**Introduction and Importance**

The CMU Delphi group supports and advises the CDC’s community-driven COVID-19 forecasting effort by collecting data about COVID cases, symptoms, and surveys across the United States. Delphi group is working to create alternative data sources for information related to the pandemic which can be leveraged by policy makers and researchers. Their curation of data signals, called covidcast, contains several resources for covid-relevant information and many of these data sources measure the same concept. However, Delphi has noticed some disagreement across different data signals that supposedly seek to measure the same concept. This could indicate a difference in data quality that could negatively impact forecasts or analysis that consumers of the covidcast API conduct. Thus, our role in this capstone project is to investigate data disagreements in the covidcast data suite and provide relevant insights into the reason behind this disagreement in order to improve the quality of Delphi’s data.

Our team will pick different pairs of signals perhaps from different sources and calculate the correlation between two signals across states or counties with various approaches: Exact numeric agreement, Spatial / temporal correlations, Rank correlation, Pearson correlation, and etc. Moreover, our team will do research on any other effective method analyzing the correlation between these signals and statistical investigation on data segmented by region or time.

We pick two pairs of signals: test positivity rate from CTIS and Quidel and COVID symptoms from “sore throat” from CTIS and Google Search. We use rank correlation to compare each pair of signals across states and counties. For two signals on test positivity rate, we find that there is less agreement in California and more agreement in Pennsylvania. Rank correlation between two signals on COVID symptoms indicates that there is less agreement on New England/MidWest and more agreement on West/Central. Some possible reasons for disagreements between signals across states and counties might be that:

1. Policy change: Democratic party has conducted several policy change from 2021 to 2022
2. Missing data: Much data in Test positivity rate from CTIS and Quidel is missing
3. Biasness - For small sample sizes, the response might be affected

Our next steps are to try more approaches to measure the agreement and disagreement between signals and to have a deeper analysis on why these disagreements happen.