Statement of Purpose

Ameya Daigavane

I am interested in machine learning for spatiotemporal data, especially because of its broad applicability to solve complex modelling problems in the computational sciences.

I was first exposed to this exciting line of research during a wonderful internship at NASA Jet Propulsion Laboratory in the summer of 2019, funded by the Caltech Summer Undergraduate Research Fellowship (SURF). I worked with Dr. Kiri Wagstaff and Dr. Gary Doran at the Machine Learning and Instrument Autonomy group to detect environmental transitions in data collected by the plasma spectrometer onboard spacecraft. The plasma spectrometer provides deep insight into the composition of planetary magnetospheres. Our solutions could enable future instruments to automatically switch operating modes based on the immediate spacecraft environment. Such responsive instrumentation would not only collect higher fidelity data but could also possibly image novel scientific phenomena in unseen worlds. We realized that identifying transition events between different environments could be cast as an anomaly detection problem in multi-dimensional time-series data. With this insight, I implemented and evaluated multiple techniques to identify magnetospheric transition events in observations collected by the Cassini Plasma Spectrometer on the recently concluded Cassini mission to Saturn.

One of my key contributions in this project was a novel extension of the Matrix Profile algorithm to identify anomalous subsequences in multi-dimensional time-series. Previously, the Matrix Profile could only be employed for anomaly detection in unidimensional time-series. Of all the methods we investigated, our extension had the highest recall at bounded precision when detecting Saturn ‘bow shock’ transitions in Cassini data. I explored time-series visualization techniques to better understand our data, and performed exhaustive analyses to understand modelling errors. Our research found that generalization across mission years was a challenge for all of the algorithms: models often failed when spacecraft orbits, and hence, interactions with planetary environments changed significantly. A promising approach to handling such domain shifts is better modelling of the underlying spatio-temporal dynamics in the spacecraft environment. This has driven my desire to continue studying such complex time-evolving systems in graduate school.

I have had the opportunity to present this research at the Second AI and Data Science Workshop for Earth and Space Sciences at Caltech, and the 6th Workshop on Mining and Learning from Time Series at KDD, where ours was one of five papers accepted for oral presentations. Our latest work is now under review at Computers and Geosciences and IEEE Transactions on Aerospace and Electronic Systems, both top journals in their respective fields. In 2020, I was one of three winners of the ACM SIGBED Scholars Award from all over the world, recognizing the impact of our research on autonomous embedded and cyber-physical systems.

As a Pre-Doctoral Researcher at Google Research, I am currently working on multiple applications of graph neural networks (GNNs), which have recently become incredibly popular for modelling structured spatial data. Despite their popularity, I quickly noticed a lack of accessible resources to understand the mathematical building blocks of GNNs. In response, I began to create an expository article filled with interactive visualizations that would motivate these building blocks as natural

generalizations of image filters found in Convolutional Neural Networks. My first-authored paper on the challenges of designing and creating these visualizations was accepted with strong reviews at the Rethinking ML Workshop at the International Conference for Learning Representations (ICLR) 2021. The complete article was later accepted at Distill, an ML journal that focuses on lucid explanations of ML research ideas.

One of the new challenges that spatiotemporal models such as GNNs introduce when being deployed in real-world applications is preserving the privacy of the data they are trained on. Current differential privacy techniques for securely learning neural network parameters do not translate directly to GNNs. This is because the neighbourhood aggregation operation in standard GNN architectures induces significant correlations across the predictions of neighbouring nodes, greatly increasing the possibility of privacy leakage. With Dr. Prateek Jain, Dr. Abhradeep Thakurta, Dr. Gaurav Aggarwal and other collaborators at Google Research, I led a study of this challenging setting of privately learning GNN parameters for node-level tasks. My contributions here were multifold: I performed a sensitivity analysis of existing GNN models, derived a non-trivial extension of the 'privacy amplification by subsampling' theorem, and set up an efficient distributed pipeline to enable private training of GNN models. We demonstrated that node-level differentially-private GNNs outperform traditional non-graph based models on multiple benchmark datasets, while still respecting node-level privacy. I am the first author on a paper detailing this research, currently under review at the International Conference for Learning Representations (ICLR) 2022.

At Google, I collaborated with the Google Accelerated Science team to improve Plateviewer, a data visualization tool for cell imaging experiments by adding visualization capabilities and data filtering choices. Plateviewer is used to analyze cell cultures imaged by biologists at the New York Stem Cell Foundation (NYSCF), allowing visual identification of both experimental failures and novelties. Plateviewer's multi-faceted visualizations bolster confidence that downstream models trained on these images are not picking up on spurious features. As a result, we could successfully train models to identify signatures of Parkinson's disease from cell-painted human fibroblasts. I enjoyed working on such multidisciplinary research that blends together different algorithmic techniques to solve important problems in the life sciences. Our article on this research has recently been accepted at Nature Communications.

Spatiotemporal systems show up everywhere across multiple subdisciplines: computational chemistry, physics, biology, astronomy, geometric learning, computer vision and robotics, to name a few. At MIT, I am particularly interested in Prof. Tess Smidt's research on equivariant models for predicting molecular properties, Prof. Connor Coley's work on learning 3D molecular representations for reaction predictions and virtual drug screening, Prof. Regina Barzilay's research on machine learning for drug discovery (as part of the Machine Learning for Pharmaceutical Discovery and Synthesis consortium at MIT), and Prof. Mina Konaković Luković's research on computational material design and fabrication. In the future, I want to tackle the important open problems of rectifying concept drift and understanding how to better integrate inputs across multiple modalities in spatiotemporal models.

MIT's rich interdisciplinary academic culture aligns strongly with my own research motivations. With my ability to dive into both theory and empirical methods, I am confident that I can succeed as a graduate student at MIT.