**MILESTONE – 2** (PROJECT GROUP #111)

Forecasting Covid-19 cases using sars-cov-2 titers in wastewater

**TEAM MEMBERS:** *V Bhatia, P Bhatt, L Bondili, R Soori*

# Background

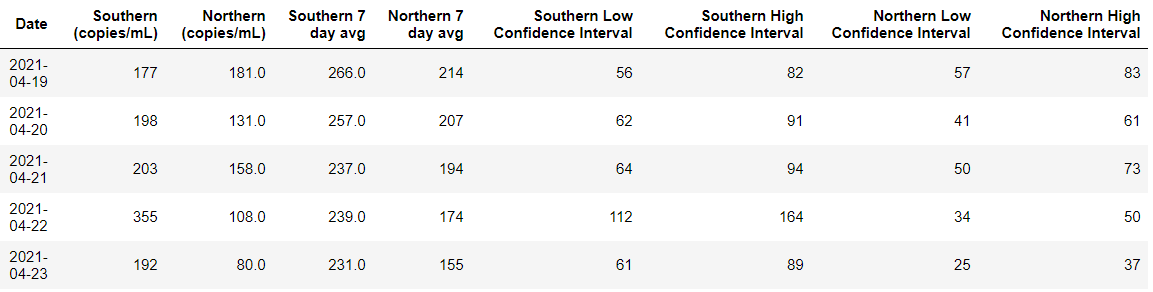
Conventionally COVID-19 is tracked via symptomatically and clinically diagnosed cases. This method has inherent drawback as many of the patients do not report the disease. Recent developments in detecting COVID-19 cases en masses have led to techniques such as measuring viral titers in urban wastewater as a method to track and forecast COVID-19 cases at the local level as suggested in the recent work by Wu et al[1]. The authors were successful in uncovering a trend that shows the presence of the virus in urban wastewater is a strong indicator of the current rate of infections and leads the gold standard clinical data with by a margin of 4-10 days. The work suggests that longitudinal wastewater analysis can be used to predict COVID-19 cases more accurately and more instantaneously than clinical data.

# Description of Data

We intend to utilize two distinct datasets for this project.

1. **MWRA Wastewater COVID-19 Tracking :** This is a publicly available dataset collected and maintained by the Massachusetts Water Resources Authority [2] measuring the quantity of COVID-19 viral RNA samples in sewage wastewater for Boston’s Southern and Northern districts. The dataset includes daily titer measurements starting from beginning of pandemic circa March-2020 up until the present data. For each of the district’s counties, the data also provides moving average values and confidence intervals.

## Table showing Sample data



2. **Massachusetts official COVID-19 Case Count :** This is also a publicly available dataset provided by the Massachusetts Department of Public Health [3] which includes official counts of COVID-19 cases and deaths reported in the state of Massachusetts. The dataset also includes additional information such as number of hospitalizations, vaccinations and other information relevant to the outbreak. At present, we only intend to use the daily number of confirmed case count per county for our project.

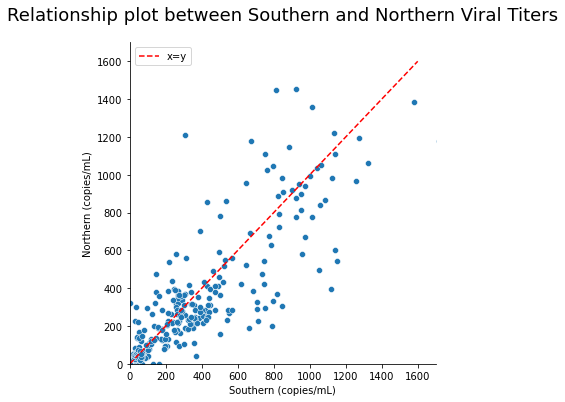
# Project Proposal

Given the two datasets above, we intend to explore and establish the relationship between wastewater surveillance data and the outbreak of COVID-19 in the neighboring communities. We intend to use supervised or self-supervised learning techniques to develop models which can accurately predict the current count of COVID-19 cases in the community using solely the data available from wastewater viral titers.

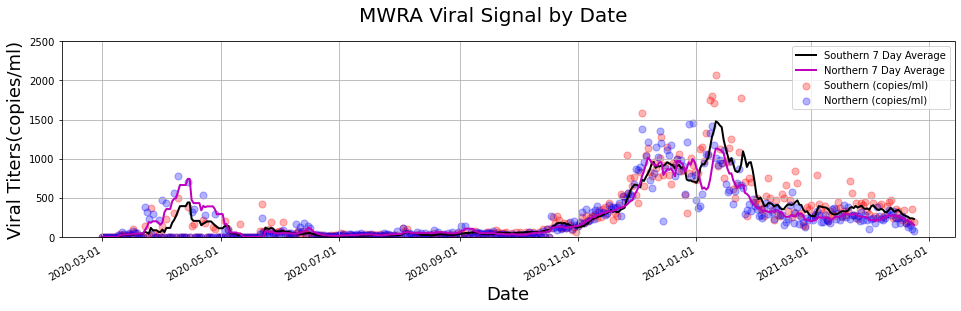
# Exploratory Data Analysis – Insights Gained

## Relation between *Northern* and *Southern* samples

Based on the plot shown below, we see a strong linear relationship between measurements from the Southern and Northern districts which shows consistency and provides some evidence that quality of measurements is high.

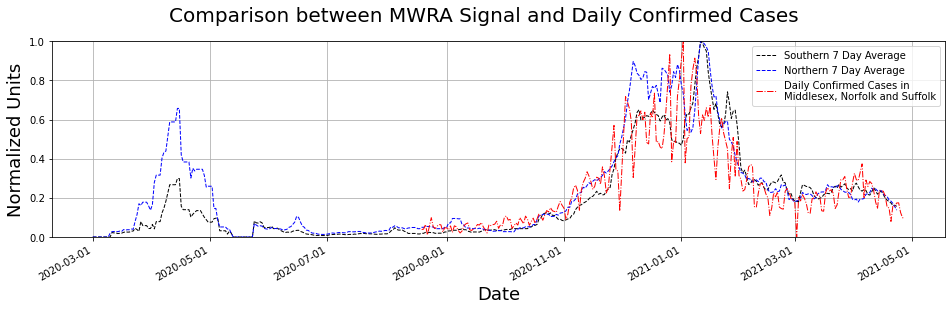


The above-mentioned point is also clearly explained in the time series data shown below which again shows a strong correlation between trends for both Southern and Northern districts.



## Comparing Viral RNA signal to Daily COVID-19 Confirmed cases

The following time series plot overlays the viral RNA signal from the two districts along the daily number of confirmed COVID-19 cases in the Mass counties of Middlesex, Norfolk, and Suffolk. Data has been normalized to visualize the trend in the curves. It is also important to note that daily case count in Massachusetts was only available from the beginning of August-2020, hence the viral RNA signal preceding that time will not be available for model training.



# Model Implementation Plan

We envision the use of following models for forecasting COVID-19 cases.

1. Naive Model – Our baseline model would be developed using the persistent models described from Influenza predictions [4] which simply uses a moving average from current data to predict case count for the following week.
2. Baseline Model – A simple baseline model for this problem can be implemented a Linear Regression model with the COVID-19 cases as a response variable and the viral titer measurements as the predictors.
3. Enhanced Models -
   1. GAM – Generalized Additive Models
   2. Gradient Boosting Classifiers
   3. Deep Neural Networks
      1. Gated Recurrent Units
      2. LSTM

# Additional Notes

We acknowledge the limited number of samples available as part of this dataset and its limit to Boston counties only. In the early discussion with Dr. Mauricio Santillana we were made aware that additional such datasets from around the world including Singapore and the Netherlands might be publicly available. We will work with the project mentors to include such datasets and refine our project proposal; however, those datasets were not available to us at the time of writing.

# References

1. Wu et al. SARS-CoV-2 titers in wastewater foreshadow dynamics and clinical presentation of new COVID-19 cases (2020 doi: https://doi.org/10.1101/2020.06.15.20117747**)**
2. <https://www.mwra.com/biobot/biobotdata.htm>
3. <https://www.mass.gov/info-details/covid-19-response-reporting#covid-19-interactive-data-dashboard->
4. Yang et al. Accurate estimation of influenza epidemics using Google search data via ARGO (2015, https://doi.org/10.1073/pnas.1515373112)