International Journal of Data Science and Analytics
https://doi.org/10.1007/s41060-017-0093-7

TRENDS OF DATA SCIENCE

Sports analytics and the big-data era

Elia Morgulev[1,2] · Ofer H. Azar[1] · Ronnie Lidor[2]

Received: 9 August 2017 / Accepted: 28 December 2017

Abstract
The explosion of data, with large datasets that are available for analysis, has affected virtually every aspect of our lives. The sports industry has not been immune to these developments. In this article, we provide examples of three types of data-driven analyses that have been performed in the domain of sport: (a) field-level analysis focused on the behavior of athletes, coaches, and referees; (b) analysis of management and policymakers' decisions; and (c) analysis of the literature that uses sports data to address various questions in the fields of economics and psychology.

Keywords Big data · Data analytics · Decision making · Sports · Psychology · Economics

1 Introduction

Sport is an important endeavor in the lives of many people. One reason is that many of them are engaged in sport as a way of exercising and improving their health and life style. Another reason is that watching and keeping track of professional sports is a major activity shared by both young and adult individuals. People all around the globe watch sports on television, many on a daily basis. In addition, sports fans tend to be highly involved, reflecting on coaches' decisions, comparing players' metrics, and predicting outcomes of games and the final ranks of individuals and teams playing in competitions. Many newspapers contain a regular sports section, and entire television channels are devoted to both individual and team sports. Major sport competitions, such as the Olympic Games, the World Cup in soccer, or the World Championships in basketball and swimming, are among the most popular events worldwide. Billions of dollars are involved in the various aspects of the sports industry, from the cost of game tickets to payment for broadcasting licenses, salaries of top players, and advertising.

In his book on the promise and the pitfalls of big data, Nate Silver [51] elaborated on the possibilities of performance assessment and sport scouting that have been unleashed in the big-data era. Silver initially gained his reputation when he succeeded in determining causality and separating skill from luck, aggregating extensively large datasets on major league baseball players' performance. Variables that predict future performance were elicited by Silver based on analysis of thousands of players during more than five decades in the major leagues. This effort enabled him to estimate predictors' parameters and to devise a forecasting model that tends to outperform expert scouts. Silver speculated that baseball may offer the world's richest dataset, where just about everything that has happened on a major league field in the past 140 years has been accurately recorded and is now available for analysis.

Since data from different sport events have been regularly recorded for many years, and are often available to the public at large (entire games are recorded on video in addition to the quantitative data that are retained in datasets), the domain of sport provides a uniquely authentic arena for exploring research ideas. In addition to the big-data characteristic of sport, other factors also contribute to sport being an excellent source for analytics and research, particularly those concerning certain aspects of human behavior. The rules of the games in sport are clear and well defined. The players in professional sports are considered to be experts, and are offered large incentives to perform the best that they can. Differences between sports (for example, individual sports versus team sports) create a variety of situations, each allowing the assessment of different aspects of performance. Sports being prevalent worldwide allows for global analysis, either

✉ Ofer H. Azar
azar@som.bgu.ac.il

1 Department of Business Administration, Guilford Glazer Faculty of Business and Management, Ben-Gurion University of the Negev, Beer-Sheva, Israel

2 The Academic College at Wingate, Wingate Institute, Netanya, Israel

Published online: 09 January 2018

using comparisons between countries, or aggregating data from different parts of the world. Patterns of behavior in sport can often provide insights about all types of human behavior, because universal phenomena that affect human behavior in general will often be reflected in sports behavior as well.

The data for research in sports are provided in many cases by companies that specialize in measuring and coding sport performance with an eye for selling customized packages of information for clubs, associations, broadcasters, and academic researchers. For example, these companies may assist the team's scouting staff with detailed information on soccer or basketball players' performance in every league on the globe that plays on a professional or semi-professional level. The purpose of this article is to provide examples of data-driven analysis in sports, sometimes with implications outside sports as well.

2 Sports analytics defined

Sports analytics is the investigation and modeling of sports performance, implementing scientific techniques. More specifically, sports analytics refers to the management of structured historical data, the application of predictive analytic models that use these data, and the utilization of information systems, in order to inform decision makers and enable them to assist their organizations in gaining a competitive advantage on the field of play ([1], see also [39,40]). Historical data can be either quantitative or qualitative; these data are typically collected from multiple sport-relevant resources, among them biographical data, films/videos, box-score performance data, medical reports of the athletes, and scouting reports. The collected data are standardized, centralized, integrated, and analyzed using different metrics. It is assumed that a reliable and systematic analysis of the data will enable coaches, athletes, and policymakers to strengthen their decision-making processes. A sports analytics framework is described in Fig. 1.

3 Development of field-level oriented analysis

Sports analytics originated in the 1960s in the USA, where American football and basketball were analyzed using coded notes (i.e., notational analysis) [25]. Notational analysis is an objective way of recording performance, so that critical events in that performance can be quantified in a consistent and reliable manner [25]. Such analysis enables the coach and the manager to objectively assess competitive performance, and therefore to improve it.

Another popular American sport is the game of baseball. Baseball is a less dynamic game than football or basketball, and as a result it is more convenient to break down into distinct events to be analyzed. It was in baseball where the first platform for statisticians to work with individual and team performance data (box score) was developed, during the second half of nineteenth century. By 1971, a group of baseball analysts founded the *Society for American Baseball Research* (*SABR*) [21].

In the 1950s, in England, a retired Royal Air Force Wing Commander and an amateur statistician named Charles Reep began to analyze the number of passes in soccer that led to a goal, alongside the field positions where those passes originated. Reep's work led to the one of the first scientific publications in sports analytics [48] and constituted the "long ball" style of play which for decades stamped its mark on English soccer.

Yet gathering sports' data and conducting a comprehensive analysis was an extremely time-demanding task during the pre-computerized era. For example, the first hand notational system developed for tennis was never actually used due to its complexity, whereas another analysis system developed for squash took five to eight hours to master and an additional 40–50 h to analyze the data from a single game [30].

Technological developments in the 1980s enabled gradual computerization both of the data-gathering process and

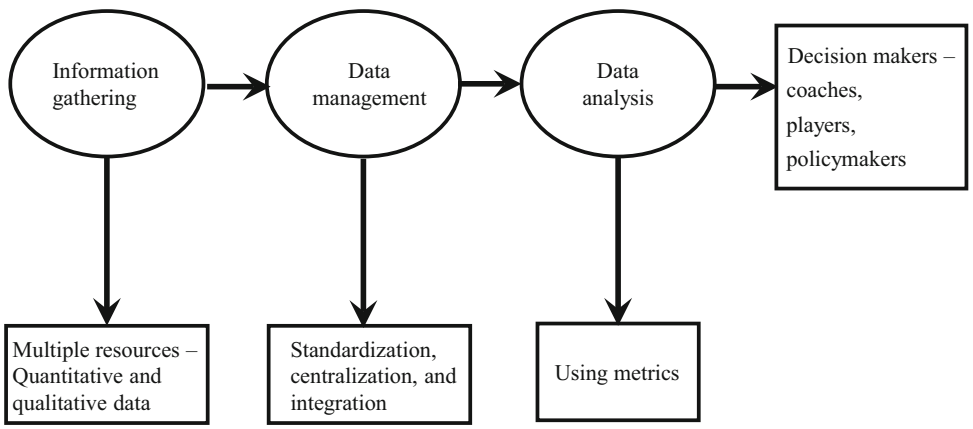

Fig. 1 A sports analytics framework

of its analysis. Computerized versions of notational analysis for tennis and squash were implemented by the end of that decade [30]. In 1989 David Smith, a biology professor, founded *Retrosheet*, a nonprofit organization aimed at computerizing the box score of every major league baseball game ever played, in order to analyze the statistics of the game [19]. Smith drew on the previous work of another well-known baseball enthusiast, Bill James, who became frustrated about the major league administration's refusal to publish play-by-play game accounts, and therefore initiated the *Scoresheet* project—a network of fans who would collect and distribute this information. Insights from this collaborative effort by Smith, James, and other members of *SABR* society were implemented by the professional staff of the Oakland Athletics franchise, where a more quantitative approach to baseball was put to use in the 1990s. Billy Beane took over as general manager of the Oakland Athletics in 1997, and capitalizing on statistical approaches he was able to assemble a highly efficient team. These events were later popularized in the best-selling book [35] and movie, *Moneyball*.

From the mid-1990s, professional sports gradually entered the big-data era. As an example, throughout the 1995–1996 National Basketball Association (NBA) season, *Advanced Scout* software was distributed to 16 NBA teams. The raw data from games were initially collected using a unique system designed for logging basketball data. Information was collected on various defensive and offensive variables of the game, among them the number of players' shot attempts, type of shots taken, and the number of rebounds taken by players. At the end of each game, the data were uploaded to an electronic bulletin board, and a team could download their own data or the data of any other team from this billboard. The *Advanced Scout* software was able to seek out and discover meaningful patterns in the game [9]. A number of years later, in the 2003–2004 NBA season, data on players' shot attempts were already publicly available on sites such as *espn.com* [49], and therefore could be used for systematic analytics. For example, researchers who analyzed 1270 shot attempts made by one player in the above-mentioned season discovered useful predictors of both shot location and field goal percentage, and proposed a new statistical model for analyzing basketball shot charts [49].

In soccer, teams in the *English Premier League* (EPL) became relatively advanced in terms of performance analytics, and some of them have even made performance data available to fans for open-source analysis [21]. However, when compared to basketball, assessing players' ability to score in soccer is hindered by the low frequency of scoring events. Tactical factors, such as number and length of possessions, passing sequences, and spatial analysis of the territory played are aggregated in order to optimize performance in offense [21]. A specific example of how players and coaches may benefit from the assessment of large samples of events in soccer is the information on probable directions of penalty shots, based on the shooters' previous statistics provided by the analysts to the goalkeepers before critical matches [42].

In the 2010s analysis of video data became possible across many professional sports. For example, in the NBA a camera system (SportVU), originally based on Israeli missile-tracking technology, became mandatory in all arenas. This system is hung above the court, and records the ball and players' movement data. In baseball, PITCHf/x, HITf/x, and FIELDf/x video systems are used to capture and analyze pitching, hitting, and fielding, respectively. Video analytics systems in basketball are able to produce huge datasets of players' movements, ball touches, rebounds, and shot locations. Nowadays, the amount of data is apparently larger than the capability to extract possible insights from it [21].

3.1 Implications of big-data analytics in the field

One notable impact of studies analyzing big-data files is the transition toward the three-point shooting style of play, evident nowadays in the NBA. Annual data on shot location have enabled analysts to develop a model of expected points per shot from each location on the court. This model revealed that decisions on long-range two-pointers are inferior to the selection of three-pointer shots. Visualization of ball movements and shot outcomes [29] allowed the players not only to optimize the allocation of the ball between the team members in an attack, but also to learn about the best positioning for a defensive rebound, depending on the player who attempted the shot, and the spot from which the shot was taken.

Combining modern statistical projections with traditional scouts' insights leads to more accurate assessments of a player's prospects at the professional level. For instance, the Boston Celtics were able to pick future all-star Rajon Rondo late in the 2006 NBA draft because they identified rebounding by guards as an undervalued skill in the NBA. Other teams at that time did not realize the potential value of this performance indicator as the more analytical Celtics did [1].

Based on comprehensive datasets of on-court performances in basketball, analysts were able to develop a number of sophisticated performance game-related indicators, allowing them (a) to account for the number of minutes played by the players when comparing points scored by starter and backup players; (b) to distinguish shooting accuracy from shooting selection when deciding on the acquisition of players competing in different leagues; and (c) to control for overall rebound opportunities when assessing a player's rebound ability.

Alamar [1] described the problem of projecting a player's development in a position which differs from the one he has played so far. For example, the NBA player Russel Westbrook played predominantly as a shooting guard and attracted the attention of the Supersonics who were in need of a point guard

(a player who passes rather than shoots) and not a shooting guard (player who shoots rather than passes). The ordinary "number of assists" metric was found to be insufficient in the assessment of Westbrook's passing ability, and therefore a new performance indicator, which measured the change in the team's shooting percentage when a specific player made a pass to the shooter, was created by Alamar. This analysis revealed that Westbrook's effect on his teammates' shooting ability was of the same caliber as of the top point guards in the NBA.

Similar processes in football led to the development of metrics that account for the quality of the catcher, the strength of the defense, and the effectiveness of the linebackers, while evaluating the quarterback's passing skill. In baseball, machine learning is being leveraged to predict the pitching behavior of players to better inform in-game decision making. The outcome of this is a model that can predict upcoming pitches using real-time game statistics, with an accuracy of 74.5%. This model incorporates factors such as the type of pitches thrown by particular pitchers/teams, the number and position of the players on base, the ball-strike count, and the number of innings played [26].

In soccer, bias toward seeing "what is there" and ignoring "what is not there" makes evaluating defense difficult. Attacking in soccer has one simple best outcome: scoring a goal. But defending is quite the opposite: there, the best outcome is a goal that is not conceded—an event that does not actually happen. This may be because of a shot that did not come, a cross that was not made, or a through ball that could not be passed properly. As a result, for instance, in 2001 Sir Alex Ferguson, one of the most successful managers in British soccer history, decided to sell the Dutch international defender Jaap Stam to Lazio. The sale was prompted partly by match data. Studying the numbers, Ferguson had spotted that Stam was tackling less often than before. He presumed that the defender, then twenty-nine, was declining. So he sold him. Ferguson has called this decision the biggest mistake of his career [2].

Advanced information systems in soccer (e.g., *Opta*, *Prozone*) provide the decision maker with heat maps and visualizations of ball movements on the pitch. While looking at such data, the coach is able to identify patterns for the defender's ability to prevent passes and penetrations to the area that he is in charge of, even without producing clears and tackles. Such a skill has remained under the radar so far.

Another important source of information on players' and teams' performance is locational and biometric devices (e.g., GPS devices, radio frequency devices, accelerometers). Such devices are most frequently used to assess the total physical activity undertaken by players in games and practices [21]. Devices worn by players provide a rich source of objective information on their external workloads and movement patterns. In addition to location-based (x–y coordinates) and distance–time (speed) data, GPS units are fitted with accelerometers, gyroscopes, and magnetometers, providing data on accelerations, decelerations, change of direction movements, and vertical jumps performed. Strength and conditioning coaches can modify training intensity according to this objective information, and use it to help them decide about in-game substitutions and player rotations. Based on these data, sport analysts have developed models for predicting the risk of injury and are now able to alert the coach when player workloads are mismanaged [26].

4 Management and policymakers' decision-oriented analysis

The business-oriented analysis of sports is excessively versatile and addresses areas from economic assessment of the impact of mega-sport events to allocation of scarce resources while building professional teams' rosters, to optimal ticket pricing via evaluation of the fans' level of interest. In this section, we present various examples of business-oriented research of sports, with an eye to briefly introducing the reader to this wide spectrum of scientific endeavor.

Researchers are now able to produce a detailed cost–benefit analysis of hosting mega-sport events (e.g., Olympic Games, FIFA World Cup). For instance, Billings and Holladay [10] examined whether hosting the Olympic Games improves a city's long-term growth. The researchers matched the host cities with cities that were finalists for the Olympic Games, but were not selected by the International Olympic Committee. An examination of post-Olympic impacts for host cities between 1950 and 2005 indicated no long-term impacts of hosting an Olympics on population, real Gross Domestic Product per capita, or trade openness.

Baade and Matheson [5] suggested three major categories on the cost side of hosting an Olympic Games: (a) general infrastructure such as transportation and accommodation; (b) the specific sports infrastructure required for competition venues; and (c) operational costs. On the benefit side, they proposed immediate tourist spending during the Games, long-term benefits (i.e., an Olympic legacy) that might include improvements in infrastructure, and increased trade, foreign investment, or tourism after the Games, as well as intangible benefits such as the "feel-good effect" and civic pride. Each of these costs and benefits was assessed, and the main conclusion was that in most cases the Olympic Games are a money-losing enterprise; they result in positive net benefits only under very specific circumstances in developed countries. As early as 1956, Rottenberg [50] proposed that the output of teams depends on player skills, training facilities, the stadium, the management, and other owner-supplied resources. Rottenberg formulated the "uncertainty of outcome hypothesis," which suggests that fans receive more

utility from viewing competitions with an unpredictable outcome. This principle implies that teams should possess even playing abilities to some extent, in order for the game's outcome to be less certain.

On these grounds, league authorities often invoke outcome uncertainty as a rationale for intervention measures. For instance, in the National Football League (NFL), ticket revenue sharing, equal broadcast revenue sharing, and a salary cap are all combined steps meant to secure a certain degree of competitive balance across the teams; that is, to prevent large market teams from acquiring excessive talent relative to the rest of the league [14]. Interestingly, a more recent analysis of big datasets of annual league-level attendance both in Europe and USA, controlling for a large number of plausible influences on game/match day attendance, provided evidence contradicting the uncertainty of the outcome hypothesis [14,20,43].

An acute issue confronting team managers since the inception of the salary cap is how to optimally distribute limited resources across their team members. Borghesi [12] examined the relationship between compensation and performance during 10 consecutive seasons in the NFL, and provided evidence that productivity, draft, and experience variables are significantly related to the levels of base and bonus salaries. Teams that compensate players the most inequitably were found to be the most likely to perform the worst. Borghesi pointed out that superstar players on a roster can be disruptive, even if their output on the field justifies their compensation levels.

Optimization of ticket pricing is a legitimate way for teams to increase their revenues. For decades, the seat location was the sole determinant of price. However, since the 2010s a dynamic pricing strategy, where ticket prices fluctuate daily based on changing market conditions, was introduced in sports. Analysis of the price determinants provides ticket sellers a basis on which to set prices, and therefore is critical for revenue generation [52]. For example, the Boston Red Sox baseball team has monitored the flow of fans into the stadium. Analyzing the entrance used by fans relative to their seat location, they optimized the location of concession and memorabilia stands in order to minimize the lines and the distance fans would have to cover to find food and souvenirs [26].

Another example of using a big-data approach to improve managerial decisions is the customer relationship management (CRM) systems, which gather information on customers at various touch points. The data collected are then used to guide a sport team's relationship with its customers, to build a loyal fan base and to increase revenues. The English Premier League's Manchester City Football Club performs customer-based, data-driven marketing very successfully. The club provides its supporters with member cards to be used for buying tickets, entering the stadium, making purchases at the stadium, and so on. The stadium has a system that interacts with the members' cards and gathers data throughout their visit to the stadium. The data is stored in a CRM system and allows the club to understand their fans' behavior in great detail. The club uses the insights derived from the data analyses to engage with their fans, to build deeper long-term relationships with them, and to add value to their relationships with the club [26].

Recently, sports franchises have become interested in developing targeted approaches to marketing based on a fan's history and past purchases. At the same time, social media has come to be an important source of ample data, and if used appropriately a new lens for many new types of studies [37]. In this regard, the use of social media analytics is a practice at the frontier of measuring fan engagement [21]. For example, Bagić Babac and Podobnik [6] analyzed user comments published on the Facebook pages of the top five 2015–2016 Premier League soccer clubs. They shed light on who, how, and why fans participate in social media sport websites, and suggest that outcomes from social media mining bring insights about human behavior patterns that are not visible otherwise. Such results have the potential to influence soccer marketing and to encourage organizations to develop new strategies, for example in targeting women as growing consumers of soccer-related products.

The sports betting industry is another solid sector of sports business where both customers and suppliers attempt to correctly predict outcomes of future events. Consequently, researchers continue to introduce a variety of models that are formulated by diverse forecast methodologies. These models may be targeted on the prediction of results of individual matches [17] or on tournament outcomes [38]. Due to the extremely competitive nature of the gambling market, it has become an arena for the implementation of advances in computing and machine learning, with cutting-edge predictive algorithms (see [17,18,38,41,56]).

5 Analysis of sports data to learn about human behavior

Over 30 years ago, Gilovich et al. [28] caused a stir by debunking a widely accepted belief in the tendency of a hitting player to produce more and more hits (i.e., "hot hand"). The analyzed data provided no evidence for a positive correlation between the outcomes of successive shots. The primary focus of this study was by no means the assessment of basketball players' performance, but rather to demonstrate a common bias caused by implementation of the representativeness heuristic [55]. The data were taken from records of 48 home games of the Philadelphia 76ers, collected by the team's statistician. The authors stated that records of consecutive shots during basketball games for individual players

were not available for other teams in the NBA during that time. Gilovich et al. suggested that the failure to detect evidence of streak shooting might also be attributed to the selection of shots by individual players and the defensive strategy of the opposing teams. However, they also reached the conclusion that there is no correlation between the outcomes of successive shots, both after analyzing free-throw records of the Boston Celtics and from a controlled shooting experiment with the males and females of Cornell's varsity teams.

The study by Gilovich et al. is a showcase of research relying on data from the field of sports. It is one of the most notable instances of sports being used as a laboratory for assessment of important psychological-economic theories. On Google Scholar, for example, this study is cited over 1300 times, and the literature on the hot hand is very broad today (see, for example, the review article [7]).

Almost 30 years later, another study showed how technological developments and the much wider availability of data today are reflected in current research. Bocskocsky et al. [11] used novel metrics provided by the optical tracking system SportVu alongside play-by-play data recorded for each game in the NBA. These researchers analyzed a dataset of over 83,000 shots from the 2012–2013 season, combined with data of both players and ball position in each shot attempt. Relying on such a rich database, they were able to construct a comprehensive model of shot difficulty, and by this to demonstrate that players who exceeded their expectations in shots over recent attempts took shots from significantly further away, faced tighter defense, and were more likely to take their team's next shot.

Economists noted that the essence of game theory is to facilitate understanding and to predict behavior in economic, social, and political contexts [34]. However, testing game theory predictions has proven to be extremely difficult, and as a result even the most fundamental premises have not yet been supported empirically in real situations [46]. One of the fundamental basic tools in analyzing behavior in games (in the field of game theory) is the notion of mixed strategies, where players may play some of their strategies with certain probabilities rather than pursue a single pure strategy. Basic concepts in strategic situations that require the player to be unpredictable are von Neumann's Minimax Theorem and the mixed-strategy Nash equilibrium ("MSNE"). These two concepts were examined empirically in laboratory experiments, but given the advantages of field data, especially when decision makers are experts and face large incentives, several attempts have been made to examine these concepts with sports data as well. These studies attempted to conclude whether professional players seem to be playing according to the MSNE, and thus to obtain insights about the usefulness of this solution concept. In another study, Walker and Wooders [57] used data on championship (Grand Slam) professional tennis matches and found that win rates in the serve and return play are consistent with the Minimax hypothesis.

Penalty kicks in soccer are a good context to examine the MSNE concept, since they present a situation with large incentives (due to the small number of goals in soccer, a penalty kick can determine the entire game), and start in the same way every time (as opposed to complex situations during play), with only two players involved, and with simultaneous play and simple rules and potential strategies. As such, a number of studies used data on penalty kicks to examine the MSNE concept and its applicability to behavior in real games. In one study, Chiappori, Levitt, and Groseclose [16] collected data on penalty kicks in Division 1 (the highest division for competitive soccer) in France and Italy, and developed a theoretical model of the penalty kick game as a simultaneous 3X3 (each player can play right, left, or center) game between the kicker and the goalkeeper. Chiappori et al. made some assumptions about the payoffs in the game, and derived predictions on which strategies or combinations of strategies should be more common than others if the players play the MSNE. The predictions, and therefore the MSNE concept, were supported by the data.

In another study, Palacios-Huerta [46] collected and examined a different dataset, which consisted of 1417 penalty kicks from various countries (mostly Italy, England, and Spain). He performed most of the analysis on a simplified 2X2 game (right vs. left without center) and found that the winning probabilities of each strategy of each player are similar and that players' choices are serially independent, demonstrating that professional soccer players and goalkeepers behave as predicted by the Minimax Theorem when deciding on the directions of shots and jumps in penalty duels. In a more recent study, Azar and Bar-Eli [4], using a different dataset of penalty kicks, compared the predictions of the MSNE to other prediction methods and found that the MSNE predictions were the closest to the data, even though some other prediction methods used information on the marginal distribution of kicks or jumps whereas the MSNE did not.

The desired characteristics of simplicity, constant situation, and large incentives, and the fact that penalty kicks in soccer are sometimes taken as a series of shootouts after a tied game (in certain competitions), have made penalty kicks a source for addressing additional research questions. For example, Bar-Eli et al. [8] analyzed 286 penalty kicks in top leagues and championships worldwide and found that given the probability distribution of kick direction, the optimal strategy for goalkeepers is to stay in the goal's center. Goalkeepers, however, seem to behave non-optimally and almost always jump to the right or left. The authors suggest how this can be explained by the norm theory [32]. Because the goalkeepers' norm is to act (jumping to a side), a goal scored yields worse feelings for the goalkeeper following inaction (staying in the center) than following action (jump-

ing to a side), leading to a bias for action. The more common omission bias, a bias in favor of inaction, is reversed in this context, since the norm is reversed—to act rather than to choose inaction. The claim that jumping is the norm was supported by a survey conducted with 32 top professional goalkeepers.

In another study on penalty kicks in soccer, Apesteguia and Palacios-Huerta [3] revealed a surprising phenomenon that takes place during penalty shootouts. The purpose of a penalty shootout is to decide the winning team where competition rules require one team to be declared the winner after a drawn game. The referee tosses a coin and the team whose captain wins the toss decides whether to take the first or the second kick. Five kicks are taken alternately by the teams. The explicit randomization mechanism used to determine which team goes first in the sequence, in a situation where both teams have exactly the same opportunities to perform a task, suggests that we should expect the first and second teams to have exactly the same probability of winning the shootout. Yet, using data on 1343 penalty kicks from 129 penalty shootouts over the period 1976–2003, Apesteguia and Palacios-Huerta found that teams that take the first kick in the sequence win the penalty shootout 60.5% of the time. Given the characteristics of the setting, the researchers attributed this difference in performance to psychological effects (e.g., pressure) resulting from the kick consequence. In particular, most kicks are scored, and this puts the team that kicks second behind in its score most of the time. Such a finding provides a fertile ground for the intervention and help of a sport psychologist, and is valuable for influencing the decisions of coaches and teams' captains. Interestingly, another study [33] that increased the sample size from 129 shootouts to 540 suggests a much smaller first-mover advantage (53.3% for the first-kicking team to win), which is no longer statistically significant. This demonstrates the importance of relying on as much relevant data as possible, which reinforces the advantages we have today in the big-data era.

In another study on soccer, Misirlisoy and Haggard [44] examined all 361 kicks from the 37 penalty shootouts performed in the World Cup and Euro Cup games over the period 1976–2012, and suggested that goalkeepers displayed a sequential bias: following repeated kicks to the same direction, goalkeepers became more likely to jump to the opposite direction in the next kick. This is an illustration of the phenomenon of the gambler's fallacy, a famous psychological bias. Surprisingly, kickers did not seem to exploit these goalkeeper biases.

However, a new observation on the same issue by other researchers yielded different findings: Braun and Schmidt [13] suggested that even with the original data, but using a different statistical analysis (a binomial test instead of bootstrapping), the results of Misirlisoy and Haggard are no longer statistically significant, even at the 10%-level with a one-tailed test (equivalent to 20% in a two-tailed test). The results of Misirlisoy and Haggard turn out to be very sensitive, due to the small number of cases on which their findings are based. Although the entire dataset consists of a reasonable number of kicks (361), their main finding emerges from only 16 kicks that represent a situation where the three previous kicks of that team went to the same direction (right or left). This shows the challenge that still exists even in the big-data era: if the event being considered is special and relatively rare, then even a large dataset may provide only a few relevant events, making the analysis problematic. Moreover, the sensitivity of the results turns out to be susceptible not only to changing the statistical test but also to expanding the sample. In particular, Braun and Schmidt found that adding some additional competitions, which increased the number of relevant sequences (three previous kicks to the same direction) from 16 to 26, reduced the percentage of cases in which the goalkeeper jumps to the opposite direction (compared to the last three kicks) from 69% to only 58%; the difference between 58% and 50% is not statistically significant with only 26 observations.

Large datasets have come to be especially useful in detecting systematic biases in the decision making of referees. For example, using field data from the Spanish soccer league, Garicano et al. [27] examined the amount of extra time a referee adds after 90 min (regular time in a soccer game) and provided clear evidence that referees add significantly more extra time in the case where the home team is behind. On a related topic, Dohmen [23] analyzed the neutrality of referees during 12 German premier league soccer seasons, and documented evidence that social forces influence agents' preferences and decisions. Dohmen reported that referees tend to favor the home team in decisions to award goals and penalty kicks. The insights generated by these studies have been incorporated in referees' training programs.

In one study on basketball, Morgulev et al. [45] combined an examination of referees' decisions with the analysis of the related behavior of players and its impact on the team. More specifically, the researchers examined the behavior of professional referees and players in the context of offensive fouls in basketball. Over 500 incidents that had the potential to meet the criteria of an offensive foul were recorded and analyzed by basketball experts. Falling intentionally to improve the chances to get an offensive foul was found to be a very common behavior of defenders (almost two-thirds of the recorded falls). At first, it seems helpful, increasing the chances to be given an offensive foul. However, an additional statistical analysis suggests that the overall impact of an intentional fall on the team seems to be negative (e.g., because a fallen player is less helpful for the team than a standing player). The authors argue that both rational reasons and biased decision making lead players to act against their

team's interest by falling. In addition, the authors reported that referees almost never call an offensive foul if the player remains on his feet, and generally call fewer fouls than the number judged by experts as appropriate. They explain the referees' behavior as being partially biased by the representativeness heuristic, but also partially reflecting officiating mistakes that are rational given the referees' incentives.

An additional central concept in economics that was examined by data collection from the sport arena is the Prospect Theory, and the effect of loss aversion derived from this theory. Pope and Schweitzer [47] turned to the field of golf in order to put this psychological mechanism to a test. In golf, every hole has a par number of strokes associated with it, and the par number provides a reference point for a satisfying performance. For a professional golfer, a birdie (one stroke under par) is a gain, and a bogey (one stroke over par) is a loss. The researchers compared a situation where the player is putting to avoid a bogey, with a more favorable setting where the player is aiming to achieve a birdie. A hypothesis dictated by loss aversion suggests that players will try harder when putting for a par (to avoid a bogey) than when putting for a birdie. Their analysis of more than 2.5 million putts supported this prediction. Nobel Laureate Daniel Kahneman referred to this finding in his 2011 book: *These fierce competitors certainly do not make a conscious decision to slack off on birdie putts, but their intense aversion to a bogey apparently contributes to extra concentration on the task at hand* [31, p. 304].

In a study on professional basketball players, Staw and Hoang [53] used the player market in the NBA to elicit field-originated data on the existence of the well-known sunk-cost effect. These researchers tested whether the amount teams spent on players influenced how much playing time the players got and how long they stayed with the NBA franchises. This study was one of the first quantitative field tests of the sunk-cost effect.

Another area in which big-data analysis provided interesting insights is corruption in sports. Duggan and Levitt [24] analyzed the results of all "critical" sumo matches from January 1989 until January 2000, and found strong evidence for match rigging in professional sumo. This study was replicated and extended by Dietl et al. [22], who discovered more intriguing trends, relying on even bigger datasets. In basketball, Taylor and Trogdon [54] analyzed the winning percentages of NBA teams that were eliminated from the playoffs. They found that to gain higher draft positions, these teams were 2.5 times more likely to lose than teams that were still trying to secure their place in the playoffs. Recently, Elaad, Kantor, and Krumer (2016)[1] began to utilize data from crucial soccer games between a team in immediate danger of being relegated to a lower division and a team not much affected by the results in the respective game, on the last day of a season. Based on data from 75 countries between 2001–2013, they found that the odds of the team in danger to avoid relegation are significantly higher when the country is more corrupt according to the Corruption Perceptions Index (CPI).

6 Challenges of big data in sports

Silver [51] summarizes his book with the realization that prediction in the era of big data is not going very well. The author admits that his success in building functioning forecasting systems for baseball and politics is in large part due to his ability to choose his "battles" well. We mentioned previously that baseball is an exceptional domain with a defined set of distinct actions that can be counted, and have actually been counted for more than a century. In contrast, the game of soccer is far more susceptible to chance, with success rates of pre-game favorites only slightly above 50% [2].

Daryl Morey gained his reputation as the first truly analytical general manager in the NBA. His staff gathered original data by measuring items that had previously gone unmeasured. In 2008, Morey used his predictive model to select a center player that was soon revealed to be a bust, and failed to detect DeAndre Jordan, future dominant NBA center and the second-best player in the entire draft class. Digging deeper into the matter, Morey revealed that his model disregarded the prospects' age. He realized that an entire class of college players existed who played better due to the fact that they were much older than the players they were playing against. The failure to detect DeAndre Jordan proved to be far more complex: He had played a single year of college basketball, hated his coach, and did not even want to be in school—it is impossible to see this prospect's future in his college statistics. The analytical model would always miss DeAndre Jordan; however, one of Morey's scouts had wanted to draft Jordan on the strength of what appeared to him Jordan's undeniable physical talent [36].

Due to this reason, Alamar [1] emphasized the importance of integrating different data sources, both quantitative and qualitative, into the data management system. For example, if the quantitative data give a different picture from the scouting reports, integrated medical information about the player's medical history may explain the differences. If not, then integrated and linked videos lets the decision maker see the player in action, and he can then conclude which source of information is the most relevant. The installation of such comprehensive and integrated information systems requires significant financial investments in technology, alongside the cooperation and change of behavior of the different departments in the organization.

[1] Elaad, G., Kantor, J., & Krumer, A. (2016) Corruption and Contests: Cross-Country Evidence from Sensitive Soccer Matches, mimeo.

Assembling and organizing all of the quantitative and qualitative data is a monumental task, and strong and determined leadership is essential in order to move from a culture of data silos to a centralized system. The whole organization needs to understand the importance of the new data management system. Confident leadership is crucial when the data-driven approach does not lead to immediate success and it becomes open to attack in a way that the old approach to decision making was not [36].

Some of the data in sports are quantitative (e.g., points gained by the teams during the game, league scores, and other objective performance criteria), but much else of what can be analyzed is more complex as it comes from situations during the games. Except for a few situations that are generally the same each time (e.g., a penalty kick in soccer [4,8,44,46] or a basketball free throw), most other situations are complex and involve many variables such as the identity of the involved players and their abilities, how much times remains until the game end, previous fouls of the players in basketball or yellow cards in soccer, the current score, the identity of the home team, the number of spectators, etc. In addition, some variables are hard to quantify in a manner that can be used in regressions or a similar statistical analysis, for example the location of multiple players and the ball, the location of the referee, etc. Moreover, the importance of the game may be different for the teams based on the situation in the league or the championship. Often data collection requires to obtain multiple sources. When the game situation is important for the analysis, a common approach is to collect video clips of game events and categorize them based on some relevant characteristics (e.g., whether the player with the ball in soccer is inside the 16-meter area of the opponent, or whether a collision in basketball may be an offensive foul [45]), sometimes using expert judges for this purpose [45]. Because game events can differ in so many ways, one often has to make compromises and put together in the analysis events that have some common features although they differ in other aspects. For example, we may categorize soccer situations based on the location of the player who holds the ball, ignoring the location of other players, which is different across the events.

Another important challenge in sport analysis is that it may be hard to conclude about causality even when a correlation is found, due to endogeneity problems. That is, we are usually considering in sports complex systems where multiple players make decisions that affect each other and are influenced by various factors, making it hard to conclude what is the reason for what. For example, finding that the chances to make the shot in basketball are higher when the player is blocked by an opponent player, probably does not mean that being blocked improves accuracy. Whether the opponent team blocks a player is endogenous, i.e., it is determined within the system (the game dynamics). It makes sense that the opponent team blocks a player in situations that are more dangerous, and this is why we find a correlation between being blocked and having better chances of making the shot. To infer about causality one needs to find ways to neutralize the endogeneity problem, which is often quite difficult.

7 Conclusion

The big-data era, where large datasets are available for analysis in many domains of life, provides various unique opportunities for research (see [15] for a review). In this article, we focus on the case of analyzing sports data. We provide some examples, from various sports, such as the world's highly popular team sports of soccer and basketball, the individual sports of golf and tennis, and the more unique and local sport of sumo, for studies that used sports data to address a variety of research questions. In many cases, although the data are collected from the domain of sport, the lessons learned are more general and have implications for other fields. We hope that this short literature review may lead some readers to develop their own ideas on how to use sports data to address interesting research topics.

Compliance with ethical standards

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

1. Alamar, B.C.: Sports Analytics—A Guide for Coaches, Managers, and Other Decision Makers. Columbia University Press, West Sussex (2013)
2. Anderson, C., Sally, D.: The Numbers Game. Why Everything You Know About Soccer is Wrong (2013)
3. Apesteguia, J., Palacios-Huerta, I.: Psychological pressure in competitive environments: evidence from a randomized natural experiment. Am. Econ. Rev. **100**, 2548–2564 (2010)
4. Azar, O.H., Bar-Eli, M.: Do soccer players play the mixed-strategy Nash equilibrium? Appl. Econ. **43**, 3591–3601 (2011)
5. Baade, R.A., Matheson, V.A.: Going for the gold: the economics of the Olympics. J. Econ. Perspect. **30**, 201–218 (2016)
6. Bagić Babac, M., Podobnik, V.: A sentiment analysis of who participates, how and why, at social media sport websites: how differently men and women write about football. Online Inf. Rev. **40**, 814–833 (2016)
7. Bar-Eli, M., Avugos, S., Raab, M.: Twenty years of "hot hand" research: review and critique. Psychol. Sport Exerc. **7**, 525–553 (2006)
8. Bar-Eli, M., Azar, O.H., Ritov, I., Keidar-Levin, Y., Schein, G.: Action bias among elite soccer goalkeepers: the case of penalty kicks. J. Econ. Psychol. **28**, 606–621 (2007)
9. Bhandari, I., Colet, E., Parker, J., Pines, Z., Pratap, R., Ramanujam, K.: Advanced scout: data mining and knowledge discovery in NBA data. Data Min. Knowl. Discov. **1**, 121–125 (1997)

10. Billings, S.B., Holladay, J.S.: Should cities go for the gold? The long-term impacts of hosting the Olympics. Econ. Inq. **50**, 754–772 (2012)
11. Bocskocsky, A., Ezekowitz, J., Stein, C.: The hot hand: a new approach to an old "fallacy". In: Proceedings of the 8th MIT Sloan Sport Analytics Conference (2014)
12. Borghesi, R.: Allocation of scarce resources: insight from the NFL salary cap. J. Econ. Bus. **60**, 536–550 (2008)
13. Braun, S., Schmidt, U.: The gambler's fallacy in penalty shootouts. Curr. Biol. **25**, R597–R598 (2015)
14. Buraimo, B., Simmons, R.: Do sports fans really value uncertainty of outcome? Evidence from the English Premier League. Int. J. Sport Finance **3**, 146 (2008)
15. Cao, L.: Data science: a comprehensive overview. ACM Comput. Surv. (CSUR) **50**, 43 (2017)
16. Chiappori, P.A., Levitt, S., Groseclose, T.: Testing mixed-strategy equilibria when players are heterogeneous: the case of penalty kicks in soccer. Am. Econ. Rev. **92**, 1138–1151 (2002)
17. Constantinou, A.C., Fenton, N.E., Neil, M.: Profiting from an inefficient Association Football gambling market: prediction, risk and uncertainty using Bayesian networks. Knowl. Based Syst. **50**, 60–86 (2013)
18. Constantinou, A., Fenton, N.O.R.M.A.N.: Towards smart-data: improving predictive accuracy in long-term football team performance. Knowl. Based Syst. **124**, 93–104 (2017)
19. Costa, G.B., Huber, M.R., Saccoman, J.T.: Understanding Sabermetrics: An Introduction to the Science of Baseball Statistics. McFarland (2007)
20. Cox, A.: Spectator demand, uncertainty of results, and public interest: evidence from the English Premier League. J. Sports Econ. 1527002515619655 (2015)
21. Davenport, T.H.: Analytics in sports: the new science of winning. Int. Inst. Anal. **2**, 1–28 (2014)
22. Dietl, H.M., Lang, M., Werner, S.: Corruption in professional sumo: an update on the study of Duggan and Levitt. J. Sports Econ. **11**, 383–396 (2010)
23. Dohmen, T.J.: The influence of social forces: evidence from the behavior of football referees. Econ. Inq. **46**, 411–424 (2008)
24. Duggan, M., Levitt, S.D.: Winning isn't everything: corruption in sumo wrestling. Am. Econ. Rev. **92**, 1594–1605 (2002)
25. Franks, I., Hughes, M.: Notational Analysis of Sport: Systems for Better Coaching and Performance in Sport. Routledge, London (2004)
26. Fried, G., Mumcu, C. (eds.): Sport Analytics: A Data-Driven Approach to Sport Business and Management. Taylor & Francis, New York (2016)
27. Garicano, L., Palacios-Huerta, I., Prendergast, C.: Favoritism under social pressure. Rev. Econ. Stat. **87**, 208–216 (2005)
28. Gilovich, T., Vallone, R., Tversky, A.: The hot hand in basketball: on the misperception of random sequences. Cogn. Psychol. **17**, 295–314 (1985)
29. Goldsberry, K.: CourtVision: New visual and spatial analytics for the NBA MIT Sloan Sports Analytics Conference. In: MIT Sloan Sports Analytics Conference (2012)
30. Hughes, M., Hughes, M.T., Behan, H.: The evolution of computerised notational analysis through the example of racket sports. Int. J. Sports Sci. Eng. **1**, 3–28 (2007)
31. Kahneman, D.: Thinking, Fast and Slow. Macmillan, London (2011)
32. Kahneman, D., Miller, D.T.: Norm theory: comparing reality to its alternatives. Psychol. Rev. **93**, 136–153 (1986)
33. Kocher, M.G., Lenz, M.V., Sutter, M.: Psychological pressure in competitive environments: new evidence from randomized natural experiments. Manag. Sci. **58**, 1585–1591 (2012)
34. Kreps, D.M.: Game Theory and Economic Modelling. Oxford University Press, Oxford (1990)
35. Lewis, M.: Moneyball: The Art of Winning an Unfair Game. WW Norton & Company, New York (2004)
36. Lewis, M.: The Undoing Project: A Friendship That Changed the World. Penguin, London (2016)
37. Liu, H., Morstatter, F., Tang, J., Zafarani, R.: The good, the bad, and the ugly: uncovering novel research opportunities in social media mining. Int. J. Data Sci. Anal. **1**, 137–143 (2016)
38. Lopez, M.J., Matthews, G.J.: Building an NCAA men's basketball predictive model and quantifying its success. J. Quant. Anal. Sports **11**, 5–12 (2015)
39. Martin, L.: Sports Performance Measurement and Analytics. Pearson, Old Tappan (2016)
40. Miller, T.W.: Sports Analytics and Data Science. Pearson, Old Tappan (2016)
41. Martins, R.G., Martins, A.S., Neves, L.A., Lima, L.V., Flores, E.L., do Nascimento, M.Z.: Exploring polynomial classifier to predict match results in football championships. Expert Syst. Appl. **83**, 79–93 (2017)
42. Memmert, D., Hüttermann, S., Hagemann, N., Loffing, F., Strauss, B.: Dueling in the penalty box: evidence-based recommendations on how shooters and goalkeepers can win penalty shootouts in soccer. Int. Rev. Sport Exerc. Psychol. **6**, 209–229 (2013)
43. Mills, B., Fort, R.: League-level attendance and outcome uncertainty in US pro sports leagues. Econ. Inq. **52**, 205–218 (2014)
44. Misirlisoy, E., Haggard, P.: Asymmetric predictability and cognitive competition in football penalty shootouts. Curr. Biol. **24**, 1918–1922 (2014)
45. Morgulev, E., Azar, O.H., Lidor, R., Sabag, E., Bar-Eli, M.: Deception and decision making in professional basketball: is it beneficial to flop? J. Econ. Behav. Organ. **102**, 108–118 (2014)
46. Palacios-Huerta, I.: Professionals play minimax. Rev. Econ. Stud. **70**, 395–415 (2003)
47. Pope, D.G., Schweitzer, M.E.: Is Tiger Woods loss averse? Persistent bias in the face of experience, competition, and high stakes. Am. Econ. Rev. **101**, 129–157 (2011)
48. Reep, C., Bernard, B.: Skill and chance in association football. J. R. Stat. Soc. Ser. A (Gen.) **131**, 581–585 (1968)
49. Reich, B.J., Hodges, J.S., Carlin, B.P., Reich, A.M.: A spatial analysis of basketball shot chart data. Am. Stat. **60**, 3–12 (2006)
50. Rottenberg, S.: The baseball players' labor market. J. Polit. Econ. **64**, 242–258 (1956)
51. Silver, N.: The Signal and the Noise: Why so Many Predictions Fail-but Some Don't. Penguin, London (2012)
52. Shapiro, S.L., Drayer, J.: An examination of dynamic ticket pricing and secondary market price determinants in Major League Baseball. Sport Manag. Rev. **17**, 145–159 (2014)
53. Staw, B.M., Hoang, H.: Sunk costs in the NBA: why draft order affects playing time and survival in professional basketball. Adm. Sci. Q. **40**, 474–494 (1995)
54. Taylor, B.A., Trogdon, J.G.: Losing to win: tournament incentives in the National Basketball Association. J. Labor Econ. **20**, 23–41 (2002)
55. Tversky, A., Kahneman, D.: Judgment under uncertainty: heuristics and biases. Science **185**, 1124–1131 (1974)
56. Ulmer, B., Fernandez, M., Peterson, M.: Predicting Soccer Match Results in the English Premier League. Doctoral dissertation, Ph. D. dissertation, Stanford (2013)
57. Walker, M., Wooders, J.: Minimax play at Wimbledon. Am. Econ. Rev. **91**, 1521–1538 (2001)