Statement of Purpose (CMU Machine Learning & Public Policy)

Joshua Fan - jyf2@uw.edu

Machine learning has emerged as a powerful tool to identify patterns and make predictions from data; it has been applied to impactful problems such as diagnosing diseases, food insecurity, and alleviating poverty. I want to pursue a PhD in Machine Learning and Public Policy in order to extend cutting-edge **machine learning** innovations to benefit people. In particular, I am very interested in exploring applications of machine learning to improve **public policies** and **healthcare systems** for the **developing world**. I am also interested in **probabilistic models** and **network science**, as they can help us better understand the complex social and biological systems responsible for challenges such as poverty and diseases. Because machine learning can unfortunately lead to discriminatory and harmful effects if misapplied, I am also very concerned about ensuring **fairness** and **interpretability** in machine learning models.

I began pursuing my interest in using technology to address global development issues as a research assistant at the Information and Communication Technology for Development (ICTD) Lab during my third year as an undergraduate at the University of Washington. At the ICTD Lab, my work focused on redesigning the mPneumonia app. The app helps public health workers in India and Ghana collect data from patients and recommends treatments based on their responses. I collaborated with PATH, a global health company, to update the app's logic to match complex new medical protocols. Since this app was intended to replace cumbersome paper-based forms, I also worked with field testers to address their concerns and make the app easier to use. From this experience, I saw firsthand how technology has the potential to transform healthcare systems in developing regions by making sure that patients are treated correctly.

Working on the mPneumonia app showed me how technology and policy can go hand-in-hand to improve people's lives in the developing world. However, the app used brittle hard-coded rules to diagnose disease, and did not use the data it was collecting to improve future recommendations. When I took a course in machine learning, I was intrigued by how it can automatically learn patterns from data, instead of requiring humans to come up with hard-coded rules. I wanted to explore this further through research. In the spring and summer of 2017, I worked with Prof. Sreeram Kannan to identify patterns and cell types from noisy single-cell RNA-seq gene expression data using machine learning techniques. The previous methods for doing this relied on very slow optimization procedures that were unable to handle large datasets. In an effort to make this process more scalable, I implemented several online optimization algorithms which converge faster in general, and compared their speed and accuracy.[1]

However, these algorithms were still not fast enough, and we looked for new approaches. We noticed that there has been a lot of work to speed up topic modelling algorithms for texts. While these algorithms are from a different application domain, they still address the problem of finding low-dimensional structure in a noisy high-dimensional dataset. On the theoretical side, we showed that the underlying probabilistic models behind topic models and our gene expression dataset were very similar, so it makes sense to apply topic modeling algorithms here. I then

---

[1] http://joshuafan.github.io/files/CSE_547_Final_Report.pdf

adapted these algorithms to work with genomic datasets. I also derived a technique to intelligently initialize the algorithm with prior information. The algorithms I added were able to run on large genomic datasets (up to 1 million cells) efficiently, which previous approaches were unable to handle. The cell archetypes we identified were usable for downstream applications, such as tracking the progression of cell states over time. This work contributed to a paper in *Bioinformatics*.[2] Through machine learning, we were able to extract valuable information from a very noisy single-cell sequencing dataset, which has the potential to improve our understanding of how different cell types function and potentially inform the design of new medicines.

During my internships at Facebook, I again had the opportunity to use machine learning to transform data into insights that can inform decisions. In Fall 2018, I applied a convolutional clip-based neural network to predict whether a video contains violent content. The model was able to detect violent videos more accurately than existing approaches, preventing users and Facebook's content reviewers from being exposed to traumatizing and graphic content. Since violent content can desensitize people to violence and lead to real-world harm, it is especially critical to use tools like machine learning to protect communities from such dangerous content.

In addition to doing research that benefits others, I am also passionate about directly serving the community around me. For example, as a lead Teaching Assistant for courses in Probability and Discrete Math, I took initiative to design additional resources on confusing topics, organize review sessions, and mentor newer TAs. I made sure to understand students' thought processes, which was especially important in topics such as combinatorics where there can be a bewildering array of possible approaches. I care deeply about helping others understand difficult concepts, so that they can stretch their thinking and apply these concepts in fascinating ways down the line.

Ultimately, I hope to pursue a PhD so that I can benefit and serve people more effectively. I've seen first hand that machine learning is not just a set of interesting algorithms that can generate revenue for large corporations, but has significant potential to improve society. My aspiration is to become a professor who works closely with policy-makers and organizations to create machine learning techniques that can inform better policies and solutions. I would like to continue my studies at Carnegie Mellon University because of its interdisciplinary strengths at the intersection of machine learning and public policy. While I am interested in a wide range of topics, the research of Professors **Roni Rosenfeld, Alexandra Chouldechova, Artur Dubrawski,** and **George Chen** are especially exciting to me, as they apply machine learning to socially impactful problems such as forecasting epidemics,[3] targeting rural villages for development,[4] and fighting child abuse in a fair way.[5] The machine learning techniques developed at Carnegie Mellon University have made a large difference on important policy problems that touch so many lives, and I would love to continue this progress through the joint PhD program in Machine Learning and Public Policy.

---

[2] Sumit Mukherjee, Yue Zhang, **Joshua Fan**, Georg Seelig, and Sreeram Kannan. "Scalable preprocessing for sparse scRNA-seq data exploiting prior knowledge." *Bioinformatics*, 34, 2018, i124–i132. http://joshuafan.github.io/files/UNCURL_paper.pdf
[3] Logan Brooks, et al. "Nonmechanistic forecasts of seasonal influenza with iterative one-week-ahead distributions." *PLOS Computational Biology,* 2018.
[4] Kush Varshney, et al. "Targeting Villages for Rural Development Using Satellite Image Analysis." *Big Data*, March 2015.
[5] Alexandra Chouldechova, et al. "A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions." *Proceedings of Machine Learning Research 81:1-15*, 2018.