Special Issue Article

Sociological Methods & Research
2022, Vol. 51(4) 1868–1885

DOI: 10.1177/00491241221123086
journals.sagepub.com/home/smr

From Text Signals to Simulations: A Review and Complement to *Text as Data* by Grimmer, Roberts & Stewart (PUP 2022)

James Evans[1]

Abstract

Text as Data represents a major advance for teaching text analysis in the social sciences, digital humanities and data science by providing an integrated framework for how to conceptualize and deploy natural language processing techniques to enrich descriptive and causal analyses of social life in and from text. Here I review achievements of the book and highlight complementary paths not taken, including discussion of recent computational techniques like transformers, which have come to dominate automated language understanding and are just beginning to find their way into the careful research designs showcased in the book. These new methods not only highlight text as a signal from society, but textual models as simulations of society, which could fuel future advances in causal inference and experimentation. *Text as Data*'s focus on textual discovery, measurement and inference points us toward this new frontier, cautioning us not to ignore, but build upon social scientific interpretation and theory.

Keywords

text analysis, machine learning, deep learning, social science methodology, content analysis, data mining, neural networks

[1] Sociology, University of Chicago, Chicago, IL, USA

Corresponding Author:
James Evans, Sociology, University of Chicago, 1126 E. 59th Street, Chicago, IL, USA; Santa Fe Institute, 1399 Hyde Park Rd. Santa Fe, NM, USA.
Email: jevans@uchicago.edu

Text as Data by Justin Grimmer, Margaret Roberts and Brendan Stewart represents a major, welcome addition to the scant and aging offerings for teaching introductory to advanced classes in Computational Social Science, Text Analysis for the Social Sciences, Content Analysis, or Data Science with Text. The book directly addresses concerns held by social scientists, humanists, and data scientists seeking to use established natural language processing techniques to enrich their descriptive and causal analyses. The book comes from three top research performers who have individually and together published leading work with these methods on topics that range across several areas of American and international politics (Feder et al. 2021; Grimmer 2009; Grimmer and Stewart 2013; King, Pan, and Roberts 2013). With these methods, they have generated and validated a number of new insights regarding the character, causes and consequences of political communication that demonstrate the power of text as data to enrich many fields of social inquiry. In this essay, I review and discuss the achievements of the book, highlight some complementary opportunities, and sketch out the world of recent computational text techniques that have emerged over the last decade to dominate computational approaches to language understanding and analysis, which are just beginning to find their way into the kind of crisp social scientific research designs showcased in *Text as Data*. It was not necessarily a poor choice to exclude them from the book as they point to opportunities that do not all neatly fit within the "text as data" frame, shifting from models that extract signals from text, to models that simulate human-like text generation. Flattening these models to produce "data" from text is possible, increasingly common, and fits within the framework that Grimmer, Roberts and Stewart sketch out, but this misses their generative potential, which I expect will take another decade or more to find their way into social scientific practice. It will require substantive demonstrations like Grimmer, Roberts and Stewart have made for topic models and associated methods that has brought them to well-deserved prominence (Grimmer 2009; Roberts, Stewart, and Tingley, n.d., 2019; Roberts et al. 2014).

In December 2021, when I learned that *Text as Data* was coming out in March, 2022, I requested a copy and was told that the Kindle edition would be available in January, so I downloaded an edition the day it came online and reorganized my graduate class on "Computational Content Analysis" overnight, which was scheduled for January through March (2022) at the University of Chicago in the Computational Social Science program. The first two-thirds of the course drew heavily on the book as a support, along with many substantive papers (some referenced by the authors) and code to run them, followed by more recent developments in

textual modeling, and methods that analyze content beyond text (e.g., images, video, audio). I was not disappointed by the book, as I detail below, and I expect to assign this book as an important part of my course's framework for years to come.

There were several things that made this book distinctive and valuable for a range of audiences who might seek to analyze text as data: (1) support for alternative audiences—social science, digital humanities, and data science through a task-focused framing on the use of text for social scientific discovery, measurement, and inference; (3) attention to interpretive judgment for qualitative validation and in support of multi-method analysis; (4) careful attention to sampling and bias correction (for inference), (5) consistent formalisms that enabled a consistent conceptualization of different methods and highlighted new conceptual connections between them, and (6) an extensive discussion of causality in textual analysis, which is unique among treatments of text as data. What was most notably absent from the book was just as valuable as what was distinctively present: (7) sustained restraint from offering hard and fast rules unbacked by the science and engineering underlying contemporary textual methods and scientific principles. Turning 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, including sampling strategy and corpus size. The authors are sensitive to these complex dependencies ("it depends" is often used); they convey appropriate ambiguity about those choices, the contexts that favor one over another, and approaches to validation and evaluation throughout. There are no quick fixes or one-size fits all, which might seem bothersome to those seeking assurances or to ignore complex dependencies, but these are the correct recommendations for the development of expertise, scientific and technical judgment.

The weakest part of the book was its lack of imagination regarding text embedding methods for discovery and measurement, its failure to consider sophisticated formal models of text generation (e.g., language and conditional language models), and the relationship between the two, which underlies modern transformer models that dominate textual computation—either as encoders for text understanding or decoders for text generation. Instead of highlighting these models, the book introduced supervised learning with antiquated but easy-to-explain naive Bayes and other non-text-specific machine learning models for classification. This choice educates the reader new to machine learning about the overarching concepts involved, but only weakly prepares them to understand and grapple with the formidable computational challenges and opportunities associated with modern text processing.

Nevertheless, the authors' approach to "model-skepticism" highlights the cumulative nature of social science and suggests a slow and careful adoption of methods as they are validated for social scientific purpose, avoiding the boom and bust cycles of AI fad and fashion. In this, the authors cannot be accused of not practicing what they preach.

In what follows, I detail the many strengths of the book, and explore alternative organizations and content that augment it, including a gloss of the last decade of neural text methods that will likely influence future versions of this book, or other books to come. Such books might consider not only text as signals from society, but textual models as simulations of society, as I suggest below, which could further fuel advances that follow from the authors' commitment to causal inference and experimentation.

Notable Strengths

Support for Alternative Audiences through a Focus on Tasks

Text as Data is written for and by social scientists and it respects the full range of activities social scientists perform. Social scientists include anthropologists and sociologists interested in characterizing the discourse of a social group or social class. They include political scientists and economists attempting to causally identify the impact of a publicized phase or characterization on subsequent action and interaction. The book intends not to discriminate between these social science audiences, and suggests research tasks relevant to each, moving from the representation of text, to discovery through text, to measurement of concepts in text, to inference of general patterns and processes from text. Grimmer, Roberts and Stewart also highlight the generative interrelationship between these tasks. They highlight the potential for repurposing unsupervised discovery methods (e.g., clustering, topic modeling) for measurement while recommending a careful approach to splitting data. This allows them to recommend discovering patterns on some, then testing those patterns on the remainder, continually revalidating models and associated measurement for social scientific advance.

The authors speak even-handedly and respectfully about these overlapping tasks, crystallizing a useful vocabulary for categorizing work in this area. This allows them to naturally demonstrate their selective relevance to other scholarly and professional communities committed to gaining insight from text as data. These include digital humanists with their historical focus on discovery and measurement within texts of cultural importance (e.g., the canon of accepted nineteenth Century French novels), without a necessary

commitment to extend insights to other texts. Their valued audiences also include data scientists focused on generating predictions, who may support domains of action benefiting from causal inference with text data. Consider analysts managing a company's social media presence or running a high-stakes political campaign. I anticipate that the authors' clear presentation of tasks will increase conversations between humanists, social scientists and data analysts who will see not only reasons to collaborate, but glosses of published social scientific work that invite admiration and may persuade the relevance of new tasks (e.g., causal analysis) for a humanistic or data scientific endeavor. Their presentation also encourages that tasks pursue open science goals of open access and reproducibility ("Source material should always be identified and ideally made public"; "the coding process should be explainable and reproducible") in order to facilitate conversation at the level of the data and models, not just framings and inferences.

Attention to Interpretive Judgment

There is so much complexity in turning text—a high-dimensional artifact of meaningful communication—into lower-dimensional data for quantitative evaluation and analysis. The authors distill this complexity into aphorisms or "principles" that artfully boil down and in some cases powerfully theorize that process. For text analysis in general these include self-descriptive gems like "text analysis does not replace humans—it augments them", "social science theories and substantive knowledge are essential for research design", "the best method depends on the task", and "validations are essential and depend on the theory and the task". In other words—the methods we will teach you are powerful, but they are simply not enough, and they neither replace a powerful, interpretive brain nor the accumulated intellectual culture embedded in social theory. Amen! Saying these in front of the mirror makes me feel good about myself—as a human and as a social scientist. Thank you, authors. This self- and theory-confidence is taken further than I would have taken it, but it suggests a consistent epistemic commitment to incremental, robust growth from a place of scientific certainty.

Grimmer, Roberts and Stewarts' principles for text representation echo the importance of beginning the text analysis journey without laying aside analyst intelligence or disciplinary education: that there is "no values-free corpus construction", but that text collections should be "question-specific". They highlight that there is "no right way to represent text", a principle they manifest with classics of political communication analysis, like the Mosteller and Wallace analysis of stylistics—analysis of the distribution of function words

(e.g., prepositions, articles) in text—for identifying the authors of the anonymous Federalist Papers (1963). This "function word" representation would not have been used to analyze the content of those documents. Depending on your purpose, the particular words themselves may be important, or the meanings simmering beneath them, or the ambiguity that their expression in-context conveys, or the underlying stances and sentiments that gave rise to them (Evans and Aceves 2016).

In addition to presenting design principles for the appropriate representation of text and selection of methods, the authors highlight the importance of expert content-coding, and carefully describe that process in a way that guides those with limited experience in either interpretive assessment or social scientific replication. They promote reasoned and interpretive skepticism throughout, arguing for the continual validation of measures and exploration of algorithmic limitations.

Careful Attention to Sampling

I loved *Text as Data*'s unique take on sampling. Rather than focus on the precise algorithm for sampling (e.g., random, systematic, convenience), the authors attend to the biases that any algorithm must avoid or explicitly counter. These include resource, incentive, medium and retrieval biases. Encoding ideas in text and preserving them over time is a sign of wealth, represents human intention, and interacts with the medium of communication (e.g., the length of a tweet) and the protocols used to retrieve analyzed text. In other words, they help us realize how nonrandom sampling of meanings has already occurred in the production, preservation and availability of texts for which no random-sampling in retrieval can compensate. The authors assert that biases must be investigated, measured and countered directly when making inferences about the existence, prevalence, or influence of meanings in text.

Consistent Formalisms

Many discussions throughout the book benefitted from a careful effort to create consistent conceptual and mathematical formalisms. These enabled a conceptualization of similarities underlying distinct models, drawing new conceptual connections between them. At a high-level, I appreciate the way in which the book summarizes text processing techniques such as lowercasing, removing punctuation and stopwords, stemming and lemmatizing, and dropping infrequent words as acts of complexity reduction. The same

might be said of part-of-speech tags and recognized named entities, introduced in the chapter on sequential representations.

The discussion and formal comparison of similarity metrics used to assess different positions of documents within a vector space model elegantly articulate how each weights the importance of different factors.

The authors' early rendering of the multinomial language model set up a smooth transition to introduce topic models, an approach that continues to perform very well on corpora of modest size, which are particularly prevalent in the social sciences. These authors are the strongest and most creative users of topic models in the social sciences. They helped to formalize the structural topic model framework for incorporating covariates, ranging from time and author to topical persistence and geographic coherence (Grimmer 2009; Roberts, Stewart, and Tingley, n.d., 2019; Roberts et al. 2014). Some of the authors also collaborated on the popular and excellent R software most frequently used to estimate such models (Roberts, Stewart, and Tingley 2019). The book's formalisms are most useful in clearly articulating the control of moving parts required for causal analysis, one of the emerging frontiers of text analysis development in both social science and the computer and information science communities.

Causal Analysis of Text

Perhaps the most unique contribution from *Text as Data* is in the penultimate section on inference, or drawing reasoned insights from observed text to that which is unobserved (e.g., written in the future, from other places or for other purposes). The authors, leaders in causal analysis with text (Feder et al. 2021), clearly distinguish between prediction, or the anticipation of future events, and causal inference that requires conditional or counterfactual prediction. Their principles of prediction are clear and elegant (e.g., "predictive features do not have to cause the outcome", "it can be difficult to apply prediction to policy-making"). The authors' attention to prediction and its practical uses demonstrate their commitment to an agnostic approach that values multiple ways of knowing. They are not causal fascists who elevate such insights above all others, while still articulating the power of causal inference for understanding and policy.

Grimmer, Roberts and Stewarts' chapter on causal analysis highlights critical principles for causal identification and articulates challenges associated with the accuracy of causal modeling assumptions, like the Stable Unit Treatment Value Assumption (SUTVA), which requires that the response of a particular unit depends on the treatment to which it was assigned, not

the treatments of those around it. They emphasize the importance of experiments, which they illustrate with some of their own pathbreaking work in this area (Egami et al. 2018; Grimmer 2015). I argue below that modern transformer methods could contribute most to this area, by enabling the automated contruction of text manifesting precise semantic variation with generative models, and the simulation of text understanding and reception with encoding models.

The authors productively walk through scenarios in which text: (1) plays the role of a caused outcome, (2) plays the role of a causal treatment, and (3) plays the role of a cause-confusing confounder, which may covary and interfere with the treatment. In many cases, of course, text plays multiple roles simultaneously. In analyzing the causal influence of social media insults on subsequent topics and tone, text might play all three roles at once—the treatment (e.g., an insult) on the outcome (e.g., subsequent tone) clouded by confounders (e.g., language surrounding the insult). Nevertheless, Grimmer, Roberts and Stewart's clear separation of these roles into distinct chapters clarifies them and enables the reader to understand and decompose a causal research design.

Potential Augmentations

Despite the great strengths of the volume and its presentation, there were necessarily "paths not taken", connections not articulated, and material not included, some for good reason. One cannot write all possible books on a subject, and our universe has gifted us with only one. Here I highlight alternative framings and additional content that may address and augment what some might view as limitations.

An Alternative Sequence of Tasks

Grimmer, Roberts and Stewarts' organization of text analysis into tasks feels natural. Without intention (one section is titled "an agnostic approach to text analysis"), however, it could be understood to reinforce a hierarchy of research activities and social scientific epistemologies or ways of knowing. The order of discovery then measurement then inference subtly suggests that adventurers discover things analysts can subsequently count, which scientists can then use to infer causal claims. This is one natural flow of analytical hand-offs, the most popular one, and probably the way I and most would have organized the book for clarity.

An alternative pathway, however, moves in the opposite direction with causally identified treatments or tightly measured patterns provoking discovery. Such discovery might focus on the identification of candidates for underlying mechanisms that drive treated impact. Discovered mechanisms could be semantic and psychological or rational and game theoretic. Discovery might identify complex configurations of meaning and intention that characterize inscrutable patterns. In short, discovery is not just about things to be counted, and it does not and should not necessarily precede, in logic or time, other text analytic tasks. Moreover, discovered, complex relationships may elude causal identification not only in practice, but in principle, as when they involve system-wide dependencies, such as multiple voices parroting a shared, but unmeasured (e.g., historical) source of information. The authors do not deny this and their emphasis of multiple, cumulative qualitative validations of measurement suggests that they would not disagree.

There were many other places where specific sub-tasks were ordered in a particular way that may unintentionally blind students to the potential for varying them, with outputs from some models profitably feeding inputs to others. For example, even though the authors' excellent discussion of similarity metrics capturing differences between documents in a vector-space model was described before presentation of distributed representations of words (i.e., word embeddings), such similarity metrics are particularly useful for analyzing the distance between words and documents based on embedding model outputs (e.g., 300-dimensional vectors). Another example involves clustering, which the authors thoroughly and thoughtfully discuss for text. Often clustering too is done atop not only word or coded features, but topic loadings from topic models or vectors from embeddings, processes which effectively pre-cluster words into meanings-in-context. The authors' use of a general formalism referencing text features suggests that any model output can serve as a feature. Nevertheless, the sequence of topics could unintentionally inhibit student imagination about how fine-grained model outputs are routinely fed as inputs for coarse-grained models to reduce their dimensionality and improve interpretation. The power of many of these methods comes from their ability to plug-and-play in recursive designs crafted to the research task.

Attention to Scale

The current volume does highlight how massive, contextual models like Bidirectional Encoder Representations from Transformers (BERT) require proportionally massive textual corpora, processing energy and time. There were other opportunities, however, that might further have drawn out the

relationship between model resolution and corpus size throughout that could be instructive. For example, reductions in complexity, like lowercasing, stemming and lemmatizing words, are necessarily more desirable when analyzing small textual corpora to increase individual word frequency and context variation for analysis. With more text, textual variation may be sufficient to analyze more subtle, high-resolution patterns, and such reductions in complexity (i.e., stemming, lemmatizing) become less desirable. As such, different methods are best suited for different scales of research. With very little text, word-coding methods may be most feasible; with more, word counting becomes available; with still more, word collocation methods (e.g., topic models) become useful; with large text data, word context methods (e.g., BERT) may produce stable and reliable outputs. Higher-resolution methods will simply not work on limited text. These trade-offs were hinted at, but not directly articulated and dimensionalized throughout.

Corpus Comparison

I appreciated *Text as Data*'s discussion of similarity metrics capturing differences between words within a modeled corpus. The book did not consider classes of distances between not only documents, words and phrases within a corpus, but between corpora (e.g., tweets in 2016 versus those in 2020). These most prominently feature the divergences and distances between probability distributions, where normalized word (or word vector) distributions are treated as draws from a distribution. These include the asymmetric Kullbach-Liebler Divergence, the symmetric Jensen-Shannon Distance, the Wasserstein Distance, the χ^2 Divergence and many others, each with distinct theoretically relevant implications for different inferences the analyst may hope to make. For example, the information-theoretic KL divergence identifies the asymmetric surprise experienced by observing one probability distribution of words (or word vectors) with respect to another (Barron et al. 2018; Murdock, Allen, and DeDeo 2017), which differs from the stable Wasserstein Distance defined by symmetric transport of probability from one part of the word (or vector) distribution in order to make them converge, or the χ^2 Divergence to ascertain whether the same or different statistical inferences could be drawn from the two word distributions.

Embeddings

Text as Data presents text embeddings early, in the representation section (the vector-space model and distributed word representations chapters), but

neglects them for the rest of the book, including the chapters on discovery and measurement. I admit that these are methods I have spent professional time and attention on, and I mention here examples of ways in which they can be used for the purposes of textual discovery and measurement. Embeddings can be reduced and reused for discovery following the approach of Arora et al. (2015, 2018) and Arseniev-Koehler et al. (2022) to form semantically coherent topics. The approach assumes that the underlying meanings of a corpus traverse the same high-dimensional space in which words are embedded such that if a random walk over that space spends disproportionate time within one restricted area, it represents a topic. Using *k*-SVD, a matrix factorization method that deploys singular value decomposition (SVD) to decompose the embedding into *k* "discourse atoms" or topics that balance R^2, which measures how well the atoms predict word vectors, and topic diversity, which measures how distinct they are from one another. Resulting topics represent a "coordinate system" of the semantic space, near-orthogonal axes capturing the essential clusters of meaning among the documents in question. With sufficiently large text, words loading on these atoms are extremely semantically coherent.

In terms of metrics, distances described in the representation section are relevant to calculating the separation not only of documents across words, but words across documents. There are many other indirect distances that have been operationalized into semantically coherent measures. One increasingly common metric derivable from word embedding spaces involves tracing the location of a word relative to a semantic dimension of interest. Embedding models can solve analogical reasoning tasks with vector algebra such that for embedding models built on many English corpora, $\overrightarrow{king} - \overrightarrow{man} + \overrightarrow{woman} \sim \overrightarrow{queen}$, and many similar ones. This architecture can be used to define dimensions of interest within a semantic space. Consider how a "gender" dimension is operationalized through the $\overrightarrow{man} - \overrightarrow{woman}$ and $\overrightarrow{king} - \overrightarrow{queen}$ vectors. Several have detailed alternative methods to construct such dimensions (Ahn 2019; An, Kwak, and Ahn 2018; Kozlowski, Taddy, and Evans 2019), typically with vector subtraction or division, followed by a projection of other concepts onto the semantic dimension by calculating its cosine distance with respect to relevant words.

There are also powerful formal connections between Latent Dirichlet Allocation (LDA) topic models and embeddings. The former are sparse to maximize interpretability, such that each document loads nontrivially on only a few topics and each topic loads on only a few words, while the latter is dense where it loads nontrivially on them all to maximize precise distances between

documents and words. Embeddings' dense representations have improved the canonical language model $p(y_1, y_2, \ldots, y_n) = \prod_{t=1}^{n} p(y_t|y < t)$ that assesses the probability of each word (y_t) as a function of those that have come before (y_t). By reducing the complexity and dimensionality of a corpus to the embedding dimension, which is far lower than the number of unique word types, embeddings have dramatically improved the accuracy of these models and the conditional language models on which they are built. Conditional language models highlight the sequence to sequence (seq2seq) architecture $p(y_1, y_2, \ldots, y_n|x) = \prod_{t=1}^{n} p(y_t|y < t, x)$ that models the probability of each word or word vector (y_t) conditional on the language (y_t) or context (x) to be translated or transformed. And this seq2seq architecture lays the foundation for modern transformers, the very deep neural network methods that have become ubiquitous in both discriminative models of textual meaning, such as BERT, but also generative, "causal" models for producing text, including Generative Pretraining Transformers (GPT-*n*—GPT, GPT-2, GPT-3). These models use a "self-attention" architecture that allows higher layers of the neural network to attend to different parts of the encoder or decoder outputs from each prior layer, discovering the relevance of the specific part of an input sequence for an ultimate encoding (e.g., BERT) used to classify meanings, or for an ultimate decoder used to select words for text generation (e.g., GPT*). For example, self-attention can link a pronoun "it" in one sentence with the noun phrase "Marxist ideology" from a prior sentence, or can generate the appropriate "it" in text to avoid tedious repetition of a long noun (phrase).

Virtually every current expert, question-answering, or chat(bot) system is based on discriminating or generating models built atop this architecture, which effectively learns the rules of the language in question from very large text, often not specific to the social scientific domain in question. These generative models could dramatically influence causally-identifying experiments to be run in future, just as encoding models could potentially identify and analyze natural experiments based on precise word and sentence semantic differences.

I mention some of these specifics to augment Grimmer, Roberts and Stewart's use of older (and easier to explain!) methods like the Naive Bayes' model to illustrate supervised learning. Their discussion of machine learning more generally stopped short of providing intuition for deep learning (multi-layer neural network) models, like the transformer, that have come to dominate text processing. For example, their formalism for adding a regularization *term* to enable model optimization while reducing overfitting may not help analysts consider the much broader range of regularization approaches for deep learning models, which are not directly optimized, but lie outside

as "hyperparameters". Such choices include the depth and breadth of a neural network, the density of connections between internal nodes, and limits to the number of training epochs or random samples of data used to train the model. While these hyperparameters can be placed inside a meta-model, and optimized with active learning, this is prohibitively expensive for very large models like transformers.

Other Modeling Opportunities

Finally, there are additional places in which models can be used to augment textual analysis. For example, *Text as Data* does a very strong job of describing the process of hand-coded meanings, managing coders, and checking reliability. There are additional opportunities to leverage modeling to estimate coder accuracy and dynamically compensate for specific coder limitations. This is especially critical in the context of crowdsourcing, where management is more difficult and expertise more varied. One may simply filter "trolls" and "imbeciles" based on sufficient disagreement with a gold standard. Nevertheless, coders and crowds are uneven in their skill such that modeling coder and code accuracy, and weighting their contributions in proportion enables an analyst to more effectively use hand-coded data (Dawid and Skene 1979; Rzhetsky, Shatkay, and Wilbur 2009).

Risk and Epistemic Preference

A final augmentation or alternative involves a shift in the epistemic standard. Grimmer, Roberts and Stewart's "method agnostic", "theory-centered", "model skeptical" approach to text analysis trusts human intuition and social scientific theory over model outputs when they disagree. While a reasoned and consistent approach, this may lead some analysts to miss learning new theoretical principles from unexpectedly successful methodologies. For example, when latent semantic analysis methods, based on factorization of word-document matrices, were largely displaced by text embedding in the early 2010s (Mikolov et al. 2013; Pennington, Socher, and Manning 2014), this held strong implications for linguistic and communication theory. The new methods were formally equivalent to the old (Levy and Goldberg 2014), but restricted information to local semantic contexts from small, surrounding word windows rather than entire documents. This single change, shifting from global to local context, resulted in a punctuated increase in accuracy on a wide range of tasks (Johnson et al. 2017; Taddy 2015; Zhou et al. 2015) and represented a theoretical surprise and contribution. In

short, an alternative version of the book might convey that social theory should not be so privileged that it cannot be informed and updated by functional performance and methodological experimentation.

The results of such experimentation may shift our perspective such that even if there is no correct representation, neither are all representations equal. The agnostic nature of methodological presentation does not fully acknowledge that some methods provide more information than others and can be used in many ways for many tasks. Consider the transformer, glossed above, which produces high-resolution context-sensitive representations of text. In such a model, a word *in context* is a high-dimensional vector; and a word out of context is a point cloud (a sample of vectors) with position and variance that are highly discriminative for many supervised and unsupervised computational and social science tasks. A lot of text is required to train these models: this resolution does not come for free, but these principles should not be interpreted as restatements of the "No free lunch" theorem that any two algorithms are equivalent when their performance is averaged across all possible problems (Wolpert and Macready 1997). Human language has so much structure that some representations are superior for many things social scientists and analysts seek to accomplish.

The authors close by restating the importance of iterative, sequential, replicative social science, model skepticism and how to avoid the "cycle of creation and destruction" in social science methodology, as new methods are hailed as overcoming challenges of those earlier before their own limitations and flaws have been revealed. *Text as Data* soberly invites us to take things slow, with the liability that new and emerging text methods may take decades before incorporation into social science methodology. As a personal preference, I hope that social scientists will take a few more risks, try (and publish) a few more failures, which will necessarily create some periodic boom/bust dynamics as methodological strengths are selectively reported before their failures are known. Innovation in methodology, as in all things, involves surfing the phase transition between order and chaos, and this book helped me discover my own modest preferences for novelty and risk as I respect the authors' calls for law and order.

Conclusion

Among the metaverse of possible books on *Text as Data* that could have been published in 2022, I was pleased that my universe produced this one. I will assign this book as a critical part of my own course on content analysis for years to come, and it has already altered and improved the coherence of

my own vocabulary and articulation for several critical choices underlying the process of turning text into data. The book is certain and uncertain in the right places, and it points, skeptically but hopefully at the neural text methods that will likely influence future versions of this book, and other books to come. Such books will likely not only consider text as signal, but textual models as simulations—text as a generative process. This will likely expand the opportunities for causal inference from experiments where flesh-and-blood human subjects are exposed to automatically generated information. But it will also enable the construction of digital doubles of humans and human institutions that will enable unprecedented simulations of human response, allowing us to engage *in silico* experiments we could neither field nor afford without them. *Text as Data*'s focus on textual discovery, measurement and inference points us toward this new frontier, and cautions us to walk towards it slow and steady, remembering our wits and respecting our social scientific theories and history. I highly recommend this book.

Declaration of Conflicting Interests

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

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

James Evans https://orcid.org/0000-0001-9838-0707

References

Ahn, Yong Yeol. 2019. "FrameAxis: Characterizing Framing Bias and Intensity with Word Embedding." ACL.

An, J., H. Kwak, and Y. Y. Ahn. 2018. "SemAxis: A Lightweight Framework to Characterize Domain-Specific Word Semantics Beyond Sentiment." arXiv Preprint arXiv:1806.05521. http://arxiv.org/abs/1806.05521.

Arora, Sanjeev, Rong Ge, Behnam Neyshabur, and Yi Zhang. 2018. "Stronger Generalization Bounds for Deep Nets via a Compression Approach." Pp. 254-63 in *Proceedings of the 35th International Conference on Machine Learning*, edited by Jennifer Dy and Andreas Krause, Vol. 80. Proceedings of Machine Learning Research. PMLR.

Arora, Sanjeev, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. 2015. "Random Walks on Context Spaces: Towards an Explanation of the Mysteries of Semantic Word Embeddings." arXiv Preprint arXiv:1502. 03520, 385-99.

Arseniev-Koehler, A., S. D. Cochran, V. M. Mays, K.-W. Chang, and J. G. Foster. 2022. "Integrating Topic Modeling and Word Embedding to Characterize Violent Deaths." *Proceedings of the National Academy of Sciences of the United States of America* 119(10):e2108801119.

Barron, Alexander T. J., Jenny Huang, Rebecca L. Spang, and Simon DeDeo. 2018. "Individuals, Institutions, and Innovation in the Debates of the French Revolution." *Proceedings of the National Academy of Sciences* 115(18):4607-12.

Dawid, A. P. and A. M. Skene. 1979. "Maximum Likelihood Estimation of Observer Error-Rates Using the EM Algorithm." *Journal of the Royal Statistical Society. Series C, Applied Statistics* 28(1):20.

Egami, Naoki, Christian J. Fong, Justin Grimmer, Margaret E. Roberts, and Brandon M. Stewart. 2018. "How to Make Causal Inferences Using Texts." arXiv [stat.ML]. arXiv. http://arxiv.org/abs/1802.02163.

Evans, James A. and Pedro Aceves. 2016. "Machine Translation: Mining Text for Social Theory." *Annual Review of Sociology* 42(1):21-50.

Feder, Amir, Katherine A. Keith, Emaad Manzoor, Reid Pryzant, Dhanya Sridhar, Zach Wood-Doughty, Jacob Eisenstein, et al. 2021. "Causal Inference in Natural Language Processing: Estimation, Prediction, Interpretation and Beyond." arXiv [cs.CL]. arXiv. http://arxiv.org/abs/2109.00725.

Grimmer, J. 2009. "A Bayesian Hierarchical Topic Model for Political Texts: Measuring Expressed Agendas in Senate Press Releases." *Political Analysis: An Annual Publication of the Methodology Section of the American Political Science Association*. https://academic.oup.com/pan/article-abstract/18/1/1/1446901.

Grimmer, Justin. 2015. "We Are All Social Scientists Now: How Big Data, Machine Learning, and Causal Inference Work Together." *PS, Political Science & Politics* 48(1):80-3.

Grimmer, J. and B. M. Stewart. 2013. "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts." *Political Analysis: An Annual Publication of the Methodology Section of the American Political Science Association*. https://academic.oup.com/pan/article-abstract/21/3/267/1579321.

Johnson, M., M. Schuster, Q. V. Le, M. Krikun, and Y. Wu. 2017. "Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation." *Transactions of the*. https://www.mitpressjournals.org/doi/abs/10.1162/tacl_a_00065.

King, Gary, Jennifer Pan, and Margaret E. Roberts. 2013. "How Censorship in China Allows Government Criticism but Silences Collective Expression." *The American Political Science Review* 107(2):326-43.

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(5):905-49.

Levy, Omer and Yoav Goldberg. 2014 "Neural Word Embedding as Implicit Matrix Factorization." Pp. 2177-85 in *Advances in Neural Information Processing Systems 27*, edited by Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger. Red Hook, NY: Curran Associates, Inc.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013 "Distributed Representations of Words and Phrases and Their Compositionality." Pp. 3111-19 in *Advances in Neural Information Processing Systems 26*, edited by C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger. Red Hook, NY: Curran Associates, Inc.

Mosteller, Frederick and David L. Wallace. 1963. "Inference in an Authorship Problem: A Comparative Study of Discrimination Methods Applied to the Authorship of the Disputed Federalist Papers." *Journal of the American Statistical Association* 58(302):275-309.

Murdock, Jaimie, Colin Allen, and Simon DeDeo. 2017. "Exploration and Exploitation of Victorian Science in Darwin's Reading Notebooks." *Cognition*. https://doi.org/10.1016/j.cognition.2016.11.012.

Pennington, Jeffrey, Richard Socher, and Christopher Manning. 2014. "Glove: Global Vectors for Word Representation." Pp. 1532-43 in Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP).

Roberts, Margaret E., Brandon M. Stewart, and Dustin Tingley. 2019. "stm: An *R* Package for Structural Topic Models." *Journal of Statistical Software*. https://doi.org/10.18637/jss.v091.i02.

Roberts, Margaret E., Brandon M. Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson, and David G. Rand. 2014. "Structural Topic Models for Open-Ended Survey Responses." *American Journal of Political Science*. https://doi.org/10.1111/ajps.12103.

Roberts, Margaret E., Brandon M. Stewart, and Dustin Tingley. n.d. "The Structural Topic Model and Applied Social Science." *On Topic Models* …. https://mimno.infosci.cornell.edu/nips2013ws/slides/stm.pdf.

Rzhetsky, Andrey, Hagit Shatkay, and W. John Wilbur. 2009. "How to Get the Most out of Your Curation Effort." *PLoS Computational Biology* 5(5):e1000391.

Taddy, Matt. 2015. "Document Classification by Inversion of Distributed Language Representations." Pp. 45-9 in Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers). Stroudsburg, PA, USA: Association for Computational Linguistics.

Wolpert, D. H. and W. G. Macready. 1997. "No Free Lunch Theorems for Optimization." *IEEE Transactions on Evolutionary Computation* 1(1):67-82.

Zhou, Guangyou, Tingting He, Jun Zhao, and Po Hu. 2015. "Learning Continuous Word Embedding with Metadata for Question Retrieval in Community Question Answering." Pp. 250-59 in Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Vol. 1. aclweb.org.

Author Biography

James Evans is Max Palevsky Professor of Sociology, Founding Faculty Director of Computational Social Science, and Director of Knowledge Lab at the University of Chicago and the Santa Fe Institute. His research uses (and fuses) large-scale text, network, image and other data, machine learning, generative models and interactive crowd-sourcing and experiments to understand how collectives think and what they know, with a special focus on emergence and innovation in science, technology and society.