Zhenghong Lieu, M.S. in Statistics (Data Science)

I want to learn advanced computational and statistical methods from the Data Science program, then work in industry as a data scientist before bringing best practices from industry to the civil service. My domain knowledge in politics and economics and my position as a scholar in a fast-moving statutory board put me in a powerful position to shape social policy; specifically, I hope to combine data science methods with social science research to combat unfair political practices and build tight-knit social communities.

I possess strong mathematical and computer science skills and a good work ethic

I have a strong mathematical background, having taken the most quantitative courses available to me in Oxford. I have learned probability and statistics in Quantitative Economics, multivariate calculus and constrained optimisation in Microeconomics, and experimental design from Behavioural Economics. I will be taking courses in Game Theory and Microeconomic Analysis, both courses with a heavy mathematical component. Microeconomic Analysis will cover linear algebra in-depth. I have specifically requested and taken an extra Mathematical Analysis course in Oxford normally taught only to mathematicians, learning how to anchor calculus to a set of rigorous foundations. I performed well in it, receiving a good report from the professor.

I have also engaged in extensive self-study, specifically choosing to complete the most theoretically rigorous MOOCs. I have completed courses in Linear Algebra and Multivariate Calculus. In computer science, I have built my own operating system in a computer architecture course and completed Jennifer Widom's Databases and Tim Roughgarden's Algorithms courses on Stanford Lagunita. Tim Roughgarden has said on the Algorithms course that "the course is rigorous... after completing this course, you will have a greater mastery of algorithms than almost anyone without a graduate degree in the subject."

My previous software engineering internships and self-learning have made me a proficient programmer. I am most comfortable in Python, having built and maintained multiple medium-sized (~5000 LoC) architectures with it, but I am also familiar with JavaScript, R and C++. I have worked on deep learning projects in industry and am therefore well acquainted with libraries like Keras and scikit-learn. I have also built blockchain-enabled applications with Hyperledger, bespoke building inspection software with JavaScript and Electron, and an item-tracking and SMS-reminder service for the Singapore Army.

I work hard to go beyond what is taught. I won the Raff Prize for best second-year Economics performance in my college in part because I researched a specific economic growth model deeply, and could therefore answer a question on it very well (I was awarded the highest mark). I also won the essay prize for best quantitative essay of my Oxford politics cohort (~300 people) due to my grasp of advanced econometric methods not taught in the course. My self-study allowed me to formulate sophisticated econometric criticisms and support them by running extensive panel regressions; I will devote the same effort to learn as much as possible in the program.

I have relevant data science experience from my research interest and previous work

My passion is combating unfair political practices and building tight-knit social communities. As such, my primary research interest is using computational and statistical techniques to detect and prevent gerrymandering. In order to "pack" voters into safe seats, gerrymanderers often lump together people who live very far away and have nothing in common. This greatly damages democracy: if people's votes don't matter and they have nothing in common with co-districtors, they become disenfranchised and disillusioned. If

we can detect gerrymandering and create fair districting plans, we directly empower millions.

Therefore, I am currently co-authoring a paper with professors Jonathan Rodden (Stanford) and Nick Eubank (Duke), where I use travel times between voters to better detect gerrymandering. In my own thesis, I build a novel computational metric that allows me to draw districts with increased social interaction and better democratic outcomes. This brought to bear many technical skills: I've had to work with geospatial datasets using Python, learn data science and geospatial libraries like GeoPandas, and calculate and store large amounts of data (5 billion travel times between voters) efficiently using SQL. I've also learned the basics of stochastic processes, as I use Monte Carlo Markov Chain (MCMC) methods to generate distributions of redistricting plans. At Stanford, I would love to continue building upon the work that I have already done with Prof Jonathan Rodden. I want to learn advanced stochastic methods to improve both the scope and robustness of my districting plans, and fight to replace partisan districting with automated districting in Singapore and the US.

My previous data science internship and self-learning have given me a strong foundation in the data science skills necessary to excel in the course. I highlight three projects that showcase my strong learning ability and familiarity with data science. First, I performed distributed data analysis on a Raspberry Pi cluster using Scala and Apache Spark, calculating average speeds of over 50,000 trips and visualising the results. Having average-speeds-per-road gave the company a more accurate metric of dangerous driving---feeding directly into their risk prediction models. In this project, I learned the functional programming paradigm in Scala and distributed computing with Apache Spark from the Big Data Specialisation by EPFL. At Stanford, I aim to learn new ways to analyse massive data sets.

To increase the accuracy of my company's risk prediction model, I built an end-to-end deep learning pipeline. The pipeline did everything: reading and cleaning the data, extracting and normalising the feature set, and finally training, checkpointing and saving the model. I learned about---and experimented with---various architectures from simple 1D CNNs to RNNs and LSTM models. The company took my model and is using it to secure a contract with the world's second largest dashcam manufacturer, opening up a new revenue vertical for the company. For this project, I benefited greatly from Andrew Ng's Machine Learning course and Deep Learning Specialisation on Coursera. At Stanford, I want to learn how to build more powerful predictive models while being sensitive to ethical problems like privacy and racial bias.

Bringing my training in experimental economics and probability together, I ran a randomised controlled trial (RCT) where I sent loss-aversive SMSes to see if they were more effective at "nudging" customers. I solved this multi-armed bandit problem optimally using Bayesian posterior sampling. I built a software architecture that sent customised SMSes to customers and tracked the clickthrough rates, then used that data to update the next wave of SMSes. This intervention increased the client's monthly active users by 20%, directly increasing revenue. For this project, I self-studied Bayesian statistics and also picked up SQL from Jennifer Widom's Databases course. At Stanford, I want to increase my knowledge of Bayesian statistics and experimental design to design and run more robust evaluations.

I am certain that the Master's program will give me the knowledge and skills to achieve my goals in social policy, and I wholeheartedly hope to be able to take it. Thank you.