Equity Portfolio Optimization using ML Techniques

Graduate Project

ECE/CS 498 DSG

Spring 2020

Team

Akhilesh Somani : somani4

Gowtham Kuntumalla : gowtham4

Manan Mehta : mananm2

Gowtham

Outline

- Why Personal Finance?
- Problem Statement
- Exploring S&P500 Dataset
- Challenges & Solution
- Results and their significance and/or insight
- Conclusion / Feedback

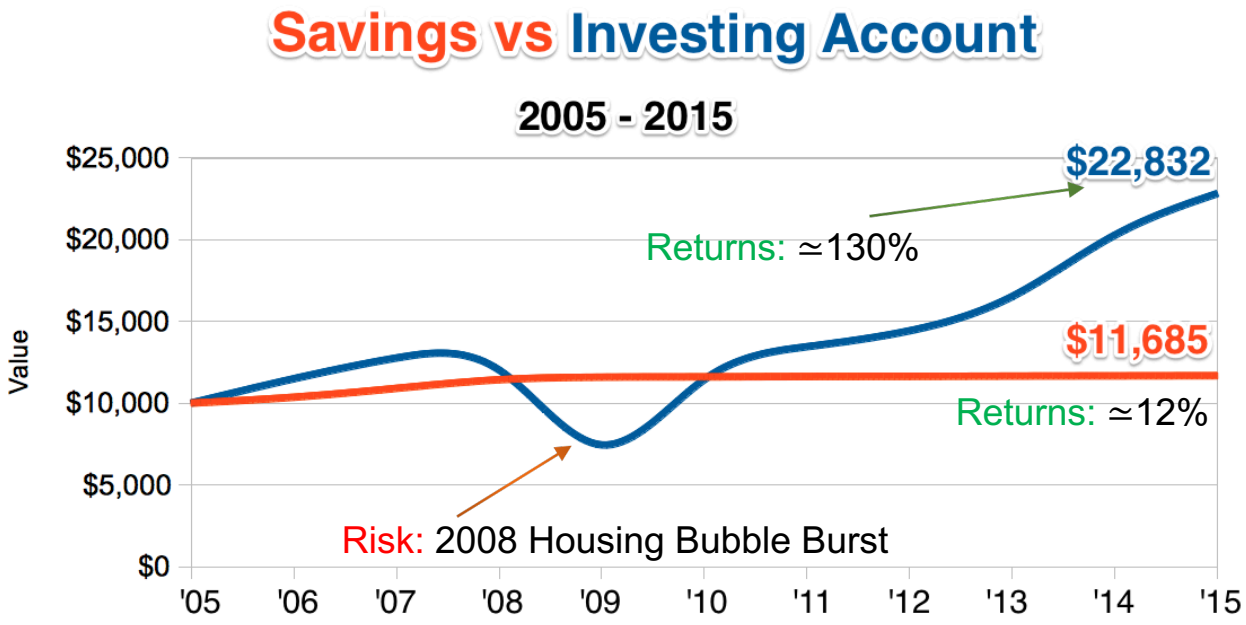
Gowtham

Why Personal Finance?

- Personal Finance
  - Wealth management at an individual or family scale

- Saving vs. Investing (vs. Trading)
  - Value investing is the best way for passive investors looking for good returns

- Focus: Equity – Stocks
  - Choose diverse stocks and build a portfolio
  - Return vs. Risk



**** Why Investing is better than saving***

Gowtham

Problem Statement

How to choose stocks in a portfolio to **maximize returns** while **minimizing risk** (aka volatility) over a fixed timeframe?

Gowtham

Data

- S&P 500 stocks data from 02/2013 – 02/2018
  - ≃ 500 exchange traded stocks
  - Daily Open, High, Low, Close Prices & Volume (OHLCV)

	Date	Name	Open	High	Low	Close	Volume
One stock	2013 – 02 – 08	AAPL	67.7142	68.4014	66.8928	67.8542	158168416
	…	…	…	…	…	…	…
	2018 – 02 – 07	AAPL	163.085	163.4	159.0685	159.54	51608580
500 such stocks	…	…	…	…	…	…	…
	2018 – 02 – 07	ZTS	72.7	75	72.69	73.86	4534912

- Custom Feature Engineering
  - OHLCV – not sufficient to accurately study trends
  - Nonlinear features – Technical Indicators for momentum, volatility, trend, etc.
  - Studied ~ 111 indicators used for trading (domain knowledge)
  - Implemented ~ 22 indicators out of them which represent macrotrends

Akhilesh

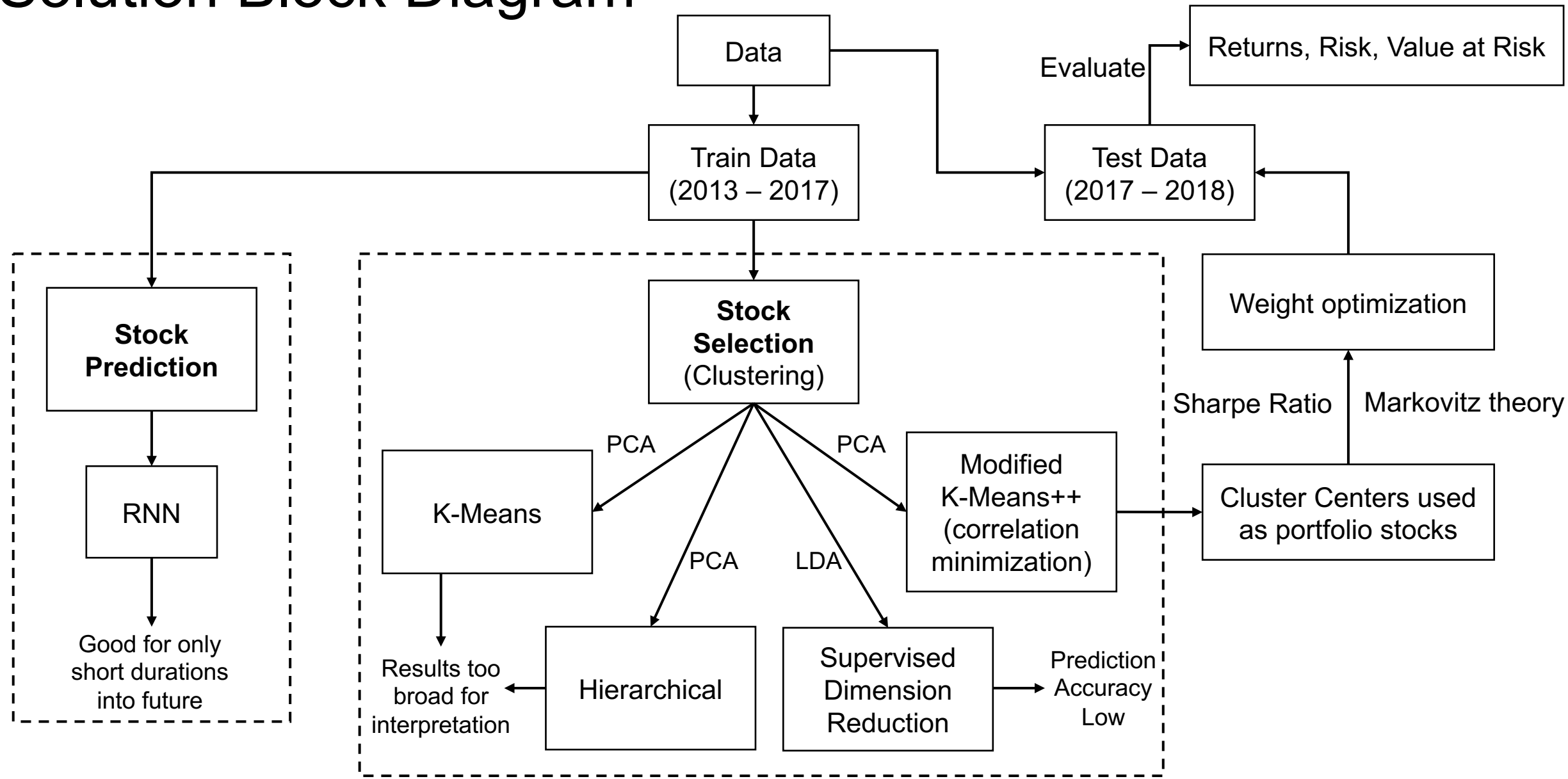
Challenges

- Huge Time Series Dataset
  - 1200 data points for each stock (5 year data)
  - 500 such stocks

- Historical prices alone do not guarantee returns
  - News related events influence prices
  - Quarterly earnings of companies may be correlated

- Complex interactions between stock returns
  - Seemingly unrelated stocks from different industries may vary together
  - Randomly picking portfolio stocks from different industries may not reduce risk directly

Solution Block Diagram



Akhilesh

RNN – Long Short Term Memory (LSTM) neural network

Note: This prediction for 1 year was made at data at beginning of Feb 2017
(No peeking into the future, Rolling prediction)

- Doesn't predict long term variations well
- It is decent for very short term

RNN Conclusions:

1. Past prices alone cannot predict long term future

2. New information continually affects price significantly
   a. company news,
   b. policy changes,
   c. disaster info

Gowtham

Stock Clustering

- Correlation Minimized Selection Algorithm:
  - Step 1: Provide a seed stock
  - Step 2: Perform PCA on the seed stock, store PC1
  - Step 3: Perform PCA on all other stocks **using seed stock as the fitter**
  - Step 4: Find $\rho_i = \rho(PC1_{seed}, PC1_i)$ for all i
  - Step 5: Select $stock_i$ such that $\sum_j \rho_{i,j}$ is minimum, (j : stocks already selected)
  - Step 6: Seed stock = $stock_i$, repeat steps 2 – 5 till 'n' stocks selected
- Ensures least correlated stocks are selected based on training data

Sample run with seed stock 'FB':

Iteration	1	2	3	4	5	6	7	8	9	10
Stock Selected	Facebook (seed)	Extra Space Storage, Inc	Constellation Brands	Fiserv	OR Auto	Acuity	American Water Works	Global Payments, Inc	Nasdaq, Inc	Ecolab
Industry	Tech	Real Estate	Consumer Staples	Finance	Industrials	Real Estate	Water	Finance, Tech	Finance	Energy

Baseline Model

- Portfolio performance is a (strong) function of weights assigned to each security

- Minimum variance (risk) portfolio: Markovitz Theory

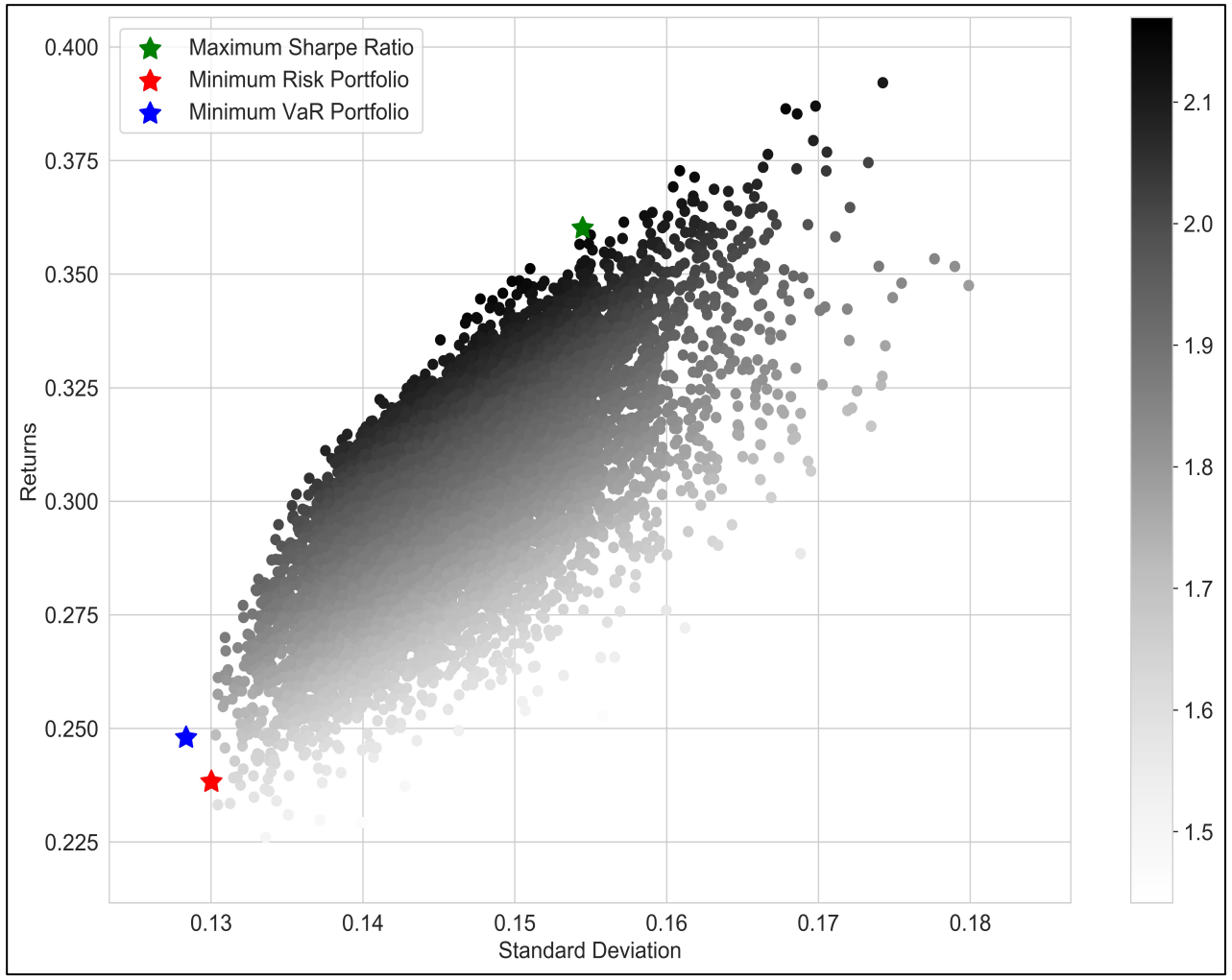
- Markovitz optimization:

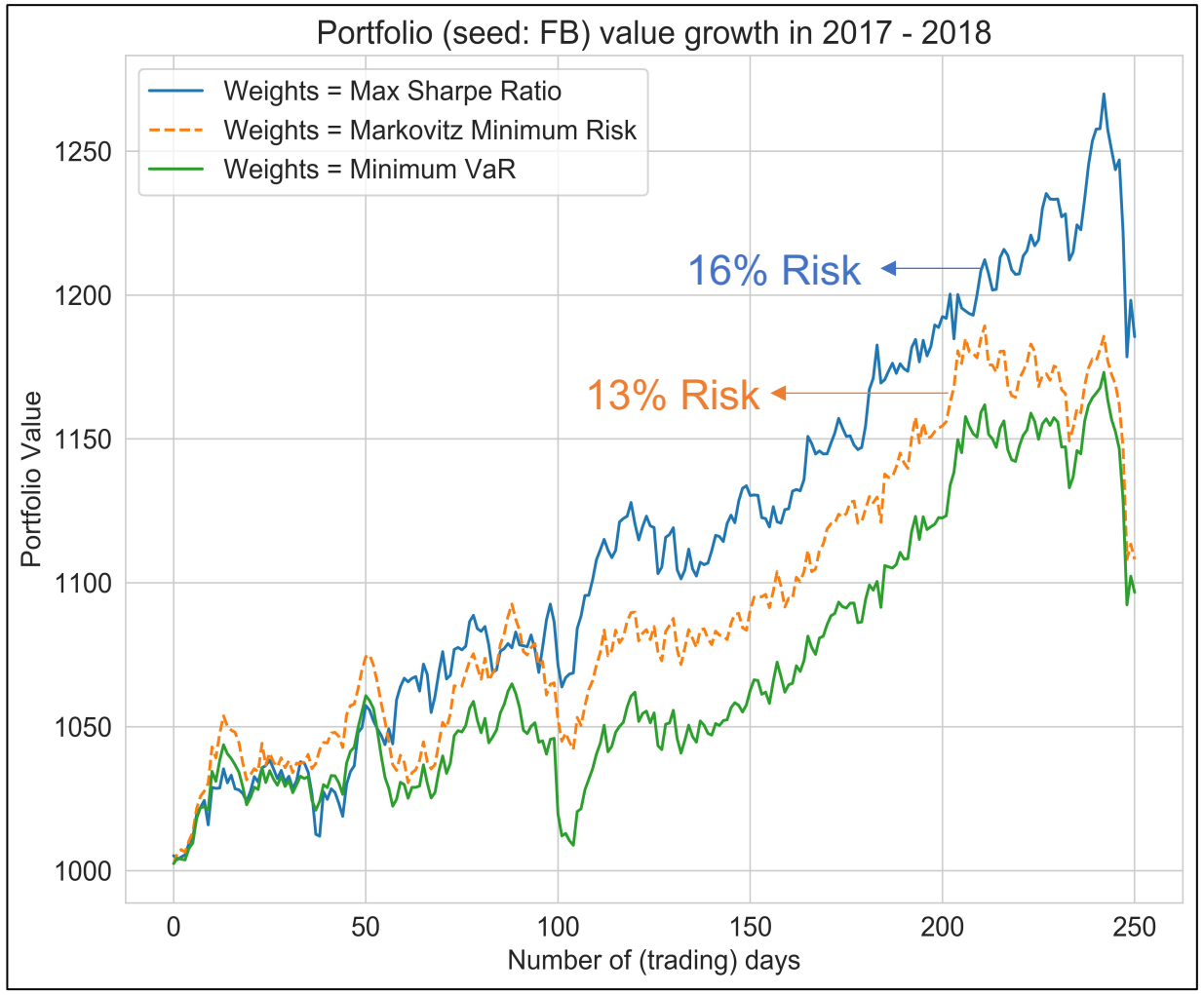
Minimize	$\frac{1}{2}\, w^T \sum w$	(risk)
subject to	$m^T w \geq \mu_d$	(a minimum return)
and	$e^T w = 1$	(weights summing to 1)

where $m_{n\times 1}$ is the weight vector, $w_{n\times 1}$ is the mean vector, $e_{n\times 1}$ is the vector of ones and $\sum_{n\times n}$ is the covariance matrix

Results

Seed = Facebook (NYSE: FB)

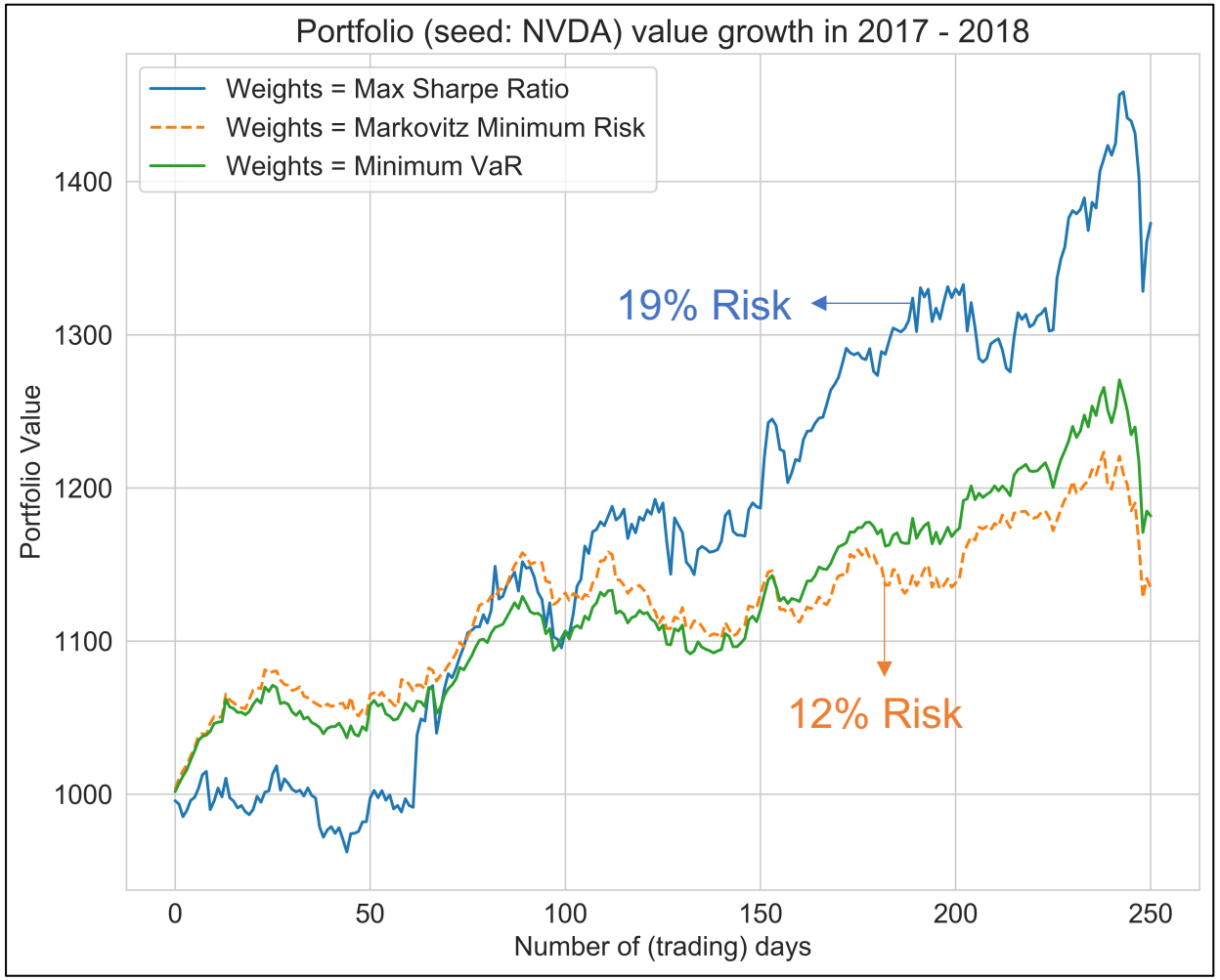




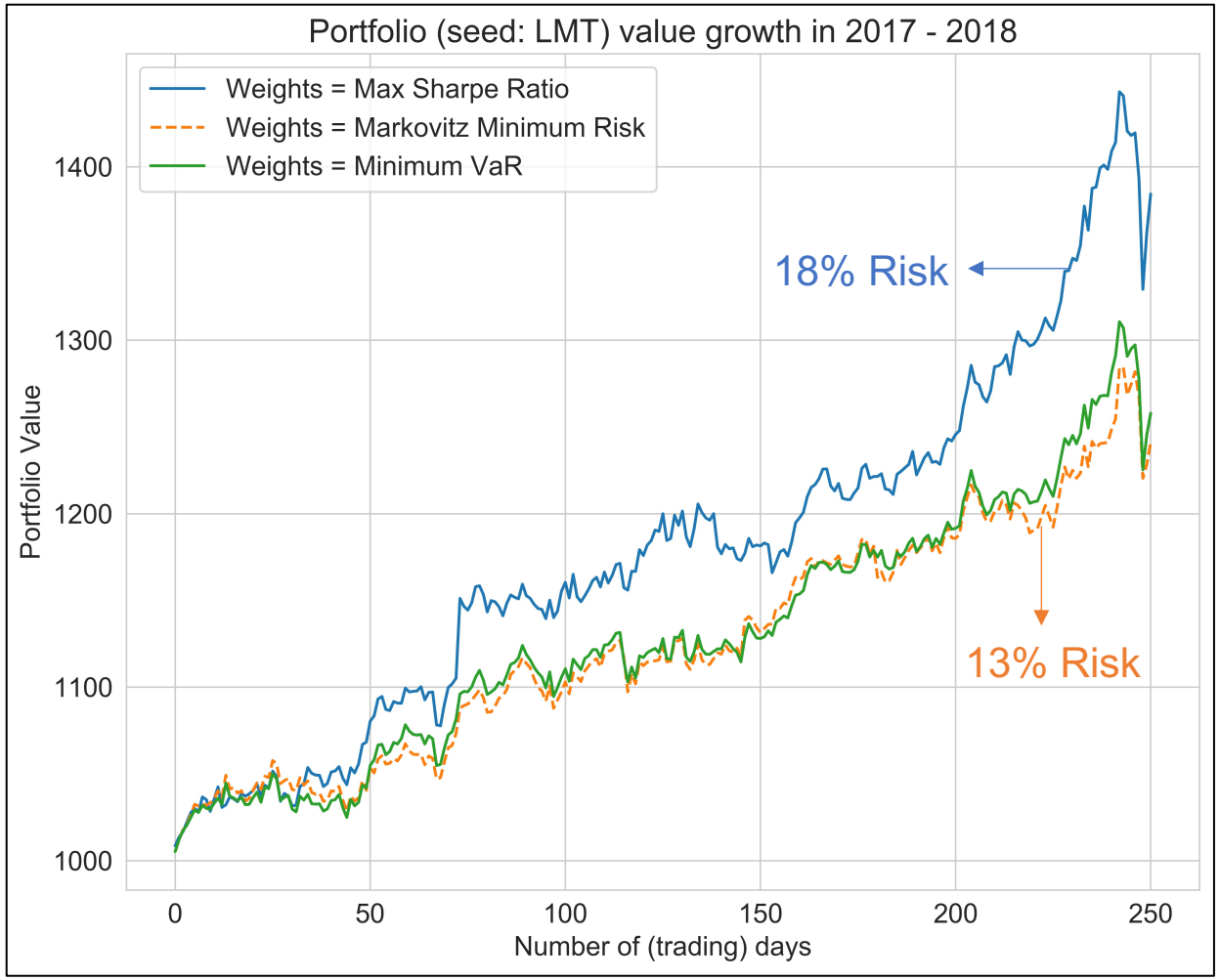
Inferences (weights) from training data ⟶ Portfolio Performance in test data

Manan

Results – Other Good Predicted Portfolios



Seed = Nvidia (NYSE: NVDA)



Seed = Lockheed Martin (NYSE: LMT)

Manan

Conclusions

- No ‘right’ portfolio, but most portfolios tested perform better than the underlying stocks in terms of risk-return compromise.

- The Maximum Sharpe ratio portfolio outperforms others, but at higher risk. The minimum Value at Risk (95% confidence interval) portfolio is suggested for the average investor.

- Limitations
  - No comparable metric to quantify clusters – clustering algorithm is unsupervised
  - A predictive model can help in quarterly portfolio rebalancing to have even higher returns. We tried this using RNNs but couldn’t get good results.

- Feedback for instructors
  - The progress reports were well scrutinized and feedback was thoughtful
  - The Checkpoints kept the project on track, well organized timeline

Contribution

- Akhilesh: Feature Generation, Principal Component Analysis, Linear Discriminant Analysis, KMeans Clustering, LSTM-RNN

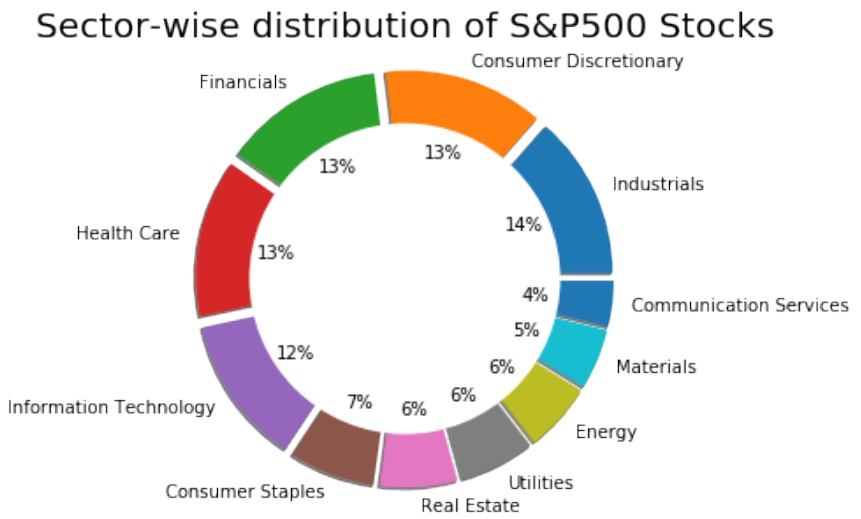
- Gowtham: Feature Generation, LSTM-RNN, Hierarchical Clustering, Portfolio management technical discussions

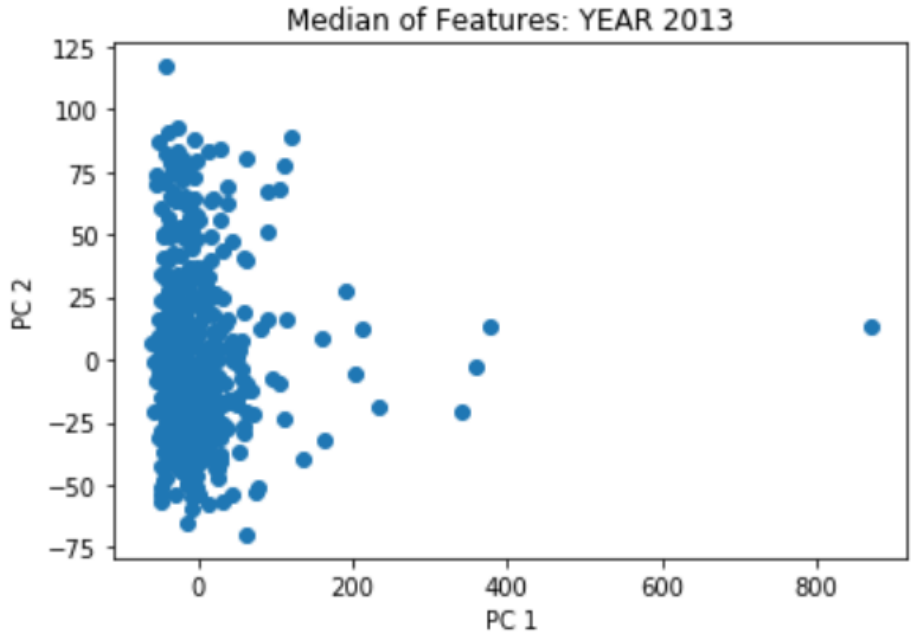
- Manan: Feature Generation, PCA, KMeans++ clustering, correlation minimization selection algorithm, Markovitz optimization, portfolio evaluation (Sharpe ratio, future returns, risk, VaR, Monte-Carlo weight plots)

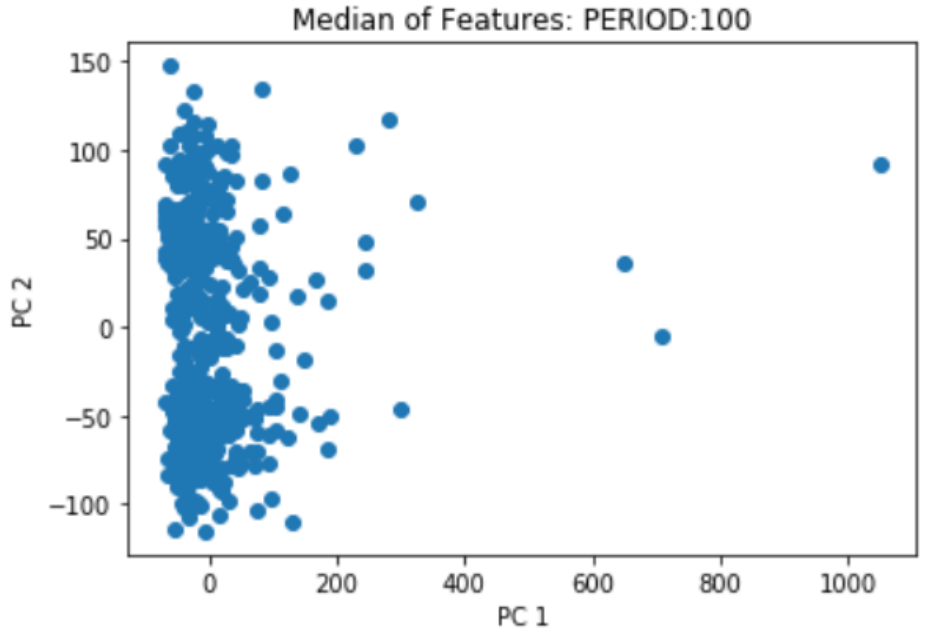
Appendix

PCA Approach

- Novel approach to achieve sector-wise segregation
- Challenges:
  →3D time-series data
  →PCA works on 2D data
- Approach: Condense the time-series data
- Results: →PC1 and PC2 cover ~ 85 percent variance

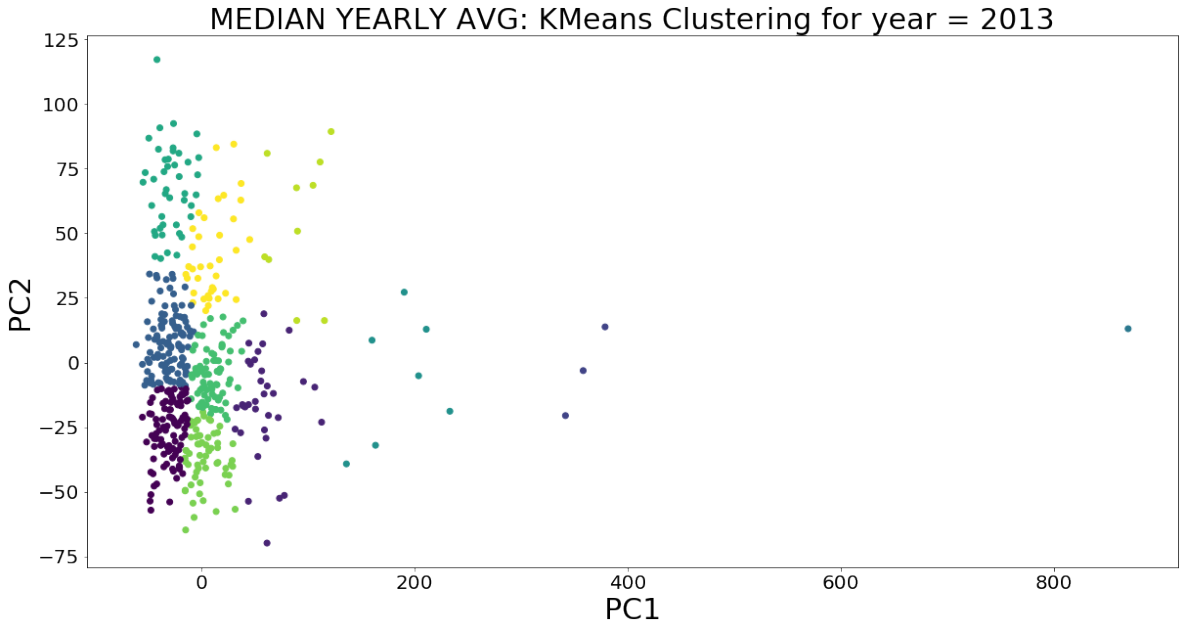




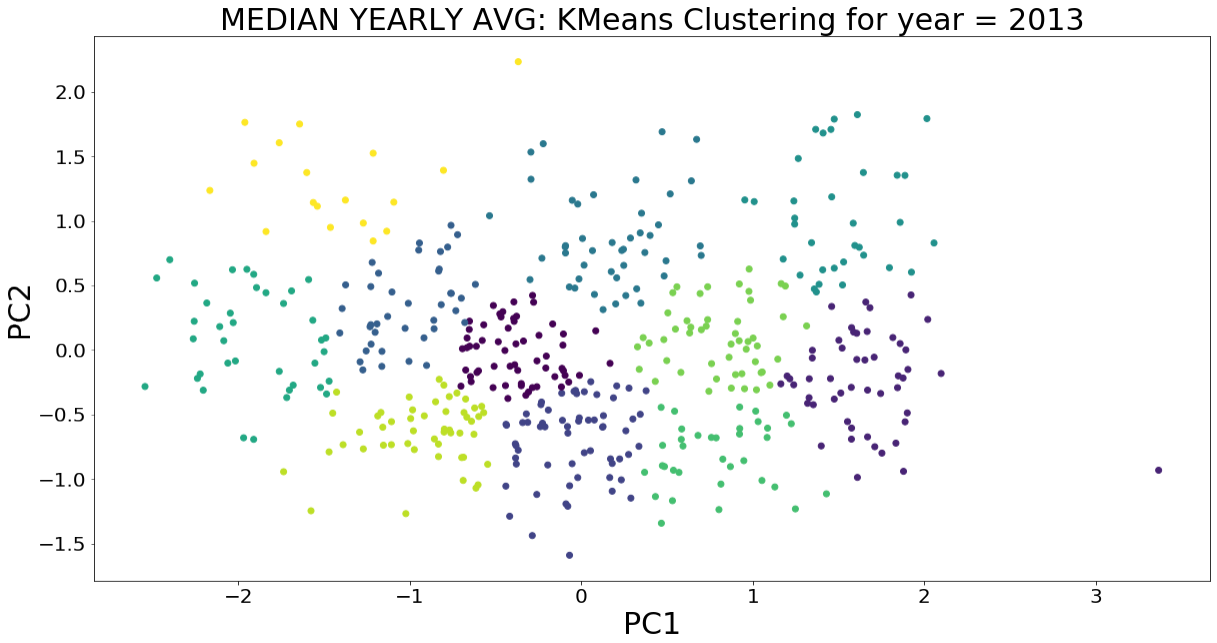


Clustering Approach

KMeans Clustering (using K = 11) on PCA transformed data



KMeans Clustering – Unscaled Data

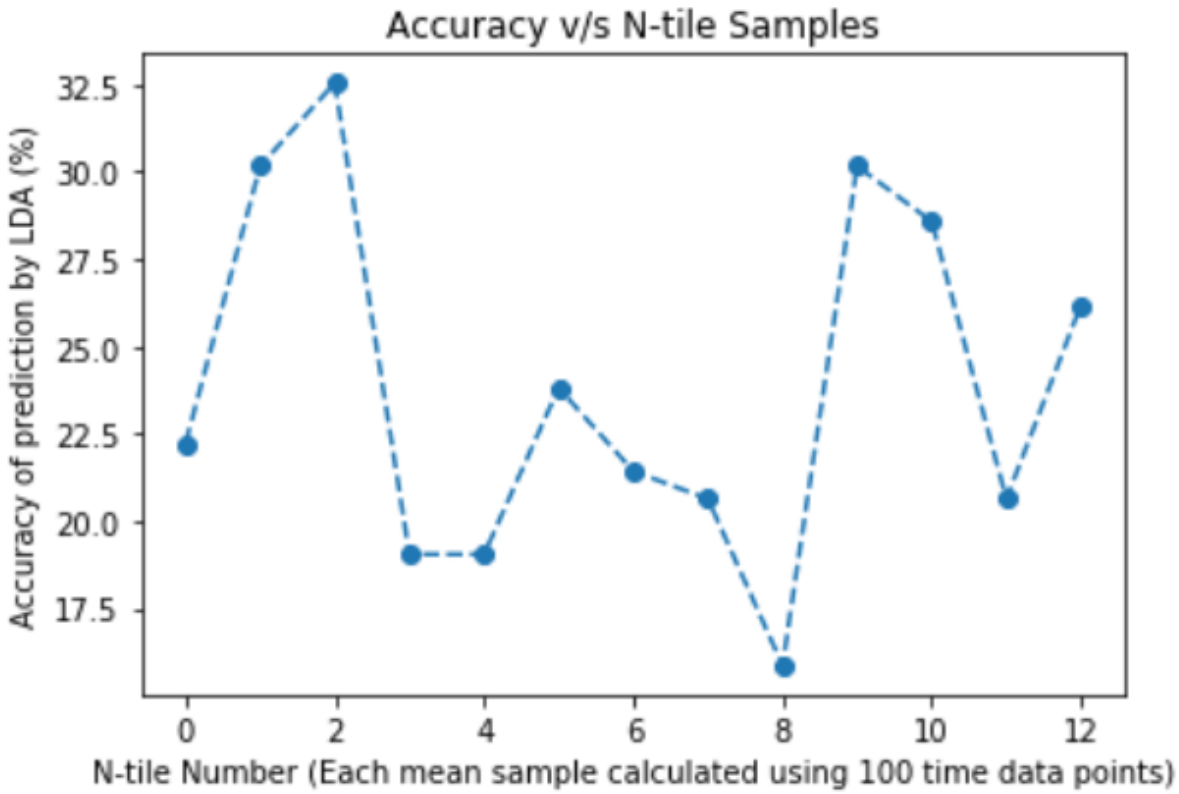


KMeans Clustering – Scaled Data

LDA Approach

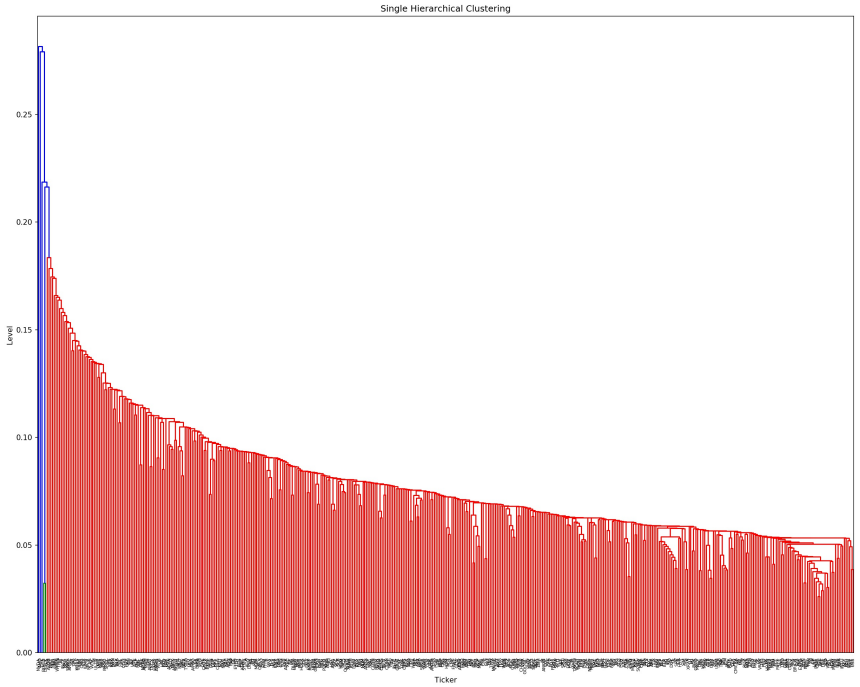
- Linear Discriminant Analysis to reduce dimensions and achieve class separation

- A supervised way of learning

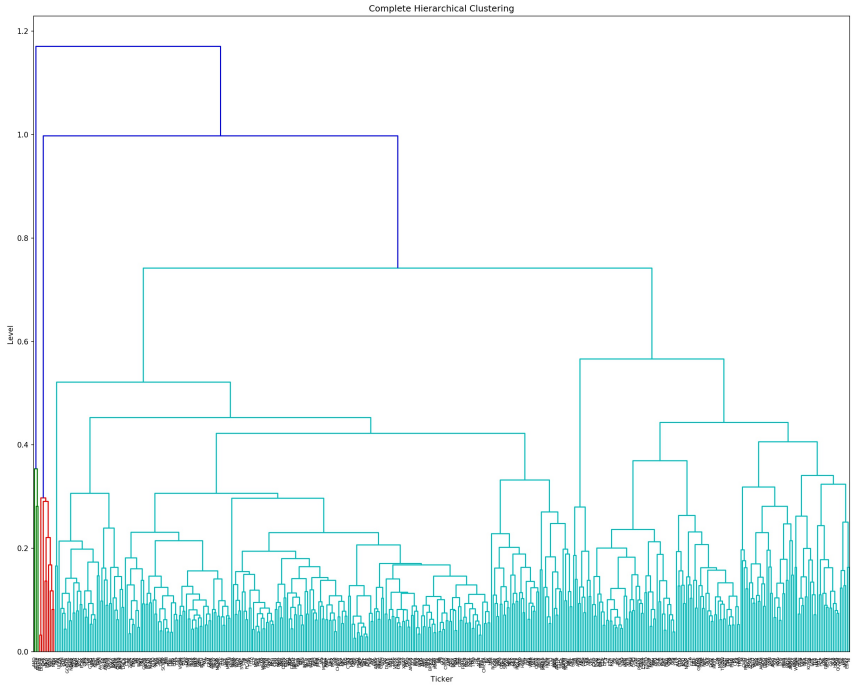


Hierarchical Clustering

1. Decent Performance:
   a. Ex: AMD and NVIDIA are in same cluster.
   b. Though Facebook and Google are farther than expected.
2. Too sparse for quantitative interpretation.
3. Feature: close prices



Single Link



Complete Link