the words with the lowest difference of proportions (i.e., the largest negative difference) are distinct to heterodoxy.

[b]Words with highest (CWLU) and lowest (Redstockings) difference of proportions, CWLU literature - Redstockings literature. Words with the highest difference of proportion are distinct to CWLU, while the words with the lowest difference of proportions (i.e., the largest negative difference) are distinct to Redstockings.

Table 3. Top Topics with Highest Weighted Words by Organization.

Hull House (Chicago, First Wave)			Heterodoxy (New York City, First Wave)			CWLU (Chicago, Second Wave)			Redstockings (New York City, Second Wave)		
Public Institutions[a] (28%)[b]	Hull-House Social Activities (27%)	Hull-House Practical Activities (18%)	Sanger and Birth Control (22%)	Women's Resistance (22%)	Women's Lives (9%)	Liberation School (8%)	Antiwar (7%)	Women's Sexual Health (6%)	Movement History (11%)	Movement Theory (9%)	Forms of Resistance (9%)
hullhous[c,d]	club	hullhous	sanger	woman	one	women	vietnam	women	movement	radic	women
miss	year	play	one	women	love	liber	vietnames	gonorrhea	women	liber	men
school	member	year	will	man	will	work	people	doctor	feminist	women	liber
children	miss	given	public	will	mother	cwlu	war	infect	histori	polit	movement
hous	boy	greek	birth	suffrag	life	union	american	can	radic	feminist	male
chicago	social	italian	inform	men	day	chicago	south	pain	femin	attack	group
resid	mrs	lectur	can	one	littl	call	north	treatment	liber	movement	organ
year	hullhouse	build	time	life	man	offic	nixon	drug	polit	group	struggl
work	even	meet	year	great	know	peopl	bomb	diseas	new	consciousnessrais	revolut
offic	parti	organ	new	world	billi	center	govern	patient	lesbian	left	oppress
summer	meet	dramtic	give	sex	woman	chang	will	caus	even	issue	work
public	room	music	life	home	work	will	prison	pill	one	power	fight
citi	two	present	control	suffragett	came	legal	one	bacteria	first	action	now
street	given	audienc	mrs	like	take	can	can	penicillin	time	person	right
open	committee	mani	law	say	well	societ	peac	vagina	idea	peopl	chang
made	present	school	book	vote	ladi	womankind	agreement	tube	origin	problem	equal
neighborhood	danc	russian	make	can	like	problem	forc	symptom	now	psycholog	polit
investig	entertain	club	pamphlet	new	mickey	come	militari	uterus	year	theori	must

(continued)

Table 3. (continued)

Hull House (Chicago, First Wave)			Heterodoxy (New York City, First Wave)			CWLU (Chicago, Second Wave)			Redstockings (New York City, Second Wave)		
Public Institutions[a] (28%)[b]	Hull-House Social Activities (27%)	Hull-House Practical Activities (18%)	Sanger and Birth Control (22%)	Women's Resistance (22%)	Women's Lives (9%)	Liberation School (8%)	Antiwar (7%)	Women's Sexual Health (6%)	Movement History (11%)	Movement Theory (9%)	Forms of Resistance (9%)
associ	mani	one	woman	social	time	togeth	viet	birth	media	interest	radic
visit	one	chicago	case	never	hand	abort	saigon	examin	write	oppress	issue

Note: Top 3 most prevalent topics from each organization calculated using the 40-topic Structural Topic Models. The words are the top weighted words or most distinctive words, for each topic, which suggest the content of the topic. The topic was labeled by me, done by examining the top words and representative documents for each topic to suggest the content of each topic. See Online Appendix A for excerpts from the top weighted documents for each of these topics. CWLU = Chicago Women's Liberation Union.

[a]Topic labels chosen by me.

[b]Percentage of words from the organization's literature aligned with the topic. This was calculated for each organization and each topic by multiplying the topic weight by the number of words for each document, summing this result across all documents, and dividing by the total number of words.

[c]Top weighted words for the topic.

[d]Words were stemmed using the Porter stemmer.

Table 4. Top Weighted Words for One Topic across Multiple Structural Topic Models.

	20-Topic Model	30-Topic Model	40-Topic Model	50-Topic Model
Top weighted words	abort	abort	abort	abort
	women	law	women	hospit
	law	women	law	center
	can	state	doctor	medic
	doctor	medic	medic	matern
	will	doctor	hospit	women
	control	legal	will	doctor
	one	right	woman	chicago
	infect	will	can	care
	woman	court	state	new

Note: These are the top weighted words for one topic from four different Structural Topic Models, with 20, 30, 40, and 50 topics specified. This topic, which these words suggest is about abortion, is roughly similar across all four models evidenced by similar top weighted words such as *abort*, *law*, *state*, and *doctor*. Words were stemmed using the Porter stemmer algorithm.

chosen a different model (see Table 4 for an example of a topic that was comparable across all models).

Table 3 shows the highest weighted words for the 12 most commonly occurring topics represented in the literature from each of the four organizations in my study produced via the 40-topic STM. As is common practice, the topics were labeled by me, but this is not a necessary step. This model also outputs a distribution for each document across all topics, which enables researchers to quickly identify the most representative document for each topic. This output is important for step 2 below.

The weighted lists of words, alternatively, are important for step 1. By examining the lists from the difference of proportions analysis and the STM (as shown in Tables 2 and 3), I found the words that scored high in the New York City literature were often abstract and general (e.g., *history*, *liberation*, *feminist*, *will*, and *like*), namely, they did not refer to things that can be experienced through the senses. Alternatively, the words that scored high in the Chicago literature were more concrete and specific (e.g., *abortion*, *Nixon*, *hospital*, *school*, and *members*); they referred to entities or things that can be directly experienced through the senses.

These word patterns, abstract and general versus concrete and specific, suggest a different type of political discourse in each city. This pattern was not apparent to me when reading through the text sentence by sentence

without doing any computational work, but when these words are taken out of context and grouped into the above lists, they revealed this potential pattern. Against the commonly accepted account of U.S. women's movements that would predict more similarities within each wave and differences between the two waves, this analysis suggested that there was more similarity in the politics within each city over time and more differences between the two cities within each wave.

I come back to this inductively identified pattern in the next two steps below, the pattern refinement and pattern confirmation steps.

Computational Exploration in Sum. Taken together, this first, computational pattern detection step serves two purposes. First, as these computational methods decontextualize and simplify text in key ways, they can reveal patterns in the text not immediately available to human readers and they can encourage researchers to either view their data in new and perhaps surprising ways and/or reveal new directions to take an analysis. Computers, in effect, make visible that which humans do not see. Second, as these techniques parse text quickly and reliably, they are both more efficient than a researcher reading through the text and they are completely reproducible. These techniques can simplify and reveal patterns in data of almost infinite size without much added work on the part of the researcher. They also structure the text to allow for a reengagement with the data, which is step 2 of the computational grounded theory process.

Step 2: Hypothesis Refinement Using Human-centered Interpretation

Grounded theory involves moving back and forth between the results of the analysis and the data. Computational grounded theory involves the same process. In the second step, researchers return to the data via a structured qualitative analysis to do three things: confirm the plausibility of the patterns identified via an analysis of the computationally driven results, add interpretation to the analysis, and potentially modify the identified patterns to better fit a human, and holistic, reading of the data.

Computationally Guided Deep Reading. Most corpora of interest to sociologists are too large to read everything or to read in a sustained and systematic way. It is through deep reading that researcher bias can often seep into an analysis, as researchers can, consciously or unconsciously, give more weight to passages that confirm their previously held belief about the data, and ignore passages that challenge their beliefs. Deep reading, however, is a

necessary step toward an interpretive understanding of text. A benefit of using topic modeling is that the text is now thematically (and reproducibly) classified. Through these algorithms, the researcher can mathematically identify texts that are representative of a particular theme or category, and they can additionally be used to calculate the relative prevalence of that category. By carefully using these models to choose representative texts, researchers can "read" and interpret any amount of text without the burden of reading the full text. Both the researcher and reader, additionally, can trust that when a quote is chosen as an example of something, it is not an outlier but is indeed representative of some theme in the text.[15]

Through this kind of guided deep reading, researchers can check their interpretations of the groups of words produced in the quantitative step, and they can also better determine how those groups of words translate into full sentences or arguments. The guided reading step can additionally either confirm or revise the patterns identified in the first step. Because this step is both guided and backed up by numbers (including the pattern confirmation step described below), both the researcher and the reader can be more confident in the particular interpretation developed by the researcher. This makes the process more efficient, but it also ensures the researcher will not skip over important passages because of fatigue or bias. Conversely, because this step involves a human actually reading the text, the numbers are given context and interpreted in a meaningful, more traditional sociological and theory-informed fashion.

Example: Women's Movements. In the women's movement organizations research, I followed the inductive computational analysis with a guided deep reading of the text. This reading contextualized the patterns identified from the results of the computational analysis, showing, for example, how abstract words translated into abstract political arguments but also added to the pattern.

Computationally guided deep reading utilizes the topic distribution for each document. To determine which documents are most representative of a topic, the researcher can simply sort the output in descending order for the topic of interest. Table 5 shows example output sorted according to the *movement history* topic from the 40-topic STM calculated in step 1 (see the *Movement History* column in Table 3 for the list of words associated with this topic). The researcher can then easily read the top documents for each topic, knowing their reading is targeted toward that topic, and can then calculate, and present to the reader, truly representative quotes.

Table 5. Sample of Structured Dataframe Sorted by *Movement History* Topic Weights.

File Name	Text (Head)	Movement History Topic Weight	Antiwar Topic Weight
nyc.redstockings.1973.sarachild.powerofhistory-28.txt	THE ARCS OF HISTORY tific and fearless writers of her day" and Elizabeth Cady Stanton, too, "the matchless writer." [. . .]	0.967030332	0.00031311
nyc.redstockings.1973.sarachild.powerofhistory-27.txt	Ten any idea of what that work was all about "it's purpose and the breadth of its contents and even its method-we [. . .]	0.963596868	0.000113575
nyc.redstockings.1973.sarachild.powerofhistory-29.txt	Paraging depiction of the History in the bibliography, it suddenly struck m_e that Stanton, Anthony and Gage's [. . .]	0.947334513	0.000125144
nyc.redstockings.1973.sarachild.powerofhistory-30.txt	Said it was. And those who did take action in the area of history "especially for the present record"did not [. . .]	0.940307645	0.000347795
nyc.redstockings.1973.sarachild.powerofhistory-26.txt	And so their absence from history books meant there weren't any. We had to discover the problem of [. . .]	0.91508104	2.90019E-05
nyc.redstockings.1973.sarachild.powerofhistory-18.txt	Opposite of Beauvoir's book which was disappearing from the lists. I soon learned, even without reading it, [. . .]	0.864953051	0.000107619
nyc.redstockings.1973.sarachild.powerofhistory-31.txt	POSTSCRIPT The history question had a lot to do with the leadership question. History, after all, is all about [. . .]	0.864581881	0.000165923
nyc.redstockings.1973.sarachild.powerofhistory-25.txt	Most of the theories citing history that we encountered from both the right and the left essentially counseled [. . .]	0.816334809	9.56399E-05
nyc.redstockings.1973.sarachild.powerofhistory-02.txt	The Power Of History I am obnoxious to each carping tongue Who says my hand a needle better fits, A poet's [. . .]	0.756516182	0.000251796
nyc.redstockings.1973.sarachild.powerofhistory-17.txt	Of the key elements of Beauvoir's analysis upon which the WLM later built its work included: 1. Women have [. . .]	0.70522738	0.000726348

Note: This is a sample of a structured dataframe (see Figures 1 and 2) structured via a 40-topic Structural Topic Model. The topic weights indicate the percentage of total words from each document related to each topic (e.g., 97 percent of the total words in the first document are related to the *movement history* topic, as indicated by the "Movement History Topic Weight" column, while close to 0 are related to the *antiwar* topic). Doing a descending sort by the *movement history* topic quickly indicates which documents are most representative of this topic. As the text is included in the dataframe, the researcher can quickly read these documents to better understand the content of each topic. In this particular example, all of the top documents for the *movement history* topic are from the Redstockings literature, starting with the article "The Power of History."

In my analysis of women's movements, I first printed excerpts from the top two documents for each of the top 12 topics from the STM (see Online Appendix A), so I and the reader can get a sense of each topic.[16] I then did a qualitative deep reading of the top 10 documents for each of these topics as well as representative documents from the remaining topics. As I read, I looked for ways in which the patterns identified in step 1 (abstract and general words in the New York City literature, and concrete and specific words in the Chicago literature) translated into full sentences and political statements, but I also interrogated those patterns to determine if my interpretation of the computational output was valid and to look for additional patterns in the data.

Through this reading, I found that women in Chicago in both waves had a particular approach to politics in which they worked to identify concrete needs of women in their community, such as childcare or legal counseling, and then they worked to either directly meet those needs or lobbied the state to meet those needs. Women in New York City, on the other hand, were more concerned with detailing experiences of individual women in order to generalize and abstract from those experiences to make claims about social structures affecting women's lives. To illustrate these different approaches, I provide representative quotes from the documents around one of these themes.

The documents produced by Hull House and CWLU detailed a variety of campaigns these organizations pursued and services they offered to make women's lives easier. Hull House, for example, initialized a number of community services that they eventually incorporated into official city institutions:

> We had maintained three shower baths in the basement of the house for the use of the neighborhood, and they afforded some experience and argument for the erection of the first public bath-house in Chicago, which was built on a neighboring street and opened under the care of the Board of Health. It is immediately contiguous to a large play-ground which was under the general management of Hull-House for thirteen years and has lately been incorporated in a city play-ground. The Reading Room and Public Library Station which was begun in the house is continued only a block away. The lending collection of pictures has become incorporated into the Public School Art Society of Chicago. The summer classes in wood work and metal, formerly maintained at Hull-House, are discontinued because they are carried on in a vacation school maintained in the Dante Public School. (Hull-House 1907 :54)

The public bath house was eventually attached to the Board of Health, the Hull House playground became a city playground, and summer classes

offered by Hull House were moved into the Dante Public School system. We know that this quote is representative of around one fifth of the documents produced by Hull House by using the document distribution over topics and the topic distribution over the corpus. Ninety-eight percent of the words in the above document are related to the *Public Institutions* topic (see Table 3), and 28 percent of the total number of words in all of the documents produced by Hull House are related to this topic.

We see similar types of documents in the CWLU literature, which outline a slew of services the organization offers women:

> The Edgewater Women's Center of the Chicago Women's Liberation Union has open rap sessions every Thursday evening at 8 pm. There is also a pregnancy testing program every Saturday from 11 am to 1 pm at the Center 5412 N. Clark . . . . The Pregnancy Testing Service has provided pregnancy tests at cost to a large number of women. Now in a new location at Augustana Lutheran Church, 55th and Woodlawn, we plan to expand our services in the fall to include VD testing, pap smears, more extensive medical referrals, and several courses on women and their bodies, one of which might be for high school women. Eventually, we hope to organize a political action project which will work to win better services from the medical professionals. (CWLU 1972:2)

Like Hull House, CWLU worked to meet the concrete needs of the community, with the goal of transferring these services to official city institutions. To put this quote in perspective, 98 percent of this document is related to this, *Liberation School* topic, and 8 percent of the total words produced by CWLU are related to this topic.

As the above demonstrates, I used the output from the STM to dig deeper into a few of the most prevalent topics in my data. In practice, any topic can be analyzed in this fashion—including topics that are not prevalent—to interrogate and interpret the patterns identified in step 1. In my example, I claim the two passages quoted above demonstrate how concrete words are translated into a complete political discourse. I did the same with the New York City articles, better understanding and demonstrating to the reader how the pattern identified in step 1 indicates different underlying approaches to politics in these two cities.

In addition to translating and interpreting the output from the first step in light of the data as a whole, my reading in this step indicated another pattern across these texts: In addition to being more abstract, the literature in New York City more often mentioned individuals, while in addition to being more

concrete, the literature in Chicago more often mentioned specific institutions and organizations. I return to both of these patterns one more time below.

Pattern Refinement in Sum. Step 2 in this framework brings the researcher and interpretation back to the data, replicating the traditional approach to grounded theory, but with a computational twist. With steps 1 and 2 combined, the researcher can move between the analysis, or the output from computational techniques, and interpretive readings of the text to refine their analysis of the data. To recap, the computational portions of these steps serve three purposes: (1) They utilize the potential for computers to extract patterns that may not be immediately obvious to humans, or to make visible that which humans do not see; (2) the output is immediately reproducible, allowing other researchers to reproduce the computational portion of the analysis so they can test the interpretation of the output without having to laboriously reproduce a coding scheme; and (3) the techniques are fully scalable and can thus incorporate "big" data. The interpretive portion translates the computational output into sociologically meaningful concepts to enable researchers to draw more abstract conclusions about the social world that produced the data. Following the grounded theory framework, these two steps can also point researchers to the need to collect additional, or different, data in order to draw appropriate conclusions about their subject of interest.

Once data-driven patterns are identified and refined through these two steps, computational techniques assist the researcher in the all-important final step of pattern confirmation.

Step 3: Pattern Confirmation

Through the first two steps, researchers identify patterns in their data by interpreting computational output and through guided deep reading. To ensure the identified patterns are not an artifact of a specific algorithm, or are based on a biased interpretation of the output and deep reading, step 3 deductively tests whether these patterns hold throughout the corpus. This important, final step mitigates some of the challenges inherent in the first two steps. While computational text analysis has many benefits, there are many aspects of natural language, such as humor, irony, or sarcasm, that may not be captured in the output. Step 2 provides a check on the computational output to ensure the more interpretive aspects of language are taken into account, but this interpretive step is based on reading a subset of the full corpus and is still subject to human bias. This third step tests whether the patterns identified in the first two steps are generalizable to the entire

corpus and provides a final reliability test to the grounded theory process. Like causal inference, or causal identification, the conclusions drawn from this step should be done with caution. Often, this step serves to confirm identified patterns rather than to definitively or causally confirm relationships in the text. Nonetheless, this step is an essential check on the first two inductive steps. This step additionally challenges the researcher to operationalize the patterns identified through the first two steps in measurable terms, formalizing patterns and concepts in a way not always done in purely qualitative analyses.

There are a number of ways to computationally confirm patterns within text. Supervised machine learning (Burscher, Vliegenthart, and Vreese 2015; Hanna 2013; King et al. 2013) is a common method used to confirm and calculate themes or patterns in text. To test patterns using supervised machine learning, the researcher would need to code a random sample of documents according to the patterns identified in the first two steps and then use a supervised machine learning algorithm to code the remaining documents. The researcher could then test their hypotheses from the first two steps using these coded documents. As supervised machine learning relies on hand-coding text, this method can be applied to most patterns identified by the researcher. Hand-coding, however, is difficult to reproduce. In some cases, testing patterns can also be done using dictionary methods (e.g., Tausczik and Pennebaker 2010) and natural language processing tools. Three of these tools are surveyed here, but there are more available. Additionally, new techniques are rapidly becoming available that may eventually aid in this step of the process.

In my example, I identified two patterns in the corpus which I will test in this step. *Pattern 1*: The Chicago texts contained more concrete and specific words compared to the New York text, which contained abstract and general words. *Pattern 2*: The Chicago texts more often mentioned organizations and the New York text more often mentioned individuals.

To test pattern 1, I used two techniques, word hierarchies using WordNet (Princeton University 2010) and a crowd-sourced dictionary. The lexical resource WordNet organizes words together based on different types of linguistic relationships. Relationships include synonyms (march and demonstration), super–subordinate relations or hyponyms (nongovernmental organization is a hyponym of organization), part–whole relations or meronyms (kitchen and house), and antonyms (suffrage and disenfranchisement). The goal of WordNet is to map all words (within one language) through these various relationships into one big word network. I use the hypernym relationship in WordNet to measure the level of specificity in a text. In WordNet,

each sense of each noun has set paths of hypernyms (words that are increasingly more broad) to reach 25 root (general) words and each verb the equivalent to 9 root verbs. For example, the word *furniture* has this path to its root word, *entity*:

furniture → furnishing → instrumentality → artifact → whole → object → physical_entity → entity

Compare this to chair:

chair → seat → furniture → furnishing → instrumentality → artifact → whole → object → physical entity → entity

The length of the path to the root word is equivalent to a word's number of hypernyms—the more hypernyms a word has, the more specific it is. *Furniture* has seven hypernyms while *chair* has nine hypernyms, making *chair* more specific than *furniture*. The specificity score of a text is then the average number of hypernyms over all of the nouns and verbs in a text. I used this calculation to demonstrate a difference in distribution of the number of hypernyms per word across the texts produced by the women's movement organizations in New York City versus Chicago. On average, the New York City literature was 1.1 percent more general on the overall hypernym scale. Compare this to a difference of 3 percent of the overall hypernym scale between a sample abstract text (Kant's *Metaphysical Elements of Ethics*) and a sample specific text (the Wikipedia page on Germany). While the difference in the women's movement texts was smaller compared to these sample texts, it was in the expected direction and was statistically significant using an independent samples *t*-test.

To test the contention that Chicago organizations used more concrete language compared to Chicago, I used a crowd-sourced database that contains a human-rated concreteness score for close to 40,000 English lemmas—the dictionary form of a word (Brysbaert, Warriner, and Kuperman 2014). The researchers who created this database used Amazon's crowdsourcing Web site Mechanical Turk to recruit workers to rate groups of lemmas on a concreteness scale, providing an average concreteness rating for each lemma. Researchers have used this database to measure the concreteness of various texts and their relationship to social processes (Snefjella and Kuperman 2015). I used this database to calculate the average "concreteness score" for each publication. The New York City texts were, on average, 6 percent more abstract on the overall concreteness score scale. This was again smaller than the percentage of difference for the sample texts (the difference in the sample

texts, described above, was 16%) but was again in the expected direction and was statistically significant. These two tests are evidence that pattern 1 indeed holds throughout the texts.[17]

To test the hypothesis that the New York City literature mentioned more individuals while the Chicago literature mentioned more organizations, I used named entity recognition. Named entity recognition is a subset of part of speech taggers, which utilize the grammatical structure of a sentence to categorize individual words into a part of speech, including different types of named entities such names of persons (e.g., Shirley Chisholm), organizations (e.g., National Woman's Party), locations (e.g., New York City), and other categories like monetary expressions (e.g., 2,000 dollars). I simply counted the number of individuals versus organizations mentioned in the texts and found that the New York City organizations mentioned more individuals compared to organizations (2,775 individuals compared to 1,799 organizations), and the Chicago organizations mentioned more organizations compared to individuals (5,567 organizations compared to 3,421 individuals), confirming the inductively identified pattern.

In my research, I used various natural language processing techniques as well as one dictionary method to confirm and provide more evidence to support the patterns identified in the first two steps. There is a certain amount of creativity in this last step, however, and successfully applying this step to other research projects requires general knowledge of the range of text analysis and natural language processing tools available. The three methods mentioned above worked on my data, but for others, different tools are needed. Researchers should be aware of the range of tools available as they construct this crucial final step.

The Result: Inductively Identified Political Logics and Theory Generation

With reliable patterns identified computationally backed up with expert interpretation from the guided deep reading and then confirmed using further computational techniques, the researcher can now bring the analysis together to summarize their more abstract, and in some cases, theory-building conclusions about social reality. In my example, I ended the analysis by constructing the different political logics underlying the women's movements fields in New York City and Chicago from 1865 to 1975. The women's movement organizations in Chicago over both waves, I claim, shared a political logic that assumed social change happens through institutions and the state and is achieved through short-term goals around particular issues

that win concrete changes that affect women's lives. The organizations in New York City followed an alternative political logic, one that assumed social change happens through individuals, and is achieved through building solidarity based on generalizing the experiences of individual women and mobilizing individual consciousness through abstracting from these experiences to make claims about social structures. I concluded that political models institutionalized in the first wave did not disappear when the women's movement retreated, and the same model guided the politics of women's movement organizations in the second wave. This conclusion was based on a combination of computational techniques and qualitative deep reading, and, importantly, the entire analysis is easily and quickly reproducible (see Figure 2 for a diagram of the entire computational grounded theory process).

Conclusion

Researchers in other social science disciplines are using computer-assisted text analysis to complement traditional content analysis, and some are relying solely on computer-assisted text analysis. Sociologists, in particular, sociologists who study meaning, often ask questions from their data that require interpretation, a task that computers have not yet been successfully programmed to accomplish. In this article, I propose a general framework to incorporate computational methods into inductive sociological content analysis, a framework I call computational grounded theory. This mixed-methods framework adapts the theory-building and interpretation-rich tradition of grounded theory to contemporary data-rich questions, by incorporating methods to make it more efficient, reliable, and reproducible. In addition, because this framework provides a method to calculate how prevalent or representative each pattern is within the larger corpus, I argue this method is more valid than traditional grounded theory, which asks the reader to simply trust the representativeness of particular examples or quotes.

As with any method, researchers using computer-assisted text analysis techniques should understand the range of methods available and choose ones that are best suited to the research question and available data. In addition to the range of techniques proposed in this article, from lexical selection to classification and computational natural language processing, computer scientists and computational linguists are continually adding more. As sociology increasingly incorporates computer-assisted text analysis methods into the content analysis umbrella, best practices around the use of these techniques in sociology should continue to develop and evolve. These best

practices, however, should always remain grounded in the disciplinary knowledge developed within sociology.

Author's Note

Readers can find a replication repository at https://github.com/lknelson/computational-grounded-theory.

The repository includes replication data as well as all of the code used in the analyses in this article, with accompanying README files.

Acknowledgments

The author would like to thank Leslie McCall, Kim Voss, Brayden King, Charles Seguin, Brandon Gorman, Aliza Luft, and Robert Braun for their helpful comments on drafts of this article. The author would also like to thank Berkeley Research Computing and D-Lab at UC Berkeley for providing computational resources to carry out the analyses in the examples used.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The author received support from the National Science Foundation and the University of California, Berkeley, for the research, authorship, and/or publication of this article.

Supplemental Material

Supplemental material for this article is available online.

Notes

1. More often, researchers are providing supplemental material describing their coding process, which is greatly helping to demystify qualitative text analysis (see, e.g., H. E. Brown 2013).
2. Natural language processing is, in the technical literature, an umbrella term incorporating any analysis of natural languages including counting and machine learning. In practice, however, I have found that text analysts in the social sciences tend to identify "natural language processing" with techniques that incorporate some aspect of the structure of language or words. They also distinguish this from machine learning or dictionary methods that treat words (or other features such as a word's part of speech) simply as strings of meaningless

characters. I am therefore using natural language processing to refer to what linguists call morphology: techniques that incorporate language structure, such as grammar or syntax, into the analysis. As morphology is a very technical term that is not commonly used by sociologists, I thought it was preferable to maintain the less technical term natural language processing to refer to techniques that incorporate language structure.

3. Refer to note 2 for an explanation of why I am not using natural language processing as an umbrella term.
4. Qualitative data analysis software are starting to incorporate simple automated text analysis tools into their packages, such as word frequency counts and clustering techniques based on word similarity (see, e.g., NVivo). These packages, however, assume that you will be coding the documents yourself (unlike the framework presented below) and do not yet offer more sophisticated techniques such as topic modeling and part of speech tagging. While this software may eventually incorporate all the techniques covered in this article, scripting languages such as R and Python, at least for now, remain the most flexible and adaptable languages to apply the widest variety of computer-assisted text analysis techniques. Software changes rapidly, however, so researchers should investigate the most up-to-date software before embarking on a research project.
5. I did the majority of my data analysis work using the programming language Python, a free and open-source language. While the learning curve for Python is steep, once learned it provides a wide array of flexible tools to do almost any form of text analysis. Researchers new to this language could start with this tutorial to get a basic understanding (https://www.codeschool.com/courses/try-python). R is another option, with equally simple online tutorials (https://www.codeschool.com/courses/try-r). For the specific computer code used throughout this article, see the replication repository at https://github.com/lknelson/computational-grounded-theory.
6. Other methods include term frequency-inverse document frequency (tf-idf) scores (Salton and Buckley 1988) and Dunning (1993) log likelihood.
7. See Lancichinetti et al. (2014) for a critique of the reliability of topic modeling algorithms.
8. Another common technique is clustering, such as k-means, which assign documents to only one category. These techniques are almost always more accurate than topic modeling algorithms when tested on a corpus labeled with predefined categories, for example, correctly clustering or topic modeling a multilingual corpus into its distinct languages (see, e.g., Lancichinetti et al. 2014). Clustering is thus a good option for shorter, thematically focused texts such as Tweets. Complex documents, however, which are often what sociologists encounter, will

contain multiple topics, and as such topic modeling algorithms are more appropriate. Some qualitative data analysis software are incorporating clustering techniques into their packages.

9. See the December 2013 issue of *Poetics* for examples of uses of topic modeling in the social sciences. The journal *Signs* also presents an interesting use of topic modeling found at http://signsat40.signsjournal.org/topic-model/. There are new visualization packages for topic modeling, including LDAvis. An interactive example using LDAvis is found at http://cpsievert.github.io/LDAvis/reviews/vis/#topic=3&lambda=0.6&term=cop.
10. An alternative approach is to run a model with a large number of topics, say 100 or 200, and then hierarchically cluster those topics to find topic clusters of substantive interest to the researcher (Grimmer 2010).
11. Stop words are words that are generally thought to not contain much information, like "an," "the," and "and." There are many different lists of stop words, and some contain words, like "she," "he," and "they," that are important for some questions.
12. Stemming words will combine different tenses of a word into the same "stem." For example, stemming will change "politics" and "political" into the stem "polit," so they will be considered the same word.
13. The data consisted of the public bulletins, journals, or articles produced by four core women's organizations, one in each city and from each wave. This included Hull House's *Bulletin* printed between 1900 and 1920; articles about women and feminism published in the journal *The Masses* and written largely by women in the feminist organization Heterodoxy; all of the articles from the journal *Womankind* written and distributed by the Chicago Women's Liberation Union; and articles from *Notes from the First Year*, *Notes from the Second Year*, and *Feminist Revolution* written and compiled by women from Redstockings. Together, the literature comprises around 1 million words (just over 1,000 pages), so this does not constitute big data, but the method I use to analyze the literature can scale up almost indefinitely.
14. I used the Structural Topic Model (STM) library in R, another free and open-source language. This library is the only software I am aware of that implements this particular algorithm. The Latent Dirichlet Allocation library in R is another topic modeling library, as is the scikit-learn library in Python.
15. Or, alternatively, the researchers can purposively choose an outlier in order to find a counterexample, or counterfactual, in the text.
16. The STM library in R has a command for this, *findThoughts*, which will print the top *n* documents for each specified topic. Alternatively, researchers can output the document by topic distribution to any standard dataframe, including a delimiter separated values file, and then sort the dataframe by the topic of interest (see Table 5).

17. If a dictionary that is relevant to the patterns identified does not already exist, the researcher can create their own dictionary by identifying lists of words relevant to their categories. These manually created dictionaries, of course, should always be validated (see, e.g., Schwartz and Ungar 2015).

References

Alexa, Melina and Cornelia Zuell. 2000. "Text Analysis Software: Commonalities, Differences and Limitations: The Results of a Review." *Quality and Quantity* 34: 299-321.

Armstrong, Elizabeth A. 2002. *Forging Gay Identities: Organizing Sexuality in San Francisco, 1950-1994*. Chicago, IL: University of Chicago Press.

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* 77:855-79.

Bail, Christopher A. 2014. "The Cultural Environment: Measuring Culture with Big Data." *Theory and Society* 43:465-82.

Bearman, Peter S. and Katherine Stovel. 2000. "Becoming a Nazi: A Model for Narrative Networks." *Poetics* 27:69-90.

Biernacki, Richard. 1997. *The Fabrication of Labor: Germany and Britain, 1640-1914* . Berkeley: University of California Press.

Biernacki, Richard. 2012. *Reinventing Evidence in Social Inquiry: Decoding Facts and Variables*. New York, NY: Palgrave Macmillan.

Biernacki, Richard. 2015. "How to Do Things with Historical Texts." *American Journal of Cultural Sociology* 3:311-52.

Blei, David M. 2012. "Probabilistic Topic Models." *Communications of the ACM* 55: 77-84.

Bonilla, Tabitha and Justin Grimmer. 2013. "Elevated Threat Levels and Decreased Expectations: How Democracy Handles Terrorist Threats." *Poetics* 41:650-69.

Brown, Hana E. 2013. "Race, Legality, and the Social Policy Consequences of Anti-immigration Mobilization." *American Sociological Review* 78:290-314.

Brown, Peter F., Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. 1993. "The Mathematics of Statistical Machine Translation: Parameter Estimation." *Computational Linguistics* 19:263-311.

Brysbaert, Marc, Amy Beth Warriner, and Victor Kuperman. 2014. "Concreteness Ratings for 40 Thousand Generally Known English Word Lemmas." *Behavior Research Methods* 46:904-11.

Burscher, Bjorn, Rens Vliegenthart, and Claes H. De Vreese. 2015. "Using Supervised Machine Learning to Code Policy Issues Can Classifiers Generalize across Contexts?" *The ANNALS of the American Academy of Political and Social Science* 659:122-31.

Carley, Kathleen. 1994. "Extracting Culture through Textual Analysis." *Poetics* 22: 291-312.

Charmaz, Kathy. 2014. *Constructing Grounded Theory*. London, UK: Sage.

Chicago Women's Liberation Union (CWLU). 1972. "Chicago Women Are Moving." *Womankind*, September, pp. 2-3.

Chuang, Jason, Sonal Gupta, Christopher Manning, and Jeffrey Heer. 2013. "Topic Model Diagnostics: Assessing Domain Relevance via Topical Alignment." Pp. 612-20 in *International Conference on Machine Learning (ICML)*. Accessed September 21, 2017 (http://machinelearning.wustl.edu/mlpapers/papers/icml2013_chuang13).

Chuang , Jason, Daniel Ramage, Christopher D. Manning, and Jeffrey Heer. 2012. "Interpretation and Trust: Designing Model-driven Visualizations for Text Analysis." Pp. 443-452 in *Proceedings of the ACM Human Factors in Computing Systems (CHI)*. Available at http://vis.stanford.edu/papers/designing-model-driven-vis.

Cott, Nancy Falik. 1987. *The Grounding of Modern Feminism*. New Haven, CT: Yale University Press.

DiMaggio, Paul. 2015. "Adapting Computational Text Analysis to Social Science (and Vice Versa)." *Big Data & Society* 2: 1-5.

DiMaggio, Paul, Manish Nag, and David Blei. 2013. "Exploiting Affinities between Topic Modeling and the Sociological Perspective on Culture: Application to Newspaper Coverage of U.S. Government Arts Funding." *Poetics* 41:570-606.

Dunning, Ted. 1993. "Accurate Methods for the Statistics of Surprise and Coincidence." *Computational Linguistics* 19:61-74.

Evans, John H. 2002. *Playing God? Human Genetic Engineering and the Rationalization of Public Bioethical Debate*. Chicago, IL: University of Chicago Press.

Evans, Sara. 1980. *Personal Politics: The Roots of Women's Liberation in the Civil Rights Movement and the New Left*. New York, NY: Vintage Books.

Ferree, Myra Marx, William Anthony Gamson, Jürgen Gerhards, and Dieter Rucht. 2002. *Shaping Abortion Discourse: Democracy and the Public Sphere in Germany and the United States*. New York, NY: Cambridge University Press.

Franzosi, Roberto. 2004. *From Words to Numbers: Narrative, Data, and Social Science*. Cambridge, UK: Cambridge University Press.

Franzosi, Roberto. 2010. *Quantitative Narrative Analysis*. Thousand Oaks, CA: Sage.

Friedland, Roger, John W. Mohr, Henk Roose, and Paolo Gardinali. 2014. "The Institutional Logics of Love: Measuring Intimate Life." *Theory and Society* 43: 333-70.

Glaser, Barney G. and Anselm L. Strauss. 1999. *The Discovery of Grounded Theory: Strategies for Qualitative Research*. Chicago, IL: Aldine Transaction.

Glaser, Barney G. and Anselm L. Strauss. 2005. *Awareness of Dying*. New Brunswick, NJ: Aldine Transaction.

Grimmer, Justin. 2010. "A Bayesian Hierarchical Topic Model for Political Texts: Measuring Expressed Agendas in Senate Press Releases." *Political Analysis* 18:1-35.

Grimmer, Justin. 2013. "Appropriators Not Position Takers: The Distorting Effects of Electoral Incentives on Congressional Representation." *American Journal of Political Science* 57:624-42.

Grimmer, Justin and Brandon M. Stewart. 2011. "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts." *Political Analysis* 21:267-97.

Griswold, Wendy. 1987. "The Fabrication of Meaning: Literary Interpretation in the United States, Great Britain, and the West Indies." *American Journal of Sociology* 92:1077-117.

Hanna, Alexander. 2013. "Computer-aided Content Analysis of Digitally Enabled Movements." *Mobilization: An International Quarterly* 18:367-88.

Hinton, G., L. Deng, D. Yu, G. E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury. 2012. "Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups." *IEEE Signal Processing Magazine* 29:82-97.

Hirschberg, Julia and Christopher D. Manning. 2015. "Advances in Natural Language Processing." *Science* 349:261-66.

Hull-House. 1907. *Hull-House Year Book: September 1, 1906- September 1,1907*. Chicago: Hull-House.

King, Gary, Jennifer Pan, and Margaret Roberts. 2013. "How Censorship in China Allows Government Criticism but Silences Collective Expression." *American Political Science Review* 107:1-18.

Krippendorff, Klaus H. 2013. *Content Analysis: An Introduction to Its Methodology*. Los Angeles, CA: Sage.

Lancichinetti, Andrea, M. Irmak Sirer, Jane X. Wang, Daniel Acuna, Konrad Körding, and Luís A. Nunes Amaral. 2014. "A High-reproducibility and High-accuracy Method for Automated Topic Classification." *arXiv:1402.0422 [Physics, Stat]*, February. Accessed September 21, 2017 (http://arxiv.org/abs/1402.0422).

Lee, Monica and John Levi Martin. 2015. "Coding, Counting and Cultural Cartography." *American Journal of Cultural Sociology* 3:1-33.

Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. New York, NY: Cambridge University Press.

Martin, John Levi. 2000. "What Do Animals Do All Day? The Division of Labor, Class Bodies, and Totemic Thinking in the Popular Imagination." *Poetics* 27:195-231.

Mische, Ann and Philippa Pattison. 2000. "Composing a Civic Arena: Publics, Projects, and Social Settings." *Poetics* 27:163-94.

Mohr, John W. 1998. "Measuring Meaning Structures." *Annual Review of Sociology* 24:345-70.

Mohr, John W. and Petko Bogdanov. 2013. "Introduction—Topic Models: What They Are and Why They Matter." *Poetics* 41 (6): 545-69.

Mohr, John W. and Vincent Duquenne. 1997. "The Duality of Culture and Practice: Poverty Relief in New York City, 1888–1917." *Theory and Society* 26: 305-56.

Mohr, John W., Robin Wagner-Pacifici, Ronald L. Breiger, and Petko Bogdanov. 2013. "Graphing the Grammar of Motives in National Security Strategies: Cultural Interpretation, Automated Text Analysis and the Drama of Global Politics." *Poetics* 41:670-700.

Monroe, Burt L., Michael P. Colaresi, and Kevin M. Quinn. 2008. "Fightin' Words: Lexical Feature Selection and Evaluation for Identifying the Content of Political Conflict." *Political Analysis* 16:372-403.

Moretti, Franco. 2013. *Distant Reading*. London, England: Verso.

Nardulli, Peter F., Scott L. Althaus, and Matthew Hayes. 2015. "A Progressive Supervised-learning Approach to Generating Rich Civil Strife Data." *Sociological Methodology* 45:148-83.

Nelson, Laura K. 2015. "Political Logics as Cultural Memory: Cognitive Structures, Local Continuities, and Women's Organizations in Chicago and New York City. *Working Paper. Kellogg School of Management, Northwestern University*.

Neuendorf, Kimberly A. 2001. *The Content Analysis Guidebook*. Thousand Oaks, CA: Sage Publications.

Pachucki, Mark A. and Ronald L. Breiger. 2010. "Cultural Holes: Beyond Relationality in Social Networks and Culture." *Annual Review of Sociology* 36:205-24.

Princeton University. "About WordNet." WordNet. Princeton University. 2010. <http://wordnet.princeton.edu>

Reed, Isaac Ariail. 2015. "Counting, Interpreting and Their Potential Interrelation in the Human Sciences." *American Journal of Cultural Sociology* 3:353-64.

Rosen, Ruth. 2000. *The World Split Open: How the Modern Women's Movement Changed America*. New York, NY: Penguin Books.

Saldana, Johnny. 2015. *The Coding Manual for Qualitative Researchers*. 3rd ed. Los Angeles, CA: Sage.

Salton, Gerard and Christopher Buckley. 1988. "Term-weighting Approaches in Automatic Text Retrieval." *Information Processing & Management* 24:513-23.

Schwartz, H. Andrew and Lyle H. Ungar. 2015. "Data-driven Content Analysis of Social Media a Systematic Overview of Automated Methods." *The ANNALS of the American Academy of Political and Social Science* 659:78-94.

Snefjella, Bryor and Victor Kuperman. 2015. " Concreteness and Psychological Distance in Natural Language Use." *Psychological Science* 26:1449-60.

Spillman, Lyn. 2015. "Ghosts of Straw Men: A Reply to Lee and Martin." *American Journal of Cultural Sociology* 3:365-79.

Tausczik, Yla R. and James W. Pennebaker. 2010. "The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods." *Journal of Language and Social Psychology* 29:24-54.

Tilly, Charles. 1997. "Parliamentarization of Popular Contention in Great Britain, 1758-1834." *Theory and Society* 26:245-73.

Yu, Liang-Chih and Chun-Yuan Ho. 2014. "Identifying Emotion Labels from Psychiatric Social Texts Using Independent Component Analysis." Pp. 837-847 in *Proceedings of COLING 2014, Stroudsburg, PA*.

Author Biography

Laura K. Nelson is an assistant professor of sociology at Northeastern University. She uses computational methods and open-source tools to study social movements, culture, and gender.