

orcid.org/0000-0002-1435-8535

scholar.google.com/citations?user=0Dd6IAEAAAJ

Interests

Data-centric AI: large language models, the role of training data, and evaluation.

AI for computational social science: persuasion, opinion change and media ecosystems.

Socially aware AI: graph deep learning and models for social network settings.

Education

Massachusetts Institute of Technology

Ph.D., Media Arts and Sciences (MIT Media Lab)

(expected 2025)

Cambridge, MA

- Thesis: Language Models as Opinion Models – Techniques and Applications
- Committee: Deb Roy (advisor), Jacob Andreas, John Horton

Massachusetts Institute of Technology

M.S., Media Arts and Sciences (MIT Media Lab)

2020

Cambridge, MA

- Thesis: Mapping U.S. Talk Radio: A Textual Survey at Scale
- Advisor: Deb Roy

College of William & Mary

B.S. (*summa cum laude*), Mathematics and International Relations

2011

Williamsburg, VA

- Exchange student, University of Münster, 2010

Publications

Journal Papers

1. [A Large-Scale Audit of Dataset Licensing and Attribution in AI](#)
Shayne Longpre, Robert Mahari, Anthony Chen, Naana Obeng-Marnu, Damien Sileo, **William Brannon**, Niklas Muennighoff, Nathan Khazam, Jad Kabbara, Kartik Perisetla, Xinyi Wu, Enrico Shippole, Kurt Bollacker, Tongshuang Wu, Luis Villa, Sandy Pentland, Sara Hooker.
Nature Machine Intelligence, 2024. Accepted — to appear.
2. [The Speed of News in Twitter \(X\) versus Radio](#)
William Brannon, Deb Roy.
Nature Scientific Reports, 2024.
3. [Dubbing in Practice: A Large-Scale Study of Human Localization With Insights for Automatic Dubbing](#)
William Brannon, Yogesh Virkar, Brian Thompson.
TACL, 2023.

Conference Papers

4. [AudienceView: AI-Assisted Interpretation of Audience Feedback in Journalism](#)
William Brannon, Doug Beeferman, Hang Jiang, Andrew Heyward, Deb Roy.
CSCW, 2024. Accepted — to appear.

5. *Bridging Dictionary: AI-Generated Dictionary of Partisan Language Use*
Hang Jiang, Doug Beeferman, **William Brannon**, Andrew Heyward, Deb Roy.
CSCW, 2024. Accepted — to appear.
6. *Position: Data Authenticity, Consent, & Provenance for AI Are All Broken: What Will It Take to Fix Them?*
Shayne Longpre, Robert Mahari, Naana Obeng-Marnu, **William Brannon**, Tobin South, Katy Ilonka Gero, Alex Pentland, Jad Kabbara.
ICML, 2024. **Spotlight**.
7. *RadioTalk: A Large-Scale Corpus of Talk Radio Transcripts*
Doug Beeferman, **William Brannon**, Deb Roy.
Interspeech, 2019.

Workshop Papers

8. *ConGraT: Self-Supervised Contrastive Pretraining for Joint Graph and Text Embeddings*
William Brannon, Suyash Fulay, Hang Jiang, Wonjune Kang, Brandon Roy, Jad Kabbara, Deb Roy.
TextGraphs at ACL, 2024. Accepted — to appear.
9. *The Data Provenance Project*
Shayne Longpre, Robert Mahari, Niklas Muennighoff, Anthony Chen, Kartik Perisetla, **William Brannon**, Jad Kabbara, Luis Villa, Sara Hooker.
GenLaw at ICML, 2023.

Internships

Amazon Web Services 2022
Applied Scientist Intern Santa Clara, CA

- Research on automatic dubbing of video content between languages.

Professional Experience

Democratic National Committee 2015 - 2017
Data Scientist Washington, DC

- Led DNC Data Science work on fundraising analytics, built predictive models of donation and voting behavior, and owned stakeholder relationships.

Analyst Institute 2013 - 2015
Director of Analytics Washington, DC

- Managed a team of five analysts in implementing and analyzing 100+ political field experiments and building models of treatment responsiveness.

Democratic Senatorial Campaign Committee 2013
Analytics Manager Washington, DC

- Modeling and analysis of digital and major-donor fundraising data, including coordinating an A/B testing program for email.

Presidential Inaugural Committee 2012 - 2013
Lead Digital Analyst Washington, DC

- Responsible for all digital data and reporting, including an email list of more than 1 million addresses, and A/B testing of email sends.

Obama for America 2012
Deputy Data Director Columbus, OH

- Developed and produced daily reports for senior leadership on early voting, including overseeing data management and quality control.

Obama for America 2012
Associate Analyst Chicago, IL

- Digital analytics: analysis, reporting and infrastructure.

William & Mary Global Research Institute 2011 - 2012
Project Manager Williamsburg, VA

- Managed two international surveys on the foreign policy opinions of academics and senior policy-makers, including supervising a team of research assistants.

Awards and Grants

MIT IGNITE Generative AI Entrepreneurship Competition. MIT Martin Trust Center, 2023. Finalist; top ~10% of teams. \$5,000 unrestricted award.

Generative AI Impact Award. MIT, 2023. \$70,000 research grant, awarded for the Data Provenance Initiative.

SMA Fellowship. MIT Institute for Data, Systems and Society, 2017 - 2018.

Phi Beta Kappa. College of William & Mary, 2010.

Elsa Diduk Fellowship. College of William & Mary, 2010. Support for study in Germany.

Monroe Scholar. College of William & Mary, 2008. Top 7% of students; \$3,000 research support.

Word Power Challenge Scholarship. Reader's Digest, 2007. \$25,000 undergraduate scholarship.

Conference Presentations and Invited Talks

ACL: Association for Computational Linguistics Annual Meeting (2023). *Dubbing in Practice.*

IC2S2: International Conference on Computational Social Science (2024). *Speed of News in Twitter (X) versus Radio.*

Mozilla Data Futures Lab (2024). *Data Provenance Initiative.* [\[Link\]](#)

MIT Algorithmic Alignment Group (2023). *Data Provenance Initiative.*

Selected Media Coverage

IEEE Spectrum (2023). *Public AI Training Datasets Are Rife With Licensing Errors.* [\[Link\]](#)

TechCircle (2023). *MIT, Cohere for AI, others launch platform to enhance transparency in AI data.* [\[Link\]](#)

VentureBeat (2023). *MIT, Cohere for AI, others launch platform to track and filter audited AI datasets.* [[Link](#)]

Washington Post (2023). *AI researchers uncover ethical, legal risks to using popular data sets.* [[Link](#)]

Cohere Blog (2023). *Data Provenance Explorer Launches to Tackle Data Transparency Crisis.* [[Link](#)]

PaperCup Blog (2023). *Lip sync not a dubbing priority, Amazon study finds.* [[Link](#)]

Slator (2023). *Large Amazon Study on Human Dubbing Has 'Surprising' Implications for AI Dubbing.* [[Link](#)]

Service

ICWSM: International Conference on Web and Social Media (2024). *Program Committee.*

ITIF: Instruction Following & Finetuning Workshop @ NeurIPS (2023). *Reviewer.*

MIT Center for Constructive Communication (2021). *Application Reviewer.*

MIT Center for Constructive Communication (2021). *Research & Rigor Working Group.*

International Speech Communication Association (2019 - 2020). *Student Advisory Committee.*

Monitor Journal of International Studies (2009). *Publicity Director.* [[Link](#)]

Teaching

College of William & Mary (2009). *Teaching Assistant, German 101.*

Open-Source Software

sqlscore: R utilities for database-backed scoring of generalized linear models. [[Link](#)]

twclient: A Twitter API client focused on bulk ingest of data and loading to relational databases for analysis. [[Link](#)]

Memberships

Association for Computational Linguistics

Updated July 10, 2024. ■