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

Tracking Spread of Covid-19 from Sewage data

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

# Background

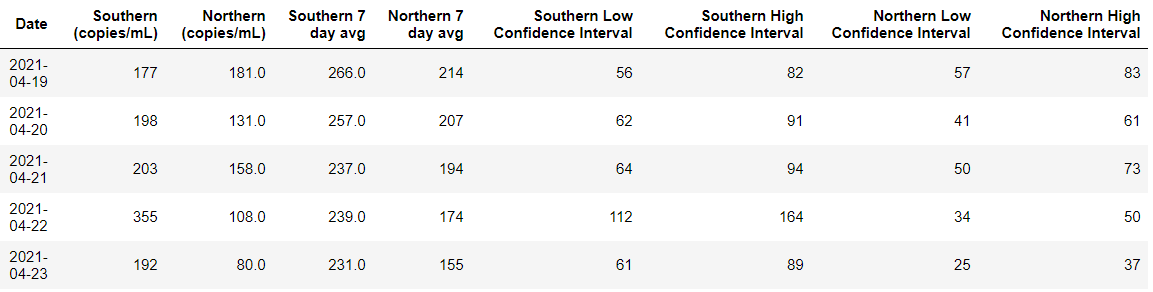
Conventionally COVID-19 is tracked via symptomatically and clinically diagnosed cases. This method has inherent drawback as many of the patients do not report the disease. In the paper titled “SARS-CoV-2 titers in wastewater foreshadow dynamics and clinical presentation of new COVID-19 cases” authors gathered data of presence of SARS-CoV-2, the virus responsible for COVID-19, in wastewater. The paper was published in <https://www.medrxiv.org/content/10.1101/2020.06.15.20117747v2> . The authors were successful in uncovering a trend that showed the presence of the virus in wastewater is a leading indicator as in 4-10 days the clinical data follows the wastewater data. The work suggests that longitudinal wastewater analysis can be used to predict COVID-19 cases more accurately than clinical data.

# Description of Data

We are using the data set in an excel file **MWRAData20210424-data.xlsx**. This data is collected by company called BioBots who specialize in detecting presence of viruses in wastewater.

* The measurements are collected daily, and data is collected from *Northern* and *Southern* counties around Boston.
* The dataset has more than one-year worth of daily data, with Southern county having 419 days and Northern having 417 days of data.
* The measurement is in **copies/ml**.
* For each of these southern and northern counties, the data also provides **7-day average, low confidence interval,** and **high confidence interval**.

## Table showing Sample data



# Project Question:

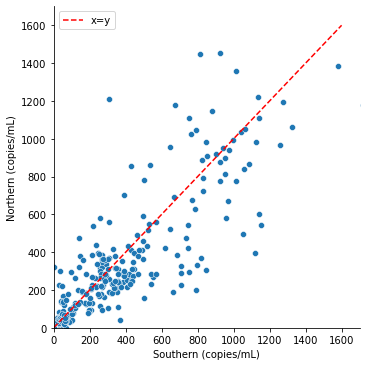
With the above explained dataset, the question we are trying to answer is:

**“CAN we figure out a relation between spread of Covid-19 and the contents of Sewage water? If yes, then HOW can we get a better short-term forecast of expected Covid-19 cases using most recent data from Sewage Water?”**

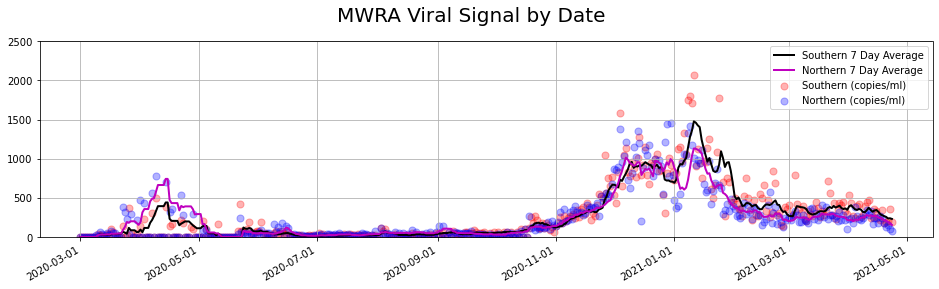
# Exploratory Data Analysis – Insights Gained

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

Based on the plot shown below, we can see a clear relationship between *Northern* and *Southern* data. This shows that, on any given day, if copies/mL are higher (or lower) in one county, it is higher (or lower) in the other county as well.

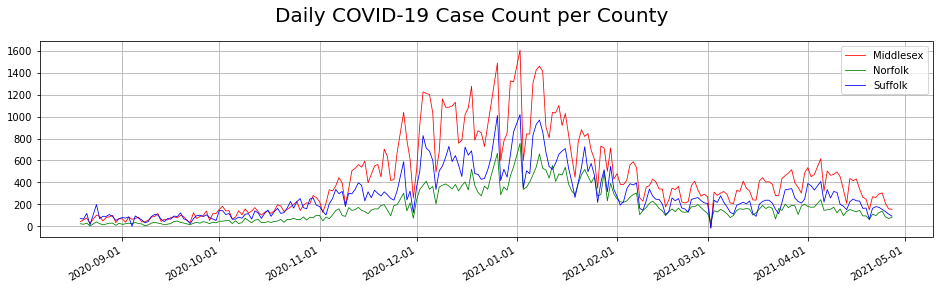


The above-mentioned point is also clearly explained in the plot shown below, where 7-Day average of Northern and Southern values seem to increase or decrease together.



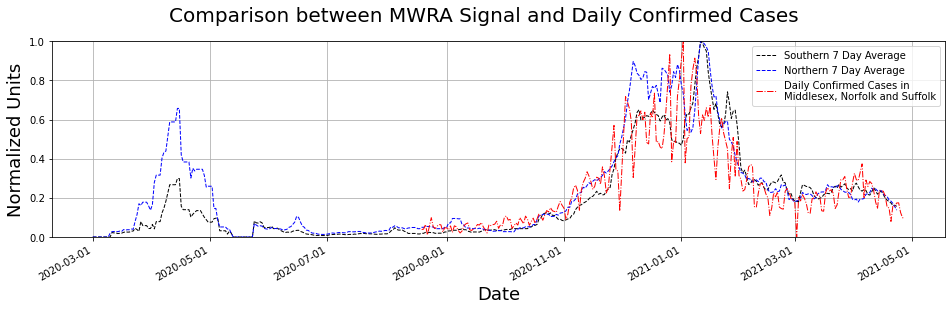
## Covid-19 cases per county

The following plot shows the Covid-19 cases per county. This plot shows a similar trend seen in the plot above for copies/mL, where there is an increase in the number of cases around the same timeline where copies-mL increased in the Sewage water.



## Comparing copies/mL to confirmed cases

The following plot compares both the plots shown above in one place which helps us to see a clear relation between the copies/mL in sewage water and Covid-19 cases.



# Model Implementation Plan

We propose to forecast the future increase/decrease in Covid-19 cases based on the past historical data. We would like to:

* Build various forecasting models using techniques like RNN, LSTMs, Decision Trees, GAMs, etc.
* Get new data from other counties, cities or states (if possible) to see if there is a similar/different trend and understand reasons behind them.

# References

1. <https://www.medrxiv.org/content/10.1101/2020.06.15.20117747v2>