From Worst to Most Variable?
Only the worst performers may be the most informative

Bradly Alicea
Orthogonal Research
bradly.alicea@outlook.com

ABSTRACT

What makes a good prediction good? Generally, the answer is thought to be a faithful accounting of both tangible and intangible factors. Among sports teams, it is thought that if you get enough of the tangible factors (e.g. roster, prior performance, schedule) correct, then the predictions will be correspondingly accurate. While there is a role for intangible factors, they are thought to gum up the works, so to speak. Here, I start with the hypothesis that the best and worst teams in a league or tournament are easy to predict relative to teams with average performance. Data from the 2013 MLB and NFL seasons plus data from the 2014 NCAA Tournament were used. Using a model-free approach, data representing various aspects of competition reveal that mainly the teams predicted to perform the worst actually conform to expectation. The reasons for this are then discussed, including the role of shot noise on performance driven by tangible factors.

Introduction

What is the link between performance and *a priori* predictions? At the individual level, this question has important implications for areas as diverse as sports [1] and job [2] performance potential, including the so-called "Moneyball" approach. This may also be useful for understanding the effects of exercise, education, and technological augmentation on human populations.

This study was approached with a working hypothesis: performance of sports teams that are perennial winners or losers are much easier to predict than teams that exhibit parity. The notion of outlier is important here. For example, if a team is predicted most likely to "win it all" on the basis of its roster and schedule, the role of unexpected factors such as injuries or streaky play should be minimized. While this sort of prediction noise is expected to at least be nominal, it should leave a statistical fingerprint. As we shall see,

In an attempt to form theoretical insights based on this question, I conducted a rudimentary analysis on how informative prognostications of the 2013 MLB season, 2013 NFL season, and 2014 NCAA basketball championship were with respect to the final regular season standings [3]. A similar analysis was done on NFL data to see if these results hold across types of data. The MLB and NCAA datasets constitute predictions made by the PredictWise (see Methods, Section 1) website (http://www.predictwise.com) and regular season/tournament outcomes. The NFL dataset consists of predictions made by the Sporting News and regular season outcomes.

To initially test this hypothesis and then replicate it in a slightly different context (multi-tiered tournament), I used information from PredictWise. PredictWise is an aggregator of likelihoods for purposes of betting on outcomes. Their predictions include contests in the realm of politics, sports, and entertainment. The likelihoods are updated as the event unfolds, but the

comparison of *a priori* predictions provides interesting comparisons with the final outcome. These predictions are not entirely naive, but do rely upon a fair number of assumptions. I used information from the *Sporting News* NFL data as a means to replicate the results using a different source of predictions, which provides robustness with respect to prior information.

2013 MLB Regular Season and World Season

The first graph shows the difference in rank-order position between the likelihood of winning the World Series (generated *a priori*) and the regular-season won-loss record. The "difference from prediction" was then calculated for the top, middle, and bottom tercile on teams based on their regular-season record (Figure 1).

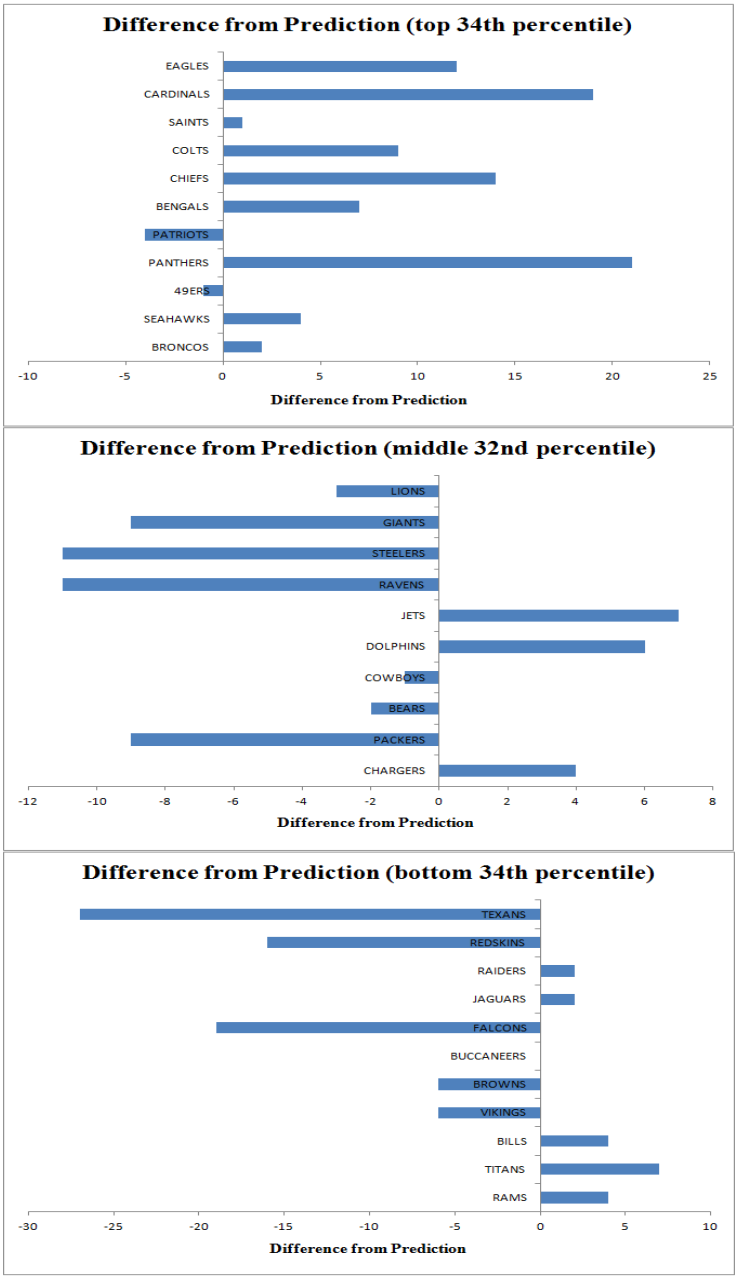


Figure 1. Difference from Prediction for the top, middle, and bottom terciles of the MLB.

Interestingly, many of the winningest teams were not predicted to finish strongly. By contrast, the bottom tercile was equally represented by teams that had the least chance of winning it all and teams that were supposed to finish more strongly. With a few exceptions, the middle tercile was represented by underachieving teams, and the most consistent performances (smallest deviations from prediction) were among the lowest achieving teams (Supplemental Figure 1).

The next two graphs (Figures 2 and 3) show the magnitude of deviation from a given prediction. This is used as an indicator of predicted performance vs. observed performance. This results in an index (value: 0-1) based on a team's deviation from prediction relative to the maximum and minimum of all teams in the league. The third graph (two panels) breaks this down into teams that finished better than and worse than expected.

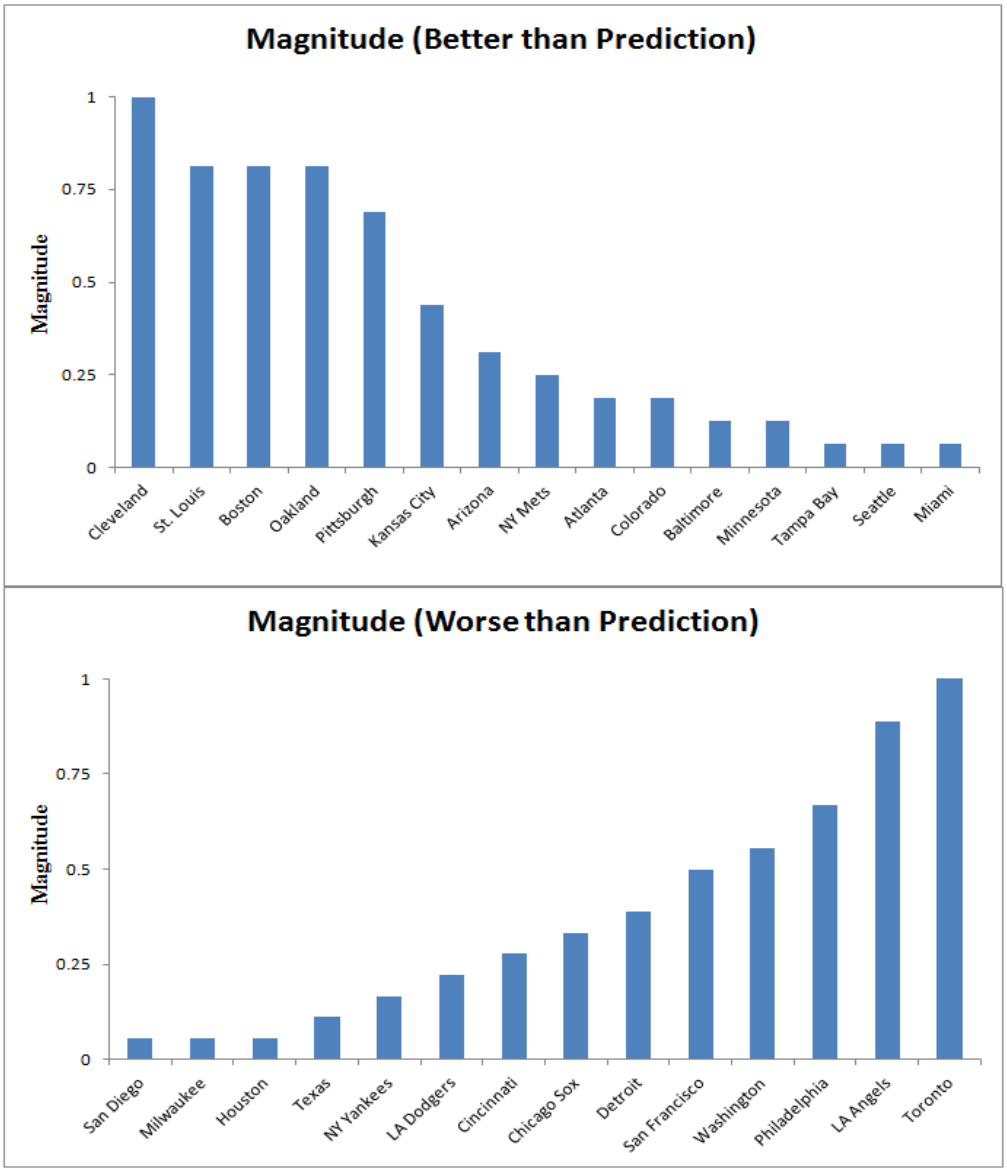


Figure 2. Reformulation of Supplemental Figure 1 broken out by teams that did better than prediction (top) and worse than prediction (bottom).

Finally, Figure 3 demonstrates how the deviation from prediction is related to the total number of wins a team had during the season. While the plot in Figure 3 lends no additional support (R^2 = 0.1609) to the initial hypothesis, but is consistent with the notion of "worst performers, best predictors".

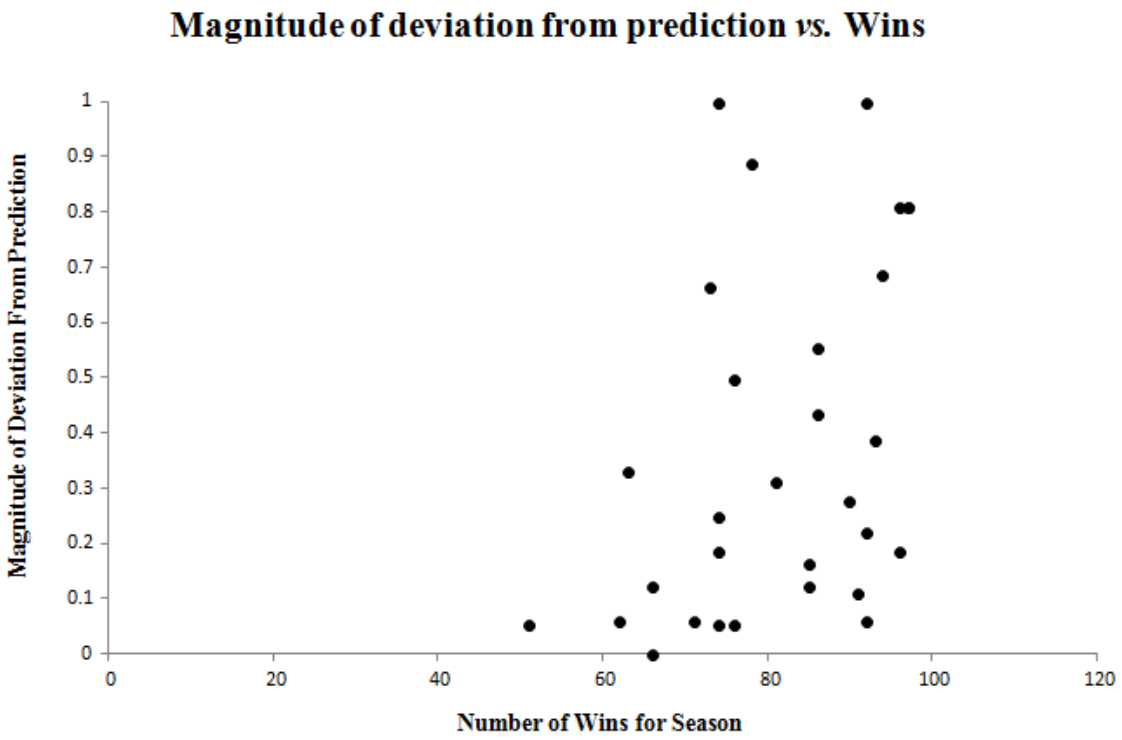


Figure 3. The number of wins in the 2013 season as predicted by the magnitude of deviation for each team.

NFL 2013 Season Performance

To compare these tendencies across sports and odds-making enterprises, I used the Sporting News *a priori* predictions for the NFL 2013 season [4]. In this example, I compared a team's *n*-to-1 odds of winning the Super Bowl with the final season standings. This analysis used a methodology similar to the MLB analysis, but used a different source of predictions (Sporting News). While such a methodology ignores the detail that not all teams with the best regular season record win the championship, pre-season (e.g. *a priori*) predictions are made without regard for this distinction.

From the exploratory graph in Figure 4, a similar trend of "worst performers, best predictors" emerges, albeit with more outliers on the lower end. Recapitulating the difference from prediction analysis done for the MLB data (Figure 5), the NFL data shows more deviations from prediction for every stratum of the dataset. However, again, there is a slight tendency for the bad teams to be predicted correctly and the best performing teams to be poorly-predicted. In the case of the NFL data, there is a countervailing "dynasty" effect as well: teams that have been winning consistently were also predicted to do well. As they met this expectation, they were easier to predict correctly.

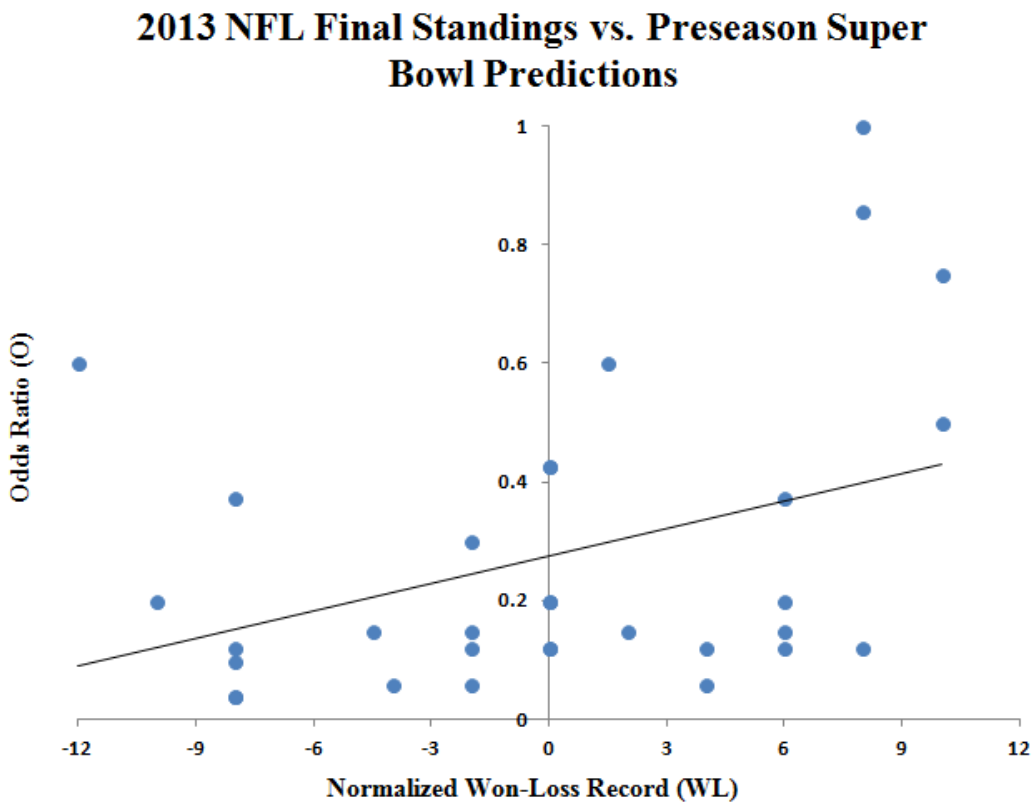


Figure 4. Performance for NFL teams during the 2013 season against preseason predictions. For the mathematical definitions of each measure used, please see Methods, Section 2.

NCAA Tournament Data

Our working hypothesis was replicated using 2013 NCAA Tournament prediction from PredictWise. NCAA bracket predictions were used because they notoriously hard to predict correctly. For our purposes, correct prediction is secondary to better characterizing why predictions are mostly incorrect. In this dataset, predictions were staggered across the tournament duration, and track the unexpected triumph and defeat of teams at five different stages of the tournament. Rather than a simple comparison of prediction and performance, we have an iterative record of how prediction and performance can unpredictably fluctuate (Supplemental Figure 2).

The change in the chance of a team winning across stages of the tournament can be further examined in terms of its initial seeding. In Figure 6, changes in the likelihood of tournament victory can be contrasted across four seed groups (1-4, 5-8, 9-12, and 13-16). Teams of seed 9 and below begin with a very low chance of winning, and this trend remains across the tournament span. Teams that being the tournament with a seed anywhere from 1 to 8 exhibit far more fluctuation in their predicted performance over the course of the tournament. The teams that either went farthest n the tournament or were eliminated early were from these two seed groups.

Again, the basic pattern emerges from the data: teams predicted to finish at the bottom tend to fulfill expectations, while teams predicted to finish above the 50th percentile are much

harder to accurately predict. This can be shown in MLB season-long data, NFL season-long data, and NCAA tournament data. But what would explain this pattern of outcomes given a set of competing teams and a set of informed predictions across so many different contexts?

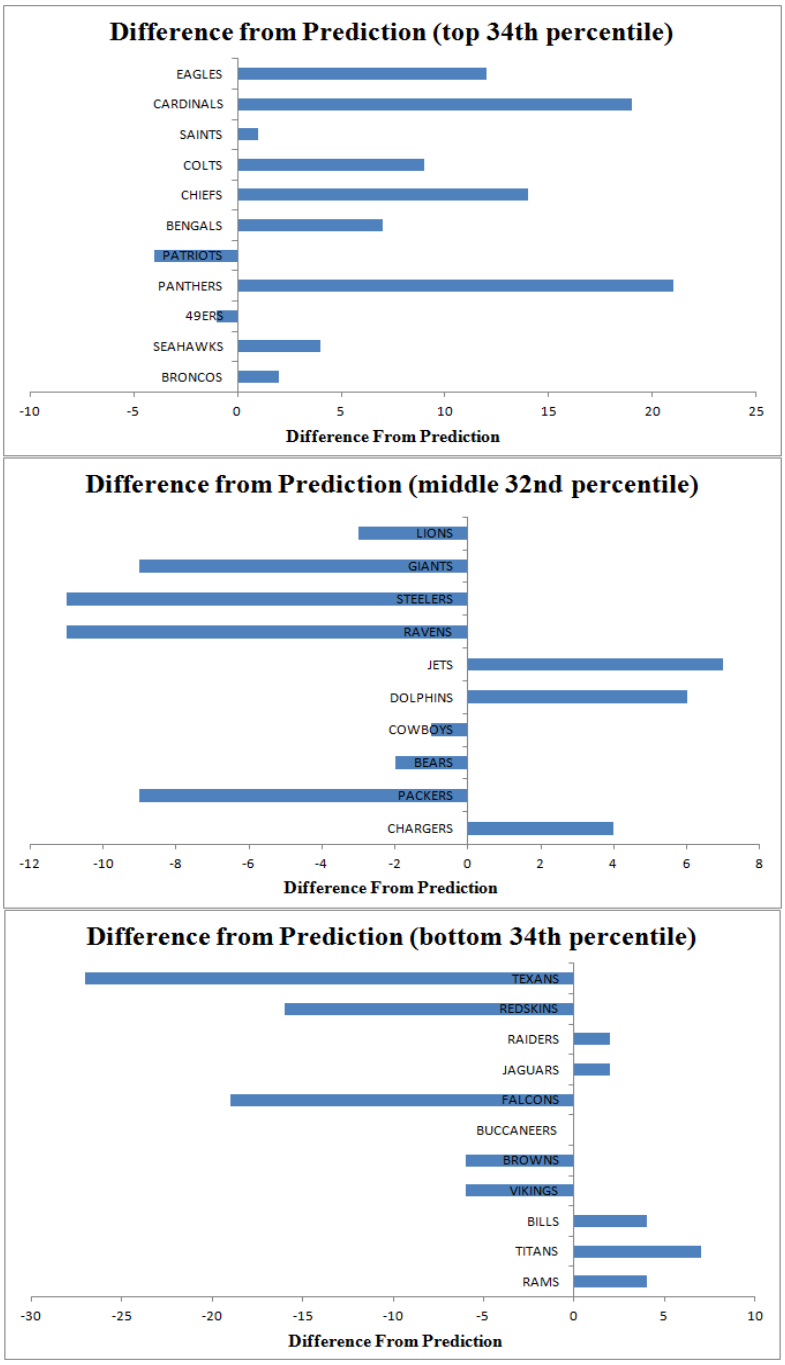


Figure 5. Prediction vs. performance for the 2013 NFL season by tercile.

A noise-driven model of performance

To better understand this, we will use the laws of small and large numbers to contrast the effects of noise on the performance of perennial losers vs. competitive teams. Consider that significant noise has a much larger effect on small numbers than large numbers. This noise,

represented by fluctuations in the small number over time, has a larger effect on performance when the number is small. Most of this is due to the scale of the noise relative to the number of objects (size) being affected. In the case of performance, the reverse may be true. This is due to noise in performance resembling shot noise, which is a form of random fluctuation which increases in its effect size with the size or intensity of the corresponding system.

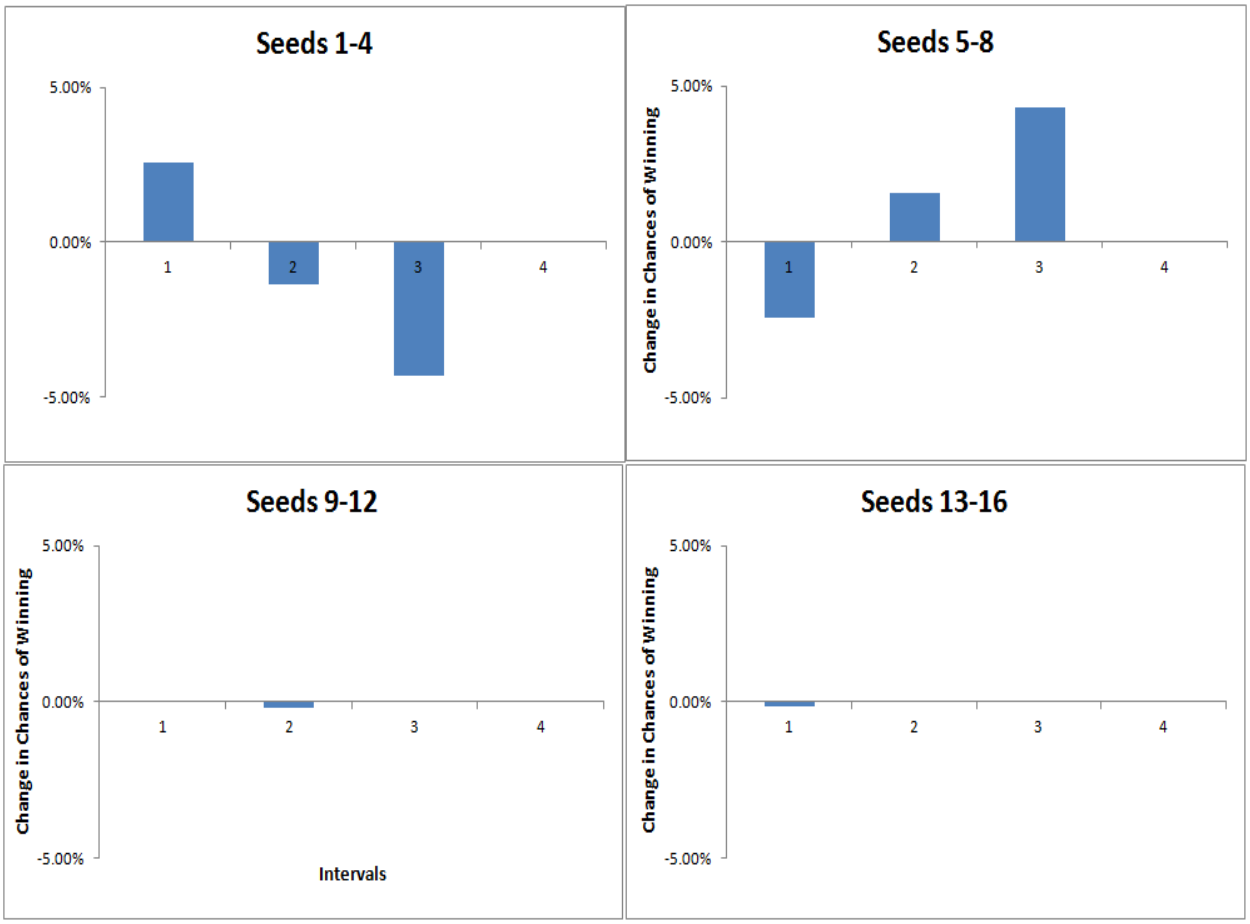


Figure 6. Change in the chances of winning by group of seeds. Each seed group consists of 16 teams (4 per seed number, and 4 seed numbers).

Consider a case in which the initial informed prediction is close to zero. Even when performance is dominated by noise (ranging from 2P to P/2), fluctuations such as winning/losing streaks, hot/cold hands, or being "in the zone" versus outside of it have almost no ability to add competitiveness. This can be called "worst-deterministic" performance. However, when the initial informed prediction is above the 50th percentile of all individuals/teams being predicted (e.g. average to above-average performance), the same kind of noise that was unhelpful for the teams close to zero initial prediction becomes useful for giving teams predicted to be merely average a shot at championships. However, noise can also make a team predicted to do very well underperform. The best way to think of this is as a form of "*competitive-shot noise*" performance, and can be contrasted with "*worst-deterministic*" performance. Overall, the very factors that drive predictions of high performance are also susceptible to fluctuations (day-to-day performance) and noise (extenuating circumstances).

Discussion

So are there better means to predict outcomes than making odds? PredictWise uses a combination of *a priori* odds-making and individual wagering. When people are willing to wager on an outcome, a diversity of mental models are used to inform the prediction. We can also use real-time surveys that make predictions in a manner similar to a logistic regression model [5]. However, whether such approaches can ameliorate the "surprise" factor of unexpected levels of performance (good or bad) is questionable.

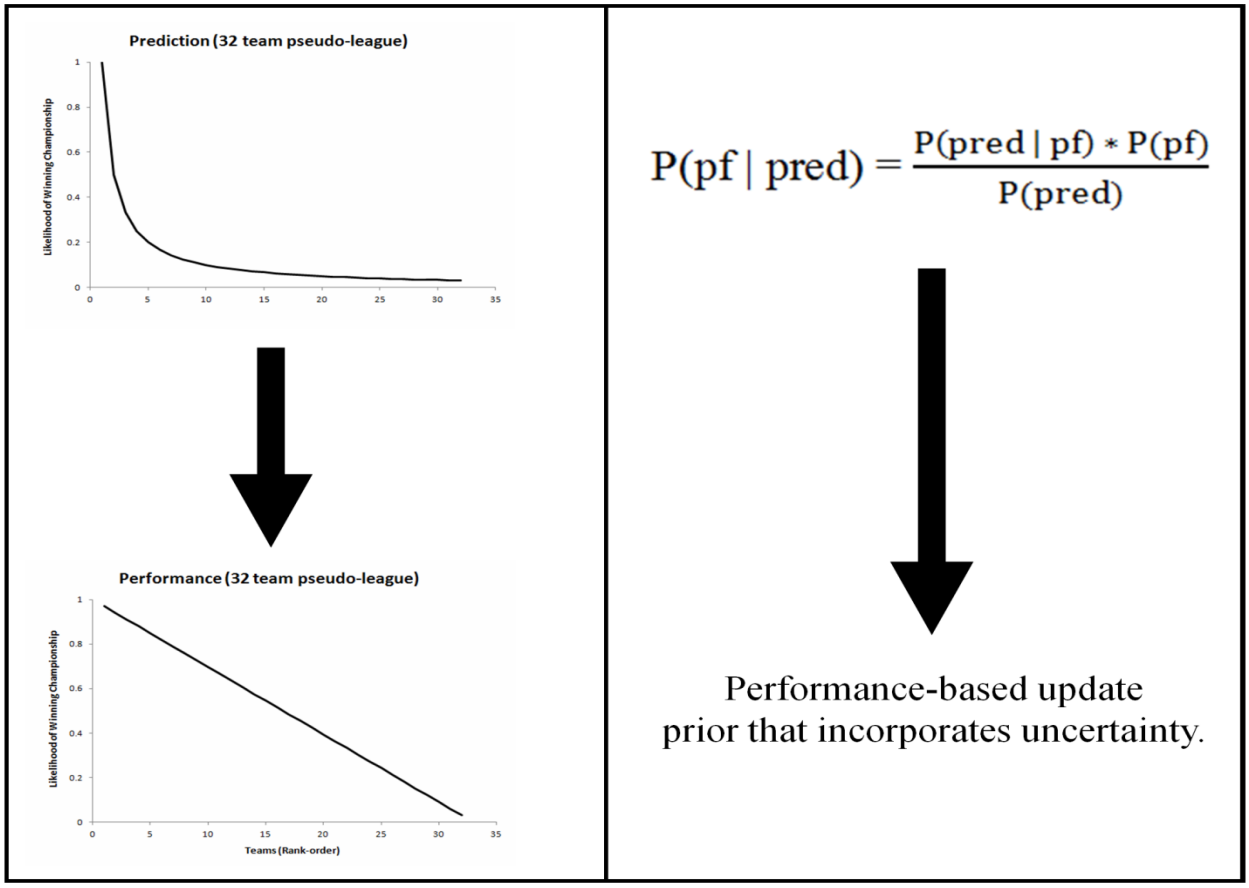


Figure 7. Towards a predictive model which maps prediction to actual performance given noise (left). While a Bayesian model (right) would perhaps be the most informative approach, the Bayesian approach does not incorporate wideband noise (the basis of the proposed model in Methods, Section 4) effectively.

While these data are suggestive, they alone do not resolve whether only the worst teams conform to expectation or if the very best teams are also easy to predict. These outcomes might be due to a secondary phenomenon (e.g. dynastic effect). But they might also be an artifact of prediction methodology, the competition structure, or something else entirely. One way to clarify these results is, of course, to include more data from various contexts. However, another option is to construct a predictive model (see Figure 7). Given the role of intangibles in performance, one possibility is a model for bounded prediction that incorporates wideband noise that treats intangible factors as a source of systematic noise See Methods, Section 4 for formalism. The bounded prediction model is suitable alternative to a Bayesian formulation in that potential prior

distributions cannot properly account for the inherent presence of systematic noise (see Methods, Section 3 for a more detailed explanation). It also supplements network-based and combinatorial predictive models by providing the conditions for victory in a given tournament structure [6].

In conclusion, we can extend this hypothesis from sports performance to realms such as educational and other forms of competitive performance where diversity and innumerable intangible factors dominate performance. It cannot be stressed enough that the results presented here argue against systems that incentivize or otherwise reinforce elitism through the all-or-nothing reward of top performers. In fact, the data suggest that most competitors have enough potential value to be incentivized regardless of outcome for a particular instance or competition. More work needs to be done to understand this tendency in more detail, but still may serve as an instructive guiding principle for many areas of competitive performance.

METHODS

Full dataset for all seasons, tournaments, and statistical models available on Figshare (doi:10.6084/m9.figshare.944542).

Section 1: PredictWise and Betfair predictions

The predictions for PredictWise and Betfair were made on the basis of their specific methodologies. Likelihoods are made in part on the basis of popular wager: given a particular proposition (e.g. "will the Tigers win the World Series?"), a wager was elicited. The entire set of teams being wagered on (T) constitutes 100% of all wagers. A team's likelihood to win, then, was the proportion of wagers that team (in the form of a binary proposition to win the championship) received.

Section 2: Methodology for calculating magnitude of deviation from prediction

The graph in figure 4 includes two measures: x and y. The variable *WL* is defined as the final won-loss record centered upon a .500 (e.g. 8-8) record. This can be described mathematically as

$$\mathbf{WL = W - L + T(0.5)} \quad \textbf{[1]}$$

where W is wins, L is losses, and T is ties. The variable O is an index based on the odds ratio, where the lowest odds are set to 1.0. This can be defined mathematically as

$$\mathbf{O = (O_i \,/ O_{min})^{-1}} \quad \textbf{[2]}$$

where O_i are the odds for a particular team or individual, and O_{min} are the lowest odds. Rank-orderings of these metrics (and distances between these rank-orderings) were also used to generate the graphs.

Section 3: PredictWise NCAA tournament prediction sampling

The predictions made by PredictWise (and subsequent tournament performance) were sampled at the following points: 64 teams, 16 teams, 4 teams, 2 teams, 1 team (champion). PredictWise bases the likelihood of tournament victory on a team's share of 100%. For example, at the 4 team sampling interval, the likelihoods for tournament victory might be 45%, 25%, 20%, and 10%. While these numbers are contingent upon their previous likelihoods, they also tend to skew the team's performance at the 4 and 2 team sampling intervals.

Section 4: Model for bounded prediction

The equation for calculating change in rank (CR) between the predicted likelihood of coming out on top and the overall rank-order position of performance can be stated as

$$\mathbf{CR = pred - obs} \quad \textbf{[3]}$$

Change in rank can also be expressed as a likelihood that describes how likely it is that performance is worse than expected. This can be stated mathematically as

$$\mathbf{p(CR)} = \frac{\mathbf{1}}{\boldsymbol{n}*|\boldsymbol{LR}|} \quad \textbf{[4]}$$

Since change is rank is an intangible performance factor between teams, adaptively changing the prediction due to performance information can be modeled using uniform wideband noise. The overall change in rank reformulated as a posterior update of the prediction given performance is

$$\boldsymbol{\delta_w} = \boldsymbol{pred_u} - \boldsymbol{pred_L} \quad \textbf{[5]}$$

where $pred_u$ is the upper bound and $pred_L$ is the lower bound of the noise, respectively. From this distribution of noise, we can draw a range of new predictions.

This range can be conceptualized as a binomial distribution. The likelihoods of selecting a point from the mass of this distribution are as follows.

$$\boldsymbol{BP} = \begin{cases} \mathbf{pred\ (1 + CR) = \ \delta_{w1}} \\ \mathbf{pred\ (1 - CR) = \ \delta_{w0}} \end{cases} \quad \textbf{[6]}$$

with the parameter δ_{wn} describing the location of this prediction in the distribution.

REFERENCES:

[1] Morey, D. The elephant on the court. Economist, April 20 (2012).

[2] Armstrong, J.S. Predicting Job Performance: the Moneyball factor. Foresight, Spring (2012).

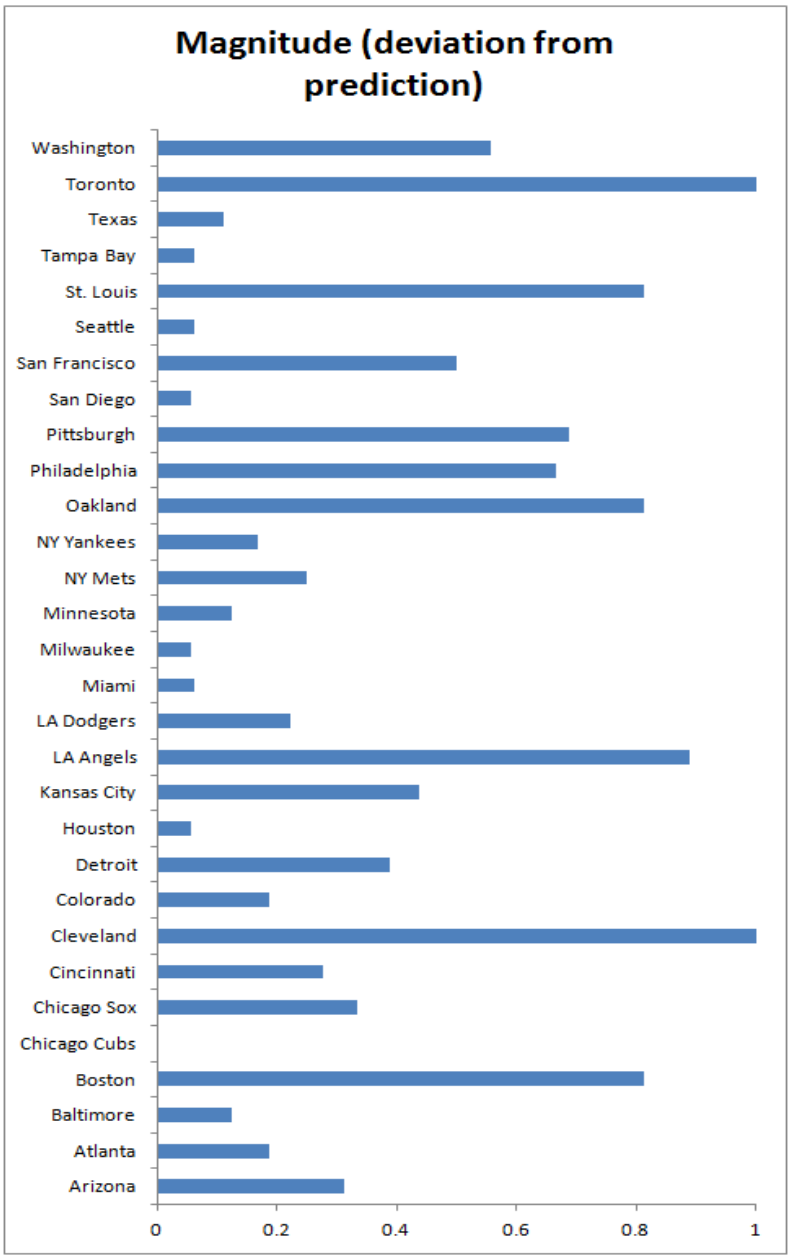
[3] Sohmer, S. PredictWise: aggregating the wisdom of crowds. Hypervocal, October 11 (2011).

[4] The Linemakers Odds to win 2014 Super Bowl. Sporting News, February 4 (2013).

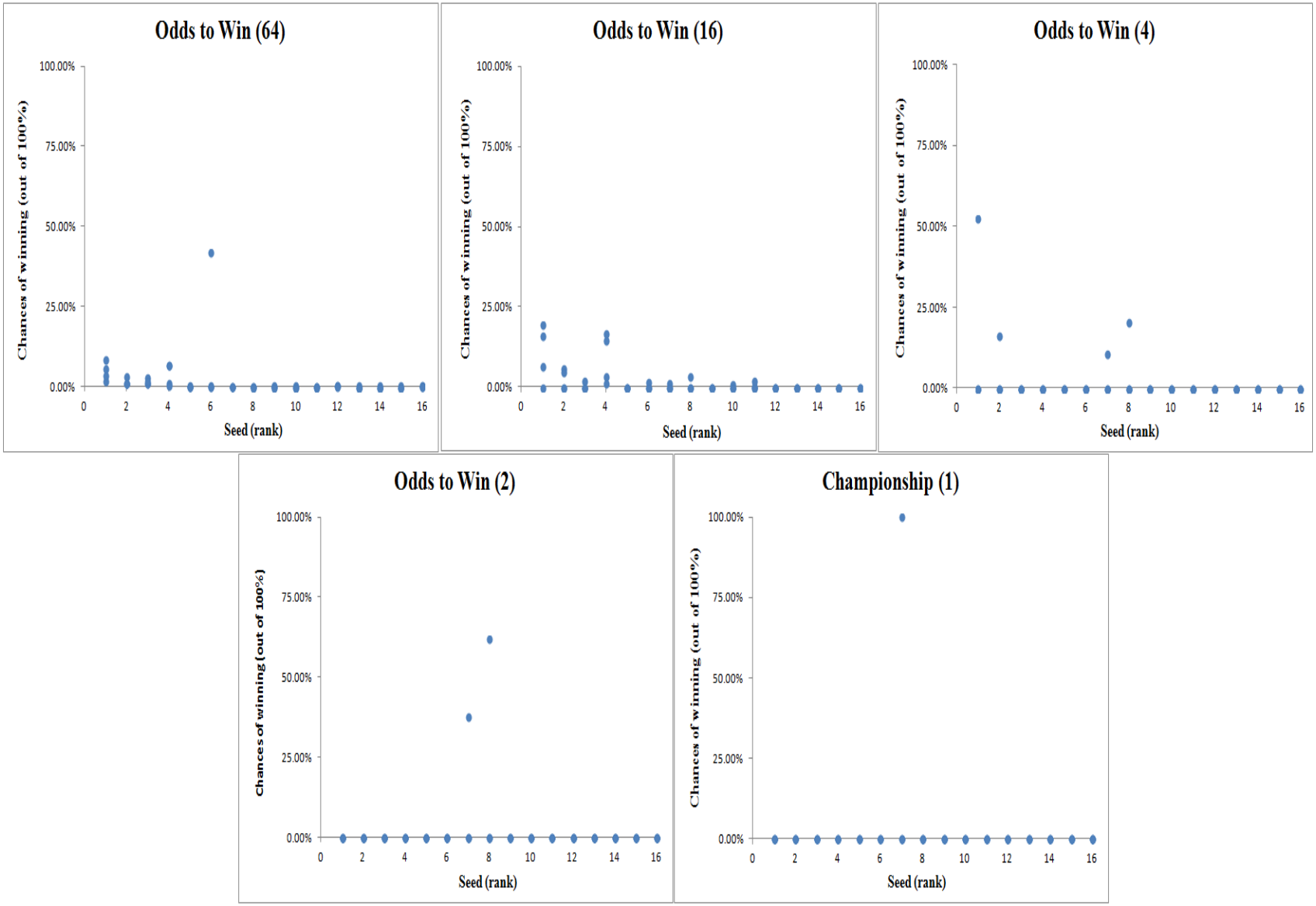
[5] Ulfelder, J. Using Wiki surveys to forecast rare events. Dart-throwing Chimp blog, August 11 (2013).

[6] Ezekowitz, J. Quantifying Intangibles: a new way to predict the NCAA tournament. The Harvard Sports Analysis Collective, May 18 (2011).

Supplemental Figures:



Supplemental Figure 1. The magnitude of how far away the prediction is from actual performance. See Methods, Section 2 for details.



Supplemental Figure 2. Odds to win by NCAA tournament interval. See Methods, Section 3 for details.