Data Collection Report (Ryan Rigby, Scott Gale)

1. To analyze stock market performance in US Presidential election years, we gathered stock market data from companies represented in the Dow Jones Industrial Average (DOW) from 1972 to 2016 (12 election cycles). Although the DOW index is only representative of 30 companies; these 30 companies provide a comprehensive cross section of market segments and industries within the United States. We used the Yahoo Finance API to gather the data.

2. The bulk of our data is stock market data. We obtained End-Of-Day (EOD) market data for 30 stocks for the 12 election years from 1972 - 2016 (~ 240 trading days per year). This equates to around 700,000 unique combinations of stocks and dates. The final size of the data will be a multiple of this number since each EOD quote consists of the open, high, low, and closing prices for the stock that day, as well as the number of shares traded (volume) and the date. The remaining data relates to electoral candidates, election outcomes, and market sectors associated with each stock, which is small in comparison to the stock data.

3. The data will be stored in an array / matrix format containing a mixture of floating point (price information) and integer (volume) values. There will be one .csv file for each election year (12 total) with all corresponding data contained therein. We will create a class for each election year containing 240 dimensional matrices (representing each trading day of the year) for each of the following features:

Open (float), High (float), Low (float), Close (float), Volume (int)

This structure will allow us to efficiently access, cluster, and manipulate the data. We will also include fields containing election and market sector information.

4. For this project, we will not need to convert between format types. We will likely need to normalize the price data to a standard range so that values can be accurately compared against each other. We will structure the data into a smaller set of files that will be convenient to analyze and implement different techniques taught in the course.

5. Due to the unpredictable nature of the stock market, simulation is fairly simple using random numbers. We could derive a distribution from historical stock market data that models the statistical variations of gains and losses of stocks over a given period of time. From that distribution we could randomly select values to represent a stock's daily performance (positive or negative). Then we could pull a number from another gaussian distribution that would determine the magnitude of the positive or negative shift in value. This procedure could be repeated indefinitely to produce as much data as desired.