**COVID Challenge**

# Plan

03/29/20- 04/04/20

1. Clean up the data (text preprocessing, preparing the abstract text and title) – Panayiotis
2. Identify the knowledge base: what we are interested in
   1. Define the details of the knowledge base (step 1) – Johnny, Panayiotis
   2. Define the keyword map (step 2) - Johnny
   3. Implement and classify each paper into different knowledge pool (step 1&2) – Johnny
3. Extract the information (step 2)
4. Visualize (step 3)

# Overview

**Step 1: knowledge base: what are we interested in and looking at in the literature**

1. Identify a list of subjects of interests from stable country+time-point and then find information of them globally along different time points
   * Approach: we look, for example, Wuhan, papers and then checkouts out the papers that published after the diseases is stablished/death rate start to decrease, and identify the list of risk factors
2. Subjects of interests
   * Risk factors: factors that increase the risk for the disease
     + Age, disease comobility, habit (smoking, exercise), personal social risk factors (e.g., occupation), government-driven social risk factors (e.g., containment measure), demographics (e.g., weather, transpiration), gender
   * Disgnostics
     + Imaging variables related to covid-19 (clinical findiings the MR/CT imaging)
     + Physical appearance symptoms of covid-19 (dry cough, fever, lose of sense and smell)
     + Chemical testing for disease
       - RNA test performance (time, cost, effective, fpr, tpr)
   * Treatment & vaccines
     + Medicines that we used, vaccines that we used, medical procedures on the diseased patients.
     + Clinical trial
       - What are the clinicial trials initiated in the country, how effective, what stage.
   * Patient outcomes
     + Death rate, recovery rate, outcome measure (mobility, lung capacity), length of survival, length of recovery rate, length of hospitability, covid-19 induce clinical complexity (indice some of the disease like sepus)
3. Approach: classify each paper whether they contains subjects of interests information (multi-label classification)
   * 1-level: keyword match (synonym)
   * 2nd-evel: machine learning model
   * **Data source: only title and abstract at this moment**

Step 2: Subject of interest information extraction

1. Information (give us some y-axis number, 0/1, or a scale, or whatever)
   1. Risk factor, diagnostics, treatment & vaccines
      1. Positively or negatively or neutral to the risk for the disease
   2. Patient outcomes
      1. Fact
2. Meanwhile, we need to store different paper information (meta data)
   1. All the paper will have this meta information
      1. Timestamp: when the paper publishes, when the paper is submitted, the time range of the data (e.g., “we sample the data from Jan 1 to Jan 31”), origin of the data (e.g., China, Italy), Demographic information, gender

Step 3: Data Visualization in different dimensions

03/28/20 Ideas

Goal and method

* Separate information into different country
* And subgroup by important clinical information: (may be key word match for now, later we can use smarter system)
  + Age
  + Gender
  + Clinical history Etc
* We want to present the scientific finding crossing different timepoints and different dimensions (e.g., age) and identify if there are aligned information or missing information so that we can guide the next important research questions and public health guidance.
* This will clear out whether some papers, although they are scientifically sound, they may not represent the truth because of different factors
* We can also observe the temporal change of variables (e.g. risk factors) along time and compare it with the death trend and outbreak.

Details step

1. Classify the paper into multiple domain information, like whether they provide information about coronavirus in iddfnerent areas. We probably need to define an ecology type of map to define how many layers of information we need, etc. At first, we can do keyword match, but later can use machine learning to do a better job
   1. Risk factor
   2. Medical condition
   3. Demographic infomration
   4. Weather, etc
2. Once identify the papers into different category (can be a multi-class & multi-label classification), then we need to start doing the information extraction. Like we can have different level of understanding of the information. For example, the basic level can be “age is postively/negatively impacting the coronavirus survival rate” , “an athletic is postiviely/negatively impacting the coronavirus survival rate” <— this is still pretty abstract to me. We need to define what we are looking for, so that we can have a good summary of the finding. Well, although, the basic version we have can just link the paper to a time point and let the reader to read it
3. Once the information is extracted, we can present them in time-series plot. Each time series plot represent a category we define in step 1. Like risk factor. The y-axis will be the scale we define in step 2 (positive or negative at least or in scale), and x-axis will be timestamp. These time series plots can be grouped by counties because seem like countries research finding does matter. (Like the virus behave differently in different region).