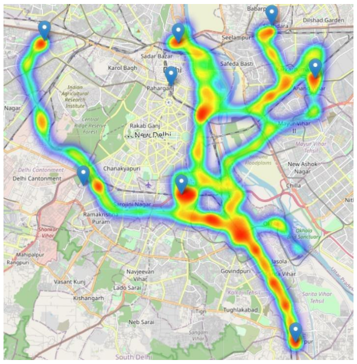
Machine Learning Problem: Prof. Rijurekha Sen's team has been working to scale up Particulate Matter (PM) measurement using vehicle mounted sensors[1]. Her team has a pilot deployment of pollution measuring sensors on 10 DIMTS buses (regions covered shown on the left) in Delhi that are able to record fine-grained particulate matter (PM) values (PM1.0, PM2.5 and PM10) along with GPS coordinates. While mobile sensors mean values are not collected at all places at all times, with state of the art spatio-temporal interpolation methods like Gaussian Processes [2] and Graph Convolutional Neural Networks [3], values where no sensors are present at a particular time can be estimated. The mobile sensors, along with interpolation methods, can give much more fine-grained information, than the current ~35 static pollution sensors, deployed by CPCB and DPCC to cover the huge geographical areas of Delhi-NCR. Thus the Machine Learning Problem of interest here is that of training spatio-temporal interpolation models from pollution data.



Privacy Problem: Privacy preserving ML is a very important research topic these days. For example, hospitals trying to train cancer detection using patient images, cannot share patient images with other hospitals to preserve patients' privacy. However, the trained model can be much better, if images of all hospitals could be fed into the training algorithm. In scaling the vehicle deployed pollution sensing to other fleet companies like Ola and Swiggy, Prof. Rijurekha's team is faced with a similar privacy concern. These fleet companies do not want to share their real time or historical GPS location traces, as that information is of high business value. Ola doesn't want Uber to know the Ola vehicle positions, Swiggy doesn't want Zomato to know the Swiggy vehicle positions. But for the training the spatio-temporal interpolation models for pollution, location-time and PM data collected by the fleets is necessary. Thus how to combine the data collected by different vehicle fleets for interpolation model training, while keeping the GPS location information private will be explored in this MTech Thesis.

Privacy Preserving ML: Facebook has recently released the Crypten framework [4] for privacy preserving machine learning. It implements a Secure Multi-Party Computation (MPC) library using Secret Sharing protocols. We will implement our pollution interpolation training mechanism within Crypten's framework, which will need significant coding efforts as Crypten currently supports few ML functions that need to be extended. We will analyze the accuracy-vs-latency trade-offs of our implementation, as pollution is a dynamically changing phenomena, and the interpolation has to run periodically, instead of a single shot ML model training. Thus latency of model training using Crypten is important. Based on the trade-off analysis of whether ***computation*** in the cryptographic protocol is the bottleneck or ***communication*** among the different fleet companies' servers holding the data, we will try appropriate optimizations. The work will this involve understanding of cryptographic protocols, ML and systems.

References[1]
https://timesofindia.indiatimes.com/city/delhi/iit-team-to-mount-devices-on-200-cluster-buses-to-check-air-pollution/articleshow/69965333.cms, [2] https://gpytorch.ai/, [3] http://snap.stanford.edu/graphsage/, [4] https://ai.facebook.com/blog/crypten-a-new-research-tool-for-secure-machine-learning-with-pytorch/