Data Mining in Sports:
A Research Overview

Osama K. Solieman
MIS Masters Project
August 2006

Table of Contents:

Chapter 1: Data Mining in Sports – Applications and Opportunities

1.1 What is Data Mining?

There is valuable information hidden in data. Since the underlying data is generated much faster than it can be processed and made sense of, this information often remains buried and untapped. It becomes virtually impossible for individuals or groups with limited resources – specifically technological – to find and gain any insight from the data.

Data Mining encompasses tools and techniques for the "extraction or 'mining' [of] knowledge from large amounts of data" (Han & Kamber, 2001). It is about finding patterns and relationships within data that can possibly result in new knowledge. Furthermore, these relationships can also result in predictors of future outcomes.

The importance of data mining has been established for business applications, criminal investigations, bio-medicine (Chen et al, 2005), and more recently counter-terrorism (Chen, 2006). Most retailers, for example, employ data mining practices to uncover customer buying patterns – Amazon.com uses purchase history to make product recommendations to shoppers. Data mining can be applied wherever there is an abundance of data available for and in need of analysis.

1.2 Sports Applications

The sports world is known for the vast amounts of statistics that are collected for each player, team, game, and season. There are also many types of statistics that are gathered for each – a basketball player will have data for points, rebounds, assists, steals, blocks, turnovers, etc for each game. This can result in information overload for those trying to derive meaning from the statistics. Hence, sports are ideal for data mining tools and techniques.

Sports organizations, due to the extremely competitive environment in which they operate, need to seek any edge that will give them an advantage over others. It would appear that the culture has long encouraged analysis and discovery of new knowledge exhibited by its longstanding utilization of scouting. However, traditionally sports knowledge has been believed to be contained in the minds of its experts – the scouts, coaches, and managers. Only recently have sports organizations begun to realize that there is also a wealth of knowledge contained in their data. Currently, most team sports organizations employ in-house statisticians and analysts to retrieve meaning and insight for the scouts who evaluate future prospects and talent, the coaches who are in charge of the team on the playing surface, as well as the general managers who are in charge of drafting or signing players.

Scouting

Scouting has been a staple of the sporting world since professional sport emerged as a legitimate money-making enterprise nearly a century ago. There are two primary types of scouting efforts that are used by sports organizations.

The first, which serves primarily a human resources role, is the scouting of potential talent. To do this, scouts often travel across the nation, and increasingly across the globe, to evaluate prospects. These scouts compile reports and evaluations detailing each prospect's abilities, strengths, and weaknesses. The second form of scouting, referred to as 'advance scouting', is the assessment of upcoming competitors. These scouts travel to watch competitors and compile reports that are used to help determine strategies and approaches when facing the competition.

Traditionally, advance scouts in baseball were sent to games to collect data, chart pitches and the like, and create reports concerning team and player abilities. With the availability of statistical analysis tools and video the realm of scouting has forever changed. In an article by Paul White in USA Today (White, 2006), advance scout Shooty Babbitt states that scouting has become "more about tendencies, pitch to pitch" (Section 7E). Scouting has gone beyond the strengths and weaknesses of the opponent to analysis of typical player, coach, and team strategies and behaviors in certain situations.

Predictions from Data

Data mining can be used by sports organizations in the form of statistical analysis, pattern discovery, as well as outcome prediction. Patterns in the data are often helpful in the forecast of future events. A pilot program begun in 2002 by European soccer club AC Milan uses software to help predict player injuries by collecting data from workouts over a period of time (Flinders, 2002). The

biomedical tool created by Computer Associates produces predictions from the medical statistics amassed for each player. Since athletes are their biggest investments, teams are hoping that prediction of injury will help save millions of dollars.

Similarly, researchers indicate that data mining can be used on physical aptitude test data in order to predict future physical performance (Fieltz & Scott, 2003). Data mining software was used to link test data of cadets at the United States Military Academy and their actual performance in a required fitness class. This type of analysis would have significant implications to sports organizations – which put prospects through rigorous examination.

Prior to each annual draft, the National Football League (NFL) holds an event referred to as the Combine in which eligible college players perform various performance tests in the presence of various team personnel including scouts, coaches, and general managers. Among the included physical tests are the 40 yard dash, vertical and broad jump, as well as physical measurements. Throughout the years, NFL teams and experts have developed common consensus on what are considered poor, good, and excellent results in the test based on the performance of the athletes throughout the previous years.

In terms of the mental aspect of potential players, the Combine also allows teams to interview players and each player is expected to take the Wonderlic test. The fifty minute pre-employment examination is used by teams to assess the intelligence of prospects. Similar to the physical tests, NFL teams have

developed expected Wonderlic scores based on amount of intelligence presumed necessary to play particular positions. For example, quarterbacks – the position believed to require the most brains – score an average of 24 while running backs average 16 (Zimmerman, 2006).

Measuring Performance

As most coaches would agree, statistics themselves can be very misleading. Certain players are able to build impressive stats but have little effect on a game. On the other hand, there are players who make a significant impact on the game without having impressive statistics. In American football, take for instance the comparison of the defensive back that is more prone to taking risks to one who plays solid cover defense. The former may accumulate more interceptions than the latter – a statistic often used to indicate a defensive back's value – but will also allow the opponent greater success when the gamble does not pay off.

Jeff Van Gundy, the coach of the NBA's Houston Rockets, called defensive rebounds off of missed free-throws "the biggest selfish glut of all time" in an article by Sports Illustrated's Chris Ballard about the inaccuracies of traditional statistics (Ballard, 2005). The player who helps block out the opposition from grabbing the basketball, Van Gundy argues, is just as important as the player who collects the rebound. Some players have become adept at using certain situations to boost their statistics because higher statistics, even if misleading, will eventually result in bigger contracts. For this reason, pro teams hire internal statisticians to help make sense of all the information.

Data mining is not meant to take the place of general managers, coaches, and scouts. Rather, it is a tool that can be used to aid in the decision-making processes that they undertake. In today's business world, a CEO or executive would not make any important decision without hard numbers and figures to back it up. Sports organizations must be run similarly to their counterparts in other industries.

1.3 Advantages and Benefits

The advantage for sports organizations when it comes to data mining is in the resulting performance of their respective teams and players. Some sports are currently more advanced than others. This is especially true in the case of baseball and its current use of statistical analysis.

Moneyball

In his influential book *Moneyball* (Lewis, 2003), author Michael Lewis details how Oakland Athletics general manager Billy Beane and his staff used statistical analysis to design a low-budget team that could compete with teams in bigger markets with larger payrolls.

In order to accomplish this, Billy Beane would need to take advantage of the fact that drafted players – from high school and college – can earn only a fraction of what veterans on the free agent market can. Like most teams, the A's could not afford to make mistakes in the draft which would result in the commitment of

critical funds to unproductive players. Beane and his staff would go against the traditional view of what made a future major league baseball star, the basis of which was the perception of scouts. This method often resulted in bad decisions because scouts were prone to fall in love with players for their physical attributes rather than any past success they may have had. Furthermore, scouts had a tendency of preferring high school kids to those who competed at the college level.

Beane, whose experience as a failed baseball prodigy with ‘unlimited potential’ gave rise to his methodology, saw baseball less as an athletic endeavor and more as a skill. Some people had it and others did not. Hence the ability to play baseball was not based on certain physical attributes but rather on an inner-talent which could not be determined merely by having a player workout or watching a few games.

The answer, of course, was in the numbers. The Oakland management would use statistics as the basis for their selections. The first step was to eliminate high school prospects because it is hard to quantify their performance. This is primarily due to the inconsistency of the different high school leagues across the country. The statistics for college players, on the other hand, were more solid since the competition is at a consistently high level. While physical ability still had value, the focus of the Oakland Athletics’ draft approach was based on what the player had already accomplished not what they could potentially accomplish.

From their own analysis as well as building upon the research of others, Billy Beane and his management team concluded that traditional baseball statistics were misleading. They found that RBI – runs batted in – totals for example that are coveted by most general managers are also deceptive. In order to have high RBI stats, you need the players ahead of you to get on base. The statistics found that the most influential in showing a player's ability to score runs were on-base and slugging percentages. The Athletics sought to build a team based on the ability to get people on base and then bring them in to score. The management team sought players who did well in these categories and also were walked a lot. This showed the discipline at the plate that would be necessary for future team success.

The result was that Oakland got many of the players they had sought to draft because they were overlooked by other teams who preferred high school kids and players with higher perceived promise. On the baseball diamond, where it mattered most, Oakland was known for always fielding a competitive team. Their draft selections, ridiculed at first, often became baseball stars and award winners. Barry Zito, who was drafted in the 1999 draft, won the Cy Young Award – given annually to the best pitcher in each league – only three years later.

The Aftershock

In 2001, Theo Epstein at age 28 was hired to take over the general manager position for the Boston RedSox. With the combination of objective analysis and the ability to compete with the rival New York Yankees in terms of player

spending, Epstein helped field a champion in 2004. It was the organizations first in 86 years.

While Lewis's book brought national exposure to the existence and value of statistical analysis, it also brought attention to similar movements occurring in other sports. In the National Basketball Association, Dean Oliver – viewed as one of basketball's best statistical analysts, has been a consultant for the Seattle Supersonics for over half a decade. Oliver, who has also published the influential book *Basketball on Paper: Rules and Tools for Performance Analysis* (Oliver, 2005), helped bring a unique perspective to the organization that was driven by hard data. Similarly, other NBA teams have recently begun to hire analytical minds that can bring about new insights from the available data.

On April 3rd, 2006 the Houston Rockets hired Daryl Morey – who previously served as Senior VP of Operations and Information with the Boston Celtics – as assistant GM with the intention that he take over the general manager position in 2007. One of Morey's main responsibilities with the Celtics was the development of analytical methods and technological tools that would help in making basketball decisions. Analytical minds until that time had only been hired as consultants or staff. However, the hiring of Morey as general manager, the organizational authority on basketball decisions, marked the entry point of the NBA into the 'Moneyball era'.

Sports have become big business. There is a lot of money to be made and winning is the key to attaining it. In addition, leagues like the NBA and NFL

implement salary caps to level the playing field. With player salaries soaring, any wrong decisions are potentially disastrous to an organization. Even in baseball, where there is no league enforced salary caps, the majority of teams are limited by the size of their respective markets in terms of how much they can afford to pay their players.

For the above mentioned reasons, it is no surprise that there would be a revolution in terms of the approach to running sports organizations. The traditional approach of using intuition or 'gut feeling' to make decisions is no longer favored. Rather, assessments need to be based on strong analysis and while statistics had always been a consideration among decision-makers in sports organizations they had never been approached in the scientific manner of today.

1.4 Research Opportunities

There are a number of independent research organizations devoted to the increase of knowledge and understanding of their respective sport. These associations serve as central hubs for idea exchange and collaboration among sports experts and researchers. Many of them provide online databases and publish journals or newsletters. As the majority of these organizations are non-profit, their work is not driven by monetary gain but rather by a pure passion for sport and its research.

The Society for American Baseball Research (SABR)

Established in Cooperstown, NY in 1971, the Society for American Baseball Research (SABR) was created to foster research about the game of baseball. Among its services, the organization hosts a number of unique databases such as a free catalog of baseball literature. The Statistical Analysis Committee, founded by Pete Palmer and Dick Cramer in 1974, focuses on the study and evaluation of the game of baseball from an analytical perspective. The committee publishes quarterly issues of its newsletter which are publicly available (http://www.philbirnaum.com). The group seeks to analyze the existing baseball data – a historical view – as well as produce models to study the future of the game. Furthermore, the term "sabermetrics" – referring to the objective and scientific analysis of baseball often through statistics – was coined from the SABR acronym.

Association for Professional Basketball Research (APBR)

Founded in 1997 by Robert Bradley, the Association for Professional Basketball Research (APBR) was created to promote interest into the history of professional basketball. This not only includes the National Basketball Association (NBA) but also other professional leagues – many of which are now defunct. The APBR provides a "central library" and databases for access by researchers, authors, as well as fans. Similar to sabermetrics, the APBR has played a role in the development of "APBRmetrics". These measures seek to go beyond traditional basketball statistics which do not necessarily reflect the events and outcome of the game. Among the chief philosophies is the importance of examining

basketball games based on possessions since the team that is able to more efficiently use its possessions is often the victor. Likewise, researchers have found per-minute statistics to be more valuable than per-game statistics as per-minute averages give a better sense of a player's contributions while on the court.

Professional Football Researchers Association (PFRA)

Started in the mid-1980s, the Professional Football Research Association (PFRA) is devoted to maintaining and, when necessary, amending pro football history. Among the articles that are published on their site are football statistical analysis articles contributed by various authors (http://www.footballresearch.com/frpage.cfm?topic=articles3&categoryID=9).

The International Association on Computer Science in Sport (IACSS)

The main goal of the International Association of Computer Science in Sport (IACSS) is to facilitate cooperation among researchers and experts in the fields of computer science and sports. The association serves two main functions – organization of events and the publication of a journal. Since 1997, the IACSS has hosted the International Symposium on "Computer Science in Sport" in various European cities. Since 2002, the association has also published its electronic journal bi-annually (http://www.iacss.org/ijcss/iacss_ijcss.html).

The International Association for Sports Information (IASI)

Founded in September of 1960, the International Association for Sports Information (IASI) is a center for collaboration among information experts as well as sports scientists, researchers and libraries. Their goal is illustrated by their mission statement: “To Develop and Promote the Value of Sports Information”. The organization hosts an annual meeting of its members and publishes its newsletter two to three times per year (http://www.iasi.org/publications/newsletter.html). The official site contains links to various online sports databases from around the world as well as links to information technology (IT) aspects of sports such as software.

Chapter 2: Statistical Analyses Research in Traditional Sports

2.1 New-Age Statistical Analysis

History and Inherent Problems of Statistics in Traditional Sports

In the 1800s, Henry Chadwick, a sportswriter and recreational statistician often dubbed as the “Father of Baseball”, developed some of the common baseball statistics such as batting and earned run averages based on experience from the game of cricket. The Batting Average is merely the number of hits a player has accumulated divided by the total number of opportunities (at-bats) a player has received. The Earned Run Average (ERA) indicates how many earned – i.e. those not produced by errors by fielders – a pitcher averages per 9 innings of work. These statistics, in addition to raw batting and pitching numbers, would be the main tools by the general managers of baseball organizations to rate player abilities.

Similarly in sports such as basketball and American football, the traditional statistics were made up of raw data which provided little meaning or insight. In basketball, the main statistics are points, rebounds, assists, turnovers, and the like. Additional value can be obtained by looking at the data relative to statistics such as field goal percentage, fields goals made divided by the number attempted, and assist-to-turnover ratio. For example, a player who scores a lot of points but shoots a low field goal percentage can be said to be inefficient. Likewise, a player who has high assist totals but also commits an excessive

number of turnovers, a player with a low assist-to-turnover ratio, is considered ineffective. In American football, common statistics are merely totals and per-game averages of yards accumulated, touchdowns, receptions, interceptions, and so forth. These types of data independently are not indicators of team and player abilities.

The problem with these traditional forms of statistics is the lack of context with which they are processed. Often they are merely low level data which in and of themselves do not provide great meaning. A baseball player's batting average, for instance, does not take into account the type of hits a player accumulates. For example, is a player who hits three singles, one base reached, in four tries (a batting average of .750) better than a batter who has two homeruns in four tries (a batting average of .500)? The batting average itself would have you believe that the former were far superior to the latter. There are many questions that cannot be answered by traditional statistics and certainly they are not the best methodology for measuring player achievement and effectiveness.

Pioneers of statistical analysis, unknowingly involved in a data mining process, sought to extract deeper knowledge from the statistics of their sports. Since the traditional beliefs within sports organization were questioned, statistical analysis would not be immediately accepted. However, in recent times, it has not only become appreciated by sports organizations but also a staple of their operations.

Bill James

Bill James is widely heralded as one of the foremost statistical analysis pioneers in the sport of baseball. He coined the term sabermetrics – based on the acronym for the Society of American Baseball Research – to represent his mathematical and scientific approach to the collection and processing of baseball data. In 1977, James published the first of many editions of his *Bill James Baseball Abstract*. The works included unique perspectives on the game of baseball as well as its teams and players that were driven by statistics. Since selling only fifty copies of his first edition, the annual *Abstract* eventually gained mass appeal among baseball enthusiasts.

Recently, in 2001, James has published *Win Shares* (James, 2001), a book that offers a new methodology of assessing each player's contribution toward wins. Michael Lewis's bestselling *Moneyball* brought even more attention to James as the book chronicled how his works influenced Oakland management in making their decisions.

In 2003, the Boston Red Sox, whose owner John Henry was a fan, hired Bill James as a consultant. After years as an outsider in the baseball world, the official appointment of James as an expert further cemented the acceptance of statistical analysis in the baseball world. In combination with owners and a general manager, Theo Epstein, who appreciated the value of information held in hard data, the Boston Red Sox would win the World Series becoming champions for the first time in 86 years.

2.2 Major League Baseball Research – Sabermetrics

Origins of the Game of Baseball

Referred to as the “national pastime”, Baseball has been a staple of American culture since the 1800s. Professional baseball emerged after the mid-point of the century, with the National League emerging in 1876. Its counterpart, the American League, was founded 25 years later in 1901. The rival leagues eventually made peace and had their champions duel in the annual World Series. Baseball remains one of the most beloved sports in the United States and has gained popularity throughout the world especially in the Far East and Central and Southern America.

Building Blocks

The basic statistics should not be viewed as determinants of player and team accomplishment but rather as building blocks that can be used to find real and useful knowledge. As far as batting, the fundamental statistics are walks and hits which vary by the number of bases reached – a single resulting in one base and a home run resulting in all four bases reached and a run scored. The On Base Percentage (OBP) was developed to measure a player’s ability to get on base allowing subsequent hitters to drive them in for scores. Similarly, the slugging percentage was developed to reward players who hit for more bases (i.e. doubles, triples, and homeruns). The slugging percentage is merely the total number of bases reached divided by the total number of at-bats.

The value of these two percentages resulted in the creation of the OPS (On-Base plus Slugging) percentage which is merely the combination of both formulas. The OPS is considered one of the best measures of a player's offensive capabilities.

Runs Created

In seeking a better method for assessing a player's ability to manufacture runs for his team, Bill James developed the Runs Created formula in 1982. There are two essential factors in generating runs (Albert, 1994). The sum of the hits and walks amassed by a team reflect the team's ability to get players on base. The total bases reached by a team shows the teams ability to move and score runners who are already on base. These dynamics can be seen in James's formula:

$$\text{RUNS} = \frac{(\text{HITS} + \text{WALKS})\ (\text{TOTAL BASES})}{\text{AT} - \text{BATS} + \text{WALKS}}$$

This formula, according to James, was better suited to measure a particular player's contribution to runs scored than a simple batting average. After all, in order for a team to win against its opposition it must score more runs, not have a higher average.

Win Shares

In his 2002 book of the same name, Bill James introduced the concept of Win Shares. The complex formula takes the number of wins a team has accumulated and awards win shares (a third of a win) to players based on their statistical,

offensive and defensive, contributions. Unlike other ratings which attempt to rate individual player abilities, the Win Share system strives to rate a player's value to the team based on past performance. The formula has been the source of much debate within the sabermetrics community and has seen its share of criticism.

Linear Weights

In 1963, George Lindsey developed a formula which assigned weight (run values) to each at-bat possibility. Through his analysis of past baseball data and probability theory (Albert), Lindsey created the system:

$$\text{RUNS} = (.41)\ 1\text{B} + (.82)\ 2\text{B} + (1.06)\ 3\text{B} + (1.42)\ \text{HR}$$

The formula includes the various types of hits where 1B refers to a single and HR refers to a homerun. However, it does not take into account the other way in which a player can get on base such as a walk (BB) or being hit by a pitch (HBP). Similarly, the formula does not take into account that a player can advance while on base by stealing a base (SB). An adjusted formula includes such factors where predicted runs is equal to

$$.47\ 1B + .78\ 2B + 1.09\ 3B + 1.40\ HR + .33\ (BB + HBP) + .30\ SB - .60\ CS - .25\ (AB - H) - .5\ (OutsOnBase).$$

The linear weights formula is best used to assess a player's offensive impact on the baseball diamond.

Pitching Measures

The basic statistics for measuring a pitcher's performance are the number of wins and Earned Runs Average (ERA). Both are problematic because, like most simple statistics, they lack any contextual perspective. A pitcher may accumulate many wins despite performing poorly if his team is strong offensively. The opposite is also true where a good pitcher may gain many losses due to the ineffectiveness of his team. While the ERA does effectively measure a pitcher's usefulness, it does not provide the context of the benefit a pitcher is provided for a whole season (Albert, 1994). The pitching runs formula developed by Thorns and Palmer places the pitcher's ERA relative to that of all the pitchers in the major leagues. The formula is:

$$\text{Pitching Runs} = \text{Innings Pitched} \times \frac{\text{League ERA}}{9} - \text{ER}$$

where the innings pitched by a pitcher multiplied by league earned run average per 9 innings (a complete game) reflects the average number of runs a pitcher would allow. The earned runs allowed by the pitcher are then subtracted. If the resulting value is 0 then the pitcher is average. A result over 0 shows that the pitcher has performed better than the average pitcher and a result below 0 shows that the pitcher is worse.

2.3 National Basketball Association Research – ABPRmetrics

Similar to the sabermetrics revolution in baseball, basketball – specifically the National Basketball Association (NBA) – has had its own form of statistical

analyses innovation. This movement is often referred to as ABPRmetrics, named for the Association of Professional Basketball Research (ABPR).

ABPRmetrics revolves around the idea that basketball is a team sport in the truest sense of the word and cannot be measured by individual statistics alone. There are many intangibles such as team chemistry – how well certain players perform together – and other factors that are not accounted for in the traditional numbers. According to ABPRmetricians, as analysts are often referred to, the game of basketball must be viewed differently. One of their influences has been the development of statistical analysis based on team concepts such as possessions (when a team has the ball on offense) and how efficiently they are able to score points. In fact, the new statistical analysis has often proven that certain individual players with good stats can actually have negative impacts on their respective teams.

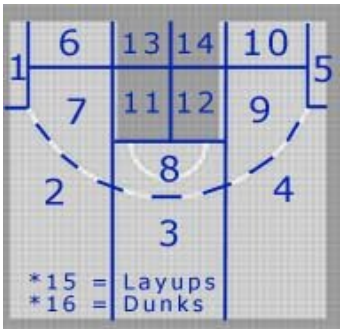
In 2004, then Golden State Warrior center Erick Dampier was having by far the best statistical season of his career. With his contract expiring, the Warriors and many other NBA teams saw Dampier as a must-have free agent (Craggs, 2004). According to the research of free-lance analyst Roland Beech, Dampier was anything but the player his statistics would have you believe. Rather, the team performed significantly worse when he was on the court then when he was not. On the other hand, there are players who seem to make little to no contribution when viewed in light of traditional statistics who can in fact provide great value when on the court.

Beech is the founder of 82games.com, a public online resource featuring various types of NBA statistical analysis and related articles, which went online in the fall of 2003. A basketball outsider, Beech as well as fellow peers provide new and unique insights into basketball players' value and contributions and team performance and efficiency by searching for patterns within raw statistics.

Nearly all NBA organizations currently employee in-house statistical analysts to aid the general manager and coaches in their decisions and some have made a significant impact on the sport in general as has Dean Oliver of the Seattle Supersonics. Oliver's influence within the Sonics organization has developed greater respect within the basketball community for the value of statistical analysis.

Shot Zones

One of the unique ways in which basketball field goal attempts may be viewed and analyzed is based on shot zones, divisions of the half-court area in which offensive possessions occur. 82Games.com used shot data to examine field goal percentage based on shot zones. The court is divided into 16 areas (figure below).



(82Games.com)

The value of this type of analysis is in seeing which areas a particular player may shoot best from. This is has especially great benefit to head coaches who may wish to know where best to use their players on offense. Conversely, by finding out where opposing players shoot the worst, coaches can plan their defensive strategies to try to force players into shooting from locations they are uncomfortable with.

The NBA Shot Zone article analyzes which players had the most field goal attempts and were most accurate at each specific zone. Players with a high number of field goal attempts at a particular zone favor that area. Players with high field goal percentage perform best in certain areas. For example, from the tables below we see that San Antonio Spurs forward Bruce Bowen attempts the majority of his shots from the corner three pointers. During games, it can be seen that when the Spurs are on offense Bowen, who is not a main scoring option for the team, will position himself in one of the corners ready to receive a pass if a teammate is double-teamed or he is left open for the shot. His field goal percentage of .454 from the left corner proves that he has become very efficient and comfortable in this area.

Zone 1: Left corner three-pointers

Most attempts

Tm	Player	FGM	FGA	FG%
SA	Bowen	49	108	.454
TOR	Peterson	38	93	.409
PHO	Johnson	38	81	.481
PHI	Korver	34	77	.442
MIA	E.Jones	29	69	.420

Zone 5: Right corner three-pointers

Most attempts

Tm	Player	FGM	FGA	FG%
PHO	Johnson	46	90	.511
TOR	Marshall	38	90	.422
SA	Bowen	34	89	.382
HOU	Wesley	27	60	.450
NY	Crawford	27	59	.458

(82Games.com)

PER (Player Efficiency Rating)

John Hollinger, a basketball statistical analyst and author of the annual *Pro Basketball Forecast* (Hollinger, 2005), developed the PER as a per-minute rating of a player's effectiveness. This methodology takes into account both the positive contributions of a player as well as the negative impacts. The basic formula, referred to as the unadjusted PER (uPER) is:

```
uPER = (1/MP)*
        [ 3P
        + (2/3)*AST
        + (2 - factor*(tmAST/tmFG))*FG
        + (FT*0.5*(1 + (1 - (tmAST/tmFG)) + (2/3)*(tmAST/tmFG)))
        - VOP*TO
        - VOP*DRBP*(FGA - FG)
        - VOP*0.44*(0.44 + (0.56*DRBP))*(FTA - FT)
        + VOP*(1 - DRBP)*(TRB - ORB)
        + VOP*DRBP*ORB
        + VOP*STL
        + VOP*DRBP*BLK
                -   (PF*((lgFT/lgPF) - 0.44*(lgFTA/lgPF)*VOP)) ]
```

Where:

factor = (2/3) - (0.5*(lgAST / lgFG)) / (2*(lgFG / lgFT))
VOP = lgPTS / (lgFGA - lgORB + lgTO + 0.44*lgFTA)
DRBP = (lgTRB - lgORB) / lgTRB (Basketball-Reference.com)

Plus / Minus Rating

A simple way to measure a player's importance and value to their team is through the on-court vs. off-court plus / minus rating. These ratings merely indicate the team point differentials when a particular player is on the court compared to when the player is off. A positive plus / minus rating indicates that the when the player is on the court the team performs better. Likewise, a negative rating indicates that the team performs better when a player is off the court.

2004-05 Regular Season

#	Player	Team	On Court +/-	Off Court +/-	Net Team +/-
1	Duncan	SAS	+15.1	-1.4	+16.6
2	Kidd	NJN	+4.7	-11.3	+16.0
3	Ginobili	SAS	+14.7	-0.8	+15.5
4	Nowitzki	DAL	+9.3	-6.0	+15.3
5	Nash	PHO	+12.4	-2.6	+15.0
6	Brand	LAC	+2.9	-11.8	+14.7
7	Marion	PHO	+10.1	-4.4	+14.5
8	Prince	DET	+6.9	-6.0	+12.8
9	Marbury	NYK	-0.4	-12.4	+12.0
10	Hamilton	DET	+7.0	-5.0	+11.9

(82Games.com)

Measuring Player Contribution to Winning

While plus / minus ratings give a sense of a player's value to the team when on and off the court, they can be somewhat misleading. The rating can be seen as a comparison of a player's contribution relative to their substitute. Additionally, a player's performance on the court does not occur in a vacuum. Rather, the other

players on the court have a significant impact. The abilities of the players one is playing with as well as those being played against effects the ability of any player to perform.

A more accurate method with which to measure a player's contributions towards victory would be to 'adjust' the plus / minus ratings to take into account the talent of the other players on the court – those on the same team and the opposition (Rosenbaum, 2004). To do this, Dan Rosenbaum, a basketball consultant, developed the following regression:

(1) MARGIN = $\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_K X_K + \varepsilon$, where

MARGIN = 100 * (home team points per possession – away team points per possession)

X_1 = 1 if player 1 is playing at home, = -1 if player 1 is playing away, = 0 if player 1 is not playing

X_K = 1 if player K is playing at home, = -1 if player K is playing away, = 0 if player K is not playing
ε = i.i.d. error term

β_0 measures the average home court advantage across all teams
β_1 measures the difference between player 1 and the reference players, holding the other players constant
β_K measures the difference between player K and the reference players, holding the other players constant

The regression serves to account for the other players who are on the court at the same time as the player being measured as well as the effect of the home court. This formula results in what is referred to by Rosenbaum as 'pure' adjusted plus / minus ratings.

Table 1: Pure Adjusted Plus/Minus Ratings for the Top 20 Players in 2002-03 and 2003-04

Rank	Name		Pure Adj. +/-	
	First	Last	Rating	SE
1	Kevin	Garnett	19.3	3.0
2	Richie	Frahm	17.3	6.3
3	Nenê		11.9	2.7
4	Vince	Carter	11.1	2.5
5	Andrei	Kirilenko	11.1	2.6
6	Dirk	Nowitzki	10.6	2.7
7	Tim	Duncan	10.3	3.3
8	Jason	Hart	10.1	5.6
9	Mike	Sweetney	10.0	5.8
10	Shaquille	O'Neal	9.9	3.0
11	Rasheed	Wallace	9.6	2.1
12	Mickael	Pietrus	9.5	4.3
13	Ray	Allen	9.0	2.3
14	Tracy	McGrady	8.6	2.7
15	Earl	Watson	8.1	4.5
16	Jeff	Foster	7.7	2.7
17	Baron	Davis	7.6	2.5
18	John	Stockton	7.3	5.3
19	Eric	Williams	7.1	2.1
20	Carlos	Arroyo	7.1	5.3

(82Games.com)

The pure adjusted ratings, however, show some strange ratings resulting both from standard error and not enough emphasis on statistical contributions. Quite a few of the players in the top 20 list are players who are not well known and do not play significant minutes to be rated so highly. Rosenbaum developed another formula to link game statistics and the pure adjusted plus / minus ratings.

$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_{13} X_{14} + \varepsilon$, where

Y is the pure adjusted plus minus statistic (the βs from the first regression)

Where X1 through X14 are various basketball game statistics such as points, assists, and the like measured per forty minutes of play. Using the coefficient estimates from the above table, Rosenbaum was able to produce a cleaner version of his adjusted plus / minus ratings which takes into account the statistical contributions of the players.

Table 3: Statistical Plus/Minus Ratings for the Top 20 Players in 2002-03 and 2003-04

Rank	Name		Statistical	
	First	Last	Rating	SE
1	Tracy	McGrady	15.8	2.6
2	Kevin	Garnett	15.3	2.5
3	Andrei	Kirilenko	13.2	2.5
4	Kobe	Bryant	12.6	2.4
5	Shaquille	O'Neal	12.4	2.6
6	Tim	Duncan	12.3	2.5
7	Dirk	Nowitzki	9.5	2.5
8	Elton	Brand	9.1	2.4
9	Arvydas	Sabonis	8.7	3.8
10	Brad	Miller	8.0	2.3
11	Ray	Allen	8.0	2.4
12	Baron	Davis	7.7	2.5
13	Brent	Barry	7.4	2.4
14	Brian	Cardinal	7.1	2.3
15	Paul	Pierce	7.1	2.4
16	Shawn	Bradley	6.9	2.7
17	Jason	Kidd	6.8	2.4
18	Shawn	Marion	6.7	2.2
19	Ron	Artest	6.6	2.3
20	Sam	Cassell	6.5	2.3

(82Games.com)

In addition, Rosenbaum combines the pure adjusted and the statistical plus / minus ratings to develop an ‘overall’ rating which gives a more complete sense of a player’s contributions to a team’s ability to win.

Rating Player Performance in the Clutch

The period in the game, specifically the fourth quarter and overtime, in which the game is in the balance and the teams are close in score is referred to as the clutch. The legends of the NBA often have cemented their legacies by performing best in these situations. For NBA organizations, players who can make plays in the clutch have additional value because more often than not effective performance in the final minutes of the game is the difference between winning and losing.

In a three-part series of articles, the analysts at 82Games.com sought to measure player effectiveness in the clutch. To do this, they defined the "clutch" as the last five minutes of the fourth quarter and the entire overtime period where no team led by more than 5 points. The reason for this choice, rather than stricter settings, was to ensure a large enough sample size with which to measure players. In the third and final article, the authors use the Player Efficiency Rating (PER) described above to assess how NBA players performed during the final period of the game where the game is decided.

Top 20 Players in the Clutch through March 15th of 2004-05 Season

#	Player	Team	FGA	eFG%	FTA	iFG	Reb	Ast	T/O	Blk	PF	Pts	PER*
1	Nash	PHO	21.1	.578	17.8	21%	4.6	17.8	5.3	0.7	4.0	39.5	55.1
2	Stoudemire	PHO	15.4	.800	20.4	60%	14.2	1.2	3.1	2.5	4.3	38.9	53.7
3	Ginobili	SAS	18.6	.450	30.4	60%	8.7	5.6	5.0	1.2	8.7	41.6	49.2
4	Nowitzki	DAL	27.3	.383	24.8	18%	17.1	3.0	2.1	2.6	6.4	42.7	44.7
5	Marion	PHO	18.0	.657	8.7	37%	12.3	0.0	1.5	2.1	5.1	31.3	43.0
6	Duncan	SAS	22.2	.395	15.8	47%	19.3	3.5	1.8	1.2	5.3	29.8	40.9

7	James	CLE	25.3	.415	14.3	30%	7.2	9.1	2.9	1.9	1.9	32.9	38.9
8	Hughes	WAS	20.1	.543	19.3	21%	7.9	8.3	1.3	0.0	4.4	38.1	36.6
9	Francis	ORL	29.9	.411	22.3	36%	9.8	4.9	2.6	0.8	9.4	43.5	36.2
10	Stackhouse	DAL	18.6	.522	20.3	30%	8.1	2.4	1.6	0.0	3.2	34.8	36.1
11	Kidd	NJN	21.7	.565	11.3	15%	7.1	9.0	2.8	0.0	2.4	33.5	35.8
12	Lewis	SEA	20.0	.620	8.2	19%	10.0	1.7	1.3	1.3	4.3	30.0	35.6
13	Griffin	MIN	16.3	.722	4.5	33%	15.3	3.6	0.9	5.4	5.4	27.1	35.6
14	O'Neal	IND	30.3	.500	17.5	34%	17.5	3.5	2.9	3.5	6.4	43.1	35.1
15	Gordon	CHI	27.7	.607	14.3	21%	5.9	2.0	2.0	0.0	5.4	45.5	34.1
16	Terry	DAL	15.0	.692	9.3	30%	1.7	4.6	1.2	0.6	3.5	28.9	33.0
17	Allen	SEA	23.3	.451	12.8	25%	6.4	4.1	1.4	0.0	1.8	32.0	31.8
18	Camby	DEN	9.8	.625	6.7	37%	13.5	1.8	0.6	6.1	4.9	18.4	31.8
19	Boykins	DEN	15.5	.643	15.0	21%	3.3	5.0	2.2	1.7	7.8	32.7	31.7
20	Marshall	TOR	16.5	.681	6.0	19%	16.1	0.0	0.9	1.8	6.4	27.1	30.7

(82Games.com)

In measuring a player's performance in the clutch, it is critical to take into account all aspects of the game. The above PER measure accounts only for the offensive contributions of the respective players. An alternative measure would be needed to gauge the defensive impact a player has during the critical time in the game. One way to do this is to measure the PER ratings of a particular player's counterpart on the other team. The basic idea is that if a player performs well defensively in the clutch, their counterpart should have a poor efficiency rating. The assumption is that a point guard will cover the opposing point guard and so on. While this method is faulty – it does not take into defensive switches, zones, players guarding different positions – it nonetheless gives a valuable perspective.

Top 20 Player "Counterpart" Defenders in the Clutch through March 15th of 2004-05 Season

#	Player	Team	FGA	eFG%	FTA	iFG	Reb	Ast	T/O	PF	Pts	PER*
1	Wade	MIA	12.4	.258	5.3	33%	5.3	3.0	2.6	6.4	9.8	1.4
2	Claxton	GSW	16.6	.217	10.5	23%	7.2	3.9	2.8	9.4	13.8	2.5
3	Chandler	CHI	19.9	.282	7.7	38%	10.2	1.5	1.0	10.2	15.3	2.8

4	Hassell	MIN	18.5	.210	6.6	25%	7.1	4.2	1.2	3.0	11.9	2.8
5	Brand	LAC	14.5	.220	9.6	36%	9.0	1.7	0.9	7.3	13.6	3.5
6	Gasol	MEM	13.2	.400	8.8	26%	10.6	0.9	5.3	12.3	15.0	3.9
7	Camby	DEN	15.4	.240	5.5	48%	9.2	3.1	2.5	8.6	11.1	3.9
8	Davis	BOS	14.2	.282	4.6	20%	7.0	4.6	2.6	4.1	11.6	4.7
9	Jeffries	WAS	17.0	.350	8.5	5%	5.1	2.5	0.8	7.6	17.0	4.8
10	Deng	CHI	12.7	.395	4.0	5%	6.7	2.7	2.7	8.0	12.7	4.9
11	Jackson	HOU	9.3	.167	6.2	25%	8.5	2.3	0.8	3.1	8.5	5.1
12	Anthony	DEN	15.0	.212	4.0	34%	8.6	4.6	2.3	4.0	9.8	5.1
13	Maggette	LAC	13.1	.341	8.0	19%	6.1	3.8	3.2	7.0	15.6	5.4
14	Iverson	PHI	14.9	.282	8.0	10%	5.3	5.3	6.1	7.2	16.0	5.6
15	Van Horn	MIL	17.5	.375	2.6	35%	7.9	0.0	0.9	4.4	14.9	6.6
16	Carter	NJN	15.3	.310	8.7	33%	8.7	1.5	1.5	7.3	15.3	6.7
17	Wesley	NOH	15.9	.333	6.0	12%	5.3	6.0	2.7	4.0	13.9	6.7
18	Hinrich	CHI	14.2	.284	6.2	16%	7.3	6.9	1.9	7.7	11.9	6.9
19	D.Jones	MIA	16.2	.380	3.6	24%	4.2	4.2	2.9	5.2	15.6	6.9
20	Miller	IND	12.5	.385	5.3	15%	5.8	1.9	3.8	4.8	13.5	7.1

(82Games.com)

The final element missing in order to develop an overall rating for players in the clutch would be to rate their respective team performance in that period. An easy and effective way to measure the team's performance in the clutch would be to use the plus / minus rating. By measuring the team's point differential when a particular player is in the game, we get a sense of whether or not a player is contributing to the team's ability to outscore their opponent. To account for the varying amount of clutch minutes measured for each player, the plus / minus rating is evaluated per-48 minutes.

The final formula for an overall player clutch rating developed by the analysts at 82Games is the difference between the offensive and defensive PER plus 1/5 times the per-48 minute plus / minus rating. The assumption is that because there are five players on the court at a time for a team, a particular player should

only be awarded one-fifth of the point differential when they are on the court to account for the contributions of the other players.

Top 20 Overall Player Clutch Rating through March 15th of the 2004-05 Season

#	Player	Team	PER	dPER	Diff	+/-	Rating
1	Ginobili	SAS	49.2	11.3	37.9	+9	39.0
2	Stoudemire	PHO	53.7	19.5	34.2	+16	36.2
3	Nash	PHO	55.1	23.8	31.3	+31	35.4
4	Wade	MIA	29.2	1.4	27.8	+79	33.7
5	Nowitzki	DAL	44.7	16.0	28.7	+43	32.3
6	Camby	DEN	31.8	3.9	27.9	+22	30.6
7	Griffin	MIN	35.6	15.8	19.8	+58	30.2
8	James	CLE	38.9	12.4	26.5	+13	27.8
9	Stackhouse	DAL	36.1	9.2	26.8	+5	27.6
10	Allen	SEA	31.8	10.0	21.7	+58	27.0
11	Gordon	CHI	34.1	10.1	24.0	+26	26.5
12	Terry	DAL	33.0	12.8	20.2	+52	26.2
13	Hughes	WAS	36.6	14.6	22.0	+42	25.7
14	Lewis	SEA	35.6	15.4	20.2	+59	25.4
15	O'Neal	IND	35.1	14.6	20.5	+18	22.6
16	O'Neal	MIA	26.0	9.7	16.3	+67	21.4
17	Daniels	SEA	30.5	16.6	13.9	+65	21.4
18	Kidd	NJN	35.8	18.9	17.0	+38	20.5
19	Tinsley	IND	22.4	7.9	14.5	+44	19.9
20	Ridnour	SEA	28.0	12.2	15.9	+22	19.2

(82Games.com)

2.4 Emerging Research in Other Sports

Other leagues, such as the NFL, have only yet begun to have the type of analysis and research that has become a staple of the MLB and NBA.

NFL DVOA

The common argument used to explain the fact that statistical analysis has yet to have the impact in the NFL as it has had in other sports is the relative lack of

statistics, especially when measuring individual players. Baseball and basketball have more individual statistics collected. Furthermore, the NFL season is only 16 games paling in comparison to the many games in MLB and NBA seasons. The importance is to treat individual plays in each NFL game as a unique event.

The Football Outsiders, authors of the annual *Pro Football Prospectus* (Schatz, 2006) and whose website (http://www.footballoutsiders.com) applies unique statistical analyses to football, developed the Defense-adjusted Value Over Average formula (DVOA) which measures the success offensive players have had in specific situations relative to league averages. This is done by analyzing each individual play throughout an NFL season. If two running backs run for three yards are their performances equal? Not necessarily. Assume running back A ran for three yards on first down and 10 yards to go accomplishing a fairly average, if not below average, result. Running back B, on the other hand, gains 3 yards on third down with 2 yards to go thus gaining a new first down. The second running back, while gaining the same amount of yards, has a made a greater contribution to his team than the first.

The first part of the formula is the Value Over Average (VOA) measures the success an offensive player had in a particular situation compared to the league averages of other offensive players in the same situation. The value of a player's performance is measured by the total number of yards gained as well as the number of yards towards a first down. In football, it is critical to maintain drives (offensive possessions), which is done by being able to consistently gain first

downs marching towards the end zone and avoiding fourth down. Based on the research of Bob Carroll, Pete Palmer, and their counterparts in their book *The Hidden Game of Football* (Carroll et al, 1998), the system considers a play successful if 45% of the needed yards are gained on 1st down, 60% on second down, and all the needed yards are gained on either third or fourth down. Successful plays are given one point with additional points awarded for the bigger (more yards gained) a play is.

An additional element that needs to be considered is how good the opposing defense is since not all teams play the same schedule or the same level of opponents. Hence, each situation needs to be measured against the defensive ability of the other team. Adjusting the VOA based on the defense's average success in stopping similar plays throughout the season results in the DVOA. Additionally, the DVOA of the entire team can be used as a way to measure an entire team's effectiveness on offense, defense, and special teams.

NFL Team Efficiency Ratings (DVOA) for the 2004 Season

	TEAM	TOTAL DVOA	LAST YEAR	NON-ADJ TOTAL VOA	W-L	OFFENSE DVOA	OFF. RANK	DEFENSE DVOA	DEF. RANK	SPECIAL DVOA	S.T. RANK
1	NE	35.6%	2	32.1%	14-2	26.3%	4	-9.1%	6	0.2%	16
2	PIT	35.4%	16	36.3%	15-1	16.6%	7	-15.0%	4	3.8%	7
3	IND	34.7%	4	37.0%	12-4	38.9%	1	2.3%	18	-1.8%	22
4	PHI	28.7%	9	24.0%	13-3	17.7%	6	-3.5%	13	7.4%	3

5	BUF	28.6%	22	30.4%	9-7	-5.1%	21	-24.5%	1	9.2%	1
6	NYJ	23.8%	18	26.7%	10-6	23.4%	5	2.8%	19	3.1%	11
7	DEN	19.6%	12	18.7%	10-6	8.2%	10	-14.2%	5	-2.8%	23
8	SD	19.2%	25	22.0%	12-4	16.4%	8	-6.6%	11	-3.8%	29
9	BAL	18.9%	6	12.7%	9-7	-2.5%	16	-16.6%	2	4.7%	5
10	KC	12.1%	1	3.8%	7-9	29.0%	2	16.2%	28	-0.8%	19
11	CIN	9.8%	23	-0.1%	8-8	7.6%	11	1.1%	17	3.2%	10
12	CAR	4.2%	20	8.9%	7-9	0.1%	15	-7.1%	10	-3.0%	25
13	MIN	1.5%	10	4.6%	8-8	28.0%	3	22.8%	31	-3.7%	28
14	JAC	-0.1%	19	-3.9%	9-7	-7.5%	25	-8.5%	8	-1.0%	20
15	HOU	-1.2%	29	-5.2%	7-9	0.2%	14	-2.4%	14	-3.7%	27
16	TB	-2.0%	8	-0.5%	5-11	-6.1%	23	-8.8%	7	-4.7%	31

(FootballOutsiders.com)

NCAA Score Card and Dance Card

Using analytic software from SAS, Jay Coleman – a professor of operations management at the University of North Florida – and Allen Lynch – a professor of Economics at Mercer University – have developed a formula to predict “at-large” teams for the annual NCAA tournament. This system is called the “Dance Card” and has a 94 percent prediction success rate. The pair has also developed a formula for predicting the outcomes of the NCAA tournament games called the “Score Card”. This system has a 75 percent success rate in predicting winners.

The exact formulas for the Dance Card and Score Card have not been made public. The Score Card formula was devised using the results of the NCAA tournament games from 2001 through 2004 (256 games), around 50 pieces of

information for the two teams involved in each game such as records and RPI. There are four statistics which are viewed as very valuable by the system:

- **Ratings Percentage Index (RPI)**

  The RPI is a rating system used by the NCAA to rank its basketball teams. It is comprised of three facets. The first is the team's winning percentage which covers 25% of the formula. The second and highest weighted facet is the average opponents' winning percentage which is half of the formula. The third part of the formula, weighted at 25%, is the opponents' opponents' winning percentage. Hence, the RPI not only takes into how well a team is winning but also places great emphasis on the quality of the opponents being played through their relative win percentages.

- **Conference Ranking**

  The ranking of the conference a team comes from gives a sense of the value of the competition that a team faces on a regular basis. A team that plays in a highly-ranked ranked and, thus, difficult conference is considered more battle-tested than a team that plays in a relatively weak conference. This value is measured using the non-conference RPI ranking of each conference. In other words, the performance of the conference's teams when

playing teams from other conferences determines the ranking of that conference.

- **Regular Season Conference Championship**

  Whether a team has won its regular season championship is a factor that shows how well a team performed against its conference opponents. Teams in the same conference often are very familiar with one another and, hence, being able to win a conference shows the ability of a team to be successful.

- **Win Percentage Against Non-Conference Opponents**

  Conversely to the ability to win against teams within the same conference, the winning percentage of a team against non-conference opponents is also particularly important, especially when the non-conference schedule is fairly difficult.

Analysis of various statistics and data is a fundamental aspect of sports data mining. This analysis provides new ways with which to view and measure sports team and player performance.

Chapter 3: Tools for Sports Data Analysis

3.1 Data Mining Tools

There are yet to be a wide variety of commercial products which are advertised as data mining tools for sports uses. Currently, most sports organizations that are interested in data mining applications to their respective sport do their analysis in-house. With the prominence that data mining has gained in other fields and especially with the increase in public awareness of its usefulness, there will no doubt be an increase in third-party companies seeking to apply data mining to sports for commercial purposes.

Advanced Scout

Developed in partnership with IBM, the Advanced Scout program was designed to aid NBA coaches in identifying hidden patterns in game statistics and data. Since the mid-90s, Advanced Scout (AS), which uses business intelligence and data mining techniques, has provided NBA teams with insights that may have otherwise gone unseen. The application has two data sources, one structured – event data from a courtside collection system – and the other unstructured – multimedia data from NBA game tape. The program can be used by coaches to prepare for upcoming opponents as well as to analyze team post-game performance.

The raw data that will be processed by the application is collected by a specially-designed system. The types of data include events such as which player took the

shot, the outcome, possible resulting rebounds, and the like associated with time codes (Bhandari et al, 1996). In terms of pre-processing, the data is cleaned by AS through a series of consistency checks that reduces, if not clears, the number of errors made during data collection. Data collection errors include missing or impossible events. If two shots are taken consecutively without any player being given credit for a rebound, the application assumes that the player who took the first shot recovered his own rebound and shot again. If this is not believed to be the case, the game footage can be used as verification.

Once the data has been cleaned, Advanced Scout reformats the events from the raw data into a play sheet – a list of events listed in chronological order based on the game time. The play sheets are especially useful for coaches as they provide a snapshot of the game. In the final stage of the transformation process the individual events are grouped into possessions. As seen in the previous chapter, statistical analysis has shown that those teams that make the most efficient use of their possessions are nearly always the victors.

The last stage of the pre-processing is data enrichment where additional value is given to the data through the use of inference rules and supplementary data. Advanced Scout uses information from a player-role table – e.g. power forward, center – to make inferential analyses concerning player-role relationships.

An NBA coach is able to use Advanced Scout to run data mining queries to find insightful patterns relating to shooting performance or for possession analysis which is useful in determining best player combinations on the court. AS uses a

data mining technique known as Attribute Focusing (AF) where “an overall distribution of an attribute is compared with the distribution of this attribute for various subsets of data” (Bhandari et al, 1996). In other words, for any focus attribute if there exists a subset of data with a distinctively different distribution than the overall distribution then that subset is possibly indicative of an interesting pattern.

Patterns marked by AS are presented to the user, presumably a coach, in two forms – a text description and a graphical representation, of which a sample has not been made publicly available. The common text description example used states:

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.

While this illustration is dated, the players referred to have been retired from the NBA for nearly a decade, it shows the user-friendly way the information is presented to users who are most likely not very computer-proficient. It is now up to the coach (domain expert) to determine why this pattern occurred. Relating to the past example, it was found that when Mark Price was double-teamed by the

opposing New York Knicks, he was able to find forward John “Hot Rod” Williams for wide-open jump shots.

In addition to its data mining functionalities, Advanced Scout can be used as a traditional query and reporting tool by coaches and general managers. Also, since AS provides a video time stamp for each identified pattern, coaches have the ability to view those sections of the tape to gain a clearer picture of and apply their own knowledge to the events in question.

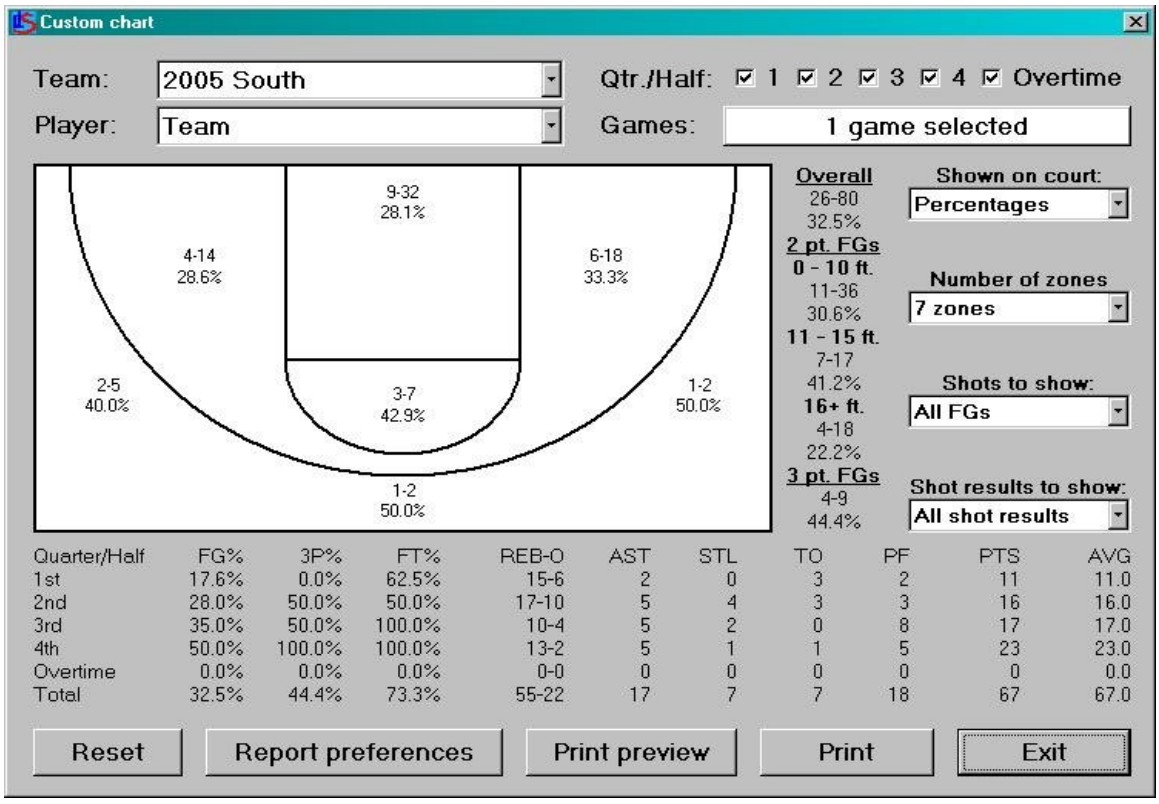
In 1997, Dr. Inderpal Bandari, developer of the Advanced Scout program while at IBM, founded the company Virtual Gold which now controls the application. Virtual Gold has developed similar applications for other sports leagues such as Major League Baseball and the Professional Golf Association (PGA).

3.2 Scouting Tools

Digital Scout

Digital Scout provides scouting software and tools for baseball, basketball, and football among others. The software can be used to collect and analyze statistics as well as to generate various reports. For example, baseball reports include custom hit charts for batters and pitchers. While basketball reports include player combination reports in addition to player and team shot charts.

Custom Chart – Shooting Percentages by Zone



(DigitalScout.com)

Digital Scout products are extremely popular and their baseball tool is used by Team USA as well as the Little League World Series. Additionally, the software has used by a number of NBA teams (Dudzik, 2000). However, because their software and tools are relatively inexpensive, any team can afford to have this valuable tool.

Inside-Edge

Inside-Edge is widely acclaimed for its exceptional baseball services and solutions which include various tools and reports. A number of MLB teams and media use their scouting service – six consecutive baseball World Champions utilized them significantly. Additionally, Inside-Edge sells scouting data collection

and analysis software including its PalmScout tool. Designed to have the same look, feel, and portability of a PDA, the PalmScout has won several awards.

Included in the types of reports provided by Inside-Edge to its major league clients are hitter and pitcher profiles which give an overall perspective of a player's strengths, weaknesses, tendencies, and the like. For example, a batter report can be used by a manager or pitcher when devising a strategy when facing a particular batter. Also, these reports have been known to be used by players who wish to gain a better understanding of themselves and how to improve on their weaknesses. The Batter report is divided into three subsections:

Edge Notes

This section provides an overall description of a player's tendencies as well as their strengths and weaknesses backed by statistical data (batting percentages). It can be seen from the example that Carlos Delgado struggles with curveballs and will take many (64%) for strikes. Hence, a pitcher armed with this knowledge will have a significant advantage when facing this batter.

Edge Notes from Batter Profile for Carlos Delgado

(Inside-Edge.com)

Pitch Tips

Inside Edge

Hitter Profile

Delgado, Carlos

Brewers vs. RHPs

Bats: Left
Stance: N.A.

Pitches Charted: 5688
Pitches per Pl. App.: 3.98

EDGE NOTES

Covering Scouted Games from 3/31/01 to 6/23/05

1st Pitch
- He doesn't chase 1st pitch Curves (6%)

Overall Tendencies
- Chases Splitters (35%)
- He's taken a lot of Curves (64%) for strikes.

Strengths
- Fastballs (.338), espec. early (.416) & when he's ahead (.471) & with runners on base (.357)
- Changeups early (.471) & when he's ahead (.485)
- Splitters early (.350) & when he's ahead (.500) & with runners on base (.345)

Weaknesses
- Curves (.137), espec. early (.167) & with 2 strikes (.127) & when he's behind (.091)
- Sliders with 2 strikes (.158) & when he's behind (.152)
- Changeups with 2 strikes (.145) & when he's behind (.145)
- Splitters with 2 strikes (.138) & when he's behind (.151)

The Pitch Tips section of the batter profile gives the pitcher a sense of which pitches are best in which situations. For example, we can see from the chart below that the best pitch to throw first for a strike is a curve ball as Delgado has not seen this often (4 attempts) and was unable to get a hit on any.

Pitch Tips from Batter Profile for Carlos Delgado

Pitch TIPS	Scouted BAs	All Counts	First Pitch	Early Counts	Two Strikes	Hitter Ahead	Hitter Behind	With RISP	Chase% (balls)	Take% (strikes)
	Fastballs	.338 250/739	.430 58/135	.416 123/296	.210 67/319	.471 81/172	.218 62/285	.357 66/185	21%	30%
	Curves	.137 14/102	.000 0/4	.167 3/18	.127 10/79	.250 2/8	.091 6/66	.182 4/22	19%	64%
	Sliders	.211 23/109	.333 3/9	.321 9/28	.158 12/76	.400 4/10	.152 10/66	.200 6/30	27%	40%
	Changeups	.305 47/154	.550 11/20	.471 32/68	.145 10/69	.485 16/33	.145 9/62	.303 10/33	26%	26%
	Splitters	.241 21/87	.429 3/7	.350 7/20	.138 8/58	.500 6/12	.151 8/53	.345 10/29	35%	30%

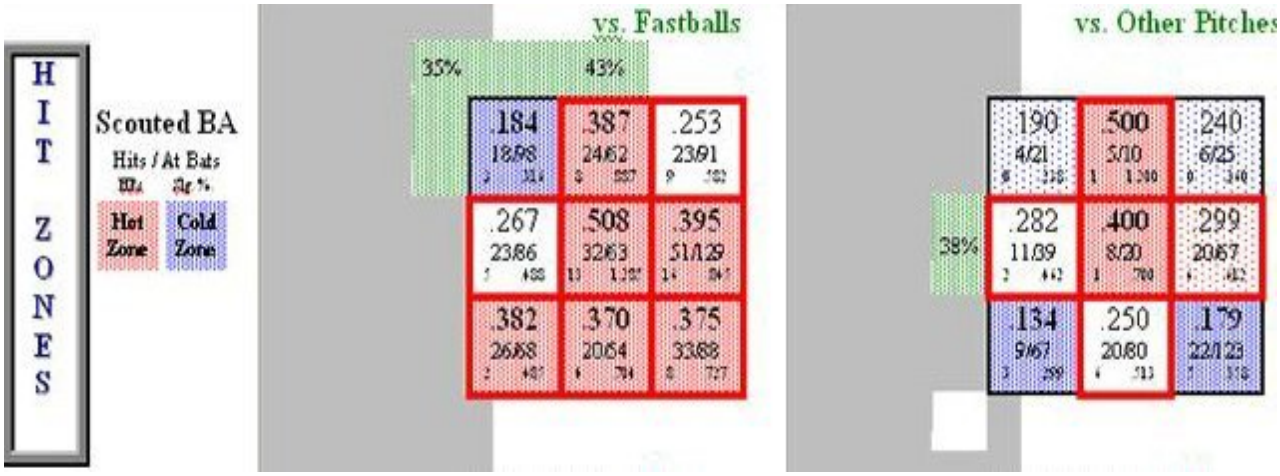
Very High / High / Avg. or unsure / Low / Very Low

Hit Zones

The Hit Zone analysis divides the strike zone – the area in which a pitcher must throw the ball over the home base in order to get a strike call – into nine subsections. The graphical representation, which is based on statistical

performance, allows pitchers to visually grasp which areas of the strike zone a player is best at.

Hit Zone Performance from Batter Profile for Carlos Delgado



(Inside-Edge.com)

3.3 Simulation Software

BBall

The BBall basketball simulation software developed by Bob Chaikin, a consultant for the Miami Heat and basketball researcher, was designed specifically for NBA coaches, general managers, directors of player personnel, and scouts to evaluate their teams and players using statistics to simulate games. Some of the uses are:

- *Determining a team's optimum substitution pattern*

  NBA personnel can simulate entire seasons using various substitution patterns to find the one that results in the most wins for the team.

- *Whether signing a free agent or making a trade will beneficial to the team*

  Trades and free-agent acquisitions can be done in the software and games can be simulated with the potential roster changes. Depending on whether the team performs better or worse in simulations, general managers can gauge whether an acquisition would be beneficial to the team.

- *How valuable a player is to the team*

  The simulation software can be used to simulate how a team would perform without a particular player, thus showing that player's value to the team and its ability to win.

- *Determining which factors are needed by the team to improve performance*

  With the ability to simulate entire NBA seasons and view a variety of resulting statistics, a general manager can view what dynamics (scoring, rebounding, etc.) their team lacks and is in need of. Hence, the team's strategy in obtaining new or additional players can be directed towards those needs.

Chaikin, whose software is used by many NBA teams, developed the simulation program which applies APBRmetric principles to statistics (provided in the form

of historical databases). By allowing users to simulate many games – entire seasons – in a number of minutes, BBall is a powerful tool for NBA personnel.

3.4 Baseball Hacks

In his book *Baseball Hacks: Tips & Tools for Analyzing and Winning with Statistics* (Adler, 2006), Joseph Adler presents a series of techniques with which anyone can analyze baseball statistics. Adler targets passionate baseball fans especially those who take part in fantasy leagues. The tools used to do the analysis are primarily open-source and, hence, free for use. Additionally, Adler provides detailed instruction as well as sample scripts and code that allow even novices to do their own analysis.

Historic and Current Baseball Database Set Up

In order to be able to do statistical analysis, the right tools must be available. The backbone of any data mining endeavor is the DBMS (Database Management System) which stores the data to be analyzed. The purchase of a commercial DBMS such as Oracle would not fit into the budget of most individuals and would be impractical. To this point, Adler begins by showing users how to download and install MySQL, a free DBMS, and by also providing instruction to those who have Microsoft Access, which is available with the Professional version of Microsoft Office. Additionally, the book shows readers how to download free baseball player and team statistic databases from past seasons online for

MySQL (http://www.baseball-databank.org) as well as Access (http://www.baseball1.com).

As for current baseball statistics while a season is still being played, Adler provides methods for the extraction of statistics from online sources such as MLB.com, Major League Baseball's official site. Using the fairly new web query feature with Microsoft Excel, the popular spreadsheet program, up-to-date player and team statistics can be downloaded directly into the spreadsheet in order to be manipulated and analyzed.

Statistics Visualization

The average person probably has never heard of R, an open-source statistical and graphical programming language and software environment. In fact, most would be intimidated by the previous description itself. Adler presents straightforward and uncomplicated ways with which to use R's intuitive nature to analyze baseball statistics and view graphical representations in the form of spray charts and data cubes. For those less adventurous, instruction is provided for the creation of graphs and charts using Microsoft Excel.

Sabermetric Analysis

The earlier sections of this work helped introduce the concept of Sabermetrics and the various formulae which have been developed by its advocates. Adler not only presents many of these same formulae and concepts but also provides

practical ways, using SQL code, to calculate them using the player and team databases. In addition to common Sabermetric computations, such as measurements such as batting or pitching using linear weights discussed previously (2.2 – Linear Weights), Adler also provides techniques for finding "clutch" players and measuring the odds of the best regular season team winning the World Series.

Baseball Hacks provides any baseball fan the technical understanding necessary to do the same types of statistical analysis that professional baseball organizations have become famous for. Likewise, the various tools that have been summarized above provide users with similar capabilities for their respective sport.

Chapter 4: Prediction Research for Traditional Sports and Horse / Dog Racing

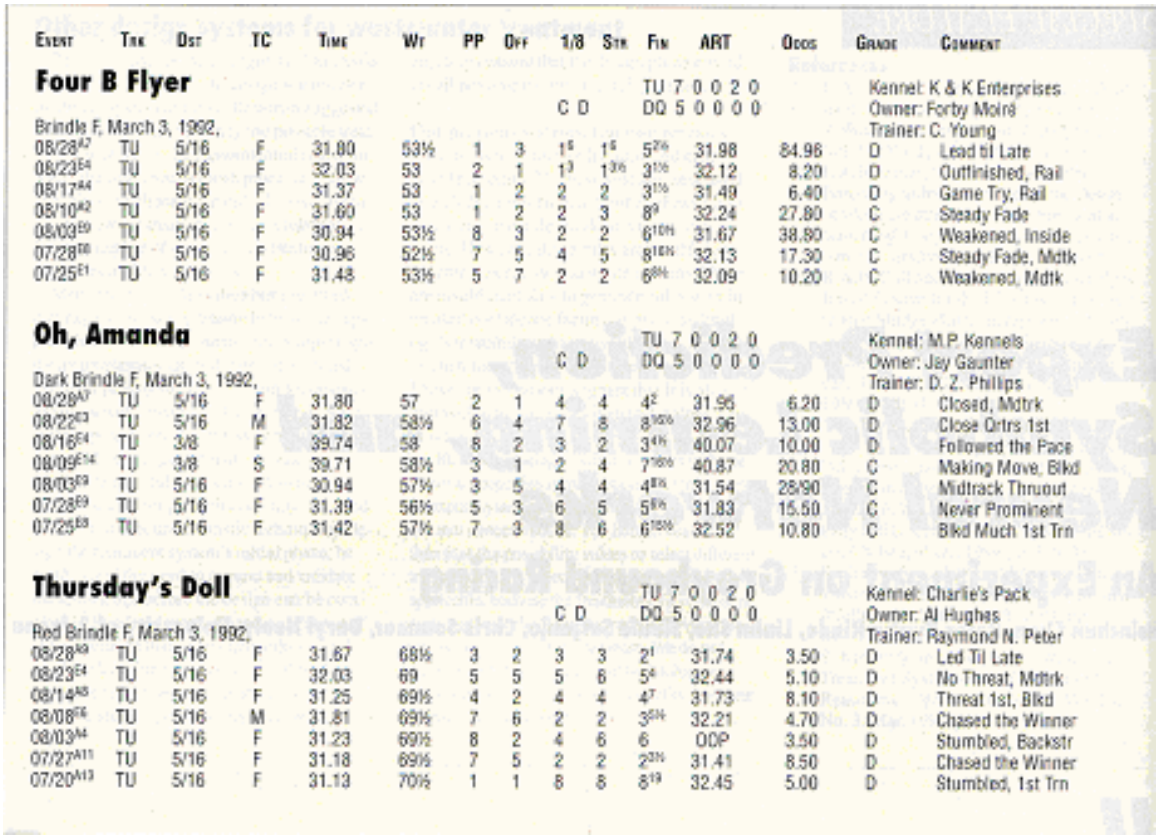
4.1 Case Study: Greyhound Racing

Data mining techniques can be used for predictive purposes in sports. One way to do this is to use machine learning, which covers a variety of solutions such as decision trees, production rules, and neural networks, to make predictions based on the hidden information within data (Chen et al, 1994). In their paper *Expert Prediction, Symbolic Learning, and Neural Networks*, Dr. Hsinchun Chen and his counterparts test the predictive capabilities of machine learning against those of human experts in greyhound racing.

The two machine learning techniques used in the study are ID3, a decision-tree building algorithm, and backpropogation, a neural network learning algorithm. An artificial neural network, to put it as simply as possible, is a non-linear data modeling tool that can be used to find hidden patterns and relationships within data. The data set used revolved around information available to bettors consisting of each dog's historical performance records. Additionally, each race program contained information about each of the eight dogs competing such as their fastest time, total number of races, as well as the amount of first, second, third, and fourth place finishes. Also, the programs displayed a detailed view of the last seven races the dog had competed in. These listings showed the starting and finishing position of the dog and its position in the first (break), second, and

third turn. Finally, the dog's total race time and the grade of the race (indicating its competitiveness) are given.

Figure: Sample Greyhound Racing Program



Event	Trk	Dst	TC	Time	Wt	PP	Off	1/8	Str	Fin	ART	Odds	Grade	Comment
Four B Flyer										TU 7 0 0 2 0				Kennel: K & K Enterprises
								C D		DQ 5 0 0 0 0				Owner: Forby Moira
Brindle F, March 3, 1992,														Trainer: C. Young
08/28[A7]	TU	5/16	F	31.80	53½	1	3	1^{5}	1^{5}	$5^{2¼}$	31.98	84.90	D	Lead til Late
08/23[E4]	TU	5/16	F	32.03	53	2	1	1^{3}	$1^{3¼}$	$3^{1½}$	32.12	8.20	D	Outfinished, Rail
08/17[A4]	TU	5/16	F	31.37	53	1	2	2	2	$3^{1½}$	31.49	6.40	D	Game Try, Rail
08/10[A2]	TU	5/16	F	31.60	53	1	2	2	3	8^{9}	32.24	27.80	C	Steady Fade
08/03[E9]	TU	5/16	F	30.94	53½	8	3	2	2	$6^{10¾}$	31.67	38.80	C	Weakened, Inside
07/28[E6]	TU	5/16	F	30.96	52½	7	5	4	5	$8^{16¾}$	32.13	17.30	C	Steady Fade, Midtk
07/25[E1]	TU	5/16	F	31.48	53½	5	7	2	2	$6^{8¾}$	32.09	10.20	C	Weakened, Midtk
Oh, Amanda										TU 7 0 0 2 0				Kennel: M.P. Kennels
								C D		DQ 5 0 0 0 0				Owner: Jay Gaunter
Dark Brindle F, March 3, 1992,														Trainer: D. Z. Phillips
08/28[A7]	TU	5/16	F	31.80	57	2	1	4	4	4^{2}	31.96	6.20	D	Closed, Midtrk
08/22[E3]	TU	5/16	M	31.82	58½	6	4	7	8	$8^{16¾}$	32.96	13.00	D	Close Qrtrs 1st
08/16[E4]	TU	3/8	F	39.74	58	8	2	3	2	$3^{4½}$	40.07	10.00	D	Followed the Pace
08/09[E14]	TU	3/8	S	39.71	58½	3	1	2	4	$7^{16¾}$	40.87	20.80	C	Making Move, Blkd
08/03[E9]	TU	5/16	F	30.94	57½	3	5	4	4	$4^{8¾}$	31.54	28/90	C	Midtrack Thruout
07/28[E9]	TU	5/16	F	31.39	56½	5	3	6	5	$5^{8¾}$	31.83	15.50	C	Never Prominent
07/25[E8]	TU	5/16	F	31.42	57½	7	3	8	6	$6^{15¾}$	32.52	10.80	C	Blkd Much 1st Trn
Thursday's Doll										TU 7 0 0 2 0				Kennel: Charlie's Pack
								C D		DQ 5 0 0 0 0				Owner: Al Hughes
Red Brindle F, March 3, 1992,														Trainer: Raymond N. Peter
08/28[A9]	TU	5/16	F	31.67	68½	3	2	3	3	2^{1}	31.74	3.50	D	Led Til Late
08/23[E4]	TU	5/16	F	32.03	69	5	5	5	6	5^{4}	32.44	5.10	D	No Threat, Midtrk
08/14[A5]	TU	5/16	F	31.25	69½	4	2	4	4	4^{7}	31.73	8.10	D	Threat 1st, Blkd
08/08[E6]	TU	5/16	M	31.81	69½	7	6	2	2	$3^{5¾}$	32.21	4.70	D	Chased the Winner
08/03[A4]	TU	5/16	F	31.23	69½	8	2	4	6	6	ODP	3.50	D	Stumbled, Backstr
07/27[A11]	TU	5/16	F	31.18	69½	7	5	2	2	$2^{3¾}$	31.41	8.50	D	Chased the Winner
07/20[A13]	TU	5/16	F	31.13	70½	1	1	8	8	8^{19}	32.45	5.00	D	Stumbled, 1st Trn

Based on the opinions of regular bettors, track experts, and park management, the researchers chose the performance variables believed to be the most foretelling of future performance. The resulting ten attributes recommended by the experts were:

- Fastest Time: The fastest time in seconds for a 5/16 mile race.

- Win Percentage: The number of first place finishes divided by the total number of races.
- Place Percentage: The number of second place finishes divided by the total number of races.
- Show Percentage: The number of third place finishes divided by the total number of races.
- Break Average: The average dog's position during the first turn for the seven most recent races.
- Finish Average: The average finishing position for the seven most recent races.
- Time 7 Average: The average finishing time for the seven most recent races.
- Time 3 Average: The average finishing time for the three most recent races.
- Grade Average: The average grade of the seven most recent races the dog competed in.
- Up Grade: Weight given to a dog when dropping down to less competitive racing grade.

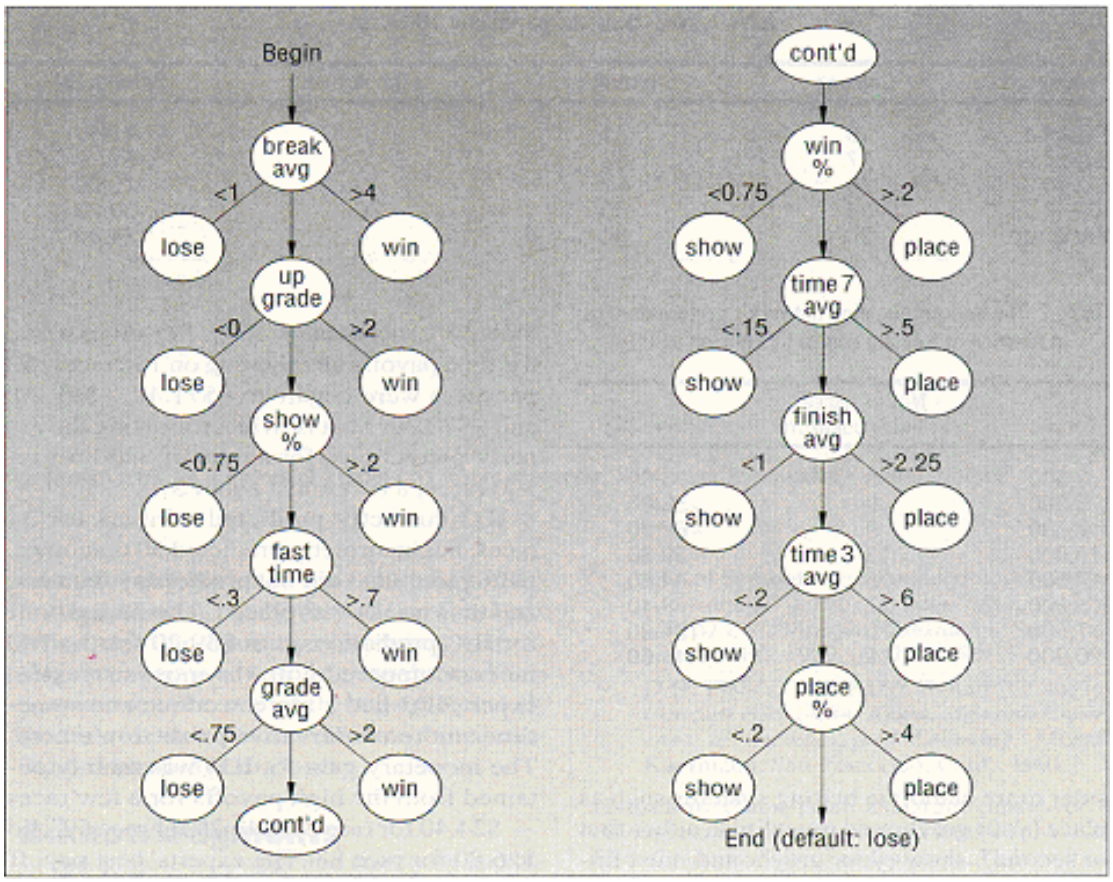
To ensure that the algorithms could properly value the race grade attribute, each race type is assigned a weight. Races of grade A (the most competitive) down to those of grade D (the least competitive) received values of four down to one respectively. Other races, such as those for training purposes, received a value

of zero. Similarly, race times for each dog are relative to the other dogs competing in that race. Dogs in higher race grades will often have faster overall times than those in lower graded races. Hence, relative scaling of the race times was done by assigning the slowest time a value of zero and each other dog the difference between its time and the slowest time. For example, if the lowest time for the race was 32.00 seconds and another dog in the same race finished in 31.00 seconds then that dog would receive a value of 1.00.

For this study, the researchers used two-thirds of the total data – 200 races and 1600 greyhounds – for algorithm training purposes and the remaining one-third – 100 new races and 800 greyhounds – in order to test the predictive capabilities of the algorithms.

The ID3 algorithm employed in the investigation used the attribute values, such as fastest time and others, to classify each greyhound as a winner or loser.

Figure: ID3 Decision Tree



The predictions made by the algorithm were tested by comparing to those of the track experts. If an expert or the algorithm predicted a winner, $2.00 was wagered on that dog. If an algorithm predicted multiple winners, a bet was placed for each predicted winner. No bet was placed in the case that no prediction was made. The payoff odds given in the result sheets were used in order to calculate the total winnings of each expert in addition to the two algorithms. Only payoffs for first place finishes were considered. The table below summarizes the results of the experiment based on 100 races.

Table: Summary of experimental results

Technique	Correct	Incorrect	Did not bet	Payoffs ($)
Expert 1	19	81	0	-71.40
Expert 2	17	83	0	-61.20
Expert 3	18	82	0	-70.20
ID3	34	50	16	69.20
Backprop.	20	80	0	124.80

The experts averaged 17 correct predictions and were incorrect in the remaining races. Each of the three experts had a negative payoff at the end of all betting. On the other hand, the ID3 and backpropogation algorithms actually finished with significantly positive payoffs. ID3 in effect predicted double the amount of winners that were predicted by the experts while also avoiding betting on situations that were too close to call. The neural networks (backpropogation) algorithm, on the other hand, while predicting less winners showed a propensity for correctly picking longshots – those dogs with the odds heavily against them. Hence, the neural network approach saw nearly double the winnings of the ID3 algorithm. Nonetheless, both algorithms radically outperformed the human experts by using data mining techniques in order to make predictions.

4.2 Neural Network Prediction Research: Football

Traditional College Football Rankings

Unlike other college sports such as baseball and basketball which have their annual tournaments to determine the national champion, NCAA football has long used the bowl system as its postseason format. In the old bowl system, the

different bowls – each with its own organizers and sponsors – selected teams to compete in their event. The final national champion was determined through the use of polls. The AP poll is based on the compiled rankings of a number of media members. The other poll, which has been associated with USA Today, CNN, and ESPN over the years, is a compilation of the rankings of a sample of college football head coaches. This, however, often resulted in controversy in the form of accusations of bias, split national champions, and fan frustration.

In *Ranking College Football Teams*: *A Neural Network Approach* (Wilson, 1995), Rick Wilson proposes neural networks as a solution to the bias that is widespread in college football. The human rankings which were the determinants in the awarding of the national championship were based heavily on what games and teams the 'experts' had seen, where they were from, and what schools they may have had associations with. Neural networks would provide a fair and unbiased data-driven approach to the creation of true rankings for college football teams.

In the neural network application to football, the neurons (processing elements) would be the various college teams and the value of each neuron would represent the overall strength of that team. Links between neurons would represent a game played between the two teams and the weight would indicate a win, loss, or tie and its relative magnitude. The algorithm calculates a value for each game from each team's perspective based on the game outcome (incoming connection weights) and opponent's strength (output values of linked neurons).

The overall strength of a team is the averaged value of the total game weights for that team.

Wilson's ranking system presupposes a number of college football ranking characteristics that were developed using ranking theories and the author's own beliefs:

1. Defeating a team is worth more than losing to that same team
2. If a team defeats two teams A and B, and A is stronger (higher value) than B, than more value should be awarded for defeating team A than defeating team B.
3. If a team loses to two teams A and B, and A is stronger (higher value) than B, than more value should be awarded for losing to team A than losing to team B.
4. A team should receive more value for beating an improving team (one whose value increases) than for losing to that same team.
5. There are certain scenarios in which defeating a weak team would provide less value than losing to a strong team.

Wilson applied this neural network approach to ranking the 1993 college football season using a neuron for each of the 106 teams in Division I-A. While some Division I-A teams will often play schools from lower divisions at the start of each season, only games between Division I-A teams were included in the implementation. The connection weights were valued based on the point differential in the game with a maximum of 15 to account for the fact that some

coaches try to run-up the score while others prefer the sportsmanship route and avoid further humiliating opponents.

Since the NCAA has long refused to implement the playoff system similar to college basketball that fans and media have sought, neural networks may present an effective technique for ranking college teams in order to determine the teams that should compete for the national championship. Over a decade ago, Wilson proposed a technique designed to free college football from the controversy and unfairness that have long plagued the traditional bowl system.

Further Neural Network Applications to College Football

In 1998, the NCAA implemented the Bowl Championship Series (BCS) as a solution to the problems that had previously plagued college football. The BCS would be comprised of the four most prominent bowl games – the Rose Bowl, Sugar Bowl, Orange Bowl, and Fiesta Bowl. One of the bowl games, rotated annually, would host the two teams competing for the national championship. Of the eight teams that would compete, six would be the conference champions from the six BCS Conferences – Pac 10, Big Ten, Big XII, ACC, SEC, Big East – and the remaining two would be at-large selections based on 'merit'.

The formula itself is a combination of a number of factors which include a) the average of the two human polls, b) the average of seven computer polls, c) a schedule ranking where two-thirds of the overall value came from the win/loss

percentage of the opponents and the remaining one-third came from the win/loss percentage of the opponents' opponents, d) a point added for each loss suffered, e) and a 'quality win component' that awarded a team additional points for defeating a quality opponent weighted by the opponent's BCS ranking.

The BCS system, though, did not provide a solution to the problems that had plagued college football. In the 2002-03 season, USC finished the season ranked as the number one overall team in both human polls. However, due to a low ranking in the computer polls, it finished third in the final regular season BCS rankings. Louisiana State University (LSU) and Oklahoma, the top two BCS teams, played for the national championship. LSU, the victor in the BCS National Championship Game, was awarded the title. USC which defeated Michigan in the Rose Bowl was awarded the AP National Championship resulting in another split title. This was not the only controversy to hit the BCS, as previous seasons also had their share of issues.

In his research project entitled *An Artificial Neural Network Approach to College Football Prediction and Ranking* (Pardee, 1999), Michael Pardee builds on the research of Rick Wilson concerning the forecasting abilities of neural networks. The scope of the project included only the teams composing the Big Ten conference using the 1998 season statistics for training the algorithm and the following season's statistics for testing the neural network's effectiveness in predicting winners. Similar to the greyhound racing study (Chen et al, 1994), Pardee chose to use backpropogation for the college football prediction analysis.

The results for training on the 1998 season data and testing on the subsequent (1999) season in over 100 trials resulted in an average prediction rate of 76.2%. For the same games, a computer ranking system used by the BCS formula called the Massey Index Ratings exhibited a 68.2% prediction rate. Additionally, reversing the process by training on the 1999 season data and testing on the previous season (1998) in 20 trials resulted in a 74% success rate.

Since college players are required to stay at least three years in school in order to be eligible for the NFL, there is some level of consistency year to year in terms of talent for each team. This may explain the relative success of Pardee's approach of using previous seasons to predict subsequent season results. However, there are a variety of factors that are not considered in this methodology that may impact the actual games such as loss of players to graduation or the NFL, player injury, the ability of younger players to improve, and the like.

Wilson's approach, on the other hand, is perhaps a perfect fit for the bowl system. At the end of each regular season, teams with winning records are supposed to be selected to the appropriate bowl game based on merit. Only the top two teams in the nation should be allowed to compete in the national championship game without regard to regional bias and politics. Currently, this is rarely the case. Wilson's neural network approach, where each game is an entity with a unique strength value for each team involved, essentially treats a team's

regular season as a portfolio. The algorithm ranks these portfolios based on their strength and merit. Hence, using neural networks each team is given its fair shake.

NFL Game Prediction

In another similar study entitled *Neural Network Prediction of NFL Games* (Kahn, 2003), Joshua Kahn examines the predictive capabilities of neural networks using the various football statistics for the competing teams in a game. For the purposes of his study, Kahn used a back-propagation network in order to find relationships and associations that would forecast future NFL team performance based on past performances.

NFL box scores, such as those found in the sports section of most newspapers, were used as the data source from which the statistics would be collected. Box scores from every game during the first fourteen weeks of the 2003 season were gathered resulting in 208 distinct data sets.

In his first attempt to understand the statistics and their connection to team performance, Kahn plotted the outcomes of each of the 208 games from the first fourteen weeks. From this initial examination, he determined that the determining factors that should be analyzed by the algorithm were:

- Total Yardage Differential: The difference between the yards gained by the offense and those given up by the defense.
- Rushing Yardage Differential: The difference between the rushing yards gained by the team and those given up by the team.
- Time of Possession Differential: The difference between the amount of time the team had the ball and the time the opponent had the ball.
- Turnover Differential: The difference between the turnovers committed and the turnovers gained.
- Home or Away: This accounts for the home-field advantage.

The previously mentioned data set was used in order to train the algorithm and the prediction set was applied to the neural network. For each game, two outputs were generated – a predicted outcome for the home team and one for the away team. For each game, the data was applied three times and the results were averaged to produce a final prediction.

For the following weeks (14 and 15) of the NFL season, the data sets for the teams from the season up until that point were effectively used in predicting game outcomes. The neural network correctly predicted 75% of the game winners for each of the two weeks. However, results using data only from the

previous three weeks was less successful than those of the season-long data sets. As a baseline, Kahn compared the results to those of eight NFL experts who make predictions on game outcomes at ESPN.com. The results for the two weeks were 57% the first week and 87% the following week – an average of 72%. While there seems to be only a minor difference in performance, the neural network was far more consistent than the experts. Kahn's study indicates that data mining tools, such as neural networks, can be used not only to find patterns in the data but also for predictive and forecasting purposes in sports.

4.3 Neural Network Tools

There are a number of available general neural network analysis software applications that also advertise their applications to sports. In particular, they promote the predictive capabilities of neural networks, which have been exhibited in the pervious sections, primarily targeting those bettors who wish to profit from gambling activities. Personally, I do not condone this use of neural network tools nor are they being promoted for such behavior. Rather, I have decided to include the description of one specific type of tool that can and should only be used to test the predictive qualities of a data mining technique, neural networks, in relation to sports events.

NeuroXL

The Predictor software provided by NeuroXL (http://www.neuroxl.com) is an add-in to Microsoft Excel which can be used to predict the outcome of sporting events as well as horse and dog races. A user must supply input data – the NeuroXL website uses the example of a team's performance statistics – which is analyzed by the neural networks application resulting in an outcome such as the probability of a team winning. NeuroXL advertises the ease-of-use of its software and the affordable price for those seeking to apply neural networks to their own data sets.

4.4 Other Sports Prediction Research

Horse Racing

The analysis of data and uncovering of patterns and relationships among different pieces of data and information can result in knowledge that can be used to help predict future performance. In their paper *The Relationship of Subsequent Racing Performance to Foreleg Flight Patterns* (Seder & Vickery, 2005), Jeffrey Seder and Charles Vickery discovered that the foreleg motion of 2-year-old thoroughbred race horses who had yet to compete measured at auctions had a direct relation to its future success measured in race earnings.

According to the study, horses with "good" foreleg motion, the determination of which is done by veterinary experts and is beyond the scope of this review, earned 83% more money than those who were determined to have "bad" foreleg motion. The difference in foreleg motions is difficult to assess first-hand or using

regular video tape. However, through the use of stop-action video analysis, buyers are able to better gauge the quality of a horse's foreleg before making an investment.

Additionally, the predictive quality of this relationship was used to debunk a popular perception among horse racing experts. It is commonly believed that horses with bad foreleg motion are better equipped to race on turf than those with good foreleg motion. However, it was found that for all horses that raced on turf, those with good foreleg motion, were more likely to win and finish in the top three.

Hidden within data are often indicators of future performance. Using the right algorithms, such as neural networks and decision trees, and identifying the key pieces of information, historical data can be used to make accurate predictions.

Chapter 5: Sports Video Analysis

5.1 Sports Video Tools: Synergy Sports Technology

In 2004, Garrick Barr, a former Phoenix Suns video coordinator, and Nils Lahr, a former Microsoft engineer who is considered one of the fathers of streaming media, founded Synergy Sports Technology (Sandoval, 2006). The fledgling company currently provides video analysis services to NBA teams. During the 2005-06 season, four NBA teams including Boston, Indiana, as well as the two NBA Finals participants Dallas and eventual-champion Miami paid the five-figure fee for the video services.

Before this technology, NBA team decision makers would often have to wait days for video to be edited and transferred onto digital media such as DVDs. Synergy's online service allows coaches and scouts to log in and access streaming video results depending on the search query. Scouting players through video, which previously required a tedious amount of work just to get the footage desired, has become a quick and painless task.

A team of 'loggers' are employed by Synergy in order to link each possession footage with the critical statistical information for the search capability including the player with the ball as well as the type and result of the play. Based on his background as both a college basketball player and NBA video coordinator, Barr understood that in most sports, especially the NBA, statistics can often be misleading. By linking the statistics to the actual event footage, NBA

organizations can become more systematic in their analysis and consequently more confident in their decisions.

The Synergy video analysis tool's capabilities are not confined to the National Basketball Association. Recently, the company began offering college basketball services. The application could also have a strong impact on other sports organizations and leagues such as the NFL and the Major Leagues.

5.2 Video Analysis Research

SoccerQ

Researchers at the University of Miami and Florida Atlantic University have developed a system that can store, manage, and retrieve video clips with their associated information and features for soccer footage (S.-C. Chen et al, 2005). The program, SoccerQ, takes input queries from the user and returns the relevant video clips. However, the framework of the research is intended to be applicable to most types of sports video.

The tool supports both basic queries and relative temporal queries. The basic queries are comprised of three statement types for video retrieval:

select *video* from *search_space [where* condition];

select *shot* from *search_space [where* condition];

select *variable* from *search_space [where* condition];

where variable may refer to the name of a team involved, for example. The user display (Fig. 5.2.1) is divided into video browsing and event query panels. A graphical interface, where buttons and drop-down lists represent the various filters and operators, is used for the relative temporal queries.

The screenshot below displays the interface query settings and results for the query:

> "Find all the *corner kick* shots from all the female soccer videos where the *corner kick* resulted in a *goal event* occurring in 2 minutes."

Figure: SoccerQ Screenshot

In the screenshot, we can see that the two events selected are “Corner Kick” (Event A) and “Goal” (Event B). Since the query specified clips from women’s games, the “Female Soccer Videos” search space is selected. Additionally, the corner kick in the search is expected to be the precursor of the goal scored hence the temporal relationship chosen is “A starts B”. Finally, the time between the end of event A and event B (RBf – Af) is set to 2 with “Minutes” selected as

the unit of measurement. The resulting video clips are displayed at the bottom of the screen. The user merely selects one the returned clips in order to view it.

Sports video analysis is an emerging field. Research and tools that link data mining techniques with sports video are few and far between. However, the two summarized above show that the sports video analysis field has great potential.

Chapter 6: Conclusions and Future Directions

The application of data mining concepts and techniques has yet to blossom to its full potential in the sports world. In many respects, most sports organizations have only begun to unearth the information and knowledge hidden in their data.

Organizations, from commercial to government, now seek a competitive edge by introducing data mining initiatives and projects. The Australian Institute of Sport (AIS) recently introduced two initiatives aimed at making the best of use of the abundance of information, relating to the various sports the comprise the body, that currently exist and will be produced in the future (Lyons, 2005). The first initiative is the creation of a digital repository to house the various video, audio, and data files. This repository will be centralized such that the various sports programs can access needed information. The second initiative is aimed at using data mining techniques in order to uncover new knowledge that may be held in the vast amounts of data contained in their databases.

As professional sports teams grow from multi-million to billion dollar organizations and the stakes continue to rise, information technology and data mining will correspondingly play larger roles. Similarly, governments will also further use these techniques for national teams, performance measures, and predictive purposes. All in all, data mining has shown great value to organizations in the sports world and the recent future will undoubtedly bring increased research as well as commercial opportunities.

References:

- Adler, Joseph (2006). Baseball Hacks: Tips & Tools for Analyzing and Winning with Statistics. O'Reilly Media Inc.
- Albert, Jim, "An Introduction to Sabermetrics", Bowling Green State University (http://www-math.bgsu.edu/~albert/papers/saber.html), 1997.
- Ballard, Chris, "Measure of Success", Sports Illustrated (SI.com), 2006-10-01.
- Basketball-Reference.com, "Calculating PER", (http://www.basketball-reference.com/about/per.html).
- BBall, "BBall: The Pro Basketball Simulation Software", (http://www.bballsports3.com/bball.html).
- Beech, Roland, "NBA Player Shot Zones", 82Games.com (http://www.82games.com/shotzones.htm), 2005-10-08.
- Beech, Roland, "NBA Clutch Players, Part III", 82Games.com (http://www.82games.com/clutchplay3.htm) , 2005-3-10
- Bhandari, Inderpal et al., "Advanced Scout: Data Mining and Knowledge Discovery in NBA Data", Data Mining and Knowledge Discovery v. 1, p. 121-125, 1997.
- Carroll, Bob and Palmer, Pete and Thorn, John and Pietrusza, David (1998). The Hidden Game of Football: The Next Edition. Total Sports.
- Chen, H. (2006). Intelligence and Security Informatics for International Security: Information Sharing and Data Mining. Springer.
- Chen, Hsinchun et al., "Expert Prediction, Symbolic Learning, and Neural Networks: An Experiment in Greyhound Racing", IEEE Expert, December 1994.
- Coleman, Jay and Lynch, Allen, "NCAA Men's Basketball Tournament Score Card", University of Northern Florida (http://www.unf.edu/~jcoleman/score.htm), 2006.
- Craggs, Tommy, "He Stats! He Scores!", SF Weekly, 2004-2-11.

- Fieltz, Lynn and Scott, David, “Prediction of Physical Performance Using Data Mining”, Research Quarterly for Exercise and Sport, March 2003 v74 i1 pA-25.
- Flinders, Karl, “Football Injuries are Rocket Science”, vnunet.com (http://www.vnunet.com/vnunet/news/2120386/football-injuries-rocket-science), 2002-10-14.
- Han, J. and Kamber, M. (2001). Data Mining: Concepts and Techniques. San Francisco, CA: Morgan Kaufmann.
- SAS, “A Method to March Madness”, SAS.com (http://www.sas.com/news/feature/01mar05/dancecard.html), 2005-3-01.
- Hollinger, John (2002). Pro Basketball Prospectus: 2002 Edition. Potomac Books.
- Justice, Richard, “Rockets see Exec Morey as Extra Edge”, Houston Chronicle (http://www.chron.com/disp/story.mpl/sports/justice/rockets/3796565.html), 2006-4-16.
- Kahn, Joshua, “Neural Network Prediction of NFL Games”, University of Wisconsin – Electrical and Computer Engineering Department, 2003.
- Lewis, Michael (2003). Moneyball: The Art of Winning an Unfair Game. W. W. Norton & Company.
- Lyons, Keith, “Data Mining and Knowledge Discovery”, Australian Sports Commission Journals, Ausport Volume 2, Number 4, September 2005 (http://www.ausport.gov.au/journals/ausport/vol2no4/24data.asp).
- MIT Sloan Alumni Profile – Daryl Morey, MBA ’00 (http://mitsloan.mit.edu/mba/alumni/morey.php).
- Oliver, Dean (2005). Basketball on Paper: Rules and Tools for Performance Analysis. Potomac Books.
- Pardee, Michael, “An Artificial Neural Network Approach to College Football Prediction and Ranking”, University of Wisconsin – Electrical and Computer Engineering Department, 1999.

- Pelton, Devin, “The Sonics Play Moneyball: Part One”, Supersonics.com, 2005.
- Rosenbaum, Dan T., “Measuring How NBA Players Help Their Teams Win”, 82Games.com (http://www.82games.com/comm30.htm), 2004-4-30.
- Sandoval, G., “A Video Slam Dunk for the NBA”, CNET.com (http://news.com.com/A+video+slam-dunk+for+the+NBA/2100-1008_3-6034908.html), 2006-2-06.
- S.-C. Chen, M.-L. Shyu, and N. Zhao, “An Enhanced Query Model for Soccer Video Retrieval Using Temporal Relationships”, Proceedings of the 21st International Conference of Data Engineering (ICDE 2005), 2005.
- Schatz, Aaron (2006). Pro Football Prospectus 2006: Statistics, Analysis, and Insight for the Information Age. Workman Publishing Company.
- Seder, Jeffrey A. and Vickery, Charles E., “The Relationship of Subsequent Racing Performance to Foreleg Flight Patterns During Race Speed Workouts of Unraced 2-Year-Old Thoroughbred Racehorses at Auctions”, Journal of Equine Veterinary Science, Volume 25, Number 12, December 2005.
- White, Paul, “Scouts Uncover a Winning Edge”, USA Today, 2006-3-03 Section 7E.
- Wilson, Rick, “Ranking College Football Teams: A Neural Network Approach”, Interfaces Volume 25 p. 44-59, 1995.
- Virtual Gold, Inc. (http://www.virtualgold.com/solutions_sports.html)