## A: Provide a written report (in pdf format), available in your class github repo containing at minimum:

- High level project description: a summary of what was in your project proposal

My topic is market prediction. This is a fundamentally simple and reasonable goal. However, depending on the data that I harvest from news sources for this goal it could have either a reasonable outcome that is highly helpful or an absurd, chaotic, misrepresentative one. In the first case this project would be more of a market analysis tool and in the other case it would be an absurd experimental data driven art piece detailing the chaotic nature of market reporting and decision making in our very fast paced world.

Markets have a treasure trove of data associated with them. They have existed for hundreds of years. They have millions, billions or trillions of data points created constantly, and they are still being updated at every millisecond. Furthermore, as they continue to be updated, they have more data generated because they expand exponentially. I think this data theme offers a lot of avenues for interesting and revealing outcomes.

Stereyl foregrounds the concepts of signal and noise. In my project the noise would be irrelevant news regarding market movements and the signal would be news that is relevant and actually operates as predictors. Even in retrospect, it's impossible to know, on an individual basis, which of the millions of headlines predicted a movement in the market, let alone, any one of the market's constituents. However, if given a large enough dataset of daily market movements and daily news collections, I think it might be possible to screen out the noise and find which keywords, topics, or news sources are the most reliable in predicting the market and therefor constitute the signal. Furthermore, I can download archives of the history of news and markets and look at how news predicted markets in the past.

Every datapoint has an associated a time tamp and sometimes this becomes misleading. This is because time is largely arbitrary when it comes to movements in prices. Usually price changes happen because of events, not times. So, time becomes an almost useless datapoint even though it is the most common. This is also why I am prioritizing collecting data over time. The more data I have, the more accurate my machine prediction is.

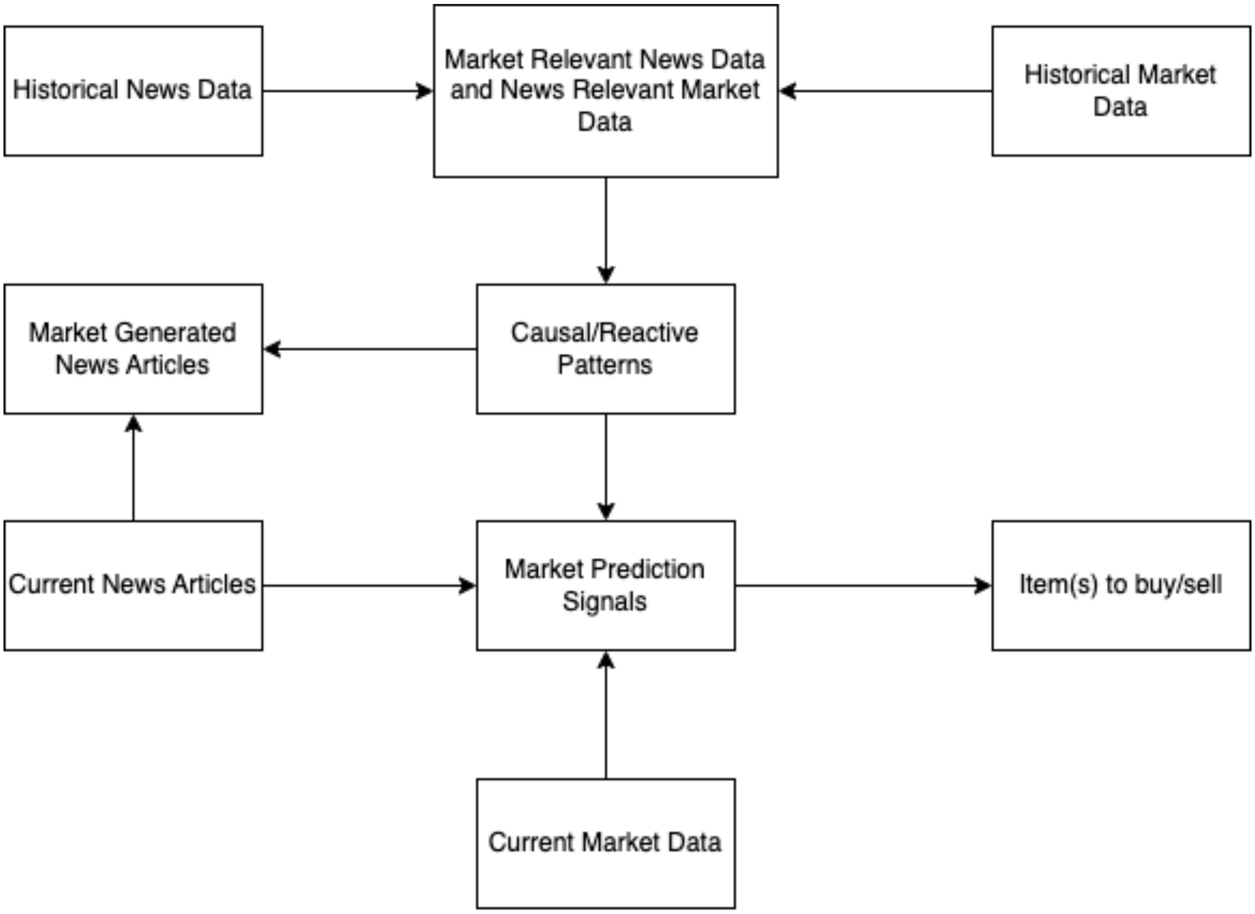
So, along with this simplistic, capitalist application, I also want to create a front-end for the news aggregation aspect of the project. Maybe an apophenic newspaper. Or machine-readable news. Or rather, a machine digested news. One that aggregates hundreds or even thousands of news sources and spits out the most euphoric/pessimistic stories at the top and most mundane ones at the bottom. Or even a newspaper that combines stories on a given topic into one article resulting in a jumbled, schizophrenic diatribe of sorts.

- Description of which stage you are at in the project: what has been completed and what is still to be completed

I have completed the foundation of both portions of the project, news and market data aggregation. For the news side, I have a script that pulls as many articles as the api allows (up to 100 requests a day each with a maximum of 100 articles). This equals 10,000 articles plus their metadata per day. Working within just the 'business' category, I should be able to archive every article published going forward. While relatively expansive, it would be beneficial to expand the scope of my data to other categories and into the past if possible.

For the market data portion, I have available a seemingly infinite amount of data requests with only certain types of data restricted. Here I may have a problem with the quality of data instead of quantity. I have written a script to pull the valuation history of every single stock

- Detailed images/diagrams of the overall system (i.e. how data flows between the various components)



- For each component/feature, provide written descriptions on the usage/purpose and how it integrates into the project

Historical news data and historical market data are aggregated to be as wide in scope as possible in order to find the most accurate correlations.  The end point, item to buy/sell, is the holy grail, and it is determined by a market prediction signal which receives data from current market data, current news articles, and reactive patterns.  Reactive patterns is determined by the coupling of historical news data and historical market data.  Market generated news articles is basically a reverse of market prediction signals, where an article is generated in response to a movement in the market instead of the other way around.

- Detailed explanations for which features/components are working and which need to be modified/adapted/scraped or reworked.

The coupling of historical news data and historical market data has been completed. After aggregating every stock in the American market and as many articles as my news api will allow, I have written a script that searches through every single news article for the occurrence of a company name and/or ticker symbol.  If an article mentions any company, a sentiment analysis is performed on the words leading up to and following this mention.  If the mentioned company's valuation deviates by a statistically significant amount before or after the time the article was published, then a record signifying a causal/reactive pattern is created.  This record contains the company, news article and metadata, as well as the details of the cause/effect: the degree and quality of the sentiment of the article and when in time the deviation occurred relative to the publishing of the article.

Once enough causal/reactive patterns are stored, the ones that record a deviation in price after the article is published are used as a filter to determine which current news articles are signals to buy/sell.  Those that record a deviation in a company's valuation before the publishing of the associated article are used in combination with current news articles as well as current market data to generate market relevant articles.

While the occurrence of buy or sell signals may be far and few between due to a lack of news data and processing power, the generated articles could influence market participants if given enough exposure.  Of course only one of these goals will word.  If the news informed model can predict markets, then news must be actually credible