# Stock Market Analysis

Group 4: Stephen Kay, Rachel Mamich, Levi Nickerson, and Brandon Rajkowski

Our goal is to analyze historical stock market prices and trading volume in order to develop an investment strategy.  The investment strategy will be tested by creating a mock portfolio and comparing it to the S&P 500.

# Questions We Aim to Answer

- What stocks are commonly bought and sold together for a profit?
- Are there patterns in stocks prior to stock values crashing or skyrocketing?
- Are there any predictable patterns that can be exploited for profit?
- Are there pairs of stocks whose prices typically trend in the same or opposite directions?

# Prior Work

- Heavy previous use of data mining in the stock market industry
- One popular example is using decision trees to purchase the most optimal stock options
- In General it is very hard to outplay the stock market with data mining but it is possible
- Previous papers have shown modest returns of ~$10,000 when using data mining over other available techniques
- Clustering and Decision Trees seem to be two of the more valuable techniques for analysis

# Datasets

- Kaggle US Stocks and ETF’s
  - https://www.kaggle.com/borismarjanovic/price-volume-data-for-all-us-stocks-etfs
  - Yes on Brandon’s machine

- NASDAQ Historical Quotes
  - https://www.nasdaq.com/quotes/historical-quotes.aspx
  - Yes on Brandon’s machine

- NYSE Historical Quotes
  - https://www.kaggle.com/dgawlik/nyse
  - Yes on Brandon’s machine

# Proposed Work

# Data Cleaning

- Convert date/time formatting to consistent format
  - Different datasets use different date/time format

- Clean missing values by interpolating between the two nearest data points

- Adjust return rates when a stock split occurs

# Data Integration

- Datasets will be combined into one data warehouse
  - Some datasets have all stock data in one file, others have separate files for each stock

- Stock price date/time alignment
  - Datasets contain different time periods
  - Stock price data needs to be aligned temporally

# Data Processing

- Basket of goods analysis using Apriori algorithm
  - Use volume as a support indicator
  - Determine what stocks are bought together over a time period
- Frequent pattern mining
  - Two or more stock prices that typically move symmetrically and asymmetrically over a time period
  - Frequent patterns that occur before large price swings in a single stock
- Cluster analysis based on stock price and volume to determine similar performing stocks
- Decision tree to decide if the stock should be bought and to decide if its a good time to sell the stock if already owned

# Tools

- Python
  - Numpy, Pandas, Scipy, Matplotlib
  - Able to handle millions of rows of data
- Orange
  - Open source Data Mining Tool
  - Useful visualization features
  - Functionality to join two data sets
- Druid IO
  - OLAP queries

# Results Evaluation

- Build a mock portfolio based on the results of the data mining
- Track the portfolio’s returns over the course of a few weeks
- Compare portfolio’s final return with that of the S&P 500 for the same time frame
- Outperforming the S&P 500 indicates an above average portfolio and investment strategy

# Questions?