# 1 Data

For this project, we decided to use real life financial market data aggregated from the Yahoo Finance website. It provides a more interesting result and tests the accuracy of the algorithm. We begin by describing the conditions imposed on the data before collection, our data-mining algorithm, and lastly what methods were used to deal with missing data.

## 1.1 Collection

Below is a list of criteria and their justification chosen to fulfil the project requirements.

1. Years 2008-2012
   - the stock data has daily open and closing prices from Jan 2008 to Dec 2012
   - the time period was chosen to avoid any large financial movements in the market (ex. the 2008 financial crisis and the 2007-past bubble before)
2. Stock by Sector
   - stocks were organized by the following sectors: Basic Materials, Conglomerates, Consumer Goods, Financial, Healthcare, Industrial Goods, Services, Technology, and Utilities
   - this was done since the MVP algorithm would create a portfolio from each sector and then create a portfolio from those portfolios

## 1.2 Mining Algorithm

All the data was collected via a web-scraping algorithm programmed with R. The central idea of the program is to determine the URLs associated with each Yahoo Finance sector page with those the program can identify the stock symbol by inspecting the HTML content. See stocklist.R in the appendix.

After a list of stock symbols were created, the program then uses the function *getYahooData* from TTR: (Technical Trading Rules) package in R to collect the historical prices. This process yields in total approximately 25,000 stocks.

## 1.3 Missing Data

Unsurprisingly, there was missing data in some of the stock historical prices. This could be a due to a variety of factors:

- Stock was split during some time in the year
- Stock was newly created during some time in the year

- Different exchanges operate in different counties, thus some stocks that only operate on specific exchanges might have a lag of information due to holidays

To resolve the problem, stocks who only had a minimum of 70% of data were put in the selection pool. Here 70% of the data is defined for the stock to have 180 entries since there are typically 256 working days in a given year.

For those stocks with empty entries, it was replaced with either the past day before closing price or the next day opening price depending on information available. This method is commonly used when repairing financial stock data.