Stephanie Buongiorno
Website: https://stephbuon.github.io/

Education

PhD, ABD, Applied Science in Engineering, Southern Methodist University — Expected 2023
MA, English, West Virginia University — 2017
BA, English & Linguistics, University of Texas at Arlington (Summa Cum Laude) — 2014

Recent Professional Experience

Post Doctoral Research Fellow at the Human Trafficking Project — (forthcoming) 2023-25
Guildhall, Southern Methodist University

- I am the technical lead designing a machine learning system that leverages human computation gaming. This system will improve predictive named entity recognition and the detection of subtext.

Technical Lead at the Human Trafficking Project — 2022-23
Congressional Funding Bill (H.R. 2471), Southern Methodist University

- I am collaborating with researchers from four university departments and in conversation with Congress and not-for-profis.
- While offering guidance to two CS master's students, I am designing the structure of a video game narrative for players to train our system.
- I am augmenting a neo4j graph database to improve analysis of relationships (like social networks).
- I perform various leadership roles including: a) creating new documents (e.g. data sharing agreements), b) hiring, and c) training researchers.

Teaching Fellow at Foundations and Applications of Humanities Analytics — 2022-23
National Endowment of the Humanities Grant (no. HT-272418-20), Santa Fe Institute

- I plan, teach, and evaluate workshops on text mining and data analytics for humanities data.
- I foster an inclusive and intellectually rigorous classroom for researchers with diverse backgrounds, from senior faculty to early career graduate students.

Technical Lead for Global Urbanization and Housing Affordability: Poverty, Property and the City — 2018-21
National Science Foundation Grant (no. 1520103), Southern Methodist University

- I developed the first app that applies advanced natural language processing techniques to the c19 British Parliamentary debates for research. I used frameworks like gensim and spaCy.
- I pioneered an extraction method (for analyzing subject-object relationships) with higher accuracy than existing methods (re: Jo Guldi's forthcoming book, *The Dangerous Art of Text Mining*).
- I stored and processed data sets of 100-200 years of historical data in the form of tables, matrices, and n-grams for time-series analysis. I ran language models locally and on a Unix supercomputer.
- I hired, trained, and managed a team of 3-6 undergraduates from CS, public policy, and history to create a system for disambiguating speakers in 100 years of historical data.

Research Assistant for "Toward a History of the Associative-Developmental State" — 2021-22

- I hired and collaborated with an ArcGIS specialist to make historical maps of New York and its labor movements over time. I geocoded and mapped addresses using geopandas, geopy, and folium.

Research Assistant for "A History of Early and Modern China" — 2021

- I stored and processed csv data for time series analysis. I topic modeled debates from Hong Kong's legislative council to show change throughout the 20th-century.

Skills

- 6 years experience collecting, processing, and modeling large amounts of data (up to ~200 GB) using Python, R, SQL, Shell, and Git.
- 5 years of leadership experience on multidisciplinary research teams (domestic and international), giving talks to technical and non-technical audiences, and directing student work.
- 5 years of experience creating new documents for grant-funded research.

Data Sets:

- Steph Buongiorno; Robert Kalescky; Eric Godat; Omar Alexander Cerpa; Jo Guldi, 2022, "The Hansard 19th-Century British Parliamentary Debates with Improved Speaker Names: Parsed Debates, N-Gram Counts, Special Vocabulary, Collocates, and Topics", https://doi.org/10.7910/DVN/ZCYJH8, Harvard Dataverse
  (used by researchers in the U.S., Sweden, and Finland).
- Steph Buongiorno; Omar Alexander Cerpa; Wardah Alvi; Noah Meyer; Jo Guldi, 2022, "The Hansard 19th-Century British Parliamentary Debates with Improved Speaker Names: Speaker Metadata", https://doi.org/10.7910/DVN/Z3LTVV, Harvard Dataverse, FORTHCOMING 2022
- Human trafficking Neo4j data dump for machine learning FORTHCOMING 2023

Notable Open Source Software:

- **dhmeasures (R):** https://github.com/stephbuon/dhmeasures
  Optimized “white box” statistical functions for analyzing change over time. In collaboration with undergraduate RA, Ryan Schaefer.
- **usdoj (R):**
  For fetching open data from the United States Department of Justice API such as press releases, blog entries, and speeches. Optional parameters give users the ability to specify the number of results starting from the earliest or latest entries, and whether these results contain keywords.
- **posextract (Python):** https://github.com/stephbuon/posextract
  Grammatical information extraction methods designed for the analysis of historical and contemporary textual corpora. Outperforms existing methods on historical text, like Stanford’s OpenIE *(pypi package under development)*
- **posexrtractr (R):** https://github.com/stephbuon/posextractr
  Grammatical information extraction methods designed for the analysis of historical and contemporary textual corpora. *(package in progress – will be submitted to CRAN)*
- **hansard-speakers (Python):** https://github.com/stephbuon/hansard-speakers
  A data processing pipeline to disambiguate speakers in the 19th-century British Parliamentary debates.
- **democracy-lab (Python/R):** https://github.com/stephbuon/democracy-lab
  Materials used for scholarly research.

Public Apps: (Demo Apps used for NSF grant proposals–to be replaced with a high performance app with distributed back end)

- **Congress Viewer (R-SQL, JavaScript, HTML, CSS):**
  App: https://shinyviz.smu.edu/shiny/public/congress-viewer-demo/
  Code: https://github.com/stephbuon/congress-shiny
- **Hansard Viewer (R-SQL, JavaScript, HTML, CSS):**

App: https://shinyviz.smu.edu/shiny/public/hansard-shiny/
Code: https://github.com/stephbuon/hansard-shiny
These apps offer an array of data-mining and statistical tools to explore the nature and evolution of language change in a political institution. Citizens and scholars using our toolset can navigate from an overview showing change over time. They can inspect how different candidates talk about the same issue. They can also compare the context of how particular words were used in different periods. For each view, the user is offered a variety of different statistical and machine-learning tools, offering comparison of how different methods produce different insight into language use.

Manuscripts:

- "The Hansard 19th-Century British Parliamentary Debates with Improved Speaker Names: Parsed Debates, N-Gram Counts, Special Vocabulary, Collocates, and Topics" co-authored with Rob Kalescky, Eric Godat, and Jo Guldi (revise and resubmit)
- "Speaker Name Disambiguation in the Hansard Parliamentary Debates" co-authored with Omar Cerpa and Jo Guldi (revise and resubmit)
- *Text Mining for Historical Analysis* (textbook co-authored with Jo Guldi)

Dissertation

Chapter 1: "Method" - a computer science article comparing my method of grammatical triples extraction to existing methods.
Chapter 2: "Application" - I apply my triples extraction method to news articles from NPR and FOX to see how actions are attributed to different people, organizations, and events.
Chapter 3: "Pedagogy" - Includes excerpts from *Text Mining for Historical Analysis*, a textbook co-authored with Jo Guldi).

Student Mentorship

- I offered guidance to 2 Master's students in video game design.
- II have offered project-based mentorship to 25+ students under the NSF grant no. 1520103.
- I directed 3 student internships for course credit.
- I advised the thesis of 1 undergraduate in computer science for honors distinction.
- I supported Jo Guldi's "Text Mining for Historical Method" course.
- I was a teaching fellow for "Foundations and Applications of Humanities Analytics" with the Santa Fe Institute.