**The Dataset**

In this assignment, we will use data taken from [the Gapminder organisation.](https://www.gapminder.org/data/) The organisation collects and maintains global datasets related to important macro socioeconomic variables such as education, healthcare, environment change, etc. In this assignment, we will use the carbon dioxide emissions dataset to predict the factors that contribute the most to global emissions.

The dataset (attached below) contains various files each representing a variable that, you suspect, can contribute to (and thus predict) carbon dioxide emissions. For example, the file 'cars\_trucks\_and\_buses\_per\_1000\_persons' contains the number of cars, trucks and buses per 100o persons for each country for a number of years. Similarly, the file 'industry\_percent\_of\_gdp' contains the industrial GDP of each country as a percentage of the total GDP. You can look at the data dictionaries of all the files on [the Gapminder website here](https://www.gapminder.org/data/) (use the table at the bottom of this page). Some other important predictor variables are (look at the website for units of measurements etc.):

* forest\_coverage\_percent: The percentage of land area covered by forests per country
* oil\_production\_per\_person: Oil production per capita per country
* electricity\_generation\_per\_person: Electricity production per capita per country
* etc.

**The Task**

In the dataset, there is a file named 'co2\_emissions\_tonnes\_per\_person.csv' which contains the per capita CO2 emissions of various countries across many years. This is your **target variable.**Your task is to build a regression model to predict per capita CO2 emissions using all the other variables as predictors. The objective is to understand how much each predictor constitutes to the global  CO2 emissions.

Note that all the files contain information across a number of years. Your task is to **build a model only for the year 2014** (that's the year most of the data is available). Also, note that **each country will be one data point.**

**Data Preparation and Modelling**

First, import all the files (target and predictor variables) and **filter them for the year 2014.**Then merge them using 'countries (or geo)' as the common column. Get a data frame where each row represents a country and each column represents a variable (for the year 2014), including the target variable. Finally, after appropriate data cleaning and preparation, build a model **using Lasso for feature selection.**Interpret the results and conclude which variables contribute the most to global emissions.