**COVID Challenge**

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

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 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).