# Proposal

Names/NetID:

- Justin Chan: jachan
- Tambre Hu: thu53
- Oat Sukcharoenyingyong: sukcharoenyi
- James Ma: yma255
- Shawn Riemer: seriemer

## Dataset

URL: https://www.kaggle.com/kaggle/world-development-indicators

This dataset is a collection of some of the indicators of economic development the World bank collects. We think this dataset is good because it is a rich dataset so there would be plenty of interesting patterns to find. Also, the conclusions will be grounded in measures of economic development (i.e. "Adolescent fertility rate negatively correlates with GDP"). Since the dataset is so rich we could find both simple and complex patterns (to use more advanced techniques).

## Statistical Questions

1. What interesting or strong correlations between variables are there?
2. What kinds of claims can we make from this data? How can we confirm them?
3. What is the largest indicator of economic growth?
4. What is the largest indicator of economic stabilitiy?
5. Can we find any intersting groupings using GDP as the distance metric? What about other indicators?
6. Is there any way to measure geopolitical or geographical patterns? (A form of clustering)

## Variables

This dataset has 118 columns for each country (so ~200 rows). There are a lot of interesting colums such as:

- SystemOfNationalAccounts (economic activity)
- Adolescent fertility rate (births per 1,000 women ages 15-19)
- Age dependency ratio (% of working-age population)
- Birth rate, crude (per 1,000 people)
- CO2 emissions (metric tons per capita)

It's hard to choose what variables we think are the most important right now because we are unsure what the most significant indicators are of economic growth, success, or stability. However, we feel confident that these inicators *will* result in interesting patterns that will be more than sufficient to do analysis on.

# Methods

- Monte Carlo: We want to use Monte Carlo simulation to create confidence intervals of our estimates of our data. Knowing the mean, variance, etc. of our data would be quite useful, and we could use Monte Carlo to create these estimates.
- Clustering: Clustering by different distance measures and seeing how they compare. Some intersting ones we will try first will be using GDP and metrics that have clear correlations (like high inflation).
- Regression: Use regression to fit models to our data, experimenting with different options (linear, logistic, multiple, etc.) to discover relationships between different variables.
- Correlation Matrix: We think that cleaning the data and then running a correlation matrix that includes all the data will be a good way to start exploring the data. This is especially true since this dataset has so many variables from various domains of knowledge (economics, global health, environmental sciences, etc.) so it might be easier to filter out the important variables this way.