Special Issue Article

Sociological Methods & Research
2022, Vol. 51(4) 1469–1483

DOI: 10.1177/00491241221123088
journals.sagepub.com/home/smr
SAGE

From Ends to Means: The Promise of Computational Text Analysis for Theoretically Driven Sociological Research

Bart Bonikowski[1,*] **and Laura K. Nelson**[2,*]

Abstract

As the field of computational text analysis within the social sciences is maturing, computational methods are no longer seen as ends in themselves, but rather as means toward answering theoretically motivated research questions. The objective of this special issue is to showcase such research: the use of novel computational methods in the service of advancing substantive scientific knowledge. In presenting the contributions to the issue, we discuss several insights that emerge from this work, which hold relevance not only for current and aspiring practitioners of computational text analysis, but also for its skeptics. These concern the central role of theory in designing and executing computational research, the selection of appropriate techniques from a rapidly growing methodological toolkit, the benefits—and risks—of methodological bricolage, and the necessity of validating all aspects of the research process. The result is a set of broad considerations concerning the effective application of computational methods to substantive questions, illustrated by eight exemplary empirical studies.

[1] Department of Sociology, New York University, New York, NY, USA
[2] Department of Sociology, University of British Columbia, Vancouver, Canada

*Both authors contributed equally to this article.

Corresponding Author:
Bart Bonikowski, New York University.
Email: bonikowski@nyu.edu

Keywords

computational text analysis, computational social science, applied methods, research design, logic of inquiry

Computational text analysis is in its heyday. The availability of unprecedented volumes of digitized content, the development of sophisticated methods for extracting meaning from such data, and the rapid upgrading of the relevant technical expertise among practitioners—all fueled by the cross-pollination of ideas between computer science, engineering, linguistics, and the social sciences, as well as between academia and industry—have brought this mode of empirical inquiry into the scholarly mainstream. Much early scholarship yielded by these advances prioritized methodological interventions, in which illustrating the power of new analytical tools took precedence over their application to important theoretical problems. As computational social science has matured, however, so has the orientation of the research. Using a large corpus or an exciting method as an end in itself no longer carries the same cachet as it did several years ago; instead, a growing number of empirical studies are employing computational text analysis to address theoretical puzzles of central disciplinary relevance.

We conceived of this special issue of *Sociological Methods & Research* to showcase such innovative, theory-driven research. In all of the featured articles, the authors' methodological contributions were designed in the service of theoretically and substantively relevant research questions spanning culture, the economy, politics, organizations, social movements, gender, health, immigration, and inequality. Scholars interested in these topics should find the articles relevant on substantive and not just methodological grounds. To the degree that some readers remain skeptical of computational methods, they should be encouraged by the studies' rich findings.

In the course of demonstrating new ways of using text analysis to investigate vexing sociological problems, the featured articles point to broader insights about the particular advantages of computational text analysis—and possible ways to avoid some of their pitfalls. In what follows we focus on four such insights: how to incorporate theory into text analysis and use text analysis to build theory, how to select appropriate methods from the growing range of available techniques, how to fruitfully combine different methodological approaches, and how to validate the findings of computational studies.

Theory In, Theory Out

Given the emphasis on methodological proofs-of-concept in early computational text analysis studies, casual observers could have understandably concluded that this emerging tradition lends itself to atheoretical research that prioritizes technical mastery over substance. As this field has developed, however, it has become increasingly clear that access to greater volumes of data—and more complex data, such as text (as well as maps, images, and audio)—has necessitated more theory, not less. This is so because theory inevitably enters the computational research process at multiple steps.

All of the articles in this special issue lead with theoretically motivated questions and then select the appropriate analytical tools for answering them. In so doing, they convincingly demonstrate the utility of computational methods for core sociological concerns, including what machine learning can teach us about the development of human cognition (Areseniev-Koehler and Foster), how migration alters shared conceptions of social inclusion among ethnoracial majorities (Voyer, Kline, Danton, and Volkova), how social attributes influence the direction of agency in social relations (Stuhler), how and when collective actors become imbued with agency in public discourse (Knight), how symbolic boundaries within organizations become reconfigured in the aftermath of corporate mergers (Bhatt, Goldberg, and Srivastava), how radical-right candidates make use of discursive frames previously legitimated in mainstream politics (Bonikowski, Luo, and Stuhler), what knowledge is excluded from the history of women's movements (Nelson, Getman, and Haque), and how different modes of quantification and marketization transform the production of scientific knowledge (Pardo-Guerra and Pahwa).

In computational research, perhaps more than in other traditions, theory is needed not only to decide what data are appropriate to explore a given research question, but also to reflect on what the data themselves represent, not least because computational analyses typically use data that are found, not collected by the researcher (McFarland and McFarland 2015). This can be an advantage: the data generation process is typically non-reactive and takes place outside of a controlled research setting (Salganik 2019). At the same time, however, "found" textual data are often limited by platform architecture (as in social media posts), substantively relevant missingness (as in historical archives), absence of extra-textual context (as in political texts generated in the course of public performances), and non-representativeness.

With these constraints in mind, several authors in this special issue explicitly theorize their data as a central step in their analysis. In their exposition of

the extended computational case method, for instance, Pardo-Guerra and Pahwa explain how and why theory is required when constructing and interpreting a corpus: "patterns observed within a collection of texts acquire meaning not only through an internal analysis and coding of said textual elements but rather through their contraposition to other narratives available to the researcher that speak to a 'well identified' theoretical concern." They argue that a corpus should not be seen as a bounded (if incomplete) universe of inquiry but rather as a product of social practice that must be juxtaposed with other sources of data.

Nelson et al. make data "missingness" the substantive object of their inquiry. By comparing primary documents generated by activists and participants in the early-20th-century women's movement with the descriptions of the movement on Wikipedia, they show that the latter (representing the dominant historical record accessible to most lay readers) is prone to consequential omissions. This theoretically driven analysis reveals processes of historical knowledge production, while also calling into question the completeness of a corpus that others may treat as unproblematic.

Data preparation in computational text analysis also necessitates myriad preprocessing decisions. Although this has parallels in other traditions, like survey methods, in text analysis it often involves more than the subsetting of data on existing variables. A corpus must be actively constructed, and therefore the scholar must theorize what the data represent and how they should be manipulated to match the research question. For instance, working with newspaper data, Knight selects only those documents that focus on organizations, but to do so, she must first construct a theoretically justified keyword list that, when combined with named entity recognition, indexes references to organizations in the texts. Similarly, Arseniev-Koehler and Foster use keywords to limit their newspaper corpus to articles about health and obesity before examining their content. Bhatt et al. also incorporate ethical and privacy concerns into their text processing pipeline, as their corpus contained personal email communications. In addition to removing identifying information, they represent the linguistic style (itself a theoretical construct) of each email as a vector of feature counts and mask the original text to make it unrecoverable. These steps tend to blur the boundary between data construction, ethical and privacy concerns, and analysis, and in so doing, they highlight the importance of theory throughout the research design and execution.

Theoretically informed analytical decisions are especially relevant—and visible—in the process of operationalizing constructs central to a given research question. Indeed, the bulk of the methods used in computational

text analysis are designed and deployed with this objective in mind: to identify meaningful features in the corpus, which can then be counted and correlated with other phenomena that are either observed in the text itself or in meta-data (such as document attributes). Examples of fruitful theorization of measurement decisions in the special issue are too numerous to list here. From Stuhler's use of dependency parsing to capture actor relations (hearkening back to the work of John Mohr [1994] and Roberto Franzosi [1998]) and Voyer et al. use of concept mover's distance to identify text chunks referencing—often only implicitly—immigrant groups, to Bonikowski et al.'s discussion of the alignment between neural language models and sociological insights about the relationality and polysemy of meaning and Bhatt et al.'s use of lexicons to measure group communication styles and the symbolic boundaries they demarcate, every contribution to this special issue engages theoretically with how meanings should be measured to address a given research question. Without such theorization, computational text analysis is, at best, incomplete and at worst, inadequate.

Finally, the logic of specific text analysis methods themselves—and not only the operationalization enabled by these approaches—also needs to be theorized, not least to effectively adapt them to social scientific applications. Most computational text analysis techniques are borrowed from computer science and computational linguistics. Applying tools built in other disciplines to answer sociological questions requires care, validation, and, importantly, a grounding in theory. Such discussions are particularly notable in Stuhler's and Arseniev-Kohler and Foster's articles.

Sthuler notes that dependency parsers, a suite of methods that have a long history in linguistics, are rarely used in sociology, despite their promise for extracting semantically rich relations from text, a core concern for scholars of culture. He posits that this is due to a mismatch between the typical output of dependency parsers (e.g., the full set of syntactic relations between all features in a sentence) and the kind of information that is relevant to social science practitioners. To overcome this disjuncture, Stuhler formulates a series of translation rules that can be applied to parser output: he begins with an understanding of how sociologists theorize social relations and then extracts only those semantic relationships that are relevant for building sociological theory.

The manner in which computer science tools capture meaningful relationships in texts can become a source of theoretical insight in itself. We know, for example, that word embeddings work; they capture cultural associations in texts that reflect the associative beliefs held by the texts' authors (which in turn reflect public culture more broadly) (Kozlowski, Taddy, and Evans

2019). But why do they work? And what can we learn about cultural processes in answering this question? These are the issues raised by Arseniev-Kohler and Foster, who use theories of distributed cognition to investigate why word embeddings are able to "learn" cultural associations in text, and in turn, what this can teach us about how humans acquire cultural associations through language. In particular, they point to structural affinities between the diachronic reweighting of neural pathways in artificial neural networks and the adaptive updating of human cognitive associations over successive exposure to linguistic stimuli. They empirically examine these parallels by using word embeddings to extract schemas of body weight in the New York Times, revealing persistent cultural biases shared by the models and the texts used to train them.

The preceding examples demonstrate that computational text analysis not only makes use of theory—to pose research questions, construct and understand textual corpora, effectively measure meaning in text, and translate methods borrowed from computer science for use in social science—but it is also manifestly capable of helping practitioners build theory. By answering broadly relevant research questions, applied computational studies, such as those featured in this special issue, generate theoretical insights that promise to drive forward social scientific knowledge. By interrogating data, they shed light on the social processes that generate texts in the first place. By theorizing how meanings can be effectively measured in textual corpora, they bring research questions and methods into alignment, enabling the identification of robust, valid, and reproducible answers. And by examining the architecture of the methods themselves, they can not only translate those methods for a social science audience, but also make contributions to our understanding of fundamental cultural processes.

Multiple Methods, Multiple Guidelines

In the decade since the field-defining special issue of the journal *Poetics* that popularized topic models (Mohr and Bogdanov 2013), sociologists have gained access to a growing set of well established and thoroughly tested computational text analysis methods. This large toolkit can be intimidating to practitioners: which methods are best for solving which theoretical puzzles? Topic models, dictionaries, word embeddings, key phrase extraction, supervised machine learning, language models, and dependency parsers, among other approaches, all illuminate some aspects of text and language while obfuscating others. The studies featured here offer scholars guidance for selecting methods from the many techniques available—and as

importantly, for combining multiple methods—to address substantive questions. In so doing, they also illustrate the need to carefully validate not only the algorithmic output from each method, but also the corpus and the multiple interpretive steps necessitated by computational research.

Sifting Through the Toolkit

The ever-expanding range and complexity of available methods makes choosing the appropriate analytical approach more challenging than in other, more mature fields. Indeed, mismatches between the question and the method are a frequent pitfall of computational studies, especially among less experienced practitioners. To make sense of the motley of techniques available, we offer some guidelines here, inspired by the special issue contributions.

At the most general level, the distinction between unsupervised and supervised methods maps reasonably well onto that between, respectively, inductive and deductive research. As new techniques are developed and added to the computational toolkit, scholars can make sense of them by grounding them in their chosen logic of inquiry. For example, having found instances of "agent talk" through a combination of supervised methods, Knight then used topic models, an unsupervised method, to inductively discover the semantic contexts in which this discursive phenomenon occurred most frequently in early 20th-century newspaper articles, revealing its association with specific social, legal, and political domains. Arseniev-Koehler and Foster relied on word embeddings, an inductive method that models word associations across the corpus (based on local co-occurrence) rather than within documents, to examine cultural associations with body weight, whereas Voyer et al. used a variant of this approach to identify paragraphs associated with different immigrant groups and, in a more deductive step, to measure the association between these immigrant groups and the concepts of "normal" and "strange."

For scholars interested in measuring social relations in texts, dependency parsers are an effective alternative. Stuhler demonstrates the utility of this approach for detecting relations between actors and provides an R package for easily extracting them from textual data. Similarly, Knight uses dependency parsers to capture subject-verb relations that are indicative of attributions of agency to organizations. Both approaches are unsupervised and follow an inductive logic.

Alternatively, key phrase extraction works well when the objective is to inductively identify important or meaningful phrases in a corpus and to

then track those exact phrases across documents. Doing so may be useful, for example, in studies of cultural diffusion, or, in the case of Nelson et al., when building a historical "ground truth" against which contemporary accounts can be compared.

Supervised methods, in contrast, may be more appropriate when the researcher seeks to measure specific predefined categories, particularly when they occur infrequently in the corpus. Dictionaries, for example, group words into categories representing a phenomenon of interest. Several of the contributions to this issue employ predefined or custom dictionaries for corpus subsetting (e.g., Knight, Arseniev-Koehler and Foster) and the classification of their focal phenomena (e.g., Knight, Stuhler). In a variant of the latter approach, Bhatt et al. use the Linguistic Inquiry and Word Count (LIWC) dictionary to capture the group-level linguistic styles—e.g., "use of abstract versus concrete language, the expression of positive versus negative sentiment, and orientations toward the past, present, and future"—observed in email communication prior to and following a corporate merger in order to reveal shifts in the construction of employees' symbolic boundaries.

Supervised methods play a central role in Bonikowski et al.'s contribution as well. To measure the presence of a set of predefined frames—specifically, populism, exclusionary and declinist nationalism, and authoritarianism—in U.S. presidential campaign speeches, the authors rely on supervised classifiers based on neural language models. By hand-coding a sample of paragraphs from their corpus and using it to fine-tune a RoBERTa model that had been pretrained on 161 gigabytes of textual data, they are able to identify documents featuring the polysemic, vague, and rarely occurring frames of interest to their study.

The relationship between induction and deduction and unsupervised and supervised methods is only a rough heuristic, of course. Sometimes, studies combine inductive and deductive logics, and computational text analysis methods can accommodate this. Unsupervised methods can, for example, be used as input into supervised analysis that serves deductive ends, as when word embeddings serve as the basis for the measurement of specific cultural dimensions (Voyer et al.) or language models based on contextual embeddings are trained to recognize specific phenomena in a corpus (Bonikowski et al.).

Methodological Bricolage

The above examples point to another insight about computational text analysis: most substantive research questions cannot be effectively answered with a single method, but rather require the concatenation of multiple methods into a broader workflow. In performing this form of methodological

bricolage, computational studies, even more so than other methodological approaches, often defy the very distinction between induction and deduction. They instead engage in an iterative process of data exploration, formalization, and interpretation that more closely conforms to the abductive logic of inquiry (Tavory and Timmermans 2014). Thus, for instance, identifying a known concept in the corpus (a deductive step) may be best accomplished with a supervised method like dictionary analysis or a machine learning classifier, but exploring the themes that accompany instances of the concept in the text (a subsequent inductive step) may be most appropriately carried out via an unsupervised approach like topic modeling.[1]

The use of multiple interconnected methods to solve complex analytical puzzles is a common theme across the articles featured in this issue. Nelson et al.'s question about what is left out of the dominant historical record concerning women's movements, for instance, cannot be effectively addressed with a single analytical tool. Instead, the authors first parsed primary data from the movement using RAKE, an unsupervised keyword extraction method, to compile 32,295 domain-relevant unique phrases; they then generated a baseline of their usage in the English language by applying a similar RAKE procedure to the Brown book corpus; they finally used Elasticsearch to perform a fuzzy keyword search for the extracted phrases in Wikipedia articles relevant to women's movements (which had to be identified in the first place as well). Having obtained these results, they interpreted them by closely reading a selection of the relevant Wikipedia documents and developing a theoretical typology of factual omissions.

Similarly complex workflows are found, for instance, in contributions by Knight (who combines optical character recognition, dependency parsing, dictionary methods, named entity recognition, topic modeling, and regression), Bhatt et al. (who use dictionary methods, supervised random forest classifiers, and regression), and Stuhler (who brings together part of speech tagging, dependency parsing, hand coding, dictionary methods, and supervised naïve Bayes classifiers). Importantly, the objective of these concatenations of methods is not to simply illuminate the same data from different angles. Rather, the point is to create a cohesive workflow where one set of methods accomplishes a discrete goal in the manipulation of the data and its output is then fed forward into the next analytical step in the pipeline. This process often begins with the acquisition and preprocessing of the corpus, but then extends through corpus subsetting, the selection of the relevant units of analysis, drilling down into measurable entities in the text that correspond to a study's main concepts, and finally identifying patterns and associating them with other variables (in either a descriptive or causal mode).

In most substantive applications, methodological bricolage serves the purpose of gaining a better purchase on the phenomena of interest within the corpus itself (or possibly within multiple corpora). Pardo-Guerra and Pahwa, however, challenge the notion of a bounded corpus. In advocating for the extended computational case method, they urge scholars to not only combine multiple quantitative methods along with close reading—as many studies do—but to also bring the results generated with a specific corpus into conversation with other forms of data, such as qualitative interviews and ethnographic field work.

Whether in its traditional or extended-case form, methodological bricolage carries inherent risks that must be carefully navigated by scholars. Most notably, the overall analytical pipeline and the discrete steps it comprises can be difficult to evaluate and replicate. As James Evans writes in his review of Grimmer, Roberts, and Stewart's (2022) book *Text as Data* (featured in the special issue), "[t]urning text into data involves a cascade of interlocking path-dependent choices and so it is rare that particular choices will be right or wrong without the context of choices made before and after." The onus is therefore on the researcher to transparently communicate the full range of analytical steps to the reader so that each step can be evaluated in its broader context, as well as to make code and data readily available to the scholarly community for purposes of reproducibility. The contributions to this special issue consistently put these principles into practice: each includes a detailed summary of its computational workflow and all provide replication materials (in keeping with *SMR*'s policy).

Validating the Research Pipeline

Just as computational methods and complex textual data have necessitated more, not less, theory, the contributions to this issue demonstrate that computational methods also require extensive validation at every step of the research process: not only do the results themselves need to be validated, but so does the corpus construction process, the methods used to analyze the corpus, and the multiple interpretative decisions made along the way.

Pardo-Guerra and Pahwa point out that one essential step in the computational analytic process, and one that is often overlooked, is validating the corpus itself. How the corpus is selected, subset, and processed, they argue, matters for what claims it enables. Each of these steps needs to be carefully scrutinized and the consequences of the specific decisions for downstream analyses (especially compared to alternative specifications) need to be considered. Moreover, scholars need to pay special attention to the

provenance of the corpus and the purposes for which it was produced in the first place, as well as to structural biases that may be coded directly into it. Understanding the texts through deep reading, being aware of what is omitted from the data, and taking into account the data generation process are all essential data validation steps.

Once the corpus is constructed, the appropriateness of specific methods for the data at hand must often be validated as well. Knight, for instance, questions the ability of established text analysis methods to effectively analyze historical data, as most existing models are trained on contemporary corpora, such as social media and web content. To overcome these limitations, Knight trained her own custom classifier and used lists of company names from the historical period she studied to locate organizational actors in her historical texts.

Validation in text analysis studies often requires qualitative interpretation and, sometimes, the introduction of additional data or methods to ensure that one's conclusions are defensible. Voyer et al., for example, demonstrate the benefits of combining qualitative reading with computational methods. They provide full passages from the text to illustrate to the reader the validity of their models, a step that all researchers should consider when using complex text analysis methods. Arseniev-Koehler and Foster, in contrast, point to survey-based findings to validate the cultural associations revealed via their word embedding models.

As the above examples suggest, not only is validation crucial in computational research, but it often requires more than concern over statistical model fit. In *Text as Data*, Stewart, Grimmer, and Roberts (2022) make a distinction between structural and agnostic approaches to computational text analysis. The structural approach assumes the existence of a true underlying data-generating process, which text analysis methods are intended to model. The agnostic approach alternatively assumes that in most, or even all, cases it is impossible to use computational methods to model the true data generation process due to the complexity and high-dimensionality of textual corpora. Rather than solely optimizing the mathematical fit of the model to the data, the agnostic approach encourages researchers to choose a model that maximizes interpretability and the potential to yield substantive knowledge and generalizable theory. These tradeoffs may be moot when fit statistics and interpretability concerns align in favor of the same model, but when they do not, scholars must weigh the relative importance of model fit against the identification of patterns that are substantively meaningful and yield appropriate interpretations.[2] One way to potentially balance these competing considerations is to compare multiple imperfect models and select those that jointly

optimize *reasonable* fit with the richness of the resulting theoretical insights (cf. Bonikowski and DiMaggio, 2022).

Validation is all the more essential when combining multiple techniques in the same analysis. The concatenation of methodological steps increases the analytical distance between the raw data and the final algorithmic output, as each successive data transformation procedure moves the corpus further away from directly readable text. This is particularly true of methods that aggregate data to higher levels of analysis, such as various forms of clustering. Not only can such abstraction make it difficult for readers to judge whether a study's results are valid, but it also complicates the validation process for the analysts themselves. It is all too easy to accept highly distilled analytical results at face value without doing the painstaking work of verifying all the individual steps that generated those results. Yet, a study based on opaque and uncertain foundations is a house of cards, even if the final results look plausible. The contributions featured in this issue are notable for their careful and detailed validation of the entire analytical process and the transparency with which they guide the readers through the multiple stages of data manipulation. They do so not only by describing the steps involved, but also by linking their final results back to the raw text data (for instance, through examples of the cumulatively measured phenomenon in the original corpus). This goes a long way in increasing readers' confidence in the studies' findings.

Conclusion

The research featured in this special issue puts the insights we have discussed into practice. The studies are theoretically motivated and they build theory in the course of their inquiry. They carefully select computational research designs that are most suitable for their data and research questions and in so doing, they frequently rely on methodological bricolage—the purposeful weaving together of multiple methods into a powerful computational workflow. At the same time, they avoid the pitfalls of this mode of research by clearly articulating their analytical steps, illustrating their cumulative results with examples that link to the original data, and making their code and data publicly available. Finally, they engage in the careful validation not only of their final results, but also of corpus construction, of the methods' appropriateness for the data at hand, and of the interpretive choices made at various steps of the research process. In short, they are exemplars of the effective application of computational text analysis methods to sociologically relevant substantive and theoretical puzzles.

Our hope is that readers of this special issue will develop a greater appreciation of the utility of computational methods for social scientific inquiry, as well as of the unique capabilities—and inevitably, limitations—of this rapidly growing research tradition. For those interested in incorporating computational tools into their own work, the articles featured here will offer useful examples for how to solve concrete research problems, as well as a broader set of epistemological practices that commonly characterize the computational approach to text as data. As the field continues developing at its characteristically rapid pace, the repertoire of available methods will grow, but the principles that underlie their successful adaptation to social scientific ends and their effective use in scholarly practice are likely to remain unchanged. Chief among them is the central principle guiding this special issue and its individual contributions: a method is only as good as the theoretical insights it is able to generate for the purposes of advancing scientific knowledge.

Acknowledgments

We are grateful to Chris Winship for encouraging us to propose this special issue and to Felix Elwert and Lisa Charron for their sage guidance and patience throughout the editorial process. Special thanks are owed to the anonymous reviewers who engaged so deeply with the contributions to this issue, invariably strengthening them all. Finally, we express our deep appreciation to the authors, who not only produced a set of exemplary articles, but also handled an exacting peer review process with poise.

Declaration of Conflicting Interests

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

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

ORCID iDs

Bart Bonikowski https://orcid.org/0000-0003-0107-0644
Laura K. Nelson https://orcid.org/0000-0001-8948-300X

Notes

1. Even seemingly deductive methods can be put to abductive use, as in Bonikowski et al.'s study, in which iterative supervised machine learning (i.e., active learning)

revealed new, substantively meaningful patterns to the researchers. The ability of algorithms to not only mimic human cognition but to enhance it by revealing surprising patterns is among the underexplored promises of computational methods identified by Evans in his review of Grimmer et al. (and in his prior work, e.g., Evans and Aceves 2016).

2. Model fit metrics can also serve purposes other than model selection. In a particularly creative application, Bhatt et al. use them as substantively meaningful measures of cultural fit, where poorer performance by their classifier is indicative of greater cultural distance from the training set.

References

Bonikowski, Bart and Paul DiMaggio. 2022. "Mapping Culture with Latent Class Analysis: A Response to Eger and Hjerm." *Nations and Nationalism* 28:353-65.

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

Franzosi, Roberto. 1998. "Narrative Analysis–or Why (and How) Sociologists Should Be Interested in Narrative." *Annual review of sociology* 24:517-54.

Grimmer, Justin, Margaret E Roberts, and Brandon M Stewart. 2022. *Text As Data: A New Framework for Machine Learning and the Social Sciences*. Princeton, NJ: Princeton University Press.

Kozlowski, Austin C., Matt Taddy, and James A. Evans. 2019. "The Geometry of Culture: Analyzing the Meanings of Class Through Word Embeddings." *American Sociological Review* 84:905-49.

McFarland, Daniel A. and H. Richard McFarland. 2015. "Big Data and the Danger of Being Precisely Inaccurate." *Big Data & Society* 2:1-4.

Mohr, John W. 1994. "Soldiers, Mothers, Tramps and Others: Discourse Roles in the 1907 New York City Charity Directory." *Poetics* 22:327-57.

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

Salganik, Matthew J. 2019. *Bit by Bit: Social Research in the Digital Age*. Princeton, NJ: Princeton University Press.

Tavory, Iddo and Stefan Timmermans. 2014. *Abductive Analysis: Theorizing Qualitative Research*. Chicago, IL: University of Chicago Press.

Author Biographies

Bart Bonikowski is Associate Professor of Sociology and Politics at New York University. Relying on survey methods, computational text analysis, and experimental research, his work applies insights from cultural sociology to the study of politics in

the United States and Europe, with a particular focus on nationalism, populism, and the rise radical-right parties.

Laura K. Nelson is Assistant Professor of Sociology at the University of British Columbia. She uses computational methods to study social movements, gender, culture, and institutions. Her work has appeared in outlets such as the *American Journal of Sociology, Sociological Methods & Research, Mobilization, Poetics,* and *Gender & Society.*