David Bamman

Assistant Professor
School of Information
University of California, Berkeley
102 South Hall #4600
Berkeley, CA 94720-4600

dbamman@berkeley.edu
http://ischool.berkeley.edu/~dbamman/

Research interests

Natural Language Processing, Machine Learning, Digital Humanities, Computational Social Science

Education

2015 **Carnegie Mellon University**
PhD, School of Computer Science, Language Technologies Institute. Advisor: Noah Smith.

2006 **Boston University**
MA, Applied Linguistics

1998 **University of Wisconsin–Madison**
BA, Classics

Experience

2015– **University of California, Berkeley**
Assistant Professor, School of Information

2017– **University of California, Berkeley**
Assistant Professor (affiliated), Electrical Engineering and Computer Science (EECS)

2011–2015 **Carnegie Mellon University**
Graduate Research Assistant, School of Computer Science, Language Technologies Institute

2013 **Amazon**
Research Scientist Intern, X-ray for Kindle

2005–2011 **Perseus Project, Tufts University**
Senior Researcher

Awards

2020 National Science Foundation, "CAREER: Using Fiction to Improve Real-World Information Systems," $450,000

National Endowment for the Humanities, "Multilingual BookNLP: Building a Literary NLP Pipeline Across Languages," $324,874

2019 Hellman Fellowship (UC Berkeley), "Representation Learning for the Discovery of Musical Influence," $60,000

2018 National Science Foundation, "Building Subjective Knowledge Bases by Modeling Viewpoints," $500,000 (co-PI with Brendan O'Connor, UMass)

Amazon research award, "Natural Language Processing for Literary Texts," $80,000

Digital Humanities Collaborative Research Grant (UC Berkeley), $20,000

2017 National Endowment for the Humanities, "Text in Situ: Reasoning about Visual Information in the Computational Analysis of Books," $325,000 (co-PI with Taylor Berg-Kirkpatrick, CMU)

2016 Amazon AWS Cloud Credits for Research Grant, $12,000

NVIDIA GPU Hardware Grant, Tesla K40

Digital Humanities Collaborative Research Grant (UC Berkeley), $10,000

Digital Humanities Course Development Grant (UC Berkeley), $11,890

2015 Honorable mention, CMU School of Computer Science Dissertation Award

2014 Winner, Alan J. Perlis SCS Graduate Student Teaching Award, for designing and co-teaching the interdisciplinary *Digital Literary and Cultural Studies* at Carnegie Mellon University

2011–2014 ARCS Foundation Fellowship

2010 Winner, Best Paper Award at the ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL) for David Bamman, Alison Babeu, and Gregory Crane (2010), "Transferring structural markup across translations using multilingual alignment and projection"

Teaching

2020 Instructor, "Natural Language Processing" (INFO 159/259), UC Berkeley

2019 Instructor, "Applied Natural Language Processing" (INFO 256), UC Berkeley

Instructor, "Information Organization and Retrieval" (INFO 202), UC Berkeley

2018 Instructor, "Natural Language Processing" (INFO 159/259), UC Berkeley

Instructor, "Information Organization and Retrieval" (INFO 202), UC Berkeley

2017 Instructor, "Natural Language Processing" (INFO 159/259), UC Berkeley

Instructor, "Information Organization and Retrieval" (INFO 202), UC Berkeley

Instructor, "Deconstructing Data Science" (INFO 290), UC Berkeley

2016 Instructor, "Deconstructing Data Science" (INFO 290), UC Berkeley

Instructor, "Natural Language Processing Seminar"' (INFO 290/CS 294), UC Berkeley

Instructor, "Information Organization and Retrieval" (INFO 202), UC Berkeley

2014 Teaching assistant, "Natural Language Processing" (11-411/611), Carnegie Mellon University

2013 Instructor, "Digital Literary and Cultural Studies" (76-429/829), Carnegie Mellon University

Press

2019 Wired, "Machine learning is totally changing what we think of as literature"

2018 The Guardian, "Women better represented in Victorian novels than modern, finds study"

Smithsonian Magazine, "Women were better represented in Victorian novels than modern ones"

Economist, "Machines are getting better at literary analysis"

Washington Post, "How computational analysis is teaching us to read in new ways"

2016 Popular Mechanics, "Teaching Siri to snark"

CBC Spark, "Can an algorithm detect sarcasm better than you?"

2015 Washington Post, "Inside the surprisingly high-stakes quest to design a computer program that 'gets' sarcasm online"

2012 Boston Globe, "How Twitter language reveals your gender — or your friends"'

New Scientist, "Revealed: How China censors its social networks"

BBC, "China's social networks hit by censorship, says study"

2011 New York Times, "The jargon of the novel, computed"

Service

2020 Co-organizer, Workshop on Natural Language Processing and Computational Social Science (EMNLP)

Area chair (social media and computational social science), ACL

Area chair (social media and computational social science), EMNLP

2019 Co-organizer, Workshop on Natural Language Processing and Computational Social Science (NAACL)

Co-organizer, Workshop on Narrative Understanding (NAACL)

Area chair (social media and computational social science), EMNLP

2017 Co-organizer, Workshop on Natural Language Processing and Computational Social Science (ACL)

Area chair (social media and computational social science), EMNLP

2016 Co-organizer, Workshop on Natural Language Processing and Computational Social Science (EMNLP/WebSci)

Co-organizer, Algorithms in Culture conference (UC Berkeley)

Area chair (social media), ACL

Program committees and reviewing

NLP: *Transactions of the ACL* (2014ff.), ACL (2014ff.), NAACL (2015), EMNLP (2013ff.), EACL (2013ff.).

Computational humanities: *Journal of Cultural Analytics* (2017), Digital Humanities (2011ff.)

Machine learning: NIPS (2015), ICML (2015), *Journal of Artificial Intelligence Research* (2015).

General computer science/web: WWW (2015), *IEEE Internet Computing* (2012).

Workshops: ACL 2015 Workshop on Noisy User-generated Text (W-NUT); NAACL 2015 Workshop on Computational Linguistics for Literature (CLfL); Workshop on Language Technology for Cultural Heritage Data (LaTeCH) (2008ff.), Workshop on Treebanks and Linguistic Theories (TLT) (2009-2010; 2015).

Refereed Conference Papers

Matthew Sims and David Bamman, "Measuring Information Propagation in Literary Social Networks," EMNLP 2020.

Matthew Jörke, Jon Gillick, Matthew Sims and David Bamman, "Attending to Long-Distance Document Context for Sequence Labeling," *Findings of EMNLP* 2020.

David Bamman, Olivia Lewke and Anya Mansoor (2020), "An Annotated Dataset of Coreference in English Literature," LREC 2020.

Matthew Sims, Jong Ho Park and David Bamman (2019), "Literary Event Detection," ACL 2019.

Jon Gillick, Adam Roberts, Jesse Engel, Douglas Eck and David Bamman (2019), "Learning to Groove with Inverse Sequence Transformations," ICML 2019.

Jon Gillick, Carmine-Emanuele Cella and David Bamman (2019), "Estimating Unobserved Audio Features for Target-Based Orchestration," ISMIR 2019.

David Bamman, Sejal Popat and Sheng Shen (2019), "An Annotated Dataset of Literary Entities," NAACL 2019.

Jon Gillick and David Bamman (2018), "Please Clap: Modeling Applause in Campaign Speeches," NAACL 2018.

Kimiko Ryokai, Elena Durán López, Noura Howell, Jon Gillick, and David Bamman (2018), "Capturing, Representing, and Interacting with Laughter," ACM CHI Conference on Human Factors in Computing Systems (CHI).

Lara McConnaughey, Jennifer Dai and David Bamman (2017), "The Labeled Segmentation of Printed Books," Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP).

Yi Wu, David Bamman and Stuart Russell (2017), "Adversarial Training for Relation Extraction," Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP).

David Bamman, Michelle Carney, Jon Gillick, Cody Hennesy, and Vijitha Sridhar (2017), "Estimating the Date of First Publication in a Large-Scale Digital Library," Proceedings of the ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL).

David Bamman (2017), "Natural Language Processing for the Long Tail," Digital Humanities (DH).

Smitha Milli and David Bamman (2016), "Beyond Canonical Texts: A Computational Analysis of Fanfiction" (EMNLP).

David Bamman and Noah Smith (2015), "Open Extraction of Fine-Grained Political Statements," Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP).

David Bamman and Noah Smith (2015), "Contextualized Sarcasm Detection on Twitter," Proceedings of the 9th International Conference on the Web and Social Media (ICWSM).

David Bamman, Ted Underwood and Noah Smith (2014a), "A Bayesian Mixed Effects Model of Literary Character," Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL).

David Bamman, Chris Dyer and Noah Smith (2014b), "Distributed Representations of Geographically Situated Language," Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL).

David Bamman, Brendan O'Connor and Noah Smith (2013), "Learning Latent Personas of Film Characters," Proceedings of the Annual Meeting of the Association for Computational Linguistics

(ACL).

David Bamman, Adam Anderson, and Noah Smith (2013). "Inferring Social Rank in an Old Assyrian Trade Network," Digital Humanities.

David Bamman and Gregory Crane (2011), "Measuring Historical Word Sense Variation," Proceedings of the ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL).

David Bamman, Alison Babeu, and Gregory Crane (2010), "Transferring Structural Markup Across Translations Using Multilingual Alignment and Projection," Proceedings of the ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL). **Winner, Best Paper Award.**

Boschetti, Federico, Matteo Romanello, Alison Babeu, David Bamman, and Gregory Crane, "Improving OCR Accuracy for Classical Critical Editions," in Proceedings of the 13th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2009).

David Bamman and Gregory Crane, "Building a Dynamic Lexicon from a Digital Library," in: Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL 2008).

David Bamman, Marco Passarotti, Roberto Busa, and Gregory Crane (2008), "The Annotation Guidelines of the Latin Dependency Treebank and Index Thomisticus Treebank: The Treatment of Some Specific Syntactic Constructions in Latin," in: Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC).

Alison Babeu, David Bamman, Gregory Crane, Robert Kummer and Gabriel Weaver, "Named Entity Identification and Cyberinfrastructure," in: Proceedings of the 11th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2007), pp. 259-270.

Gregory Crane, David Bamman, Lisa Cerrato, Alison Jones, David Mimno, Adrian Packel, David Sculley and Gabriel Weaver, "Beyond Digital Incunabula: Modeling the Next Generation of Digital Libraries," in: Proceedings of the 10th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2006), pp. 341-352.

Refereed Journal Articles

Ted Underwood, David Bamman, and Sabrina Lee (2018), "The Transformation of Gender in English-Language Fiction," *Cultural Analytics*.

David Bamman and Noah Smith (2014), "Unsupervised Discovery of Biographical Structure from Text," *Transactions of the Association for Computational Linguistics (TACL)*.

David Bamman, Jacob Eisenstein and Tyler Schnoebelen (2014), "Gender Identity and Lexical Variation in Social Media," *Journal of Sociolinguistics* 18.2.

David Bamman, Brendan O'Connor, and Noah A. Smith (2012), "Censorship and Deletion Practices in Chinese Social Media," *First Monday*, 17(3).

David Bamman and David Smith (2012), "Extracting Two Thousand Years of Latin from a Million Book Library" ACM *Journal on Computing and Cultural Heritage*, 5(1).

David Bamman and Gregory Crane (2009), "Computational Linguistics and Classical Lexicography," *Digital Humanities Quarterly*, 3(1).

David Bamman, Marco Passarotti and Gregory Crane (2008), "A Case Study in Treebank Collaboration and Comparison: Accusativus cum Infinitivo and Subordination in Latin," *Prague Bulletin of Mathematical Linguistics* 90.

Gregory Crane, David Bamman, Alison Babeu, "eScience and the Humanities," *International Journal of Digital Libraries*, 7(1), 2007.

Refereed Workshop Papers

Jon Gillick and David Bamman (2019), "Breaking Speech Recognizers to Imagine Lyrics," NeurIPS Workshop on Machine Learning for Creativity and Design.

Jon Gillick and David Bamman (2018), "Telling Stories with Soundtracks: An Empirical Analysis of Music in Film," NAACL 2018 Storytelling Workshop.

David Bamman, "Interpretability in Human-Centered Data Science," CSCW Workshop on Human-Centered Data Science, 2016.

Schneider, Nathan, 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 (LAW 2013), Sofia, Bulgaria, August 2013.

Brendan O'Connor, David Bamman, and Noah A. Smith (2011), "Computational Text Analysis for Social Science: Model Assumptions and Complexity," Proceedings of the Second Workshop on Computational Social Science and the Wisdom of the Crowds (NIPS 2011).

David Bamman, Francesco Mambrini, and Gregory Crane (2009), "An Ownership Model of Annotation: The Ancient Greek Dependency Treebank," The Eighth International Workshop on Treebanks and Linguistic Theories (TLT 8).

David Bamman and Gregory Crane, "The Logic and Discovery of Textual Allusion," LREC Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2008).

David Bamman and Gregory Crane, "The Latin Dependency Treebank in a Cultural Heritage Digital Library," in: Proceedings of the 2007 ACL Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2007), pp. 33-40.

David Bamman and Gregory Crane, "The Design and Use of a Latin Dependency Treebank," in: Proceedings of the Fifth International Workshop on Treebanks and Linguistic Theories (TLT2006) (Prague, Czech Republic: 2006), pp. 67-78.

Book Chapters

David Bamman (2020), "LitBank: Born-Literary Natural Language Processing," in: Jessica Marie Johnson, David Mimno, and Lauren Tilton (eds.), *Computational Humanities*, Debates in Digital Humanities.

Gregory Crane, David Bamman, and Alison Babeu, "ePhilology: When the Books Talk to Their Readers," in: Blackwell Companion to Digital Literary Studies (Oxford: Blackwell, 2007).

Open-source software

2017 book-segmentation. Labeled segmentation for the document structure of printed books
`https://github.com/dbamman/book-segmentation`

2014 book-nlp. Natural language processing pipeline that scales to book-length documents.
`https://github.com/dbamman/book-nlp`