In my introductory biology class freshman year in high school, we did a unit on evolution, including how to figure out how closely related species are, using some genetic information. Our teacher called it "creating an evolutionary tree." As the final project of the unit, our teacher gave us a list of 15 species, each with 12 codons. Our job: list them in order of genetic similarity to humans. He wanted us, by hand, to compare 540 letters, and assign a difference quotient to each species from humans, then array them from smallest to largest according to that integer. It took members of my class weeks to finish, even when they created groups to share the workload. I finished the project in half an hour. The solution was to write a simple computer algorithm. Once I had the algorithm, and was able to enter the codons, it took my computer milliseconds to do flawlessly what would take hours of painstaking comparisons for a human. That was my introduction to the power of harnessing computers for solving problems.

Nearly a decade later I'm doing the same thing. I'm searching out confounding problems and trying to bring computing to bear on unraveling them. I work across genetics, sociology and political science, and I see profound connections between the methods that best inform each of these fields. In the end, everyone wants to be able to ask questions about their data and get answers they can be confident in. **That's why I'm drawn to problems of representation learning and robust inference: they seem almost universal; if we can solve them, we can use the solutions in all kinds of arenas.**

Research Background and Intellectual Merit

Part I, Genetics: Representing Regulatory DNA

Freshman year I started in the genetics lab of Prof. Anshul Kundaje, where I work on deep generative models for regulatory sequence design. Currently, the way genetic circuit design works is that eight bioengineers sit around a table with their laptops, pulling from the known list of DNA elements with known impacts to fit into their systems. This is hugely limiting, because it means that genetic systems (1) have to be hand-designed and (2) cannot really be fine-tuned. The end goal of regulatory sequence design is empowering biologists to say "I need a piece of DNA that will do this" — and have a generative model spit it out.

When I started in the Kundaje lab, there was no one working on generative genomics, so the work I did was largely independent. My first project in the lab was a class assignment for Stanford's machine learning class, CS229, and it revolved around developing a method for this type of optimization, essentially a conditional GAN with a continuous input, as opposed to the categorical inputs common to the literature. **My paper was selected as the best among 400 projects completed for CS229 that quarter.** In the course of that first foray, I realized that using deep generative models to engineer regulatory sequence is really a process of building high quality learned representations (using GANs or VAEs), and then tuning those latent spaces to only represent sequences exhibiting a given property of interest, most often maximizing or minimizing expected gene expression. I also realized that the literature didn't need more one-off methods; it needed baselines. There was no systematic review of existing methods, in realistic testing conditions, comparing their performance.

My next project focused on two things: first replicating all existing methods for sequence design and building a typology to compare, contrast and combine them, and second building some non-parametric tests for the "realistic-ness" of maximized sequences as a way to evaluate the quality of generated sequences. **I wrote a first-author paper summarizing these finding, which was a spotlight paper at the 2020 ICML Computational Biology workshop.** A key insight from replicating all of these methods was that sequences with the property of interest will always be sparse in the training data – otherwise you could just sample a generative model, no optimization required. This sparsity means that the quality of the generated sequences is upper-bounded by the quality of some analyzer model which makes predictions about the property of interest. Combined with the sparsity of training samples, that makes sequence design an out-of-distribution prediction problem. Worse, many methods in the sequence design literature are extraordinarily similar to the Fast Gradient Sign Method developed for generating adversarial examples in the vision literature. This only compounds the issue, meaning the analyzer needs to both make good predictions on data with minimal support in the training distribution, but also must be robust to adversarial examples (although these are of course related issues).

My most recent work in the lab is focused on developing these theoretical insights into some methods for quantifying analyzer uncertainty. As a part of this project, I worked extensively with simulated datasets, using the lab's internal simulation framework. I ended up rewriting large chunks of the codebase internally, parallelizing simulated sequence generation. I also commented and generally cleaned up the API. **Eventually I was invited to co-first author a manuscript explaining the software framework, which we are hoping to submit as a software paper to Bioinformatics.**

Part II, Politics: Representing Voters

At the same time I was joining Prof. Kundaje's lab, I was also becoming involved with a think tank called Data for Progress. There I have worked on getting high quality polling data to understand what the American public thinks about key issues, and getting that data into the hands of members of Congress and Senators, so the views of Americans are accurately reflected in the policymaking process. Polling requires a tremendous amount of infrastructure and a lot of methodological rigor to do well. I singlehandedly built the infrastructure that forms the heart of the organization's data operations, including automatic report generation, data visualization, and an SQL backend to support it all. But the infrastructure is useless unless the polling and analysis are accurate. Dr. Colin McAuliffe (a co-founder of Data for Progress) and I have been working closely for a year on two interrelated issues: (1) building a weighting scheme to best generalize our finite samples to the general population, (2) ecological inference, or trying to impute individual vote probabilities from aggregated election outcomes.

We developed a number of strategies for weighting polling data in an effort to correct for the polling issues of 2016, which primarily emerge out of failures to control for key variables like education and geography. But adding these additional controls tends to blow up the variance in polls, so we developed a novel weighting scheme using maximization subject to a variance constraint, allowing us to tune the tradeoff between matching our surveyed population to the electorate and the resulting variance.

It was also through Data for Progress that I met Prof. Nicholas Davis (University of Alabama) and started working on a paper representing Americans' political beliefs as Gaussian graphical models, or pairwise Markov random fields with Gaussian joint and sparse covariance. This parameterization of survey responses is becoming more popular in political science and psychology, and we were interested in studying it as a representation of populations of voters, and its relative stability and predictive value over time. In particular we wanted to explore the common practice of assuming that centrality within this "belief network" was predictive of change in element centrality over time. We concluded that changes in Gaussian graphical models do not, in fact, necessarily reflect changes in voters' beliefs. **I am the first author on this paper, which has been accepted at the American Journal of Political Science, the top journal in political science.** Prof. Davis is not really a methodologist, so I did all the quantitative analysis in our collaboration, with him providing the domain-specific theory and qualitative analysis (as well as details on the 2008 election, when the data is from, which I am not old enough to really remember).

These Gaussian graphical models actually formed the basis for the next phase in the development of Data for Progress's weighting scheme. We realized building a generative model of voter beliefs could allow us to account for more complex interaction effects, and when paired with our regulation scheme, the variance from inverse probability weighting could be kept minimal with only a small amount of bias. High quality voter representations seemed like an incredible candidate for improving the generalizability of our survey results. It was at this point that Dr. McAuliffe and I realized there was a deep connection between our work on weighting and our work on ecological inference. We realized we could use ecological inference to build voter representations which we could use in our weighting.

Ecological inference is the problem of predicting individual labels from aggregates. We developed a custom loss function for approximating the Poisson Binomial and Poisson Multinomial likelihood, allowing us to use neural networks to learn voter representations. The flexibility of neural networks also enabled us to do multi-task learning of multiple elections, and we found that this led to higher quality learned representations. We also found that the learned representations were better at making out-of-distribution predictions than the raw covariates across a range of prediction tasks on survey

questions. **This work culminated in a paper that I am first author on, which is under review at the International Conference on Learning Representations.**

Part III, Causal Inference: Gaussian Processes and Uncertainty

Last year I took my first causal inference course. I was immediately hooked. Being able to precisely estimate the effect of an intervention seems like the holy grail everywhere from basic science, to clinical applications, to public policy. Later that school year I took Prof. Susan Athey and Prof. Stefan Wager's course Causal Inference and Machine Learning, and it opened up the world of causal inference to the tools I had been using in my other research for the past several years. I reached out to Prof. Kosuke Imai (Harvard IQSS), and we developed a proposal for a project using Gaussian processes to estimate heterogeneous treatment effects as functions of continuous variables. I applied for and received a Major Grant ($7,500) from Stanford University to pursue this project this summer. Focusing on estimating treatments as a function of distance, Prof. Imai and I decided that enforcing monotonicity constraints would give us additional power and were a plausible assumption. **I spent the summer independently developing a new method for enforcing monotonicity constraints in Gaussian processes by adding a monotonicity term to the stochastic variational loss function.** We are planning to take this work to publication.

Broader Impacts

Academic In college I have also done direct work around disability rights and sexual assault. I built a system for students to document professors' possible violations of the ADA, which the student senate is preparing to deploy. I am also part of an ongoing project to dehouse fraternities because of their role in sexual assaults. Going forward I want to continue to work with the disabled community, who are often excluded or actively harmed by scientific pursuits, and to support survivors of sexual assault in any way I can. I also went to a public high school in Washington, DC. I did not know the Intel Science Fair existed until I got to college. I only learned about the opportunity to work as an NIH intern through a family friend who worked there. As an academic, I hope to bridge that gap in the pipeline by working to get under-served high school students' exposure to research opportunities.

Societal For several years I have been a software engineer doing infrastructure and computational modeling at the think tank Data for Progress (DfP), and through that work I have supported DfP's collaborations with organizations from the ACLU and the Human Rights Campaign, to the Sunrise Movement, Indivisible, and The Justice Collaborative. We consulted with Sen. Elizabeth Warren on corruption reforms, Sen. Bernie Sanders on a medical innovation prize, Sen. Kamala Harris on ending cash bail, Sen. Cory Booker on housing expansion, Sen. Tammy Baldwin on employee governance. We worked closely to develop HR3 (Rep. Elijah Cumming's "Lower Drug Costs Now Act") and were a driving force behind the adoption of automatic voter registration in New York state. I also do research with Prof. David Grusky and the Center for Inequality and Poverty, which has been illuminating about the depth and complexity of issues surrounding economic injustice, and how we might begin to repair those deep rifts in society.

Future Goals

I am currently applying to PhD programs in statistics and computer science. A PhD will give me a stronger mathematical foundation for understanding the theory behind my work, and a language to formally express when it works and why. In an ideal world I would lead an interdisciplinary research group at a university working on statistical methods with wide ranging applications across the hard and the social sciences. I want to keep working on genetic engineering without having to give up my work on economic inequality. I do not want my work to be purely academic: I want to work with industry and with government to translate research into substantive products and policies that change people's lives for the better. An NSF Graduate Research Fellowship would raise my work to the next level and give me the flexibility to tackle the interdisciplinary set of problems I expect to devote my life to.