# Cardiovascular Death Rate Capstone: Data Wrangling and Cleaning

Zach Wimpee

November 5, 2020

## 1 Overview

The primary goal of this project is to build a model that predicts county level changes in annual cardiovascular death rate. To accomplish this, a data set is needed that contains county level information on population demographics and cardiovascular mortality counts over a range of years. This document describes the construction of such a data set using the U.S. Census Bureau API and the CDC WONDER databases. The associated code is available on GitHub

## 2 CDC Data

The first thing needed to construct this data set is county level cardiovascular death rate data over some range of years. The CDC WONDER online databases provide this information through a convenient interface. Using the interface, a request was made for cardiovascular death count for every county in the United States from 2010 through 2018. The response was a tab-delimited text file in which each row corresponds to a single county for a given year, and records both the county population and total cardiovascular related mortality count. This file is saved as `data/cdc_data.txt` in the GitHub repository linked above.

# 3 Combining with Census Data

The CDC data provided the information used to calculate the crude cardiovascular death rate (CDR), and equivalently the change in county CDR between years. Data is needed that can serve as the feature space over which CDR change will be predicted. The U.S Census Bureau has a number of APIs that can provide social and economic data at the county level. Two sets of variables were requested for all counties from 2011 through 2018 using ACS 1-year Comparison profiles data. The resulting data sets were then combined into a dataframe, and saved as `data/cdr_data.csv`. These procedures were performed in the notebook `data_wrangling.ipynb`