**6643 Longitudinal Project**

**Guannan Shen**

**Data Source**

<https://www.kaggle.com/c/two-sigma-financial-news>

**Description**

This project is a Kaggle competition launched by “Two Sigma”, which is designed to use the predictive power of the news to predict stock price performance.

The dataset has two separate parts: 1) Market Data (2007 to 2016-12-31) provided by Intrinio (Stock Price API) contains financial market information such as opening price, closing price, trading volume, calculated returns, etc. 2) News Data (2007 to 2016-12-31) Source: Thomson Reuters - contains information about news articles/alerts published about assets, such as article details, sentiment, and other commentary. Each asset is identified by an assetCode (note that a single company may have multiple assetCodes). There is also an assetName tagged for each asset, however assetName could be “unknown”.

**Key research questions of interest**

Using previous stock price performance and News reports data (2007 to 2016-12-31) to predict the stock price performance from 2017-01-01 to 2018-07-31. (The stage 1 goal of this competition)

**Correlation in the Dataset**

1. Time series data (Market data), thus there is correlation within single stock overtime.

2. Correlation between stocks, such as on industrial level.

3. Correlation in News data overtime.

**Challenges and Motivation**

1. This is a Kernels-only, time-based competition, I cannot interact directly with the data files as I would in a standard data analysis. Thus, I must use Python code to access the competition data on the platform called “Kaggle Kernels”, in which runs the Jupyter notebooks in the browser. In this way, the whole process is up on the Cloud by python, instead of on my local machine. Python could be the problem but I really want to give it a try.

2. The dataset is huge with several million rows in each part of the dataset.

3. How to join the information from two parts of dataset? I might start from the easier approach, using only previous stock price data (Market Data), which can be treated as multiple time series data.

4. Missing data.

5. Left and Right - censoring, as there may be instruments that enter and leave this subset of data (Market Data). There is no doubt that not every company has news reported daily.

6. I might try fancy machine learning methods later. However, I will definitely build linear mixed model and time series model (Maybe Market data only to simplify) in Python first as the benchmark.

7. This is an open-ended real-world problem, which is very exciting.