# CIS 545: Project Proposal

## Team members

Sisun Cheng, Shimin Tu, Lan Xiao

* Sisun Cheng: Exploratory data analysis
* Shimin Tu: Time series analysis
* Lan Xiao: Classification model

## Data Source

[Health Nutrition and Population](https://www.kaggle.com/theworldbank/health-nutrition-and-population-statistics)

## Project Plan

As population aging becomes one of the most critical issues in many developed countries while developing countries can not afford their growing population pressures, we expect to study and predict the birth population and the degree of aging for each country.

1. Objective:

1. First and foremost, we would like to **predict birth rate** for each country with **time series models (eg. ARIMA)**, considering that there is a demographic cycle.
2. Also, we will judging aging levels for countries based on UN standards, and **predict 10-year-later aging levels** by **classification models** (eg. Decision Tree, XGBoost, Random Forest) and various predictors.

2. EDA:

1. Cycle/Time Pattern of birth rate for each country, White Noise test
2. Correlations - aging level

## Why Project?

Nowadays, a lot of countries, especially developed countries, are facing the problem of low birth rates, while some developing countries are suffering from high birth rates. We will explore the contributing factors of birth rates, which will be used to predict the future situation. Since population aging is to some extent related to low birth rates, we are also going to predict the level of population aging five years from now based on the prediction outcome of our first model. According to the prediction of population aging, governments are able to put forward reasonable policies to respond to the lack of medical resources and economic stagnation in the future or even slow down the aging of the population.

## Challenges

1. **Identify useful indicators.** Given that there are such a huge amount of indicators in our dataset, we need to carefully review the data and dive deep into the EDA process to identify useful predictors for our time series analysis and prediction model.
2. **Deal with null values.** Due to complex dimensions of our dataset, many records are missing and we have to carefully deal with the NAs in our dataframe. What we need to do is to classify and take use of the existing predictors for different countries.
3. **Identify appropriate classification models.** There are quite a lot of machine learning models that are commonly used for classification, including Decision Tree, Random Forest, SVM, etc. We need to find out the most appropriate model with a satisfying model accuracy.

## TA Request

Craol Li

Ernest Ng

Kailin Zheng