**Goal:** Predict the S&P 500 index based upon anticipated news

**Data Sources:** Reddit, major news sources, economic reports, and individual stock prices of the S&P (will use weighted addition to group these together according to sector and market caps)

**Procedure Steps:**

1. Identify potential sources of information
   1. include reasoning steps as to why information sources makes sense
      1. Such as r/politics, or the economist
2. Scrape the relevant information
   1. Use an api if possible
   2. Don’t get hung up on a source
3. Create a bias normalized dataset consisting of public data
   1. Identify if a source leans right or left if political, or if a source is overly protective/supportive of a company, or if a source …
   2. The bias can be accounted for in a key phrase vectorizer; if one key phrase occurs more often than other key phrases, then account for it (maybe random selection
4. Find sector/megacap company pricing relationships with the public data
   1. Build tree models, and
5. Check the credibility of the relationships by using a testing model; for time series data you do backtesting