**CAPSTONE PROJECT PROPOSAL FOR DATA SCIENCE INTENSIVE (SPRINGBOARD)**

**ARUN BHARADWAJ, 28 AUGUST COHORT**

## INTRODUCTION

Capstone project titled 'Understanding Portfolio Theory using data science and natural language processing' aims to satisfy the requirements of the Springboard Data Science Intensive certificate. The goal of the capstone project is two-fold: to understand the relationship between stock prices and fundamental company data/macroeconomic indicators, and to develop a simple algorithm that can classify a market participant`s investing philosophy.

## UNDERSTANDING RELATIONSHIP BETWEEN FUNDAMENTAL DATA, MACROECONOMIC INDICATORS AND STOCK PRICES (PART 1)

This is done by using historical company financial statements, economic datasets and identifying their relationship with the company`s stock price. Company financial statements are divided into P&L (profit and loss), balance sheet and cash flow. Economic datasets may include inflation, GDP, national debt, trade deficit, consumer confidence index, S&P 500 index etc. Taken together, individual company performance and macroeconomic indicators are the biggest determinants of stock price movement.

## CLASSIFYING INVESTING PHILOSOPHY (PART 2)

Investors are classified based on their investing styles. There are growth investors, value investors, quantitative investors, macroeconomic investors, short sellers etc. Growth investors look for higher risk stocks that have the potential to generate large returns. Value investors look for less liquid, undervalued stocks that have a difference between their market value and intrinsic value. Quantitative investors use trading and momentum patterns to generate investment ideas. Macroeconomic investors use a top down approach to investing and switch between different international markets to achieve market outperformance. Short sellers profit from falling stock prices. An investor`s investing philosophy is gleaned from the letters they write to their clients. Clients invest capital in a fund run by the investor and receive constant communications from the investor. By using these communications, natural language processing and text analytics, the capstone project tries to classify investors into different types.

## WHY IS THIS PROBLEM STATEMENT IMPORTANT

The two parts of this problem statement constitute, at a very high level, some of the most pressing questions in investment finance, with respect to stock market investing.

Some academics claim that stock prices are a random function that have only a minor relationship with other economic variables. In contrast, many market participants and other academics claim that stock price movements can be predicted using company specific and macroeconomic variables. If stock prices are a random function, investors will be better off buying index funds since stock selection will be a futile exercise. If stock prices are a function of some underlying variables, stock picking is likely to be more profitable than buying an index fund.

The investor classification algorithm is significant since it allows future researchers to understand the effect of investing philosophy on the returns generated by investors. An algorithm, compared to a manual classification, allows us to classify the thousands of investors operating in stock markets in an efficient manner. This classification can then be used to identify the aggregate returns generated by different types of investors.

**FOR WHOM IS THIS PROBLEM STATEMENT IMPORTANT**

This problem statement is important for academics, hedge funds and students of portfolio theory. Academics are always interested in understanding whether humans can generate returns better than the S&P 500 index. If humans are better, academics would further like to understand the behavioural theory behind this outperformance. Hedge funds are also interested in these topics since an understanding of stock price movement will help them outperform the index and generate better returns. Today`s students are tomorrow`s academics and hedge fund managers. So, they are interested in this field of study.

**ACQUIRING REQUIRED DATASETS**

Stockpup.com has an extensive library of company fundamental data. It has 739 CSV files containing quarterly fundamental information about 739 listed companies in the US over many years. It also contains stock price high and low for the day of quarter ending. This is a very extensive dataset and contains all attributes needed to perform our analysis.

Quandl.com has an extensive library of US macro-economic data under the classification of futures data, federal reserve economic data, Yale department of economics, University of Michigan etc.

For the investor classification, we use letters sent by famous investors to their clients. For example, all of Warren Buffet`s letters to shareholders are available in public domain. Similarly, letters written by different types of investors will be acquired from different sources to generate the training dataset for the classification algorithm.