Sports Knowledge Management and Data Mining

Robert P. Schumaker[1], Osama K. Solieman[2] and Hsinchun Chen[3]

[1]Information Systems Dept, Iona College, New Rochelle, New York 10801, USA
rschumaker@iona.edu

[2]6015 N. Mardelle Circle, Tucson, Arizona 85704, USA
osolieman@gmail.com

[3]Artificial Intelligence Lab, Department of Management Information Systems
The University of Arizona, Tucson, Arizona 85721, USA
hchen@eller.arizona.edu

Word Count: 17,721

Introduction

Vast amounts of sports data are routinely collected about players, coaching decisions and game events. Making sense of this data is important to those seeking an edge. By transforming this data into actionable knowledge, scouts, managers and coaches can have a better idea of what to expect from opponents and be able to use a player draft more effectively. With millions of dollars riding on the many decisions made within a sports franchise (Lewis, 2003), the sports environment is ideal for data mining and knowledge management approaches. While the application these approaches to the sports environment may be unique and the focus of this chapter, the topics of data mining and knowledge management should certainly be well known to the reader and form the basis of the approaches we discuss.

Background and Motivation

Before the advent of data mining and knowledge management techniques, sports organizations relied almost exclusively on human expertise. It was believed that these domain experts (coaches, managers and scouts) could effectively convert their collected data into usable knowledge. As the different types of data collected grew in scope, these organizations sought to find more practical methods to make sense of what they had. This led first to the employment of in-house statisticians who created better measures of performance and better decision-making criteria. One way that these measures were used was to augment the decision-making of domain experts with additional knowledge and provide them with a competitive advantage. Armed with this knowledge, it was not a far step for sporting organizations and fans alike to begin harnessing more practical methods of extracting knowledge using data mining techniques. These newer techniques allowed organizations to begin to predict particular player matchups and/or forecast

how a player may perform under specific conditions. Sports organizations were sitting on a wealth of data and needed ways to harness it.

The primary knowledge management and data mining techniques that can be used by sports organizations include statistical analysis, pattern discovery and outcome prediction. A variety of non-typical sports data can be similarly monitored including injury likelihood. One such example is a biomedical tool piloted by AC Milan, an Italian professional soccer club, which uses software to monitor workouts that helps to predict player injuries (Flinders, 2002). Another example is software used to monitor sports betting locales for unusual bets which may signal corrupt officiating or players that are compromised (Audi & Thompson, 2007). Similarly, data mining researchers have found that physical aptitude correlates to anticipated physical performance (Fieltz & Scott, 2003). Every year the National Football League (NFL) conducts a "Combine" where prospective college draft players are run through a series of physical drills in front of team scouts and coaches. The Combine also includes a mental evaluation of players called the Wonderlic Personnel Test, which assesses the intellectual capacity of prospects. The NFL has developed expected Wonderlic scores based on amount of intelligence required to play a particular position; e.g., a quarterback who has to make a myriad of on-field decisions should have a higher Wonderlic score (24), than a halfback (16) whose job is to run the ball (Zimmerman, 1985).

Sport statistics, by themselves can be misleading without an understanding of their fundamental meaning. This comes from either imprecise measurement of an event or the sports community's misuse and over reliance on particular statistics. As evidence, consider the fact that certain players can build impressive individual statistics yet have little impact on the performance of the team. The impreciseness of sports statistics can be best illustrated by

baseball's Runs Batted In (RBI) statistic which has been long heralded as a cornerstone of evaluating player contribution. Developed by British-born journalist Henry Chadwick during the mid-1800s, the RBI was an attempt to quantify game events and attribute them to particular players (Lewis, 2003). While Chadwick was more familiar with cricket than baseball and had an incomplete understanding of the game, he managed to popularize his statistics which were never seriously questioned until the latter half of the 20th century. The RBI's imprecise measurement can be summed up in the following thought experiment. Suppose two players had the same batting average, meaning that they hit the ball with the same percentage of success. Further suppose that both players are not power hitters but routinely hit for singles, advancing themselves and their teammates one base at a time. The RBI is then dependent upon the actions of those who that batted before them. If team members were able to routinely get on base for one of these players and not for the other, then the first of our hypothetical players would be credited with RBIs when their teammates crossed home plate as a consequence of the player's hits. The second of our hypothetical players would not receive any RBIs, even though both players performed the exact same actions. Basing a player's value on RBI statistics alone would be a misleading indicator of performance. Besides impreciseness in measuring player productivity, the sports community has overvalued the RBI as a measurement of performance in contract negotiations and player comparisons. It wasn't until pioneering baseball statistician Bill James began questioning the RBI, that better measurements arose such as the On Base Percentage (OBP) which measures how often a player gets on-base.

Another difficulty with the use of sports statistics is how to measure risk. In American football, a defensive back can either stay in mid-field and attempt to intercept the ball or play solid cover defense. In the first instance, the player is taking a risk which can quickly change the

momentum of the game whereas in the second instance, the player is playing it safe. However, by being successful at taking risks and making interceptions, there is a greater perceived player value. Quantifying risk taking behavior is a difficult problem.

Another example of statistical imprecision is the measurement the number of defensive rebounds off missed free-throws in Basketball. In order to get a defensive rebound, teammates must block out opposing players and in doing so, they typically cannot get the rebound although their actions arguably make them just as important in the accomplishment (Ballard, 2006). However, given the way in which rebounds are measured; only the player who gets the ball is credited with the rebound.

In this chapter, we propose a Sports Knowledge Management framework to categorize the different methods sports organizations use to uncover new knowledge and better value player contributions. From this, we will highlight measurement inadequacies and showcase techniques to make better usage of data collected in a wide domain of sport and sport-related specialties. Properly leveraging Sports Knowledge Management techniques can result in better team performance by matching players to certain situations, identifying individual player contributions, evaluating the tendencies of the opposition and exploiting any weaknesses.

For these reasons, there should be no surprise that many sports organizations are revolutionizing themselves. The traditional decision-making approach of using intuition or gut instincts is falling out of favor. Instead, assessments are being made on the basis of strong analysis and scientific exploration. With more and more sports organizations embracing the digital era, it may soon become be a battle of the better algorithm or measurement used, where back-office analysts may become just as important as the players on the field.

Significance of this Survey

The knowledge management revolution in organized sports began with the book *Moneyball* which was a case study of the successes enjoyed by the Oakland Athletics, a professional baseball team (Lewis, 2003). The Oakland Athletics, commonly known as the A's, had long been at the low end of major league baseball's payroll with salaries much lower than the league average. This made it difficult for the A's to acquire talented players from other teams and impossible for them to retain any of their good players. Rather than accepting their situation team management adopted a radical approach. By carefully selecting players in the 2002 draft, the A's could lock players that were oftentimes overlooked by other clubs, into long contracts that didn't pay much money and thus develop a strategy to compete with larger payroll teams. When the players become good and their contracts were about to expire, the A's would then have the option of trading or selling them to larger market teams and getting a return on their investment. The trick was to pick the "right" players.

Up until that time, the player draft was seen as a type of crap shoot because teams never really knew what they were going to get. Teams generally did not spend too much time on the draft and left the bulk of the work up to scouting departments, whose scouts would travel the country to view new talent and make recommendations. Billy Beane, the general manager of the Oakland A's, questioned this old approach and began to use a systematic method of statistically analyzing draft picks by the numbers they generated throughout their careers. It was reasoned that if the A's were careful in the selection process, that they could get a few productive years out of the players before the higher salary teams would take them.

From this strategy, the Oakland A's began to field a competitive team that bucked the trend of all the other low salary ballclubs. Relying instead on computers and algorithms to pick talent, the A's produced such star-studded players such as Barry Zito, Mark Mulder, Tim Hudson, Jason

Giambi, Miguel Tejada, Eric Chavez, Nick Swisher and Mark Teahan to name a few. The first step in their selection process is to eliminate all high school players from consideration. This is a significant departure from the old way of doing things, where high school players were seen as valuable commodities. However, stars at the high school level rarely panned out and making comparisons between high school players and leagues at that level was difficult. Instead, Beane focused on the statistics generated for college players, adopting different evaluative metrics. For example, the RBI does not measure the ability of a player to get on-base so they tested different formulas and found that on-base and slugging percentages were the most influential indicators of run production. Beane and his colleagues would then use these metrics to rank order the draft players. Most of the players who were rated highly using these methods were overlooked by other clubs. The other organizations would even tease Oakland for picking players that they saw as worthless, especially so high in the draft.

The results soon became clear when Oakland fielded competitive teams year after year in spite of its low payroll. Competitors and commentators did not understand how Oakland was able to win consistently. Even Major League Baseball's Blue Ribbon Panel of economic experts that was investigating the salary inequities in baseball, concluded that Oakland's performance was a statistical anomaly (Levin et al., 2000). At that time, knowledge management techniques were not widely understood in sports.

Baseball was not the only sport that was undergoing transformation. During the 1980s and 1990s, Dean Oliver was applying statistical analysis techniques to basketball; many years before the *Moneyball* revolution. A contemporary of baseball's Bill James, Oliver focused more on creating statistics that would showcase team behavior rather than individual performance, and began to publish his thoughts for the rest of community (Oliver, 2005). Oliver and James were

eventually both hired as statistical consultants to the Seattle Supersonics and Boston Red Sox respectively, cementing the inclusion of data analysts in the new foundation of sports competitiveness.

Sports organizations are big business. Advancing to playoffs and winning championships can tap into lucrative television revenue and vast marketing opportunities. The key is winning. With so many competitive forces lining up against a professional sports organization, such as larger salaried teams, salary caps and revenue sharing schemes, it becomes of paramount importance that the right decisions are made to maintain a competitive advantage. These decisions come from the hard facts and data already acquired. It is just a matter of finding ways to discover knowledge trapped within the data.

Chapter Scope and Methods

This chapter investigates a number of sports knowledge management techniques and related research about data mining methods, with a special emphasis on those sports with the most interesting applications of technology. However, novel and insightful techniques from lesser known sports and sports outside of the US are also explored and included in this survey.

While the coverage provided in broad in scope covering a multitude of sports organizations, sports research centers, academia and private industry within the United States; readers are encouraged to follow the broad subject matter themes from prior ARIST publications on data mining and knowledge management topics.

Chapter Structure

This chapter is arranged as follows. Section 2 provides an analytic framework for Knowledge Management and positions the domain of Sports Data Mining within it. Section 3 examines the data sources that fans and organizations can access. Section 4 examines the use of

statistical analyses as methods of knowledge extraction. Section 5 provides an overview of the systems and tools that are used to gather both data and knowledge. Section 6 details the predictive aspects of sports knowledge systems. Section 7 inspects the emerging trend of multimedia and video analysis as methods of obtaining a competitive advantage. Finally, section 8 delivers our conclusions and a brief discourse on future research directions.

Analytic Framework for Knowledge Management and Data Mining in Sports

Knowledge Management first appeared in academia in 1975 as a way to encompass a range of tools, technologies and human expertise (Davenport & Prusak, 1998) that can give an organization a competitive advantage (Lahti & Beyerlein, 2000) and a method for maintaining the continuity of knowledge in the organization (Serenko & Bontis, 2004). By retaining and sharing knowledge within the organization, businesses are discovering increased productivity and innovation (O'Reilly & Knight, 2007). However, before getting to the stage of useable knowledge we must examine the intermediate levels of data and knowledge that are represented by the Data-Information-Knowledge-Wisdom (DIKW) hierarchy (Ackoff, 1989). The DIKW hierarchy is a widely accepted concept in knowledge management groups as a way to represent the different levels of what people can see and what they can know (Cleveland, 1982; Zeleny, 1987). Each successive level; data, information, knowledge and wisdom, builds upon prior levels and provides an increased awareness of surroundings (Carlisle, 2006) where meaning can be found within this DIKW continuum (Chen, 2001; Chen, 2006).

Data are the observable differences in physical states (Boisot & Canals, 2004) that are acquired from stimuli and examination of the world around us. By themselves, data are generally overwhelming and not entirely useable. In the framework we are developing here, data can be thought of as all of the individual events that occurred in the sporting event. If applied to

baseball, this data would contain a record of pitch sequences, at-bat events and defensive moves which by themselves provide little interest or value.

In order to be of practical value, data must be transformed by identifying relationships (Barlas et al., 2005) or limited only to that which is relevant to the problem at-hand (Carlisle, 2006). This transformation results in information, or meaningful, useful data (Bierly et al., 2000). Using our baseball example again, information could be focused on only the pitch sequences by a particular pitcher. Although not too useful at this stage, abstracting it to the next level of the hierarchy, knowledge, can provide us additional meaning by identifying patterns within the data.

Knowledge is the aggregation of related information (Barlas et al., 2005), that forms a set of expectations or rules (Boisot & Canals, 2004) and provides a clearer understanding of the aggregated information (Bierly et al., 2000). At this level of the hierarchy rule-based systems are developed which can allow individuals to expand their own knowledge while also benefiting the organization (Alavi & Leidner, 2001). Returning to our baseball example, analysts can evaluate the pitching information and look for tendencies or expectations as to the types of pitches to be encountered. Data mining is the hunt for knowledge within the data.

While the precise definitions of data, information and knowledge are still a matter of debate; wisdom can be viewed as a grasp of the overall situation (Barlas et al., 2005), that uses knowledge and knowledge alone (Carlisle, 2006) to achieve goals (Bierly et al., 2000; Hastie et al., 2001). In our baseball example, we have 1. knowledge of the types of pitches to be encountered, 2. knowledge of effective strategies to combat specific types of pitches and 3. knowledge that a successful at-bat can help to win a game. Putting all of this disparate knowledge together into wisdom; the batter has a chance to positively influence the game in their

favor. Uncovering this truth rests in the capabilities of cognition and human understanding (Carlisle, 2006), as a computational wisdombase is currently difficult to imagine (Barlas et al., 2005).

Data Mining involves procedures for uncovering hidden trends and developing new data and information from data sources. These sources can include well-structured and defined databases, such as statistical compilations, or unstructured data in the form of multimedia sources.

The DIKW framework then sets the stage for disambiguating data from knowledge and sets definitional boundaries for what data, information and knowledge are. Applying this to the sports domain, certain activities and techniques serve at the data level (i.e., data collection, data mining and basic statistics). Other techniques and algorithms are more suited to the knowledge end of the spectrum, such as strategies and simulations. Throughout this chapter, the DIKW framework can be used to identify the set of relevant tools that can be used depending whether data or knowledge is desired.

Sports Knowledge Management Framework

Sports data can come from a myriad of structured and unstructured sources. The process of transforming these data into useful and interesting sports knowledge can be categorized by the techniques used; expert examination, statistics and machine learning techniques, as shown in Figure 1.

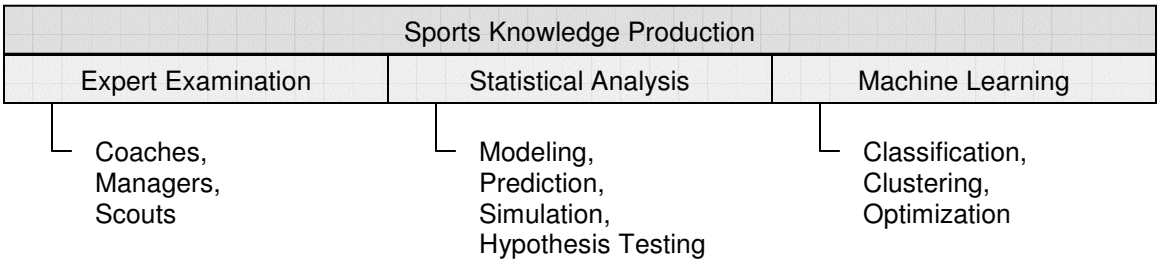

Figure 1. A Sports Knowledge Framework

In expert examination, human domain experts, (i.e., coaches, general managers and scouts) make decisions based upon their experience and the data presented to them. Sometimes these decisions can be fraught with gut reaction and instinct, that run counter to the available data (Lewis, 2003; Page, 2005). The use of sports experts as the sole repository of knowledge has been declining with the advent of computational knowledge acquisition techniques in sports. These systems can give an organization an analytical edge that few domain experts can compete against.

Statistical techniques are often used in sports knowledge discovery. While statistics in sports have been around for a long time, there has been a recent overhaul in the way performance is measured. Newer and more sophisticated algorithms are being used to find interesting patterns in player tendencies and team strengths/weaknesses (Dong & Calvo, 2007). Sub-areas such as prediction, simulation and hypothesis testing can be used as an augmentation tools (Hirotsu & Wright, 2003) by players, coaches and general managers to make better decisions. Statistical techniques lay at the heart of data mining, distinguishing between something interesting and random noise and allowing researchers to test hypotheses and make predictions (Piatetsky-Shapiro, 2008).

The third area of sports knowledge gathering is machine learning. Machine learning techniques differ from statistics, by allowing an algorithm to learn patterns from the data and apply that knowledge in real-time to previously unseen data. Leveraging pattern-matching algorithms can uncover many hidden trends that domain experts and statisticians never thought to pursue. Sub-areas such as classification, clustering and optimizations allow analysts to maximize their teams effectiveness by conducting a series of what-if analyses on the data by changing one or more variables (Berry, 2005; Chen & Chau, 2004).

While Expert Examination can suffer from biases and unquantifiable "gut instincts," statistical and machine learning methods also have their form of weaknesses in generalizing its results to future activity. Each study must take into account these limitations.

Taken together, these knowledge discovery methods offer a powerful tool to sports organizations. For more details about machine learning algorithms, readers are referred to Chen and Chau (2004).

Summary of Applications in Sports Data Mining

It was only until the past few years that sports knowledge was believed to reside only in the minds of domain experts such as scouts, coaches and managers. These experts were the sole group responsible for translating the gathered data into actionable knowledge. However, with problems of data overload from traditional and newer multimedia sources, these experts quickly became overwhelmed, leading to the hiring of technologically sophisticated analysts to make sense of their data. Their focus was on discovering better methods of performance measurement. They soon created many new formulas such as baseball's On Base Plus Slugging (OPS) (Thorn & Palmer, 1984) and basketball's Player Efficiency Rating (PER) (Hollinger, 2002) to name two. Similarly, progress was also made in the area of event prediction through various tools such as neural networks.

Scouting has been the backbone of sports organizations' knowledge collection for nearly a century. Scouts serve two primary roles, the first to seek out and evaluate new talent and second to prepare assessments of opposing teams.

To seek out new pools of talent, scouts will often travel to the locations of potential draft picks and evaluate their skills during practice and regular games. The reports generated usually focus on the strengths and weaknesses of the potential draft pick as well as the overall

impression of the draftee within the organization. These reports are important because they affect a player's draft position and indicate the organization's expectation for that player's success (Page, 2005).

The second type of scout, an advance scout, observes competing teams and compiles reports on player weaknesses, opposing teams' strategies and other useful tidbits that may lead to a competitive advantage.

Traditional scouting involves the collection of hard data and expert opinions about the potential for draftees and opponents' strategies and performances alike. However, these opinions could oftentimes form biases where a scout may "fall in love" with a certain player's skills or overlook others which can lead to questionable recommendations (Lewis, 2003).

Following the recent Moneyball revolution, scouting has witnessed two fundamental changes. One of the first changes was to adopt a more scientific and statistically-based strategy to compare players against one another in an unbiased manner. Using data mining tools on the data already gathered, players and opponents could be evaluated without the usual scouting biases. From there, scouting moved away from simply identifying the strengths and weaknesses and into a more in-depth study of situations and tendencies (White, 2006). The second major change was the advent of more automated and fine-grained data gathering and analysis, including multimedia and video analysis techniques.

Data Sources for Sports

Data on sport performance can come from a variety of sources. The most typical method is in-house statisticians. Statistics are generally kept for team-level and individual player performances. However, most organizations keep such information to themselves which has

opened the door to professional societies and application-specific companies to fill the gap of data sources for sports.

Professional Societies

There are a number of professional societies dedicated to exploring new facets of knowledge within their particular sport. They serve as centralized repositories where members can share insights and explore further research. Many of these societies collect, evaluate, store and disseminate sport-related data for members and maintain periodical newsletters and journals. Their main activities involve discovering and sharing knowledge within the sporting community.

The Society for American Baseball Research (SABR) was formed in Baseball's Hall of Fame Library in 1971 (Society for American Baseball Research, 2008). Its purpose is to foster research about baseball and create a repository of baseball knowledge not captured in the box scores. In 1974, SABR founded the Statistical Analysis Committee (SAC) with the goal of carefully studying both the historical and modern game of baseball from an analytical point of view. Its research became known as Sabermetrics and the SAC Committee publishes its research on a quarterly basis (Birnbaum, 2008).

The Professional Football Researchers Associations (PFRA) was started in 1979 with the goal of preserving and reconstructing historical game day events (Professional Football Researchers Association, 2008). The PFRA also publishes articles on a bi-monthly basis which cover statistical analyses as well as new methods of performance measurement.

The Association for Professional Basketball Research (APBR) was formed in 1997 with the objective of promoting the history of professional basketball (Solieman, 2006). While their research concentrates on NBA-related statistics, they also include rival basketball leagues, many of which are now defunct (The Association for Professional Basketball Research, 2008). Similar

to Baseball's Sabermetrics, the APBR has developed the APBRmetrics which are used to create better measurements and statistical yardsticks for comparison purposes.

The International Association on Computer Science in Sport (IACSS) was founded in 1997 to improve the cooperation amongst researchers interested in applying techniques and technologies from the field of Computer Science to sport-related challenges (International Association on Computer Science in Sport, 2008). The IACSS focuses on disseminating the research of their members through periodic newsletters, journals and organized conferences.

The International Association for Sports Information (IASI) was founded in 1960 with the goal of standardizing and archiving the world's sports libraries (International Association for Sports Information, 2008). The IASI is a worldwide network of sport experts, librarians and document repositories. The Association's information dissemination comes in the form of a tri-annual newsletter and an organized World Congress every four years.

Special Interest Sources

In addition to professional sport-related societies, there are other organizations that collect and analyze sport-related statistics. Oftentimes these sources offer traditional statistics as well as augmented data in the form of player biographies, records and awards. Examples of these sources include Baseball-Reference.com which portrays itself as a one-stop shop for all basic statistics, current standings, player and team rankings by various categories, draft picks, and historical box score data (Baseball-Reference.com, 2008). Pro-Football-Reference.com compiles player, team and league stats along with historical game data (Pro-Football-Reference.com, 2008) and 82games.com positions itself as Basketball's innovative data source for fans, coaches and the media (82games.com, 2008).

Statistical Analyses Research in Sports

Once the data have been gathered, the next steps involve a process of finding the knowledge locked within. Statistical analyses of many different types can be applied to the data from statistically intense sports such as baseball and basketball to less data-intense sports such as Curling. Other types of analyses can be used to measure player performance, team balance, opposition weaknesses and even the possibility of a debilitating injury.

Statistical Analysis

While a myriad of statistics have been kept as records of sports events, the statistics themselves were not called into question for nearly a century. The basic question became, *are we measuring what we think we are measuring*. Early pioneers of statistical analysis such as Bill James and Dean Oliver, not only asked these questions, but began to offer new statistics and insights.

History and Inherent Problems of Statistics in Sports

The origins of early baseball statistics are often traced to Henry Chadwick, the 19th century sportswriter and statistician (Lewis, 2003). Chadwick created many of today's familiar statistics, e.g., batting average and earned run average, based on his experience with the game of cricket. This is one of the reasons why walks (i.e., advancing to a base without a hit) are not included in these formulae, because the walk had no equivalency in cricket.

Batting Average, defined as the number of hits a player collects divided by the number of times at-bat is one such example of a statistic that ignores walks. If a player manages to draw a walk during a time at bat, then the at-bat is not counted. This leads to imprecision when rating players, because if the goal is to get on-base, hits and walks should be both counted. Players who walk often may have lower batting averages therefore using Batting Average as a sole

measurement of performance will lead to an unfair comparison and may underestimate a player's contribution to team performance.

Similarly, the Earned Run Average (ERA) is another cornerstone of baseball's performance metrics. The ERA is the number of earned runs against a pitcher per nine innings. The term "earned run" is important because it is a run that is achieved through a hit. Other means of getting on base and scoring, such as getting hit by a pitch (when the batter is awarded first base after being hit by the ball during an at bat), a balk (an illegal motion by the pitcher which results in base runners being awarded the next base), a dropped third strike (normally a batter strikes out after a third strike but can attempt to run to first base if the catcher drops the pitch), fielding errors and walks, do not count towards the ERA. Again the over-emphasis on hitting tends to skew ERA values. These two statistics alone, Batting Average and ERA were used as the primary performance indicators by scouts, coaches and general managers for well over a century.

American football also has some imprecision in its measurements such as the number of receptions and yards per carry to name two. Defining the number of receptions as the number of times a player catches a forward pass, is misleading. Receptions do not indicate success in terms of touchdowns, but may instead indicate a preference for a particular player by a quarterback and thus inflate the reception total of the preferred receiver. Yards per carry is another example, where success is not predicated on scoring points. Should one player who ran for 40 yards in one play be valued more than another that runs an average of 3 yards per play? While one obvious solution would be to only compare those players with a minimum number of carries, the process of setting arbitrary thresholds ignores the issue that yards per carry does not take into account the points scored, and thus the statistic leads to inexact comparisons.

Basketball uses similarly imprecise statistics such as field goal percentage and rebounds. The field goal percentage is the number of field goals made divided by the number attempted. A player who scored a high number of points yet who has a low field goal percentage might be rated as unsuccessful. Likewise the rebound statistic, or the number of times a player gets the ball after a missed shot attempt, does not imply that points will be scored. Nevertheless basketball experts have long valued these statistics as adequate measures of performance.

The problem with these traditional formulae lies in what the statistic is intended to measure. Oftentimes, data are gathered and used in ways that cannot be meaningfully interpreted. The data itself is not at fault, it is the methods that are used for comparing player performances. This also leads us to the realization that there are some problems that cannot be answered through statistical examination alone. The questioning of statistics that were held as truths, the very foundations of modern sports, brought about new techniques and measures which have rapidly become commonplace within modern sports organizations.

Bill James

The fundamental shift from traditional statistics into knowledge management can be credited to Bill James. In 1977, James published the first of many "Bill James Baseball Abstracts" in which he began to openly question traditional statistics and offer his unique insight about remedying the problems he was encountering. While only selling 50 copies at the outset, James was not deterred and continued to publish his annual compendium of insights, new statistical measures, which he called *sabermetrics*, and strange ranking formulae. Readers of the Bill James Baseball Abstracts became interested in the new way of computing performance and began to make their own contributions. Soon sporting enthusiasts and fantasy baseball team owners began applying this newfound knowledge with overwhelming success. Even with the

tidal wave of fan excitement for this revolution in thinking, sports organizations were quite resistant to these new ideas for several decades because scouting was so entrenched within organization as the sole vessel of knowledge (Lewis, 2003).

In 2002, Oakland A's General Manager Billy Beane became the first Bill James disciple in Major League Baseball to adopt sabermetrics when selecting draft picks. Beane's use of data mining and knowledge extraction tools landed the A's in either the playoffs or playoff contention for five straight years (Lewis, 2003).

That same year, the Boston Red Sox hired another Bill James disciple, Theo Epstein. Epstein, a Yale graduate, similarly appreciated the hard facts that could be gleaned from reams of data. He hired Bill James as a consultant in 2003 and went on to engineer the Red Sox World Championships in 2004 and 2007.

Dean Oliver

Dean Oliver is to basketball what Bill James is to baseball. Asking some of the same types of questions throughout the 1990s, Oliver sought to better quantify player contribution and began popularizing *APBRmetrics*, basketball's answer to sabermetrics. Oliver focused a lot of attention on the proper usage of the possession statistic, where possession is defined as the period of time one team has the ball. Part of Oliver's contribution was to evaluate team performance on how many points they scored or allowed opponents to score per 100 possessions. In 2004, Dean Oliver was hired as a consultant to the Seattle Supersonics, ushering basketball into the Moneyball era. Seattle then went on to win the Division title in 2005.

Also in 2005, the Houston Rockets hired Daryl Morey as assistant general manager. Morey, an MIT graduate and believer in knowledge management principles, had previously worked with

the Boston Celtics and STATS Inc where he invented and refined several new Basketball statistics (MIT Sloan Alumni Profile, 2008).

Baseball Research

Baseball has been called the "National Pastime" and has been a part of the American culture for nearly two centuries. The first professional baseball team, the Cincinnati Red Stockings, was founded in 1869 and played amateur teams across the country, amassing an impressive 81 game winning streak (Voigt, 1969). Baseball was further organized into sustainable leagues in 1876 with the creation of the National League. The National League, which quickly became the premier baseball league, withstood competition from startup leagues including the Player's League, the American Association and the Federal League; Wrigley Field in Chicago was originally built as a Federal League ballpark (Fetter, 2003). However, in 1901 a startup league called the American League was not so easy to vanquish. Both leagues competed intensely with one another and poached talent to the detriment of the game, until 1903 when both sides agreed to recognize each other as major league and also began the tradition of World Series competitions between the two rival leagues (The New York Times Staff, 2004).

Building Blocks

Statistics by themselves should not be primary means by which player performance is determined, but rather the beginning of a process during which useful knowledge can be discovered. For instance, one of baseball's fundamental statistics has been hits. As discussed earlier, this statistic does not account for other means that a player can get on-base and these additional means do not count in either the player's batting average or hits total. Because of this, On-Base Percentage (OBP) was developed to better measure the player's ability to get on base by including these different methods. Furthermore, another statistic that can better measure

offensive player productivity is slugging percentage. With slugging percentage, the number of bases reached is divided by the number of at-bats, and rewards players who hit doubles and triples instead of singles, or hit home runs. By contrast the hits statistic treats doubles, triples and homeruns as equivalent to singles.

Building upon both of the fundamental statistics of OBP and Slugging percentage, we can derive the On-Base Plus Slugging (OPS) statistic which is the summation of these two statistics and provides a better representation of a player's ability to get on base and hit with power. OPS is considered to be one of the most effective measures of a player's offensive capabilities.

Runs Created

In his third *Baseball Abstract*, James reasoned that players' performances should be measured based upon what they are trying to accomplish, scoring runs, rather than baseball's predominant indicator of the day, batting average (James, 1979). James recognized the disconnect between the two concepts and questioned how run production could be better measured. From this, James developed the Runs Created (RC) formula which was ((Hits + Walks) * Total Bases) / (At Bats + Walks) (James, 1982). The Runs Created formula reflected a team's ability to get on-base as a proportion of its opportunities through at bats and walks. James then evaluated historical baseball data using his model and found that Runs Created was a better model at predicting the number of runs that a major league team would accomplish than other predictors (Lewis, 2003). This formula was found to be a better measure of a player's offensive contribution than batting average, because wins are decided on the team with the highest number of runs, not the highest batting average.

Further instantiations of Runs Created led to Runs Created Above Average (RCAA) (Sinins, 2007) which compares Runs Created to the league average (Woolner, 2006) and Runs Created

per 27 Outs (RC/27) (James & Henzler, 2002) which takes into account sacrifice flies, where the player hits the ball to the outfield with the expectation that the ball will be caught by the opposition in order to advance one of the base-runners to the next base, and sacrifice hits, where the batter hits the ball to the infield with the expectation that he will be thrown out at first base in order to advance one of the base-runners to the next base. RC/27 is simply RC / (at-bats – hits + number of times caught stealing + number of times a player hits into a double play + sacrifice flies + sacrifice hits). The RC/27 is a comparison to model complete offensive player performance over the course of an entire game (27 outs). From further analysis, it was found that bench players (i.e., players that do not start but come into the game later) will typically have 80% of the offensive capability of the starter, with the exception of catchers at 85% and first basemen at 75% the starter's ability (Woolner, 2006).

Win Shares

In 2001's *Baseball Abstract*, Bill James introduced the concept of Win Shares, where players are assigned a portion of the win based upon their offensive and defensive input and further explained it in a follow-up book of the same name (James & Henzler, 2002). Win Shares is a complicated formula that takes into account many constants and educated guesses, primarily because some of the measures were never captured in the historical data. While still a matter of debate within the sabermetric community, Win Shares attempts to assign players credit for winning a game based upon their performance. Assuming a team has equal offensive and defensive capabilities, defense is credited with 52% of a win whereas offense is only 48%. This seemingly arbitrary division is justified as a way to even out the public's perception that offense is the more important component of a win. While the formula itself is still being refined in the crucible of the sabermetric community, its results are difficult to argue with. Players with

seasonal Win Shares of around 20 are typically all-stars, Win Shares of 30 indicates an MVP season and Win Shares of 40+ point to a historic season. For example, Barry Bonds had a Win Share of 54 in 2001 when he set the record of 73 homeruns in a single season.

Linear Weights and Total Player Rating

The Linear Weights formula calculates runs based upon the actions of the offensive player. Using the formula of 0.47(1B) + 0.78(2B) + 1.09(3B) + 1.40(HR) + 0.33(BB + HBP) + 0.30(SB) – 0.60(CS) – 0.25(AB – H) – 0.5(Outs on Base), George Lindsey used this as an alternative to simple batting average (Albert, 1997). Recognizing that there were three ways to get on base, hits (1B, 2B, 3B and HR), walks (BB) and being hit by a pitch (HBP), Lindsey further extended his model to reward those players that advanced through base stealing (SB), punish players that were caught stealing (CS) and punish those that were called out on the basepaths (Outs on Base).

Pushing the idea of linear weights further, Total Player Rating (TPR) is a little more complicated and builds into itself comparisons for the position played and the ballpark (Schell, 1999). These comparisons allow statisticians to compare the performance of players as above or below average based on the player's defensive position and the ballpark they are playing, because some ballparks may be more difficult for a position player than others. TPR has ratings of Batting Runs, Pitching Runs and Fielding Runs which are 1) summed, 2) adjusted for player position and ballpark and 3) divided by 10 such that players can be compared against league averages. However, this statistic is also undergoing scrutiny where an average player is assumed to have a TPR of zero and Bill James claims it should be substantially positive (James & Henzler, 2002).

Pitching Measures

So far we have analyzed offensive measures meant to capture the true value of a player's offensive performance. Pitching is another important staple of baseball and the performance of pitchers are closely watched by fans and sports organizations alike. Earned Run Average (ERA) which measures pitching performance over nine innings against the number of earned runs, runs that come from hits, is one of the most relied upon pitching statistic. This statistic is usually coupled with a Win/Loss record and can be deceptive. Take for instance a poorly performing pitcher that plays on a team with a high powered offense. The pitcher will have a high ERA but also a deceptively high Win/Loss record. Similarly, an excellent pitcher playing on a team without run support will have a good ERA, but a poor Win/Loss record. In order to adjust to these situations, the Pitching Runs statistic was developed to more directly compare pitchers to league performance. In Pitching Runs, the number of Innings pitched is divided by nine innings then multiplied by the league's ERA and then the earned runs allowed is subtracted out. The result of this formula gives the anticipated number of runs a pitcher would allow over the course of a complete game. Average pitchers would have a Pitching Runs score of zero while the Pitching Runs for poorly performing pitchers would be negative.

Another pitching measure recently put forth by Bill James is the Component ERA (ERC) which breaks out the different components of pitching outcomes and figures them in with the ERA. ERC is (((H + BB + HBP) * PTB) / (BFP * IP)) * 9 – 0.56 where BFP is number of batters faced by the pitcher, IP is number of innings pitched and PTB is 0.89(1.255(H – HR) + 4 * HR) + 0.56(BB + HBP – IBB) where IBB is intentional walks (Baseball Info Solutions, 2003). However, the ERC goes into more complicated formulation under certain conditions and other organizations offer differing models.

Football Research

Advances in statistical techniques in football have not reached the levels of those collected in both baseball and basketball. For this reason, there is a lack of statistical data on individual players. While some basic statistics are obviously collected such as number of touchdowns, receptions and interceptions, these aggregate counts do not rise to the level of their sabermetric counterparts. The other reason for the lack of data comes from the number of games played. The NFL plays 16 regular-season games compared to baseball's 162 and basketball's 82 games. Despite this, there are several metrics meant to bridge these deficiencies.

Defense-Adjusted Value Over Average

The Defense-Adjusted Value Over Average (DVOA) is a comparative measure of success for a particular play (Schatz, 2006). This statistic treats each play as a new event and measures the potential of success versus the average success of the league. Certain variables are taken into account such as time remaining, the down, distance to the next down, field position, score and quality of opponent. These variables carry different rewards if met and can be used to measure a particular player's contribution or aggregated to highlight team-based performance. A DVOA of 0% indicates that the defense is performing on par with the league average. Whereas positive and negative DVOA values indicate that the defense is performing above or below league averages respectively.

In football, possession is broken into four downs (or plays) with a sub-goal of exceeding a set number of yards before the expiration of downs. DVOA considers that in order to meet this sub-goal, 45% of the required yards should be gained on the first down, 60% on the second down and 100% by the third or fourth down (Carroll et al., 1998). If the play is deemed to be successful DVOA assigns it one point. If the play is successful early (e.g., attaining the sub-goal in the early downs), more points are awarded.

Defense-Adjusted Points Above Replacement

Defense-Adjusted Points Above Replacement (DPAR) is a player-based statistic that is compiled over the course of a season (Schatz, 2006). DPAR is used to determine the point-based contribution of a player as compared to the performance of a replacement player. If a player is said to have a +2.7 DPAR, it means that the team should score 2.7 points because of the player's presence in the lineup, whereas that 2.7 points would be lost if the player was substituted by a typical replacement.

Adjusted Line Yards

Adjusted Line Yards (ALY) is a statistic to assign credit or responsibility to an offensive line in relation to how far the ball is carried (Schatz, 2006). This statistic attempts to separate the running back from the contribution of the offensive line and is measured per league averages. If a running back is brought down behind the line of scrimmage (i.e., takes a loss of yards), the offensive line will be penalized heavily for the failure. If the same running back manages to break free and make a long gain (i.e., picked up many more yards than usual), the offensive line is given minimal credit, because the offensive line can only make so much of a contribution and much of it is up to the running back. The ALY is also adjusted to league averages.

Basketball Research

Basketball experienced its own sabermetric revolution shortly after the book *Moneyball* began to be circulated among baseball enthusiasts (Pelton, 2005). With their own wealth and depth of statistics, several pioneers of basketball statistics set about to better quantify and assign credit through the creation of ABPRmetrics, named for the Association of Professional Basketball Researchers (ABPR). ABPRmetrics is fundamentally different from sabermetrics, in that ABPRmetrics attempts to view statistics in terms of team rather than individual performance. One such example of this is team possession and how effective the team is at

scoring points. The thinking is that since teams must function as cohesive units, they should be analyzed as such because quantifying team chemistry (how well players perform with one another) at the individual level is currently unattainable.

To back up this point, players can have either a positive or negative impact on team performance. As an example, during the 2004-2005 season, Stephon Marbury had a -0.4 point negative impact on the team while he was on the court. At the outset, this statistic would indicate that Marbury was performing at or slightly below average. However, when Marbury was off the court and not helping his team, the team had a -12.0 point deficit. This 11.6 point differential when Marbury was on the court versus off the court, illustrates that Marbury can best improve his team's performance when he is on the court.

Shot Zones

The basketball court can be divided up into 16 areas where a player on offense might be inclined to shoot a basket. By analyzing the percentage of player success from each of these zones, defensive adjustments can be made to limit scoring while offensively, coaches may try to maximize these types of shots (Beech, 2008). Figure 2 illustrates the different shot zone locations.

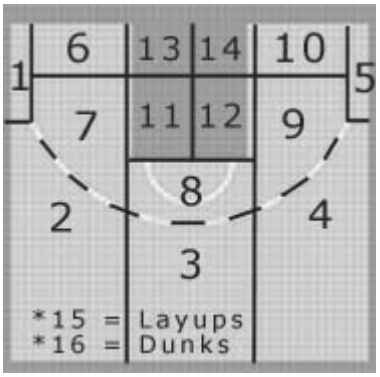

Figure 2: Shot Zone Layout (82games.com)

From this analysis on the 2004-2005 season, 82games.com found that for 3 point shots from the corner (Zones 1 and 5), Golden State's Dunleavy had the highest accuracy of 0.571 from the left corner and Sacramento's Mobley had 0.600 accuracy from the right corner. Likewise, shot zones can portray player tendencies. For example, under the basket (Zones 13 and 14), Miami's Shaquille O'Neal made the most attempts in the league but was not very successful, 0.416 from the left and 0.424 from the right. Knowing where players are successful and comfortable can lead to better strategies.

Player Efficiency Rating

Player Efficiency Rating (PER) is a per-minute rating a player effectiveness that rewards positive contribution and punishes negative ones (Hollinger, 2002). This formula takes on many variables including assists, blocked shots, fouls, free throws, made shots, missed shots, rebounds, steals and turnovers among others and tries to quantify player performance in regards to their pace throughout the game and the average performance level of the league. However, PER is still a matter of debate as Hollinger admits that it does not take into account all of performance related criteria, such as hustle and desire (Hollinger, 2002).

Plus / Minus Rating

Another method of calculating performance is through the Plus / Minus Rating system in which each player is evaluated by calculating the number of points the team makes with that player on the field minus the number of points the opposing team receives. This calculation is done for each team player while they are on court and while they are on the bench. Player contribution can then be measured as the differential between their on and off court presence (Rosenbaum, 2004). Positive values indicate the player is making a positive point-based contribution to the team whereas negative values would point towards detrimental activity. Take

for example Dwight Howard during the 2004-2005 season. Howard had a Plus / Minus rating of -2 when he is on the court versus an even rating when he is not (Rosenbaum, 2005). This would seem to indicate that the team is better off without Howard's presence. However, the Plus / Minus Rating system is not without its own share of quirks. Critics of the system point to its over-valuing of players that take a high number of shots and commit a large number of turnovers, which is not a beneficial team activity.

Measuring Player Contribution to Winning

A further metric to evaluate player contribution versus a substitute player is to adjust the Plus / Minus Rating system to account for the talent level of teammates (Rosenbaum, 2004). The reasoning is that player performance does not occur within a vacuum, but rather is a function of the overall team effort. The Adjusted Plus / Minus is a regression estimate where the constant is the home court advantage against all teams, the k^{th} order constants are the Plus / Minus differences between Player K and the players of interest, holding all others constant. The x values, x_1 through x_{14} refer to game level statistics per 40 minutes of play: points, field goal attempts at home, field goal attempts on the road, three point attempts, free throw attempts, assists, offensive rebounds, defensive rebounds, turnovers, steals, blocks, personal fouls, (points * assists * rebounds)$^{1/3}$ and minutes per game. Regressing these 14 values together nets the Adjusted Plus / Minus Rating.

Rating Clutch Performances

The reason that 40 minutes is typically studied rather than the standard game time of 48 minutes, is the belief that the final minutes of the game are completely different from the rest of the game (Ilardi, 2007). In instances where one team is ahead by several points, the lagging team may institute fouling in order to retrieve possession of the ball. This behavior tends to skew

these statistics from normal game behavior. The final 8 minutes of game time and any overtime if necessary, providing that the scores of the two teams are within 5 points of one another, is referred to as the clutch. Some players tend to excel during this period, leading APBRmetricans to study the clutch performances of players. Some would argue that player contribution during the clutch is more important than during the rest of regulation play, because the prospect of winning or losing is hanging in balance. Others point towards legends of the game that have defined themselves with clutch performances (e.g., Bill Walton and Michael Jordan). Using PER during this period can identify new insights into offensive capability. To test the defensive match-up in man to man coverage, it assumed that one player's PER (i.e., looking at PER's of opposing positions), will be superior to that of their counterpart if a clutch performance is occurring. While PER is limited to man to man coverage and cannot be used in other types of defenses such as zone (where the defensive player is confined to a specific area of space), this method can still provide valuable insight into player execution.

Another strategy for measuring clutch performance is to evaluate the performance of the team as a whole by using the Plus / Minus rating system and aggregating clutch points amongst the entire squad on the court during the last eight minutes of the game. This provides additional insight into a player's clutch abilities by showing both on-court and off-court results in terms of point contribution.

Emerging Research in Other Fields

Aside from baseball, football and basketball, many other sports are experiencing their own statistical renaissance. Soccer is pioneering work in predicting the likelihood of injury based on biomedical monitoring as well as isolating the features that lead to tournament wins. NCAA College Basketball researchers are predicting tournament matchups and victories with impressive

accuracy. Two other sports, Olympic Curling and Cricket, are similarly gathering data on their opponents and analyzing the factors that contribute to winning. It is not a far stretch to adapt any these techniques to other sports.

Soccer

Soccer arguably garners the most passionate fans worldwide. With such devotion to the sport, it is understandable that many researchers and fans alike have an interest in predicting prestigious tournament outcomes. While one such study found that time of possession is an important factor in game outcomes (Papahristoulou, 2006), other studies have noted that country of origin and home field advantage were sizable factors in predicting team success (Barros & Leach, 2006). From this later study of the teams comprising the UEFA Tournament, researchers used a myriad of factors including league win/loss records, tournament win/loss, shots, team record at home and on the road and past tournament performance to predict not only who the strongest teams will be, but also to forecast which team should win the tournament.

Another important contribution is the ability to forecast when a player may be experiencing the onset of an athletic impairment through injury prediction. Oftentimes a player, regardless of their sport, will try to play despite their injury or performance degradation. AC Milan has been piloting predictive software that monitors the workouts of their players (Flinders, 2002). This software compares an athlete's workout performance against that of a baseline, and drops in performance may indicate that the player may be disposed to experiencing an injury soon. Other biomedical methods employ a series of weighted variables including injury rate, odds of injury and history of injury to compile a risk likelihood measure (Hopkins et al., 2007). Another method looks at 17 various risk factors, such as previous injuries, playing characteristics,

endurance and game-time preparation among others, and it was found that inadequate warm-ups were the usual factor in injury-related events (Dvorak et al., 2000).

NCAA Basketball

NCAA basketball has its own share of research. One notable figure is Jeff Sagarin who publishes his basketball rating system based on a team's win/loss record and the strength of their schedule (USA Today, 2008). However, more research exists concerning the NCAA Men's Basketball Tournament. Every March, college basketball enters into March Madness – a tournament where 64 Division I teams will compete for the title of National Champion. While the exact selection process for the 64 teams is not made public, a Selection Committee makes the determinations and the 64 teams are selected on a "Dance Card." Two researchers that were interested in this process, developed a method of predicting the at-large bids with a 93.3% success rate over the past 14 years (Coleman & Lynch, 2008a). This would seem to indicate that the Selection Committee uses similar selection techniques every year, even though the membership of the committee changes from year to year (SAS, 2005). The technique weights 42 pieces of information on each team, including their RPI ranking (or relative strength against other teams), win/loss record, conference win/loss record, etc. and forms a rank order score called the "Dance card score" (Coleman & Lynch, 2001).

Once the teams have been selected, this same team of researchers has devised a second algorithm, "Score Card," to predict the winners (Coleman & Lynch, 2008b). Using data from the 2007 tournament, their system was able to correctly predict the winners for 51 of the 64 games, an accuracy of 79.7%. The Score Card algorithm is remarkably simpler than its counterpart Dance Card, because only 4 variables are necessary; the team's RPI value, RPI value

of the team against non-conference opponents, whether the team won the conference title and the number of wins in their previous 10 games.

NCAA Football

NCAA Football also uses data mining and knowledge management techniques to rank collegiate teams. Because NCAA football does not enter into a tournament style of play like basketball does, disputes routinely break out regarding which two teams should compete for the National Championship. The Bowl Championship Series or BCS, was created to address these problems, however, it became part of the controversy in 2004 when the University of Southern California (USC) was rated number one by the Associated Press poll and number three by the BCS. Following 2004, the BCS algorithm was rewritten.

The BCS is a fairness type algorithm in which many various polls are taken into account and weighted accordingly. In particular, the BCS uses the Harris Interactive College Football poll, the Coaches poll (what rankings fellow football coaches believe is fair) and computer polls including Jeff Sagarin's NCAA football poll at USA Today and the Seattle Times. Each team is then assigned points based upon their poll ranking in all of the component polls. Teams are then rank ordered based on their score.

Olympic Curling

The Curling event in the 1998 Winter Olympics would appear to be a non-typical place to find data mining and knowledge management tools at work. During the eight days of Curling competition at the Nagano Olympics, IBM was collecting plenty of data on players, strategy, the precise paths the stones took as well as outcomes (Taggart, 1998). While this data collection was not extensively used at the time, the potential still exists to isolate a Curling player's tendencies and weaknesses (Cox & Stasko, 2002).

Cricket

Similar to the wealth of statistics kept in baseball, the game of Cricket also holds an extensive store of data within the Wisden Almanack, going back to 1864 (CricInfo, 2008). This data has also been recently explored using data mining and knowledge management tools to some success. In a study of One Day Test Cricket matches, it was found that a mix of left/right batsmen and a high runs to overs ratio were both highly correlated to winning (Allsopp & Clarke, 2004). The usage of alternating left and right-handed batsmen is believed to keep the opposing team's bowler out of their typical rhythm and thus be less effective (Allsopp & Clarke, 2004). The high number of runs to overs ratio, (e.g., amount of runs scored as a proportion to the number of offensive periods) indicted that a quicker paced game (i.e., more runs) was also a factor in determining a winning team. These factors can further be used to determine team effectiveness and also tournament play.

Tools and Systems for Sports Data Mining and Knowledge Management

While still in its infancy, the proliferation of systematic data mining and knowledge management tools has been mainly constrained to in-house analyses by sports organizations. However, simpler tools using the theories of Bill James and his contemporaries have been used by fantasy team managers and rotisserie leagues before the advent of the *Moneyball* revolution. These individuals found success in using data mining and knowledge management tools leading to further development of measurement techniques and knowledge-based tools. There are growing trends of third-party vendors using data mining and knowledge management tools to sell their niche services to individuals and sporting organizations to isolate player tendencies, provide more in-depth scouting reports and uncover fraudulent activity within the sports arena.

Data Mining and Knowledge Management Tools

One area of this third-party development has been in designing tools that do not fit the traditional data mining mold. Incorporating elements of game footage that can be broken down into component pieces and queried is one of the unique ways that companies such as Virtual Gold is filling the gap. Other distinctive methods include simple graphical analysis of existing statistics, allowing domain experts to more readily identify the patterns within the data. Information visualization has long been recognized as an effective tool for knowledge management (Zhu & Chen, 2005).

Advanced Scout

Advanced Scout was developed by IBM during the mid 1990s as a data mining and knowledge management computer program. Its purpose is to glean hidden patterns within NBA game data and provide additional insights to coaches and other organization officials. Advanced Scout not only collects the structured game-based statistics during play, but also unstructured multimedia footage. With the entire NBA league having access to Advanced Scout, coaches and players can use this tool to prepare for upcoming opponents and study their own game-level performance (Shulman, 1996).

The multimedia aspect of Advanced Scout functions by collecting raw game-time footage, processing and error-checking the content and finally segmenting it into a series of time-stamped events such as shots, rebounds, steals, etc (Bhandari et al., 1997). The processing and error-checking stage is a rule-based series of processes to verify the consistency and accuracy of the data. This includes removing impossible events (events tagged incorrectly), looking for missing events and attributing plays to particular players. In cases where the rule-based strategy is unable to identify key elements, a domain-expert can use game footage to manually label the event.

Advanced Scout also possesses a knowledge management component called Attribute Focusing, where a particular attribute can be evaluated over the entire distribution of data and both textual and graphical descriptions of the anomalous subsets (i.e., those with a distinctly different statistical distribution) are set aside for further analysis by players or coaches (Bhandari, 1995). For example, consider the following textual description from Advanced Scout:

> *When Price was Point-Guard, J. Williams missed 0% (0) of his jump field-goal-attempts and made 100% (4) of his jump field-goal-attempts. The total number of such field-goal attempts was 4. This is a different pattern than the norm which shows that: Cavaliers players missed 50.70% of their total field-goal-attempts. Cavaliers players scored 49.30% of their total field-goal-attempts (Bhandari et al., 1997).*

This description illustrates an easy to read analysis of the anomalous behavior of Williams when Mark price was the Cavaliers point guard. Once a coach or player receives this information, it is up to them to determine why this is case. For the above example, it was determined that when Price was double-teamed, he would pass the ball to Williams for wide-open jump-shots.

Aside from the anomaly detection facet of Attribute Focusing, Advanced Scout can also be queried to find a relevant game-time event such as particular shots, rebounds, etc. Players and coaches alike can use this information to hone skills and better understand player dynamics.

Visualization Tools

Another way of finding interesting data is to do so graphically. SportsVis is one such tool that allows users to view a plethora of data over a selected period of time (Cox & Stasko, 2002).

This data could include team runs over an entire season or player-specific criteria such as the runs scored off professional baseball pitcher Curt Schilling over a 32 game period, as shown in Figure 3.

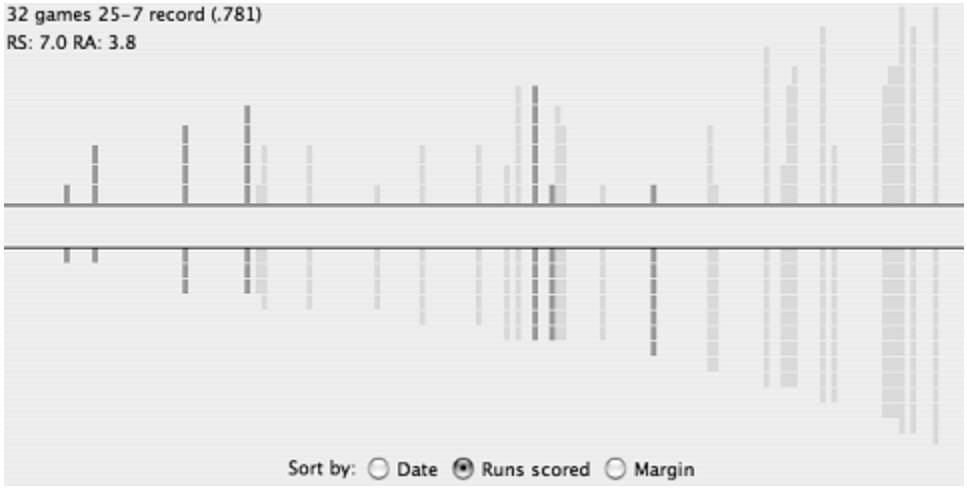

Figure 3. Curt Schilling runs scored over 32 games (Cox & Stasko, 2002)

This pictorial description may indicate trends or uncover potential problems such as injuries. Other interesting visualization techniques can be found in *Baseball Hacks*, where author Joseph Adler walks users through the process of using Excel and Access databases to view various baseball statistics (Adler, 2006). These techniques include batter spray diagrams where a hitter may favor hitting the ball to certain portions of the field under certain situations, and frequency distributions using many of the sabermetric statistics.

Scouting Tools

Scouts used to rely on manual methods to keep track of player performance. Today that power is being placed into the hands of fans and next generation scouts. Game statistics can be input on the fly and complete game reports and individual attributes can be focused on for later player improvement.

Digital Scout

Digital Scout is the digital answer to collecting statistics or filling out score cards. Fans and sports organizations alike can use this software on a palmtop, laptop or desktop machine to collect and analyze game statistics. This software can be adapted for all the major sports including volleyball. Digital Scout can also allow users to print box score results or create custom reports on particular attributes, such as baseball hit charts, basketball shots and football formation strengths (Digital Scout, 2008).

This software has been found to be very useful and has been adopted by Baseball's Team USA (Petro, 2001), Little League Baseball (Petro, 2003), and basketball tournaments (Weeks, 2006).

Inside Edge

Another scouting tool is Inside Edge which was created by Randy Istre and Jay Donchetz in 1984 and provides pitch charting and hitting zone statistics for college and professional baseball teams (Inside Edge, 2008a). Coupled with a professional scouting department, Inside Edge has been used by many MLB ballclubs including all of the World Series champions between 1996 and 2001. The strength of Inside Edge is in easy to read scouting reports that employ a host of textual and graphical elements as well as the expected opponent strengths, weaknesses and tendencies.

Reports on strengths, weaknesses and tendencies are all backed by statistical data. An example spray chart of Rafael Furcal of the Atlanta Braves is shown in Figure 4. Note the density of infield hits shown for the second baseman.

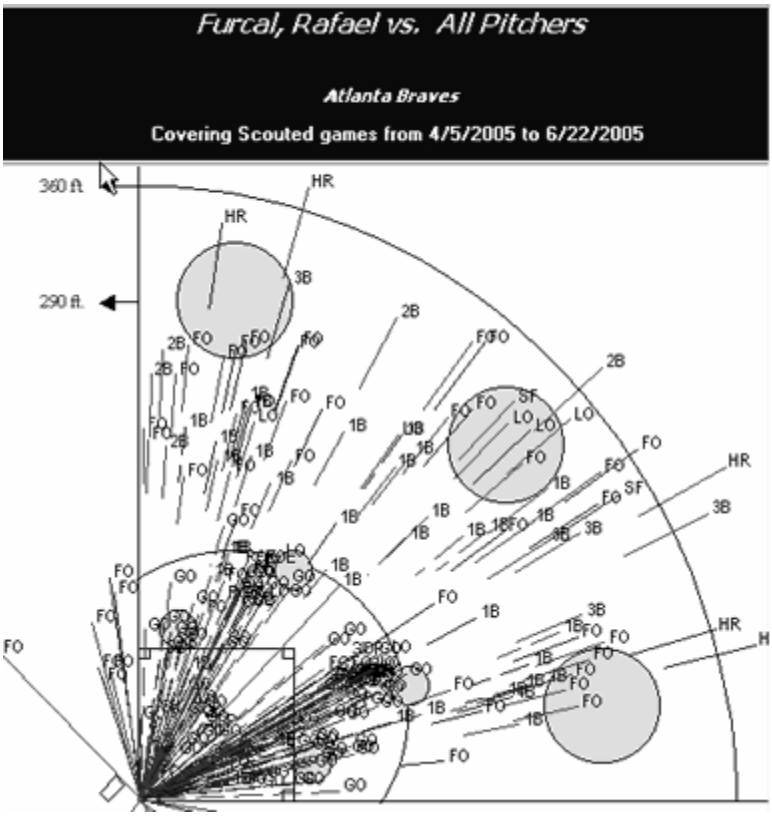

Figure 4. Spray Report of Rafael Furcal (Inside Edge, 2008b)

Another more complete report is the Pitcher Postgame as shown in Figure 5. From this output you can easily see the increases in the velocity of pitches as the game progresses (from 92 to 95 mph for fastballs) as well as pitch effectiveness (opposing left-handed batters, LHBs, perform poorly against Bartolo Colon's fastball pitch with a 0.167 batting average).

The graphical representation of pitcher performance in the strike zone, based on individual statistical performance, can allow pitchers to visually comprehend which areas of the strike zone they are most effective.

Pitcher Postgame Report
Colon, Bartolo - Angels

Inside Edge SCOUTING SERVICES

Game Information / Totals

Date: 6/15/05 **Opponent:** Nationals **Location:** Anaheim

IP	TPs	PA	AB	H	2B	3B	HR	KS	KC	BB	IBB	HBP	Pit's/PA
9.0	92	34	33	7	0	0	1	1	1	0	0	0	2.7

Average Velocity by Pitch Count

	Avg.	Low	High	Pit's 1-15	Pit's 16-30	Pit's 31-45	Pit's 46-60	Pit's 61-75	Pit's 76-90	Pit's 91-105	Pit's 106-120	Pit's 121+
Fastball	93	89	96	92	93	94	92	95	92	95	0	0
Curve Ball	0	0	0	0	0	0	0	0	0	0	0	0
Slider	85	82	89	84	84	86	84	88	0	0	0	0
Changeup	84	83	86	0	83	84	84	84	86	0	0	0
Other	0	0	0	0	0	0	0	0	0	0	0	0

Pitch Breakdown vs. RHBs

				1st Pitch		
PA : 12	% (#) Pit's	Str. %	Opp. BA	% (#) Pit's	Str. %	Opp. BA
Fastball	83% (24)	79%	.200 (2/10)	83% (10)	90%	.000 (0/3)
Curve Ball	0% (0)		(0/0)	0% (0)		.000 (0/0)
Slider	17% (5)	60%	.500 (1/2)	17% (2)	100%	.500 (1/2)
Changeup	0% (0)		(0/0)	0% (0)		.000 (0/0)
Other	0% (0)		(0/0)	0% (0)		.000 (0/0)
Total	100% (29)	76%	.250 (3/12)	100% (12)	92%	.200 (1/5)

Pitch Breakdown vs. LHBs

				1st Pitch		
PA: 22	% (#) Pit's	Str. %	Opp. BA	% (#) Pit's	Str. %	Opp. BA
Fastball	87% (55)	76%	.167 (3/18)	91% (20)	60%	.250 (1/4)
Curve Ball	0% (0)		(0/0)	0% (0)		.000 (0/0)
Slider	5% (3)	100%	.000 (0/1)	0% (0)		.000 (0/0)
Changeup	8% (5)	80%	.500 (1/2)	9% (2)	50%	.000 (0/1)
Other	0% (0)		(0/0)	0% (0)		.000 (0/0)
Total	100% (63)	78%	.190 (4/21)	100% (22)	59%	.200 (1/5)

NOTE: Total pitches include unspecified pitch types.

Location of Pitches

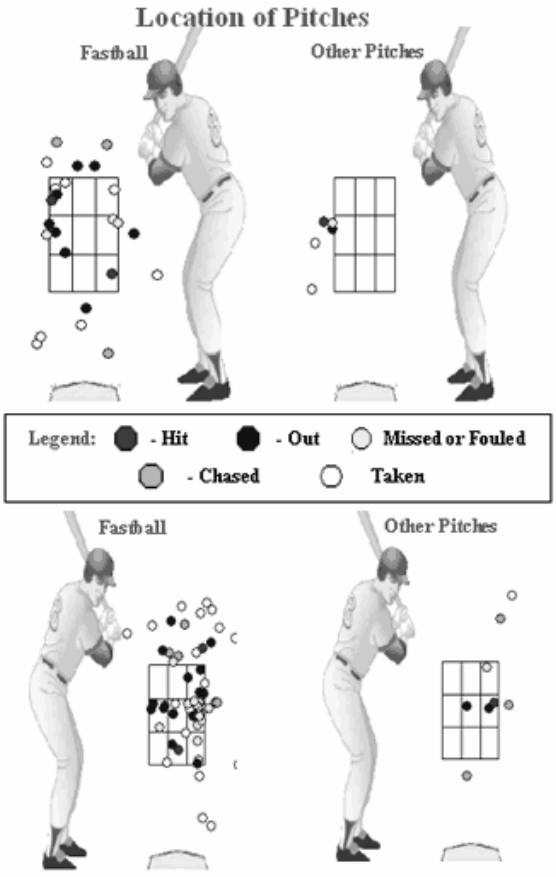

Figure 5. Pitcher Postgame report for Bartolo Colon (Inside Edge, 2008b)

Fraud Detection

Fraud in sports is nothing new. While some scandals have led to historical precedence, e.g., the 1919 White Sox throwing the World Series resulted in eight players being banned from organized baseball for life; some scandals were controversial, e.g., Pete Rose's alleged betting on the team he was managing and his subsequent banning from the game; to recent developments including the use of performance-enhancing drugs. Fraudulent activity in sports generally falls into one of three categories: poor player performance, a pattern of unusual calls from the referee and lopsided wagering (Audi & Thompson, 2007).

Poor player performance, or point shaving, is one way in which game integrity can be compromised. This involves a player or group of players that purposefully under-perform in order to affect the game's betting line. Before two teams physically meet for a match, sportsbooks set a betting line which will draw an equal dollar amount of wagers for either team, that way the losing side of the wager pays the winning side minus the sportsbook's commission. Should the line become unbalanced, the sportsbooks would be responsible for the difference and would either cause them to lose money or lose business. If one team is heavily favored, then the line will be more pronounced with one team having to achieve a larger victory in order to win the wager. Point Shaving is simply a player trying to manipulate the outcome of the game by not meeting the betting line. A recent study into NCAA basketball, found that 1% of games involve some form of point shaving (Wolfers, 2006). Being able to discover instances of point shaving is incredibly difficult (Dobra et al., 1990), especially when there is no serial correlation in betting markets from game to game (Oorlog, 1995).

A pattern of unusual calls from the referee can also influence the game outcome. Similar to Point Shaving, compromised referees can also manipulate the betting line. Referees have it in their power to make the game easier or harder for a team, and thus influence the betting line

(Igloo Dreams, 2007). A recent example of this was in the summer of 2007 when NBA referee Tim Donaghy was investigated by the FBI for compromising games to pay off his gambling debts.

Both point shaving and questionable referee calls have the same outcome in mind, making money. Thus lopsided wagering can be used as an indicator of a compromised game. This type of wagering could involve betting in excess of what is normally expected or betting heavily against the favorite. In one particular example, a gambler from Detroit made repeated bets against the University of Toledo versus Temple in football (Audi & Thompson, 2007). One of the wagers, $20,000, was four times larger than what was considered to be a typical large wager for that conference. This gambler correctly picked that Toledo would be unable to make a required number of points and suspicions were raised from Sportsbook operators. In the following game, more atypical wagers began coming in against Toledo forcing one of the Sportsbooks to cancel Toledo events from their boards. Sportsbooks make their money through evenly positioning the wagers, one side hands their money over to the other, minus a commission. When games are compromised and the wagers uneven, the Sportsbook will lose money on the event. So it becomes in the Sportsbooks best interest to keep integrity within sports and to set unbiased betting lines (Paul & Weinbach, 2005).

Las Vegas Sports Consultants (LVSC)

One of the organizations that actively looks for fraudulent sports activity is Las Vegas Sports Consultants Inc (LVSC). This group sets betting lines for 90% of Las Vegas casinos. The LVSC statistically analyzes both betting lines and player performance, looking for any unusual activity. Player performance is judged on a letter-grade scale (i.e., A-F) and takes into account variables such as weather, luck and player health (Audi & Thompson, 2007). Taken together,

games are rated on a 15 point scale of seriousness. A game rated at 4 or 5 points may undergo an in-house review, 8-9 point games will involve contact with the responsible league. Leagues are similarly eager to use the services of LVSC to maintain game honesty. The LVSC counts several NCAA conferences, the NBA, NFL and NHL as some of its clients.

Offshore Betting

Las Vegas Sportbooks are not the only gambling institutions with an interest in honest and fair sports events, offshore betting operations are starting to fill this role as well. One offshore gambling site, Betfair.com, has signed an agreement with the Union of European Football Associations (UEFA) to help monitor games for match-fixing, unusual results or suspicious activity (Jimmy, 2007). While difficult to detect all instances of gambler-tainted matches, these beginning steps can assure fans and bettors alike that vigilance is taking place and that player and referee activity is being scrutinized.

Predictive Modeling for Sports and Gaming

The ability to predict future events has been the drive for many researchers and gamblers alike. This shared drive has led to many exciting developments in the sports world, including statistical simulations and machine learning techniques. Using these concepts, trends in the data can be identified and manipulated for personal, competitive or an economic advantage.

One area of predictive research that has been comprehensively explored is that of streaky performance. This includes players with the supposed “hot-hand” effect where performance is elevated above average for an extended period of time. In research on the hot-hand effect in basketball, it was argued that if the hot-hand effect existed, making a shot would increase the chance of making another shot (Tversky & Gilovich, 2004). However, from empirical study the success of a shot was found to be independent of previous shot outcomes. Baseball though still

has its adherents to streaky behavior. In an interesting piece of research that sought to model streaky player performance, it was found that certain players do exhibit significant streakiness, more than what probability can account for (Albert, 2008). This is where simulation and machine learning comes into focus.

Statistical Simulations

Statistical simulations involve the imitation of new game data using historical data as a reference. Once this imitation data has been constructed, it can be compared against actual game play to see the accuracy of its predictive power. Simulations can be performed in a wide variety of sports domains including baseball, basketball, football and hockey.

Baseball

Baseball has long been a hotbed of simulation, with fantasy and rotisserie leagues to name two. Simulations can be made on finding the optimal pinch hitters using Markov chains, where matrices of players, inning states (top or bottom of the inning), number of outs and the on-base possibilities are all taken into account and multiplied by substitution matrices using pinch hitters (Hirotsu & Wright, 2003). This method can then be used to find the optimal pattern of player substitutions based upon the given situation.

A player-focused simulation method developed at Loyola Marymount, uses historical player data to predict future homerun totals by analyzing the frequency distributions of homeruns, where top performances (i.e., record-breaking seasons) are considered “large” events and then relating those large event frequencies to the frequencies of smaller events (i.e., individual homeruns) (Kelley et al., 2006). To put it loosely, if the ball is flying out of the park more than usual during a season, the potential exists for someone to have a terrific year, leading to the observation that such historical performances are often linked.

Another study that investigated the prediction of Division winners, those that finish first within their respective division, used a two-stage Bayesian model based on a team's relative strength, measured by winning percentage, batting averages and starting pitcher's ERAs; and home field advantage, where it is suspected that teams playing at home possess an advantage (Yang & Swartz, 2004). This study simulated MLB baseball's entire 2001 season and their method was found to be surprisingly accurate in predicting 5 of the 6 division winners by July 30th. Other Bayesian models such as predicting Cy Young winners (best pitcher in the league that season) have also netted similar accuracy results (Smith et al., 2007).

B-BALL

One popular basketball simulator is BBall. It was developed by basketball researcher Bob Chaikin, a consultant of the Miami Heat (Solieman, 2006). This software uses historical data and APBRmetrics to simulate anywhere from one game to an entire season. Developed for NBA coaches, scouts and general managers, BBall can determine a team's optimum substitution pattern over the course of a season (e.g., the pattern that produces the most simulated wins), the effect a player trade may have on the team's performance, the effect of losing one or more players to injury and the identification of the factors necessary to improve team performance (i.e., rebounds, assists, scoring, etc).

Other Sports Simulations

Other sports can benefit from using simulated data as well. In Yacht Racing, a variety of factors on boat design can be tested and winning designs can be put into practice (Philpott et al., 2004). In Boxing, an array of both physical and psychological characteristics can be used to determine match winners 81% of the time (Lee, 1997).

Hockey game simulation research involves using hidden Markov chains to pattern expected outcomes based upon where the puck is located and the team holding possession (Thomas, 2006).

Football games can be simulated using both regressive and autoregressive techniques to determine the factors most responsible for scoring events (Glickman & Stern, 1998; Willoughby, 1997), as well as Bayesian learning (Stern, 1991). Soccer has taken advantage of simulating game play by using Monte-Carlo methods (Koning, 2000; Rue & Salvensen, 2000).

Data can often hold indications of future performance. By using the right algorithms to identify the key drivers of knowledge, historical data can be used to make accurate predictions.

Machine Learning

Aside from statistical prediction, machine learning techniques are another method of providing sport-related predictions. Neural Networks are one of the most predominant machine learning systems in sports. Within neural networks, data sets are learned by the system and hidden trends in the data can be exploited for a competitive or financial advantage. Other machine learning techniques include genetic algorithm, the ID3 decision tree algorithm and a regression-based variant of the Support Vector Machine (SVM) classifier, called Support Vector Regression (SVR).

Soccer

In a predictive study of Finland's soccer championships, Rotshtein et. al. compared the forecasting ability of both genetic algorithms and neural networks (Rotshtein et al., 2005). They first set about classifying the wins into one of five categories: big loss, small loss, draw, small win and big win, where a big loss would be in the range of 3 to 5 point deficit, small loss a 1 to 2 point deficit, etc. From there, they fed past tournament data (e.g., the tournament win/loss

performance of each team over the prior 10 years) into both a genetic algorithm and a neural network for training on the most recent seven years worth of data. The results found that the neural network performed significantly better than the genetic algorithm in all five categories. Overall the neural network had 86.9% accuracy of selecting winners as compared to the genetic algorithm's 79.4% accuracy. The other finding was that the neural network required less time for training as the genetic algorithm was attempting to optimize the solution set and did not satisfy the study's stopping conditions.

Greyhound and Thoroughbred Racing

Predictive algorithms can also be conducted in other non-traditional sports, such as Greyhound and Thoroughbred racing. These types of predictions generally involve machine learning techniques to first train the system on the various data components and second to feed new data and extract predictions from it.

One machine learning technique that has been used with success in Greyhound racing, is neural networks. Feeding a back-propagation neural network (BPNN) 10 parameters gleaned from greyhound racing experts as the most important variables, Chen et. al. evaluated simulation results in two ways; accuracy of predicting a winner and payout if a bet was placed on the predicted winner (Chen et al., 1994). From this work, they found the BPNN operating with 20% accuracy and a $124.80 payout as compared to track experts which managed 18% accuracy and a payout loss of $67.60. Using the same data on the ID3 algorithm, Chen found better accuracy, 34%, but lower payout (than BPNN), $69.20. It was posited that the BPNN was better equipped to find and capitalize on the races with higher odds which led to its better payout in spite of lower accuracy.

A follow-up study that built upon Chen's work, examined the influence of predicting Wins, Quiniela (selecting the first two dogs to finish in any order) and Exacta (selecting the first two dogs to finish in order). In this study researchers tested a 4 layer BPNN with 18 parameters instead of 10, and found similar accuracy results (24.9% Win, 8.8% Quiniela, 6.1% Exacta) but differing payouts ($6.60 loss for Win, $20.30 gain for Quiniela, $114.10 gain for Exacta) (Johansson & Sonstrod, 2003). It was suspected that the extra parameters used had a substantial impact on predicting longshot races.

Another follow-up study to Chen's pioneering work, used the SVR machine learning algorithm instead of BPNN. This variant of SVM takes the hyperplane that is used to maximally separate the classes and performs regression estimates against it (Schumaker & Chen, 2008). This study was more interested in determining the factors that go into predicting long shots, and would vary its bets from just strong dogs (those that are predicted to finish first) to betting on all dogs (those that will finish above eighth place). From this simulation strategy and a study of all the various exotic wagers, it was found that this system achieved a peak 17.39% accuracy on Superfecta Box wagers (i.e., betting on the first four dogs in any order) when betting on dogs expected to finish between first and second place or better, as compared to random probability at 2.79% (Schumaker, 2007).

Predictive measures have also been performed within Thoroughbred Racing. In a study of the factors that lead to racing success, it was found that the motion of a two-year old thoroughbred's foreleg had a direct relation to its future earnings potential (Seder & Vickery, 2005). Horses that were determined by veterinary experts as having good foreleg motion (e.g., a lack of extraneous activity), earned 83% more than those with bad foreleg motion. This has a

direct impact on the racing industry as thoroughbred investments can now be screened and improve the chances of selecting a winning steed.

Another important factor in thoroughbred racing is the career length of the horse. Thoroughbreds with longer careers will understandably have better earning potential than those with shorter careers. Using simulation techniques on a pool of potential genetic parents, theoretical offspring can be modeled and their career lengths approximated (Burns et al., 2006). By using these techniques, thoroughbred owners can attempt to maximize the revenue potential of their investments.

Multimedia and Video Analysis

Traditional sports statistics are quickly becoming dwarfed by advances in multimedia technologies for sports. Within the past several years, video capture and isolating particular events in sports for later analysis has become more mainstream. Baseball players can go into the locker room and study a pitcher's delivery or their own motions either to prepare before the game or make adjustments during (Lewis, 2003). One company that is leading the way in baseball's video content is Advanced Media (Ortiz, 2007). Advanced Media handles the digital content of Major League Baseball, including video streaming games to fans, and the innovative MLB Game Day tool which can allow fans to watch an abstraction of an actual game, with only the basic information sans video. MLB's Advanced Media became such a lucrative success that substantial revenues are distributed to all 30 teams.

Basketball players can use similar video footage services to query the system for particular shots or defensive moves (Sandoval, 2006). Before this technology became available, teams would often have to wait several days to receive game footage, but now footage is almost instantaneously streamed to players, coaches, and scouts alike. It used to be a tedious process to

retrieve particular video sequences, but now queries such as “corner kicks resulting in a goal in the final two minutes” will return the appropriate footage for further analysis.

Searchable Video

The process of automated video searching on the surface may appear as a daunting task. While broadcast video may be considered a wide domain, sports events can generally be sequenced in a particular order of actions (Roach et al., 2002). As a baseball example to find a particular player’s at-bat sequence, many facts about the game are known in advance, such as the batting order. Similarly, when transitioning from one batter to the next, the first batter will either get on-base or head for the dugout, both of which can be identified as separate events by a multimedia system and tagged as such. Then the process of retrieving specific video sequences becomes a matter of querying the tagged material.

Tagging sports events in real-time can be done manually by domain experts or automatically by identifying a sequence of events (e.g., a basketball appearing near the rim can be tagged as a shot). Automatic tagging typically takes advantage of changes in the video, such as pans, zooms, fades and cuts which signals that a new video segment has begun (Truong & Dorai, 2000).

Tags can be metadata or a simple descriptor of events. Consider the following metadata tagging using XML (Babaguchi et al., 2007):

```
<AudioVisualSegment>
   <StructuralUnit>
      <Name>at-bat</Name>
   </StructuralUnit>
   <TextAnnotation>
      <FreeTextAnnotation>Arias</FreeTextAnnotation>
      <StructuredAnnotation>
         <Who>
            <Name>Arias</Name>
            <Name>Kudou</Name>
         </Who>
      </StructuredAnnotation>
      <KeywordAnnotation>
         <Keyword>SoloHomeRun</Keyword>
         <Keyword>OpenTheScoring</Keyword>
      </KeywordAnnotation>
   </TextAnnotation>
```

```
</AudioVisualSequence>
```

In this sequence, one can easily understand that the event is a solo homerun by Arias which led to the team's initial score of the game.

One notable multimedia tool is SoccerQ, which allows users to store, manage and retrieve soccer game video sequences (Chen et al., 2005). This program supports basic queries such as: select video/shot/variable from search_space [where condition]. The "variable" term can refer to a particular team, where "search_space" can be limited to certain sub-categories, such as men's or women's soccer, etc. As an example, a query may be "select all corner kick shots from all female soccer videos where the corner kick resulted in a goal event occurring in 2 minutes," as shown in Figure 6.

Figure 6. SoccerQ Video Retrieval (Chen et al., 2005)

From Figure 6, the left side of the SoccerQ application has only “female soccer” selected to limit the search space. The right side shows two events selected, “Corner Kick” and “Goal.” To the far right, the temporal precedence of Event A (Corner Kick) starts Event B (Goal) is selected. At the right center, 2 minutes is selected. The far bottom shows the 4 scenes that match the criteria.

The field of searchable sports video is emerging. Tools that link multimedia retrieval to data mining are promising, but few. It will be interesting to see where the field goes in the next several years.

Motion Analysis

Motion Analysis in sports research is generally concerned with object tracking and trajectory. Baseball is a hotbed of motion analysis research where not only pitching mechanics and ball trajectory are analyzed, but also the motion of batters as they approach different pitches. Object tracking and trajectory analysis usually starts with a baseline video image where the item of interest is selected or some reference point of known size is recognized (e.g., a human can be of an estimated size for comparison to the size of a ball) (Chang & Lee, 1997). Once the system has been calibrated, it can track the motion of the intended target.

Another method in motion analysis is to break videos into 3 component parts: the background or camera motion, the foreground object motion, and shot or scene changes (i.e., a different camera view) as a result of an external edit (Roach et al., 2001). From this approach, videos can be segmented and foreground motion tracked.

One of the techniques to identify trajectory is to filter the video frames. Using baseball as an example, you would filter the video to identify all the white objects in the frames, white being the color of the ball. From there, you eliminate any of the white elements that do not conform to the size or shape of the baseball. Now that a candidate list of balls have been identified, the process is repeated on the following frames and the white speck that exhibits motion between the pitchers mound and the batters box is the baseball (Chu et al., 2006). Once the trajectory has been identified, the type of pitch can be extracted as well. Curve balls typically exhibit a high arching motion which stands in contrast to a slider with lateral or sharp downward motion.

These systems have fairly high accuracy, around 90% (Chen et al., 2007). Aside from baseball, these techniques can be used on other sports as well such as soccer to track ball location, player location, shot distance, the distance a player has run and ball speed (Bialik, 2007; TRACAB, 2007; Yow et al., 1995), in football (Ding & Fan, 2007) and even tennis (Takagi et al., 2003).

Challenges and Future Directions

There are several challenges within the domain of sports knowledge management and data mining that we feel should be addressed. Presently, not many organizations are taking full advantage of advanced sports knowledge management or data mining techniques. This resistance to change is firmly ingrained in older sports such as baseball and more than likely stems from the old axiom "if it ain't broke don't fix it." However, organizations should realize that those that do embrace these sophisticated instruments, generally perform better (e.g., Oakland Athletics, Toronto Blue Jays, Boston Red Sox and the Cleveland Indians).

Secondly, the individual sports organizations that have recognized the potential competitive advantages of knowledge management and data mining systems, typically hold results in-house and do not share technologies or lessons learned with their fans or peer organizations. While this could be considered capitalistic in nature, other sports organizations take a different approach and house all data in a central sport-related repository where individuals and teams all can have equal access. Both approaches have their respective advantages, however, what is missing here are hybrid approaches in which a significant amount of material is housed collectively and teams are still free to exploit any advantages found therein. We see the beginning of such a hybrid approach for various sport-related interest groups.

The Australian Institute of Sport (AIS) has recently unveiled two initiatives designed to effectively house data from various sports (Lyons, 2005). The first of which is the creation of a

digital repository to store various sport-related video, audio and data files. This centralized repository will allow players and teams to access relevant data. The second initiative provides data mining techniques on the data repository in order to gather new insight into obvious patterns. We would suggest that further knowledge could be extracted from such a repository and that individual teams and players should take advantage by implementing their own proprietary tools in the pursuit of a competitive advantage.

The full application of sports knowledge management and data mining is still in its infancy. While several pioneering organizations are beginning to harness their data through advanced statistical/predictive analyses, many are struggling with the prospect of adopting such systems let alone using them as a competitive advantage. As larger market professional sports organizations increase payrolls to meet demands for talent, knowledge management and data mining can be leveraged by the smaller market teams to remain competitive. This competitive balance has begun to return parity to sports leagues. However, as more organizations begin to embrace these knowledge eschewing principles, it won't be long before an arms race of sorts develops, where two teams emerge; the players on the field and the analysts in the back office. Both of which will work together to propel the organization forward. Also, future advances such as distributed Artificial Intelligence that uses multiple agents or new applications of existing algorithms borrowed from computer science or physics, may revolutionize sports data mining. Similarly, centralized public data repositories constructed either by governments or a collective of fans, will also allow for the continuation of these techniques for teams, performance measures and predictive purposes. It will be interesting to see where the next few years will take us.

References

82games.com 2008. A visitor's guide to 82games.Com. Retrieved Feb 20, 2008, from http://82games.com/newuser.htm.

Ackoff, R. 1989. From data to wisdom. *JOURNAL OF APPLIED SYSTEMS ANALYSIS* 16: 3-9.

Adler, J. 2006. *Baseball hacks*. O'Reilly, Beijing.

Alavi, M. & D. E. Leidner 2001. Review: Knowledge management and knowledge management systems: Conceptual foundations and research issues. *MIS QUARTERLY* 25(1): 107-136.

Albert, J. 1997. An introduction to sabermetrics. Retrieved Jan 30, 2008, from http://www-math.bgsu.edu/~albert/papers/saber.html.

Albert, J. 2008. Streaky hitting in baseball. *JOURNAL OF QUANTITATIVE ANALYSIS IN SPORTS* 4(1).

Allsopp, P. & S. Clarke 2004. Rating teams and analysing outcomes in one-day and test cricket. *JOURNAL OF THE ROYAL STATISTICAL SOCIETY: SERIES A* 167(4): 657-667.

Audi, T. & A. Thompson 2007. Oddsmakers in vegas play new sports role. The Wall Street Journal. A1.

Babaguchi, N., J. Ohara, et al. 2007. Learning personal preference from viewer's operations for browsing and its application to baseball video retrieval and summarization. *IEEE TRANSACTIONS ON MULTIMEDIA* 9(5): 1016-1025.

Ballard, C. 2006. Measure of success. Sports Illustrated.

Barlas, I., A. Ginart, et al. 2005. Self-evolution in knowledgebases. *IEEE AutoTestCon*, Orlando, FL.

Barros, C. P. & S. Leach 2006. Performance evaluation of the english premier football league with data envelopment analysis. *APPLIED ECONOMICS* 38(12): 1449-1458.

Baseball-Reference.com 2008. Baseball-reference. Retrieved Feb 20, 2008, from http://www.baseball-reference.com/.

Baseball Info Solutions 2003. *The bill james handbook*. ACTA Publications, Chicago.

Beech, R. 2008. Nba player shot zones. Retrieved Jan 30, 2008, from http://www.82games.com/shotzones.htm.

Berry, S. 2005. Introduction to the methodologies and multiple sports articles. In *Anthology of statistics in sports*, J. Albert, J. Bennett & J. Cochran. Cambridge University Press, Alexandria, VA.

Bhandari, I. 1995. Attribute focusing: Data mining for the layman. Research Report RC 20136. IBM TJ Watson Research Center.

Bhandari, I., E. Colet, et al. 1997. Advanced scout: Data mining and knowledge discovery in nba data. *Data Mining and Knowledge Discovery* 1(1): 121-125.

Bialik, C. 2007. Tracking how far soccer players run. The Wall Street Journal.

Bierly, P. E., E. H. Kessler, et al. 2000. Organizational learning, knowledge and wisdom. *JOURNAL OF ORGANIZATIONAL CHANGE MANAGEMENT* 13(6): 595-618.

Birnbaum, P. 2008. Sabermetrics. Retrieved Feb 16, 2008, from http://philbirnbaum.com/.

Boisot, M. & A. Canals 2004. Data, information and knowledge: Have we got it right? *JOURNAL OF EVOLUTIONARY ECONOMICS* 14(1): 43-67.

Burns, E., R. Enns, et al. 2006. The effect of simulated censored data on estimates of heritability of longevity in the thoroughbred racing industry. *GENETIC MOLECULAR RESEARCH* 5(1): 7-15.

Carlisle, J. P. 2006. Escaping the veil of maya - wisdom and the organization. *39th Hawaii International Conference on System Sciences*, Koloa Kauai, HI.

Carroll, B., P. Palmer, et al. 1998. *The hidden game of football: The next edition*. Total Sports Inc., New York, NY.

Chang, C.-W. & S.-Y. Lee 1997. A video information system for sport motion analysis. *JOURNAL OF VISUAL LANGUAGES AND COMPUTING* 8(3): 265-287.

Chen, H.-S., H.-T. Chen, et al. 2007. Pitch by pitch extraction from single view baseball video sequences. *IEEE International Conference on Multimedia and Expo*, Beijing, China.

Chen, H. 2001. *Knowledge management systems - a text mining perspective*. The University of Arizona - Dept of Management Information Systems, Tucson.

Chen, H. 2006. *Intelligence and security informatics for international security: Information sharing and data mining*. Springer, New York, NY.

Chen, H. & M. Chau 2004. Web mining: Machine learning for web applications. *ANNUAL REVIEW OF INFORMATION SCIENCE AND TECHNOLOGY (ARIST)* 38: 289-329.

Chen, H., P. Rinde, et al. 1994. Expert prediction, symbolic learning, and neural networks: An experiment in greyhound racing. *IEEE EXPERT* 9(6): 21-27.

Chen, S.-C., M.-L. Shyu, et al. 2005. An enhanced query model for soccer video retrieval using temporal relationships. *International Conference on Data Engineering*, Tokyo, Japan.

Chu, W.-t., C.-W. Wang, et al. 2006. Extraction of baseball trajectory and physics-based validation for single-view baseball. *IEEE International Conference on Multimedia and Expo*, Toronto, Ontario.

Cleveland, H. 1982. Information as a resource. *THE FUTURIST* 16(6): 34-39.

Coleman, J. & A. Lynch 2001. Identifying the ncaa tournament "Dance card". *INTERFACES* 31(3): 76-86.

Coleman, J. & A. Lynch 2008a. Dance card rankings for 2008. Retrieved Feb 19, 2008, from http://www.unf.edu/~jcoleman/dance.htm.

Coleman, J. & A. Lynch 2008b. Score card rankings for 2008. Retrieved Jan 30, 2008, from http://www.unf.edu/~jcoleman/score.htm.

Cox, A. & J. Stasko 2002. Sportsvis: Discovering meaning in sports statistics through information visualization. *IEEE Symposium on Information Visualization*, Baltimore, Maryland.

CricInfo 2008. Wisden. Retrieved June 19, 2008, from http://content-www.cricinfo.com/wisdenalmanack/content/current/story/almanack/.

Davenport, T. & L. Prusak 1998. *Working knowledge*. Harvard Business School Press, Cambridge, MA.

Digital Scout 2008. Digital scout. Retrieved Feb 20, 2008, from http://www.digitalscout.com/.

Ding, Y. & G. Fan 2007. Segmental hidden markov models for view-based sport video analysis. *IEEE Conference on Computer Vision and Pattern Recognition*, Minneapolis, MN.

Dobra, J., T. Cargill, et al. 1990. Efficient markets for wagers: The case of professional basketball wagering. In *Sportometrics*, B. Goff & R. Tollison. Texas A&M University Press, College Station, TX, 215-249.

Dong, D. & R. Calvo 2007. Integrating data mining processes within the web environment for the sports community. *IEEE International Conference on Integration Technology*, Shenzhen, China.

Dvorak, J., A. Junge, et al. 2000. Risk factor analysis for injuries in football players: Possibilities for a prevention program. *THE AMERICAN JOURNAL OF SPORTS MEDICINE* 28(5): 69-74.

Fetter, H. 2003. *Taking on the yankees - winning and losing in the business of baseball - 1903-2003*. W.W. Norton & Co., New York.

Fieltz, L. & D. Scott 2003. Prediction of physical performance using data mining. *RESEARCH QUARTERLY FOR EXERCISE AND SPORT* 74(1): 1-25.

Flinders, K. 2002. Football injuries are rocket science. Vnunet.com. London.

Glickman, M. & H. Stern 1998. A state-space model for national football league scores. *JOURNAL OF AMERICAN STATISTICS ASSOCIATION* 93: 25-35.

Hastie, T., R. Tibshirani, et al. 2001. *The elements of statistical learning: Data mining, inference, and prediction*. Springer-Verlag, New York.

Hirotsu, N. & M. Wright 2003. A markov chain approach to optimal pinch hitting strategies in a designated hitter rule baseball game. *JOURNAL OF OPERATIONS RESEARCH* 46(3): 353-371.

Hollinger, J. 2002. *Pro basketball prospectus: 2002 edition*. Brassey's Inc., Dulles, VA.

Hopkins, W., S. Marshall, et al. 2007. Risk factors and risk statistics for sports injuries. *CLINICAL JOURNAL OF SPORTS MEDICINE* 17(3): 208-210.

Igloo Dreams 2007. Data mining: The referees. Retrieved Feb 6, 2008, from http://igloodreams.blogspot.com/2007/08/data-mining-referees.html.

Ilardi, S. 2007. Adjusted plus-minus: An idea whose time has come. Retrieved Sept. 25, 2008.

Inside Edge 2008a. About us. Retrieved Feb 20, 2008, from http://www.inside-edge.com/about_us.htm.

Inside Edge 2008b. Sample reports. Retrieved Feb 20, 2008, from http://www.inside-edge.com/minor/sample_reports.htm.

International Association for Sports Information 2008. About iasi. Retrieved Feb 16, 2008, from http://www.iasi.org/about/index.html.

International Association on Computer Science in Sport 2008. Iacss - objectives. Retrieved Feb 16, 2008, from http://www.iacss.org/iacss/iacss_obj.html.

James, B. 1979. *The bill james baseball abstract*. Self Published.

James, B. 1982. *The bill james baseball abstract*. Ballantine Books, New York.

James, B. & J. Henzler 2002. *Win shares*. STATS Publishing, Morton Grove, IL.

Jimmy, D. 2007. Point-shaving ref: Nba betting scandal borne of hypocritical anti-betting stance. Retrieved Feb 22, 2008, from http://www.jimmydsports.com/fantasy-sports-columns/nba-point-shaving-ref-july242007.aspx.

Johansson, U. & C. Sonstrod 2003. Neural networks mine for gold at the greyhound track. *International Joint Conference on Neural Networks*, Portland, OR.

Kelley, D., J. Mureika, et al. 2006. Predicting baseball home run records using exponential frequency distributions. Retrieved Jan 15, 2008, from http://arxiv.org/abs/physics/0608228v1.

Koning, R. 2000. Balance in competition in dutch soccer. *THE STATISTICIAN* 49: 419-431.

Lahti, R. & M. Beyerlein 2000. Knowledge transfer and management consulting: A look at the firm. *BUSINESS HORIZONS* 43(1): 65-74.

Lee, C. 1997. An empirical study of boxing match prediction using a logistic regression analysis. *Section Statistics Sports, American Statistical Association, Joint Statistical Meeting*, Anaheim, CA.

Levin, R., G. Mitchell, et al. 2000. The report of the independent members of the commissioner's blue ribbon panel on baseball economics. Major League Baseball.

Lewis, M. 2003. *Moneyball*. W.W. Norton & Company, New York.

Lyons, K. 2005. Data mining and knowledge discovery. *AUSTRALIAN SPORTS COMMISSION JOURNALS* 2(4).

MIT Sloan Alumni Profile 2008. Daryl morey, mba '00. Retrieved Jan 30, 2008, from http://mitsloan.mit.edu/mba/alumni/morey.php.

O'Reilly, N. & P. Knight 2007. Knowledge management best practices in national sport organizations. *INTERNATIONAL JOURNAL OF SPORT MANAGEMENT AND MARKETING* 2(3): 264-280.

Oliver, D. 2005. *Basketball on paper: Rules and tools for performance analysis*. Brassey's Inc., Dulles, VA.

Oorlog, D. 1995. Serial correlation in the wagering market for professional basketball. *Quarterly JOURNAL OF BUSINESS AND ECONOMICS* 34(2): 96-109.

Ortiz, J. L. 2007. Mlb's online venture is big hit. USA Today. 5C.

Page, G. 2005. Using box scores to determine a position's contribution to winning basketball games. Dept of Statistics. Brigham Young University.

Papahristoulou, C. 2006. Team performance in uefa champions league 2005-2006. MRPA Paper No. 138.

Paul, R. & A. Weinbach 2005. Bettor misconceptions in the nba: The overbetting of large favorites and the hot hand. *JOURNAL OF SPORTS ECONOMICS* 6(4): 390-400.

Pelton, D. 2005. The sonics play moneyball: Part one. Retrieved Jan 30, 2008, from http://www.nba.com/sonics/news/moneyball050119.html.

Petro, N. 2001. Digital scout to provide statistical analysis for USA baseball tournament. Retrieved Feb 20, 2008, from http://www.digitalscout.com/news/news_tournament.php.

Petro, N. 2003. Digital scout signs two-year agreement with little league baseball. Retrieved Feb 20, 2008, from http://www.digitalscout.com/news/news_littleleague.php.

Philpott, A., S. Henderson, et al. 2004. A simulation model for predicting yacht match race outcomes. *OPERATIONS RESEARCH* 52(1): 1-16.

Piatetsky-Shapiro, G. 2008. Difference between data mining and statistics. Retrieved Oct 2, 2008, from http://www.kdnuggets.com/faq/difference-data-mining-statistics.html.

Pro-Football-Reference.com 2008. Pro-football-reference. Retrieved Feb 20, 2008, from http://www.pro-football-reference.com/.

Professional Football Researchers Association 2008. Welcome!! Retrieved Feb 16, 2008, from http://www.profootballresearchers.org/index.htm.

Roach, M., J. Mason, et al. 2001. Video genre classification using dynamics. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Salt Lake City, UT.

Roach, M., J. Mason, et al. 2002. Recent trends in video analysis: A taxonomy of video classification problems. *International Conference on Internet and Multimedia Systems and Applications*, Kauai, Hawaii.

Rosenbaum, D. 2004. Measuring how nba players help their teams. Retrieved Jan 30, 2008, from http://www.82games.com/comm30.htm.

Rosenbaum, D. 2005. A statistical holy grail: The search for the winner within. The New York Times.

Rotshtein, A., M. Posner, et al. 2005. Football predictions based on a fuzzy model with genetic and neural tuning. *CYBERNETICS AND SYSTEMS ANALYSIS* 41(4): 619-630.

Rue, H. & O. Salvensen 2000. Prediction and retrospective analysis of soccer matches in a league. *THE STATISTICIAN* 49: 399-418.

Sandoval, G. 2006. A video slam dunk for the nba. Retrieved Jan 30, 2008, from http://www.news.com/A-video-slam-dunk-for-the-NBA/2100-1008_3-6034908.html.

SAS 2005. A method to march madness. Retrieved Jan 30, 2008, from http://www.sas.com/news/feature/01mar05/dancecard.html.

Schatz, A. 2006. *Pro football prospectus 2006: Statistics, analysis, and insight for the information age*. Workman Publishing Company.

Schell, M. J. 1999. *Baseball's all-time best hitters: How statistics can level the playing field*. Princeton University Press, Princeton, NJ.

Schumaker, R. P. 2007. Using svm regression to predict greyhound races. *Information Systems Dept. Research Seminar*, New Rochelle, NY.

Schumaker, R. P. & H. Chen 2008. Evaluating a news-aware quantitative trader: The effects of momentum and contrarian stock selection strategies. *JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE* 59(1): 1-9.

Seder, J. & C. Vickery 2005. The relationship of subsequent racing performance to foreleg flight patterns during race speed workouts of unraced 2-yr-old thoroughbred racehorses at auctions. *JOURNAL OF EQUINE VETERINARY SCIENCE* 25(12): 505-522.

Serenko, A. & N. Bontis 2004. Meta-review of knowledge management and intellectual capital literature: Citation impact and research productivity rankings. *KNOWLEDGE AND PROCESS MANAGEMENT* 11(3): 185-198.

Shulman, K. 1996. Data mining in the backcourt: Advanced scout gives coaches an assist. Retrieved Feb 6, 2008, from http://www.dciexpo.com/news/archives/scout.htm.

Sinins, L. 2007. Complete baseball encyclopedia: Version 8.0.

Smith, L., B. Lipscomb, et al. 2007. Data mining in sports: Predicting cy young award winners. *JOURNAL OF COMPUTING SCIENCES IN COLLEGES* 22(4): 115-121.

Society for American Baseball Research 2008. About sabr. Retrieved Feb 16, 2008, from http://www.sabr.org/sabr.cfm?a=cms,c,110,39,156.

Solieman, O. 2006. Data mining in sports: A research overview. Dept. of Management Information Systems. The University of Arizona. Tucson.

Stern, H. 1991. On probability of winning a football game. *JOURNAL OF AMERICAN STATISTICS ASSOCIATION* 45: 179-183.

Taggart, S. 1998. Olympic data crunching. Retrieved Feb 6, 2008, from http://www.wired.com/science/discoveries/news/1998/08/14175.

Takagi, S., S. Hattori, et al. 2003. Sports video categorizing method using camera motion parameters. *International Conference on Multimedia and Expo*, Baltimore, MD.

The Association for Professional Basketball Research 2008. Apbr.Org. Retrieved Feb 16, 2008, from http://apbr.org.

The New York Times Staff 2004. *The new york times guide to essential knowledge: A desk reference for the curious mind*. St. Martin's Press, New York.

Thomas, A. 2006. The impact of puck possession and location on ice hockey strategy. *JOURNAL OF QUANTITATIVE ANALYSIS IN SPORTS* 2(1).

Thorn, J. & P. Palmer 1984. *The hidden game of baseball*. Doubleday, Garden City, NJ.

TRACAB 2007. Tracab in champions league. Retrieved April 4, 2008, from http://www.tracab.com/news.asp?id=25.

Truong, B. T. & C. Dorai 2000. Automatic genre identification for content-based video categorization. *International Conference on Pattern Recognition*, Barcelona, Spain.

Tversky, A. & T. Gilovich 2004. The cold facts about the "Hot hand" In basketball. In *Preference, belief, and similarity: Selected writings*, A. Tversky & E. Shafir. MIT Press, Cambridge, MA.

USA Today 2008. Jeff sagarin ncaa basketball ratings. Retrieved Sept. 27, 2008, from http://www.usatoday.com/sports/sagarin/bkt0708.htm.

Voigt, D. 1969. America's first red scare - the cincinnati reds of 1869. *Ohio History* 78: 13-24.

Weeks, C. 2006. Digital scout software powers the stats for the inaugural arizona cactus classic. Retrieved Feb 20, 2008, from http://www.digitalscout.com/news/news_az_cactus_classic_06.php.

White, P. 2006. Scouts uncover a winning edge. USA Today. New York: 7E.

Willoughby, K. 1997. Determinants of success in the cfl: A logistic regression analysis. *National Annual Meeting to the Decision Sciences Institute*, Atlanta, GA.

Wolfers, J. 2006. Point shaving: Corruption in ncaa basketball. *AEA Papers and Proceedings* 96(2): 279-283.

Woolner, K. 2006. Why is mario mendoza so important? In *Baseball between the numbers*, J. Keri. Basic Books, New York.

Yang, T. Y. & T. Swartz 2004. A two-stage bayesian model for predicting winners in major league baseball. *JOURNAL OF DATA SCIENCE* 2(1): 61-73.

Yow, D., B.-L. Yeo, et al. 1995. Analysis and presentation of soccer highlights from digital video. *Asian Conference on Computer Vision*, Singapore.

Zeleny, M. 1987. Management support systems: Towards integrated knowledge management. *HUMAN SYSTEMS MANAGEMENT* 7(1): 59-70.

Zhu, B. & H. Chen 2005. Information visualization. *ANNUAL REVIEW OF INFORMATION SCIENCE AND TECHNOLOGY (ARIST)* 39: 139-178.

Zimmerman, P. 1985. *The new thinking man's guide to pro football*. Simon and Schuster, New York.