Machine Translation: Mining Text for Social Theory

James A. Evans and Pedro Aceves

Department of Sociology, University of Chicago, Chicago, Illinois 60637;
email: jevans@uchicago.edu

Annu. Rev. Sociol. 2016. 42:21–50

First published online as a Review in Advance on June 1, 2016

The *Annual Review of Sociology* is online at soc.annualreviews.org

This article's doi:
10.1146/annurev-soc-081715-074206

Keywords

content analysis, big data, natural language processing, machine learning, text analysis, computational methods, grounded theory

Abstract

More of the social world lives within electronic text than ever before, from collective activity on the web, social media, and instant messaging to online transactions, government intelligence, and digitized libraries. This supply of text has elicited demand for natural language processing and machine learning tools to filter, search, and translate text into valuable data. We survey some of the most exciting computational approaches to text analysis, highlighting both supervised methods that extend old theories to new data and unsupervised techniques that discover hidden regularities worth theorizing. We then review recent research that uses these tools to develop social insight by exploring (*a*) collective attention and reasoning through the content of communication; (*b*) social relationships through the process of communication; and (*c*) social states, roles, and moves identified through heterogeneous signals within communication. We highlight social questions for which these advances could offer powerful new insight.

INTRODUCTION

A vast expanse of information about what people do, know, think, and feel lies embedded in text. Textual traces range from the world's life on the web, social media, instant messaging, and online commerce to automatically transcribed YouTube videos, medical records, digitized libraries, and government intelligence. The rise of literacy, and more recently computers, scanners, the Internet, and cell phones, has conditioned an exploding supply and demand for textual information. This provides sociologists access to a greater variety of texts that reach deeper into the contemporary social world than ever before. Simultaneously, massive semiautomated archival projects (e.g., Google Books) are making vast caches of historical text digitally available for analysis.

This unfolding universe of digital text has generated a call for new information representations, natural language processing (NLP), and information retrieval (IR) and extraction (IE) tools that can filter, search, summarize, classify, and extract information from text. Moreover, data sets representing not only the increased prevalence of text but also audio, visual, and heterogeneous sensor data (e.g., click streams on the web, "likes" on Facebook, movements via cell phone) have supported the rapid growth of a new engineering paradigm, machine learning (ML). An offspring of statistics and artificial intelligence, ML devotes itself to learning from data and to predicting and extending human perceptive accuracy and understanding. Many general and text-specific ML techniques have now proven powerful for translating text and related communicative traces into sociologically valuable data (Grimmer & Stewart 2013).

In this article, we briefly review the history of content and text analysis in sociology and the social sciences. Text is sometimes layers removed from the "social games" that sociologists seek to illuminate.[1] Computational approaches are sometimes less subtle and deep than the reading of a skillful analyst, who interprets text in context. Nevertheless, we show that recent advances in NLP and ML are being used to enhance qualitative analysis in two ways. First, supervised ML prediction tools can "learn" and reliably extend many sociologically interesting textual classifications to massive text samples far beyond human capacity to read, curate, and code. Second, unsupervised ML approaches can "discover" unnoticed, surprising regularities in these massive samples of text that may merit sociological consideration and theorization.

Next, we review some of the most exciting computational approaches to the large-scale analysis of text for "translation" into sociologically relevant data. These include techniques from NLP, IR, IE, and ML that exploit language structure and context to extract meaning. Recent developments in ML, like the rise of "deep learning" or multilayer neural networks, can change the technical machinery underlying these tools and improve their accuracy (Manning 2015). As a result, we focus on the persistent language tools and tasks (e.g., disambiguation, parsing) whose designs have proven valuable for the production of sociologically relevant data, despite changes in implementation.

The bulk of our article reviews recent research that uses these approaches to develop social insight. This research groups itself into work that explores (*a*) collective attention and reasoning through examining the content of communication; (*b*) social relationships through analyzing the process of communication; and (*c*) social states, roles, and moves identified through heterogeneous

[1] We do not use the term social games to imply that human nature is primarily playful (Huizinga 1971) or that human action reflects exclusively rational, ends-oriented competition as in game theory (Myerson 2013, von Neumann & Morgenstern 1944). Rather, like Bourdieu's field theory (2013) in which social agents play high stakes games of status on an established field, we intimate that social games comprise a board, pieces, conventional rules, established moves, and widely but not universally shared objectives that include both playing and winning. Wittgenstein's "language game" (2010) incorporates several of these aspects, but in a game restricted to the conveyance meaning through partially shared representations. Social life, then, is composed of overlapping games (Long 1958), each with their own rules in which players bring interests, dispositions, and strategies to their moves. Information about these games can be gleaned through the textual evidence they leave behind.

signals within communication. This work demonstrates large-scale analyses of not only human communication but also the social and cultural worlds that produced it and that become visible through it. Much of this work has not been produced by sociologists but rather by computer and information scientists committed to building new NLP and ML tools and to demonstrating them on the expanding universe of text. Together, these have generated a wealth of sociologically relevant data about collective attention, intergroup relations, cultural associations, micro states, and behaviors, present and past. As such, we highlight classic social questions for which these advances could offer new insight. We also point to interpretive opportunities these new data hold for the development of "grounded theory" (Glaser & Strauss 1967)—using machines to mine text for social insight.[2]

CONTENT ANALYSIS AND COMPUTATION

What are the limits of text analysis? How and how well do text and other communicated content trace the social world that produced it? Here we take the "social world" to consist of individuals engaged in interaction, situated on a landscape characterized by (*a*) social structures like class boundaries, kin networks, formal organizations, and ephemeral friendships; (*b*) cultural systems of shared symbols, including the human communication protocols of language, gesture, and fashion; and (*c*) material or apparent external resources and constraints, such as capital flows and the built environment. Life in this social world constitutes the social game in which individuals are engaged. For example, lawyers and judges play the legal game, students and teachers play the education game, politicians play the diplomacy game, and generals play the war game. Actors also participate in a myriad of less specialized games, including the parenting game, the courtship game, and the job market game (Long 1958). As such, social games comprise characteristics of the social players, the environment, and their communicative engagement.

Individuals possess interests and drives that partially derive from the positions they assume within the social world and that condition an unfolding stream of social action and interaction. These actions and interactions produce other- and self-communicated content, which is sometimes in the form of text but also in audio, video, and even image recordings that can be translated with more or less accuracy into text. Consider transcripts of the tapes tracing deliberations among President Kennedy and his advisors surrounding the botched Bay of Pigs invasion and the Cuban Missile Crisis (Gibson 2012) or the Nixon White House tapes that detailed strategic discussion of the Watergate break-in and cover-up. It is not surprising that the games of presidents have long been recorded, but the ubiquity of online communication, automated speech-to-text translation, and mobile sensors have made these traces available for a much wider range of social games and players than ever before.

This increase in available text has the potential to increase its relevance for many areas of sociological scholarship. Texts and communicative traces reveal more about some social games than others. A personal journal entry may uncover the state of the writer, whereas instant messaging banter reveals the intensity with which conversationalists initiate and maintain contact. Although enabling certain views into the underlying social game, genres of text restrict others, such as the modern research article, which obscures scientists' personal sensations and experiences by fixing on the referential world of experiment, observation, and shared significance (Jakobson 1960, Rodriguez-Esteban & Rzhetsky 2008, Shapin 1994).

[2]Computationally induced data structures can surprise, challenging presumptions or pre-existing theory, and lead the social analyst to abductively generate new theory by imagining what would be socially required for those patterns to exist (Timmermans & Tavory 2012).

In this way, inferences about the social, cultural, and material landscape where social games are played depend on the coupling between text and the underlying game. If details traced in text constitute substantial moves in the game, like "flirts" in an online dating website, then text may constitute a representative sample or even the complete population of relevant social moves. The more tightly coupled a text is to social moves in the game of interest, the stronger the inferences that can be made. When more loosely coupled—consider genres of memoir and hagiography as source data for historical events—stronger assumptions are required and less relevant data are available. The increase in textual data of all types allows us to reliably analyze games of greater structural and temporal complexity. These could include temporal shifts in discriminatory attitudes from news and novels, or the subtle interplay of social roles involved in recorded surgical teams or high school cliques viewed through social media. Inferring differences, change over time, or variation along some other dimension (e.g., gender, status, race) within social games or worlds requires more text than inferring stable, universal patterns. The deluge of historical and contemporary text digitally available today opens the possibility of inferring even more elaborate structures and patterns within social games, such as cycles, spatial arrays, complex hierarchies, and transitive orders (Kemp & Tenenbaum 2008).

Systematic text analysis entered sociology during the second meeting of the German Sociological Society in 1910 when Max Weber proposed a large-scale analysis of the German press to identify the influence of the news "in making modern man" and to trace temporal shifts in values (Hardt 2001, p. 136). This research program, interrupted by World War I, resurfaced in the quantitative analyses of mass media, including newspapers (Willey 1926, Woodward 1934), television, and radio (Berelson & Lazarsfeld 1948). Micro sociologists also began to analyze the content of group interactions (Bales 1950) and the structure of conversations (Sacks 1995).

Sociologists engaging in content or conversation analysis often begin by qualitatively coding text or other media according to theoretically meaningful categories. They then interpretively or quantitatively analyze coded features, often in concert with raw textual elements (e.g., frequent or distinguishing words) and metadata (e.g., authorship, audience) to identify semantic or stylistic patterns. Consider Roscigno & Hodson's (2004) analysis of worker resistance through the coding of shop floor ethnographies and Stivers et al.'s (2009) analysis of transcribed interaction data across cultures to identify universal patterns in turn taking.

From the 1960s, computers have been used to assist sociologists in what a Rand Corporation paper titled "Automatic Content Analysis" (Hays 1960). Philip Stone and Robert Bales' General Inquirer System mapped text to content dictionaries that tallied disambiguated words associated with power, sentiment, and other categories tracing concepts relevant to theories in sociology, political science, and psychology (Stone et al. 1966). This has similarities to more recent systems created from subject-ranked terms, such as the Linguistic Inquiry and Word Count (LIWC) (see Pennebaker et al. 2001, Tausczik & Pennebaker 2010). Work by Carley (1994) analyzed symbols in networks of associations to identify cultural patterns or what Sedelow (1989) has called "society's collective associative memory" (p. 4).

In the last decade, however, statistical NLP has become dramatically more accurate and powerful at recovering linguistic structures and semantic associations recognized by both linguists and ordinary language speakers. Moreover, general ML models and algorithms have become much more accurate in their ability to predict a range of outcomes, including expert annotations and underlying qualities of context through unstructured and semistructured text data. Computational approaches can now make substantially improved inferences with ML methods acting as extensions of our cognitive capacity—as cognitive prosthetics.

Computation can augment our fine perception of patterns in language and their links to the social world beneath. In past sociological work, a researcher might code passages of text relating to

some underlying concept (e.g., feminism, democracy) or process (persuasion, consensus) but often without recognizing or articulating the details of language associated with those codes. They might not be able to construct a protocol that would enable a naive researcher or computer to reproduce them independently. ML approaches, coupled with a sufficiently rich set of textual features, can be trained on human codes to extend them with improved fidelity. This can allow researchers to automatically analyze many more documents than would be possible through traditional reading. For example, manually coding topics from 40 million scientific abstracts could take a thousand researcher-years, but automatic coding by a trained model might require only a few computer-days.

Moreover, ML techniques can be used to detect and predict qualities of the author, audience, and social world from textual details imperceptible to human researchers. Machine memory augments human analytical limits by holding a massive array of language features simultaneously in mind so we can associate them in reliable constant comparison (Glaser 1965). Finally, ML tools can discover novel patterns in text data on the basis of similarity, structural association, or predictive power. These patterns may surprise the social analyst and merit interpretive scrutiny and labeling, provoking new social theory to explain them (Tavory & Timmermans 2014). Because such patterns are based on substantial underlying data, if they prove meaningful, they are likely to be important. In summary, although computation cannot mimic the prior experience, vision, and unexpected associations of a gifted analyst, it can augment their reliability and provide new data—regularities, associations, and structures built from much larger text samples—which sociology can mine to deepen and expand our inferences about the social games and worlds underlying communication.

Data mining has acquired a bad reputation in the social sciences. Many see it as synonymous with the practice of algorithmically sifting through data for associations and then falsely reporting them as if confirmations of theoretically inspired, single-test hypotheses. Unreported and statistically unaccountable data mining leads to the overfitting of statistical models to data and fragile findings that neither replicate nor generalize. This, in turn, undermines confidence in published social science research (Freese 2007, Ioannidis & Doucouliagos 2013, Simmons et al. 2011a), just as it has in other fields like genetics, biomedicine (Ioannidis 2005), and even ML itself (Pentland 2012a). Unreported data mining was especially problematic in a social science era more heavily reliant on sparse, expensive data, such as in-person surveys and experiments. During such a time, not only published inferences but also potential data reuses were compromised by mining data's structure before social theories could be blind tested against it. With the greater volume and variety of text data available today, produced both passively through the natural flow of digital communication and actively through controlled online experiments, these concerns should be ameliorated: Statistically accountable data mining can be used to legitimately discover hypotheses on some data and then confirm those hypotheses on other data. As such, with the contemporary explosion of text and the socially relevant data being mined from it, we expect to see a renaissance of discovery about human communication and the myriad social structures and processes reflected in it.

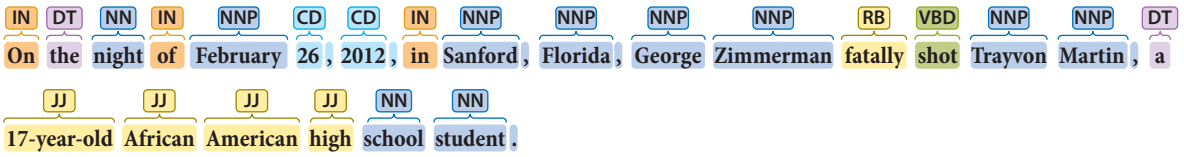
DATA IN TEXT

Text analysis attends to a range of language features, each of which suggests modes of analysis with techniques from NLP, IR, and IE. We cannot provide more than a dense, cursory treatment of these in this review; interested readers could consult the following book-length references for more complete (and relaxed) explication (Clark et al. 2010, Jurafsky & Martin 2000, Manning & Schütze 1999, Manning et al. 2008). In **Figure 1**, we quote a news source about the killing of Trayvon Martin alongside some prominent language features widely considered in text analysis, each extracted automatically from the Stanford CoreNLP (Manning et al. 2014).

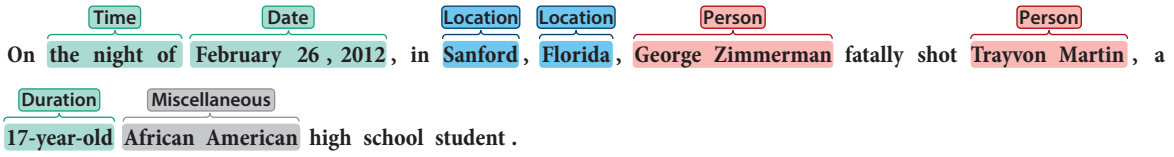
The most commonly used language features derive from the lexicon—words in the vocabulary under consideration. Many studies use some function of word frequency to make inferences about meaning and focus. For example, analysts have used topically curated word lists to identify states like ideology and emotion (Stone et al. 1966, Tausczik & Pennebaker 2010, Whissell 1989). Alternatively, the field of stylistics relies on the usage pattern of function words that carry no independent semantic information, such as articles and prepositions, to predict authorship through distinctive statistical signatures (Mosteller & Wallace 1964) or to trace power dynamics by tracking

"On the night of February 26, 2012, in Sanford, Florida, George Zimmerman fatally shot Trayvon Martin, a 17-year-old African American high school student."

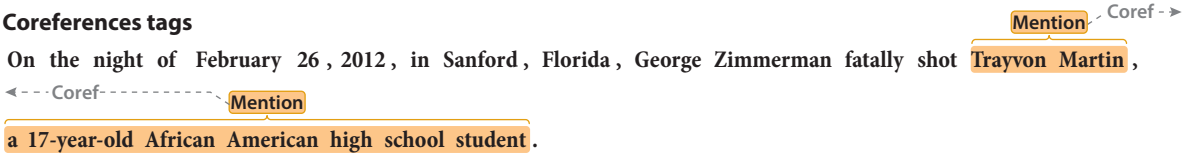
Part-of-speech tags



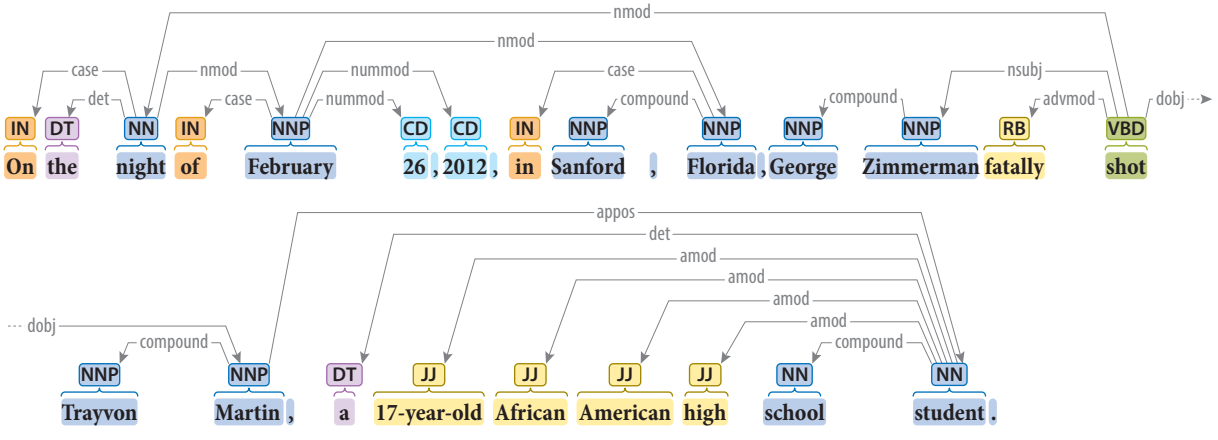
Named entities tags



Coreferences tags



Dependencies tags



mimicry (Danescu-Niculescu-Mizil et al. 2012). Words can also be associated by the role they play within a sentence through part-of-speech tagging.

When the lexicon is used without imposing any higher-order structure, the document is formally modeled as an undifferentiated "bag of words." For analysis, word instances or tokens are often stemmed for related roots (e.g., pray, prayer, and prayed all collapse to pray). Richer tokenizations also become possible by including frequent n-grams, which are common word sequences of length n (e.g., bigrams, such as prayer meeting and gangsta rap, or trigrams, such as cup of tea and Central African Republic) or skip-grams, which are n-grams with gaps of length k (e.g., 1-skip-bigrams, such as cup tea and Central Republic). Each document can then be represented as a sparse vector of counts for each token in the vocabulary, with such counts often normalized by the number of tokens within document or weighted to highlight the degree to which they distinguish the document. For example, the most common weighting scheme for identifying distinguishing words, tf-idf, uses some function of term frequency within a document, divided by the logarithmically scaled inverse fraction of documents containing the word—the total number of documents in the collection divided by those containing at least one mention. Although many text analyses, like sentiment classification, use only functions of the distributions of document words, the best performing systems include richer understandings of language structure (Hirschberg & Manning 2015).

Words refer to semantic entities, which may be referred to by many words (e.g., synonyms, pronouns). Coreference is the process by which words are linked to this underlying entity, like the resolution of anaphora introduced by an author to vary his or her writing. The resolution of a coreference can improve document vectors. Moreover, in IE, some semantic entities are considered "named entities," predefined categories, including names of persons, organizations, and locations. Named entity recognition and extraction are tasks involving the identification of these entities and extraction of details associated with their specific instances into a database (see **Figure 1**).

Syntax is the structure of words within sentences. There are many formal grammar-based approaches for uncovering sentence structure. The most common two involve parsing according to (*a*) phrase structure (or constituency) grammars and (*b*) dependency grammars. The first seeks to decompose sentences into contiguous phrases. The second identifies a network of dependencies between words across the sentence. Phrase or dependency representations may be more or less appropriate depending on the degree to which word order is critical for understanding the content of interest. Parsing sentences according to both approaches has become increasingly accurate in recent years. For example, in English text, dependency parsing has come to exceed 95% accuracy by Google researchers (Alberti et al. 2015). Nevertheless, because of the computational complexity—and often inaccuracy—of parsing complex sentences, local structure is often captured blindly through the use of n-grams and skip-grams. Alternatively, models can operate over the sequence of part-of-speech tags attached to the lexicon to robustly "chunk" a sentence into noun, verb,

Figure 1

Linguistic features for text analysis of the first sentence from the Wikipedia entry on "Shooting of Trayvon Martin," retrieved October 5, 2015, and the automated output (with punctuation tagging removed) extracted by Stanford CoreNLP (Manning et al. 2014). See **http://stanfordnlp.github.io/CoreNLP/**, or, to generate a user-submitted example, **http://nlp.stanford.edu:8080/corenlp/**. Part-of-speech tags are from the Penn Treebank tag set (Santorini 1990): CD, cardinal date; DT, determiner; IN, preposition or subordinating conjunction; JJ, adjective; NN, noun (singular or mass); NNP, proper noun (singular); RB, adverb; VBD, verb, past tense. Directed linguistic dependencies are from the Stanford Dependencies representation (de Marneffe et al. 2014): advmod, adverbial modifier; amod, adjectival modifier; appos, appositional modifier; case, case-marking, preposition, or possessive; compound, noun compound modifier; det, determiner; dobj, direct object; nmod, noun modifier; nsubj, nominal subject; nummod, numeric modifier.

and prepositional phrases. Social and computational analysts often use word copresence within a grammatical phrase, closeness in a dependency network, or proximity within a sequence of words to infer large-scale, semantically meaningful associations between words. Social analysts may sometimes use linguistic structure more directly to extract unique data within textual claims. Consider Franzosi's (2004) network analysis of subject-verb-object triples, like cops (subject) beat (verb) protestors (object), which can increasingly be semi- or fully automatically extracted as named entities related through parsed dependency relationships.

Not pictured in **Figure 1**, higher-order document structure, or discourse, has also been productively used to analyze text. Work that examines word colocation within paragraphs (Lee & Martin 2015), author-created section headings (e.g., Materials and Methods), or induced partitions of documents traced by lexical shifts (Hearst 1997) can provide valuable information about higher-order associations.

Another linguistic character that does not map unambiguously onto text is phonology, the system of speech sounds. Distinguishing dialects through phonology can reveal distinct social worlds underlying spoken interaction (Rickford et al. 2015). For example, phonological cues from interviews performed with National Longitudinal Study of Youth respondents reveal how ebonics, urban culturally marked slang, accounts for wage gaps between respondents better than race (Grogger 2011).

Improvements in all areas of NLP draw on a changing substrate of tools from probabilistic modeling, information theory, matrix factorization, and multilayer neural networks known as deep learning. Despite dramatic improvements in recent years, limitations remain. The most major is also a problem with sociological research: Most NLP resources—like most sociological studies—are only available for high-resource languages, such as English, Spanish, and Chinese, and not for low-resource languages, like Bengali, Indonesian, and Swahili, spoken by hundreds of millions of people (Hirschberg & Manning 2015). Another limitation relates to the lack of sophisticated models for higher-level linguistic discourse, such as how sentences relate to one another (Stymne et al. 2013) and aggregate into paragraphs and more or less effective arguments, although this is one of the targets of sociological text analyses (e.g., message complexity and popularity in Bail 2016).

Although sociological analyses of text sometimes benefit directly from the rich features of language induced with NLP, more often they take these features as inputs to ML models, which are themselves used to model issues of fundamental sociological significance, including collective attention, social relationships, and socially relevant states.

Theoretical Purpose, Research Design, and Machine Learning Approaches

A sociological research project employing text analysis begins with either an impulse to evaluate a preexisting theory or alternatively to explore and discover theory related to the domain from which text was sampled. **Figure 2** sketches how these distinct purposes can influence research design and shape the choice of ML methods used to create socially relevant data. Social theory breaks down into (*a*) concepts that trace social entities—natural or imagined phenomena—and (*b*) relationships that link and structure them.[3] If concepts have already been identified in text, then the researcher often uses a supervised ML approach to extend these identifications to data beyond the analyst's capacity to reliably code. If concepts are yet to be discovered, then an unsupervised method is likely to be drawn upon for assistance.

[3]If the relationship is rendered a logical predicate and the concepts its arguments, then together they formally comprise a theoretical claim. Consider the Marxist theoretical cartoon: If capitalism ≡ *c*, and destroys ≡ *D*, then *cDc*.

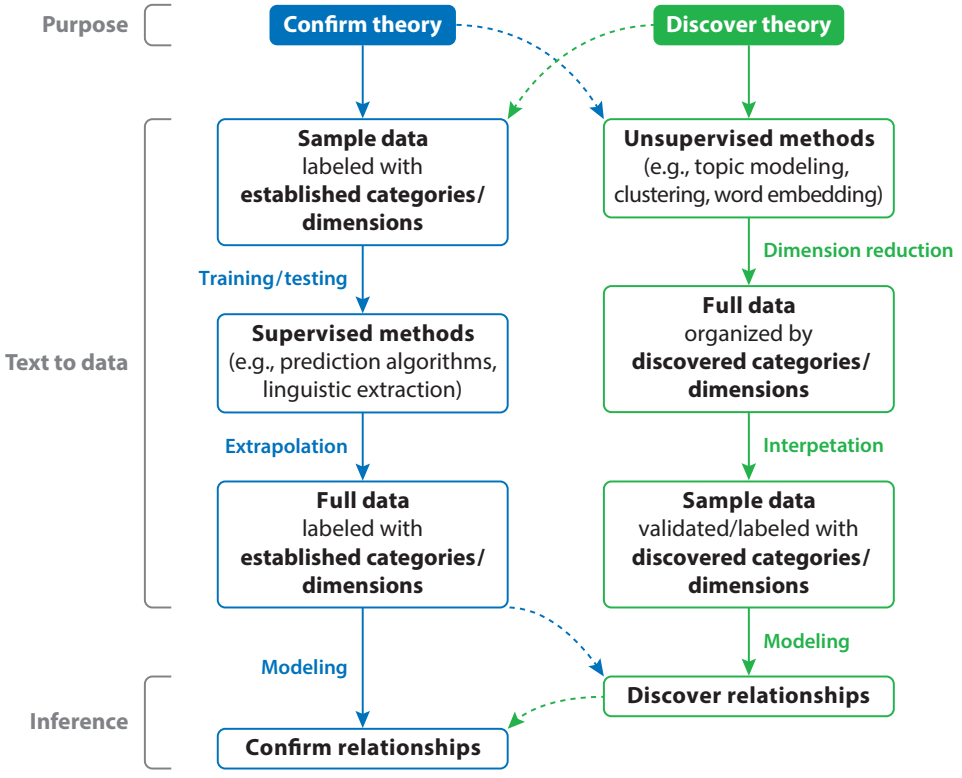


Figure 2

Text analysis and social theory. The arrows trace the text analysis research pipeline, highlighting how different motivations—for confirming versus discovering theory—influence the choice of ML methods used to construct data from text relevant for sociological inference. Solid lines represent straightforward research pipelines, and dashed lines suggest research pathways that mix research motivations for confirmation and discovery. For example, a researcher might explore novel arrangements of established categories (e.g., sentences tagged with positive sentiment). Moreover, once new patterns are discovered in one corpus of text, they may be tested in another.

Supervised and unsupervised approaches condition distinct research pipelines through which text is translated into socially relevant data. This process of transforming unstructured data, like text, into structured data that is in turn leveraged to create new forms of value has sometimes been termed datafication (O'Neil & Schutt 2013). With supervised methods, an analyst begins with a sample of text instances where concepts have been identified and coded by themselves or others. The concepts may be inherited from prior theorists, deduced from prior arguments, or discovered by an interpretive analyst in the process of coding. This sample is then divided into training and testing subsamples, and a supervised ML method draws on features associated with instances in the training sample to estimate a statistical model or tune an algorithm. The trained model or algorithm is then used to predict identified but unlabeled instances in the testing sample to evaluate its success. This division of data for training and testing can often be dynamic, as in an *n*-fold cross-validation design, where the data is split into *n* parts, and the model is successively trained on *n*-1 parts and tested on the *n*th, cycling through each split. Success is typically measured with an IR metric that captures some balance of false positives and negatives (mistaken classifications and missed classifications), such as precision, recall, or area under the receiver operator characteristic

(ROC) curve (AUC). If the accuracy of the program is not sufficient, more trained instances are identified by human coders, and the training and testing process is repeated to satisfaction. Finally, the successful model or algorithm is used to extrapolate codes to unlabeled data. With data on established codes in hand, the analyst moves on to analyze hypothesized relationships between measured constructs with appropriate statistical models.

When unsupervised methods are used to discover novel categories or dimensions from text, the allocation of human effort is typically reversed. An automated ML model or algorithm is unleashed on the complete corpus of interest, furnishing new, discovered variables for subsequent analysis. Unsupervised models need not discover new variables. They may instead be used to discover known categories, and in this case, they are built on training data and evaluated on testing data as with supervised models. Such is the case with unsupervised syntactic parsing models where expert parses are known. For a sociological example, Nelson (2015) uses unsupervised topic models to identify established differences between the discourse of feminist organizations in Chicago and New York over time. Unsupervised models are given clues about the patterns or rules they should be learning (Jurafsky & Martin 2000) and rest on assumptions about the underlying structural properties of the data, whether algebraic, combinatorial, or probabilistic (Jordan & Mitchell 2015). The algorithms learn distinguishing characteristics, such as distributional patterns or clustering properties (Clark et al. 2010). Occasionally the resulting unsupervised data structures are automatically labeled and trusted, but more often analysts formally or informally sample, peruse, and critically interpret them. For example, topics produced by a probabilistic topic model estimated on a corpus are scrutinized, then explicated, and hand labeled for easy description and reference. As with supervised models, the data that results are often subsequently used to discover or confirm theoretically significant relationships with an appropriate statistical model.

In sociological analyses, text-based variables derived from supervised or unsupervised methods usually take an explanatory role as independent variables that predict an established dependent variable from outside text. For example, Goldberg et al. (2015b) extracted the degree of an employee's cultural embeddedness within a firm from text and then used it to predict individual performance ratings and tenure. In the context of this usage, we now define some of the most recent and promising supervised and unsupervised approaches for sociological text analysis.

Supervised methods begin with a training sample of text, tagged with expert-defined codes to identify categories of interest, like positive sentiment, liberal ideology, or mention of a particular social movement strategy. This selectively tagged text furnishes positive and negative examples for a supervised model to distinguish. Supervised models include linear and logistic regression with thousands or tens of thousands of independent variables corresponding to text features, such as the word frequency vectors described above (Joshi et al. 2010). Given the high dimensionality of text data, it is not always possible to efficiently estimate these models without simplifying text variables and reducing their dimensionality. Most directly, social science applications have analyzed contingency tables to identify distinguishing n-grams from positive and negative examples (Gentzkow & Shapiro 2010, Laver et al. 2003), which are subsequently included as predictors in regression models to identify sentiment, policy positions, or ideological slant. This approach hearkens to classic approaches that relied simply on weighted counts for a predefined list of terms (Loughran & McDonald 2011, Tetlock 2007). Principal components regression (Massy 1965) and supervised latent Dirichlet allocation (LDA) topic models (McAuliffe & Blei 2008) have also been used to reduce the text dimensionality by deriving components or topics subsequently used as predictors in a regression. An integrated approach with stronger performance is multinomial inverse regression, a two-stage estimation approach in which linguistic features are first regressed on some function of the category of interest (e.g., positive sentiment) and then selectively included in a forward regression without losing responsiveness to the predicted category (Taddy 2013).

This is similar to sparse regression approaches, such as the least absolute shrinkage and selection operator (lasso), which minimizes the usual sum of squared errors with a bound on the sum of coefficient absolute values to select a sparse subset of high-signal predictors (Tibshirani 1996).

Alternative approaches use a full or partial complement of word frequencies and related linguistic features in a range of other ML classification algorithms, including *k*-nearest neighbor analysis, naive Bayes estimation, support vector machines, maximum entropy classifiers, deep learning, decision trees, and ensemble techniques that combine the judgment of multiple approaches. Some of these approaches maximize interpretability (e.g., pruned decision trees), accuracy (e.g., deep learning and ensemble techniques), or speed (e.g., support vector machines) (for examples, see Pang & Lee 2008, Srivastava & Sahami 2009, Yu et al. 2008). All of these approaches can be generalized to content beyond text, including audio, images, and video with deep learning approaches recently becoming dominant (Hinton et al. 2006, Hirschberg & Manning 2015).

Unsupervised methods begin with a corpus of unannotated text and then discover and represent novel structures for interpretation. We highlight four of the most common: clustering, network analysis, topic modeling, and vector space embedding. Each is illustrated in **Figure 3** along with the attributes of the data representations they produce. Clustering is typically used to discover coarse-grained, categorical groupings of documents through their words, whereas network analysis has typically been used to identify fine-grained topological positions of words and their underlying entities across documents. Topic modeling has been used to coarsely describe documents as sparse combinations of latent topics, and vector space embedding models spread words and documents across the high-dimensional spaces from which semantic distances can be calculated.

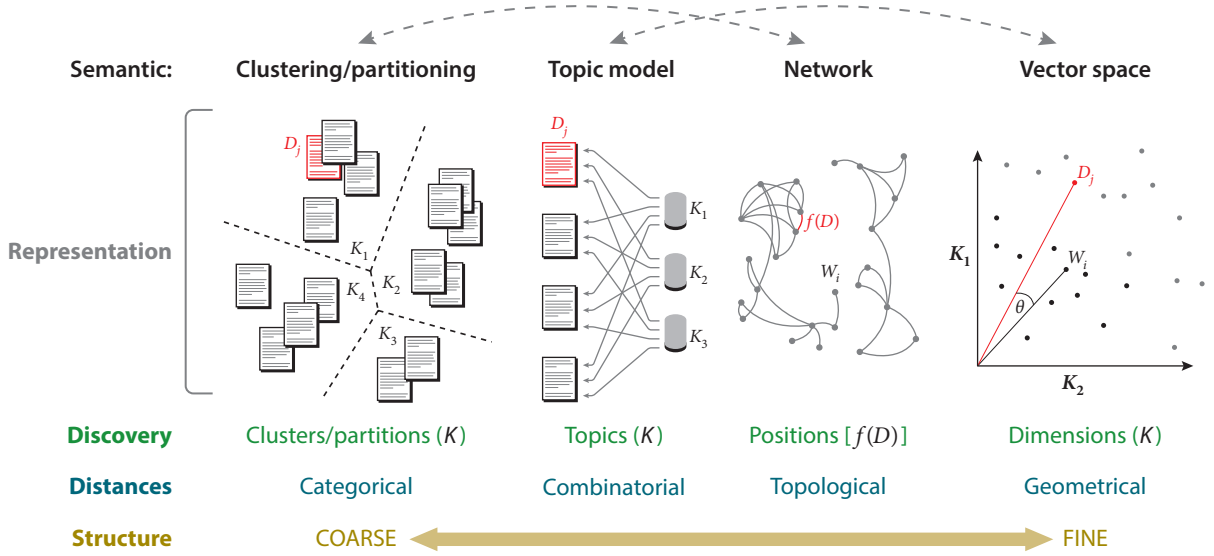


Figure 3

Properties of four common semantic representations. Four of the most widely used semantic representations of text, linked to the most frequently deployed unsupervised machine learning approaches, are clustering, topic modeling, semantic network induction, and vector space word embedding. In these representations, *D* represents documents, *W* represents words within those documents, and *K* represents discovered semantic structures. Document clusters, in which similar texts are typically grouped by shared words, closely relate to semantic networks, in which similar words often link as a function of the documents in which they co-occur, *f*(*D*). Topic models, which represent documents as sparse mixtures of induced topics, are closely related to vector space word embeddings, which define documents, words, and phrases as dense mixtures of induced dimensions, or vectors plotted in a space anchored by those dimensions.

In clustering documents, similar texts are hierarchically grouped, dissimilar texts are hierarchically divided, or some function of intracluster similarity and intercluster difference is maximized. Such algorithms are often performed on document vectors, including weighted words. Document cluster assignment can be exclusive and hard, as in all hierarchical models, or soft, allowing for degrees of membership. Different clustering rules produce different clusters, which in turn reveal different social games from the data (Grimmer & King 2011). For example, Grimmer and King cluster US Congressional press releases to reveal a common but previously unnoticed genre of partisan taunting in which one political officeholder berates another to highlight their own positions or justify political action.

A semantic network approach links words or phrases colocated within documents, sentences, clauses, and dependency parse trees. This approach can be considered the fine-grained corollary of document clustering. This unsupervised and largely model-free approach can reveal the fine structure of cognitive and cultural associations between entities through calculation of their network positions, such as word centrality, influence, structural equivalence, and constraint (Carley 1993, Carley & Kaufer 1993, Schank & Colby 1973, van Atteveldt 2008, van Atteveldt et al. 2008). These networks can also be partitioned, which is a graph-based approach to clustering. For example, Carley (1994) used semantic networks to trace shifts in culture, such as the evolution of robots from alien monsters to sympathetic companions in fiction, and Corman et al. (2002) has used aggregated betweenness centrality across noun phrases to highlight terms that channel meaning.

Topic modeling is a recent and increasingly common approach to discover semantically cohesive topics and their combination across document collections. Topic models are an influential class of generative, Bayesian probabilistic models that represent documents as drawn from a set of induced topics. Formally, each topic is a latent multinomial variable, tracing a distribution over all words in the corpus vocabulary. The first topic model was titled LDA because a Dirichlet distribution was used to draw per document topic distributions in the first stage of the model (Blei 2012, Blei et al. 2003). This distribution is typically tuned to minimize the mixture of topics describing any particular document in the collection. Estimated topics may become objects of interpretation for the sociologist, who can use them to describe document collections and to trace collective attention and reasoning for the organization, community, or culture that authored them. A recent special issue of *Poetics* was entirely devoted to topic modeling in social and cultural analysis (McFarland et al. 2013b, Mohr & Bogdanov 2013). Many topic model variants have been proposed, including those that account for local word order and syntactic dependencies (Griffiths et al. 2005, Wallach 2006), explicit document dependencies like citations or hyperlinks (Chang & Blei 2010), the temporal order of documents, and evolving topics (Blei & Lafferty 2006).

Word-embedding models construct an efficient set of dimensions or vector space from a document collection in which all documents and words can be projected. This approach is conceptually related to topic modeling; however, because stable semantic distances and not descriptions are the goal, documents are not sparsely embedded over dimensions, and dimensions are not sparsely embedded over words, as they typically are in topic models. Older approaches to word embedding include latent semantic analysis (LSA), which used singular value decomposition on the document-word matrix to identify informative dimensions (the singular values) in which words and documents could be projected. Distances are often calculated between the angles of these vectors to assess semantic distance irrespective of document size. For example, the cosine distance between a single word (e.g., America), a one-line search engine query (e.g., United States America policies laws), and an entire website (usa.gov) could be very small, suggesting a close semantic association.

More recent approaches use neural networks on a large set of linguistic features, including not only words and n-grams but also skip-grams to encode a wide range of syntactic and semantic information. The most prominent of these is word2vec, developed by a team of Google engineers.

Word2vec produces word vectors that perform well on human analogical reasoning tests requiring substantial human semantic understanding. Consider the analogy question: Madrid is to Spain as France is to ___? (Paris). In the high-dimensional geometry of a word2vec vector space model, trained on a one-billion-word corpus of Google News text, the vector for Madrid minus the vector for Spain plus the vector for France is closest to the vector for Paris. Similarly, the vector for king minus the vector for man plus the vector for woman is closest to the vector for queen, or *king* – *man* + *woman* ≈ *queen* (Mikolov et al. 2013, Pennington et al. 2014).[4] When trained on enough consistent text data, these word embedding models inscribe an enormous amount of social and cultural knowledge. A recent entrant in this class is the global vectors for word representation (GloVe) model, which combines global matrix factorization to capture word associations across each document with the rich, local context windows described above (Pennington et al. 2014). Once learned, vectors from these approaches can reveal the fine structure of words within a cultural world traced by its texts. Semantically similar words and documents show up in the same angular region of the high dimensional document space. Like the topics from topic models, dimensions and distances from trained vector spaces can also be deployed in a wide variety of supervised tasks, including sentiment analysis, document classification, and entity recognition (e.g., Taddy 2015).

MINING TEXT FOR SOCIAL THEORY

Many recent empirical articles and conference proceedings apply these NLP and ML methods to address questions central to sociological concerns. These articles sort themselves by the layer of communication on which they focus and by the depth of inferences they make about social games and the social world (see **Figure 4**). First, we review the substantial quantity of research devoted to generating knowledge from the manifest and latent content of communication in text. This research uses meaning-filled words, topics, and topic shifts to address patterns of collective framing and attention, but it also explores how society thinks by tracing the transmission, diffusion, recombination, and evolution of content. Second, an emerging stream of research has begun to study social relationships by analyzing the process of communication. This literature uses dynamic patterns of linguistic mimicry and synchrony to trace the deep and often hidden dynamics of social interaction underlying communication and the information that passes across them. Papers here focus on micro interactions, power and status dynamics, cultural embeddedness in communication, and information flow. Finally, there is burgeoning interest in using heterogeneous linguistic signals within communication to analyze social identities, states, roles, and moves. These articles attempt to access deep information within text about hidden elements of the social game being played and the social world beneath it. This work focuses on tracing sentiment, ideology, stances, and norms. Other articles in this vein identify actor roles and predict future moves. In addition, we note that some productive computational text analysis seeks to make no inferences but to draw on user-generated codes to bootstrap expanded samples of similar passages meriting qualitative investigation and inference. Such approaches directly extend qualitative text analysis by using NLP and ML to index enormous, unreadable libraries.[5]

[4]Such models perform between 70% and 80% accuracy on analogy tests. When we recently used word2vec, with 300 semantic dimensions trained on a one-billion-word Google News corpus—the same as Mikolov et al. (2013)—to probe cultural associations, we found many other associations suggestive of its ability to probe social and cultural fields, such as *hiphop* − *black* + *Mexican* ≈ *norteño*.

[5]Tangherlini & Leonard (2013) used Google books to sample passages with familiar themes (e.g., references to Darwin and evolution) in the vast archive of unread works. Shahaf et al. (2012) take a higher-level approach by using metrics of influence, coverage, and connectivity from the scientific literature to create structured summaries, or "metro maps," of information for expert perusal.

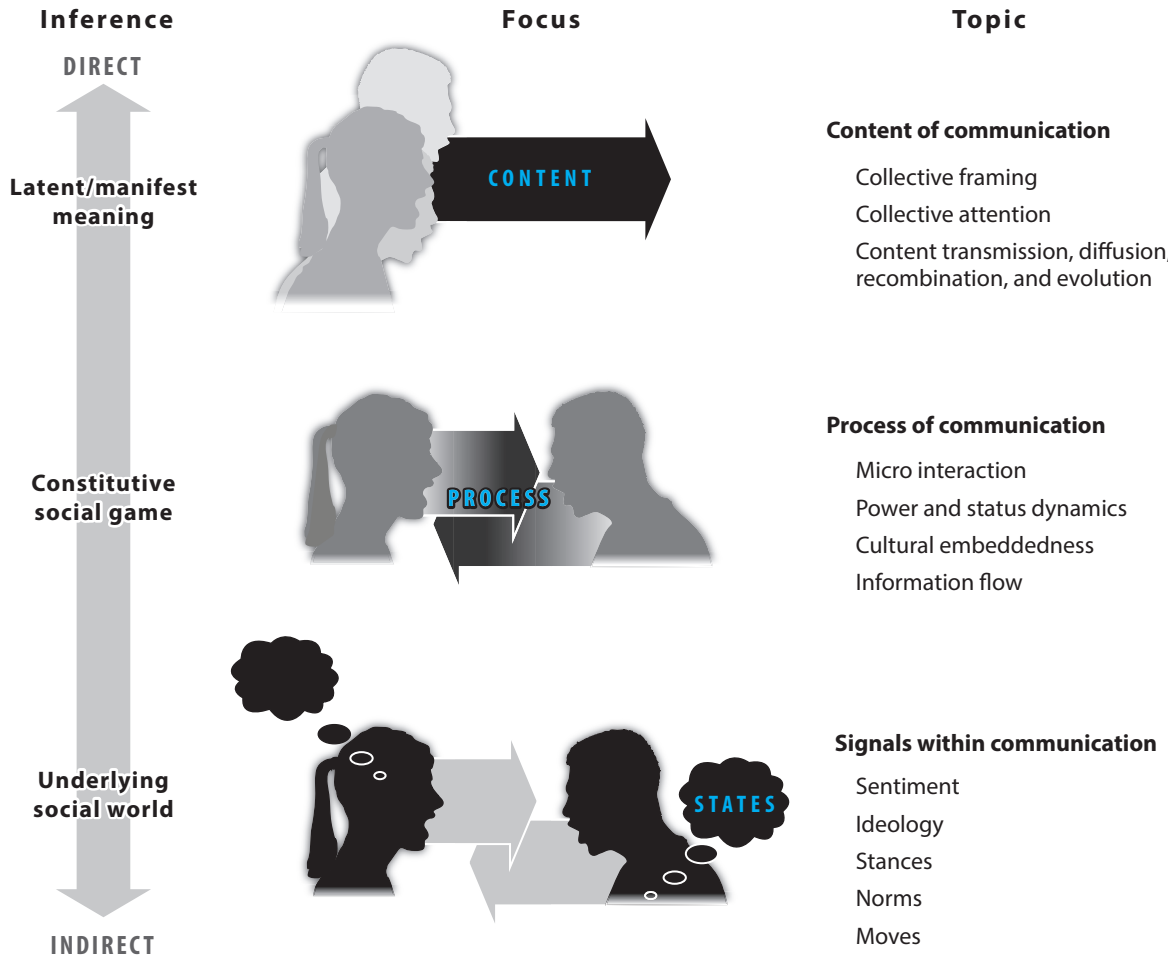


Figure 4

Social inferences from communication. Computational text analysis articles can be categorized according to the depth of inferences they make about the social world, with foci ranging from (*a*) collective attention, framing, and thinking through the manifest and latent content of communication; to (*b*) social relationships through analysis of the process of communication; and ultimately (*c*) social identities, states, roles, and moves through linguistic signals embedded in communication.

Collective Attention and Reasoning Through the Content of Communication

Articles reviewed in this section most directly extend the classical concerns of content analysis (Berelson & Lazarsfeld 1948), which first drew upon newspaper content (Woodward 1934), by focusing directly on cultural forms within communication—frames, issues, and topics. This work identifies the structure and dynamics of attention, agreement, and search across settings ranging from small groups and organizations to vast, far-flung publics, such as the Twitter sphere. Patterns of content can be stable or changing as well as unified, fragmented, or polarized. Studies tracing the dynamics of content trace how social collectives process information—how they collectively think—by tracing the growth, diffusion, mutation, recombination, and extinction of issues and topics.

Content has been used to directly investigate the structure of meaning in domains ranging from institutional logics, politics, and economics to culture and idle conversation. In a multimethod study, Nelson (2015) traced the institutional logics of women's organizations in Chicago and New York City during the early and middle twentieth century. By applying LDA topic modeling and qualitative in-depth readings to a corpus of women's movement organization publications, she showed that organizations institutionalize and embody local cognitive frames and political logics, which are then drawn upon over time as new organizations are formed, producing within-city continuities. Topic models allowed her to inductively discover sustained similarities in these logics and incorporate the findings into a historically informed understanding of the US women's movement, while extending theory on how the local institutionalization of cognitive structures persisted over time.

Other recent work has focused on political frames held by legislators, media organizations, consumers, and constituents (Grimmer & King 2011, Grimmer & Stewart 2013). Investigating senatorial press releases from 2005 to 2007, Grimmer (2013) used a topic model extension (Grimmer 2010) to simultaneously estimate the topic of each press release, the proportion of releases senators produced on each topic, and the category into which senators fell each year in office. This revealed how legislators presented themselves in a way that reflected their political alignment with constituencies. Politically aligned senators staked out political positions, whereas misaligned senators avoided controversial positions and instead claimed credit for appropriations to their districts. Gentzkow & Shapiro (2010) explored the link between politics and the market by developing a measure of media slant using phrases from the *Congressional Record* that statistically distinguished Democratic and Republican speakers. This allowed them to estimate the degree to which hundreds of American news media outlets echoed Democratic versus Republican congressional voices. When they combined slant with circulation data, Gentzkow & Shapiro (2010) found that the ideology of potential newspaper readers and not owners matched the newspaper's slant, suggesting an incentive to "tailor . . . slant to the ideological predispositions of consumers" (p. 64). In related work on the effect of Facebook's News Feed on users' consumption of discordant ideological content, Bakshy et al. (2015) directly measured Facebook users' expressions of political commitment and examined the relative influence of user choice and Facebook's algorithm at limiting exposure. They found that, although News Feed slightly limited exposure, individual choices to avoid discordant content played a much larger role.

Other recent work structured content according to the characteristics of those that produced it. For example, Jockers & Mimno (2013) studied works of nineteenth-century British, American, and Irish literature to demonstrate how factors such as the author's gender and nationality as well as the precise historical period of publication affected fluctuations in themes and word choices used to articulate them.

Still other research used computation to trace the allocation of attention across societal domains. Bail (2012) used plagiarism detection software to compare how national newspapers and television news stations distributed their attention across press releases about Islam by civil society organizations in the wake of the September 11th attacks. He found that the mass media paid more attention to fringe organizations. This then realigned organizational networks and shifted discourse surrounding terrorism. Similarly, Bonilla & Grimmer (2013) used LDA topic models to document how newspapers and nightly newscasts from major network stations allocated their attention to terrorism after elevation of the US government's color coded alert system.

The coherence and diversity of content has also been explored. In the context of political campaigns, Livne et al. (2011) analyzed tweets from federal candidates in the 2010 midterm elections to estimate content cohesiveness. They found conservative candidates portrayed the most coherent message, and Tea Party candidates displayed surprising topical and linguistic cohesiveness

despite their lack of formal organization. Tsur et al. (2015) used press releases from US representatives to investigate political framing and agenda-setting campaigns. Using LDA topic modeling and autoregressive-distributed-lag models, they found significant differences between the framing strategies of Democrats and Republicans. In the context of social movements, Bail (2014) investigated how advocacy organizations for organ donation produced different discourses in their appeal to multiple audiences. By developing a theory of cultural carrying capacity and putting it to work through structural topic modeling, Bail found an inverted-U-shaped relationship between a campaign's message diversity and social media endorsements, comments, and shares. Diverse content (e.g., social media messages that discussed religion, sports, or science) led to more public engagement but only up to a point, after which campaigns appeared incoherent, lacking shared purpose or collective identity.

Much recent work has examined shifts in content over history to identify changes in the social world underlying it. Michel et al. (2011) descriptively followed shifts in the usage and meaning of a wide range of terms and phrases from millions of books. They termed their approach culturomics or "the application of high throughput data collection and analysis to the study of human culture" (Michel et al. 2011, p. 181) and applied it to detect the rise of censorship in WWII Germany, the trajectory of fame, and shifting conventions of gender. In the political domain, Rule et al. (2015) used community detection algorithms on State of the Union transcripts from 1790 to 2014 to trace the emergence of modern political discourse through collocated terms. Likewise, Klingenstein et al. (2014) empirically documented the emergence of Elias's "civilizing process" through analysis of transcripts from the criminal proceedings in London's Central Criminal Court, the Old Bailey. Using the information theoretic Jensen-Shannon divergence, they found that from 1760 to 1910 discourse around violent and nonviolent crimes underwent a gradual but ultimately massive differentiation. Finally, Miller (2013) used LDA topic modeling on the Qing Veritable Records (a collection of Chinese documents important to an emperor's reign) to model typologies of violence held by government administrators. This typology provided insight into how different epochs understood violence and elucidated changing crime rates during the eighteenth and nineteenth centuries.

DiMaggio et al. (2013) traced discursive change by inducing topics from newspaper articles between 1986 and 1997 that discussed the National Endowment for the Arts. They found the tone of news coverage about the National Endowment for the Arts shifted after the election of George H.W. Bush from celebration to controversy and negativity. Marshall (2013) explored the topical nature and shifts within the discipline of demography in England and France. By tracing the prevalence of topics in 3,458 articles from 1946 to 2005, she delineated how research trends from each country reflected the cultural and institutional differences that shaped each country's understanding of fertility decline. Mohr et al. (2013) took a more structured approach to temporal shifts. Using named entities to identify actors, part-of-speech tagging to locate actions, and topics to capture scenes of action, the authors operationalized Kenneth Burke's "grammar of motives" across 11 national security documents produced by the US government between 1990 and 2010. Mapping rhetorical forms to structural properties, this work offered a suggestive approximation of deeper, more interpretive analysis regarding how the state legitimated its policies.

Attention and information diffusion follow stable, recurring patterns within many social games. Instant messaging conversations, fields of news production, and scientific disciplines each process information in characteristic ways. This collective reasoning results in stable patterns of attention, diffusion, mutation, and combination of content. Within mass and social media, Leskovec et al. (2009) have analyzed the temporal dynamics of the recurring news cycle on a large data set of news and social media sites. Tan et al. (2014) documented the effect of message wording on the diffusion of tweets by analyzing millions of paired tweets. Finally, Cheng et al. (2014) traced the

nature of photo reshare cascades on Facebook (over 150,000 photos) using caption and photo features as predictors. In science, Kuhn et al. (2014) have traced the diffusion of scientific memes, n-grams that propagate along the scientific citation graph, by analyzing published science in physics, biomedicine, and a broad sample of science and scholarship. Foulds & Smyth (2013) have similarly analyzed the proceedings of conferences to create a measure of topical influence that tracks the degree to which articles disseminate their topics to and through other articles that cite them.

Other studies analyze the evolution of content or how it transforms as it travels through communication channels. Much of this work looks at specific shifts in word meanings or the degraded fidelity of messages as they spread. Kulkarni et al. (2015) used word-embedding models built from Twitter, Amazon movie reviews, and a hundred years of n-grams from the Google Books corpus to find change points in word meaning and usage. Following the word "gay" across the twentieth century, they trace its trajectory from the semantic neighborhood of dapper to that of lesbian. Adamic and colleagues (Adamic et al. 2014, Simmons et al. 2011b) analyzed the dissemination of memes replicated millions of times on Facebook, discovering regularities governing the evolution of socially shared information. Simmons et al. (2011b) used the MemeTracker data set and found that mutations in quotations depend primarily on the authority of the copied source and nature of the quoting website. A paper by Gross (2014) explored the evolution of design logo content across an individual's design history to measure the creative effort behind each logo and to identify incentives for creativity within design competitions. This paper used image instead of textual content, but its edit-distance measure of difference between images was conceptually identical to Adamic et al.'s (2014) measure of difference between memes. This research suggests how shifts in content can be used to uncover deep cultural change relevant to many sociological concerns, from the evolution of gender roles to transformations in innovation and the economy.

Another class of articles explored the characteristic process by which textual elements are combined over time to suggest how collectives think, search, and discover and how they could be redesigned to do it more efficiently. In the context of financial policy making, Fligstein et al. (2014) investigated the lack of awareness at the Federal Open Market Committee of the impending economic meltdown in 2008. Through topic modeling, they showed that the Federal Reserve's primary analytic framework of macroeconomic theory prevented the Federal Open Market Committee from connecting the disparate forces of the crisis—the housing market, subprime mortgage market, and financial instruments used to package mortgages into securities—into a comprehensive picture. The common assumptions of those in charge at the Federal Reserve led to an inability to combine categories in the requisite ways to make sense of the crisis.

Other work tracing how collectives think has been carried out in science, which leaves a particularly detailed published trace of its content. Shi et al. (2015) used random walks along the network of scientific content from millions of biomedical abstracts in the US National Library of Medicine's MEDLINE corpus to reveal typical patterns of discovery. Foster et al. (2015) built on this by creating a network typology of discovery that revealed strong institutional pressures for scientists to exploit prior knowledge through incremental advances rather than explore the explosively expanding set of new opportunities. They found that rewards for high-risk innovation are not sufficient to compensate for the risk of not publishing, leading to few high-risk innovations. Rzhetsky et al. (2015) built a generative probabilistic network model to trace how molecules are typically combined in biomedicine and chemistry research, and then estimated it with extracted content on the molecules mentioned in millions of papers and patents over the latter half of the twentieth century. They then discovered optimal strategies through simulation and compared them with historical modes of collective discovery to identify inefficiencies associated with science as it is currently organized. These inefficiencies traced institutional incentives, like tenure,

which value sustained incremental productivity by an individual researcher over risky, collective advances.

These insights can also generate prescriptions, like Spangler et al. (2014) who used a similar approach, combining entity detection and graph-based information diffusion models to identify potential but untested relationships among scientific entities. They then examined these hypotheses through laboratory experiments and discovered novel relationships implied by the corpus. Research that has explored the process through which content is recombined to generate innovation has tended to focus on domains, such as molecular biology, where nouns and verbs are well-behaved (e.g., chemicals and reactions). Nevertheless, combined with robust approaches to dimension reduction, this work suggests a mode of analysis for precisely identifying and evaluating how institutions think relevant to the study of organizations, communities, cultures, and social movements. Recent movements including Occupy Wall Street, Black Lives Matter, and Fair Trade are all composed of evolving claims and shifting attention that could be better understood through the extraction of content with NLP tools at scale.

Social Relationships Through the Process of Communication

Beyond cultural forms, burgeoning attention in computational text analysis is now being paid to how such tools can be used to explore social relationships as they unfold through the process of communication. These issues have been core concerns of qualitative approaches, including ethnomethodology (Garfinkel 1967), studies of the interaction order (Goffman 2005), and conversation analysis (Schegloff 1995). Research using NLP and ML to investigate social relationships and the social games underlying communication is beginning to uncover deep regularities in interaction and is opening up exciting new areas of research. This research deals with communicative synchrony, the flow of information that results, and what this reveals about the relative positions of interaction partners and their outcomes for individuals, dyads, and groups. Data typically involve turn-taking patterns; the interactive pattern of words, phrases, and higher-order topics; and the movement of information through the ongoing flow of communication.

Recent computational research documents how social actors unequally match one another's linguistic style.[6] The imbalance reveals deep insight into positions within the social game constitutive of communication.[7] For example, Danescu-Niculescu-Mizil et al. (2012) used the micro dynamics of language coordination to trace shifting power differences between interaction partners by showing that lower-status parties to a conversation engage in conscious or unconscious mimicry of the distribution of function words (e.g., articles, prepositions, personal pronouns) expressed by those with higher status. In their analysis of conversational exchanges on Wikipedia forums, they found that low-power contributors end up coordinating much more consistently to the linguistic style of high-power administrators and that this shifts as editors are voted into and out of administrative positions. They also applied this manner of tracing language coordination to oral arguments in the US Supreme Court. Low-power lawyers coordinated much more consistently with high-power justices than the reverse.

[6] In Smith's *Theory of Moral Sentiments (TMS)*, he promoted propriety of both social and linguistic action that foregrounded sympathy or taking the other's perspective for achieving mutual correspondence and social harmony (Dascal 2006, Smith 1759). Smith mandated that sympathetic effort be equitably divided between interacting parties because it would be both improper and inefficient for either to lay upon the other "the whole burden" (Dascal 2006, Smith 1759).

[7] Later versions of Smith's *TMS* included an appendix, "Considerations concerning the first formation of languages." Scholarship shows how his ethics of social sympathy in the body of *TMS* has a direct corollary to how he describes linguistic interaction (Dascal 2006).

Even though those in favorable positions may be less likely to match the style of others, inability or unwillingness to match another's style does not improve one's position. Recent computational work using a similar design to the one described above demonstrated that polling increased for US presidential candidates who mirrored the language of their opponents in debates. Moreover, observers favored US presidents who mirrored partner language in negotiation transcripts (Romero et al. 2015). This is likely because matching an opponent's style translates one's argument for ease of understanding and reflects the ability—and flexibility—to take the other's perspective. This matches with findings from computational analysis of recorded negotiations that more communicative mirroring is positively associated with favorable negotiation outcomes (Pentland 2014).

Language style matching is not the only way to obtain domain-independent traces of power. Prabhakaran et al. (2012, 2014b) used indicators of overt displays of power (ODPs) that place constraints on the recipient (e.g., "come to my office now") to identify power dynamics within the corporate email communications of top Enron executives. Indicators were obtained using a supervised model trained on manual annotations and revealed that male superiors used significantly more ODPs compared to male and female subordinates and that female superiors used the least ODPs of all. Another approach to power measurement uses message control. Prabhakaran et al. (2014b) analyzed US presidential primary debates to identify speaker turns when conversation changed from one topic to another. Higher-powered candidates—those posting higher poll numbers—were less likely to shift topics during the course of the debate, suggesting that control of the conversation is a consequence of electoral popularity.

McFarland et al. (2013a,b) explored moves that take place during the game of courtship by analyzing audio and textual content from thousands of speed-dating encounters. Marshaling both acoustic and linguistic data from the transcribed conversations, they derived measures related to emotional intensification (prosodic attributes such as pitch, loudness, and rate of speech) and conversational synchronization (lexical, syntactic, pragmatic, and interactional attributes such as interactional targeting, interpersonal alignment, and situational alignment). They found that mutual excitement is related to social bonding and the selection of a speed-dating partner for further contact, but this was contingent on gender. For men, excitement was expressed through laughter and variance in volume, whereas for women it was expressed through the raising and varying of vocal pitch. Furthermore, participants felt a connection when men expressed sympathy and gratitude for their dates, and women engaged the situation and targeted themselves as a subject through use of the pronoun "I." It was unclear how much of this particular courtship game resulted from immutable differences in US gender identities versus the choreography of the game itself in which men were instructed to physically rotate and women stayed still, possibly giving them the upper hand in interaction.

Social relationships have also been explored in work on social and cultural embeddedness (Pachucki & Breiger 2010). Danescu-Niculescu-Mizil et al. (2013) explored this by using a decade of reviews from two review communities to discover a cultural life cycle through which users came to learn online community norms of expression. New members entered the community unwittingly introducing and perpetuating linguistic innovations but eventually synchronizing with the community. Subsequently, their language stabilized and became rigid, but community language norms continued to evolve, and they eventually fell out of sync. This disconnect with community language predicted their disengagement and ultimate exit, prompting the paper's title "No Country for Old Members." This relates to findings from Saavedra et al. (2011), who perform a nontextual, frequency-based analysis of communication embeddedness by looking at the pattern of instant-messaging activity among traders in a day-trading firm, which revealed that instant-messaging patterns among their individual networks enabled them to trade synchronously, which in turn decreased their likelihood of losing money at the end of the day.

Goldberg et al. (2015b) investigated cultural and structural embeddedness and their joint influence on individual attainment within an organization. Analyzing millions of emails from a medium-sized technology firm, they operationalized a measure of cultural embeddedness on the basis of the degree to which the language people use is more or less similar to within-organization emails they receive in a given month. This measure of cultural embeddedness was then analyzed along with their structural embeddedness—position in the email network—revealing that structural and cultural embeddedness are inversely related with respect to community advancement. Brokers did better when they demonstrated higher cultural fit but individuals in structurally cohesive positions did better when they were culturally distinct. Using the same data set, Goldberg et al. (2015a) also analyzed the enculturation trajectories of individuals, finding that employees who were slow to enculturate during the early part of their tenure with the firm were more likely to exit involuntarily compared to those who quickly adapted to organizational communication norms.

At the field level, Vilhena et al. (2014) analyzed the nature of communication within and between scientific disciplines by capturing the distribution of phrase frequencies in articles as well as the citation linkages between them from the JSTOR corpus (1.5 million scientific articles). The cultural space was mapped in terms of the communicative burden placed on interacting individuals and was measured as the ratio of entropy and cross-entropy rates between members of one community and those of another. This meant that individuals operating under very distinct phrase distributions (e.g., sociology and molecular biology) expended much more effort in understanding one another than those interacting against a backdrop of similar phrase distributions (e.g., economics and political science). Knowing the phrase frequencies of different subfields allows for the tracing of entire cultural spaces, which can then be mapped onto interactional structures (such as citation flows), leading to a topographical landscape rich in insights.

Patterns in the process of communication alter how information flows through the relationships facilitating it. Aral & Van Alstyne (2011) used hundreds of thousands of email messages at an executive recruiting firm to understand how novel information flowed through the network. They measured bandwidth as the monthly volume of communication between individuals, and novelty as the introduction of distinct new content in a vector space model of email content. They found that as recruiters communicated across structural holes, they received more diverse information, but forewent communication bandwidth, reducing the overall volume of novel information received. In this way, social relationships were traced not only through webs of interaction, but also the process and content of the communication itself.

Pentland, his Human Dynamics Laboratory within MIT's Media lab, and collaborators (Pentland 2012b, Woolley et al. 2010) have used wearable devices, (e.g., sociometric badges) to reach beyond text and to focus on nonlinguistic features of human communication, such as patterns of turn taking, tone of voice, facial movement, and gesture, to measure properties like activity, engagement, and mirroring within conversation. They have then used these qualities to predict outcomes. For example, rapid, even patterns of turn taking within groups are associated with greater problem-solving success (Woolley et al. 2010). Moreover, the presence of group conversations in which members face one another, side conversations, and energetic engagement outside meetings are all associated with greater satisfaction and collective group performance (Pentland 2012b).

Social encounters "with differentially empowered individuals, complementary parts, and mixed motives" (McFarland et al. 2013a, p. 1641) are common across many social games. Although power and status dynamics may be easily observed through markers, such as professions or institutional affiliations, understanding their subtle and various interaction markers through NLP and ML tools is adding detail in the form of social processes and mechanisms to classic sociological arguments while discovering new ones. Collectively, articles on power dynamics

(Danescu-Niculescu-Mizil et al. 2012, Prabhakaran et al. 2014a,b) presented strong evidence for domain-independent methods of inferring status relationships within communication processes. We argue that these methods can be usefully applied to a wide range of sociological inquiries where power and status dynamics are of special interest. These include contexts in which differences of race, class, gender, occupation, and other identities drive unbalanced interactions, including criminal courts, online discussion forums, town hall meetings, and anywhere with readily available sources of textual data. Many of these contexts may also generate large volumes of nondigital information, which can now be transferred to digital form through optical character recognition and speech-to-text technology, allowing scholars to trace the historical evolution of such relationships. ML approaches to the automatic classification of interactional variables from audio and video data could dramatically expand the interactional data available for conversation analysts (Stivers et al. 2009).

Social States Through Heterogeneous Signals Within Communication

Linguistic communication exhibits regularities that serve as signals tracing deeper aspects of the social world underlying it. This section reviews work that makes inferences about human and collective states, which underlie and conceptually precede both content and interaction. The research outlined below uses statistical models and ML algorithms to predict phenomena beneath human communication at or beyond the performance of human experts. Text and associated communication data are used to identify internal human states, such as sentiment and preference, social roles and stances behind communicated utterances, and dispositions that predict future strategic moves.

Human and collective states reflect their condition at a given time and place. States include sentiment, preference, ideology, stance (for or against), norms, uncertainty, and disposition to perform a predicted action in the future. Sentiment analysis has been at the center of computational content analysis since Stone's General Inquirer System (Stone et al. 1966) used dictionary-based classifiers to capture lexical traces of positive, negative, and neutral affect from speeches and news. Scholars have continued to make progress in reliably extracting sentiment from text, with existing methods that deploy neural network models over dependency trees achieving greater than 85% accuracy (Socher et al. 2013). These approaches can capture the effect of contrastive conjunctions (e.g., but, however) as well as negation (e.g., not, neither) and sentiment scope at various levels within grammar trees for both positive and negative phrases.

Research has used sentiment analysis at both individual and collective levels. At the individual level, Sudhof et al. (2014) modeled dependency paths of human emotional states within the context of product reviews, finding that reviewers are swayed by sentiment from prior reviews. Similarly, Kramer et al. (2014) documented large-scale emotional contagion in their controversial study using Facebook to demonstrate how small changes in the visibility of positive and negative content within a user's news feed influenced that user to mirror this affect in their own posts. Taking these insights to the clinic, Resnik et al. (2013) used dictionaries, supervised classification, and topic modeling to aid in the clinical assessment of neuroticism and depression through sentiment identification. At the macro level, Golder & Macy (2011) found globally characteristic patterns of mood across day, week, and season expressed on Twitter, and several studies have looked at the effect that collective sentiment can have on stock market returns (Bollen et al. 2011, Nguyen & Shirai 2015). Other work has examined the recursive effect that events, captured by news volume, have on collective sentiment (Tsytsarau et al. 2014).

Tracing ideology has become an active area of research predicting human states. Iyyer et al. (2014) applied a deep learning framework to infer the political position of sentences, and Sim

et al. (2013) learned an ideological space from a corpus of explicitly ideological books and then used a probabilistic model to predict ideological valence from politician speeches. Jelveh et al. (2014) correctly predicted the ideological leanings of professional economists through a supervised ensemble n-gram model applied to their research papers.

Stance (i.e., for or against) and disagreement have been analyzed in the context of online debates, with Sridhar et al. (2015) evaluating author and online post level stances within online debates, and Hasan & Ng (2014) going deeper to analyze not only stance but also the reasons behind it. At the level of markets, Baker et al. (2013) have used automated text search for 10 of the largest newspapers to create a measure of national-level economic uncertainty and trace it through historical news to identify its influence on capital markets. Although most psychological and social states have been examined at the individual level, research has begun to trace shared or collective states that underlie communicated content, which opens up new opportunities for the sociology of emotion, culture, and knowledge.

Clues within communication have also been used to predict strategic motivations and their influence on future moves in a variety of social games. One study used the online game Diplomacy, in which individuals engage in dyadic exchanges to form and dissolve alliances, to trace the linguistic harbingers of betrayal (Niculae et al. 2015). When the linguistic balance between partners changes by becoming more positive, polite, or focused on the future, betrayal follows soon thereafter. Similarly, Yu et al. (2015) uncovered linguistic signals of deception in the online Killer Game, where teams of killers, detectives, and citizens work to convince each other of their roles over several rounds. Their subgroup detection method outperforms humans at detecting signals of deception. In another interesting online game, Cadilhac et al. (2013) predicted player trades in Settlers of Catan, a game where players trade resources with each other to develop their regions and earn points. Their model used dialog acts to dynamically estimate player preferences as the game unfolded. These papers are indicative of the kind of work that can be carried out within sociology to better understand social action and the dispositions that precede it.

Textual clues have also been used to identify the roles of actors within social games as well as strategies to make actors more effective in their roles. Cheng et al. (2015) used data on banned members from three different online communities to predict antisocial individuals, or trolls. They found that trolls "tend to concentrate their efforts in a small number of threads, are more likely to post irrelevantly, and are more successful at garnering responses from other users" (Cheng et al. 2015, p. 1). Similarly, Blackburn & Kwak (2014) predicted toxic behavior in the League of Legends online game by using a supervised learning approach on the decisions of over 10 million user reports involving 1.46 million toxic players. In a different context, Mukherjee et al. (2014) used a Markov random field model to establish which user-generated medical statements on one of the largest online health communities were credible and trustworthy. Yang et al. (2015) studied teams in two massive open online courses to identify latent conversational roles played by students. After discovering these roles, and inferring their optimal distribution within a team, they validated the causal efficacy of this distribution by designing more efficient teams for subsequent projects on the basis of those role distributions. Finally, Wallace et al. (2013) studied the relationship between micro roles, or topic and speech act distributions, within established doctor-patient interactions that improve or harm antiretroviral medication adherence.

Together, this research used text and related communicative traces as inputs to models that detect underlying regularities in the states of individuals and collectives—the social world on which social games are played. These approaches suggest unrealized opportunities to predict other states of the social world underlying communication, including shared preferences, discriminatory biases, and a wide range of cultural assumptions that filter communication.

CONCLUSION

So much of the social world is mediated or traced by digital text today that it has come to represent a major channel through which sociologists can understand the social dynamics of the present and past. Ignoring the potential for text to illuminate our sociological understanding of virtually any contemporary social domain—from culture, courtship, and sexual encounters to commerce, politics, and science—would be closing our eyes to the primary data stream that social media, information, and big data companies use to deliver actionable insight to all sectors of the knowledge economy (Bail 2014, Golder & Macy 2014). Moreover, attempting to analyze the expanding universe of text through conventional reading exceeds the limits of human capacity. This forces the qualitative analyst to sample at rates that make it difficult to reach robust, generalizable conclusions. Here, we have surveyed some of the most exciting computational approaches to text analysis as well as their application in research that seeks sociological insight.

These tools have been used in three broad ways. Most frequently, they have been used to trace collective attention and reasoning through analyzing content within communicated text. A burgeoning collection of studies go beyond content, using interactive text to analyze social relationships revealed through the process of communication. Finally, a third collection of studies reaches deeper, making inferences about social states through signals hidden within communicated text. These three approaches point to different levels of sociological inference. Nevertheless, much of this research has not been performed by sociologists but rather by computer scientists who hold a commitment to building new tools and to demonstrating them suggestively on large but not always well-curated samples of text and related content. For example, in Tsur et al.'s (2015) analysis of differences between the framing strategies of Democrats and Republicans, they left "detailed analysis of the interplay between the different frames . . . for political scientists" (p. 1636). This interpretive handoff highlights an opportunity for sociology. By leaving careful data collection, substantive interpretation, and theoretical implications to the social scientists, this work invites sociologists to engage with this community and its models.

Beyond simply testing or extending existing sociological theory, NLP and ML approaches to text analysis have the potential to generate enormous quantities of socially relevant data from a wide range of contexts. We have shown how supervised prediction models can now reliably identify power differences, preferences, and dispositions from text and related content. Unsupervised models are generating associations of increasing complexity and accuracy, including topic models and word-embedding spaces that capture stable cultural associations known to society as a whole but not to any one person. We also point to new ML opportunities for the analysis of other traces of human behavior, interaction, and communication beyond text from images, audio, video, and other digital social data, such as likes on Facebook or swipes on Tinder. We hope to have demonstrated that, although ML and NLP cannot reproduce the subtlety of a creative researcher who brings a life of prior associations to their analysis, computational methods trained on big data can generate many suggestive, subtle associations beyond the sensitivity of human perception and the capacity of human memory.

The wealth of new data this is making available about social relationships, cultural associations, micro states, and behaviors holds the potential to surprise and inform researchers, provoking construction of new descriptive and predictive theory (Glaser & Strauss 1967, Tavory & Timmermans 2014) in virtually every substantive domain of sociological inquiry, from social, economic, and political life to organizations and social movements, inequality, race, and gender. This is why we titled our review "Machine Translation: Mining Text for Social Theory." ML is enabling the translation of text into social data, which is increasingly being mined for theoretical possibilities.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

The authors wish to thank Dan McFarland and several other authors whose work is reviewed here for constructive comments and pointers to additional, relevant research. We acknowledge funding from NSF SBE-1158803 and 0915730 and from a John Templeton Foundation grant to the Metaknowledge Network.

LITERATURE CITED

Adamic LA, Lento TM, Adar E, Ng PC. 2014. Information evolution in social networks. arXiv:1402.6792v1 [cs.SI]

Alberti C, Weiss D, Coppola G, Petrov S. 2015. Improved transition-based parsing and tagging with neural networks. *Proc. 2015 Conf. Empir. Methods Nat. Lang. Process.*, Sept. 17–21, pp. 1354–59, Seattle, WA: Assoc. Comput. Linguist.

Aral S, Van Alstyne M. 2011. The diversity-bandwidth trade-off. *Am. J. Sociol.* 117(1):90–171

Bail CA. 2012. The fringe effect civil society organizations and the evolution of media discourse about Islam since the September 11th attacks. *Am. Sociol. Rev.* 77(6):855–79

Bail CA. 2014. The cultural environment: measuring culture with big data. *Theory Soc.* 43(3–4):465–82

Bail CA. 2016. Cultural carrying capacity: organ donation advocacy, discursive framing, and social media engagement. *Soc. Sci. Med.* In press. doi:10.1016/j.socscimed.2016.01.049

Baker SR, Bloom N, Davis SJ. 2013. *Measuring economic policy uncertainty*. NBER Work. Pap. 21633, Natl. Bur. Econ. Res., Cambridge, MA

Bakshy E, Messing S, Adamic LA. 2015. Exposure to ideologically diverse news and opinion on Facebook. *Science* 348:1130–32

Bales RF. 1950. A set of categories for the analysis of small group interaction. *Am. Sociol. Rev.* 15(2):257–63

Berelson B, Lazarsfeld PF. 1948. *The Analysis of Communication Content*. Oslo, Nor.: Univ. Stud.

Blackburn J, Kwak H. 2014. STFU NOOB!: predicting crowdsourced decisions on toxic behavior in online games. *Proc. 23rd Int. Conf. World Wide Web*, Apr. 7–11, pp. 877–88. New York: Assoc. Comput. Mach.

Blei DM. 2012. Probabilistic topic models. *Commun. ACM* 55(4):77–84

Blei DM, Lafferty JD. 2006. Dynamic topic models. *Proc. 23rd Int. Conf. Mach. Learn.* June 25–29, pp. 113–20. New York: Assoc. Comput. Mach.

Blei DM, Ng AY, Jordan MI. 2003. Latent Dirichlet allocation. *J. Mach. Learn. Res.* 3:993–1022

Bollen J, Mao H, Zeng X. 2011. Twitter mood predicts the stock market. *J. Comput. Sci.* 2(1):1–8

Bonilla T, Grimmer J. 2013. Elevated threat levels and decreased expectations: How democracy handles terrorist threats. *Poetics* 41(6):650–69

Bourdieu P. 2013. *Distinction: A Social Critique of the Judgement of Taste*. Abington, UK: Routledge

Cadilhac A, Asher N, Benamara F, Lascarides A. 2013. Grounding strategic conversation: using negotiation dialogues to predict trades in a win-lose game. *Proc. 2013 Conf. Empir. Methods Nat. Lang. Process.*, Oct. 18–21, pp. 357–68. Seattle, WA: Assoc. Comput. Linguist.

Carley K. 1993. Coding choices for textual analysis: a comparison of content analysis and map analysis. *Sociol. Methodol.* 23:75–126

Carley K. 1994. Extracting culture through textual analysis. *Poetics* 22(4):291–312

Carley KM, Kaufer DS. 1993. Semantic connectivity: an approach for analyzing symbols in semantic networks. *Commun. Theory* 3(3):183–213

Chang J, Blei DM. 2010. Hierarchical relational models for document networks. *Ann. Appl. Stat.* 4:124–50

Cheng J, Adamic L, Dow PA, Kleinberg JM, Leskovec J. 2014. Can cascades be predicted? See Blackburn & Kwak, 2014, pp. 925–36
Cheng J, Danescu-Niculescu-Mizil C, Leskovec J. 2015. Antisocial behavior in online discussion communities. arXiv:1504.00680v1 [cs.SI]
Clark A, Fox C, Lappin S, eds. 2010. *The Handbook of Computational Linguistics and Natural Language Processing*. Oxford, UK: Wiley-Blackwell
Corman SR, Kuhn T, Mcphee RD, Dooley KJ. 2002. Studying complex discursive systems: centering resonance analysis of communication. *Hum. Commun. Res.* 28(2):157–206
Danescu-Niculescu-Mizil C, Lee L, Pang B, Kleinberg J. 2012. Echoes of power: language effects and power differences in social interaction. *Proc. 21st Int. Conf. World Wide Web*, Apr. 16–20, pp. 699–708. New York: Assoc. Comput. Mach.
Danescu-Niculescu-Mizil C, West R, Jurafsky D, Leskovec J, Potts C. 2013. No country for old members: user lifecycle and linguistic change in online communities. *Proc. 22nd Int. Conf. World Wide Web*, May 13–17, pp. 307–18. New York: Assoc. Comput. Mach.
Dascal M. 2006. Adam Smith's theory of language. In *The Cambridge Companion to Adam Smith*, pp. 79–111. Cambridge, UK: Cambridge Univ. Press
de Marneffe M-C, Dozat T, Silveira N, Haverinen K, Ginter F, et al. 2014. Universal Stanford dependencies: a cross-linguistic typology. *Proc. 9th Int. Conf. Lang. Resour. Eval.* **http://www.lrec-conf.org/proceedings/lrec2014/index.html**
DiMaggio P, Nag M, Blei D. 2013. Exploiting affinities between topic modeling and the sociological perspective on culture: application to newspaper coverage of U.S. government arts funding. *Poetics* 41(6):570–606
Fligstein N, Brundage JS, Schultz M. 2014. *Why the Federal Reserve failed to see the financial crisis of 2008: the role of "macroeconomics" as a sense making and cultural frame*. IRLE Work. Pap. #111-14, Inst. Res. Labor Employ., Univ. Calif. Berkeley
Foster JG, Rzhetsky A, Evans JA. 2015. Tradition and innovation in scientists' research strategies. *Am. Sociol. Rev.* 80:875–908
Foulds JR, Smyth P. 2013. Modeling scientific impact with topical influence regression. See Cadilhac et al. 2013, pp. 113–23
Franzosi R. 2004. *From Words to Numbers: Narrative, Data, and Social Science*. Cambridge, UK: Cambridge Univ. Press
Freese J. 2007. Replication standards for quantitative social science: Why not sociology? *Sociol. Methods Res.* 36(2):153–72
Garfinkel H. 1967. *Studies in Ethnomethodology*. Malden, MA: Blackwell
Gentzkow M, Shapiro JM. 2010. What drives media slant? Evidence from U.S. daily newspapers. *Econometrica* 78(1):35–71
Gibson DR. 2012. *Talk at the Brink: Deliberation and Decision during the Cuban Missile Crisis*. Princeton, NJ: Princeton Univ. Press
Glaser BG. 1965. The constant comparative method of qualitative analysis. *Soc. Probl.* 12(4):436–45
Glaser BG, Strauss AL. 1967. *The Discovery of Grounded Theory; Strategies for Qualitative Research*. Chicago: Aldine
Goffman E. 2005. *Interaction Ritual: Essays in Face to Face Behavior*. New Brunswick, NJ: Aldine Trans.
Goldberg A, Srivastava SB, Manian VG, Monroe W, Potts C. 2015a. *Enculturation trajectories: An interactional language use model of cultural dynamics in organizations.* Work. Pap., Haas School Bus., Univ. Calif. Berkeley
Goldberg A, Srivastava SB, Manian VG, Monroe W, Potts C. 2015b. *Fitting in or standing out? The tradeoffs of structural and cultural embeddedness.* Work. Pap. 3285, Stanford Univ. Grad. School Bus.
Golder SA, Macy MW. 2011. Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. *Science* 333(6051):1878–81
Golder SA, Macy MW. 2014. Digital footprints: opportunities and challenges for online social research. *Annu. Rev. Sociol.* 40(1):129–52
Griffiths TL, Steyvers M, Blei DM, Tenenbaum JB. 2005. Integrating topics and syntax. In *Advances in Neural Information Processing Systems 17*, ed. LK Saul, Y Weiss, L Bottou, pp. 537–44. Cambridge, MA: MIT Press

Grimmer J. 2010. A Bayesian hierarchical topic model for political texts: measuring expressed agendas in senate press releases. *Polit. Anal.* 18(1):1–35

Grimmer J. 2013. Appropriators not position takers: the distorting effects of electoral incentives on congressional representation. *Am. J. Polit. Sci.* 57(3):624–42

Grimmer J, King G. 2011. General purpose computer-assisted clustering and conceptualization. *PNAS* 108(7):2643–50

Grimmer J, Stewart BM. 2013. Text as data: the promise and pitfalls of automatic content analysis methods for political texts. *Polit. Anal.* 21:2667–97

Grogger J. 2011. Speech patterns and racial wage inequality. *J. Hum. Resour.* 46(1):1–25

Gross DP. 2014. *Creativity under fire: the effects of competition on innovation and the creative process*. Work. Pap., Job Market Pap., Univ. Calif. Berkeley

Hardt H. 2001. *Social Theories of the Press: Constituents of Communication Research, 1840s to 1920s*. Lanham, MD: Rowman & Littlefield

Hasan KS, Ng V. 2014. Why are you taking this stance? Identifying and classifying reasons in ideological debates. *Proc. 2014 Conf. Empir. Methods Nat. Lang. Process* Oct. 25–29, pp. 751–62. Seattle, WA: Assoc. Comput. Linguist

Hays DC. 1960. *Automatic Content Analysis*. Santa Monica, CA: Rand Corp.

Hearst MA. 1997. TextTiling: segmenting text into multi-paragraph subtopic passages. *Comput. Linguist.* 23(1):33–64

Hinton GE, Osindero S, Teh Y-W. 2006. A fast learning algorithm for deep belief nets. *Neural Comput.* 18(7):1527–54

Hirschberg J, Manning CD. 2015. Advances in natural language processing. *Science* 349(6245):261–66

Huizinga J. 1971. *Homo Ludens: A Study of the Play-Element in Culture*. Boston, MA: Beacon

Ioannidis JP. 2005. Why most published research findings are false. *PLOS Med.* 2(8):e124

Ioannidis JP, Doucouliagos C. 2013. What's to know about the credibility of empirical economics? *J. Econ. Surv.* 27(5):997–1004

Iyyer M, Enns P, Boyd-Graber J, Resnik P. 2014. Political ideology detection using recursive neural networks. *Proc. 52nd Annu. Meet. Assoc. Comput. Linguist.: Styst. Demonstr., Jun. 23–25, Baltimore, MD*, pp. 1113–22. Stroudsburg, PA: Assoc. Comput. Linguist.

Jakobson R. 1960. Closing statement: linguistics and poetics. In *Style in Language*, ed. TA Sebeok, pp. 350–77. Cambridge, MA: MIT Press

Jelveh Z, Kogut B, Naidu S. 2014. Detecting latent ideology in expert text: evidence from academic papers in economics. See Hasan & Ng, 2014, pp. 1804–9

Jockers ML, Mimno D. 2013. Significant themes in 19th-century literature. *Poetics* 41(6):750–69

Jordan MI, Mitchell TM. 2015. Machine learning: trends, perspectives, and prospects. *Science* 349(6245):255–60

Joshi M, Das D, Gimpel K, Smith NA. 2010. Movie reviews and revenues: an experiment in text regression. *Proc. Hum. Lang. Technol. 2010 Annu. Conf. N. Am. Chapter Assoc. Comput. Linguist*, pp. 293–96. Stroudsburg, PA: Assoc. Comput. Linguist.

Jurafsky D, Martin JH. 2000. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Upper Saddle River, NJ: Prentice Hall

Kemp C, Tenenbaum JB. 2008. The discovery of structural form. *PNAS* 105(31):10687–92

Klingenstein S, Hitchcock T, DeDeo S. 2014. The civilizing process in London's Old Bailey. *PNAS* 111(26):9419–24

Kramer ADI, Guillory JE, Hancock JT. 2014. Experimental evidence of massive-scale emotional contagion through social networks. *PNAS* 111(24):8788–90

Kuhn T, Perc M, Helbing D. 2014. Inheritance patterns in citation networks reveal scientific memes. *Phys. Rev. X* 4(4):041036

Kulkarni V, Al-Rfou R, Perozzi B, Skiena S. 2015. Statistically significant detection of linguistic change. *Proc. 24th Int. Conf. World Wide Web*, May 18–22, pp. 625–35. New York: Assoc. Comput. Mach.

Laver M, Benoit K, Garry J. 2003. Extracting policy positions from political texts using words as data. *Am. Polit. Sci. Rev.* 97(02):311–31

Lee M, Martin JL. 2015. Coding, counting and cultural cartography. *Am. J. Cult. Sociol.* 3(1):1–33
Leskovec J, Backstrom L, Kleinberg J. 2009. Meme-tracking and the dynamics of the news cycle. *Proc. 15th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, pp. 497–506. New York: Assoc. Comput. Mach.
Livne A, Simmons MP, Adar E, Adamic LA. 2011. The party is over here: structure and content in the 2010 election. *Proc. Fifth Int. AAAI Conf. Weblogs Soc. Media.*, *Barcelona*, *July 17–21*
Long NE. 1958. The local community as an ecology of games. *Am. J. Sociol.* 64(3):251–61
Loughran T, McDonald B. 2011. When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *J. Finance* 66(1):35–65
Manning CD. 2015. Computational linguistics and deep learning. *Comput. Linguist.* 41:701–7
Manning CD, Raghavan P, Schütze H. 2008. *Introduction to Information Retrieval*. Cambridge, UK: Cambridge Univ. Press
Manning CD, Schütze H. 1999. *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press
Manning CD, Surdeanu M, Bauer J, Finkel J, Bethard SJ, McCloskey D. 2014. The Stanford CoreNLP natural language processing toolkit. See Iyyer et al. 2014, pp. 55–60
Marshall EA. 2013. Defining population problems: using topic models for cross-national comparison of disciplinary development. *Poetics* 41(6):701–24
Massy WF. 1965. Principal components regression in exploratory statistical research. *J. Am. Stat. Assoc.* 60(309):234–56
McAuliffe JD, Blei DM. 2008. Supervised topic models. In *Advances in Neural Information Processing Systems 20*, ed. JC Platt, D Koller, Y Singer, pp. 121–28. Cambridge, MA: MIT Press
McFarland DA, Jurafsky D, Rawlings C. 2013a. Making the connection: social bonding in courtship situations. *Am. J. Sociol.* 118(6):1596–649
McFarland DA, Ramage D, Chuang J, Heer J, Manning CD, Jurafsky D. 2013b. Differentiating language usage through topic models. *Poetics* 41(6):607–25
Michel J-B, Shen YK, Aiden AP, Veres A, Gray MK, et al. 2011. Quantitative analysis of culture using millions of digitized books. *Science* 331(6014):176–82
Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26*, ed. CJC Burges, L Bottou, M Welling, Z Ghahramani, KQ Weinberger, pp. 3111–19. Red Hook, NY: Curran
Miller IM. 2013. Rebellion, crime and violence in Qing China, 1722–1911: a topic modeling approach. *Poetics* 41(6):626–49
Mohr JW, Bogdanov P. 2013. Introduction—topic models: What they are and why they matter. *Poetics* 41(6):545–69
Mohr JW, Wagner-Pacifici R, Breiger RL, Bogdanov P. 2013. Graphing the grammar of motives in national security strategies: cultural interpretation, automated text analysis and the drama of global politics. *Poetics* 41(6):670–700
Mosteller F, Wallace DL. 1964. *Inference and Disputed Authorship: The Federalist*. Boston, MA: Addison-Wesley
Mukherjee S, Weikum G, Danescu-Niculescu-Mizil C. 2014. People on drugs: credibility of user statements in health communities. *Proc. 20th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, pp. 65–74. New York: Assoc. Comput. Mach.
Myerson RB. 2013. *Game Theory: Analysis of Conflict*. Cambridge, MA: Harvard Univ. Press
Nelson LK. 2015. *Political logics as cultural memory: cognitive structures, local continuities, and women's organizations in Chicago and New York City*. Work. Pap. Kellogg School Manag., Northwestern Univ.
Nguyen TH, Shirai K. 2015. Topic modeling based sentiment analysis on social media for stock market prediction. *Proc. 53rd Annu. Meet. Assoc. Comput. Linguist. 7th Jt. Conf. Nat. Lang. Process. Int.*, Beijing, July 26–31, pp. 1354–64. **http://www.anthology.aclweb.org/P/P15/P15-1131.pdf**
Niculae V, Kumar S, Boyd-Graber J, Danescu-Niculescu-Mizil C. 2015. Linguistic harbingers of betrayal: a case study on an online strategy game. *Proc. 53rd Annu. Meet. Assoc. Comput. Linguist. 7th Int. Jt. Conf. Nat. Lang. Process.*, Beijing, July 26–31, pp. 1650–59
O'Neil C, Schutt R. 2013. *Doing Data Science: Straight Talk from the Frontline*. Sebastopol, CA: O'Reilly
Pachucki MA, Breiger RL. 2010. Cultural holes: beyond relationality in social networks and culture. *Annu. Rev. Sociol.* 36(1):205–24

Pang B, Lee L. 2008. Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.* 2(1–2):1–135
Pennebaker JW, Francis ME, Booth RJ. 2001. *Linguistic Inquiry and Word Count: LIWC 2001*. Mahwah, NJ: Erlbaum
Pennington J, Socher R, Manning C. 2014. Glove: global vectors for word representation. See Hasan & Ng 2014, pp. 1532–43
Pentland A. 2012a. Big data's biggest obstacles. *Harv. Bus. Rev.*. October. **https://hbr.org/2012/10/big-datas-biggest-obstacles**
Pentland A. 2012b. The new science of building great teams. *Harv. Bus. Rev.* 90(4):60–69
Pentland A. 2014. *Social Physics: How Good Ideas Spread—The Lessons from a New Science*. New York: Penguin
Prabhakaran V, Arora A, Rambow O. 2014a. Staying on topic: an indicator of power in political debates. See Hasan & Ng 2014, pp. 1481–86
Prabhakaran V, Rambow O, Diab M. 2012. Predicting overt display of power in written dialogs. *Proc. 2012 Conf. N. Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol.*, Montreal, Can., Jun. 3–8, pp. 518–22. Stroudsburg, PA: Assoc. Comput. Linguist.
Prabhakaran V, Reid EE, Rambow O. 2014b. Gender and power: How gender and gender environment affect manifestations of power. See Hasan & Ng 2014, pp. 1965–76
Resnik P, Garron A, Resnik R. 2013. Using topic modeling to improve prediction of neuroticism and depression in college students. See Cadilhac et al. 2013, pp. 1348–53
Rickford JR, Duncan GJ, Gennetian LA, Gou RY, Greene R, et al. 2015. Neighborhood effects on use of African-American vernacular English. *PNAS* 112(38):11817–22
Rodriguez-Esteban R, Rzhetsky A. 2008. Six senses in the literature. The bleak sensory landscape of biomedical texts. *EMBO Rep.* 9(3):212–15
Romero DM, Swaab RI, Uzzi B, Galinsky AD. 2015. Mimicry is presidential linguistic style matching in presidential debates and improved polling numbers. *Pers. Soc. Psychol. Bull.* 41(10):1311–19
Roscigno VJ, Hodson R. 2004. The organizational and social foundations of worker resistance. *Am. Sociol. Rev.* 69(1):14–39
Rule A, Cointet J-P, Bearman PS. 2015. Lexical shifts, substantive changes, and continuity in State of the Union discourse, 1790–2014. *PNAS* 112:10837–44
Rzhetsky A, Foster JG, Foster IT, Evans JA. 2015. Choosing experiments to accelerate collective discovery. *PNAS* 112(47):14569–74
Saavedra S, Hagerty K, Uzzi B. 2011. Synchronicity, instant messaging, and performance among financial traders. *PNAS* 108(13):5296–301
Sacks H. 1995. In *Lectures on Conversation*, pp. 1–131. Malden, MA: Wiley-Blackwell
Santorini B. 1990. *Part-of-speech tagging guidelines for the Penn Treebank Project (3rd revision)*. Univ. Pa. Dep. Comput. Inf. Sci. Tech. Rep. No. MS-CIS-90-47
Schank RC, Colby KM. 1973. *Computer Models of Thought and Language*. San Francisco: Freeman
Schegloff EA. 1995. Introduction. See Sacks 1995, pp. ix–lxii
Sedelow SY. 1989. The interlingual thesaurus model for global technical communication: research results. *Proc. Ann. East. Mich. Univ. Conf. Lang, Commun. World Bus. Prof., 8th, Ann Arbor, Mar. 30–April 1*
Shahaf D, Guestrin C, Horvitz E. 2012. Metro maps of science. *Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, Beijing, Aug. 12–16, pp. 1122–30. New York: Assoc. Comput. Mach.
Shapin S. 1994. *A Social History of Truth: Civility and Science in Seventeenth-Century England*. Chicago: Univ. Chicago Press
Shi F, Foster J, Evans J. 2015. Weaving the fabric of science: dynamic network models of science's unfolding structure. *Social Networks* 43:73–85
Sim Y, Acree BD, Gross JH, Smith NA. 2013. Measuring ideological proportions in political speeches. See Cadilhac et al. 2013, pp. 91–101
Simmons JP, Nelson LD, Simonsohn U. 2011a. False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol. Sci.* 22:1359–66
Simmons MP, Adamic LA, Adar E. 2011b. Memes online: extracted, subtracted, injected, and recollected. *Proc. Fifth Int. AAAI Conf. Weblogs Soc. Media, Barcelona, July 17–21*
Smith A. 1759. *The Theory of Moral Sentiments*. London: Millar

Socher R, Perelygin A, Wu JY, Chuang J, Manning CD, et al. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. See Cadilhac et al. 2013, pp. 1631–42
Spangler S, Wilkins AD, Bachman BJ, Nagarajan M, Dayaram T, et al. 2014. Automated hypothesis generation based on mining scientific literature. See Mukherjee et al. 2014, pp. 1877–86
Sridhar D, Foulds J, Huang B, Getoor L, Walker M. 2015. Joint models of disagreement and stance in online debate. *Proc. 53rd Annu. Meet. Assoc. Comput. Linguist. 7th Jt. Conf. Nat. Lang. Process. Int.*, Beijing, July 26–31, pp. 116–25. **http://www.aclweb.org/anthology/P15-1012**
Srivastava AN, Sahami M. 2009. *Text Mining: Classification, Clustering, and Applications*. Boca Raton, FL: Chapman & Hall/CRC Press
Stivers T, Enfield NJ, Brown P, Englert C, Hayashi M, et al. 2009. Universals and cultural variation in turn-taking in conversation. *PNAS* 106(26):10587–92
Stone PJ, Dunphy DC, Smith MS, Oglivie DM. 1966. *The General Inquirer: A Computer Approach to Content Analysis*. Cambridge, MA: MIT Press
Stymne S, Hardmeier C, Tiedemann J, Nivre J. 2013. Feature weight optimization for discourse-level SMT. *Proc. Workshop Discourse Mach. Transl., Sofia, Bulg., Aug. 9*, pp. 60–69. **http://www.aclweb.org/anthology/W13-3308**
Sudhof M, Goméz Emilsson A, Maas AL, Potts C. 2014. Sentiment expression conditioned by affective transitions and social forces. See Mukherjee et al. 2014, pp. 1136–45
Taddy M. 2013. Multinomial inverse regression for text analysis. *J. Am. Stat. Assoc.* 108(503):755–70
Taddy M. 2015. Document classification by inversion of distributed language representations. *Proc. 53rd Annu. Meet. Assoc. Comput. Linguist. 7th Jt. Conf. Nat. Lang. Process. Int.*, Beijing, July 26–31, pp. 45–49
Tan C, Lee L, Pang B. 2014. The effect of wording on message propagation: topic- and author-controlled natural experiments on Twitter. See Iyyer et al. 2014, pp. 175–85
Tangherlini TR, Leonard P. 2013. Trawling in the sea of the great unread: sub-corpus topic modeling and humanities research. *Poetics* 41(6):725–49
Tausczik YR, Pennebaker JW. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *J. Lang. Soc. Psychol.* 29(1):24–54
Tavory I, Timmermans S. 2014. *Abductive Analysis: Theorizing Qualitative Research*. Chicago: Univ. Chicago Press
Tetlock PC. 2007. Giving content to investor sentiment: the role of media in the stock market. *J. Finance* 62(3):1139–68
Tibshirani R. 1996. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B* 58:267–88
Timmermans S, Tavory I. 2012. Theory construction in qualitative research from grounded theory to abductive analysis. *Sociol. Theory* 30(3):167–186
Tsur O, Calacci D, Lazer D. 2015. A frame of mind: using statistical models for detection of framing and agenda setting campaigns. *Proc. 53rd Annu. Meet. Assoc. Comput. Linguist. 7th Jt. Conf. Nat. Lang. Process. Int.*, Beijing, July 26–31, pp. 1629–38
Tsytsarau M, Palpanas T, Castellanos M. 2014. Dynamics of news events and social media reaction. See Mukherjee et al. 2014, pp. 901–10
van Atteveldt WH. 2008. *Semantic Network Analysis: Techniques for Extracting, Representing, and Querying Media Content*. Charleston, SC: BookSurge
van Atteveldt WH, Kleinnijenhuis J, Ruigrok N. 2008. Parsing, semantic networks, and political authority using syntactic analysis to extract semantic relations from Dutch newspaper articles. *Polit. Anal.* 16(4):428–46
Vilhena D, Foster J, Rosvall M, West J, Evans J, Bergstrom C. 2014. Finding cultural holes: how structure and culture diverge in networks of scholarly communication. *Sociol. Sci.* 1:221–38
von Neumann J, Morgenstern O. 1944. *Theory of Games and Economic Behavior*. Princeton, NJ: Princeton Univ. Press
Wallace BC, Trikalinos TA, Laws MB, Wilson IB, Charniak E. 2013. A generative joint, additive, sequential model of topics and speech acts in patient-doctor communication. See Cadilhac et al., pp. 1765–75
Wallach HM. 2006. Topic modeling: beyond bag-of-words. See Blei & Lafferty, pp. 977–84
Whissell CM. 1989. The dictionary of affect in language. In *Emotion: Theory, Research, and Experience*, ed. R Plutchik, H Kellerman, pp. 113–31. New York: Academic

Willey MM. 1926. *The Country Newspaper: A Study of Socialization and Newspaper Content*. Chapel Hill, NC: Univ. N. C. Press

Wittgenstein L. 2010. *Philosophical Investigations*. Chichester, UK: Wiley

Woodward JL. 1934. Quantitative newspaper analysis as a technique of opinion research. *Soc. Forces* 12(4):526–37

Woolley AW, Chabris CF, Pentland A, Hashmi N, Malone TW. 2010. Evidence for a collective intelligence factor in the performance of human groups. *Science* 330(6004):686–88

Yang D, Wen M, Rose C. 2015. Weakly supervised role identification in teamwork interactions. *Proc. 53rd Annu. Meet. Assoc. Comput. Linguist. 7th Int. Jt. Conf. Nat. Lang. Process.*, Beijing, July 26–31, pp. 1671–80

Yu B, Kaufmann S, Diermeier D. 2008. Classifying party affiliation from political speech. *J. Inf. Technol. Polit.* 5(1):33–48

Yu D, Tyshchuk Y, Ji H, Wallace W. 2015. Detecting deceptive groups using conversations and network analysis. *Proc. 53rd Annu. Meet. Assoc. Comput. Linguist. 7th Int. Jt. Conf. Nat. Lang. Process.*, Beijing, July 26–31, pp. 857–66

Annual Review of Phytopathology

Volume 54, 2016

Contents

Cladosporium fulvum Effectors: Weapons in the Arms Race with Tomato
Pierre J.G.M. de Wit 1

Plant Diseases and Management Approaches in Organic Farming Systems
A.H.C. van Bruggen and M.R. Finckh 25

Replication of Tobamovirus RNA
Kazuhiro Ishibashi and Masayuki Ishikawa 55

Advances and Challenges in Genomic Selection for Disease Resistance
Jesse Poland and Jessica Rutkoski 79

Rice Reoviruses in Insect Vectors
Taiyun Wei and Yi Li 99

Mechanisms Involved in Nematode Control by Endophytic Fungi
Alexander Schouten 121

Root Border Cells and Their Role in Plant Defense
Martha Hawes, Caitilyn Allen, B. Gillian Turgeon, Gilberto Curlango-Rivera, Tuan Minh Tran, David A. Huskey, and Zhongguo Xiong 143

Using Ecology, Physiology, and Genomics to Understand Host Specificity in *Xanthomonas*
Marie-Agnès Jacques, Matthieu Arlat, Alice Boulanger, Tristan Boureau, Sébastien Carrère, Sophie Cesbron, Nicolas W.G. Chen, Stéphane Cociancich, Armelle Darrasse, Nicolas Denancé, Marion Fischer-Le Saux, Lionel Gagnevin, Ralf Koebnik, Emmanuelle Lauber, Laurent D. Noël, Isabelle Pieretti, Perrine Portier, Olivier Pruvost, Adrien Rieux, Isabelle Robène, Monique Royer, Boris Szurek, Valérie Verdier, and Christian Vernière 163

Quarantine Regulations and the Impact of Modern Detection Methods
Robert R. Martin, Fiona Constable, and Ioannis E. Tzanetakis 189

Role of Alternate Hosts in Epidemiology and Pathogen Variation of Cereal Rusts
Jie Zhao, Meinan Wang, Xianming Chen, and Zhensheng Kang 207

Multiple Disease Resistance in Plants
Tyr Wiesner-Hanks and Rebecca Nelson 229

Advances in Understanding the Molecular Mechanisms of Root Lesion Nematode Host Interactions
John Fosu-Nyarko and Michael G.K. Jones 253

Evolution and Adaptation of Wild Emmer Wheat Populations to Biotic and Abiotic Stresses
Lin Huang, Dina Raats, Hanan Sela, Valentina Klymiuk, Gabriel Lidzbarsky, Lihua Feng, Tamar Krugman, and Tzion Fahima 279

Disease Impact on Wheat Yield Potential and Prospects of Genetic Control
Ravi P. Singh, Pawan K. Singh, Jessica Rutkoski, David P. Hodson, Xinyao He, Lise N. Jørgensen, Mogens S. Hovmøller, and Julio Huerta-Espino 303

Population Genomics of Fungal and Oomycete Pathogens
Niklaus J. Grünwald, Bruce A. McDonald, and Michael G. Milgroom 323

Resistance to Tospoviruses in Vegetable Crops: Epidemiological and Molecular Aspects
Massimo Turina, Richard Kormelink, and Renato O. Resende 347

Fungal and Oomycete Diseases of Tropical Tree Fruit Crops
André Drenth and David I. Guest 373

A Multiscale Approach to Plant Disease Using the Metacommunity Concept
Elizabeth T. Borer, Anna-Liisa Laine, and Eric W. Seabloom 397

Plant-Pathogen Effectors: Cellular Probes Interfering with Plant Defenses in Spatial and Temporal Manners
Tania Y. Toruño, Ioannis Stergiopoulos, and Gitta Coaker 419

Molecular Soybean-Pathogen Interactions
Steven A. Whitham, Mingsheng Qi, Roger W. Innes, Wenbo Ma, Valéria Lopes-Caitar, and Tarek Hewezi 443

Developments in Plant Negative-Strand RNA Virus Reverse Genetics
Andrew O. Jackson and Zhenghe Li 469

Plant-Mediated Systemic Interactions Between Pathogens, Parasitic Nematodes, and Herbivores Above- and Belowground
Arjen Biere and Aska Goverse 499

Phytophthora infestans: New Tools (and Old Ones) Lead to New Understanding and Precision Management
William E. Fry 529

The Evolutionary Ecology of Plant Disease: A Phylogenetic Perspective
Gregory S. Gilbert and Ingrid M. Parker 549

DNA Methylation and Demethylation in Plant Immunity
A. Deleris, T. Halter, and L. Navarro 579

Errata

An online log of corrections to *Annual Review of Phytopathology* articles may be found at http://www.annualreviews.org/errata/phyto