**Outline**

**Topic:** Statistical Analysis of COVID-19 in the US

* **Tasks:**
  + Create a model to forecast COVID-19 in the US; this will need to be a nonlinear model so we can fit it to the whole dataset rather than only the linear parts
    - Models to consider: Random Forest, Gradient Boosting, Bayesian regression
    - **Specific quantities to forecast**
    - **Specific data on which forecasts will be based**
    - **Each model will have its unique parameterization**
  + Create models to analyze COVID-19 in different US states. States that could be of major interest are California, New York, Georgia, Florida, Texas, North Carolina, and Washington.
    - These models will most likely need to be nonlinear as well, I will need to do some exploratory analysis for each state
    - Analysis could be done to examine the growth rates over certain periods of time and compare the growth rates of different states in those timeframes
    - These models can take into account several different things:
      * The different health guidelines that each state implemented should be heavily considered, for example California has an extensive and quite strict reopening plan whereas Georgia had very few regulations which caused a massive outbreak
      * The number of people who live in each state
      * State consists of big cities or smaller cities
      * **Movement of people across states.**
      * **Superspreader events**
    - **Same notes as above: in addition will be interesting to compare an “aggregate US” model to a “high resolution state-level” model**
    - **Do we need to distinguish between “essential workers” who have to go out and “non essentials” who can work from home?**
* **Potential questions to answer:** 
  + Based on the COVID-19 forecast, how does the future trajectory of COVID-19 look in the US?
  + Based on the number of cases/deaths in various US states, did those states do a sufficient job of containing COVID-19?
  + For the states which had large COVID-19 outbreaks, is there a pattern of what they did that could have led to the outbreaks (ex. Florida, Georgia, California)? What other factors could have led to an outbreak in these states?
  + For the states which did a sufficient job of containing COVID-19, is there a pattern of what they did which could have helped to contain COVID-19 (ex. Washington, New York)? Are there any other factors that could have helped them contain the spread?
  + For the states which had large outbreaks, what could they have done differently to contain COVID-19?
  + Based on the particular state being analyzed and the current timeframe, is the current number of COVID-19 deaths and/or cases reasonable? Or should there have been less cases/deaths based on the attributes of the state? For example, based on the population and strict orders implemented in New York, is the current number of cases/deaths reasonable? Or would it be expected that the number of cases/deaths in New York should be higher or lower than what it currently is?
  + **The dynamics of restrictions: on, relaxed, off, back on, different types of restrictions (stay home, mask, etc).**

**Information from Papers**

Corona Geographic Signals

* Input factors were latitude, longitude, age, and sex. Models were between 87.44% and 90.83% accuracy
* Decision Tree model found that latitude and longitude are important factors
* These finding could give us more info about the role of environmental factors and the spread of COVID-19 based on latitude and longitude

Random Forest SIP Orders

* Analysis revealed that the March 16 presidential recommendation lowered the compound growth rate of COVID-19 for all counties in the US by 6.6%
* Random forest model was used to predict compound growth rate after a SIP (shelter in place) order and had an accuracy of 92.3%
* SIP orders were found to be effect at reducing the growth rate of COVID-19 cases in the US counties
* Most important features that the random forest found were population, longitude, and population per square mile

Corona Masks DC

* Explores evidence from a natural experiment on the effects of state government mandates in the US for face mask use in public, issued by 15 states plus DC between April 8 and May 15
* Masks clearly caused a decrease in COVID-19 cases

Corona Italy Excess Mortality

* Analysis shows good agreement with reported COVID-19 mortality for age < 70 years, but an excess in total mortality increasing with age above 70 years, suggesting there is a large population of old people missing from the official fatality statistics
* COVID-19 mortality in Italy is more than a factor of 2 higher than the official number

Corona LA Serology

* Estimated that 4.1% of LA county’s adult population has the antibodies
* With regard to statistical margin, 2.8% to 5.6% of the county’s adult population has antibody for the virus (221,000 to 442,000 adults had COVID-19)

Corona Santa Clara Serology

* 2.81% population-weighted prevalence for COVID-19 antibodies
* Ranged from 2.49% to 4.16% (95% CI)

Corona Lockdown Model Timing

* Nationwide, 61.6% (95% CI: 54.6%-67.7%) of reported infections and 55% (95% CI: 46.1% - 62.2%) of reported deaths as of May 3rd, 2020 could have been avoided if the same control measures had been implemented one week earlier
* Longer response time for reimplementing social distancing rules results in a stronger rebound of infections and death
* Early intervention and aggressive responses are essential for controlling the COVID-19 pandemic

**Initial tasks that need to be completed**

* Look at word doc in email for SIR models
* Drill down into latitude and longitude
  + US specific? Worldwide may complicate analysis
    - I think the focus of this research/thesis should definitely be on the US
* Think about what dataset we are going to use
  + Time series by county (John’s Hopkins)
  + Population data by county (census)
  + Geographic data for COVID-19
  + What kind of public health orders were implemented? County by county is ideal but statewide will work
    - Sequence of times where orders were implemented
* Imperial College in UK for modeling
* Review modeling approach/read papers (50%) and collect data sources (50%)

**Data**

* JHU time series:
  + <https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data/csse_covid_19_time_series>
* More time series, including number of new cases daily and a full dataset:
  + <https://ourworldindata.org/coronavirus-source-data>
* WHO COVID-19 data table:
  + <https://covid19.who.int/table>
* Population data (Census):
  + <https://www.census.gov/data/datasets/time-series/demo/popest/2010s-counties-total.html>
* **Public health order data and timelines:** 
  + WHO timeline: [https://www.who.int/emergencies/diseases/novel-coronavirus-2019/interactive-timeline?gclid=Cj0KCQjwtsv7BRCmARIsANu-CQckUr\_tzaFi3O2DiBIXJwyGGkLI\_77sRaC7Z\_VWvblZrcfiq53v-PQaAoFCEALw\_wcB#](https://www.who.int/emergencies/diseases/novel-coronavirus-2019/interactive-timeline?gclid=Cj0KCQjwtsv7BRCmARIsANu-CQckUr_tzaFi3O2DiBIXJwyGGkLI_77sRaC7Z_VWvblZrcfiq53v-PQaAoFCEALw_wcB)
    - This WHO timeline covers the worldwide actions taken on COVID-19
* Imperial College has some interesting data as well. A couple datasets show the number of deaths and how many hospital visits within different age groups in the USA, sorted by state: <http://www.imperial.ac.uk/mrc-global-infectious-disease-analysis/covid-19/covid-19-scientific-resources/>
* **Top 10 populated states data**
  + Top 10 populated states: <https://worldpopulationreview.com/states>
  + Wikipedia has detailed COVID-19 timelines for all states in the US; sources seem to be trustworthy (WHO, CDC, NY Times)
  + 1. California (compare sources)
    - SF Bay Area timeline: <https://abc7news.com/timeline-of-coronavirus-us-coronvirus-bay-area-sf/6047519/>
    - California timeline until July 1: <https://www.usnews.com/news/best-states/california/articles/2020-07-01/timeline-of-events-during-californias-coronavirus-outbreak>
    - Detailed California timeline: <https://en.wikipedia.org/wiki/COVID-19_pandemic_in_California>
  + 2. Texas
    - <https://en.wikipedia.org/wiki/COVID-19_pandemic_in_Texas>
    - <https://thetexan.news/covid-19-in-texas-a-timeline-of-gov-abbotts-executive-orders-and-case-numbers/>
    - <https://www.texastribune.org/2020/07/31/coronavirus-timeline-texas/>
  + 3. Florida
    - <https://en.wikipedia.org/wiki/COVID-19_pandemic_in_Florida>
    - <https://www.clickorlando.com/news/local/2020/03/20/timeline-the-spread-of-coronavirus-in-florida/>
  + 4. New York
    - <https://abcnews.go.com/US/News/timeline-100-days-york-gov-andrew-cuomos-covid/story?id=71292880>
    - <https://en.wikipedia.org/wiki/COVID-19_pandemic_in_New_York_(state)>
    - <https://www.investopedia.com/historical-timeline-of-covid-19-in-new-york-city-5071986>
  + 5. Pennsylvania
    - <https://www.inquirer.com/health/coronavirus/inq/coronavirus-cases-pennsylvania-philadelphia-nj-us-timeline-20200918.html>
    - <https://en.wikipedia.org/wiki/COVID-19_pandemic_in_Pennsylvania>
  + 6. Illinois
    - <https://en.wikipedia.org/wiki/COVID-19_pandemic_in_Illinois>
    - <https://medium.com/gdgf/timeline-how-covid-19-unfolded-in-illinois-fc8d124ae033>
  + 7. Ohio
    - <https://www.sidneydailynews.com/news/183998/a-timeline-of-covid-19-in-the-u-s-ohio>
    - <https://www.news5cleveland.com/news/continuing-coverage/coronavirus/local-coronavirus-news/interactive-covid-19-timeline-ohios-health-orders-big-events-and-covid-19-cases-visualized>
    - <https://en.wikipedia.org/wiki/COVID-19_pandemic_in_Ohio>
  + 8. Georgia
    - <https://en.wikipedia.org/wiki/COVID-19_pandemic_in_Georgia_(U.S._state)>
    - <https://ballotpedia.org/Documenting_Georgia%27s_reopening_and_path_to_recovery_from_the_coronavirus_(COVID-19)_pandemic,_2020>
  + 9. North Carolina
    - <https://en.wikipedia.org/wiki/COVID-19_pandemic_in_North_Carolina>
  + 10. Michigan
    - <https://en.wikipedia.org/wiki/COVID-19_pandemic_in_Michigan>

**Modeling**

* Autoregression, gradient boosting, random forest, Bayesian regression could potentially work; traditional ML models have a bit of a caveat
* How to take into account the population, social distancing orders, etc. in the models
* Forecasting: Tree-based models won’t work as well since they cannot extrapolate, however, there is a way to transform time series into data that can be used in a supervised learning problem for tree-based models
* I think a more ideal model for forecasting could be autoregression, ARIMA, or Bayesian regression
* Growth rate comparison: Linear regression, random forest, gradient boosting could all potentially work since we would most likely pick out certain timeframes to explore the growth rates

**Tasks for next meeting**

* Research top 10 most populated states and collect data on them
* Research Bayesian autoregression for modeling
* Look at Shea paper (R shiny app)