Capstone Project Proposal

# Goal: Implement a process of developing trading strategies in Indian Equity Markets using fundamental data of liquid listed companies that maximises out of sample average returns keeping risk under control.

# Who cares?

I do. I am planning to start my own investment firm. I have not focussed much on fundamental data before, neither have I explored advanced machine learning techniques to develop trading strategies. If there is indeed value derived out of using some subset of Data Science that I learn during this course, I am happy to invest my own money on these trading strategies.

# Data

I have subscribed to two data sets from Quandl.

1. <https://www.quandl.com/data/DEB-Core-India-Fundamentals-Data> Contains fundamental indicators derived from company financials from published balance sheet, income statements and cash flow statements from roughly 4000 listed stocks on NSE or BSE. History goes back to 2005
2. <https://www.quandl.com/data/TC1-Indian-Equities-Adjusted-End-of-Day-Prices> Contains daily prices (open, low, high, close), volume and value traded for constituents of top 500 NSE listed stocks by market capitalization. The prices are adjusted from corporate actions like stock splits, rights issues, dividends, buy backs etc.

# Approach

I wish to follow the approach of a white paper published by a Deutsche Bank research group. In that paper the authors review various machine learning algorithms and apply them in practice on Japanese Equity markets data. The Following are the steps involved

1. Data exploration to decide the date range for training set and test-set and stock universe for this study. Factors involved:
   1. Data availability
   2. Stock liquidity
   3. Avoid any bias in universe selection like survivorship
2. Investment signal creation and classification: A close look at fundamental data set to see whether it needs enrichment i.e. computing well known indicators to make a comprehensive list of possible investment signals classified into various factors like growth, value, quality, size, momentum etc
3. Visualize performance of individual investment signals If each investment signal is used only by itself as a ranker for long stock portfolio, how is the in-sample performance of the strategy.
4. Data pre-processing like filling missing data, normalizing, uniformizing, sector-neutralizing, quantizing, winsorizing etc to make variations in output variable (returns) more sensitive to variations in input variables and also make combining variables more sensible.
5. Create a benchmark by using a simple linear model (with some basic checks) for other machine learning techniques to see if they do any better
6. Test most promising models as per the paper and see if they perform any better than the simple model
7. Document the conclusions and insights