Hi Ben,

I hope everything is going well. We are writing to update you on our progress on the air pollution project so far. Here is what we have done up to this point:

- Reviewed the relevant literature published by HSPH
- Setup our collaboration infrastructure, including a forked GitHub repo and Slack channel
- Familiarized ourselves with the sensor and satellite data, including EDA (We've attached a plot demonstrating correlation between aerosol optical depth and PM2.5 concentration. It was good sanity check for us.)
- Obtained computing access and prepared to run jobs on Odyssey
- Reviewed existing GitHub code
- Started to think about data imputation and $PM_{2.5}$ prediction models. More specifically, for more accurate imputation we're interested in potentially using gaussian process regression. For more accurate pollution prediction, we're looking into alternative convolutional nets using models from more recent literature.

Here are the next steps we are looking to take:

- Obtain satellite data for areas of the United States without sensors
- Run existing GitHub code on Odyssey
- Prototype some initial data imputation methods and $PM_{2.5}$ prediction models

Looking ahead, we wanted to ask you:

1) If HSPH currently has any satellite data for areas of the United States without sensors, so that we can start thinking about how to develop continuous-area $PM_{2.5}$ models.

2) At what frequency the PM2.5 sensor data in the given .csv is collected. In a random subset we gathered, we observed that over 75% of the sensor readings were N/A.

Thank you for the help that you have provided us so far - we very much appreciate it.

Best,
Chris, Casey, Justin, Keyan