ALEXANDER SPANGHER

Curriculum Vitae

---

237 McKendry Drive, Menlo Park, CA
Phone: (631) 487-7777
Email: alexander.spangher@cmu.edu
Website: **https://alexander-spangher.com**

RESEARCH OBJECTIVES Developing discourse schemas for legal and journalistic texts. Programmatic modeling of law. Advancing natural language processing techniques for span and sentence modeling.

EDUCATION

B.S.	**Columbia University** Bachelor of Science in Neuroscience Bachelor of Science in Computer Science	*2010-2014*
M.S.	**Columbia University**[1] Master of Science in Data Science Master of Science in Journalism	*2014-2018*
Doctoral Student	**Carnegie Mellon University** Graduate Studies, Electrical and Computer Engineering	*2018-2019*
	University of Southern California[2] Graduate Studies, Computer Science	*2019-present*

EMPLOYMENT	**The New York Times**, Data Scientist	*2014-2018*
	Microsoft Research, Research Intern	*2018*
	Stanford University, Summer Research Assistant	*2019*
	Bloomberg LP, Research Intern	*2020-2021*

PUBLICATIONS

Academic Publications

1. **Aleander Spangher** and Jonathan May. *StateCensusLaws.org*: A Web Application for Consuming and Annotating Legal Discourse Learning. *Computation + Journalism. Association for Computational Linguistics.* 2021. In Submission. 2021. https://arxiv.org/abs/2104.10263.
2. **Alexander Spangher** and Jonathan May. *NewsEdits*: A Dataset of Revision Histories for News Articles (Technical Report: Data Processing). 2021. https://arxiv.org/abs/2104.09647.

[1] Masters degrees pursued part-time while working full-time at the *New York Times*.
[2] Moved with professor.

3. **Alexander Spangher**, Nanyun Peng, Jonathan May, Emilio Ferrara. "Enabling Low-Resource Transfer Learning across COVID-19 Corpora by Combining Event-Extraction and Co-Training." **Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020.** 2020.
4. **Alexander Spangher**, Nanyun Peng, Jonathan May, Emilio Ferrara. Don't quote me on that: Finding Mixtures of Sources in News Articles. *Computation + Journalism.* 2020 https://arxiv.org/abs/2104.09656.
5. **Alexander Spangher**, Nanyun Peng, Jonathan May, Emilio Ferrara. Modeling Newsworthiness for Lead-Generation Across Corpora. *Computation + Journalism.* 2020. *Southern California Natural Language Processing Conference.* 2019. https://arxiv.org/abs/2104.09653.
6. **Alexander Spangher**, Jonathan May, Sz-rung Shiang, Lingjia Deng Multitask Learning for Class-Imbalanced Discourse Classification. *Association for Computational Linguistics.* 2021. In Submission. https://arxiv.org/abs/2101.00389.
7. **Alexander Spangher**, Gireeja Ranade, Besamira Nushi, Adam Fourney, Eric Horvitz. Characterizing Search Engine Traffic to Internet Research Agency Web Properties. *The Web Conference 2020.* Taipei, Taiwan. 2020. https://bit.ly/3aCGlkT.
8. **Alexander Spangher**, Gireeja Ranade, Besamira Nushi, Adam Fourney, Eric Horvitz. Analysis of Strategy and Spread of Russia-sponsored Content in the US in 2017. *International Conference for Web and Social Media, AAAI. Revise and Resubmit.* 2018. https://arxiv.org/pdf/1810.10033.
9. **Alexander Spangher**, Berk Ustun. Actionable Recourse in Linear Classification. *Proceedings of the 5th Workshop on Fairness, Accountability and Transparency in Machine Learning, ICML. Accepted.* 2018. https://bit.ly/2FEj9pf.
10. **Alexander Spangher**, Berk Ustun. Actionable Recourse in Linear Classification in Practice. *Workshop on Ethical, Social and Governance Issues in AI, 2018, NIPS. Accepted* 2018.
11. Berk Ustun, **Alexander Spangher**, Yang Liu. Actionable Recourse in Linear Classification. (Expanded Version). *Conference on Fairness, Accountability and Transparency (FAT*), 2019, ACM. Accepted* 2018. https://arxiv.org/pdf/1809.06514.pdf.
12. Ryan L Boyd, **Alexander Spangher**, Adam Fourney, Besmira Nushi, Gireeja Ranade, James Pennebaker, Eric Horvitz. Characterizing the Internet Research Agency's Social Media Operations During the 2016 US Presidential Election using Linguistic Analyses. *Whitepaper* https://bit.ly/2SczIKt.

In Preparation

1. **Alexander Spangher**, Gireeja Ranade, Adam Fourney, Besamira Nushi. Falling into the Rabbit Hole: Browsing Patterns Among Fake News Users. 2019.
2. **Alexander Spangher**, Jia Zhang, Rahul Ramachandran, Manil Maskey, Patrick Gatlin, J.J. Miller, Sundar Christopher. Methodology for Building Scalable Knowledge Graphs using Pre-existing NASA Ontologies. 2019.

Newspaper Articles and Graphics (Selected)

1. **Alexander Spangher**. Building the Next New York Times Recommendation Engine. *The New York Times.* https://nyti.ms/2zpGG5g
2. **Alexander Spangher**. How Does This Article Make You Feel? Using data science to predict the emotional resonance of New York Times articles for better ad placement. *The New York Times.* https://nyti.ms/2PyHkcn.

3. **Alexander Spangher**. What the Paris attacks tell us about how foreign news gets made. *Columbia Journalism Review*. https://bit.ly/2DIV2TH
4. **Alexander Spangher**. 19 Countries, 43 States, 327 Cities: Mapping The Times's Election Coverage. *The New York Times*. https://nyti.ms/2ScnEbV
5. **Alexander Spangher**. Eye on the Prize: 100 years worth of Pulitzer Prize Winners by Race, Gender and Location. *Columbia Journalism Review*. https://bit.ly/2r2YEIT
6. **Alexander Spangher**. 3 Smart Data Journalism Techniques that can help you find stories faster. *Medium*. https://bit.ly/2DIUydH.
7. For more articles and graphics, see: alexander-spangher.com/data-vis.html

HONORS

Academic Honors

Bloomberg Data Science Fellow. Univ. Southern Calif. *2020-2023*
Annenberg Graduate Student Fellow. Univ. Southern Calif. *2019-2023*
Carnegie Mellon ECE Fellow. Carnegie Mellon University. *2018*
John Jay Scholar. Columbia University. *2010-2014*
Dean's List. Columbia University. *2011*

Funding

Bloomberg Data Science Fellowship. $240,000. *2020-2023*
Annenberg Graduate Student Fellowship. $80,000. *2019*
Carnegie Mellon Graduate Student Fellowship. $80,000. *2018*
Columbia School of Journalism Scholarship. $78,000. *2016-2017*
New York Times Tuition Scholarship $32,000. *2014-2018*
John Jay Scholar Summer Funding. $20,000 *2011-2012*
Intel STS Semi-finalist. $10,000. *2009*

EXPERIENCE

Research Experience

Bloomberg LP, New York City, NY *Research Intern* *2019-present*
Advisor: Sz-Rung Shiang, Yao Ming, Amanda Stent.
Discourse Analysis in Journalism.

- Development of multitask machine learning models for discourse analysis.
- Analysis of semi-supervised data augmentation techniques for sequence classification.

University of Southern California, Los Angeles, CA *Ph.D. Student 2019-present*
Advisor: Jonathan May, Emilio Ferrara, Nanyun Peng.
Discourse Analysis in Journalism and Law.

- Development of machine learning models for discourse-tagging sequences.
- Development of discourse schemas for characterizing news articles edits, source-inclusion and legal text.
- Collection of large-scale corpora with legal and journalistic value for computational linguistic analysis.

Stanford Univ., Palo Alto, CA *Research Assistant* *2019*
Advisor: James Hamilton.
Computational Journalism Field Survey.

- Practitioner interviews (51 interviews) to gain knowledge on the current innovations, methodologies and challenges relevant to journalists.
- Word-embeddings classification of U.S. patent databases and NSF grants.
- Visualization/tooling in Python/Flask, D3.js and Google App-Engine.

Carnegie Mellon Univ., Mountain View, CA *Ph.D. Student* *2018-2019*
Advisor: Jia Zhang.
Knowledge Graph Construction for NASA Earth Sciences.

- Text-modeling using hierarchical topic modeling to improve and model existing NASA concept-ontologies.
- Text-matching and word-modeling using custom lexical parsing rules to extract datasets, variables and methods from papers.
- Visualization/tooling in D3.js with an emphasis on interpretability and ease of capturing user feedback.

Microsoft Research, Redmond, WA *Research Intern* *Summer 2018*
Advisors: Gireeja Ranade, Adam Fourney, Besamira Nushi, Eric Horvitz.
Large-scale analysis of user-behavior changes in response to misinformation

- Data analysis merging data from Facebook, Twitter and Microsoft. Causal modeling using counterfactual analysis.
- Text-modeling using TF-IDF to track search query changes over time.
- Intensive fact-checking, informal Congressional briefing, contact with staff of Congressman Adam Schiff (Representative, D-CA 28th District).

Employment

New York Times, NYC, NY *Data Scientist* *2014-2018*
Advisors: Chris Wiggins, Jose Muanis Castro, Thompson Marzagao.
Collaborative Topic Models for Article Recommendations:

- Created an improved article recommendation-engine by building a topic model to incorporate information from article-text and user clicks. Scale to millions of users and provide recommendations in real-time.
- Modeling: Custom-designed Bayesian model that extends Latent Dirichlet Allocation, coded in C++.
- Collaboration: Dr. David Blei, Jake Hoffman, and Prem Gopalan, Columbia University. See https://arxiv.org/pdf/1311.1704.pdf.
- Deployment: MySQL, WSGI API, Luigi data pipelines.
- Extensions: multi-armed bandit and contextual bandit algorithms.

Project Feels:

- Modeled different emotions in *New York Times* article body text. The purpose was to predict tragic, happy and polarizing articles for downstream decision-making.
- Modeling: 7 different deep learning architectures were tested alongside ensemble methods and other linear methods.
- Data collection: Crowd-sourcing on Amazon Mechanical Turk, using active learning to select successive batches of articles to label.

- Deployment: Google Cloud Services (GKE, Datastore, BigQuery).

Newsroom Tools and Other:

- Used Latent Dirichlet Allocation and TF-IDF to build a related-articles feature for journalists doing research, directly into their publishing platform.
- Used simple character-level modeling and K-Means to cluster text-messages journalists received from Q&A sessions with readers to facilitate responding.
- Used custom Bayesian model to perform newsletter recommendations for users.
- Used Random Forest and simple decision trees to create powerful and interpretable models of user retention likelihood.

Open Source Code

Actionable Recourse Implementation for **IBM Fairness 360 Project**.

- Implements Mixed-Integer Program (MIP) for providing actionable recourse auditing (see publication section above.) CPLEX and Pyomo based optimizer.
- *Contributors:* Berk Ustun, **Alexander Spangher**. https://github.com/ustunb/actionable-recourse
- *Under development:* To be incorporated into IBM Fairness 360 project, an open-sourced project integrating different fairness and transparency algorithms.

Broca: A Battery of Natural Language Processing Methods for **Open-Source News Fellows**

- A pipeline system of organize a sequence of text-transformations. Automatic intermediate caching for time-saving and debugging.
- *Contributors:* Francis Tseng, **Alexander Spangher**. https://github.com/frnsys/broca.
- *Blog:* Francis Tseng. Introducing Broca. *OpenNews*. https://bit.ly/2DVdrwH.

Languages and Frameworks

Python, SQL, D3.js, Javascript, Node.js, JQuery, Java, Scala, C, C++, C#, CUDA, OpenCL. Google Cloud Services, Amazon Web Services. Spark, Databricks, Kubernetes, Drone, Docker. Sci-kit Learn, Tensorflow, Keras, PyOmo, CPLEX.

PRESENTATIONS and CONFERENCES

Speaking

1. **Aleander Spangher**, Ke Isherwood-Huang and Amber Lynn Scott. "What the Diff': The Narrative of a Coronavirus Crisis in the U.S. Navy. *USC Annenberg Research and Creative Project Symposium*. April 2021. https://www.instagram.com/p/CNsUlUflYV0/.
2. **Aleander Spangher** and Jonathan May. Census 2020: A Computational Law Approach. *Computation + Journalism*. March 2021. https://cj2021.northeastern.edu/schedule-and-program/.
3. **Aleander Spangher** and Jonathan May. News Discourse Patterns: A Roadmap for Computational Journalism. *Computation + Journalism*. March 2021. https://cj2021.northeastern.edu/schedule-and-program/.

4. **Alexander Spangher**, Nanyun Peng, Jonathan May, Emilio Ferrara. "Enabling Low-Resource Transfer Learning across COVID-19 Corpora by Combining Event-Extraction and Co-Training." Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020. https://www.aclweb.org/anthology/volumes/2020.nlpcovid19-acl/. May 2020.
5. **Aleander Spangher**, Jillian Kwong. The Effects of a 2020 Census Undercount: A Computational Law Approach. *USC Annenberg Research and Creative Project Symposium.* April 2020. https://www.instagram.com/p/B_VijsdFZwu/.
6. **Alexander Spangher**, Nanyun Peng, Jonathan May, Emilio Ferrara. Modeling Newsworthiness for Lead-Generation Across Corpora. *Southern California Natural Language Processing Conference.* 2019. https://arxiv.org/abs/2104.09653.
7. **Alexander Spangher**, Besmira Nushi, Adam Fourney, Gireeja Ranade, Eric Horvitz. Characterizing Search Engine Traffic to Internet Research Agency Web Properties. *The Web Conference 2020.* Taipei, Taiwan. 40+ attendees. March 2020.
8. **Alexander Spangher**. Project Feels and Actionable Recourse. Open Data Science Conference. San Francisco, California. 100+ attendees. October 2018.
9. Adam Grant, **Alexander Spangher**. *New York Times* Young Professionals Interview Series, New York City, NY. 100+ attendees. February 20th, 2018.
10. Nicholas Kristof, **Alexander Spangher**, Hannah Cassius. *New York Times* Young Professionals Interview Series, New York City, NY. 100+ attendees. March 18th, 2017.
11. **Alexander Spangher**. Project Feels: Deep Text Models for Predicting the Emotional Resonance of *New York Times* Articles. Open Data Science Conference. Boston, Massachusetts. 150+ attendees. April 30th-May 3rd, 2018.
12. **Alexander Spangher**, Adam Kelleher. Recommender Systems in Digital Media. DataEngConf Meetup. New York City. 175+ attendees. March 15th, 2018.
13. **Alexander Spangher**. Building the Next *New York Times* Recommendation Engine. Data Engineering Conference. New York City, NY. 100+ attendees. December 13th-15th, 2015.

Adhoc Reviewer

1. Association of Computational Linguistics. 2021.
2. The Web Conference. 2020.
3. Association of Computational Linguistics. 2020.
4. Automated Knowledge Base Construction 2019 Conference. Amherst, Massachusetts, May 20th-21st 2019.

Teaching

1. *Teaching Assistant.* Mark Core, University of Southern California. Applied Natural Language Processing. Spring 2020.
2. *Guest Lecture.* Mor Namaan, School of Information Science, Cornell Tech. May 20th, 2018.
3. *Guest Lecture.* Jonathan Stray, Graduate School of Journalism at Columbia University. February 20th, 2018.
4. *Guest Lecture.* Steven Coll, Graduate School of Journalism at Columbia University. December 4th, 2017.

5. *Guest Lecture.* Jonathan Stray, Graduate School of Journalism at Columbia University. February 14th, 2017.
6. *Guest Lecture.* Jonathan Stray, Graduate School of Journalism at Columbia University. February 2nd, 2015.
7. *Teaching Assistant.* Francis Champagne, The Developing Brain. Department of Psychology, Columbia University. Spring 2012.

RELEVANT COURSEWORK

Carnegie Mellon University, Relevant Courses, Doctoral Degree.

1. Osman Yagan. *Applied Stochastic Processes.* ECE 18751. *Spring 2019.*
2. Jia Zhang. *Service Oriented Computing.* ECE 18655. *Fall 2018*

Columbia University, Relevant Courses, Masters Degree.

1. Steven Coll. *Investigative Techniques.* JOUR 6018. *Fall 2017*
2. John Paisley. *Machine Learning for Data Science.* COMS 4721. *Spring 2017*
3. Adam Kelleher. *Causal Inference for Data Science.* COMS 4995. *Spring 2017*
4. Ronald Neath. *Bayesian Statistics.* STAT 4640. *Spring 2016*
5. Eleni Drinea. *Algorithms for Data Science.* CSOR 4246. *Fall 2016*
6. Zoran Kostic. *Comp. Signals & Data Processing.* EECS 4750. *Fall 2016*
7. John Paisley. *Bayesian Models for Machine Learning.* EECS 6720. *Fall 2016*
8. Mark Hansen. *Data II.* JOUR 6015. *Fall 2015*
9. James Fan. *Deep Learning and Computer Vision.* EECS 6894. *Spring 2015*
10. Flavio Bartman. *Linear Reg. and Time Series.* STAT 4440. *Fall 2014*
11. Alexandr Andoni. *High Dimensional Data Analysis.* COMS 6998. *Fall 2014*

Columbia University, Relevant Courses, Bachelor Degree.

1. Itshack Pe'er. *Machine Learning.* COMS 4771. *Spring 2014*
2. I-Han Hsiao. *Data Visualization.* QMSS 4063. *Spring 2014*
3. Xi Chen. *Analysis of Algorithms.* CSOR 4231. *Spring 2014*
4. Michael Collins. *Natural Language Processing.* COMS 4705. *Fall 2013*
5. Anargyros Papageorgiou. *Comp. Linear Algebra.* COMS 3251. *Fall 2013*
6. Bodhisattva Sen. *Probability and Statistical Inference.* STAT 4109. *Fall 2013*
7. Martha Kim.*Fundamentals of Computer Systems.* CSEE 3827. *Fall 2013*
8. Salvatore Stolfo. *Artificial Intelligence.* COMS 4701. *Spring 2013*
9. Seung Choi. *Computer Science Theory.* COMS 3261. *Spring 2013*
10. Shlomo Hershkop. *Data Structures in Java.* COMS 3134. *Spring 2013*
11. Jae Lee. *Advanced Programming.* COMS 3157. *Fall 2012*
12. John Kender. *Honors Intro to Computer Science.* COMS 1007. *Fall 2012*

OTHER EXPERIENCES

Diversity

- *Co-chair and Founder:* NYT Young Professionals.
- *Contributor:* Unpublished Black History https://nyti.ms/2KvaGDM.

Professional Music Experiences

- *Double Bassist* for critically-acclaimed off-Broadway play, *The Dybbuk*. Review: https://bit.ly/2FDSrNB.
- *Double Bassist*. New York Youth Symphony, Carnegie Hall. *2006-2010*.
- *Pianist* All-State Piano Recital and Orchestra. *2008-2010*

MEDIA MENTIONS

(Selected)

1. *BuzzFeed*. Peter Aldous. How Russia's Trolls Engaged American Voters. https://bit.ly/2TFrhsB
2. *WIRED*. Louise Matsakis. What Does a Fair Algorithm Actually Look Like? https://bit.ly/2C8uyKN.
3. *The Wall Street Journal*. Benjamin Mullin. New York Times Adapts Data Science Tools for Advertisers. *https://on.wsj.com/2sA4yof*.
4. *National Public Radio*. Le Show, February 18th, 2018. https://bit.ly/2TGEpxF. (A discussion on Project Feels.)
5. *Business Insider Japan*. Fumiaki Ishiguro. AI in Advertising at the *New York Times* (Translated). https://bit.ly/2Q57g0D.
6. *Language Log, University of Pennsylvania*. Mark Liberman. Recommended for You. https://bit.ly/2U7qktu.
7. *KnightLab*. Shakeeb Asrar. A quick look at recommendation engines and how the New York Times makes recommendations. https://bit.ly/2Sa6T15. https://bit.ly/2AIhkBQ.
8. *Women Who Code*. Ema Kaminskaya. ODSC Event Reflections:

   > "Alex's ability to captivate and connect with the audience was a sight to behold. The whole talk felt like an informal conversation between the presenter and 150+ people in the audience. That's definitely a skill and a bit of a talent to manage such a big crowd in a very conversational way, encouraging questions and sparking curiosity."