Capstone 1 Proposal:
Author: Varsha Gopalakrishnan

Problem Statement:
In densely packed cities like Oakland or San Francisco, air quality can vary wildly across neighborhoods, due to varying sources of emissions. Current methods to monitor air quality, while certainly very useful, are either too widely spread out (sparsely located monitoring systems set up by local agencies) or too localized (personal air quality monitors that have a small radius of detection). In this project, we plan to answer the following question:

> Can we predict air quality in different Oakland neighborhoods based on local meteorological conditions, local sources of emissions, and previously measured concentrations without having to rely on complex physical modeling?

We will predict air quality in the City of Oakland on a city-block basis where they are not measured using multiple sources of publicly available data such as:
- Oakland Air Pollution Monitoring Data measured by Environmental Defense Fund (EDF)
- Air Emissions data obtained from the National Emissions Inventory,
- Intersection Count from Open Street Maps,
- Local meteorological data on a 1kmx1km grid obtained from Oak Ridge National Lab
- Normalized Difference Vegetation Index (NDVI) from the Planet Labs API.

I will also build an interactive map where a user can click on a location to get its predicted concentration of Black Carbon and Nitrogen Dioxide.

Audience:

Monitoring pollutant concentration within cities is crucial for environmental management and public health policies in order to promote sustainable cities. Apart from being a resource to the general public, this project can be useful to:
- To detect if there is an anomaly in air quality on a given block.
- To detect extreme events in air quality due to anomalies.
- Local and state air agencies: Get air quality measurements on a near real-time basis
- Social fitness apps like Strava, Fitbit, Nike Run who can leverage this data to recommend routes for users to run, bike, hike, etc

Datasets:
- The air pollution monitoring data is obtained from the EDF website. The data contains latitude, longitude (points where the measurements were taken), concentration of Nitric Oxide (NO), Nitrogen dioxide (NO2) and BC. The data was collected in Oakland between June 2015 - May 2016 at 21,448 locations. These readings were taken by Google Street View cars fitted with air pollution monitoring equipment.

- Data on local sources of air emissions such as major industrial facilities were obtained from the National Emissions Inventory (NEI) for Alameda County. Data on Particulate Matter (PM2.5, PM10) and NOX were obtained from the NEI. Here PM2.5 and PM10 data are used as a proxy for BC since BC data is not available. The NEI provides annual emissions of pollutants from each industrial facility for the year 2017 along with the facility name and location.
- Local traffic data for all major traffic intersections in the Oakland-Berkeley-San Leandro neighborhoods were obtained from CalTrans for 2015. CalTrans provides intersection level traffic counts as Annual Average Daily Traffic (AADT) along with the coordinates for each intersection in their GIS web tool.
- Local meteorological data is obtained from Oak Ridge National Lab's Daily Surface Weather and Climatological Summaries here. The dataset contains gridded estimates of daily weather data including total daily precipitation, minimum and maximum surface temperature, humidity, shortwave radiation, snow water equivalent and day length for the whole of North America. Data for each coordinate of interest is obtained using The '*daymetpy*' python library provides APIs to work with the daymet dataset.

Methodology:

Details on the specific models to be used will be refined as I progress through the course and work on the analysis. At a high level, these are the following questions I would like to answer:

- Can we build an air quality prediction model that relies on publicly available datasets to get an accurate prediction of air quality per city block?
- What are the major factors that affect air quality in a neighborhood or block?
    - Does this vary by location?
    - Do neighborhoods located closer to major intersections/highways have worse air quality than neighborhoods closer to industrial sources?
    - Can we identify the top five/major industrial sources and traffic intersections that contribute most to poor air quality?
    - Which meteorological parameters impact air quality the most?
- How can local air agencies implement better measures to control air pollution coming from major industrial sources and major traffic intersections?
- How does the model built for Black Carbon perform compared to the model built for Nitrogen Dioxide?
    - What are the features that are common and different between the two models?
- How does this compare with results from physical models and air pollution monitors?

Deliverables:

The end product will be an interactive map, the code for that will be hosted on Github and a report. The Github repository will also contain a presentation slide deck summarizing the project, Jupyter Notebook containing the analyses and visualizations, supporting notebooks for any supporting analyses and the underlying datasets.

Limitations and Future Work:

- One of the limitations of this model is it relies on air pollution monitoring data measured in 2015 - 2016. Thus the tool will not give predictions in near-real time.
- The next step is to replace the 2015-2016 monitoring data with air quality measurements from localized personal monitors, and monitoring equipment setup by local agencies and apply the same methodology developed here to predict concentrations on a city block basis.