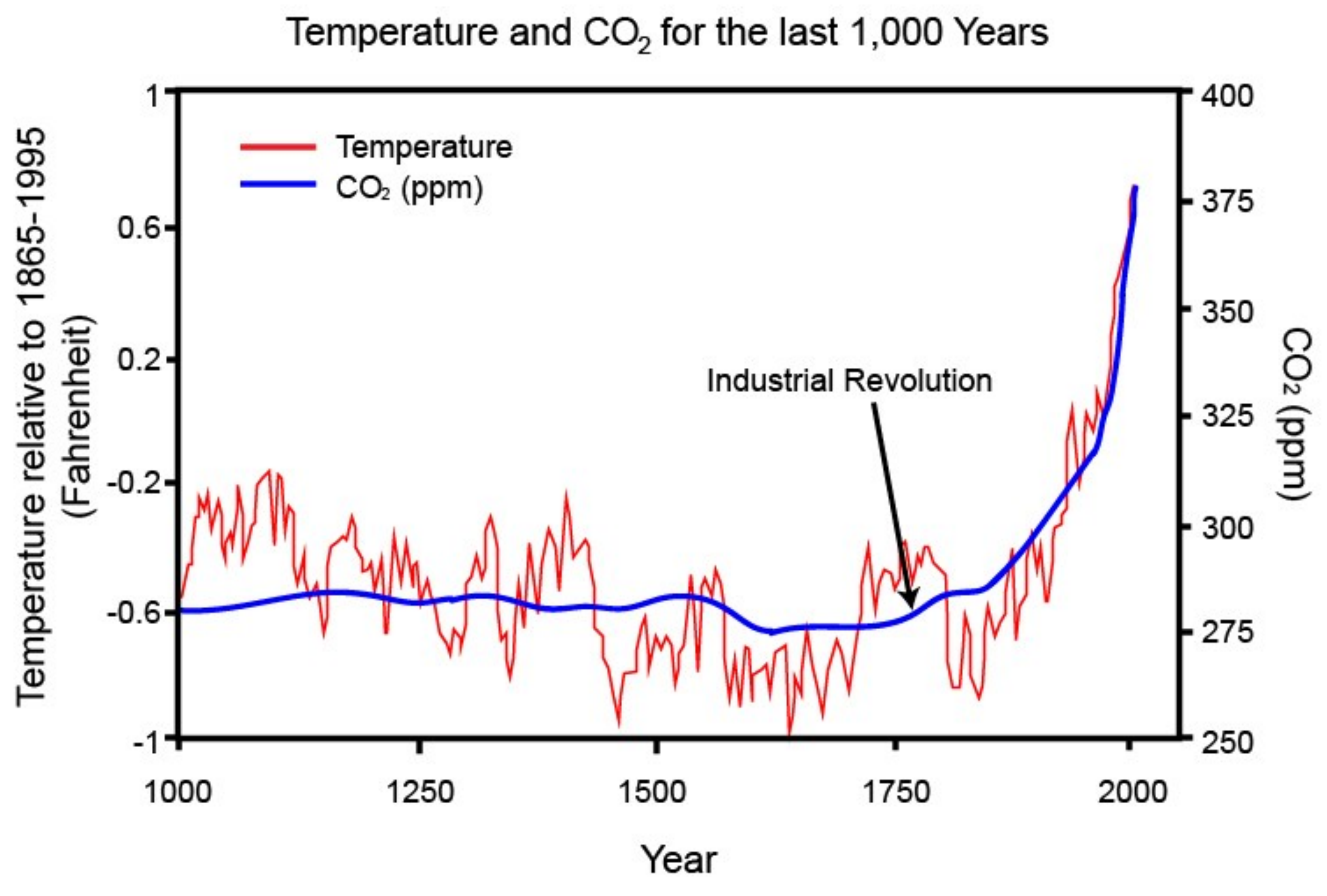
CS591 Lapets
Boston University

Modeling CO2 Emissions by Country

Yuchen Yuan
Maulik Shah

Motivation

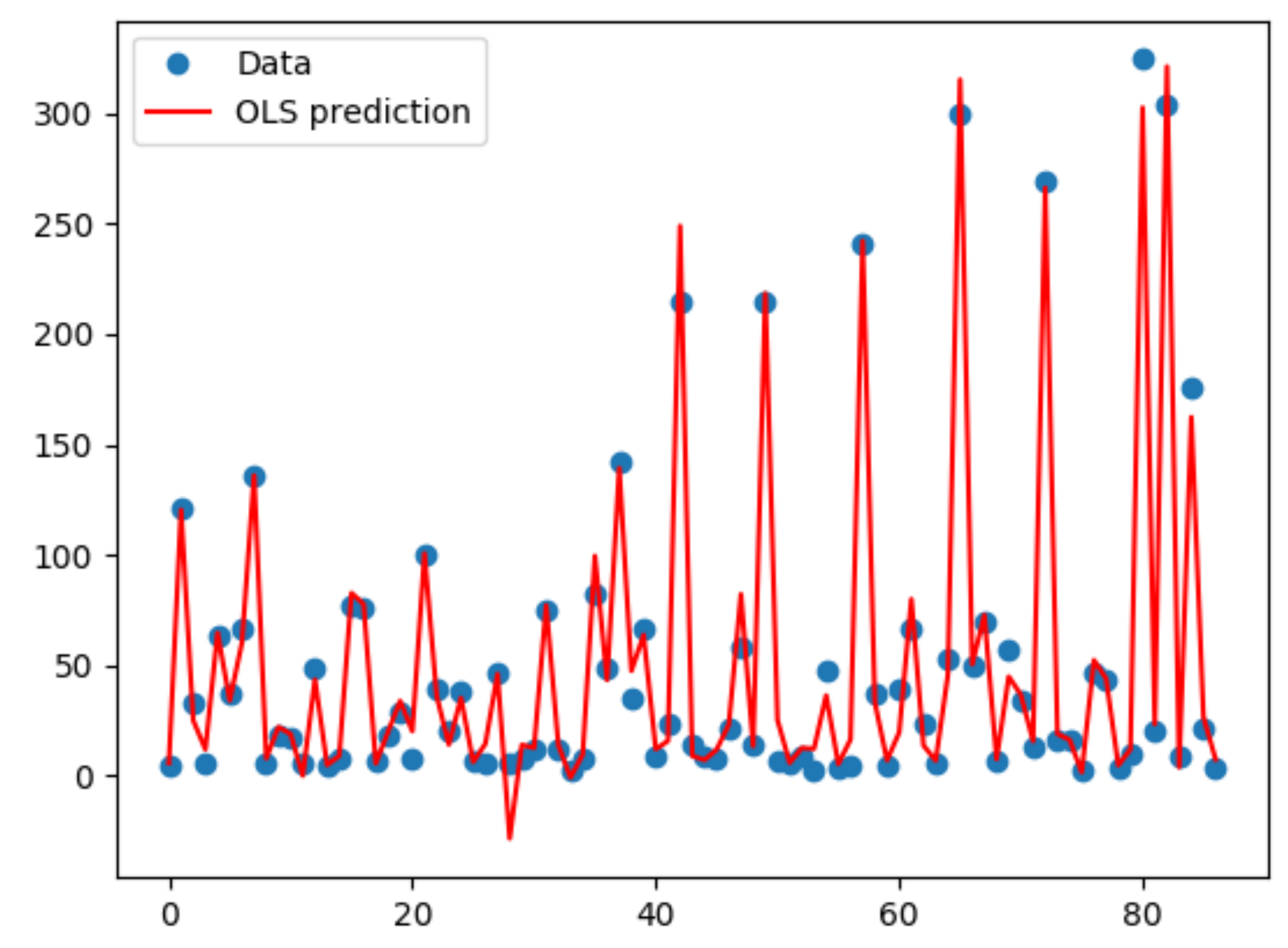
There has been a rapid increase of CO2 concentration observed in earth's atmosphere. It is the most abundant greenhouse gas in our atmosphere. We can see an increase in atmospheric temperature following the same trend as CO2 concentration. As global leaders start to address the issue and implement policies and technologies, we want to come up with an easy yet accurate way to model CO2 emissions.



Source: https://i1.wp.com/timescavengers.blog/wp-content/uploads/2017/06/1000-2000_co2temp-01.jpg?ssl=1

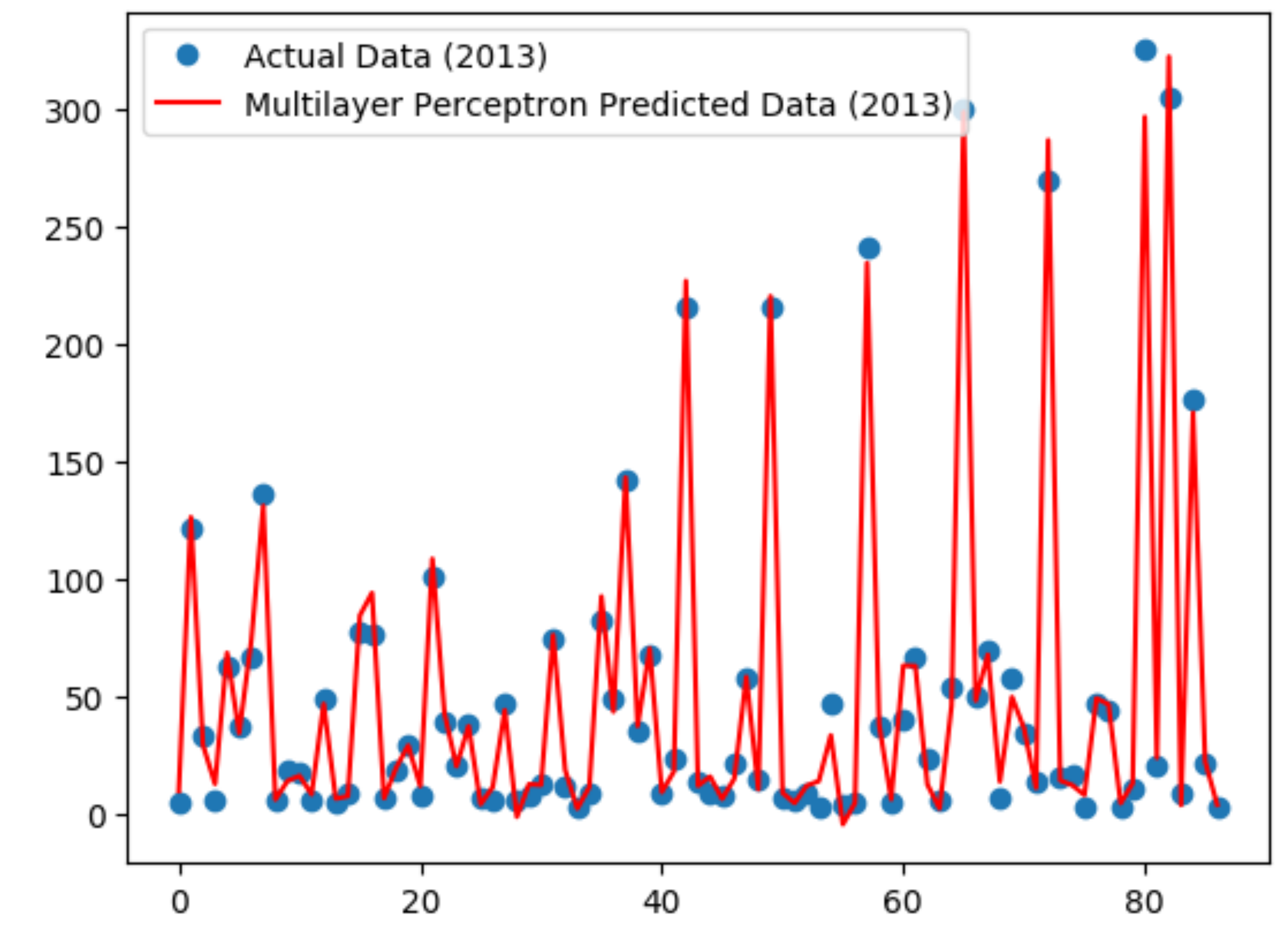
Method 1 – Regression Analysis

We tried using OLS Regression to model CO2 emissions with outliers and without outliers. We trained the model on data from 2012 and used it to predict values for 2013 and compared the predicted results to the actual results. The model works fairly well on both the datasets, which shows it generalizes well. The model was built by adding one variable at a time in descending order of their R-squared values of the linear model of CO2 emissions against just that variable. We based our formula on the kaya identity which states that CO2 emissions is a result of relation between Population, GDP per capita, carbon intensity and energy intensity. With the clean data we were able to achieve an R-squared value of 0.998919 which shows the model generalizes well for each country and that the variables we chose strongly affect the CO2 emissions. The results can be seen here:



Method 2 – Multilayer Perceptron

In method 2 we used a machine learning model called MLPRegression which generates a feedback neural network to train the model on the given data. To prevent overfitting, we had a regularization factor of 0.01 and we used n^5 hidden layers (where n was the no. of variables). We capped the model at 1000 iterations because there's not much to be gained from running it for longer time since it performs fairly well with this many iterations. The learning rate was 0.001 and momentum was 0.4. The input data was scaled using the StandardScaler and then both the training and testing datasets were scaled using the same scaler to maintain uniformity. Even without convergence, we got an R-squared value of 0.9978 which is pretty accurate and fairly quick. Ideally if allowed to converge this could go higher and tends to ~1. The results can be seen here:



Data Retrieval and Preprocessing

We were able to obtain data on CO2 emissions by country from the **Energy Information Administration** website: https://www.eia.gov/beta/international/data/browser/#/?pa=00000000000000000000000002&c=ruvvvvvfvtvnvv1urvvvvfvvvvvvfvvvou20evvvvvvvvvnvvuvo&ct=0&vs=INTL.44-8-AFG-MMTCD.A&vo=0&v=H&end=2014
Then we researched about factors that can either contribute to or be an indicator of CO2 emissions, and we downloaded data on GDP per capita by country from the **United Nations Development Programme** data portal: http://hdr.undp.org/en/data
Data on carbon intensity (kg per kg of oil equivalent energy use), energy intensity (MJ/$2011 PPP), energy use (kg of oil equivalent per capita) and total populations by country were obtained from **The World Bank**: https://data.worldbank.org/
We manually cleaned up all the data because it would be difficult to perform cleanup in python since there are different ways to write some country names. We chose data from 2012 and 2013 because they were fairly comprehensive and recent. We organized the data and uploaded them in json format to http://datamechanics.io/ in order to access them.

Conclusion

Comparison of methods In Method 1: In OLS we had to input one variable at a time based on individual R-squared values which introduces some complexity into building the model and can lead to error. It does lead to fairly good results of an R-squared value of 0.9989 but the programming complexity is huge. In Method 2: In Machine Learning Method (MLPRegression) as long as the data is scaled the neural network does most of the heavy lifting and all we had to do was set the parameters for the model which yielded a fairly good results with an R-squared value of 0.9978 even without convergence. Both of these methods worked very well in predicting 2013 CO2 Emissions. We conclude that we have chosen good indicators of CO2 emissions.