David Bamman — Research Statement

The impact of computer science on disciplines outside of itself has grown radically over the past decade, giving rise to such hybrid communities of practice as computational social science [Lazer et al. 2009], computational journalism [Cohen et al. 2011] and the digital humanities [Schreibman et al. 2008]. The research questions asked in this interdisciplinary space are fundamentally different from the classical core computer science areas of programming language theory, systems and theory of computation. They are more human, and they are messier. They involve people, with uncertain judgments, and whose answers are often not well defined or easily verified. In this human space, the empirical and often unambiguous measures of validity we are accustomed to in computer science (e.g., asymptotic or observed runtime) are rare; when present, they are often only one kind of evidence to be marshaled in support of a larger argument.

My own work explores this complex, human space from the perspective of **statistical machine learning** and **natural language processing**. I collaborate with colleagues in the social sciences and humanities, where my work often involves translating theoretical assumptions from their domain into computational models, allowing us to empirically reason about them with data. In computational social science, I published the first paper detecting censorship in Sina Weibo (the Chinese equivalent of Twitter) [Bamman et al. 2012] and explore more generally the covariates of linguistic variation in social media and web text [Bamman et al. 2014a,b; Bamman and Smith 2014]. In the computational humanities, I develop statistical models that enrich our representations of people as they are depicted in text, including learning character types that recur throughout movies and literary novels [Bamman et al. 2013b, 2014c] and inferring social hierarchies in Bronze Age civilizations [Bamman et al. 2013a]. At Carnegie Mellon, Amazon, and the Perseus Project (one of the flagships of the digital humanities), I have built NLP technologies for literary texts and datasets for humanities research. Throughout all of my work, one commonality is the use of statistical inference to **characterize human social behavior**; it is here, in this interdisciplinary space, that I believe the next grand challenges in computer science await us.

1 | MACHINE LEARNING FOR COMPUTATIONAL SOCIAL SCIENCE

1.1 Censorship in social media

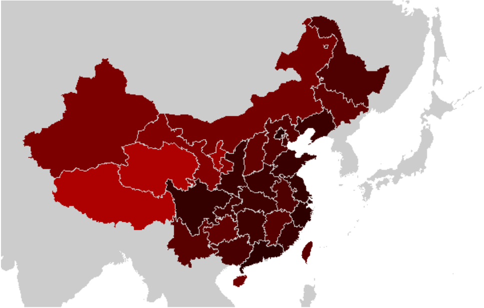

Figure 1: Censorship rate on Sina Weibo by geographical region. Brighter red signals higher rates of deletion.

Much of my work in computational social science is characterized by the use of statistical methods for large-scale data analysis. One of the places this analysis has had the most immediate impact is in uncovering evidence of censorship in Chinese social media (*First Monday* [Bamman et al. 2012]). While much work has looked at efforts to prevent access to information in China (including IP blocking of foreign websites or search engine filtering), I conducted the first large-scale analysis of political *content* censorship — i.e., the active deletion of messages published by individuals. My collaborators and I downloaded 56 million messages from the domestic Chinese microblog site Sina Weibo over a three month period and later measured which of a subset of those messages had been deleted (more than 16%). In a statistical analysis, we uncovered a set a politically sensitive terms whose presence in a message leads to anomalously higher rates of deletion (such as Falun Gong, foreign news media, and political activists like Ai Weiwei). We also note that the rate of message deletion is not uniform throughout the country, with messages originating in the outlying provinces of Tibet and Qinghai exhibiting much higher deletion rates than those from eastern areas like Beijing (see figure 1).

1.2 Computational models of linguistic variation

The sociolinguistic study of linguistic variation explores how people and groups vary in their use of language, from pronunciation to word choice. The rise of social media, with its large volume of text paired with information about its social context, has enabled the computational modeling of this kind of variation, uncovering words and topics that are characteristic of geographical regions [Eisenstein et al. 2010], learning correlations between words and socioeconomic variables [Rao et al. 2010; Eisenstein et al. 2011]; and charting how new terms spread geographically [Eisenstein et al. 2012]. These models can tell us that *hella* was (at one time) used most often by a particular demographic group in northern California, echoing earlier linguistic studies [Bucholtz 2006], and that *wicked* is used most often in New England [Ravindranath 2011]; and they have practical applications, facilitating tasks like text-based geolocation [Wing and Baldridge 2011]. My work explores how the lexical *meaning* of words is shaped by geographical influences (ACL [Bamman et al. 2014a]). In this work, I built a novel extension of a log-linear language model [Mikolov et al. 2013] that can incorporate extra-linguistic information (such as geography or time), and estimated its parameters on 1.1 billion words from Twitter. Figure 2 illustrates a sample of what this model learns: while *wicked* is used in places like Kansas to mean *bad* or *evil* ("he is a wicked man"), we learn that in New England it is used as an adverbial intensifier ("my boy's wicked smart").

Kansas		Massachusetts	
term	cosine	term	cosine
wicked	1.000	wicked	1.000
evil	0.884	super	0.855
pure	0.841	ridiculously	0.851
gods	0.841	insanely	0.820
mystery	0.830	extremely	0.793
spirit	0.830	goddamn	0.781
king	0.828	surprisingly	0.774
above	0.825	kinda	0.772
righteous	0.823	#sarcasm	0.772
magic	0.822	sooooooo	0.770

Figure 2: Terms with the highest cosine similarity to *wicked* in Kansas and Massachusetts, as learned in our model.

One practical application of computational models of linguistic variation is in inferring latent attributes of speakers (such as gender, age, and political orientation). While these efforts rest on the empirical truth that word frequencies (such as *obamacare*) often do meaningfully correlate with such demographic covariates as political affiliation, my work has advocated for a more nuanced take on this prediction task, highlighting the assumptions implicit in the case of binary gender prediction and placing limits of the scope of its interpretation (*Journal of Sociolinguistics* [Bamman et al. 2014b]). While this work uncovered a multiplicity of gendered styles on Twitter, my work on developing a statistical model to learn latent event classes from Wikipedia biographies (*Transactions of the ACL* [Bamman and Smith 2014]) uncovers variation in the *characterization* of men and women as well; though it is known that women are greatly underrepresented on Wikipedia—not only as editors [Wikipedia 2011] but also as subjects of articles [Reagle and Rhue 2011]—I find that there is a bias in their portrayal as well, with biographies of women containing up to four times as much emphasis on marriage and divorce as those of men.

Throughout all of this work, my aim has been more accurate insight into social data by better incorporating years of theoretical work from linguistics, sociology and other domains — revealing demographic variation in language use, the limits of how we can meaningfully talk about that variation, and the biases that lurk in our cultural production.

2 | MACHINE LEARNING FOR THE COMPUTATIONAL HUMANITIES

In the computational humanities, my work focuses on three aspects: 1.) methodologically, increasing the sophistication of our representation of *people* in text analysis; 2.) building concrete tools for humanists to apply computational analysis in their own work; and 3.) developing annotated corpora to occasion further research along both of those fronts. A fourth aspect of my work — **teaching** complex text analysis methods for the computational humanities — is treated in more detail in my teaching statement.

2.1 Statistical models for categories of people

In machine learning, my work has focused on developing new Bayesian models for learning categories of people (or *personas*) from their representations in text. Personal categories (such as "the Villain," "the Hardboiled Detective" and "the Surfer Dude") provide one means by which contemporary audiences and readers organize their perceptions of characters in fiction. My work attempts to learn such categories by casting the problem as one of knowledge discovery, in which we seek to uncover patterns of identity and behavior in text that are similar to those that human readers construct. My work has explored this question by inferring personas in two domains: a collection of 42,306 movie plot summaries from Wikipedia (ACL [Bamman et al. 2013b]); and a collection of 15,099 18th- and 19th-century English novels (ACL [Bamman et al. 2014c]). Both works leverage the machinery of probabilistic graphical models to learn latent entity classes from different forms of textual and extra-linguistic evidence (such as the actions those character types perform on others and the actions they have done to them, operationalized as typed syntactic dependency paths). Graphical models provide a powerful computational framework; by clearly delineating the exact relationships between all of the variables we consider (which include both observed data and presumed hidden structure), we clearly articulate our assumptions and have access to a wide range of established inference techniques, including variational methods [Jordan et al. 1999] and MCMC techniques like Gibbs sampling [Geman and Geman 1984; Casella and George 1992; Griffiths and Steyvers 2004]. The methodology I develop for fiction has general application to knowledge bases and machine learning in providing methods for **category inference**.

I've also developed Bayesian models to infer the social rank of merchants in an Old Assyrian trade network from names mentioned in a collection of 2,094 letters from a Bronze Age trade colony located in present-day Kültepe, Turkey (DH [Bamman et al. 2013b]). This work, like that described above, casts this specific case study as an instance of a general problem to be solved by statistical inference; after designing a probabilistic graphical model in collaboration with an Assyriologist to accurately encode important domain assumptions, we learned posterior estimates for the identities of names mentioned in the letters, and where they rank relative to each other. This computational analysis allowed us to confirm hypotheses put forth in prior literature and offered insight into the scope and structure of this colonial society.

2.2 Building NLP for literary texts

One of the most tangible and immediate impacts that computer scientists can have on these burgeoning interdisciplinary spaces is in developing and publicly releasing **tools** for others, especially those without computer science backgrounds, to use. To this end, I have developed an open-source natural language pipeline that scales well to book length documents (book-nlp [Bamman 2014]), building on existing tools for part-of-speech tagging and syntactic parsing, including novel components for character identification and book-length coreference resolution, and providing a common framework into which other components can easily be added. While part of the value of this work is solving engineering challenges at scale, the kind of processing I envision for literary texts does not stop simply at existing NLP technologies; there are opportunities for exciting fundamental research here as well. For literary works that contain characters and narrative, my planned work includes high-level research into extracting dominant plot points and other major events in a book (building on [Bamman and Smith 2014]) and inferring relation types between characters (building on [Bamman et al. 2013b, 2014c]). In addition to being valuable as a pre-processing step for more complex, humanistic analysis, NLP for literary books has implicit value for enhancing a reader's experience of books on digital platforms like Amazon's Kindle, iTunes Books or Oyster.

2.3 Building humanities datasets

In addition to new tools, the ongoing creation of new, in-domain **datasets** is crucial for the intellectual life of the computational humanities. Many of my published papers are accompanied by datasets for others to explore. As a senior researcher at the Perseus Project of Tufts University, I oversaw the development of major annotation projects with the Ancient Greek and Latin Dependency Treebanks (Treebanks and Linguistic Theories [Bamman and Crane 2006; Bamman et al. 2009]), developing a community standard for annotation and managing a large research group of annotators. This resource has occasioned the creation of computational tools, including part-of-speech taggers, morphological analyzers and syntactic parsers. The creation of annotated, in-domain data is one of the main boundary points where computer scientists and humanists can productively interact — the more data that experts create, the better our machine learning and NLP algorithms can be.

3 | RESEARCH AGENDA

I develop computational models to characterize human social behavior — to uncover the patterns and latent categories by which we implicitly structure the world and how we describe it in text. This broad endeavor naturally crosses many disciplinary boundaries; I'm fortunate to have established strong connections and published with co-authors whose home departments include English, Linguistics, Classics, and Near Eastern Studies; my future work relies on cultivating these bonds and developing more.

In the broadest sense, my research agenda involves winnowing out the research problems that benefit us both – that contribute a methodological advance in computer science while addressing a substantive problem in the discipline. I see two specific research areas for this, one conducive to short-term gain with immediate payoff, and one long-term vision to set the stage for years of work.

In the short term, one of the biggest barriers I see to pushing computational work further within the humanities and social sciences is the lack of good in-domain models and data in broad use. Contemporary NLP is founded on English-language newswire, and many trained models are of use only within that domain. At the same time, computationally-minded researchers in English and other disciplines have strong motivation to produce the often labor-intensive data required to improve these techniques within their own canon of works. Building on techniques like active learning, domain adaptation, and non-expert annotation (including crowdsourcing) how can we optimize the small, targeted contributions of domain experts to maximally improve the performance of our algorithms in their domains?

In the long term, I am convinced that focusing on representations of **people** in text analysis can provide a structure to organize advances for a range of disciplines. People intersect with text in multiple ways: they are its authors, its audience, and often the subjects of its content. While much current work in NLP (including named entity recognition and resolution, knowledge base inference and latent attribute prediction) approaches each of these aspects individually, I believe that developing computational models that more accurately capture the complexity of their interaction will yield deeper, socio-culturally relevant descriptions of these actors, and these deeper representations will open the door to new NLP and machine learning applications that have a more useful understanding of the world. Two concrete applications of this vision are **second-order models of information extraction** (not simply learning that Barack Obama is a member of the category Politician, but that Bill O'Reilly asserts that Obama is a Socialist); and **stereotype detection** (by which we can lean how different individuals characterize social *groups* in different ways). Both of these applications, and many more, stand at the intersection of machine learning, natural language processing, and a more nuanced view of the complexity of human social interaction that the humanities and social sciences give us. This is the point to which I see computer science advancing in the next decades; this is the point on which my research vision is fixed.

References

David Bamman. book-nlp. http://github.com/dbamman/book-nlp, 2014.

David Bamman and Gregory Crane. The design and use of a Latin dependency treebank. In *Proceedings of the Fifth Workshop on Treebanks and Linguistic Theories (TLT2006)*, pages 67–78, Prague, 2006. ÚFAL MFF UK.

David Bamman and Noah A. Smith. Unsupervised discovery of biographical structure from text. Transactions of the ACL, 2014.

David Bamman, Francesco Mambrini, and Gregory Crane. An ownership model of annotation: The Ancient Greek Dependency Treebank. In *The Eighth International Workshop on Treebanks and Linguistic Theories (TLT 8)*, 2009.

David Bamman, Brendan O'Connor, and Noah A. Smith. Censorship and deletion practices in Chinese social media. *First Monday*, 17(3), March 2012.

David Bamman, Adam Anderson, and Noah A. Smith. Inferring social rank in an Old Assyrian trade network. *Digital Humanities*, 2013a.

David Bamman, Brendan O'Connor, and Noah A. Smith. Learning latent personas of film characters. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 352–361, Sofia, Bulgaria, August 2013b. Association for Computational Linguistics.

David Bamman, Chris Dyer, and Noah A. Smith. Distributed representations of geographically situated language. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 828–834, Baltimore, Maryland, June 2014a. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P14-2134.

David Bamman, Jacob Eisenstein, and Tyler Schnoebelen. Gender identity and lexical variation in social media. *Journal of Sociolinguistics*, 18(2), 2014b.

David Bamman, Ted Underwood, and Noah A. Smith. A Bayesian mixed effects model of literary character. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 370–379, Baltimore, Maryland, June 2014c. Association for Computational Linguistics.

Mary Bucholtz. Word up: Social meanings of slang in California youth culture. In Jane Goodman and Leila Monaghan, editors, *A Cultural Approach to Interpersonal Communication: Essential Readings*, Malden, MA, 2006. Blackwell.

George Casella and Edward I. George. Explaining the Gibbs sampler. *The American Statistician*, 46(3):167–174, 1992.

Sarah Cohen, James T. Hamilton, and Fred Turner. Computational journalism. *Commun. ACM*, 54(10):66–71, October 2011. ISSN 0001-0782. doi: 10.1145/2001269.2001288. URL http://doi.acm.org/10.1145/2001269.2001288.

Jacob Eisenstein, Brendan O'Connor, Noah A. Smith, and Eric P. Xing. A latent variable model for geographic lexical variation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 1277–1287, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics. URL http://portal.acm.org/citation.cfm?id=1870658.1870782.

Jacob Eisenstein, Noah A. Smith, and Eric P. Xing. Discovering sociolinguistic associations with structured sparsity. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 1365–1374, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics. ISBN 978-1-932432-87-9. URL http://dl.acm.org/citation.cfm?id=2002472.2002641.

Jacob Eisenstein, Brendan O'Connor, Noah A. Smith, and Eric P. Xing. Mapping the geographical diffusion of new words. *ArXiv*, abs/1210.5268, 2012.

Stuart Geman and Donald Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, (6):721–741, 1984.

Thomas L Griffiths and Mark Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, 101(Suppl 1):5228–5235, 2004.

Michael I Jordan, Zoubin Ghahramani, Tommi S Jaakkola, and Lawrence K Saul. An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233, 1999.

David Lazer, Alex Pentland, Lada Adamic, Sinan Aral, Albert-Laszlo Barabasi, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, Tony Jebara, Gary King, Michael Macy, Deb Roy, and Marshall Van Alstyne. Computational social science. *Science*, 323(5915):721–723, 2009.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *ICLR*, 2013.

Delip Rao, David Yarowsky, Abhishek Shreevats, and Manaswi Gupta. Classifying latent user attributes in Twitter. In *Proceedings of the 2nd International Workshop on Search and Mining User-generated Contents*, SMUC '10, pages 37–44, New York, NY, USA, 2010. ACM.

Maya Ravindranath. A wicked good reason to study intensifiers in New Hampshire. In *NWAV 40*, 2011.

Joseph Reagle and Lauren Rhue. Gender bias in Wikipedia and Britannica. *International Journal of Communication*, 5, 2011.

Nathan Schneider, Brendan O'Connor, Naomi Saphra, David Bamman, Manaal Faruqui, Jason Baldridge, Noah A. Smith, and Chris Dyer. A framework for (under)specifying dependency syntax without overloading annotators. In *Proceedings of the ACL Linguistic Annotation Workshop*, Sofia, Bulgaria, August 2013.

Susan Schreibman, Ray Siemens, and John Unsworth. *A Companion to Digital Humanities*. Wiley Publishing, 2008. ISBN 1405168064, 9781405168069.

Wikipedia. Wikipedia editors study: Results from the editor survey, April 2011.

Benjamin P. Wing and Jason Baldridge. Simple supervised document geolocation with geodesic grids. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 955–964, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics. ISBN 978-1-932432-87-9. URL `http://dl.acm.org/citation.cfm?id=2002472.2002593`.