5-10-2023

Big Ideas in Sports Analytics and Statistical Tools For Their Investigation

Benjamin S. Baumer
Smith College, bbaumer@smith.edu

Gregory J. Matthews
Loyola University Chicago, gmatthews1@luc.edu

Quang Nguyen
Carnegie Mellon University, nmquang@cmu.edu

Recommended Citation

Baumer, Benjamin S.; Matthews, Gregory J.; and Nguyen, Quang. Big Ideas in Sports Analytics and Statistical Tools For Their Investigation. Wiley Interdisciplinary Reviews: Computational Statistics, , : 1-24, 2023. Retrieved from Loyola eCommons, Mathematics and Statistics: Faculty Publications and Other Works, http://dx.doi.org/10.1002/wics.1612

Big Ideas in Sports Analytics and Statistical Tools for their Investigation

A Preprint

Benjamin S. Baumer
Statistical & Data Sciences
Smith College
Northampton, MA 01063
bbaumer@smith.edu

Gregory J. Matthews
Mathematics and Statistics
Loyola University Chicago
Chicago, IL 60660
gmatthews1@luc.edu

Quang Nguyen
Statistics & Data Science
Carnegie Mellon University
Pittsburgh, PA 15213
nmquang@cmu.edu

January 11, 2023

Abstract

Sports analytics—broadly defined as the pursuit of improvement in athletic performance through the analysis of data—has expanded its footprint both in the professional sports industry and in academia over the past 30 years. In this paper, we connect four big ideas that are common across multiple sports: the expected value of a game state, win probability, measures of team strength, and the use of sports betting market data. For each, we explore both the shared similarities and individual idiosyncracies of analytical approaches in each sport. While our focus is on the concepts underlying each type of analysis, any implementation necessarily involves statistical methodologies, computational tools, and data sources. Where appropriate, we outline how data, models, tools, and knowledge of the sport combine to generate actionable insights. We also describe opportunities to share analytical work, but omit an in-depth discussion of individual player evaluation as beyond our scope. This paper should serve as a useful overview for anyone becoming interested in the study of sports analytics.

***K*eywords** sports analytics · R packages · sports data · pairwise comparisons · datasets

1 Introduction

Insights derived from the analysis of data have transformed the world of sports over the last few decades. While baseball—a naturally discrete sport with more than a century's worth of professional data—may be the sport with the longest relationship with sports analytics, one would be hard-pressed to identify a professional sport today in which sports analytics is not having an impact. In basketball, analytics has driven a shift in the conventional wisdom about shot selection. Most teams are shooting more three-pointers, settling for fewer long two-point shots, deploying more versatile defenders, and relying less on the strategy of pounding the ball into the paint in an attempt to get a high-percentage shot (Schuhmann, 2021). In American football, teams are going for it on fourth down far more often than in the past, a direct result of statistical analysis showing that most teams were previously overly conservative (Lopez, 2020). And, of course, in baseball, teams are using defensive shifts to maximize the probability of recording an out, encouraging hitters to improve their launch angles, and optimizing pitcher repertoires to minimize contact (Healey, 2017).

These are just the most obvious examples of strategic changes that are fueled by insights extracted from data by practitioners of sports analytics. Similar insights are now being made in less obvious settings, including esports (Clark et al., 2020; Maymin, 2021). These insights come both from academia, where researchers typically use public data to produce high-caliber, peer-reviewed scientific work, as well as from industry, where highly-trained analysts work with with players, coaches, and team officials to put new ideas into immediate

effect thanks to high-resolution, often proprietary data. A growing pool of people move seamlessly between these two worlds, leading to the formation of partnerships and the cross-pollination of ideas.

Every sport is different, with its own set of rules, strategies, methods of data collection, number of players, and the magnitude of the role of chance. At the same time, many sports are similar, either because one evolved from the other, or the structure of the games share certain attributes. Sports that are closely related historically may or may not share common applications of analytical methods. For example, despite belonging to the same bat-and-ball family, baseball and cricket differ in strategies such as batting order or sacrifice plays. Conversely, with just a few small tweaks, analytical metrics might work just as well across sports that are unrelated and quite different. For instance, an Elo rating could be equally valid for chess players and ice hockey teams.

In this paper, we explore four key ideas that have widespread applicability across many sports: the expected value of a game state (Section 2), win probability (Section 3), measures of team strength (Section 4), and the use of sports betting market data (Section 5). In each case, we define the concept mathematically, explain how it originated, and give examples of its applications in multiple sports. Our goal is to unify the conceptual threads, while doing some justice to the customizations necessary to make a metric meaningful in a particular sport. We include copious references to original works of scholarship.

Doing the work of sports analytics requires computing with data. While the sources of sports data are too numerous to list, in Section 6 we highlight a few computational tools (including a table of R packages) that make this kind of work possible. Section 7 lists several opportunities for disseminating work publicly. We conclude in Section 8 with a short discussion of some ideas that are not explored in this paper. Notably, we omit a treatment of individual player ratings for team sports, since this concept has been covered ably in these pages by Albert (2015), and its inclusion would double the length of this manuscript. We do, however, discuss individual player ratings in the context of one-person teams (e.g., chess, tennis) in Section 4.

We encourage readers to explore Cochran et al. (2017) and Albert et al. (2016) for collections of articles in sports analytics that provide broad coverage of the field.

2 The expected value of a game state

In many sports, the first step towards an analytical understanding is the estimation of the expected value of a game state at any given point in it. Mathematically, we define X to be a random variable indicating the number of points (or runs) that a team will score over some determined amount of time (e.g., remainder of game, quarter, period, or inning). Let $s \in S$ be a tuple that encodes the *state* of a game. Then our task is to estimate:

$$\mathbb{E}[X|s] = \sum_{x \geq 0} \Pr[X = x|s] \cdot x \,, \tag{1}$$

for any state $s \in S$, where $\Pr[X = x|s]$ is the probability of scoring x points given that the game is in state s and S is the set of all possible states.

The concept of a state is easier to grasp in a sport that can be modeled as *discrete* (in the sense of discrete event simulation). By discrete, we mean a sport that can be easily broken into short, distinct segments of action which are typically summarized categorically. Each of these segments might represent a state s. For example, each pitch in baseball is either a ball or a strike. If the ball is put in play, then there may be a complex sequence of movements by the players, but ultimately (within a few seconds) that sequence will end and no more action will be permitted until the next pitch. At the beginning and end of each phase of action, we will know definitively which team is on offense and defense, which runners are on which bases, the score, how many outs there are, etc. Tennis could similarly be viewed as a series of discrete actions defined by each point. To say that a player is winning 6-2, 3-1, 40-15 and serving with one fault committed is to characterize the state of the match. In American football, the game can be broken down into a discrete sequence based on each down. Contrast this to sports like lacrosse, soccer, or any variant of hockey, which feature largely running clocks and continuous player movement. In these sports, it is not obvious how to break up the action into discrete chunks.

In this Section, we illustrate how the fundamental concept of the expected value of a game state leads to compelling findings in a variety of sports.

Table 1: George Lindsey's expected run matrix. Note how (when reading across the rows) the expected runs decrease as outs increase for the same configuration of baserunners, while (when reading down the columns) expected runs generally increase as baserunners advance. 000 means no runners on base, and 110 means runners on second and third bases.

	Out		
Base	0	1	2
000	0.461	0.243	0.102
001	0.813	0.498	0.219
010	1.194	0.671	0.297
011	1.390	0.980	0.355
100	1.471	0.939	0.403
101	1.940	1.115	0.532
110	1.960	1.560	0.687
111	2.220	1.642	0.823

2.1 Discrete event analysis

First, we explore results derived from the expected value of a state in sports where discrete event analysis is common. We draw primarily on baseball and American football, but applications in other sports (e.g., tennis) are common (see for example, Kovalchik & Reid (2019)).

2.1.1 In baseball, the expected run matrix

In baseball, s is typically determined by two factors: the configuration of the runners on base (there are 8 possibilities) and the number of outs (3 possibilities). Thus, there are $|S| = 24 = 8 \cdot 3$ basic states of an inning in baseball[1], and we are often interested in the number of runs that will be scored from some state until the end of the inning. In this example using baseball, $\mathbb{E}[X|s]$ is the expected number of runs scored between now and the end of the inning given that the inning is currently in state s. The collection of estimates $\mathbb{E}[X|s]$ for all 24 states is called the *expected run matrix* [2], and it is foundational in baseball analytics.

Early work on this topic can be found in Lindsey (1963), who used play-by-play data to compute an empirical estimate for the mean number of runs scored in the remainder of the inning for each of these 24 possible states of an inning. This line of work led to analysis of all types of common baseball strategies. For example, many baseball teams elect to attempt a sacrifice bunt with a runner on first and no one out in the inning, with the goal of moving the runner to second base, at the cost of the batter being out. Figure 1 shows a reproduction of Lindsey (1963)'s original calculations, and Table 1 shows the expected run matrix in its most common form.

Tango et al. (2007) (and many subsequent analyses) conclude that the sacrifice bunt is rarely worth it, because most teams would be expected to score more runs with a runner on first and no outs than they would with a runner on second and one out.

It is worth emphasizing that the values in $\mathbb{E}[X|s]$ are estimates, and the precision of those estimates has many subtleties.

First, the values within the expected run matrix change over time. For example, any estimation of the values in the expected run matrix based on data from a high-scoring era (e.g., the early 2000s) will yield different values than equivalent analysis in a low-scoring era. In a high run-scoring environment, where there are many home runs, the value of a walk may be higher, since a player who walks is more likely to score on a subsequent home run. Conversely, in a low run-scoring environment where hits are hard to come by, stolen bases and sacrifice bunts may be comparatively more valuable. Thus, a careful estimate of $\mathbb{E}[X|s]$ would include a time parameter t, indicating when the estimate is appropriate.

[1] 25, if you include the absorbing state of 3 outs that describes the end of an inning.

[2] There is no inherent dimensionality to $\mathbb{E}[X|s]$. The *matrix* nomenclature stems from its values typically being displayed in 8×3 grid. However, when computing with $\mathbb{E}[X|s]$, it is most often convenient to treat it as a 24×1 vector.

TABLE I

DISTRIBUTION OF SCORES IN REMAINDER OF HALF-INNING

Data	B	T	$N(T,B)$	$P(0\|T,B)$	$P(1\|T,B)$	$P(2\|T,B)$	$P(>2\|T,B)$	$E(T,B)$	$\sigma/\sqrt{N}$
59/60	0	0	6561	.747	.136	.068	.049	.461	.012
	0	1	4664	.855	.085	.039	.021	.243	.011
	0	2	3710	.933	.042	.018	.007	.102	.008
59/60	1	0	1728	.604	.166	.127	.103	.813	.031
	1	1	2063	.734	.124	.092	.050	.498	.022
	1	2	2119	.886	.045	.048	.021	.219	.016
59/60	2	0	294	.381	.344	.129	.146	1.194	.083
	2	1	657	.610	.224	.104	.062	.671	.043
	2	2	779	.788	.158	.038	.016	.297	.024
59/60	3	0	67	.12	.64	.11	.13	1.39	.09
	3	1	202	.307	.529	.104	.060	.980	.072
	3	2	327	.738	.208	.030	.024	.355	.040
59/60	12	0	367	.395	.220	.131	.254	1.471	.087
	12	1	700	.571	.163	.119	.147	.939	.051
	12	2	896	.791	.100	.061	.048	.403	.032
59/60	13	0	119	.13	.41	.18	.28	1.94	.15
	13	1	305	.367	.400	.105	.128	1.115	.077
	13	2	419	.717	.167	.045	.071	.532	.054
59/60	23	0	73	.18	.25	.26	.31	1.96	.18
	23	1	176	.27	.24	.28	.21	1.56	.10
	23	2	211	.668	.095	.170	.067	.687	.080
59/60	F	0	92	.18	.26	.21	.35	2.22	.20
	F	1	215	.303	.242	.172	.283	1.642	.105
	F	2	283	.671	.092	.102	.135	.823	.085
		$\sum N = 27027$							
52/60	F	0	173	.17	.27	.17	.39	2.254	.145
	F	1	419	.310	.242	.186	.262	1.632	.080
	F	2	527	.645	.114	.110	.131	.861	.06

Figure 1: Table 1 from Lindsey's original paper. The column labeled $E(T, B)$ gives the expected run matrix as a vector, based on Lindsey's analysis of Major League Baseball data from 1959 and 1960.

TABLE I

THE EXPECTED POINT VALUES OF POSSESSION OF THE FOOTBALL WITH FIRST DOWN AND TEN YARDS TO GO FOR VARIOUS TEN-YARD STRIPS

Center of the ten-yard strip (yards from the target goal line): X	Expected point value: $E(X)$
95	−1.245
85	−0.637
75	+0.236
65	0.923
55	1.538
45	2.392
35	3.167
25	3.681
15	4.572
5	6.041

Figure 2: Table 1 from Carter and Machol's original paper. Note the monotonic increase in expected point values as the team gets closer to the endzone.

Second, the characterization of S as having 24 states is only the simplest possible. The inning, or the score of the game, or even the weather, could be incorporated into S, as those conditions might reasonably affect the estimate of $\mathbb{E}[X|s]$. More definitively, the identity of the current batter, pitcher, or batter on deck, might also affect the estimate of $\mathbb{E}[X|s]$. Indeed, Tango et al. (2007) show that when a particularly weak-hitting batter is up (i.e., the pitcher), a sacrifice bunt becomes a more effective strategy.

See Albert & Bennett (2001) for a fuller discussion of the use of the expected run matrix in baseball and Marchi et al. (2018) for examples of how to estimate the expected run matrix using Retrosheet data and the R statistical computing language (R Core Team, 2022).

2.1.2 In American football, expected points

The concept of estimating the value of the state of a game is easily extended to other sports. For example, in American football, s is determined by situational variables such as down, yardage to the next first down, time remaining in the game, and field position.

The task of estimating expected points of possession in football goes back to Carter & Machol (1971), who estimate the expected points for 1st and 10 plays in the NFL, given any yard line on the football field. Due to limitations regarding the amount of data collected, the authors divide football field into 10-yard buckets, centered at their midpoints (e.g. 5, 15, 25, 35, etc.), before averaging the value of the next scoring instance across the field to obtain the expected points. Figure 2 shows a reproduction of Carter & Machol (1971)'s estimates. As expected, the estimated expected points increases monotonically as the teams gets closer to the endzone. One limitation of this approach is the linearity assumption, which results in a high negative expected point value when the offensive team is 95 yards away from the opponent's goal line.

Early work on expected point values in American football can also be found in Carroll et al. (1988). In particular, the authors consider a similar approach to Carter & Machol (1971) and propose a linear model for expected points in the NFL. They determine that every extra 25 yards is associated with 2 more points scored on average for a football team.

Other attempts at modeling expected points in football are Goldner (2012) and Goldner (2017), who propose a Markov framework. In particular, the author considers a football drive as an *absorbing Markov chain*, consisting of distinct *absorbing states* that include touchdowns, field goals, and other possession outcomes. An absorbing state is a link in a Markov chain from which there are no possible transitions (i.e., it is the end of the chain). For any given play, the expected points are calculated using the absorption probabilities for different scoring events.

A more in-depth overview of the history of expected points in sports is provided in Yurko et al. (2019) (Section 1.1). Most importantly, Yurko et al. (2019) use publicly available data provided by the **nflscrapR** package (Horowitz et al., 2020) to model the expected points on a play-by-play level in football. The authors introduce a multinomial logistic regression approach, which takes into account the current down, time remaining, yards from endzone, yards to go, and indicators for goal down situation and whether there are less than two minutes remaining in the half. Their model estimates the probabilities of the following possible scoring outcomes after each play: no score, safety, field goal, and touchdown for both the offensive and defensive teams, all of which have a point value. The expected points for a play can then be calculated accordingly, by summing up the products of the scoring event point values and their associated probabilities (see Equation 1).

In addition, Pelechrinis et al. (2019) develop an expected points framework in the same spirit as the previous work, but account for the strength of the opponents in their method. They state that by failing to account for opponent strength appropriately, about 124.8 points per team each season (or about 3.8 wins per season) are not credited correctly. This is a substantial amount in a 16-game season.

2.1.3 In American football, 4th down strategy

The concept of expected points in American football has many applications. One of the most notable and well-studied topics is the evaluation of 4th down strategy. There is near universal consensus in the literature that NFL teams have been too conservative in the past when making 4th down decisions.

Romer (2006) examines 4th down decisions in the NFL using expected points by focusing only on examples from the first quarter of a game (to avoid issues with end-of-half and end-of-game decision making). They concluded that teams don't go for it enough if teams are trying to maximize their probability of winning the game.

Numerous other papers (see Lopez (2020) for details) use the analysis of the expected number of points to improve fourth down strategy. In addition to Romer (2006), later work by Yam & Lopez (2019) uses win probability (see Section 3), rather than expected points, and a causal inference framework to reach similar conclusions that NFL teams are too conservative in going for it on 4th down. In addition, they estimate that a better strategy would be worth about 0.4 wins per season on average, a substantial amount comparable to the effect size reported by Pelechrinis et al. (2019) above.

Lopez (2020) presents an introduction to NFL tracking data, and examines 4th down behavior as an example of the type of problem that can be more thoroughly studied with the increase in granularity of the tracking data over traditional NFL data. In the past, when looking at down and distance data to study whether NFL coaches are making good decisions about whether to "go for it" or punt on 4th down, the distance data is only a rounded approximation of the true distance "to go" (i.e. 1 yard, 1 foot, and 1 inch will all be recorded as 4th and 1. In fact, anything up to 2 yard will recorded as 4th and 1 (Lopez, 2022). However, a coach on the field during a game will be able to clearly see the difference between 1 inch and 1 yard, and this information will factor into their decision making. With tracking data, the "to go" distance can be much more accurately assessed and therefore evaluation of 4th down coaching decisions can now account for this "extra" information that is available to a coach on the field of play, but not recorded in traditional NFL data. Many past analyses of the decision to go for it or not on 4th down conclude that coaches in the NFL are too conservative in their decision making. Lopez (2020) also concludes that coaches are too conservative on 4th down decision making, but notes further that past estimates of the magnitude of how conservative coaches are on 4th down may be overstated due to the way in which to go yardage was recorded only approximately in the past.

2.1.4 Other applications of expected points in American football

Researchers have also applied the notion of expected points to investigate other aspects of the game of football, including quarterback performance and coaching decisions.

For quarterback evaluation, White & Berry (2002) present a tiered logistic regression method that can be, in general, applied to any regression setting with a polychotomous response. Using this technique, they estimate

the value of NFL plays using a simple expected points model with down, yards to go, and yards to goal as predictors. Accordingly, the model results are utilized to obtain ratings and rankings for NFL passers.

Alamar (2010) implements an expected points framework to examine play calling in the NFL. However, rather than assessing each play on its own, they evaluate the play in the context of the drive. Based on play-by-play data from 2005 through 2008, they determine that teams are under-utilizing passing plays in some situations.

Another application of expected points is to evaluate kickoff decisions made by football coaches, as demonstrated by Urschel & Zhuang (2011). Specifically, they look at surprise on-sides kicks versus regular kickoffs and the decision to accept a touchback versus returning the kickoff. Using data from the 2009 NFL season, they conclude, as many have, that coaches in the NFL tend to make conservative decisions.

2.2 Continuous event analysis

Even in sports where the concept of a state is more difficult to define, the value of a possession can be estimated with the help of tracking data. Over the past decade or so, professional sports leagues have collected tracking data which record the locations of all players and the ball (or puck) throughout a game. This high-resolution data allows researchers to produce advanced analyses of the captured spatiotemporal information and better understand the game. This is a great leap forward from older resources such as traditional box-score results and play-by-play data.

2.2.1 In basketball, expected point value

In basketball, Cervone et al. (2014) and Cervone et al. (2016) introduce expected possession value (EPV) as a means toward an assessment of a player's on-court performance. This metric is a continuous-time estimate of the expected number of points for the offensive team on a given possession using player and ball locations. The EPV takes into account all possible outcomes (a shot attempt, a pass, etc.) for a given player with the ball, with different weights being assigned to each decision. The computation of the EPV statistic is done using a (technically discrete) Markov model conditioned on spatial locations. Consequently, the authors derive a metric called EPV-Added (EPVA), measuring a player's EPV contribution in a given situation relative to a league-average player.

A demonstration of the EPV model presented in Cervone et al. (2016) is available at `https://github.com/dcervone/EPVDemo`. Figure 3 illustrates how the provided tracking data informs the evolution of EPV throughout the play. It displays a snapshot of a possession during the NBA regular season matchup between the Miami Heat and the Brooklyn Nets on November 1, 2013. Miami is the team on offense in this possession, whose outcome is a 26-foot three-point miss by Mario Chalmers. The plot consists of two elements: 1) (bottom) the player locations on the court at a particular moment in this possession: when the ball just left Chalmers's hands, and 2) (top) a line graph showing how the EPV changes continuously throughout the play until the three-point attempt. For this possession, the estimated EPV for the Miami Heat reaches its peak at 1.276 points at the moment the shot is taken.

Note that Miami starts the play with an EPV of approximately 1.0 points, which indicates their implied average points per possession. Chalmers' shot is worth 3 points, so the EPV of 1.276 points implies that the model estimate of the probability of Chalmers making this shot is 42.5%. A breakthrough in this work is that this estimate is conditional on the locations of the other 9 players on the basketball court.

Another framework for estimating expected points in basketball is proposed by Sicilia et al. (2019). The authors offer a different point of view on expected points, where they first consider a classification model which returns the probabilities for whether a player would commit a foul (shooting and non-shooting), turnover, or attempt a shot. The values associated with each of those "terminal actions" are then used to compute the expected points within a basketball play.

See also Bornn et al. (2017) for more information on how tracking data have enabled advanced statistical analyses of basketball in recent years. The strategy of maximizing expected points in basketball has led directly to the proliferation of three-point shooting in the NBA.

2.2.2 In American football, yards gained

In Section 2.1.2, we discussed advances in American football analytics based on discrete game states defined by down, yards to first down, field position, etc. The advent of player tracking data makes it possible to analyze American football using continuous states. For example, Yurko et al. (2020) use tracking data provided by the 2019 NFL Big Data Bowl (see Section 7) to model the expected yards gained for a ball-carrier during the

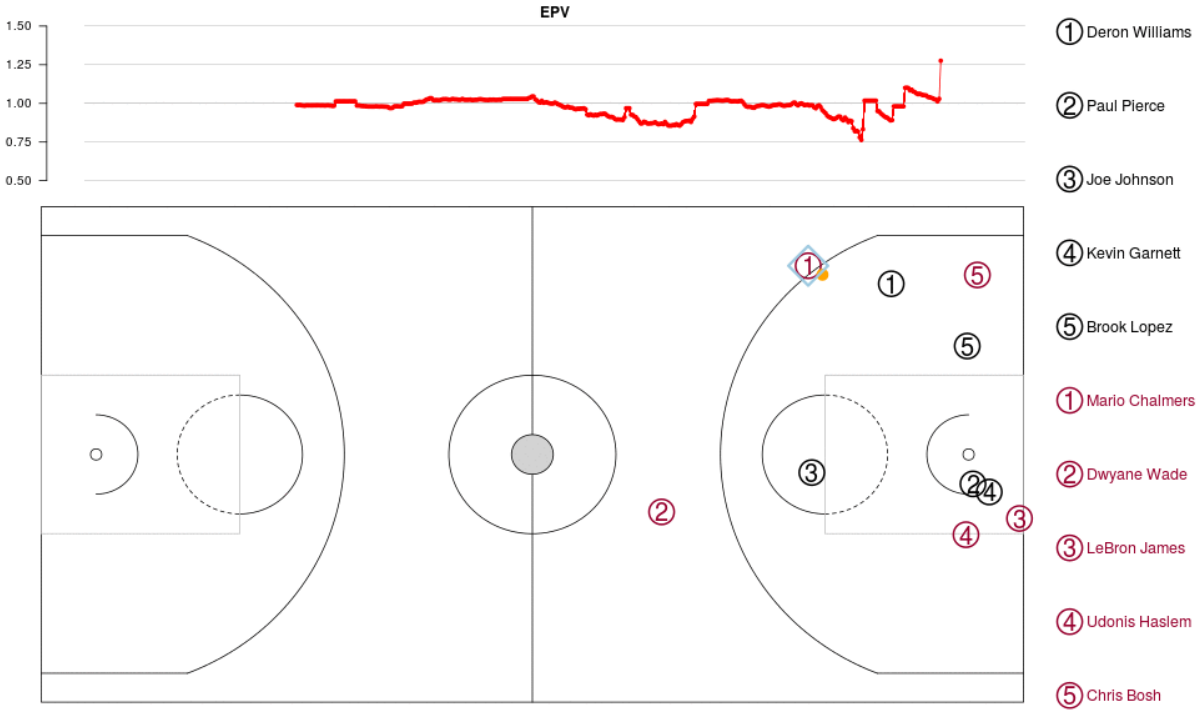

Figure 3: Player locations and estimated EPV for a possession during the Miami Heat (red) vs. Brooklyn Nets (black) NBA game on November 1, 2013. The captured moment is when Miami's Mario Chalmers just releases a three-point shot, which ends up as a missed field goal. Figure created by Cervone, et al. `https://github.com/dcervone/EPVDemo`

course of a play. As an extension to pre-existing approaches, the authors use conditional density estimation to obtain a probability distribution for the number of yards gained during the play, rather than only producing a single estimate for the expected yards gained. Accordingly, the probability of various types of outcomes at the end of a play such as a touchdown or a first-down gain can be computed from the distribution of the end-of-play yard line.

Expected point value is also the main component of a novel NFL quarterback evaluation metric introduced by Reyers & Swartz (2021). The authors take advantage of player tracking data to account for different passing and running options on the football field that are available to the quarterback. The expected points and probabilities associated with the possible quarterback options are estimated using the method of ensemble learning via stacking.

2.2.3 In other sports

The notion of expected possession value has also been extended to association football (soccer). Fernandez et al. (2021) implement deep learning methods to examine the instantaneous expected value of soccer possessions. This approach considers passes, ball drives, and shots in soccer as the main set of actions used to compute the EPV metric. Many applications can be derived from this framework, including predicting which footballer on the pitch is most likely to receive the next pass from the current on-ball player.

Macdonald (2012) uses expected goals to evaluate ice hockey players, but does not have access to player tracking data necessary to evaluate possessions. Kumagai et al. (2021) offer an EPV metric for ice hockey via a Bayesian space-time framework.

2.3 Optimal strategies that don't maximize expected points

Earlier in Section 2, we defined the expected value of a possession based on the state s of the game in terms of the expected number of points (runs) X that would be scored in the remainder of some period of time. We then showed how this value could be used to analyze the relative effectiveness of certain strategies, with

the simple idea that strategies that yield higher expected values are preferable. Generally, the goal of any sport is to score more points than the other team, which most often means trying to score as many points as possible, leading to a general strategy of maximizing expected points. However, there are situations in which maximizing the number of expected points is *not* the desired strategy.

For example, in the bottom of the ninth inning of a tied baseball game, the optimal strategy for winning the game is maximizing the probability of scoring *at least one run*, which may differ from the strategy of maximizing expected runs. If we let U be the set of all strategies, then we assert that it is not always the case that the strategy $u \in U$ that maximizes the expected number of points will maximize the probability of winning:

$$\arg\max_{u \in U} \Pr[X > 0 | s, u] \neq \arg\max_{u \in U} \mathbb{E}[X | s, u] \,.$$

Consider the situation where runners are on first and third base, and the score is tied in the bottom of the ninth inning with no one out. Information derived from Table 1 reveals that the expected number of runs scored in the remainder of the inning is 1.94 runs, while the probability of scoring zero runs is 0.13. The defense is in a tight spot, facing an 87% probability of losing the game. However, by walking the hitter to load the bases, they create the opportunity to force the lead runner at home and thus reduce the chances of scoring to 82%, even though they raise the expected number of runs scored to 2.22. In this case, the *defensive* team is wise to pursue the strategy of maximizing the expected number of runs scored, because it *minimizes* the probability of scoring at least one run.

Maximizing the probability of scoring is optimal in any sudden-death situation, which has (but currently does not) included overtime in American football (Martin et al., 2018).

The situation gets even more interesting when teams modify both their offensive and defensive strategies simultaneously. For example, in hockey teams will often pull their goalie when trailing in the final period. This strategy severely weakens their defense but strengthens their offense. The hope is to score a quick goal to get back in the game, but the risk is falling further behind. Beaudoin & Swartz (2010) show that NHL coaches do not always employ the optimal strategies, usually by waiting too long to pull their goalies. Skinner (2011) develops a general framework for these desperation strategies, which include the onside kick in American football, pulling the infield and/or outfield in baseball, and of course, the fabled Hack-a-Shaq strategy in basketball.

3 Win probability

A related, but different concept to expected points is the notion of *win probability*. Win probability is simply an estimate of the probability that a team will win the game, given its current state s. Extending the mathematical framework we defined in Section 2, let W_i be a binary random variable that indicates a win for team i. Then,

$$\Pr[W_i | s] \,,$$

is the win probability for team i in the state s.

This win probability is closely related to the expected value of a state. Albert (2015) defines the win probability as:

$$\Pr[W_i | s] = \sum_{X \geq 0} \Pr[X | s] \cdot \Pr[W_i | X, s] \,,$$

where $\Pr[W_i | X, s]$ is the probability that team i will win the game given that they score X points from state s.

Win probability is easily extended to provide a measure of the impact of sports plays and individual player contributions, as discussed in Albert (2015). Given its popularity, recent books on sports analytics often dedicate multiple chapters entirely to win probability. These include Albert & Bennett (2001), Schwarz (2004), Tango et al. (2007), Albert et al. (2016), and Winston et al. (2022).

In this section, we discuss notable previous work on win probability in baseball, American football, basketball, and several other sports.

3.1 Baseball

The notion of win probability in baseball goes back to at least as early as Lindsey (1961), who calculates the expected win probability after each inning based on the distribution of runs scored in each inning. Inspired by

Lindsey (1961)'s work, Mills & Mills (1970) utilize win probability to introduce Player Win Average (PWA), a measure of a player's contribution to the game outcome. In particular, PWA is computed as

$$PWA = \frac{Win\ Points}{Win\ Points + Loss\ Points},$$

where the win and loss points represent how much the player positively and negatively impacts their team's probability of winning after each play. In effect, the win points are the sum of the changes in $\Pr[W_i|s]$ from one state to the next.

Additionally, a mathematical model for estimating win probability in baseball is presented in Tango et al. (2007). The authors use Markov chains to look at win expectancy throughout the course of a baseball game. This approach considers different states of the game such as base, inning, outs and score, and outputs win probabilities accordingly.

See Albert (2015) for a more detailed historical overview of the use of win probability in baseball.

3.2 American football

In recent years, a number of statistical methods have been used to build well-calibrated win probability models in American football. These are flexible algorithms that have high predictability, can account for nonlinear interactions between the explanatory variables, require few assumptions, and produce feature importance scores.

Lock & Nettleton (2014) implement a random forest framework to provide a win probability estimate before each play in a football game. Covariates included in this tree-based method are the current down, score differential, time remaining, adjusted score, point spread, number of timeouts remaining for each team, total points scored, current yard line, and yards to go for a first down. According to this model, the difference in score between the two teams is the most important feature for predicting win probabilities at any given moment in an NFL game.

In addition, Yurko et al. (2019) estimate win probability in the NFL using a generalized additive model (GAM), as part of the **nflscrapR** package (Horowitz et al., 2020) and nflWAR framework. This model takes into account the estimated expected points obtained from the model described in Section 2, along with other predictors for time, current half, and timeouts. The two win probability frameworks proposed by Lock & Nettleton (2014) and Yurko et al. (2019) were also implemented in Yam & Lopez (2019) with minimal modifications. Specifically, the authors combined both approaches to estimate the win probability for each play, with an overall goal of assessing fourth down decision-making in American football.

A vital highlight of Yurko et al. (2019)'s win probability model is that it is fully reproducible and uses publicly available data. One of Yurko et al. (2019)'s goals was also to encourage researchers to "use, explore, and improve upon our work," which ultimately inspired **nflfastR** (Carl & Baldwin (2022)), now considered the successor to **nflscrapR**.

Figure 4 shows a win probability graph for the 2018 NFL Playoffs Divisional Round matchup between the New Orleans Saints and the Minnesota Vikings on January 14, 2018. We obtain the estimated probability of winning for each team using the **nflfastR** R package, which implements a gradient boosting model via the **xgboost** library (Chen et al. (2022)) for estimating win probabilities. Minnesota was leading throughout the first three quarters of the game, having win probabilities of 0.869, 0.941, and 0.742 at the end of the first, second, and third quarters, respectively. The win probabilities get close to parity late in the fourth quarter, when the Saints took the lead with 3:01 left in the game. The last play of this game—famously known as the Minneapolis Miracle—resulted in a drastic swing in win probabilities for both teams. With 10 seconds remaining in the game, the Vikings begin the final possession with a 25.3% chance of winning. Their probability increased to a perfect 1 when Stefon Diggs scored a game-winning 61-yard receiving touchdown as the game clock expired.

3.3 Basketball

Stern (1994) provides an investigation of in-game win probability and the scoring process in basketball using a Brownian motion model. Let $p(l, t)$ represent the win probability for the home team given an l-point lead after t seconds of game time. The model introduced by Stern (1994) is a probit regression model, which

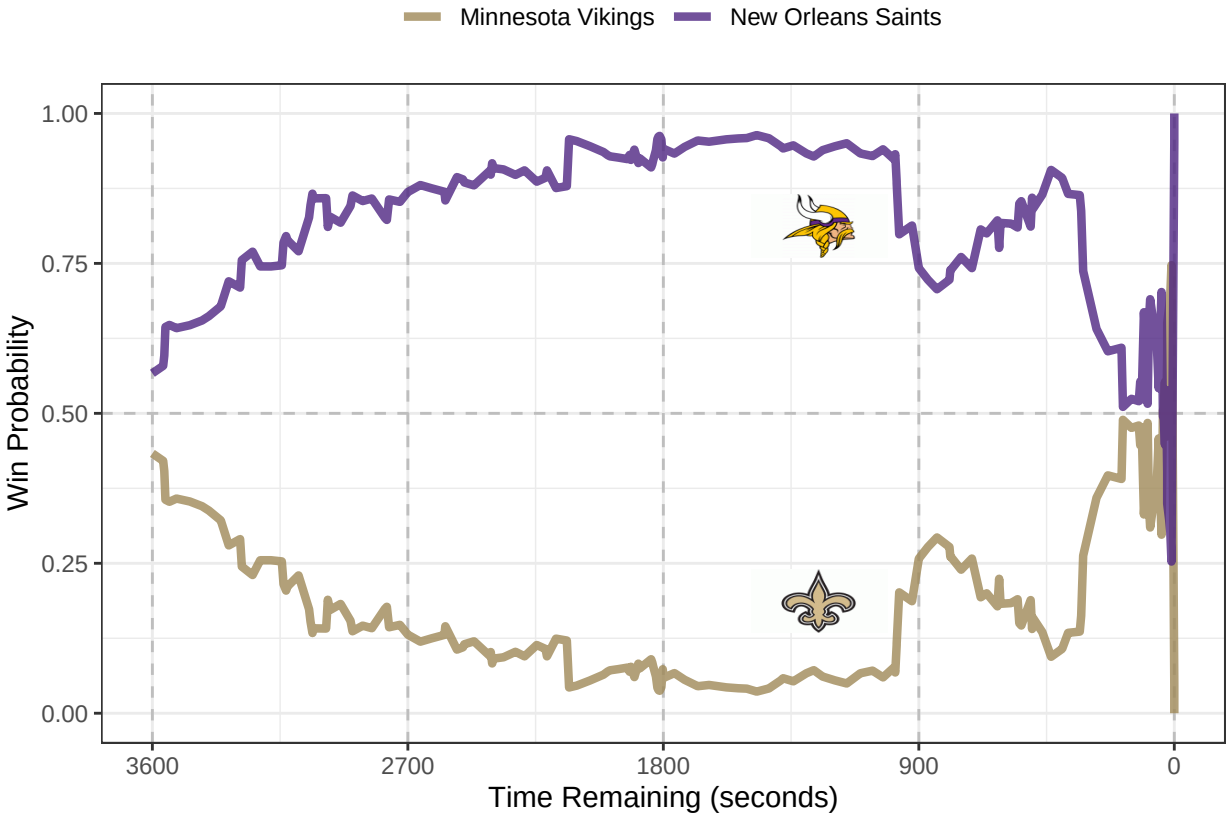

Figure 4: Win probability graph for New Orleans Saints vs. Minnesota Vikings in the 2017–18 NFL Playoffs.

provides an estimate for $p(l,t)$. Specifically,

$$p(l,t) = \Phi\left(\frac{l+(1-t)\mu}{\sqrt{(1-t)\sigma^2}}\right).$$

Here, a Brownian motion process with drift μ points advantage for the home team and variance σ^2 is used to model the score difference between the home and away teams.

On a related note, Deshpande & Jensen (2016) extend Stern (1994)'s framework by applying it in a Bayesian setting. Deshpande & Jensen (2016) propose a Bayesian linear regression model to assess the impact of individual players on their team's chance of winning at any given time of a basketball game. This model assumes independence of observations and constant variability in win probability.

Moreover, McFarlane (2019) uses logistic regression to estimate win probability for evaluating end-of-game decisions in the NBA. The approach takes into account the remaining game time, score difference, and point spread. This win probability model is then applied to the calculation of the End-of-game Tactics Metric (ETM), measuring how the chance of winning a basketball game differs between the optimal and on-court actual decisions.

3.4 Other sports

The idea of win probability is also applied to other sports, with a diverse range of statistical techniques being used to estimate the probability of winning for a player or team. Brenzel et al. (2019) use three-dimensional Markov models to estimate win probability throughout a curling match. In particular, the authors propose both homogeneous and heterogeneous Markov models for estimating the chance of winning in curling, with different independence assumptions on the relationship between performance and the current state of the game. In esports, Maymin (2021) relies on logistic regression to build a well-calibrated in-game win probability model for each specific moment during a game of League of Legends. Moreover, Guan et al. (2022) develop an in-game win probability model for the National Rugby League using functional data analysis. In this

approach, the rugby play-by-play event data are treated as functional, and the win probability is expressed as a function of the match time.

4 Team strength

A third crucial idea in sports analytics is the estimation of team strength. First, we briefly introduce a simple method for estimating team strength based on win-loss record. Next, we detail three other more sophisticated methods for estimating team strength in sports through pairwise evaluations. The methods in this Section apply equally well to multiplayer teams and single-player teams.

The impetus for all methods for estimating team strength is the realization that win-loss records are a noisy measure of team strength. As binary outcomes, and with all sports (except perhaps chess) involving some element of chance, wins and losses carry some signal of team strength, but we can do better.

4.1 Expected winning percentage

A simple method for estimating team strength that has become popular in sports analytics is expected winning percentage—often called Pythagorean expectation—developed by James (2003). Later, Miller (2007) derived the formula as a consequence of assuming that runs (in baseball) are generated by two independent Weibull processes.

Expected winning percentage is just:

$$\widehat{wpct} = \frac{X^\alpha}{X^\alpha + Y^\alpha} \,,$$

where X is the number of points (runs) that a team has scored, and Y is the number of points (runs) that they have allowed, over some specified time period. James's work was originally in baseball, and he posited the value of $\alpha = 2$. The resemblance to the formula for computing the length of the hypotenuse in a right triangle provides the nod to Pythagoras.

Subsequent analysts have tried to find the optimal value of α for various time periods. This can be done with a few lines of code, after observing that

$$\frac{X^\alpha}{X^\alpha + Y^\alpha} = \frac{1}{1 + (Y/X)^\alpha}$$

and fitting a non-linear model (see similar discussion in Baumer et al. (2021)). Figure 5 illustrates the quality of the fit in Major League Baseball since 1954, where the optimal value of α is 1.84.

Many authors have fit expected winning percentage models to other sports—too many to cite here. See, for example, Hamilton (2011) for association football (soccer), Caro et al. (2013) for Division I college football, and notably, future NBA general manager Daryl Morey for basketball (Dewan & Zminda, 1993).

4.2 Bradley-Terry models

Perhaps the most widely-used probability model for predicting the outcome of a paired comparison is the Bradley-Terry model (BTM) (Bradley & Terry, 1952). For a pair of players (or teams) i and j, let Π_{ij} denote the probability that i is preferred to j. Then the BTM is a logistic regression model with parameters β_i, β_j such that

$$\log\left(\frac{\Pi_{ij}}{\Pi_{ji}}\right) = \beta_i - \beta_j \,.$$

Here, $\exp(\beta_i)$ is often viewed as a representation of team i's ability.

The BTM can be implemented in R via the **BradleyTerry2** package (Turner & Firth, 2020). As an example, we consider the data given in Agresti (2018) (page 247) on tennis results from 2014–2018 for five men's professional players: Novak Djokovic, Roger Federer, Andy Murray, Rafael Nadal, and Stan Wawrinka. We fit a BTM to estimate the win probability for each pair of players and obtain a ranking for this group of five.

Table 2 shows the estimated coefficients of the fitted BTM. According to the abilities, between 2014 and 2018 the players are ranked as follows: 1) Djokovic, 2) Federer, 3) Wawrinka, 4) Nadal, 5) Murray. In addition to an ordering, the magnitude of the coefficients in Table 2 provide a measure of relative strength.

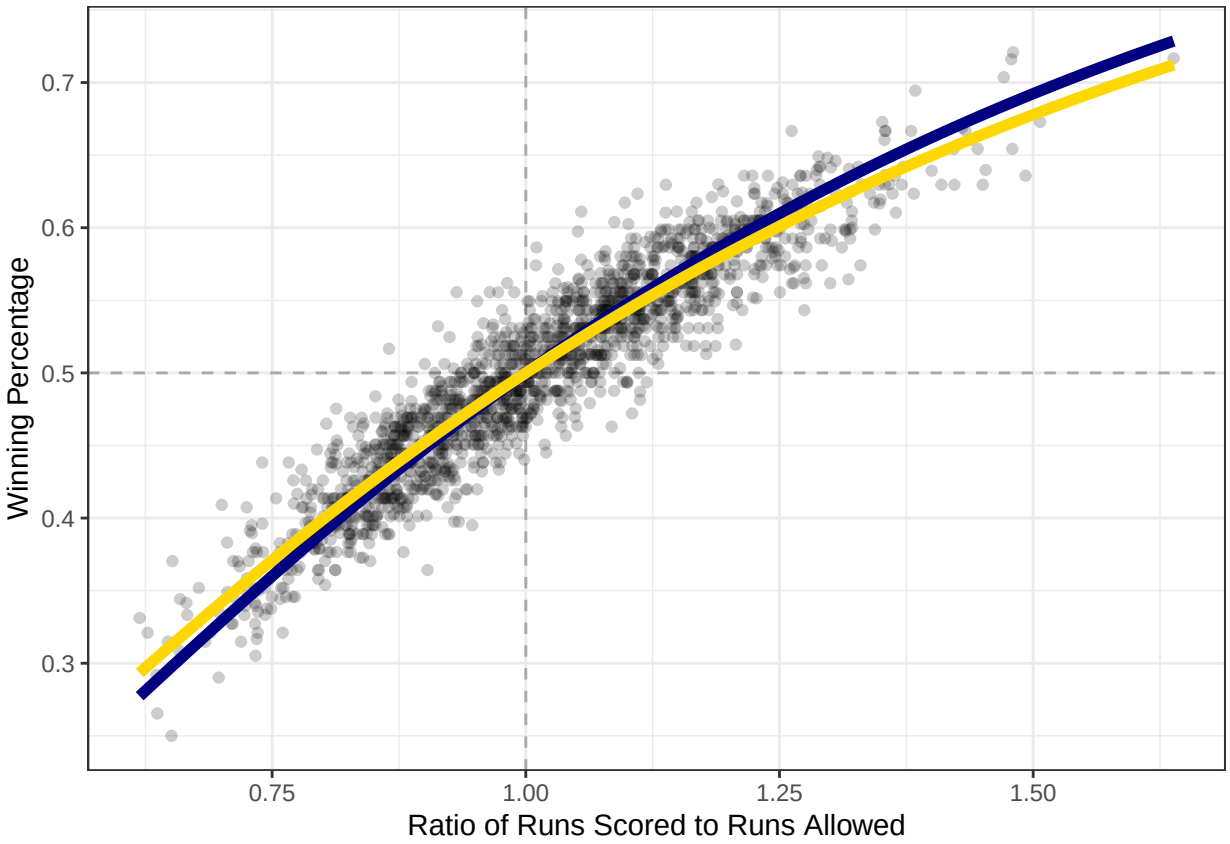

Figure 5: Winning percentages vs. runs scored and runs allowed in baseball, 1954–2021. The navy line represents the expected winning percentage model posited by James, with the exponent value of 2. The gold line shows the same model with an optimal exponent of 1.84.

Table 2: The estimated abilities (with standard errors) for each tennis player, relative to Wawrinka, obtained from the fitted Bradley-Terry model.

Player	Ability	SE
Djokovic	1.176	0.500
Federer	1.136	0.511
Wawrinka	0.000	0.000
Nadal	-0.062	0.515
Murray	-0.569	0.568

To obtain win probabilities, as an illustration, for the Federer-Nadal matchup, an estimate for the probability of a Federer victory is:

$$\hat{\Pi}_{24} = \frac{\exp(\hat{\beta}_2 - \hat{\beta}_4)}{1 + \exp(\hat{\beta}_2 - \hat{\beta}_4)} = \frac{\exp(1.136 + 0.062)}{1 + \exp(1.136 + 0.062)} = 0.768\,.$$

4.3 Elo ratings

Another widely known tool for measuring team strength is the Elo rating system (Elo, 1978), which was originally developed for chess. Given two players i and j with unknown ratings R_i and R_j, the probability Π_{ij} of i beating j is defined as

$$\Pi_{ij} = \frac{1}{1 + K^{(R_j - R_i)/400}}\,.$$

In this formula, K is commonly known as the K-factor, or development coefficient. The International Chess Federation (FIDE) uses $K = 10$ for players with any previously achieved rating of at least 2400. Finally, $K = 40$ is given to new players with under 30 games played, and players under the age of 18 with rating less than than 2300 (FIDE, 2022).

Another interpretation for Π_{ij} is the expected score of the game for player i. The scores of 0, 0.5, and 1 are associated with three possible game outcomes loss, tie, and win, respectively. After a game, the updated Elo rating R_i^* for player i is

$$R_i^* = R_i + K(S_i - \Pi_{ij}) \,,$$

where $S_i \in \{0, 0.5, 1\}$. When a tournament concludes, a post-tournament rating is obtained for each player based on the rating updates for all games played.

To illustrate, we consider a chess game played on June 1, 2022 on Chess.com by one of the authors, with data obtained from the **chessR** package (Zivkovic, 2022) (see Section 6.1). Prior to the game, the author was rated 1732, whereas his opponent was rated 1683. Since both ratings are below 2400, we apply a development coefficient of $K = 20$ to this example. The probability of the author (a) defeating their opponent (b) was

$$\Pi_{ab} = \frac{1}{1 + 20^{(1683-1732)/400}} = 0.591 \,.$$

The author won the match: that outcome is associated with a score of $S_a = 1$. The post-game Elo rating for the author is thus

$$R_a^* = 1732 + 20(1 - 0.591) = 1740 \,.$$

Besides chess, the Elo system has also been implemented to estimate team strength in other sports. See Koning (2017) for more information on applications of the Elo rating in soccer, and Kovalchik & Reid (2019) and Kovalchik (2016) for Elo ratings in tennis. Furthermore, Elo ratings are used extensively for rankings of teams in numerous sports by the data journalists at FiveThirtyEight.com.

4.4 Bayesian state-space models

Glickman & Stern (1998) propose a Bayesian state-space model for paired comparisons for predicting NFL games, allowing team strength parameters to vary over time. In particular, they model point differential in the NFL by introducing week-to-week and season-to-season as the two primary sources of variation in team strengths. See also Glickman & Stern (2017) for more discussion on estimating team strengths in American football.

More recently, Lopez et al. (2018) extend Glickman & Stern (1998)'s state-space model to understand randomness in the four major American sports leagues. Betting moneylines are used in place of point differentials in order to estimate team strengths, and this framework also accounts for home advantage. Both papers motivate the usefulness of model-based measures of team strength by demonstrating their superiority to low-resolution win-loss records. Apart from sports gambling, having an accurate estimate of team strength is useful to team officials, who are constantly monitoring and forecasting their team's ability.

In a similar Bayesian setting, Koopman & Lit (2015) study English Premier League soccer match results by assuming a bivariate Poisson distribution with time-varying team abilities. This state-space approach appears to improve on bookmaker's odds.

5 Sports betting market data

Most of the research in sports analytics is fueled by the analysis of data recorded from the outcome of sports contests. However, a growing body of literature is informed by data from sports betting markets. Since the 2018 United States Supreme Court decision in *Murphy v. National Collegiate Athletic Association*, sports gambling has exploded in the U.S. The increasing interest in sports gambling has led to increasing interest in sports gambling *data*, and that data has proven useful to researchers in at least two major ways.

First, betting market data is probably the best source for estimating the true probability of a team winning a game. The efficiency of betting market data in this respect has been demonstrated time and time again. The utility of these estimates have then informed research that has helped us learn about the sports themselves. In this sense, data generated by sports gambling has been an important source of data useful for sports analytics (see Section 5.2).

Table 3: 2023 NBA Championship odds for the top 6 and bottom 6 teams. Retrieved from FanDuel Sportsbook on January 9, 2023.

Rank	Team	Line	Odds	Prob.	Prob. Normalized
1	Boston Celtics	390	4.9	0.204	0.163
2	Milwaukee Bucks	500	6.0	0.167	0.133
3	Brooklyn Nets	800	9.0	0.111	0.089
4	Golden State Warriors	900	10.0	0.100	0.080
5	Los Angeles Clippers	1000	11.0	0.091	0.073
6	Denver Nuggets	1100	12.0	0.083	0.067
25	Oklahoma City Thunder	50000	501.0	0.002	0.002
26	Orlando Magic	50000	501.0	0.002	0.002
27	Charlotte Hornets	50000	501.0	0.002	0.002
28	Houston Rockets	50000	501.0	0.002	0.002
29	San Antonio Spurs	50000	501.0	0.002	0.002
30	Detroit Pistons	50000	501.0	0.002	0.002
Total	-	496590	4995.9	1.253	1.000

Second, sports analytics researchers have studied various types of sports gambling outlets (including fantasy sports). This research has estimated probabilities, evaluated common strategies, and offered optimized approaches for a variety of different games of chance (see Section 5.3). Some researchers have then tried to demonstrate a positive return on some of these betting strategies, with very limited success.

5.1 Example: Win probabilities from betting market data

To see how betting market data can be used to estimate team strengths, consider the betting lines posted on FanDuel Sportsbook for the 2023 NBA Champion on January 9, 2023 and shown in Table 3. This is a futures market, because the actual NBA champion will not be determined until June 2023. The Boston Celtics are the favorite to win, with a moneyline of +390, meaning that a $100 bet on the Celtics to win the championship will pay back the original bet and an additional $390 if the Celtics win it all. This style of odds are sometimes called *American odds*. The corresponding fractional odds have the Celtics at 4.9:1 to win the championship. Conversely, six teams share the lowest odds at +50000.

These moneylines (ℓ_i) can be converted into an implied probability (p_i) using the formula:

$$p_i = \frac{100}{100 + \ell_i}\,.$$

The sum of those probabilities is greater than one—this is why the sportsbook makes money regardless of who wins the championship. However, the implied probabilities can be normalized by dividing by their sum to recover true probabilities of each team winning the championship. Many different researchers have shown that these normalized implied probabilities are accurate, unbiased, and efficient estimates of the true unknowable probabilities (see Lopez et al. (2018) for discussion and an extensive list of references).

In this case, the FanDuel futures market suggests that the Celtics have a 16.3% chance of winning the championship, while the Milwaukee Bucks have the second best chance, at 13.3%. These implied probabilities can be used to fit various models for team strength, as described in Section 4.

5.2 The use of betting market data for sports analytics

While Lopez et al. (2018) use betting market data to model team strengths, they do not directly address strategies for betting or inefficiencies in betting markets. Early work by Gandar et al. (1988) examine the rationality of NFL betting markets and concludes that statistical tests fail to reject the hypothesis of rationality. Related work such as Lacey (1990) and Boulier et al. (2006) explores the efficiency of NFL betting markets in the mid-1980s and late-1990s, respectively. Neither paper finds strong evidence for inefficiencies in the markets. Boulier & Stekler (2003) compare the predictive performance of power rankings and media experts to the betting market for NFL games and found that the betting market is the best for predicting winners. Lopez & Matthews (2015) show that betting market data was most useful in predicting men's college basketball tournament outcomes.

Sports betting market data has also been used to investigate competitive behavior within leagues. Soebbing & Humphreys (2013) find evidence that sports bettors *think* tanking in the NBA is occurring, although the evidence for whether it actually is remains mixed.

5.3 Analytics for sports betting

Many different types of bets can be placed on sports. For individual contests, bets involve point spreads, moneylines (see Section 5.1 for an example), odds, or other ways of handicapping which team will win. Money can also be wagered on futures, where odds are given in advance for events that may or may not transpire (e.g., a certain team making the playoffs, or a certain player winning the MVP award). Here, we focus on betting pools, in which a group of people compete to predict winners in multiple contests (often a tournament). We also address the inevitable question of whether strategies exist that will consistently beat the market.

5.3.1 Betting pools

One popular type of betting pool is a survivor pool, in which participants stay in the competition as long as they continue to successfully pick winners. Bergman & Imbrogno (2017) present formal optimization approaches for NFL survivor pools and conclude that planning for only part of the season yields optimal results in terms of maximizing survival probability. Imbrogno & Bergman (2022) estimate the probability of having to share the winning pot in NFL survivor pools.

Perhaps the most commonly-studied sports betting market surrounds the NCAA men's college basketball tournament. Breiter & Carlin (1997) use Monte Carlo methods to study the standard "office pool." Kaplan & Garstka (2001) consider a variety of NCAA college basketball pools, and find that the simple strategy of picking the team with the better seed is generally, but not always, optimal. Metrick (1996) finds that bettors overback the heaviest favorites. Niemi et al. (2008) show an improved return on investment by picking an undervalued champion and then completing the rest of one's bracket by using published odds. Clair & Letscher (2007) develop and test strategies for maximizing expected return in both survivor and tournament-style pools.

5.3.2 Beating the market

Naturally, after studying the efficiency of sports betting markets, researchers try to find inefficiencies that can be exploited for financial gain. Not surprisingly (given the efficiency of these markets), such gains are hard to come by.

Sauer (1998) finds that while racetrack betting markets are generally efficient, information asymmetry plays a role in creating inefficient markets. Nichols (2014) concludes that the impact of travel is not completely incorporated into the betting markets, but that any effect is too small to find any profitable advantage. Paul & Weinbach (2014) investigate the less-saturated betting market for the WNBA and fail to find strategies for positive return on investment. Spann & Skiera (2009) show no way to beat the market in the German premier soccer league, given the high fees associated with placing bets.

More successfully, Buttrey (2016) explores the NHL betting market and produces a model to predict win probabilities in given games, then tests the model by placing market price bets in games where the predicted probability differs from the market. They find that their methods were able to produce a positive return on investment.

6 Tools

Analytical work in sports requires facility with an ever-changing set of computational tools for working with data. Sources of authoritative data about sports are myriad, and are too numerous to list here. Software tools for sports analytics are similarly numerous. For R, we maintain a CRAN Task View for Sports Analytics that catalogs R packages published on the Comprehensive R Archive Network (CRAN) and organizes them by sport (Baumer et al., 2022). Table 4 provides an overview of the currently available sport-specific CRAN packages. Recently, Casals et al. (2022) offer a systematic review of sport-related packages on CRAN. Further, a more general collection of software tools is being curated by the SportsDataverse initiative (Gilani, 2022).

In the remainder of this section, we highlight a few tools for sports analytics that are of general interest and illustrate a common paradigm for how these tools can be used in conjunction.

Table 4: A summary of sport-specific packages available on the Comprehensive R Archive Network (CRAN) as of October 16, 2022. While the major North American sports dominate the list, perhaps the fastest-growing collection is for esports.

Sport	Number of Packages	List of Packages
American Football	12	nflverse, nflfastR, nflreadr, nfl4th, nflseedR, nflplotR, NFLSimulatoR, fflr, ffscrapr, ffsimulator, gsisdecoder, cfbfastR
Association Football (Soccer)	9	worldfootballR, engsoccerdata, socceR, ggsoccer, footballpenaltiesBL, footBayes, itscalledsoccer, FPLdata, EUfootball
Basketball	8	BAwiR, AdvancedBasketballStats, uncmbb, BasketballAnalyzeR, NBAloveR, wehoop, hoopR, toRvik
Baseball/Softball	7	Lahman, retrosheet, pitchRx, mlbstats, baseballDBR, baseballr, runexp
Chess	5	chess, stockfish, bigchess, rchess, chessR
Esports	5	CSGo, rbedrock, ROpenData, opendotaR, RDota2
Hockey	5	hockeyR, NHLData, nhlapi, nhlscrape, fastRhockey
Cricket	4	yorkr, cricketr, cricketdata, howzatR
GPS Activity Tracking	3	trackeR, trackeRapp, rStrava
Track and Field	2	combinedevents, JumpeR
Australian Rules Football	1	fitzRoy
Swimming	1	SwimmeR
Volleyball	1	volleystat

6.1 Case study in how tools fit together: chess

Many tools in sports analytics provide the ability to read, write, and plot data stored in a sport-specific format. For example, consider chess, where the sequence of moves in games is often recorded in Portable Game Notation (PGN). Software tools can then be built around this well-defined format. The **chess** package (Lente, 2020) provides R users with the ability to read, write, display, and manipulate chess data in PGN format.

Application programming interfaces (APIs) are also a common source for data retrieval. In chess, the **chessR** package (Zivkovic, 2022) allows R users to download game data from the Chess.com API. This type of infrastructure, where one package is the "workhorse" that facilitates common generic data operations, and other packages layer on specific functionality, is common in sports analytics. See Section 4.3 for an example of how the **chessR** package can be used to compute Elo ratings.

Figure 6 shows a rendering of the starting chess board obtained via the **chess** package, along with the final position in the game won by one of the authors mentioned earlier in Section 4.3 (with data downloaded via the **chessR** package). We note how the contextual information provided by the chessboard is instrumental in helping the reader understand the data (How many of us can visualize PGN directly?). In Section 6.2, we outline a collection of graphical tools that provide similar context for different playing surfaces.

6.2 Graphical tools

Creating effective data graphics is a key component of statistical communication, and sports is no exception. We highlight a few packages that assist with the creation of data graphics about sports.

Each professional sports team has its own brand, most obviously identified by a team logo and set of colors. The **teamcolors** R package (Baumer & Matthews, 2020) provides color palettes and logos for men's and women's professional and collegiate sports teams, as well as color and fill scale functions compatible with **ggplot2** (Wickham et al., 2022). For example, the NFL teams' colors and logos shown in Figure 4 were provided by the **teamcolors** package. Figure 7 illustrates how the use of team colors, which have a natural association for many sports fans, can help to untangle what would otherwise be messy data graphics. In Figure

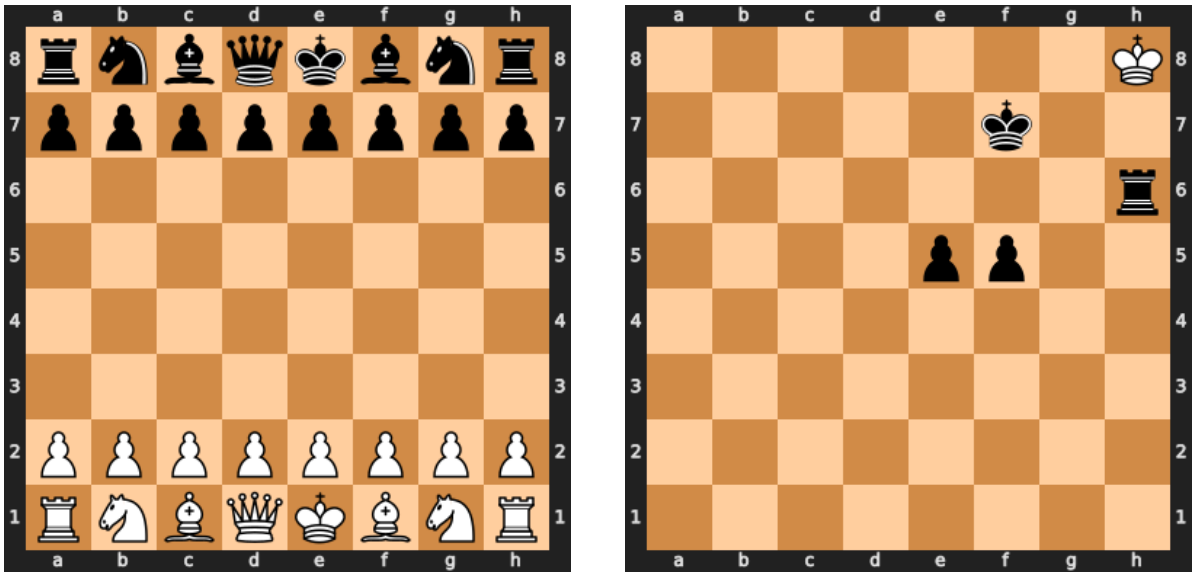

Figure 6: At left, the starting chess board printed via the **chess** package. At right, the final position for one of the authors' recent wins (a checkmate playing Black).

7, 30 different lines are plotted on top of one another, crisscrossing and intersecting in various unpredictable ways. However, the use of team colors to identify the lines makes it possible to follow the trajectory of most teams over the course of the season.

nflplotR (Carl, 2022) has a similar goal to **teamcolors**. It also provides **ggplot2** extensions but is designed specifically for the NFL. A great feature of **nflplotR** is the collection of `geom_*()` (geometric object) functions that enhance high-quality plotting of NFL team logos and player images with **ggplot2**. Figure 8 shows a scatterplot of offensive and defensive expected points added for NFL teams in the 2021 regular season. The logos of all 32 American football clubs are plotted in place of the usual dots, making it easier for the reader to identify which team each data point represents.

Player tracking data contains coordinates that reveal player movement, and these coordinates are always understood in context relative to reference points on the field, court, ice, board, or pitch for a particular sport. Orienting these points graphically may require drawing a complex set of guidelines that provide that context to readers. Thankfully, the **sportyR** package (Drucker, 2022) contains generic playing surfaces for baseball, basketball, curling, American football, ice hockey, soccer, and tennis that can be added to **ggplot** graphics with a single function call. The surfaces plotted in Figure 9 are helpful in contextualizing player tracking data (such as those shown in Figure 3) and would be laborious for each analyst to have to create on their own. With the increased availability of player tracking data, this particular tool should see increased usage.

6.3 Case study in the evolution of tools and research: baseball

As the granularity of baseball data has evolved over time, so too have the statistical methodologies for modeling that data, and the tools for working with it.

For example, before George Lindsey's work (see Section 2), most of the baseball data that was publicly available was seasonal: it showed only season totals for each player. These data, now available through the **Lahman** package (Friendly et al., 2022), were sufficient to study broad trends in baseball, and led to insights such as the value of expected winning percentage (see Section 4.1) and the importance of on-base percentage relative to batting average. These relatively simple insights fueled the "Moneyball" (Lewis, 2004) era revolution in sports analytics (B. Baumer & Zimbalist, 2014).

Over time, the resolution of baseball data has improved to include play-by-play data, pitch-by-pitch data, and now player tracking data.

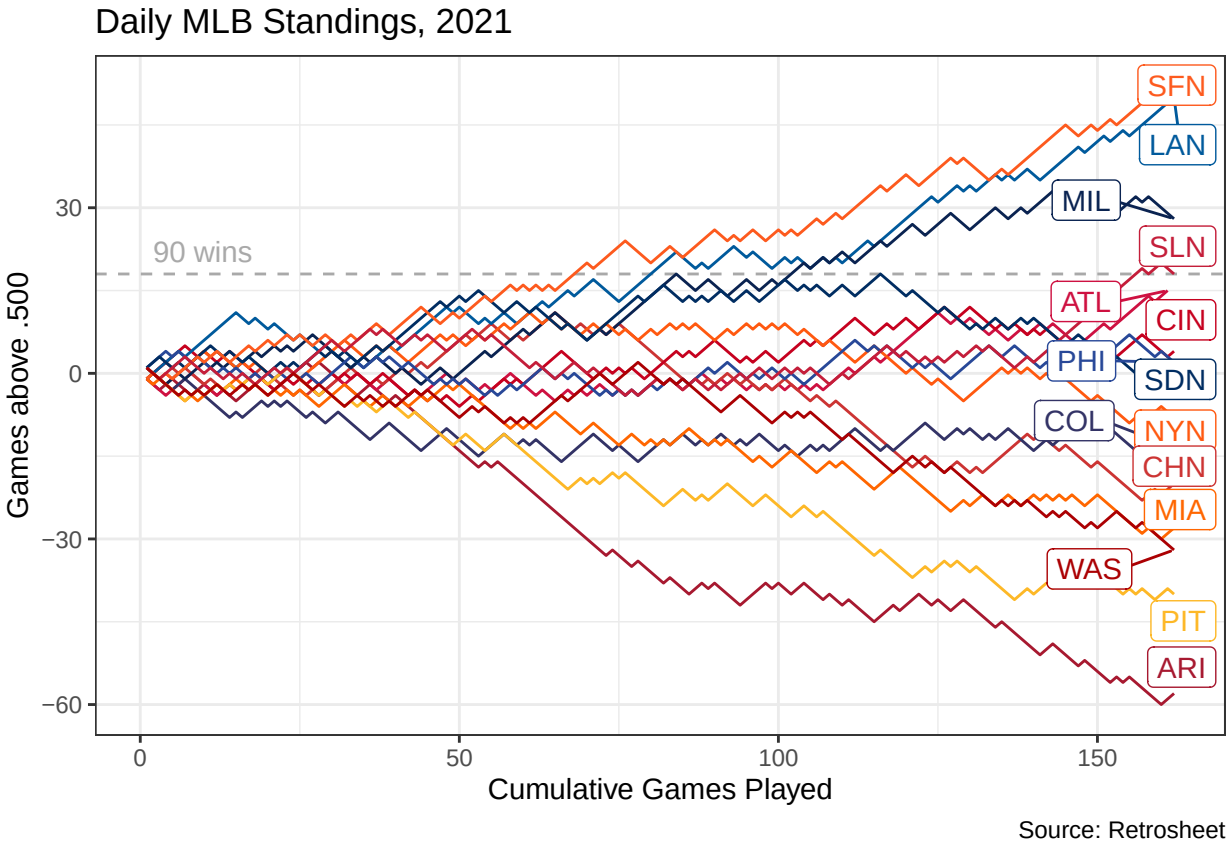

Figure 7: The progression of National Leauge team standings during the 2021 Major League Baseball season. Note how the use of team colors makes it possible to untangle what would otherwise be a messy jumble of indistinguishable lines. Data provided by **retrosheet** and colors provided by **teamcolors**.

The **retrosheet** package (Douglas & Scriven, 2021) now provides access to the historical play-by-play data available from Retrosheet (this is a comprehensive version of what Lindsay collected for his research). This play-by-play data allowed researchers to learn about strategies, like those that we discussed in Section 2. In baseball, this deepened our understanding of bunting, stolen bases, handedness, batting order, and many other aspects of the game. Play-by-play data underlies much of the analysis in Tango et al. (2007).

The **pitchRx** package (Sievert, 2015) provides access to pitch-by-pitch data that fueled innovative research into catcher framing (Deshpande & Wyner, 2017), pitch values (Healey, 2019), and pitch classification (Sidle & Tran, 2018). Catcher framing is a notable example of a concept that scouts talked about for decades, but that could not be quantified by analysts until data of the appropriate resolution became available.

While play-by-play data allows us to make valuations *between plays*, player tracking data allows us to make valuations *within plays*. The **baseballr** package (Petti & Gilani, 2022) now provides access to player tracking data from Statcast. These data have led to investigations into how defensive shifts affect batting performance (Bouzarth et al., 2021), as well as how launch angles affect the probability of hitting a home run (Marchi et al., 2018).

As we saw above with chess, the packages in baseball fit together in creative ways. In Figure 7, we showed how **teamcolors** can illuminate data pulled from **retrosheet** to make an informative data graphic. One could just as easily use **sportyR** to generate a field graphic, and then overlay player tracking data obtained from **baseballr** to depict defensive shifts.

Thus, these R packages enable research by making data more easily available. Moreover, because R is scriptable, they make it easier to share research that is reproducible. Recent conferences, such as the Carnegie Mellon Sports Analytics Conference, have included a reproducible research competition to foster these efforts (see Section 7).

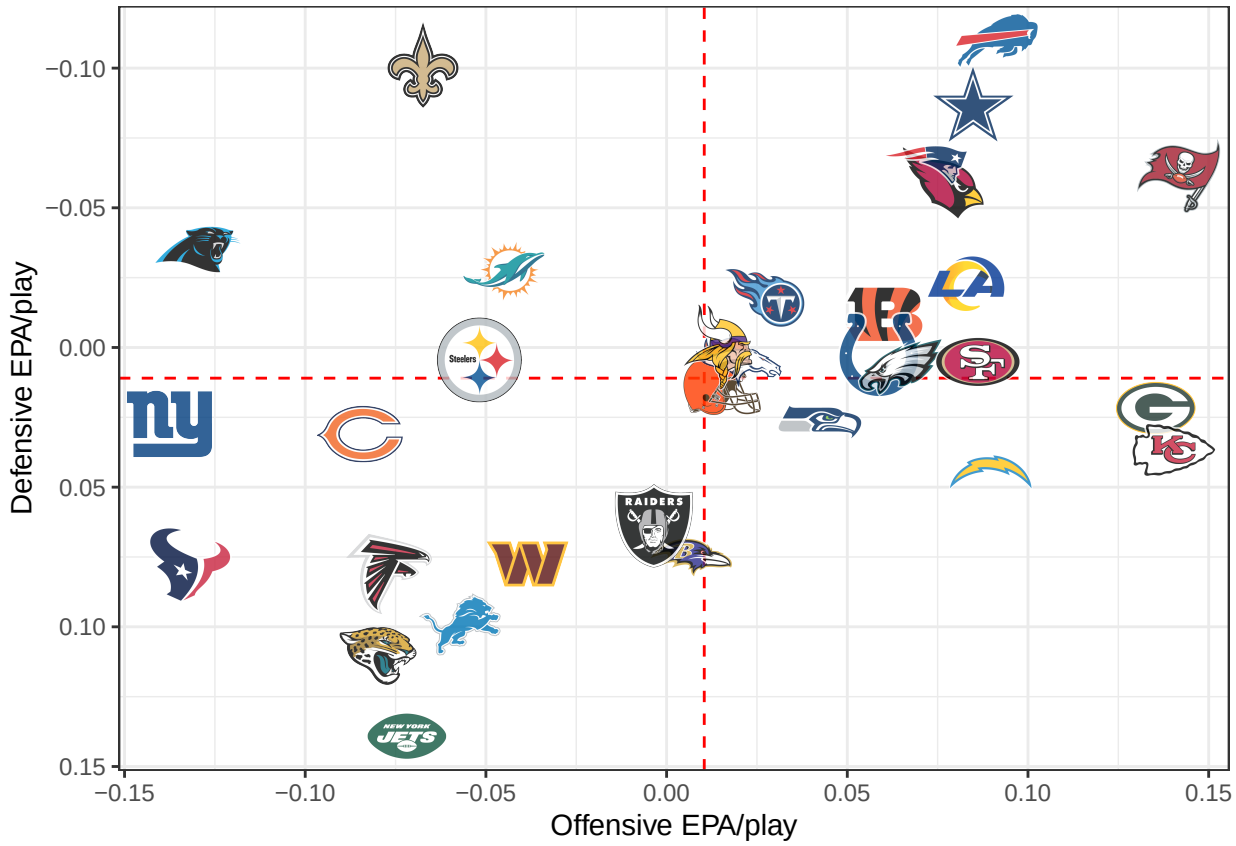

Figure 8: Offensive and defensive expected points added per play for the 2021 NFL regular season, plotted with **nflplotR** using data from **nflfastR**.

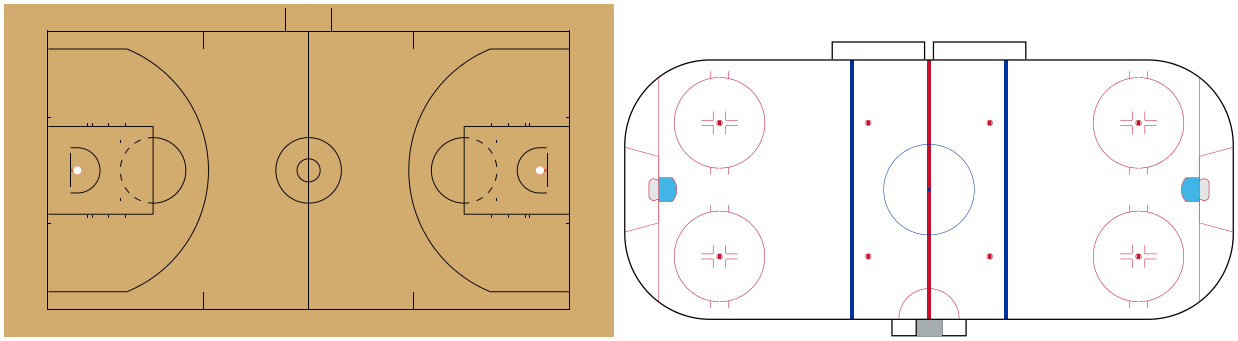

Figure 9: At left, an NBA basketball court drawn by **sportyR**. At right, an NHL hockey rink drawn by **sportyR**.

7 Opportunities

Public research in sports analytics is driven in part by several notable competitions and conferences. These venues have been an important source of new ideas and have contributed to the diversification of the field by breaking down barriers to entry.

In 2014, Kaggle launched its first March Machine Learning Mania competition for predicting the outcome of the NCAA men's basketball tournament. 243 entrants competed for the $15,000 cash prize by submitting predicted probabilities for every possible pairwise matchup among the 68 college basketball teams in the tournament (Glickman & Sonas, 2015). Subsequently, the *Journal of Quantitative Analysis in Sports* (JQAS) released a special issue on prediction methodology for the NCAA men's basketball tournament. Among the published papers, we learned that the winning entry was based on a fairly simple logistic regression model trained on betting market data (Lopez & Matthews, 2015). Thus, the competition not only sparked interest in sports analytics, but also resulted in peer-reviewed research which, in that case, demonstrated the value of quality data over sophisticated modeling.

Perhaps motivated by his success in the Kaggle March Madness competition, Michael Lopez joined the National Football League and launched the Big Data Bowl in 2019. This annual competition has similarly fueled new research directions in American football and a *JQAS* special issue on player tracking data in the National Football League (Lopez, 2020). Successful entries and their corresponding publications (Chu et al., 2020; Deshpande & Evans, 2020; Yurko et al., 2020) have launched the careers of several of the most prominent early-career researchers in sports analytics.

Similar competitions that provide opportunities for aspiring researchers to tackle sports analytics problems include the Big Data Cup for ice hockey and the SABR Diamond Dollars Case Competition for baseball, and the Big Data Derby for horse racing.

As the field of sports analytics has grown, a proliferation of sports specific and regional sports analytics conferences have arisen. The biennial New England Symposium on Statistics in Sports is likely the longest-running academic conference devoted to sports analytics. Its West Coast counterpart is The Cascadia Symposium on Statistics in Sports. Many influential results have been showcased for the first time at these conferences. Other prominent sports analytics conferences include the Carnegie Mellon Sports Analytics Conference, UConn Sports Analytics Symposium, and MathSport International.

The highest-profile sports analytics conference is undoubtedly the Sloan Sports Analytics Conference, which draws academics, industry professionals, vendors, and media organizations. While the conference holds a research competition and has certainly drawn attention to sports analytics, it has also been criticized for a variety of shortcomings. These criticisms include a lack of emphasis on reproducibility in the research competition, high ticket prices, the large salaries taken by the organizers, and the lack of diversity among attendees and presenters (Funt, 2022).

It is also worth noting that a significant, but unknown, proportion of the most innovative research is being conducted by professional sports teams. This research will likely never be published, because each team will use it to their competitive advantage. Part of what enables this research is better data. For example, professional sports teams can collect biometric data on their own players, and use that data to learn about how their workouts, sleep patterns, and diets impact their athletic performance. While this research may constitute "emerging methodologies," it unfortunately will take years, if at all, before the public benefits from it.

8 Conclusion

As an applied science, sports analytics may lack a grand unified theory that succinctly characterizes game play across all sports. However, as a maturing discipline, sports analytics has been able to address fundamental questions common to many sports. In this paper, we explore three of those big questions: Who are the best teams and how good are they? What is the likelihood of each team winning the game at any given juncture? Is there a generic framework for evaluating strategies at any given juncture in a game?

Other fundamental questions are addressed elsewhere. How significant is the element of chance in a particular sport? Given that we know who the best teams are, who are the best players and how can we quantify their relative contributions? What combinations of players work best together in a particular sport? In particular, see Lopez et al. (2018) for estimations of the element of chance across four major sports. The second question is often addressed using a formulation of *wins above replacement* (WAR)—see Baumer et al. (2015) and

Yurko et al. (2019) for details in baseball and American football. Recent work by Che & Glickman (2022) also addresses this question across sports. The third question is most compelling in sports like basketball, ice hockey, and soccer, where substitutions are common and it is obvious that different combinations of players with different sets of skills will result in squad of varying strengths and weaknesses. The concept of *plus-minus*, and then *adjusted* plus-minus is frequently applied to address this question (see Hvattum (2019) for a comprehensive overview of applications).

In drawing together these three big ideas in sports analytics, we have also drawn attention to new uses of sports betting market data, some computational tools for doing sports analytics work, and opportunities to showcase that work. Our discussion in Section 6.3 shows how the increased resolution of available data has catalyzed new research directions in baseball, but this same dynamic is playing out in all sports. It is through these exchanges of ideas, tools, models, and data that analytics moves our collective understanding of sports forward.

Acknowledgments

We are grateful to Michael Lopez and Katherine Evans for their thoughts on early versions of this paper.

The R Markdown file that generated this paper, including all R code, is available at `https://github.com/beanumber/wire21`.

References

Agresti, A. (2018). *An introduction to categorical data analysis* (3rd ed.). Hoboken, NJ: John Wiley & Sons.

Alamar, B. (2010). Measuring risk in NFL playcalling. *Journal of Quantitative Analysis in Sports*, *6*(2), 11. `https://doi.org/10.2202/1559-0410.1235`

Albert, J. (2015). Player evaluation using win probabilities in sports competitions. *Wiley Interdisciplinary Reviews: Computational Statistics*, *7*(5), 316–325. `https://doi.org/10.1002/wics.1358`

Albert, J., & Bennett, J. (2001). *Curve ball: Baseball, statistics, and the role of chance in the game*. New York: Springer.

Albert, J., Glickman, M. E., Swartz, T. B., & Koning, R. H. (2016). *Handbook of statistical methods and analyses in sports* (p. 520). New York: Chapman; Hall/CRC Press. `https://doi.org/10.1201/9781315166070`

Baumer, B. S., Jensen, S. T., & Matthews, G. J. (2015). openWAR: An open source system for evaluating overall player performance in Major League Baseball. *Journal of Quantitative Analysis in Sports*, *11*(2), 69–84. `https://doi.org/10.1515/jqas-2014-0098`

Baumer, B. S., Kaplan, D. T., & Horton, N. J. (2021). *Modern Data Science with R* (2nd ed., pp. 1–673). Boca Raton, FL: Chapman; Hall/CRC Press. `https://mdsr-book.github.io/mdsr2e/`

Baumer, B. S., & Matthews, G. J. (2020). *Teamcolors: Color palettes for pro sports teams.* `http://github.com/beanumber/teamcolors`

Baumer, B. S., Nguyen, Q., & Matthews, G. J. (2022). *CRAN task view: Sports analytics.* `https://CRAN.R-project.org/view=SportsAnalytics`

Baumer, B., & Zimbalist, A. (2014). *The sabermetric revolution: Assessing the growth of analytics in baseball* (p. 187). Philadelphia: University of Pennsylvania Press.

Beaudoin, D., & Swartz, T. B. (2010). Strategies for pulling the goalie in hockey. *The American Statistician*, *64*(3), 197–204. `https://doi.org/10.1198/tast.2010.09147`

Bergman, D., & Imbrogno, J. (2017). Surviving a national football league survivor pool. *Oper. Res.*, *65*, 1343–1354. `https://doi.org/10.1287/opre.2017.1633`

Bornn, L., Cervone, D., Franks, A., & Miller, A. (2017). Studying basketball through the lens of player tracking data. In *Handbook of statistical methods and analyses in sports* (pp. 261–286). Boca Raton, FL: CRC Press.

Boulier, B. L., & Stekler, H. O. (2003). Predicting the outcomes of national football league games. *International Journal of Forecasting*, *19*(2), 257–270. `https://doi.org/10.1016/s0169-2070(01)00144-3`

Boulier, B. L., Stekler, H. O., & Amundson, S. (2006). Testing the efficiency of the national football league betting market. *Applied Economics*, *38*(3), 279–284. `https://doi.org/10.1080/00036840500368904`

Bouzarth, E., Grannan, B., Harris, J., Hartley, A., Hutson, K., & Morton, E. (2021). Swing shift: A mathematical approach to defensive positioning in baseball. *Journal of Quantitative Analysis in Sports*, *17*(1), 47–55. `https://doi.org/10.1515/jqas-2020-0027`

Bradley, R. A., & Terry, M. E. (1952). Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, *39*(3/4), 324–345. https://doi.org/10.2307/2334029

Breiter, D. J., & Carlin, B. P. (1997). How to play office pools if you must. *Chance*, *10*(1), 5–11. https://doi.org/10.1080/09332480.1997.10554789

Brenzel, P., Shock, W., & Yang, H. (2019). An analysis of curling using a three-dimensional markov model. *Journal of Sports Analytics*, *5*(2), 101–119. https://doi.org/10.3233/jsa-180279

Buttrey, S. E. (2016). Beating the market betting on NHL hockey games. *Journal of Quantitative Analysis in Sports*, *12*(2), 87–98. https://doi.org/10.1515/jqas-2015-0003

Carl, S. (2022). *nflplotR: NFL logo plots in ggplot2*. https://CRAN.R-project.org/package=nflplotR

Carl, S., & Baldwin, B. (2022). *nflfastR: Functions to efficiently access NFL play by play data*. https://CRAN.R-project.org/package=nflfastR

Caro, C. A., Machtmes, R., et al. (2013). Testing the utility of the pythagorean expectation formula on division one college football: An examination and comparison to the morey model. *Journal of Business & Economics Research (JBER)*, *11*(12), 537–542. https://doi.org/10.19030/jber.v11i12.8261

Carroll, B. N., Palmer, P., Thorn, J., & Pietrusza, D. (1988). *The hidden game of football* (p. 415). New York: Total Sports, Inc.

Carter, Virgil., & Machol, R. E. (1971). Technical note–operations research on football. *Operations Research*, *19*(2), 541–544. https://doi.org/10.1287/opre.19.2.541

Casals, M., Fernández, J., Martínez, V., Lopez, M., Langohr, K., & Cortés, J. (2022). A systematic review of sport-related packages within the R CRAN repository. *International Journal of Sports Science & Coaching*, 1. https://doi.org/10.1177/17479541221136238

Cervone, D., D'Amour, A., Bornn, L., & Goldsberry, K. (2014). *Pointwise: Predicting points and valuing decisions in real time with NBA optical tracking data*. *28*, 3. http://www.lukebornn.com/papers/cervone_ssac_2014.pdf

Cervone, D., D'Amour, A., Bornn, L., & Goldsberry, K. (2016). A multiresolution stochastic process model for predicting basketball possession outcomes. *Journal of the American Statistical Association*, *111*(514), 585–599. https://doi.org/10.1080/01621459.2016.1141685

Che, J., & Glickman, M. (2022). Athlete rating in multi-competitor games with scored outcomes via monotone transformations. *arXiv Preprint arXiv:2205.10746*. https://arxiv.org/pdf/2205.10746

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., Mitchell, R., Cano, I., Zhou, T., Li, M., Xie, J., Lin, M., Geng, Y., Li, Y., & Yuan, J. (2022). *Xgboost: Extreme gradient boosting*. https://github.com/dmlc/xgboost

Chu, D., Reyers, M., Thomson, J., & Wu, L. Y. (2020). Route identification in the National Football League: An application of model-based curve clustering using the EM algorithm. *Journal of Quantitative Analysis in Sports*, *16*(2), 121–132. https://doi.org/10.1515/jqas-2019-0047

Clair, B., & Letscher, D. (2007). Optimal strategies for sports betting pools. *Operations Research*, *55*(6), 1163–1177. https://doi.org/10.1287/opre.1070.0448

Clark, N., Macdonald, B., & Kloo, I. (2020). A Bayesian adjusted plus-minus analysis for the esport Dota 2. *Journal of Quantitative Analysis in Sports*, *16*(4), 325–341. https://doi.org/10.1515/jqas-2019-0103

Cochran, J., Bennett, J., & Albert, J. (Eds.). (2017). *The Oxford anthology of statistics in sports, volume 1: 2000-2004* (p. 544). London: Oxford University Press. https://global.oup.com/academic/product/the-oxford-anthology-of-statistics-in-sports-9780198724926

Deshpande, S. K., & Evans, K. (2020). Expected hypothetical completion probability. *Journal of Quantitative Analysis in Sports*, *16*(2), 85–94. https://doi.org/10.1515/jqas-2019-0050

Deshpande, S. K., & Jensen, S. T. (2016). Estimating an NBA player's impact on his team's chances of winning. *Journal of Quantitative Analysis in Sports*, *12*(2). https://doi.org/10.1515/jqas-2015-0027

Deshpande, S. K., & Wyner, A. (2017). A hierarchical bayesian model of pitch framing. *Journal of Quantitative Analysis in Sports*, *13*(3), 95–112. https://doi.org/10.1515/jqas-2017-0027

Dewan, J., & Zminda, D. (1993). *STATS basketball scoreboard 1993-1994* (p. 288). HarperPerennial.

Douglas, C., & Scriven, R. (2021). *Retrosheet: Import professional baseball data from retrosheet*. https://github.com/colindouglas/retrosheet

Drucker, R. (2022). *sportyR: Plot scaled ggplot representations of sports playing surfaces*. https://github.com/sportsdataverse/sportyR

Elo, A. E. (1978). *The rating of chessplayers, past and present*. New York: Arco Publishing.

Fernandez, J., Bornn, L., & Cervone, D. (2021). A framework for the fine-grained evaluation of the instantaneous expected value of soccer possessions. *Machine Learning*, *110*(6), 1389–1427. https://doi.org/10.1007/s10994-021-05989-6

FIDE. (2022). *Rating calculator*. https://ratings.fide.com/calc.phtml

Friendly, M., Dalzell, C., Monkman, M., & Murphy, D. (2022). *Lahman: Sean lahman baseball database*. https://CRAN.R-project.org/package=Lahman

Funt, D. (2022). *At Sloan sports conference, criticism mounts over diversity, access*. The Washington Post. https://www.washingtonpost.com/sports/2022/06/13/sloan-sports-conference-diversity/

Gandar, J., Zuber, R., O'Brien, T., & Russo, B. (1988). Testing rationality in the point spread betting market. *The Journal of Finance*, *43*(4), 995–1008. https://doi.org/10.1111/j.1540-6261.1988.tb02617.x

Gilani, S. (2022). *SportsDataverse*. https://sportsdataverse.org

Glickman, M. E., & Sonas, J. (2015). Introduction to the NCAA men's basketball prediction methods issue. *Journal of Quantitative Analysis in Sports*, *11*(1), 1–3. https://doi.org/10.1515/jqas-2015-0013

Glickman, M. E., & Stern, H. S. (1998). A state-space model for national football league scores. *Journal of the American Statistical Association*, *93*(441), 25–35. https://doi.org/10.1080/01621459.1998.10474084

Glickman, M. E., & Stern, H. S. (2017). Estimating team strength in the NFL. In *Handbook of statistical methods and analyses in sports* (pp. 113–136). Boca Raton, FL: CRC Press. http://glicko.net/research/nfl-chapter.pdf

Goldner, K. (2012). A Markov model of football: Using stochastic processes to model a football drive. *Journal of Quantitative Analysis in Sports*, *8*(1). https://doi.org/10.1515/1559-0410.1400

Goldner, K. (2017). Situational success: Evaluating decision-making in football. In *Handbook of statistical methods and analyses in sports* (pp. 199–214). Boca Raton, FL: CRC Press.

Guan, T., Nguyen, R., Cao, J., & Swartz, T. (2022). In-game win probabilities for the National Rugby League. *The Annals of Applied Statistics*, *16*(1). https://doi.org/10.1214/21-aoas1514

Hamilton, H. H. (2011). An extension of the pythagorean expectation for association football. *Journal of Quantitative Analysis in Sports*, *7*(2). https://doi.org/10.2202/1559-0410.1335

Healey, G. (2017). The new Moneyball: How ballpark sensors are changing baseball. *Proceedings of the IEEE*, *105*(11), 1999–2002. https://doi.org/10.1109/JPROC.2017.2756740

Healey, G. (2019). A bayesian method for computing intrinsic pitch values using kernel density and nonparametric regression estimates. *Journal of Quantitative Analysis in Sports*, *15*(1), 59–74. https://doi.org/10.1515/jqas-2017-0058

Horowitz, M., Yurko, R., Ventura, S., & Dutta, R. (2020). *NflscrapR: Compiling the NFL play-by-play API for easy use in r*. https://github.com/maksimhorowitz/nflscrapR

Hvattum, L. M. (2019). A comprehensive review of plus-minus ratings for evaluating individual players in team sports. *International Journal of Computer Science in Sport*, *18*(1). https://doi.org/10.2478/ijcss-2019-0001

Imbrogno, J., & Bergman, D. (2022). Computing the number of winning NFL survivor pool entries. *The College Mathematics Journal*, *53*(4), 282–291. https://doi.org/10.1080/07468342.2022.2099704

James, B. (2003). *The new Bill James historical baseball abstract*. Free Press.

Kaplan, E. H., & Garstka, S. J. (2001). March madness and the office pool. *Management Science*, *47*(3), 369–382. https://doi.org/10.1287/mnsc.47.3.369.9769

Koning, R. H. (2017). Rating of team abilities in soccer. In *Handbook of statistical methods and analyses in sports* (pp. 371–388). Boca Raton, FL: CRC Press.

Koopman, S. J., & Lit, R. (2015). A dynamic bivariate poisson model for analysing and forecasting match results in the english premier league. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *178*(1), 167–186. https://doi.org/10.1111/rssa.12042

Kovalchik, S. A. (2016). Searching for the GOAT of tennis win prediction. *Journal of Quantitative Analysis in Sports*, *12*(3), 127–138. https://doi.org/10.1515/jqas-2015-0059

Kovalchik, S. A., & Reid, M. (2019). A calibration method with dynamic updates for within-match forecasting of wins in tennis. *International Journal of Forecasting*, *35*(2), 756–766. https://doi.org/10.1016/j.ijforecast.2017.11.008

Kumagai, B., Nahabedian, M., Châtel, T., & Stokes, T. (2021). *Bayesian space-time models for expected possession added value*. Hockey-Graphs. https://hockey-graphs.com/2021/07/06/bayesian-space-time-models-for-expected-possession-added-value-part-1-of-2/

Lacey, N. J. (1990). An estimation of market efficiency in the NFL point spread betting market. *Applied Economics*, *22*(1), 117–129. https://doi.org/10.1080/00036849000000056

Lente, C. (2020). *Chess: Read, write, create and explore chess games*. https://github.com/curso-r/chess

Lewis, M. (2004). *Moneyball: The art of winning an unfair game* (p. 336). New York: WW Norton & Company.

Lindsey, G. R. (1961). The progress of the score during a baseball game. *Journal of the American Statistical Association*, *56*(295), 703–728. https://doi.org/10.1080/01621459.1961.10480656

Lindsey, G. R. (1963). An investigation of strategies in baseball. *Operations Research*, *11*(4), 477–501. https://doi.org/10.1287/opre.11.4.477

Lock, D., & Nettleton, D. (2014). Using random forests to estimate win probability before each play of an NFL game. *Journal of Quantitative Analysis in Sports*, *10*(2). https://doi.org/10.1515/jqas-2013-0100

Lopez, M. J. (2022). personal communication.

Lopez, M. J. (2020). Bigger data, better questions, and a return to fourth down behavior: An introduction to a special issue on tracking data in the national football league. *Journal of Quantitative Analysis in Sports*, *16*(2), 73–79. https://doi.org/10.1515/jqas-2020-0057

Lopez, M. J., & Matthews, G. J. (2015). Building an NCAA men's basketball predictive model and quantifying its success. *Journal of Quantitative Analysis in Sports*, *11*(1), 5–12. https://doi.org/10.1515/jqas-2014-0058

Lopez, M. J., Matthews, G. J., & Baumer, B. S. (2018). How often does the best team win? A unified approach to understanding randomness in North American sport. *The Annals of Applied Statistics*, *12*(4). https://doi.org/10.1214/18-aoas1165

Macdonald, B. (2012). An expected goals model for evaluating NHL teams and players. *Proceedings of the 2012 MIT Sloan Sports Analytics Conference*. https://assets.pubpub.org/mku181yp/ee0d61ed-af35-4b1f-ba86-71c216935690.pdf

Marchi, M., Albert, J., & Baumer, B. S. (2018). *Analyzing baseball data with R* (2nd ed., p. 360). Boca Raton, FL: Chapman; Hall/CRC Press. https://doi.org/10.1201/9781351107099

Martin, R., Timmons, L., & Powell, M. (2018). A Markov chain analysis of NFL overtime rules. *Journal of Sports Analytics*, *4*(2), 95–105. https://doi.org/10.3233/JSA-170198

Maymin, P. Z. (2021). Smart kills and worthless deaths: eSports analytics for league of legends. *Journal of Quantitative Analysis in Sports*, *17*(1), 11–27. https://doi.org/10.1515/jqas-2019-0096

McFarlane, P. (2019). Evaluating NBA end-of-game decision-making. *Journal of Sports Analytics*, *5*(1), 17–22. https://doi.org/10.3233/jsa-180231

Metrick, A. (1996). March madness? Strategic behavior in NCAA basketball tournament betting pools. *Journal of Economic Behavior & Organization*, *30*(2), 159–172. https://doi.org/10.1016/S0167-2681(96)00855-4

Miller, S. J. (2007). A derivation of the pythagorean won-loss formula in baseball. *Chance*, *20*(1), 40–48. https://doi.org/10.1080/09332480.2007.10722831

Mills, E. G., & Mills, H. D. (1970). *Player win averages: A complete guide to winning baseball players*. The Harlan D. Mills Collection. https://trace.tennessee.edu/utk_harlan/6/

Nichols, M. W. (2014). The impact of visiting team travel on game outcome and biases in NFL betting markets. *Journal of Sports Economics*, *15*(1), 78–96. https://doi.org/10.1177/1527002512440580

Niemi, J. B., Carlin, B. P., & Alexander, J. M. (2008). Contrarian strategies for NCAA tournament pools: A cure for March madness? *Chance*, *21*(1), 35–42. https://doi.org/10.1080/09332480.2008.10722884

Paul, R. J., & Weinbach, A. P. (2014). Market efficiency and behavioral biases in the WNBA betting market. *International Journal of Financial Studies*, *2*, 193–202. https://doi.org/10.3390/ijfs2020193

Pelechrinis, K., Winston, W., Sagarin, J., & Cabot, V. (2019). Evaluating NFL plays: Expected points adjusted for schedule. *International Workshop on Machine Learning and Data Mining for Sports Analytics*, *11330*, 106–117. https://doi.org/10.1007/978-3-030-17274-9_9

Petti, B., & Gilani, S. (2022). *Baseballr: Acquiring and analyzing baseball data*. https://CRAN.R-project.org/package=baseballr

R Core Team. (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Reyers, M., & Swartz, T. B. (2021). Quarterback evaluation in the National Football League using tracking data. *AStA Advances in Statistical Analysis*. https://doi.org/10.1007/s10182-021-00406-8

Romer, D. (2006). Do firms maximize? Evidence from professional football. *Journal of Political Economy*, *114*(2), 340–365. https://doi.org/10.1086/501171

Sauer, R. D. (1998). The economics of wagering markets. *Journal of Economic Literature*, *36*(4), 2021–2064. https://www.jstor.org/stable/2565046

Schuhmann, J. (2021). *NBA's 3-point revolution: How 1 shot is changing the game*. NBA.com. https://www.nba.com/news/3-point-era-nba-75

Schwarz, A. (2004). *The numbers game: Baseball's lifelong fascination with statistics*. New York: Thomas Dunne Books/St. Martin's Press.

Sicilia, A., Pelechrinis, K., & Goldsberry, K. (2019, July). DeepHoops: Evaluating Micro-Actions in Basketball Using Deep Feature Representations of Spatio-Temporal Data. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. https://doi.org/10.1145/3292500.3330719

Sidle, G., & Tran, H. (2018). Using multi-class classification methods to predict baseball pitch types. *Journal of Sports Analytics*, *4*(1), 85–93. https://doi.org/10.3233/JSA-170171

Sievert, C. (2015). *pitchRx: Tools for harnessing MLBAM 'gameday' data and visualizing pitchfx*. http://cpsievert.github.com/pitchRx

Skinner, B. (2011). Scoring strategies for the underdog: A general, quantitative method for determining optimal sports strategies. *Journal of Quantitative Analysis in Sports*, *7*(4). https://doi.org/10.2202/1559-0410.1364

Soebbing, B. P., & Humphreys, B. R. (2013). Do gamblers think that teams tank? Evidence from the NBA. *Contemporary Economic Policy*, *31*(2), 301–313. https://doi.org/10.1111/j.1465-7287.2011.00298.x

Spann, M., & Skiera, B. (2009). Sports forecasting: A comparison of the forecast accuracy of prediction markets, betting odds and tipsters. *Journal of Forecasting*, *28*(1), 55–72. https://doi.org/10.1002/for.1091

Stern, H. S. (1994). A brownian motion model for the progress of sports scores. *Journal of the American Statistical Association*, *89*(427), 1128–1134. https://doi.org/10.1080/01621459.1994.10476851

Tango, T. M., Lichtman, M. G., & Dolphin, A. E. (2007). *The book: Playing the percentages in baseball*. Sterling, VA: Potomac Books.

Turner, H., & Firth, D. (2020). *BradleyTerry2: Bradley-terry models*. https://github.com/hturner/BradleyTerry2

Urschel, J., & Zhuang, J. (2011). Are NFL coaches risk and loss averse? Evidence from their use of kickoff strategies. *Journal of Quantitative Analysis in Sports*, *7*(3), 14. https://doi.org/10.2202/1559-0410.1311

White, C., & Berry, S. (2002). Tiered polychotomous regression: Ranking NFL quarterbacks. *The American Statistician*, *56*(1), 10–21. https://doi.org/10.1198/000313002753631312

Wickham, H., Chang, W., Henry, L., Pedersen, T. L., Takahashi, K., Wilke, C., Woo, K., Yutani, H., & Dunnington, D. (2022). *ggplot2: Create elegant data visualisations using the grammar of graphics*. https://CRAN.R-project.org/package=ggplot2

Winston, W. L., Nestler, S., & Pelechrinis, K. (2022). *Mathletics: How gamblers, managers, and fans use mathematics in sports* (2nd ed.). Princeton, NJ: Princeton University Press.

Yam, D. R., & Lopez, M. J. (2019). What was lost? A causal estimate of fourth down behavior in the national football league. *Journal of Sports Analytics*, *5*(3), 153–167. https://doi.org/10.3233/jsa-190294

Yurko, R., Matano, F., Richardson, L. F., Granered, N., Pospisil, T., Pelechrinis, K., & Ventura, S. L. (2020). Going deep: Models for continuous-time within-play valuation of game outcomes in american football with tracking data. *Journal of Quantitative Analysis in Sports*, *16*(2), 163–182. https://doi.org/10.1515/jqas-2019-0056

Yurko, R., Ventura, S., & Horowitz, M. (2019). nflWAR: A reproducible method for offensive player evaluation in football. *Journal of Quantitative Analysis in Sports*, *15*(3), 163–183. https://doi.org/10.1515/jqas-2018-0010

Zivkovic, J. (2022). *chessR: Functions to extract, clean and analyse online chess game data*. https://github.com/JaseZiv/chessR