Manna:

A massive data collection project led by researchers in CMU’s Delphi Research Group, COVIDcast data, is the main data resource of the project. The main question is that when two data signals are supposedly getting at the same concept, do they agree with each other? Professor Jacob Bien is an associate professor at USC and he is an expert in the field of data science and machine learning. He is a statistician, and he knows everything about the methodology of the project. The goal of the project is to investigate whether some signals like different symptoms of covid 19 are in agreement. The professor uses spatial / temporal correlations and perhaps rank correlation rather than Pearson correlation. There is so much COVID data and reports out there, and it is interesting to find effective ways to investigate how much each data source agrees or disagrees with each other.

Bhoomika:

* ● DELPHI is working to create alternative data sources for information related to the Covid-19 pandemic which can be leveraged by policy makers and researchers. For example, doctor visits, COVID-19 Trends and Impact Survey, Google Symptoms, CTIS, Quidel etc. These are housed in the covidcast package in R to make the data easily accessible.
* ● “When two different data signals are supposedly getting at the same concept, do they agree?”  
  ○ This question is motivated by the difficulty in assessing data quality. By comparing different data sources, we can use data consistency to assess the data quality. When two sources disagree, understanding nuances in the data signals can give us deeper insights.
* ● Focus is going to be on identifying effective ways to determine if two signals and hence two sources are in agreement. Suggestions - spatial/temporal correlation.

Zixuan Jin

Delphi Covid Project Presentation Report

The CMU Delphi group supports and advises the U.S. CDC’s community-driven COVID-19 forecasting effort. Professor Bien, a Delphi core member, mentioned that they gathered COVID-19 related data from different resources such as COVID-19 Trends and Impact Surveys (CTIS), Quidel, and so on. The main question here will be if two different data signals agree when these two signals are supposedly getting at the same concept. For example, we can test if the estimated percentages of people reporting they or someone in their household experienced sore throat in the past 24 hours agree with the search volume for “sore throat” symptoms on Google. Since the data came from various sources, we might consider which data is a better source and reflect real situations correctly. However, there is no “ground truth” in deciding which data source is better. Hence, it is difficult to access data quality. But, when two resources measure the same underlying concept, it is a unique opportunity to access quality via data consistency.

Moreover, Professor Bien pointed out that when two sources disagree, understanding why they are different can give us deeper insight into the data signals themselves, their strengths, and weaknesses. I believe the key ideas for this project are to find two or more similar data signals from different sources and see how they fit into each other using a specific type of statistical measurement. Professor Bien is the stakeholder for this project. Since he is a statistician, I believe he is well-informed of the technical knowledge needed for this project. The statisticians in this project should decide how we measure the agreement between two sources. There are a couple of methods such as exact numeric agreement, spatial/temporal correlations, rank correlation, etc. The statisticians could also think of new effective ways for getting whether two signals are in agreement. Even though the stakeholder has a deep technical background, the statisticians should still make the final presentation and report easy to understand and explicitly explain the statistical methods.

Ani

DELPHI

1. The primary focus of the DELPHI group’s project is around assessing quality and agreement of various data sources. Their COVIDCAST project has an assembly of various COVID-related datasets (e.g. positive cases or symptom frequency over time) where multiple datasets measure the same signal. However, the team has noticed that there is disagreement in the signal between certain data sources that supposedly measure the same variable. Thus, our goal is to quantify this agreement and use statistical analysis to understand any meaning behind the disagreement of various data sources. The general problem definition is to understand and quantify disagreement between different data sources that aim to measure the same concept through various statistical measures. The datasets are primarily publicly available sources that have been curated through the COVIDCAST project and are readily available with the group’s R package.

2. Jacob Bien is the primary stakeholder for this project and as a statistician on the DELPHI team with a phD in statistics, he has a deep understanding of the technical details needed to succeed in this project as well as the domain knowledge to provide qualitative advice on how best to succeed with the given task. Ranking and agreement methodologies are rather new to me so he will be a great contact for me personally to ground myself in this new area in preparation for the project.

3. The main issue that this project seeks to address is data quality. With several sources seeking to measure the same quantity, it is important for the DELPHI group to use the most accurate and well-balanced dataset in their COVID prediction dashboards and other services. Thus, understanding any data discrepancies that could cause quality issues in their modeling or forecasting is a critical task. Thus, our objective is to provide concrete statistical input into the quality and agreement of their various datasets in order to allow them to have the best possible understanding of the data they use for their models. Methodologies will likely include R (as their data is released and curated under an R package). Their expected postexperiment result is that we will indeed find disagreement (as they have noticed some already) and we will be able to find concrete explanations behind these discrepancies. They also believe that these explanations will allow us to provide recommendations into what data sources are more adequate for their relevant tasks.

4. The primary role of the statistician is to engineer and implement quantifications of agreement and use them to measure and interpret differences in their datasets for the data analysis step of the project (all of the data has already been collected, so we will not need to address this). Possible methods are Pearson’s rank coefficient or other sample statistics, as well as measurements of temporal/spatial correlation. The statistical problem framework is rather open ended, so it is the statistician’s role to workshop and evaluate possible approaches (e.g. simple statistics, ranking models) and determine which ones are the best to use. The statistician will also need to understand and interpret these methods in order to provide appropriate insights. DELPHI is interested in the statistics behind the agreement as long as they also provide meaning into why certain metrics behave differently, so us as statisticians must be prepared to interpret and understand the results of the quantitative analysis.

Yueni Wang:

In the project, Jacob Bien for Delphi is concerned about indicators for covid agreement. As a project concerned with main factors that enhanced COVID-19 spread rate, the team collected massive data from an R package called covidcast. The main problem that was explored is to evaluate the agreement of two signals from COVID-19 data that mainly evaluates the same concept. This approach of data streaming on massive covid-19 data is considered an open repository of covid-19 indicators, providing real-time measurements on covid-19 trends, symptoms, risks and vaccination. The technical knowledge of this project is demonstrated through the analysis of time series data and the readings of different data signals. The interpretation of two similar outcome from signals that are pointing to one concept would evaluate the data quality which is hard to measure. When two sources disagree, it is not a bad sign because we want to understand why the signals performed differently. Statisticians are important in analyzing the data and explaining why the signals are different. There are many reasons that two signals don’t agree with each other. Agreement means exact numerical agreements, space correlations, and rank correlations. They must look into practical examples and see what is the actual reason behind.