1. Can I select stocks using fundamental data?  I’m looking to predict stock performance using individual stocks’ fundamental data (such as book-value-to-price, EPS-to-price, etc.) using an ensemble of classifiers -- decision tree, random forest, neural network, etc.  I can gather monthly stock-level data over 20 to 30 years and build classifiers that predict performance using unseen data.  Each month, the regressors will be the fundamental data, and the target will be the quantile to which the stock’s next-month return belongs (e.g., maybe portfolio “1” being the worst-performing and portfolio “5” being the best-performing).  I look to train my model on the first 50 to 60% of the data, validate on the next 20% to 30%, and then test on the final 10 to 20%.  Considerations from the top-level (i.e., not at the unique model level) include (1) the nature of ensemble voting -- should it be majority or weighted; (2) bootstrapping; and, (3) bagging.  Performance will based on precision, recall, and accuracy: each month I’ll quantilize the stocks’ prior monthly returns relative to the median return, where those stocks whose returns are above the median are “winners” and those below are “losers.”  In this project, within the confusion matrix, “false positive” refers to buying losing stocks, and “false negative” refers to not buying winning stocks.

2. Can I allocate in and out of stocks using a rolling window of prior returns?  Using PCA, I want to test if I can extract a signal out of a rolling window of the S&P 500’s prior returns.  Some research exists about the fragility of markets whenever a fixed number of principal components explain a greater percentage of variation in the data.  At a fixed point in time t, I wish to look back over the past, say, two years and aggregate all stocks that belonged to the S&P 500, which proxies for the overall US equity market.  I wish to call PCA on the past two years’ worth of returns and extract that explained variance on the first n components.  The number of components and the explained variance can be tuned until a final model is settled on.  This project is given to some wrangling considerations as it relates to the covariance matrix -- given that stocks are added to and dropped from the S&P 500 with no firm periodicity, null values in covariance matrix are expected and must be addressed.  Also, intuitively, more weight should be given to more recent observations, which points to an exponentially weighted covariance matrix, where the half-life serves as a hyperparameter that might have to be tuned.  Given the unsupervised nature of this project, there exists some subjectivity as to how successful the project is.  It might be hypothesized that there exists a latent relationship between the variance explained by the first n components and the performance of the S&P 500.  Further hypotheses might develop given the findings that the project unearths.

3. How closely can I track a portfolio of stocks using as small of a subset as possible? A challenge in portfolio theory is to replicate the behavior of one instrument via a proxy. Using only historical time series, I want to see if I can replicate the performance of an index such as the S&P 500 using a subset of stocks within the index. One intuitive means of partitioning a portfolio of stocks is by each stock’s sector, such as industrial, materials, telecommunications, etc. Using unsupervised techniques such as PCA, how much variation (over a span of, say, one week, one month, etc.) of each sector is explained by as few stocks within that sector as possible? For example, if we see persistently that one or two stocks explain most of the variation within a given sector, then we can trade just those stocks as a proxy for that sector. If we then examine all sectors, perhaps we can replicate the performance of 500 stocks using only 10 or 20. The practical consideration behind “replication” rests of replicating theoretical instruments or indices that are not practical to attempt to trade, such as indices that have thousands of stocks. Trading thousands of stocks is generally not implemented due to various operational considerations.