

Do more than six stocks matter for the NASDAQ?

David Whitney (Mentor: Sameera Poduri)

Project Overview:

While thousands of American companies makeup American stock exchanges like the NASDAQ, a recent Wall Street journal article, using data compiled by brokerage firm JonesTrading, stated that just six companies accounted for more than half of the value added to the NASDAQ in 2015: Amazon, Google, Apple, Facebook, and Netflix, and Gilead [1]. This surprising finding might indicate that most of the recent changes in the NASDAQ's value have been driven by only a privileged handful of technology/ pharmaceutical-related companies designated by the acronym 'FAANG' (Facebook, Amazon, Apple, Netflix, Gilead, and Google). If such a strong relationship were to exist between these stocks and the NASDAQ, that might indicate the NASDAQ is not a reliable barometer of the overall growth of the 3,000+ companies that makeup NASDAQ, but rather reflects changes from a handful of stocks. This report investigated this possibility.

Project Motivation:

When it comes to investing sentiment matters, and investors understandably are often drawn to invest in big, popular stocks. This is showcased by the recent buzz and enthusiasm surrounding the so-called FAANG stocks, stocks which have consistently grown over the last five years [1][2]. Although investing heavily in FAANG stocks might seem attractive for the average investor, there are long-terms risks associated with not diversifying an investment portfolio. Fortunately, other NASDAQ stocks have also consistently grown during the last five years, and likely matter for understanding recent growth in the NASDAQ. And while many of these other stocks lack the decided visibility of FAANG stocks, they do constitute viable investment opportunities. Thus, this project's goal is two-fold: 1.) to understand the relationship of the NASDAQ with all of its constituent stocks and 2.) to gain awareness of a much larger group of competitive NASDAQ stocks that could supplement investments in popular stocks like the FAANG stocks.

Key findings:

- While FAANG stocks do closely co-vary with the Nasdaq Composite Index, these stocks do not behave fundamentally different than other NASDAQ stocks. In fact, 20% of other NASDAQ stocks also closely co-vary with the NASDAQ. Furthermore, this group of stocks includes most Top 100 NASDAQ stocks (i.e. those listed on the Nasdaq-100 Index).
- For the last five years (2011-2016), stocks that closely co-varied with the NASDAQ tended to show the most consistent growth.
- Stock value is fairly predictable over long-stretches of time (across multiple weeks), indicating the feasibility of forecasting future changes.
- FAANG stocks alone can largely predict how the NASDAQ has fluctuated over the last five years. Furthermore, some FAANG stocks like Apple and Google tended to be more consistent in predicting market fluctuations.
- However, other Top 100 NASDAQ stocks (excluding the FAANG stocks) can predict the fluctuations of the NASDAQ equally as well; thereby, dispelling the notion that FAANG stocks are privileged relative to other stocks.

Analysis:

Part 1 – A stock is more likely correlated with the NASDAQ if it's in the Top NASDAQ 100 stocks

For the last five years (April 2011 - April 2016), the price of FAANG stocks have closely co-varied with the value of the Nasdaq Composite Index (Figure 1). Throughout this period, most FAANG stocks increased monotonically and exhibited high correlations with the NASDAQ (ranging from 0.89 to 0.95). High correlation values are interesting because they indicate the possibility of a causal relationship between FAANG stocks and the NASDAQ. This relationship has been recently suggested in a Wall Street journal article [1]. If such a relationship were true, then the collective performance of FAANG stocks alone could be sufficient to predict how NASDAQ has changed over the last five years.

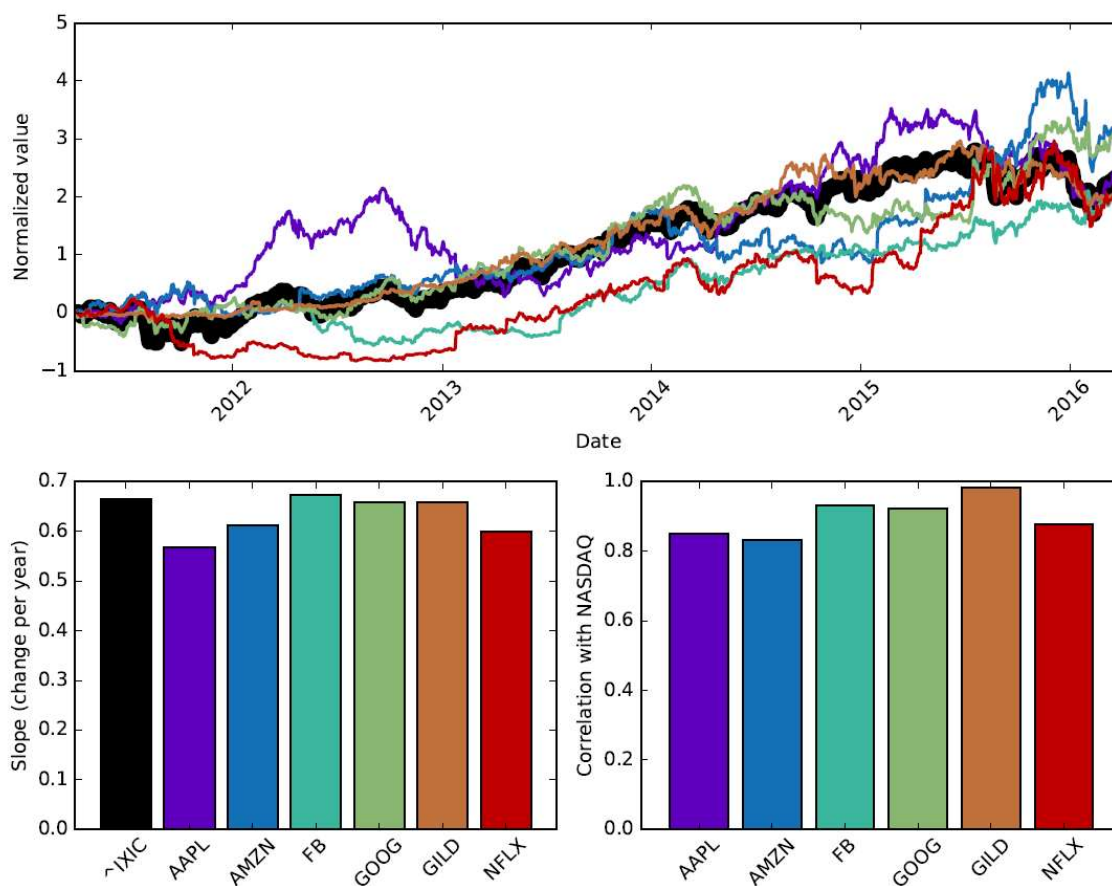


Figure 1: The relationship of NASDAQ to FAANG stocks between April 2011 – April 2016. (*top panel*) The time-varying change in normalized stock value of all the FAANG stocks relative to the normalized value of the NASDAQ (stock symbol is ^IXIC). Time-varying stock/index traces were normalized first by z-score, and then zero-shifted such that that the first datapoint for each stock/index was shifted to zero. The color for each trace matches the associated stock in the bar plots on the bottom row. (*bottom left panel*) The slope of changes for the NASDAQ and 'FAANG' stocks. (*bottom right panel*) The correlation between the NASDAQ trace and the individual FAANG stocks.

However before testing for a causative relationship between FAANG stocks and the NASDAQ, one should establish whether the FAANG stocks behave uniquely relative to other stocks that make up the NASDAQ Composite Index. Some of the largest and most influential NASDAQ stocks are listed on the Nasdaq-100 Index (which includes all of the FAANG stocks). And although correlations between FAANG stocks and the NASDAQ are consistently high ($r=0.90\pm0.02$ [SEM]), most of the other stocks included in the Top 100 Index exhibit similar

relationships to the NASDAQ (Figure 2, $r=0.80\pm0.03$ [SEM]). In fact, FAANG stocks are not significantly different from other Top 100 stocks (Rank-Sum Test: $p=0.57$). Although neither group is significantly different from the other, both groups are significantly more correlated with the NASDAQ than all other NASDAQ stocks (Rank-Sum Test: $p<0.01$). While NASDAQ stocks not within the Top 100 show systematically lower correlations with the NASDAQ as a group ($r=0.28\pm0.01$ [SEM]), ~20% of these stocks showed positive correlations with the NASDAQ that were comparable to FAANG stocks. These findings unequivocally demonstrate that FAANG stocks are not unique in being highly correlated with the NASDAQ, but rather are part of a relatively larger subset of NASDAQ stocks (~20%) that closely co-vary with the NASDAQ.

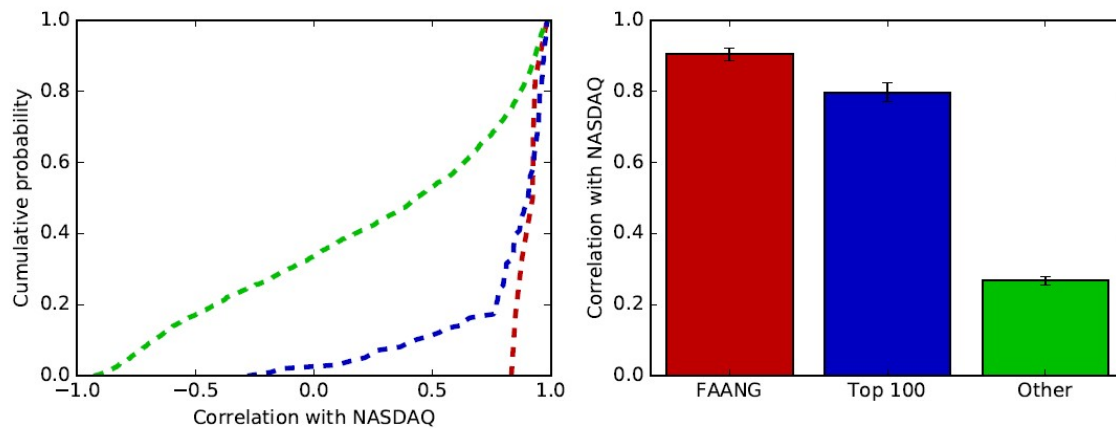


Figure 2: Correlation of different groups of NASDAQ stocks to the NASDAQ. (*left panel*) Distribution of correlations based on group type. (*right panel*) Summary correlations based on group type (Mean \pm SEM). The different stock groups are: FAANG stocks (red, $n=8$), other Top 100 stocks (blue, $n=99$), and other NASDAQ stocks (green, $n=2907$).

By definition the Top 100 stocks comprise companies with the highest market cap value in the NASDAQ. In fact, the Top 100 stocks makeup ~60% of the NASDAQ's value, and FAANG stocks alone makeup 25% of the NASDAQ's value. Thus, one uninteresting possibility is that correlations between NASDAQ and individual stocks might trivially arise based on a stock's market cap value. This is because NASDAQ is a capitalization-weighted index, where each company is weighted by its market value. However, the market cap value of a company is only marginally correlated with its stock price being correlated with the NASDAQ (Figure 3, $r=0.35$). This means that the relationship of individual stocks to the NASDAQ is more complicated and nuanced, and does not merely arise from market cap value alone.

But one practical question is whether higher correlations with the NASDAQ matter, specifically in regards to whether they are positive indicators for growth. And indeed over the last five years, higher correlations between a stock and the NASDAQ was a positive indicator for stock growth. Here stock growth is quantified as the slope of a line fit to the historical stock data (see Figure 3 - right panel). This suggests then that the stocks in the NASDAQ that are most correlated with the NASDAQ exhibited the most consistent growth between 2011-2016 (which includes the Top 100 stocks, see Figure 4). This is somewhat worrisome because if the NASDAQ's value reflects only a small growing-segment of stocks, then market growth is mostly concentrated and not broadly distributed across a larger array of companies. One important qualification is that the slopes do not measure the actual growth in a company's stock value. This is because each stock's historical data has been normalized via z-scoring; thus, highly positive slopes indicate consistency of positive growth, rather than the actual rate.

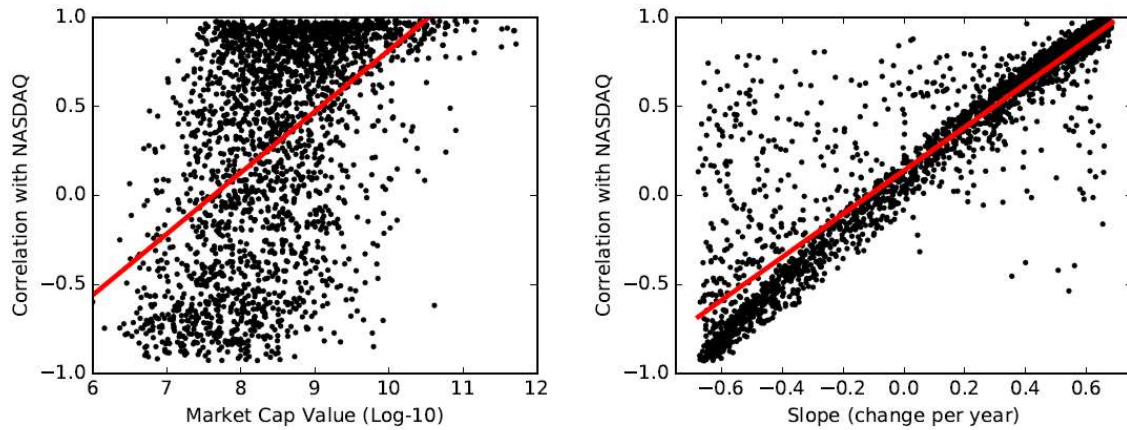


Figure 3: Relationship of each stock's correlation with the NASDAQ relative either to its market cap value or growth. (*left panel*) A weak correlation exists between the market cap value of a company (expressed logarithmically in dollars) and the correlation between a company's stock and the NASDAQ ($r = 0.50$). (*right panel*) A strong correlation exists between the growth of a company (expressed in terms of the slope for a linear fit to the company's time-varying stock trace) and the correlation between a company's stock and the NASDAQ ($r = 0.93$).

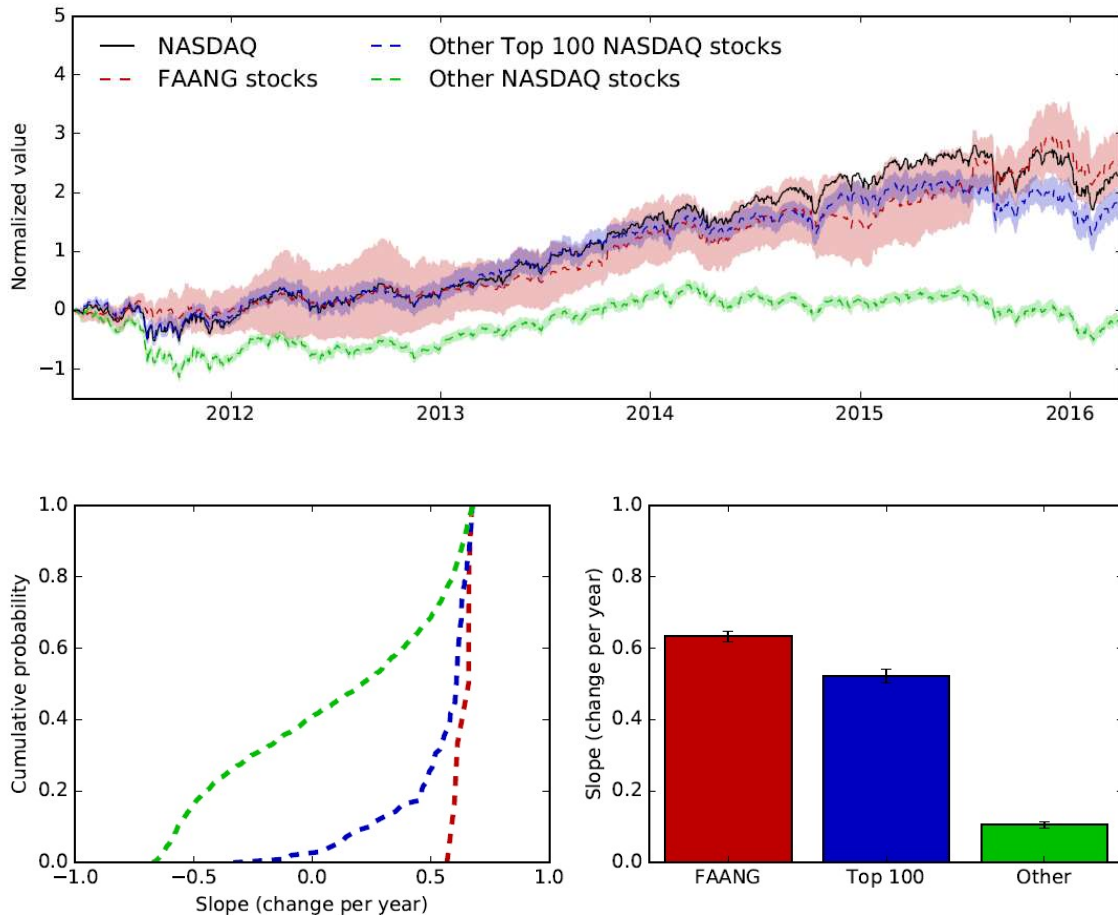


Figure 4: FAANG and other Top 100 stocks show more consistently positive growth than other NASDAQ stocks. (*top panel*) The time-varying change in normalized stock value of the NASDAQ (black), FAANG stocks (red), other Top 100 stocks (blue), and other NASDAQ stocks (excluding Top 100 stocks, black). For the three stock groups, the mean normalized stock price is shown as dashed lines, and the 95% confidence intervals are the shaded areas. Time-varying stock/index traces were normalized first by z-score, and then zero-shifted such that that the first datapoint for each stock/index was shifted to zero. (*bottom left panel*) Distribution of slopes derived from linear fits to each time-varying stock trace (separated by group type). (*bottom right panel*) Summary slopes based on group type (Mean \pm SEM).

Since stock growth is concentrated within a small subset of stocks, one might wonder if this growth is systematically reflected within certain sectors of the NASDAQ. Indeed, both technology and consumer services appear to be overrepresented in both the FAANG and Top 100 stocks (Figure 5). However, neither of these sectors appears to be drivers of growth in the NASDAQ, in which the Finance sector seems to show the most consistent growth out of any sectors (Figure 6). Nonetheless there is not a straightforward breakdown for drivers of NASDAQ growth, at least based on sector.

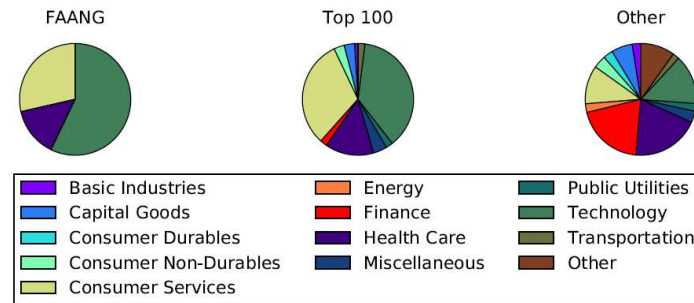


Figure 5: Sector breakdown based upon group type: FAANG stocks, other Top 100 stocks, or other stocks (excluding Top 100).

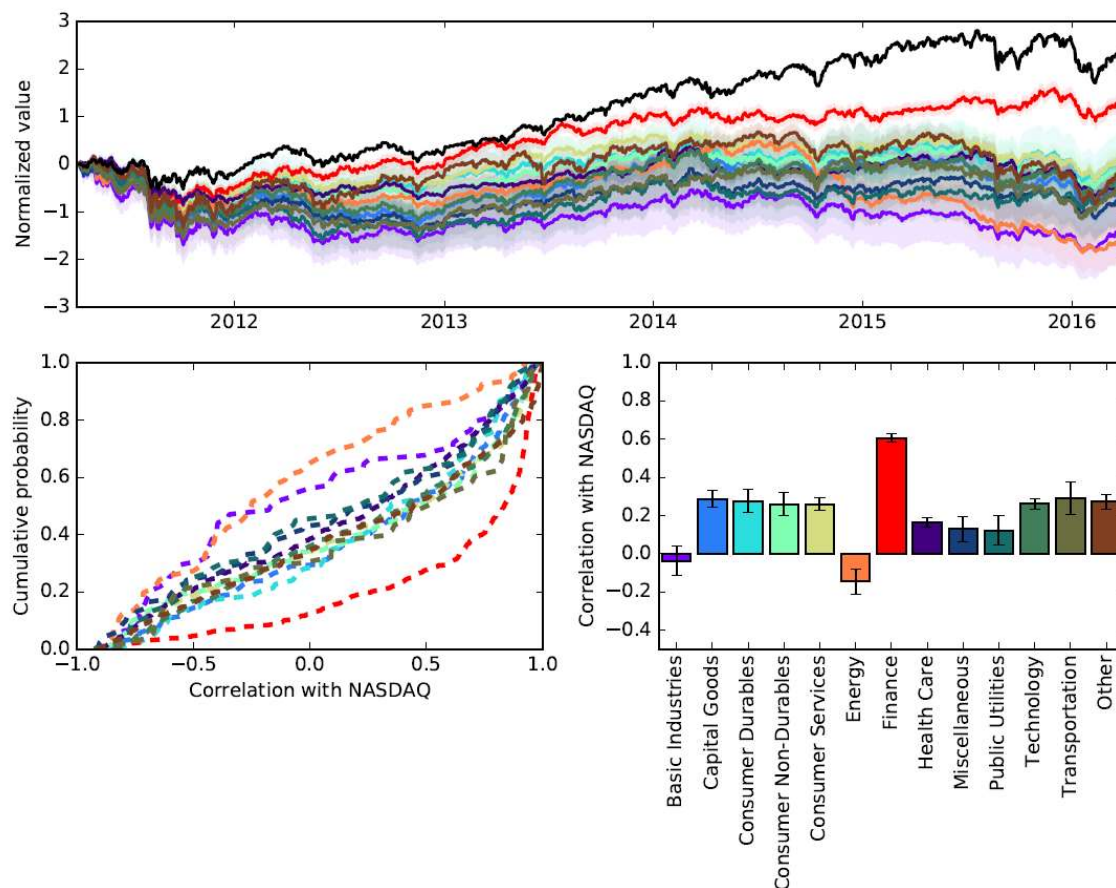


Figure 6: Only the finance sector appears to consistently be more correlated with the NASDAQ than other sectors. (*upper panel*) The time-varying change in normalized NASDAQ value (black) versus different NASDAQ sectors. For each sector, the mean normalized stock price is shown as solid lines, and the 95% confidence intervals are the shaded areas. Time-varying sector/index traces were normalized first by z-score, and then zero-shifted such that that the first datapoint for each sector/index was shifted to zero. (*bottom left panel*) Distribution of correlations with NASDAQ based on sector. (*bottom right panel*) Summary correlations based on sector (Mean \pm SEM).

If stock growth is not systematically concentrated within certain sectors, then another method is needed to reveal which stocks which behave similarly. One direct means is to use dimensional reduction via k-means clustering to reveal systematic trends across different stocks. Interestingly, when k-means is applied to most companies of the NASDAQ (including the NASDAQ itself), the majority of Top 100 stocks, along with ~20% of NASDAQ stocks, fall within the cluster containing the NASDAQ (Figure 7). Furthermore, different time-varying behaviors can be noted. First, the cluster containing the NASDAQ showed systematically positive growth (Cluster 0), while another cluster showed a delayed rise starting in 2013 (Cluster 2). In contrast, two other clusters exhibited systematic declines during 2012 and 2014 (Clusters 3 and 4), while the last cluster didn't change much across the five year period (Clusters 1). While identifying diversity across the clusters is fascinating (and we will deal with it again in the next paragraph), it is worth emphasizing that k-means clustering corroborates the idea that the most consistent stock growth is reflected in a larger subset of NASDAQ stocks (~20%).

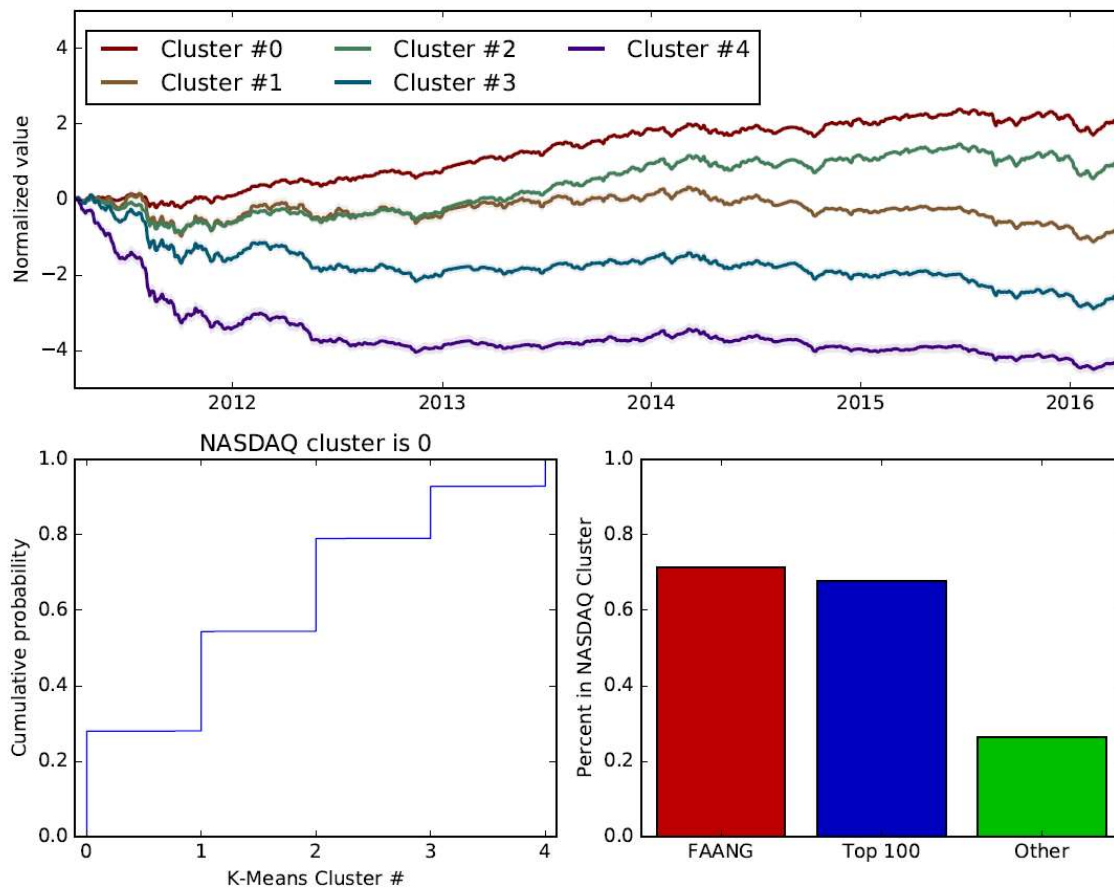


Figure 7: K-means clustering (6 clusters) of all the NASDAQ stock data reveals that different subsets of stocks display similar time-varying behavior. Most of the FAANG and Top 100 stocks are within the same cluster as the NASDAQ. (*upper panel*) The time-varying change in different K-Means clusters. For each identified cluster, the mean normalized stock price is shown as solid lines, and the 95% confidence intervals are the shaded areas. Time-varying sector/index traces were normalized first by z-score, and then zero-shifted such that that the first datapoint for each sector/index was shifted to zero. (*lower left panel*) Cumulative probability of stocks based on cluster membership. (*lower right panel*) The percentage of stock groups present in the cluster with the NASDAQ.

While K-means clustering, an unsupervised clustering algorithm, is a useful abstraction to observe the presence of clustering, a major downside is interpretability (i.e. what do these clusters really correspond to). Thus, a critical next step is determining whether such clustering can also be identified based on a directly quantifiable feature. Based on previous analysis that ~20% of NASDAQ stocks are strongly correlated with the NASDAQ, a

natural starting point would be to split the NASDAQ stocks up into five quintiles based on relative correlation with the NASDAQ. When this is done, the resulting groups qualitatively show a similar clustering to K-Means (Figure 8). The highest quintile shows the most consistent growth (81-100%, red line), whereas the lowest quintile displayed systematic declines (0-20%, blue line). These results compellingly demonstrate how the top 20% of NASDAQ stocks, based on correlation with the NASDAQ, have consistently grown over the last period. Interesting, the next leading quintile (61-80%, gold line) showed systematic increases up to 2014 and then leveled off.

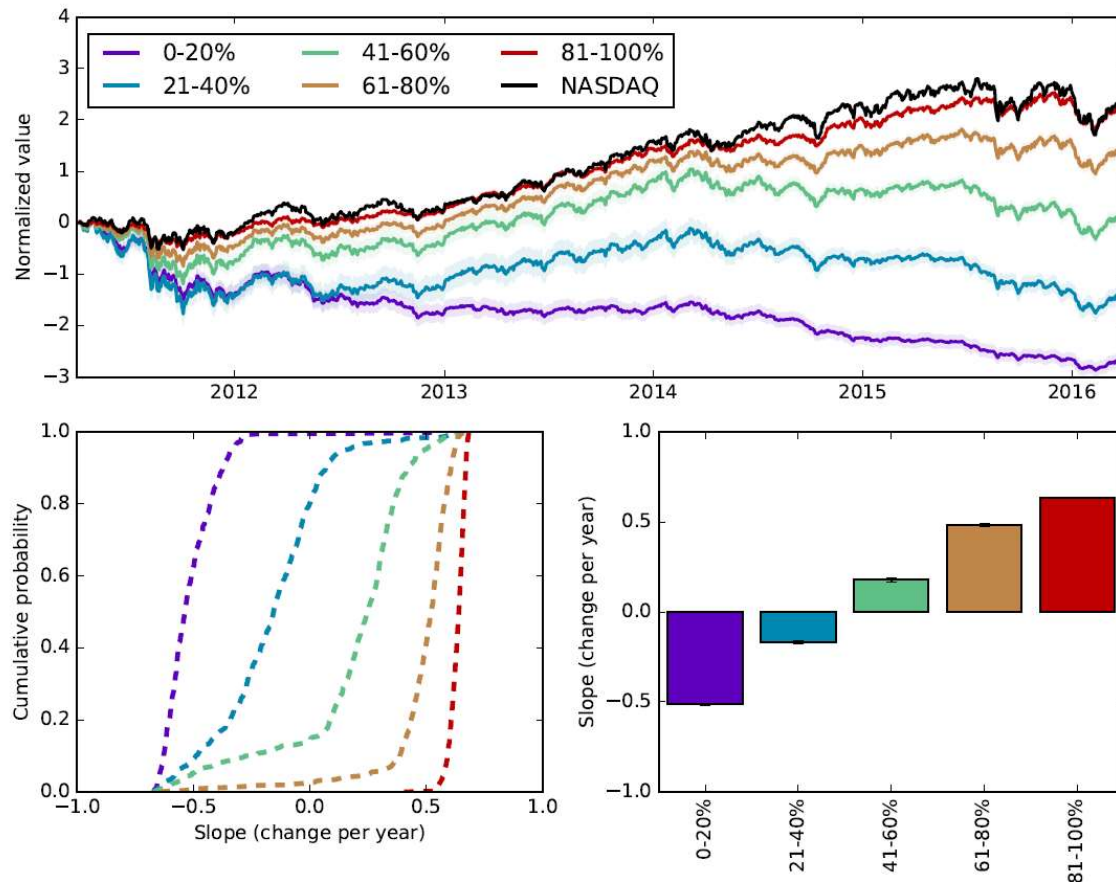


Figure 8: Segregating stocks into five quintiles based on relative correlation with the NASDAQ reveals a similar clustering to K-Means. (*upper panel*) The time-varying change in the NASDAQ and the five different correlation quintiles of stocks. For each quintile, the mean normalized stock price is shown as solid lines, and the 95% confidence intervals are the shaded areas. Time-varying sector/index traces were normalized first by z-score, and then zero-shifted such that that the first datapoint for each sector/index was shifted to zero. (*lower left panel*) Cumulative probability of slope based on correlation quintile membership. (*lower right panel*) The percentage of stock groups present in the cluster with the NASDAQ. (*bottom right panel*) Summary slopes based on correlation quintiles (Mean ± SEM).

Part 2 – Establishing the predictability of stock value

Here we more closely quantify the relationship between individual stocks and the NASDAQ. Preceding analysis established that the Top 100 stocks more closely co-vary with the NASDAQ than do other NASDAQ stocks. However, the analysis only considered how correlated the NASDAQ stocks were with the NASDAQ over the *entire* last five-year stretch, rather than during shorter time periods. When shorter time intervals are considered (i.e. correlations within 4-week intervals), a different picture emerges. Although some companies, such as Apple, show highly positive correlations with the NASDAQ over the full span of the last five years (Figure 1), the time-varying correlation over shorter periods reveals substantial variation. For example, Apple demonstrated even

highly *negative* correlations with the NASDAQ during much of 2013 and 2014, a span in which the company saw large changes (Figure 9, left panels). [3][4]

A straightforward means of quantifying this variation is to estimate the standard deviation of the correlations across the last five years for each stock. Variation was consistently high across all stocks ($r > 0.5$), regardless of whether the stock was a FAANG or Top 100 stock (Figure 10). Despite this high variation, Top 100 stocks remained consistently more correlated with the NASDAQ. Top 100 stocks (including FAANG stocks) exhibited a mean correlation (\pm SEM) of 0.581 ± 0.006 with the NASDAQ, whereas other NASDAQ stocks displayed lower correlation values of 0.250 ± 0.003 . Thus, within the time-varying average, Top 100 stocks remain typically twice as correlated with the NASDAQ than other NASDAQ stocks. Nonetheless, determining meaningful relationships between individual stocks and the NASDAQ must take into account that these relationships are always *in flux*.

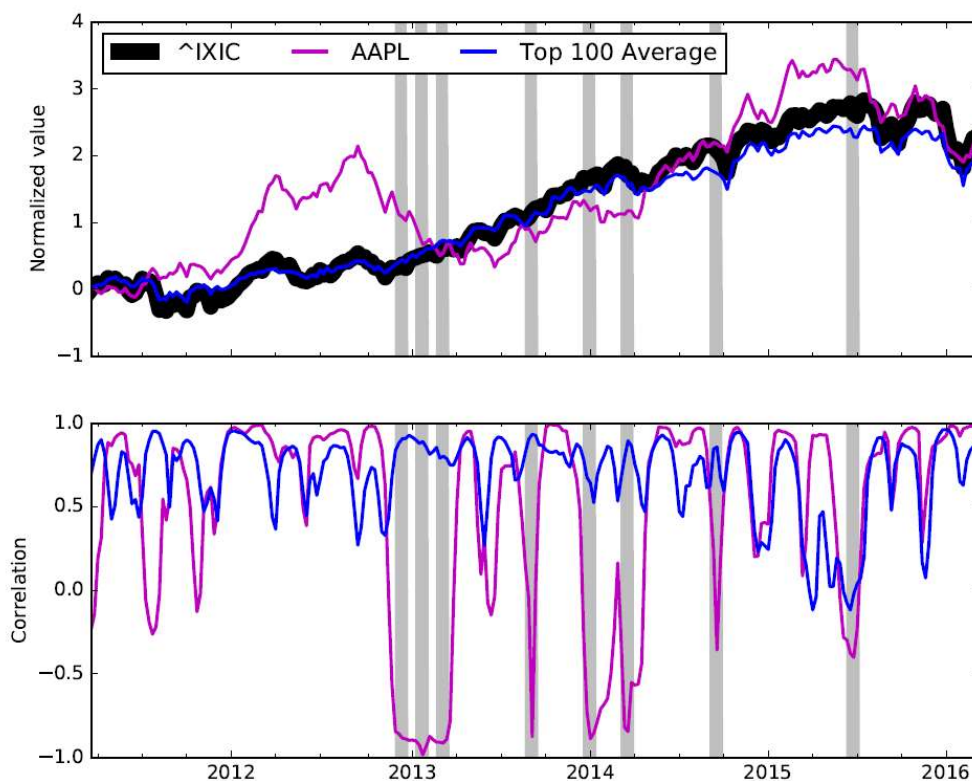


Figure 9: The relationship between an individual stock and the NASDAQ is in flux, whereas a running average of multiple stocks with the NASDAQ is steadier. (top, left panel) The normalized value of the NASDAQ (black line), a representative NASDAQ stock (Apple, blue line), and the Top 100 average (purple line). (bottom panel) The time-varying correlation of the NASDAQ with either Apple or the Top 100 average. The gray shaded regions indicate times when Apple is least correlated with the NASDAQ.

Despite individual stocks showing high variation across time with the NASDAQ, averaging across of multiple stocks tends to result in a steadier relationship with the NASDAQ (for an example see the Top 100 average in Figure 9). This can be quantitatively demonstrated via a bootstrapping approach, in which a random sampling of stocks with replacement is drawn from an increasing number of FAANG, Top 100, or other NASDAQ stocks (Figure 11). Although the correlation of the unweighted running averages with the NASDAQ progressively

improves as additional stocks are added, randomly picking FAANG and Top 100 stocks yields greater correlation with the NASDAQ. Thus, FAANG and Top 100 stocks collectively yield steadier relationships with the NASDAQ.

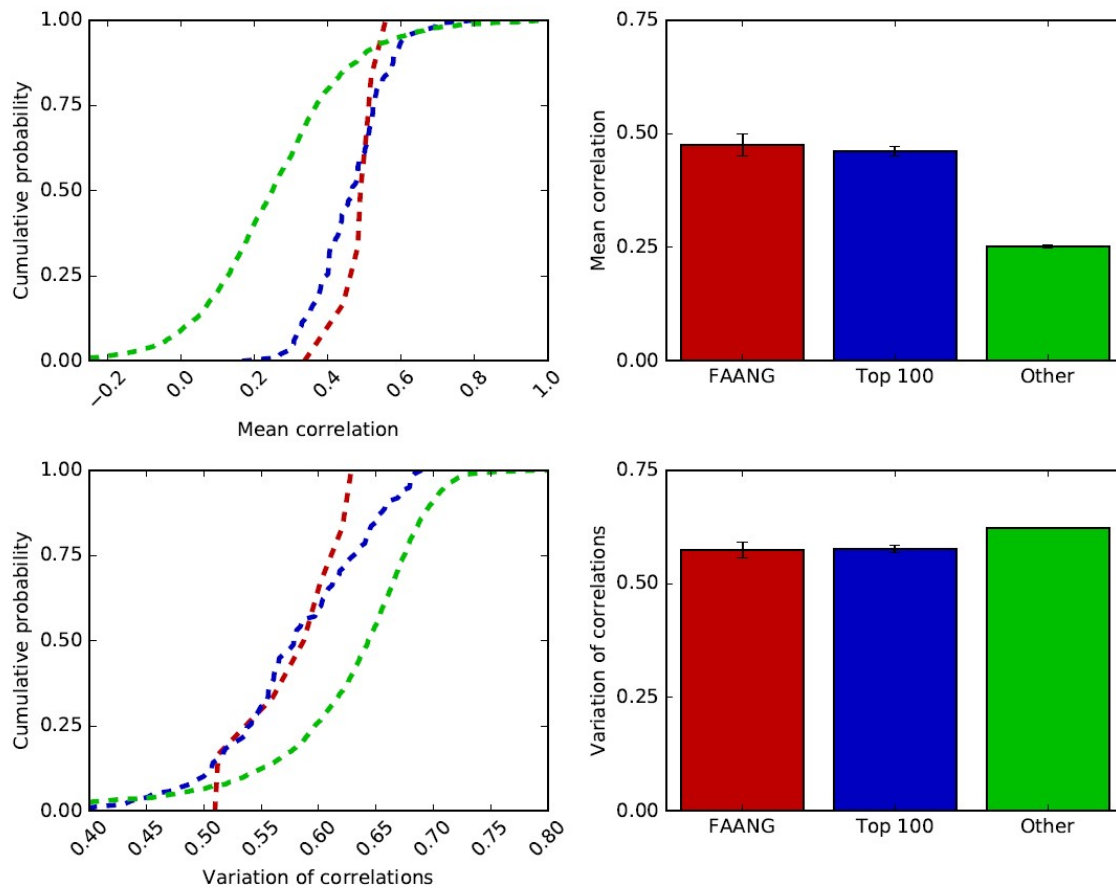


Figure 10: Top 100 stocks, including the FAANG stocks, are consistently more correlated with the NASDAQ. This was quantified by comparing the time-varying correlation within 4-week periods during the last five years between individual stock groups and the NASDAQ. The different stock groups are: FAANG stocks (red, $n=8$), other Top 100 stocks (blue, $n=99$), and other NASDAQ stocks (green, $n=2907$). (*top left panel*) Distribution of the mean time-varying correlation for each stock based on group type. (*top right panel*) Summary mean correlations based on group type (Mean \pm SEM). (*bottom left panel*) Distribution by how much the time-varying correlation varies for each stock based on group type. Variation was quantified as the standard deviation about the mean. (*bottom right panel*) Summary of the variation in correlations based on group type (Mean \pm SEM).

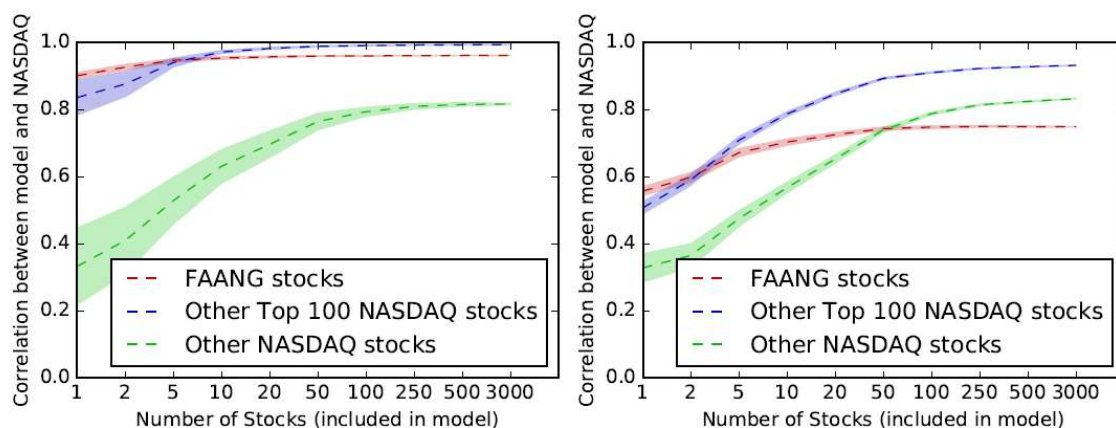


Figure 11: An unweighted running average of multiple stocks improves correlations with the NASDAQ. (*left panel*) The overall mean correlation across 2011-2015 between the NASDAQ and unweighted running averages of an increasing number of randomly selected stocks (FAANG, other Top 100 stocks, and all other NASDAQ stocks excluding the Top 100). (*right panel*) The mean time-varying correlation for 4-weeks intervals (across 2011-2015) between the NASDAQ and an unweighted average of stocks. The shaded regions indicate the 95th percent confidence intervals computed across 100 bootstrap iterations.

What might account for a steadier relationship between FAANG and Top 100 stocks with the NASDAQ? One possibility is that these stocks show more stability in stock value. To investigate this possibility, the autocorrelation of stock value was computed for each stock (Figure 12). FAANG and Top 100 stocks consistently displayed heightened auto-correlation values over longer periods of time than other NASDAQ stocks. The autocorrelations decay by about half at around 48 weeks for Top 100 stocks, whereas other NASDAQ stocks decay by the same amount in a much shorter time span of ~20 weeks. Although the autocorrelations of normalized stock value remain positive on the time-scale of weeks, the autocorrelation for day-to-day variations in stock value rapidly become zero. Because there is so little predictability in day-to-day changes of a stock, relative stock value is more suitable for longer-term forecasting of the market.

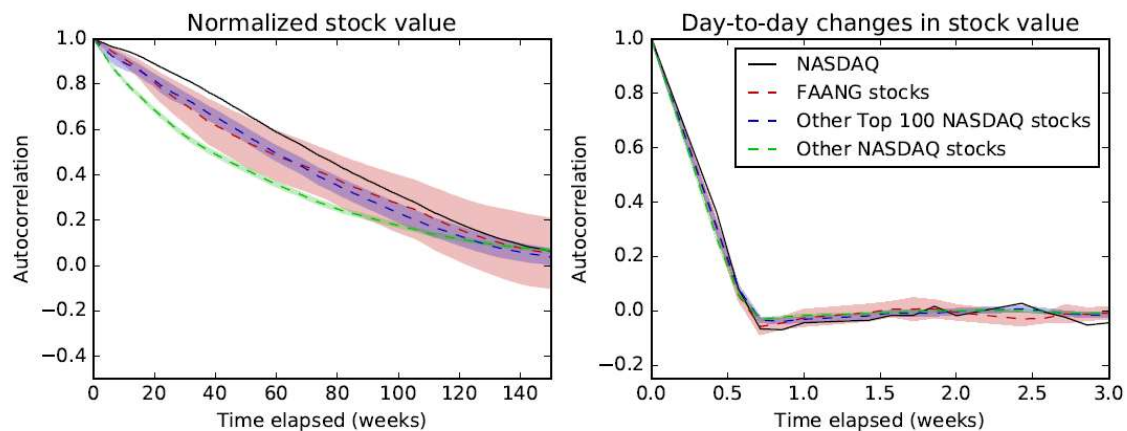


Figure 12: The value of FAANG and Top 100 stocks are more stable over longer periods than other NASDAQ stocks. (*left panel*) The mean autocorrelation of normalized stock value as a function of time and stock group. (*right panel*) The mean autocorrelation for day-to-day variations in normalized stock value as a function of time and stock group. Black indicates the NASDAQ, red indicates the FAANG stocks, blue indicates the top 100 NASDAQ stocks, and green indicates other NASDAQ stocks. The lines indicate the group means, whereas the shaded regions around the lines indicate the 95th percent confidence intervals around the mean.

How might one demonstrate that predictability in stock value can improve forecastability? A relatively straightforward algorithm that can harness historical information is the Kalman filter, a simple recursive Bayesian algorithm that probabilistically incorporates both prior knowledge and new data to make improved predictions. In the simplest implementation, a Kalman filter can make a probabilistic prediction of the NASDAQ's value, and then compare that prediction against the NASDAQ's actual value. An interesting feature of the Kalman filter is that when the filter probabilistically updates its internal estimate of the NASDAQ's true state, the algorithm does not necessarily assume that the NASDAQ's actual value is without noise. Noise in this context merely implies the possibility that the NASDAQ's day-to-day value does not accurately reflect its true state. Markets and stock prices tend to be cyclical, and the theory of mean reversion implies that the market and stocks can temporarily exhibit values deviating from the underlying pattern. This means that when the Kalman filter is optimized to be less-sensitive to short-term fluctuations in the market, the algorithm's predictions can capture longer-term trends or hysteresis in the market (Figure 13).

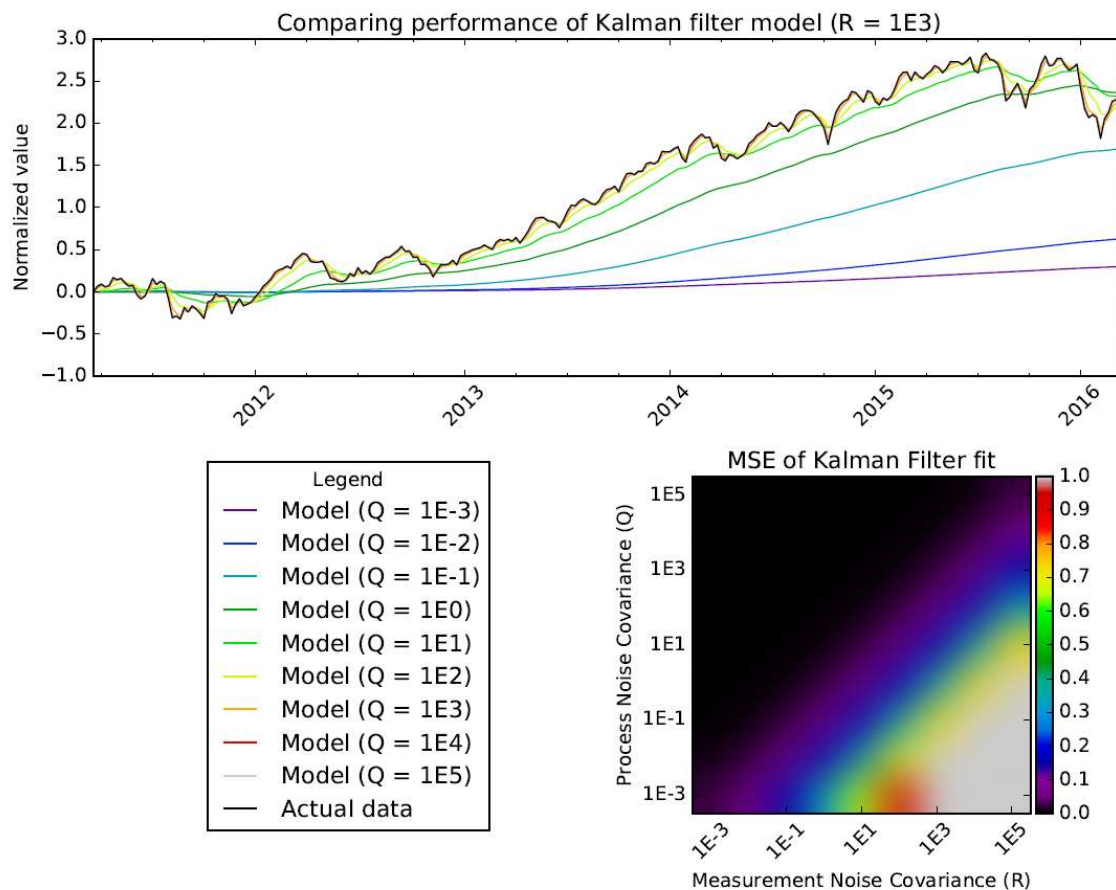


Figure 13: A Kalman filter can demonstrate hysteresis in the market and the feasibility of market forecasting. Here a Kalman filter simply uses the NASDAQ's closing value for the previous week to predict the next week's closing value. The algorithm matches the data "better" when uncertainty in the market's value (denoted by R , or the measurement noise covariance) is either too low or higher than the uncertainty of the Kalman filter's internal model (denoted by Q , or the process noise covariance). Nonetheless when R becomes greater than Q (indicating increasing uncertainty in the NASDAQ measurements), then the Kalman filter becomes less sensitive to day-to-day fluctuations in the market and reflects longer-term trends (see the green trace, $Q = 1E1$).

Part 3 – Market forecasting based on the contributions of individual NASDAQ stocks

The relationship between individual stocks and the NASDAQ have been shown to be correlative, rather than causal. However, demonstrating causal relationships is unreasonable, especially as outside, "third" economic factors like GDP, oil prices, and interest rates, could obscure any casual relationships. Nonetheless, useful information can be derived from understanding well certain stocks can explain the NASDAQ's movements relative to how well other stocks do. In the previous section, we demonstrated that increasing the number of NASDAQ stocks included within an unweighted running average systematically yields closer agreement to the NASDAQ than single stock measurements. This finding could indicate that groups of stocks, rather than individual stocks, are more likely to exhibit steadier relationships with the market by blurring out stock-specific variations. And this intuitively makes sense since NASDAQ is a capitalization-weighted index of all NASDAQ stocks.

But a key caveat is that not all stocks are equally represented by changes in the NASDAQ. So perhaps an additional improvement is to instead weight the running average of stocks. One very simple weighting scheme

could be to weight each stock by that stock's normalized correlation with the NASDAQ (i.e. transforming correlations from a range of -1 to 1 to 0 to 1, so that a correlation of 0 would yield a normalized correlation of 0.5). Unfortunately, a correlation-based weighting scheme consistently results in only a modest improvement against an unweighted average (see goodness-of-fit and MSE values in Figure 14). While other simple weighting schemes, like weighting a stock by its historical market cap data, might provide even closer agreement, simple weighting schemes ultimately assume that a single outside metric can reflect what is likely to be a complicated and dynamic relationship between individual stocks and the NASDAQ.

An alternative scheme is to abandon a simple weighting metric, and instead use linear regression to vary the stock weights systematically with the goal of minimizing the difference between the running average and the NASDAQ. As might be expected, linear regression provides a much closer agreement between the running average and the NASDAQ (Figure 14). Although linear regression can optimally minimize error between a weighted stock average and the NASDAQ, there are a number of major drawbacks to using simple linear regression. First, simple linear regression does not constrain whether the magnitude of the stock weightings can be either small or large relative to other stocks, or even if the sign of the stock weighting is negative. Second, multicollinearity between predictor variables in the model (i.e. correlations between individual stocks) can wreak havoc on the interpretability of the stock weightings. This is because highly correlated stocks are likely to provide similar information about NASDAQ movements and, as such, there is no guarantee that linear regression will equally weight these stocks to find an optimal fit. Both drawbacks mean that a simple regression approach can often yield uninterpretable weighting factors, as well as risk overfitting the model because of over-emphasizing individual stocks.

Fortunately, both drawbacks can be somewhat ameliorated by using Least Angle Regression (combined with Lasso) instead of simple linear regression [5]. In this approach, model error can be iteratively minimized within the constraints that a.) stocks weights must be positive and b.) that the stock weights of equally correlated variables will grow/shrink at the same rate. This is highly desirable, since the final solution is likely to be both more robust and interpretable. Other regression approaches like Lasso or Ridge regression address unrealistic stock weightings by explicitly penalizing weighting coefficients that become too large, but both approaches fail to deal with the problem of multicollinearity. And unfortunately, multicollinearity is likely to be a major issue, since many of the top performing stocks tend to be highly correlated with each other (Figure 15). As a note, while dimension reduction with principal component analysis might be normally a good choice to address the multi-collinearity problem [6], this would unfortunately obscure the effects of individual stocks.

Because stock weightings with Least Angle regression are interpretable, the results of fitting the last five years of stock data with Least Angle regression can reveal the relative importance of a stock in predicting the performance of the NASDAQ. For example, when the FAANG stocks are used to make predictions, the weighting of each FAANG stock is not identical. Instead, stocks like Google and Apple tend to be better predictors of the market than Facebook and Netflix (Figure 16). Nonetheless, all of the coefficients still substantially vary in time, indicating that a single FAANG stock is not a robust predictor of the NASDAQ, at least relative to other FAANG stocks.

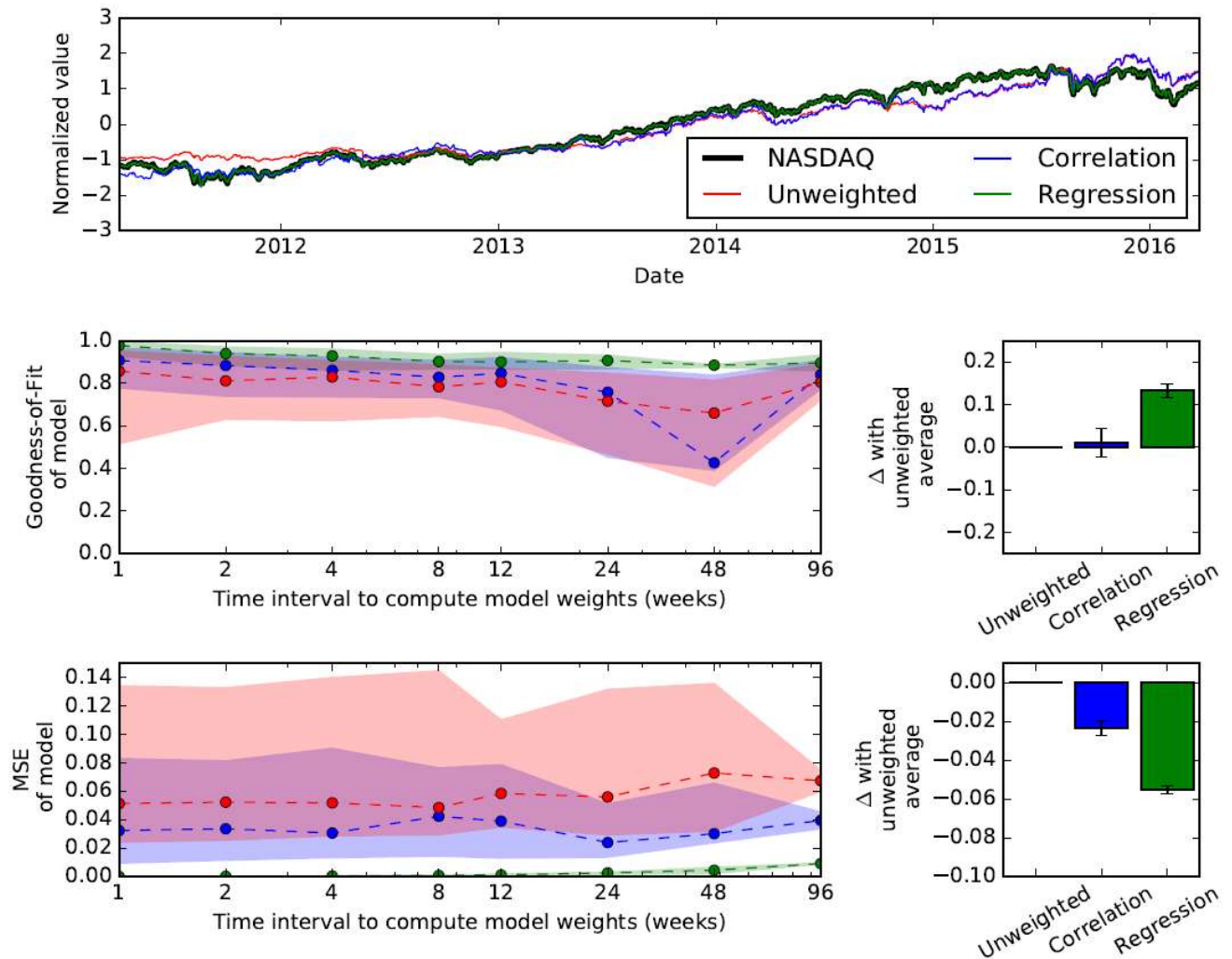


Figure 14: A linear regression-based weighting scheme better predicts the historical changes in the NASDAQ, as opposed to unweighted and correlation-based weighting schemes. (*top panel*) The running average predictions (red – unweighted, blue – correlation, green – linear regression) compared against the NASDAQ. The correlation-based and linear-regression-based weights were updated iteratively every 4-weeks. (*middle left panel*) The median goodness-of-fit comparing the running average predictions versus the actual NASDAQ value. The time-interval to compute and test the model weights was varied from 1 to 96 weeks (as opposed just to 4 weeks shown in the top panel). The shaded regions indicate the 25th to 75th percentiles. (*middle right panel*) The relative difference of the median goodness-of-fit for each running average prediction relative to the unweighted average. Error bars are SEM for the median goodness-of-fit (with respect to how the median goodness-of-fit varies with time interval). (*bottom panels*) Like the middle panels, but instead mean square error between the running average predictions and the NASDAQ are shown.

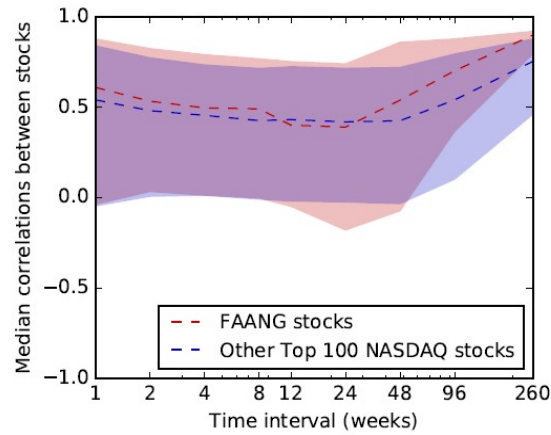


Figure 15: Correlations between top-performing NASDAQ stocks are systematically positive, and strengthen when longer time-intervals are considered. The median correlation value is shown as dotted lines for both FAANG (red) and other Top 100 NASDAQ stocks (blue). Shaded regions indicate the 25th and 75th percentiles.

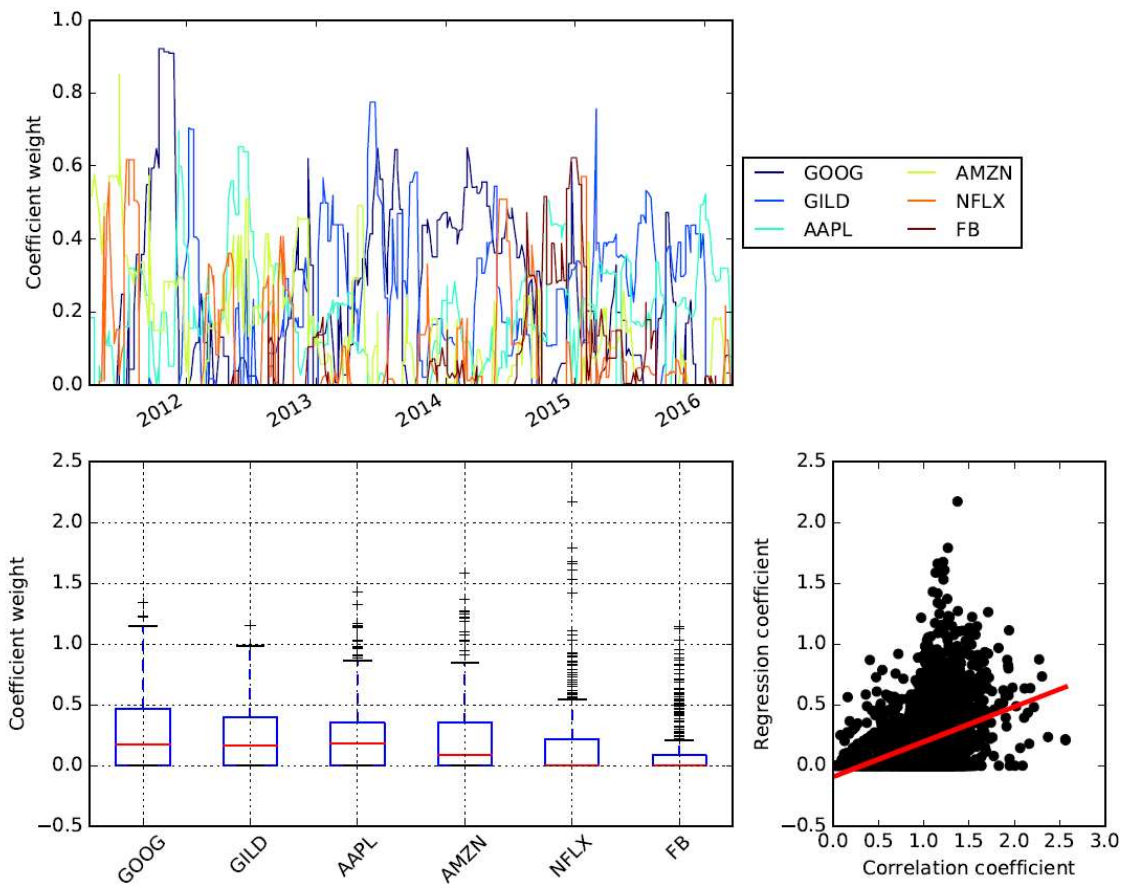


Figure 16: Coefficient weights for Least Angle regression (LARS-Lasso) demonstrate that not all FAANG stocks are equally predictive across the last five years. (*top panel*) The regression coefficients for each FAANG stock as a function of time (alpha regularization parameter = $1.4E-14$). Note: The time-interval to compute regression weights was varied from 1 to 96 weeks, and all the different weights are collapsed together in this single time trace. (*bottom left panel*) Boxplot summarizing the regression coefficients. (*bottom right panel*) A modest relationship exists between regression-based coefficients and the correlation-based coefficients ($r = 0.29$), which suggests that Least Angle regression does not completely ignore the correlative relationship between FAANG stocks and the NASDAQ. Black dots in the scatter plot indicate individual comparisons, whereas the red line represents the best fitting line.

As regression expresses the relative relationship between individual stocks and the NASDAQ in a straightforward way (albeit with some caveats), a natural question might be whether this information has any short-term forecasting power. In support of this hypothesis, the auto-correlation values of FAANG and Top 100 stocks were previously shown to be highly correlated over many weeks. Furthermore, the Kalman filter demonstrated in a practical way that historical stock information can predict future returns. This suggests that predictions made by a regression model might also be relevant across many weeks. To test this hypothesis, a model was generated to forecast the NASDAQ one business week in the future. The only information provided to the model was the stock values for the FAANG stocks at the forecast date (i.e. one week prior), and then weighted by the most up-to-date weights (either unweighted, correlation-based, or regression-based). Surprisingly, while the regression model more closely tracked the actual NASDAQ in terms of mean-square error (Figure 17), all three approaches failed to accurately predict the general trend of the market (i.e. as captured by the goodness-of-fit). Nonetheless, the general trend of the market can be captured by a regression-approach, which is proof-of-concept that the FAANG stocks can principally explain the recent historical changes of the NASDAQ.

But are FAANG stocks privileged relative to other NASDAQ stocks? Previous analysis in this paper suggests no. In support of this hypothesis, forecasting the NASDAQ one business week in the future with a regression-based weighting scheme did not yield a strong difference when other Top 100 stocks were used instead of FAANG stocks (Figure 18-19). However, there are some important differences. First, the mean-square error for the forecasting model was significantly reduced when FAANG stocks were replaced with other Top 100 stocks. But this was only true for the unweighted and correlation-based schemes, and not the regression-based scheme. Second, the regression-based approach did not outperform the correlation-based approach for Top 100 stocks. These findings then suggest the correlation-weighted scheme might actually be able to achieve a near-optimal performance as more NASDAQ stocks are included in the running average. However, when only a few stocks are considered (like the FAANG stocks), a regression-based approach is likely to achieve better results.

Discussion

A deep dive into financial data of the past five years confirms that FAANG stocks are highly predictive of changes in the NASDAQ. While this might seem to confirm the hypothesis that FAANG stocks are uniquely privileged, a more thorough analysis reveals that these stocks are not alone. Instead, a much larger subset of NASDAQ stocks (~20%), which includes FAANG and other Top 100 stocks, is highly related to recent growth in the NASDAQ. This elaboration is important, because it provides some context for what stocks might be contributing to recent market growth.

Although by larger market capitalization, FAANG stocks are definitionally more likely to influence the NASDAQ composite index, other Top 100 stocks also makeup a similar percentage of the NASDAQ (~20% market capitalization for other Top 100 stocks vs. the 25% from just FAANG stocks). Thus, while FAANG stocks as single stock entities are highly visible drivers of market growth, there is a much larger group of other relevant stocks, such as other Top 100 NASDAQ stocks, that collectively influence the NASDAQ. This observation is backed up by the observation that FAANG stocks or other Top 100 NASDAQ stocks tend to yield similarly close forecasts for future NASDAQ growth.

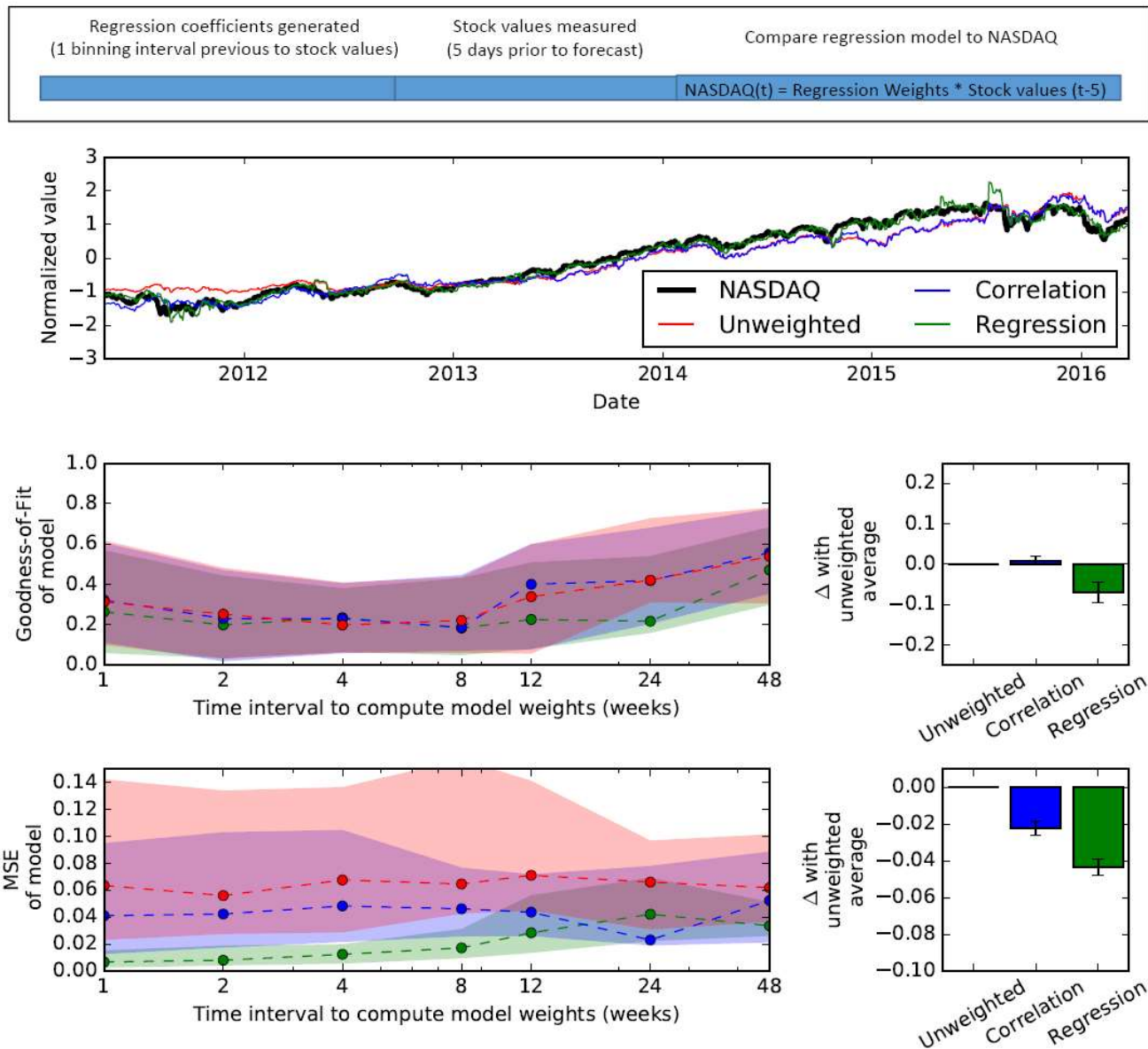


Figure 17: Performance of different running average schemes of FAANG stocks to forecast the NASDAQ one week in the future (5 business days). (top panel) Schematic to forecast the NASDAQ one business week in the future, relative to how FAANGs stocks changed during the forecast day. Stock weights were determined with the most up-to-date weights that were available at the time of the forecast day. (bottom five panels) Same as the panels in Figure 14. See Figure 14 legend for more information.

As has been mentioned multiple times throughout this report, it is worth being mindful why stock correlations with the NASDAQ are important. Based on this report, stocks which are consistently correlated with the NASDAQ have consistently grown over the last five years. Consistent growth is not a guaranteed property of stocks correlated with the NASDAQ, but rather is a by-product that the NASDAQ has also consistently grown over the last five-years. Thus, at least for now, stocks most correlated with the NASDAQ are positive drivers of growth for the last five years.

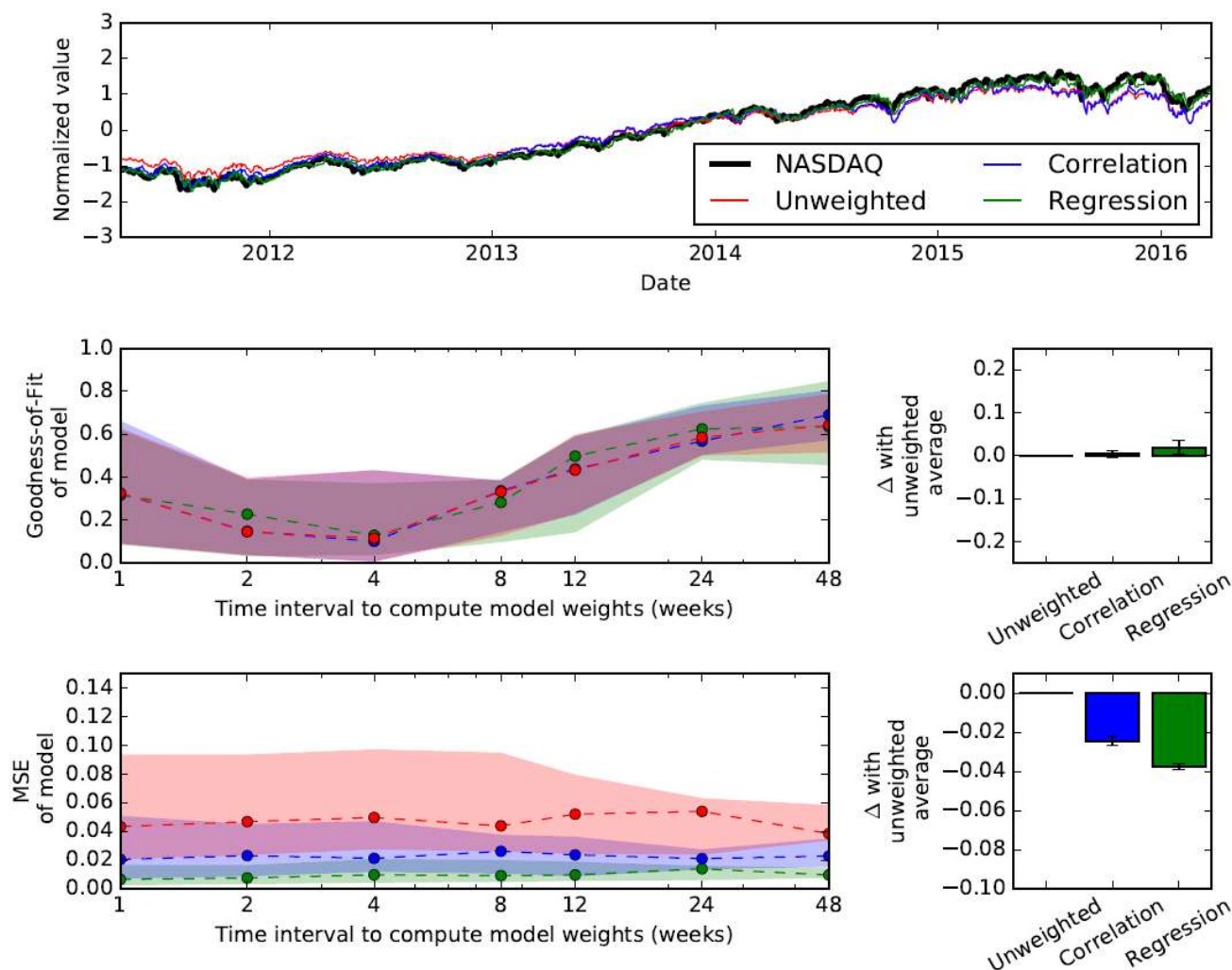


Figure 18: Performance of different running average schemes of Top 100 stocks to forecast the NASDAQ one week in the future (5 business days). (all panels) Same as the panels in Figure 14, except Top 100 stocks (excluding FAANG) are used instead of FAANG stocks. See Figure 14 legend for more information.

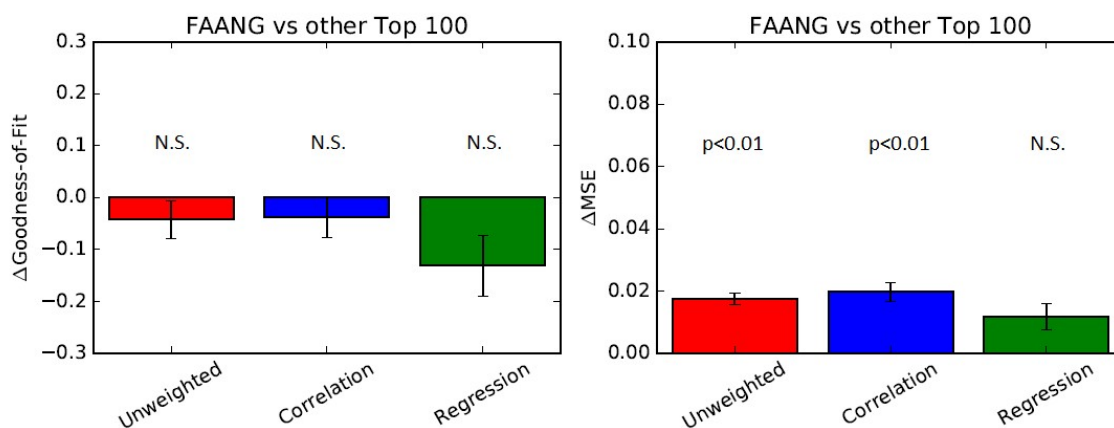


Figure 19: Top 100 stocks significantly outperformed FAANG stocks for forecasting with the unweighted and correlation-based schemes, but not the regression-based scheme. (left panel) The difference in the median goodness-of-fit for each running average prediction depending on whether FAANG stocks or other Top 100 stocks were used. Negative differences indicate better performance

with Top 100 stocks. Error bars are SEM for the median goodness-of-fit (with respect to how the difference in the median goodness-of-fit varies with time interval). Significance of the differences was assessed using a Wilcoxon rank-sum test and each significance value is noted above each bar plot (N.S. = not significant). (*right panel*) Like the left panel, except the mean-square error is plotted. Positive differences indicate better performance with Top 100 stocks.

So then what use are the findings of this report? First, they highlight that FAANG stocks are not singular. Instead, other NASDAQ stocks are highly relevant investment options because they too have consistently grown over the last five years. Second, the analysis described in this report could be just as relevant to understanding the relationship between an investor's portfolio and the underlying stocks making up the portfolio. While it is easy for an investor to look at his/her portfolio and spot stocks that are driving the strongest growth in his/her portfolio, it's much harder to determine whether the portfolio is optimally diversified. This is challenging because different stocks are likely to drive differentially changes in an investor's portfolio at any time. Fortunately, one of the tools developed in this report is determining the time-varying correlation of stocks to the NASDAQ. This tool can help in determining how well correlated individual component stocks are to each other, and an investor's portfolio.

Another tool that is likely to be useful is employing the Kalman filter. Because stock value has hysteresis, it's likely possible to forecast future changes in a stock's value. As was mentioned earlier, individual stocks experience fluctuations unrelated to the actual value of a company. The Kalman filter has been shown that it can be optimized to alert an investor when a stock begins to experience a major fluctuation that dramatically differs from the stock's recent history [7]. This information can be immensely helpful to gauge whether a stock is temporarily over-priced or under-priced relative to its recent history. This is especially true when an investor is interested in selling that stock, because the most optimal strategy is to sell a stock whenever it is temporarily over-priced. The converse is true when an investor is interesting in buying a stock when it is under-valued. Future work will be focused in this area, because ultimately the best investment tools are those that lead to actionable intelligence that an investor can use.

Data sources for capstone project:

1. Historical stock data – Data was acquired from Yahoo.com by using the built-in pandas web-reader.
2. Current market cap data – Data was acquired from Yahoo.com by using the built-in pandas web-reader.
3. NASDAQ stock list – A CSV file containing this information was downloaded from the NASDAQ website on 5/21/2016.
 - Web link: <http://www.nasdaq.com/screening/companies-by-industry.aspx?exchange=NASDAQ&render=download>

References:

- [1] D. Strumpf. Wall Street Journal. "The Only Six Stocks That Matter". <http://www.wsj.com/articles/the-only-six-stocks-that-matter-1437942926> (July 26, 2015).
- [2] A. Stevenson. CNBC. "Cramer: Investors can't resist the lure of FANG". <http://www.cnbc.com/2016/06/28/cramer-investors-cant-resist-the-lure-of-fang.html> (June 28, 2016).
- [3] B. Hein. Cult of Mac. "The Biggest Apple Stories Of 2013 [Year In Review]" <http://www.cultofmac.com/257364/entire-year-apple-news-one-post/> (December 31, 2013).

- [4] A. Heath. Cult of Mac. "The 12 biggest moments for Apple in 2014"
<http://www.cultofmac.com/307085/apple-2014-review/> (December 26, 2014).
- [5] Efron, Bradley; Hastie, Trevor; Johnstone, Iain; Tibshirani, Robert (2004). "Least Angle Regression" (PDF). *Annals of Statistics*. **32** (2): pp. 407–499. doi:10.1214/009053604000000067. MR 2060166.
- [6] Dodge, Y. (2003) *The Oxford Dictionary of Statistical Terms*, OUP. ISBN 0-19-920613-9
- [7] Rick Martinelli, Haiku Laboratories, "MARKET DATA PREDICTION WITH AN ADAPTIVE KALMAN FILTER"
<http://www.haikulabs.com/mdpwakf.htm> (1995).