# Nic Fishman

EXPERIMENTS · MACHINE LEARNING · STATISTICS

## Education

**Doctoral Student, Harvard University** — *Cambridge, MA*
Anticipated PhD in Statistics — *aug 2023 - present*

**Rhodes Scholar, Oxford University** — *Oxford, UK*
MSc in Statistics, advised by Prof. Yee Whye Teh — *oct 2021 - jun 2023*

**Undergraduate, Stanford University** — *Stanford, CA*
B.S. in Computer Science and B.A. in Sociology with Honors – GPA: 3.9 — *sept 2017 - jun 2021*

## Honors & Awards

2023 **James Mill Peirce Fellowship**, Harvard University — *Cambridge, MA*
The James Mill Peirce Fellowship recognizes excellent graduate students in the natural sciences, mathematics, and engineering at the Harvard Graduate School of Arts and Sciences.

2023 **Graduate Research Fellowship**, National Science Foundation — *Cambridge, MA*
The NSF GRFP recognizes and supports outstanding graduate students in science, technology, engineering, and mathematics. Awarded for work on machine learning and causal inference.

2021 **Rhodes Scholarship**, Rhodes Trust — *Oxford, UK*
The Rhodes Scholarship is an international postgraduate award for students to study at the University of Oxford.

2021 **Outstanding Senior Thesis Award and Firestone Medal**, Stanford University — *Stanford, CA*
Best thesis award in sociology and medal for top Stanford-wide theses. Received for work on the history of trial avoidance mechanisms and how colonialism and imperialism drove their diffusion.

2017 **G.R.E.A.T. Award**, NHGRI, National Institutes of Health — *Bethesda, MD*
The Genome Recognition of Employee Accomplishments and Talents (G.R.E.A.T) Award, given for work in the Ostrander Lab on the DOGSV system for storing and analyzing structural variants.

## Publications

### Published Papers

Fawkes, J.,* **Fishman, N.**,* Andrews, M., Lipton, Z., 2025. The Fragility of Fairness: Causal Sensitivity Analysis for Fair Machine Learning. *Advances in Neural Information Processing Systems,* 37.

**Fishman, N.**, Klarner, L., Mathieu, E., Hutchinson, M. and De Bortoli, V., 2024. Metropolis sampling for constrained diffusion models. *Advances in Neural Information Processing Systems,* 36.

Nilforoshan, H.,* Looi, W.,* Pierson, E.,* Villanueva, B., **Fishman, N.**, Chen, Y., Sholar, J., Redbird, B., Grusky, D. and Leskovec, J., 2023. Human mobility networks reveal increased segregation in large cities. *Nature*, 624(7992), pp.586-592.

**Fishman, N.**, Klarner, L., De Bortoli, V., Mathieu, E. and Hutchinson, M., 2023. Diffusion models for constrained domains. *Transactions on Machine Learning Research.*

**Fishman, N.** and Hancox-Li, L., 2022, June. Should attention be all we need? The epistemic and ethical implications of unification in machine learning. *ACM Conference on Fairness, Accountability, and Transparency* (pp. 1516-1527).

(Honors Thesis) **Fishman, N.**., 2021 Making (Global) Criminal Procedure: Empire, Justice, and Efficiency.

**Fishman, N.** and Davis, N. T., 2022. Change we can believe in: Structural and content dynamics within belief networks. *American Journal of Political Science,* 66(3), pp.648-663.

**Fishman, N.**, Shrikumar, A., Marinov, G., Kundaje, A., 2020. Systematic characterization of generative models for de novo design of regulatory DNA. *International Conference on Machine Learning, Computational Biology Workshop*, Spotlight.

### Working Papers

(In preparation for submission to Science) Furnas, Z., **Fishman, N.**, Hammond, B., Wang, D. Partisan Disparities in the Funding of Science in the United States.

**Fishman, N.** and Imai, K. Monotone Interference.

Ben-Menachem, J.,* **Fishman, N.**,* Galper, G.* The False Promise of Controlling False Positives: Ephemerality and the "Replication Crisis" in Observational Studies.

**Fishman, N.** and Rosenman R. Estimating Vote Choice with Approximate Poisson-Binomial Logistic Regression.

**Fishman, N.** Principled Models for Analyzing Max-Diff Surveys.

Shrikumar, A.,* **Fishman, N.**,* Kundaje, A. SimDNA: a library for simulating regulatory genomic sequences.

## Presentations

2022 **Invited talk, Should attention be all we need?**, Stanford Human Centered Artificial Intelligence — *Stanford, CA*

2022 **Invited talk, Should attention be all we need?**, DeepMind Sociotechnical Systems — *London, England*

2022 **Conference talk**, Law and Society Association Annual Meeting — *Lisbon, Portugal*
On 'Making (Global) Criminal Procedure'. Outlining the history of two trial avoidance mechanisms, plea bargaining and penal orders, and how colonialism and imperialism drove their diffusion.

2021 **Conference talk**, American Political Science Association Annual Meeting — *Seattle, WA*
Developed sparse-covariance Gaussian Mixture Models for studying political belief networks and the connections to Brandom's inferentialist account of belief formation. With J. Green and N. T. Davis.

## Work Experience

**Harvard Medical School, Harvard University** — *Boston, MA*
Research with Profs. Jonathan Gootenberg and Omar Abudayyeh — *may 2024 - present*

- Developed an extension for Schrödinger bridge and flow-matching to generalize experiments to new domains.
- Built a transformer-based single-cell model achieving state-of-the-art performance for "in silico experiments".
- Led project reproducing, unifying, and generalizing models for out-of-distribution prediction in single-cell sequencing. We generalized methods to allow extrapolating across perturbations and cell types. We benchmarked these models on a new suite of metrics across existing and novel single-cell datasets.

**Institute for Quantitative Social Science, Harvard University** — *Boston, MA*
Research with Prof. Kosuke Imai — *june 2020 - present*

- Constructed a non-parametric framework for causal inference under general interference using regression.
- Specifically developed efficient algorithms and theory for non-parametric monotone regression for estimation and inference in the general interference framework.

**Statistics Department, University of Oxford** — *Oxford, England*
Research in OxCSML Group under Yee Whye Teh — *sept 2021 - june 2023*

- Built novel models of cyclic peptides and antibody loop structures using constrained diffusion and flow-matching models respecting the loop constraints.
- Developed techniques for extending diffusion models to constrained domains in general Riemannian geometries.

**Center for Science of Science & Innovation, Northwestern University** — *Chicago, IL*
Research Fellow — *june 2021 - sept 2023*

- Synthesized huge international datasets on scientific grantmaking and publication to generate a picture of the landscape and political implications of scientific funding.
- Leveraged LLMs to develop fine-grained topic classifications across grants and paper abstracts to understand when grant-making institutions lead scientific investigation as opposed to following what is already being studied.

**Economics Department, Massachusetts Institute of Technology** — *Cambridge, MA*
Research Assistant for Prof. Victor Chernozhukov — *june 2021 - sept 2022*

- Developed methods for ML for dynamic treatment effect estimation in healthcare settings by extending g-estimation to non-linear blip functions using tools from minimax optimization.
- Developed asymptotic results for double/debiased machine learning in (quasi-)Bayesian estimation strategies.

**Knowledge Graph Team, Google** — *San Fransisco, CA*
Research Intern — *summer 2022*

- Research integrating LLMs and knowledge graphs (KGs) for search, particularly combining LLM representations with graph neural networks over a knowledge graph to produce KG-informed representations.
- Developed methods for identifying interpretable notions of author expertise using hybrid LLM/KG representations for topic modeling in multimedia content.

**Sociology Department, Stanford University** — *Stanford, CA*
Research Assistant for Prof. Aliya Saperstein and Prof. Michelle Jackson — *spring 2021*

- Constructed a timeline of major developments, conflicts, and characters in the history of statistics and eugenics.
- Developed the hypothesis that from 1850-1930 economics, psychology, and sociology embraced biological determinism, quantification, and professionalization to secure legitimacy and funding.
- Documented how statistical controversies drove eugenics and, perhaps more surprising, how eugenic controversies drove statistical innovation, through in-depth engagement with primary sources.

**Stanford Center on Poverty and Inequality and Stanford Network Analysis Project** — *Stanford, CA*
Research Intern for Prof. Jure Leskovec and Prof. David Grusky — *jan 2020 - june 2021*

- Used path crossings to study socioeconomic stratification in the United States, developing insights about heterogeneity in segregation across different social contexts and across different income deciles.
- Worked on algorithms to identify when individuals cross paths in time and space using massive GPS data.

**Kundaje Lab, Stanford AI Lab** — *Stanford, CA*
Research Intern for Prof. Anshul Kundaje — *jan 2018 - june 2021*

- Created modular system for generating DNA sequence using GANs/VAEs/Transformers and optimizing these models to produce sequences with specified properties, e.x. level of gene expression or chromatin accessibility.
- Developed a k-NN algorithm to evaluate (1) the fidelity of samples from generative models and (2) the robustness of neural network predictions on regression outputs (extending existing work on classification).
- Constructed neural net for predicting protein expression from whole genome ChIP-exo data. Used causal inference methods to extract and build a synthetic DNA regulatory logic for simulations in generative genomics.
- Contributed to SimDNA, a Python library for simulating DNA datasets to evaluate machine learning methods.
- Built a novel method for generating regulatory DNA to achieve targeted levels of protein expression through augmenting conditional GAN architectures (CS229 best project).

**Data for Progress** — *New York, NY*
Data Scientist — *june 2018 - may 2021*

- Led development of polling infrastructure that produced the most accurate poll results in the Democratic Primary. Built out MySQL database for storage and analysis of survey responses. Automated chart and report generation from this database. Developed search engine and website for internal use to assist in research.
- Designed, conducted, and analyzed polls used to guide policy change for the Green New Deal, Medicare for All, several HR-1 issues, and criminal justice reform, among other progressive issues.
- Developed novel poll weighting scheme achieving state-of-the-art accuracy.
- Automated argument detection so that non-technical colleagues to easily interpret open-ended survey responses in policy briefs by building hierarchical Dirichlet process models for non-parametric topic modeling.
- Created an ecological loss function and corresponding neural networks, extending ecological inference.
- Led team creating word2vec models to analyze gender/racial bias in news articles around the 2016 election.

**Sunrise Movement** — *Washington, DC*
Data Science Consultant — *fall 2020*

- Used LASSO and random forest algorithms to build interpretable heuristics for identifying low-propensity swing state voters for a large scale get-out-to-vote campaign in advance of the 2020 presidential election.

**Pritzker School of Law, Northwestern University** — *Chicago, IL*
Research Assistant for Prof. Josh Kleinfeld — *fall 2020*

- Used network models to explain plea bargaining's diffusion and its impacts on global criminal justice systems.
- Conducted research on various questions of legal/intellectual history.

**English Department, Stanford University** — *Stanford, CA*
Research assistant for Prof. Mark Grief — *apr - dec 2019*

- Conducted literature reviews on topics including social theory, psychology, and gender and sexuality studies.

**Star Lab Corporation** — *Washington, DC*
Machine Learning Specialist — *summer 2018*

- Worked on a Red Hat kernel module to log system activity and leveraged it for neural net anomaly detection.

**Ostrander Lab, National Human Genome Research Institute** — *Bethesda, MD*
Research Fellow — *dec 2015 - sept 2017*

- Responsible for building a DNA database to facilitate access and analysis of structural variants.
- Used random forest approach to identify candidate cancer risk genes from SNP arrays.