Project Proposal

Title: Mining Climate Data

Team members: Kamen Shah, Jacob Reed, Jimmy Young, Steve Su

Description: In our project we aim to verify or reject hypothesis made around the scientific community relating to global warming. One such example we aim to test is the claim that infectious disease will spread as a result of global warming. Through mining both global warming and global health data sets we can find if there is any statistical correlation between the two domains.

Prior Work: Global warming data has been explored extensively in regards to the adverse effects it has had on various facets within our world. Global warming has been correlated to an increase in health problems, coastal flooding, wildfires, and other extreme weather events. According to Annalisa Bracco, a professor at Georgia Tech, many of these prior modules rely heavily on human input than on the actual data.

Datasets:

- Historical data of global temperatures
  - https://www.kaggle.com/noaa/noaa-global-historical-climatology-network-daily
  - https://www.kaggle.com/noaa/ghcn-d
- Historical data of carbon monoxide levels
  - https://www.kaggle.com/epa/carbon-monoxide
- Historical data of human health statistics
  - https://www.kaggle.com/theworldbank/health-nutrition-and-population-statistics
- Since we are using multi-gigabyte files, we will store them locally on our hard drives.

Proposed work: what do you need to do?

- **Data cleaning:** Remove any objects that have missing data fields
- **Data preprocessing:** Remove outliers and fix data that has incompatible type, ie:) changing 370 deg to 10 deg. Since we will be merging multiple datasets to create a larger dataset which we will mine; we need to change the “time” or time related feature to match. An example of this would just just keeping the year from each dataset.
- **Data integration:** Since we have a common variables formatted correctly we can now combine the datasets based on time.
- **etc.**

List of tool(s) you intend to use : Delta Maps, Pandas, Numpy, Scipy, matplotlib

Evaluation: Since the patterns we found are due to correlations we can test using cross-validation. Using a set of data that we set aside earlier which our model didn’t encounter. Unfortunately based on previous studies done at other universities just because our model may predict correctly doesn’t mean that all the features of our model found are all correlated to the prediction.