



# Sports Analytics and Data Science

---

Winning the Game with  
Methods and Models

---

THOMAS W. MILLER

Faculty Director of Northwestern University's Predictive Analytics Program

# **Sports Analytics and Data Science**

**Winning the Game with Methods and Models**

THOMAS W. MILLER

Publisher: Paul Boger  
Editor-in-Chief: Amy Neidlinger  
Executive Editor: Jeanne Glasser Levine  
Cover Designer: Alan Clements  
Managing Editor: Kristy Hart  
Project Editor: Andy Beaster  
Manufacturing Buyer: Dan Uhrig

©2016 by Thomas W. Miller  
Published by Pearson Education, Inc.  
Old Tappan, New Jersey 07675

For information about buying this title in bulk quantities, or for special sales opportunities (which may include electronic versions; custom cover designs; and content particular to your business, training goals, marketing focus, or branding interests), please contact our corporate sales department at [corpsales@pearsoned.com](mailto:corpsales@pearsoned.com) or (800) 382-3419.

For government sales inquiries, please contact [governmentsales@pearsoned.com](mailto:governmentsales@pearsoned.com).

For questions about sales outside the U.S., please contact [international@pearsoned.com](mailto:international@pearsoned.com).

Company and product names mentioned herein are the trademarks or registered trademarks of their respective owners.

All rights reserved. No part of this book may be reproduced, in any form or by any means, without permission in writing from the publisher.

Printed in the United States of America

First Printing November 2015

ISBN-10: 0-13-388643-3

ISBN-13: 978-0-13-388643-6

Pearson Education LTD.  
Pearson Education Australia PTY, Limited.  
Pearson Education Singapore, Pte. Ltd.  
Pearson Education Asia, Ltd.  
Pearson Education Canada, Ltd.  
Pearson Educación de Mexico, S.A. de C.V.  
Pearson Education—Japan  
Pearson Education Malaysia, Pte. Ltd.  
Library of Congress Control Number: 2015954509

# Contents

Preface	v
Figures	ix
Tables	xi
Exhibits	xiii
1 Understanding Sports Markets	1
2 Assessing Players	23
3 Ranking Teams	37
4 Predicting Scores	49
5 Making Game-Day Decisions	61
6 Crafting a Message	69
7 Promoting Brands and Products	101
8 Growing Revenues	119
9 Managing Finances	133

10	Playing What-if Games	147
11	Working with Sports Data	169
12	Competing on Analytics	193
A	Data Science Methods	197
A.1	Mathematical Programming	200
A.2	Classical and Bayesian Statistics	203
A.3	Regression and Classification	206
A.4	Data Mining and Machine Learning	215
A.5	Text and Sentiment Analysis	217
A.6	Time Series, Sales Forecasting, and Market Response Models	226
A.7	Social Network Analysis	230
A.8	Data Visualization	234
A.9	Data Science: The Eclectic Discipline	240
B	Professional Leagues and Teams	255
	Data Science Glossary	261
	Baseball Glossary	279
	Bibliography	299
	Index	329

# Preface

“Sometimes you win, sometimes you lose, sometimes it rains.”

—TIM ROBBINS AS EBBY CALVIN LALOOSH IN *Bull Durham* (1988)

Businesses attract customers, politicians persuade voters, websites cajole visitors, and sports teams draw fans. Whatever the goal or target, data and models rule the day.

This book is about building winning teams and successful sports businesses. Winning and success are more likely when decisions are guided by data and models. Sports analytics is a source of competitive advantage.

This book provides an accessible guide to sports analytics. It is written for anyone who needs to know about sports analytics, including players, managers, owners, and fans. It is also a resource for analysts, data scientists, and programmers. The book views sports analytics in the context of data science, a discipline that blends business savvy, information technology, and modeling techniques.

To use analytics effectively in sports, we must first understand sports—the industry, the business, and what happens on the fields and courts of play. We need to know how to work with data—identifying data sources, gathering data, organizing and preparing them for analysis. We also need to know how to build models from data. Data do not speak for themselves. Useful predictions do not arise out of thin air. It is our job to learn from data and build models that work.

The best way to learn about sports analytics and data science is through examples. We provide a ready resource and reference guide for modeling techniques. We show programmers how to solve real world problems by building on a foundation of trustworthy methods and code.

The truth about what we do is in the programs we write. The code is there for everyone to see and for some to debug. Data sets and computer programs are available from the website for the *Modeling Techniques* series at <http://www.ftpress.com/miller/>. There is also a GitHub site at <https://github.com/mtpa/>.

When working on sports problems, some things are more easily accomplished with R, others with Python. And there are times when it is good to offer solutions in both languages, checking one against the other.

One of the things that distinguishes this book from others in the area of sports analytics is the range of data sources and topics discussed. Many researchers focus on numerical performance data for teams and players. We take a broader view of sports analytics—the view of data science. There are text data as well as numeric data. And with the growth of the World Wide Web, the sources of data are plentiful. Much can be learned from public domain sources through crawling and scraping the web and utilizing application programming interfaces (APIs).

I learn from my consulting work with professional sports organizations. Research Publishers LLC with its ToutBay division promotes what can be called “data science as a service.” Academic research and models can take us only so far. Eventually, to make a difference, we need to implement our ideas and models, sharing them with one another.

Many have influenced my intellectual development over the years. There were those good thinkers and good people, teachers and mentors for whom I will be forever grateful. Sadly, no longer with us are Gerald Hahn Hinkle in philosophy and Allan Lake Rice in languages at Ursinus College, and Herbert Feigl in philosophy at the University of Minnesota. I am also most thankful to David J. Weiss in psychometrics at the University of Minnesota and Kelly Eakin in economics, formerly at the University of Oregon.

My academic home is the Northwestern University School of Professional Studies. Courses in sports research methods and quantitative analysis, marketing analytics, database systems and data preparation, web and network data science, web information retrieval and real-time analytics, and data visualization provide inspiration for this book. Thanks to the many students and fellow faculty from whom I have learned. And thanks to colleagues and staff who administer excellent graduate programs, including the Master of Science in Predictive Analytics, Master of Arts in Sports Administration, Master of Science in Information Systems, and the Advanced Certificate in Data Science.

Lorena Martin reviewed this book and provided valuable feedback while she authored a companion volume on sports performance measurement and analytics (Martin 2016). Adam Grossman and Tom Robinson provided valuable feedback about coverage of topics in sports business management. Roy Sanford provided advice on statistics. Amy Hendrickson of T<sub>E</sub>Xnology Inc. applied her craft, making words, tables, and figures look beautiful in print—another victory for open source. Candice Bradley served dual roles as a reviewer and copyeditor for all books in the *Modeling Techniques* series. And Andy Beaster helped in preparing this book for final production. I am grateful for their guidance and encouragement.

Thanks go to my editor, Jeanne Glasser Levine, and publisher, Pearson/FT Press, for making this book possible. Any writing issues, errors, or items of unfinished business, of course, are my responsibility alone.

My good friend Brittney and her daughter Janiya keep me company when time permits. And my son Daniel is there for me in good times and bad, a friend for life. My greatest debt is to them because they believe in me.

Thomas W. Miller  
Glendale, California  
October 2015



*This page intentionally left blank*

# Figures

1.1	MLB, NBA, and NFL Average Annual Salaries	10
1.2	MLB Team Payrolls and Win/Loss Performance (2014 Season)	11
1.3	A Perceptual Map of Seven Sports	13
2.1	Multitrait-Multimethod Matrix for Baseball Measures	25
3.1	Assessing Team Strength: NBA Regular Season (2014–2015)	40
4.1	Work of Data Science	50
4.2	Data and Models for Research	52
4.3	Training-and-Test Regimen for Model Evaluation	54
4.4	Training-and-Test Using Multi-fold Cross-validation	56
4.5	Training-and-Test with Bootstrap Resampling	57
4.6	Predictive Modeling Framework for Team Sports	59
6.1	How Sports Fit into the Entertainment Space (Or Not)	72
6.2	Indices of Dissimilarity Between Pairs of Binary Variables	73
6.3	Consumer Preferences for Dodger Stadium Seating	77
6.4	Choice Item for Assessing Willingness to Pay for Tickets	79
6.5	The Market: A Meeting Place for Buyers and Sellers	80
7.1	Dodgers Attendance by Day of Week	104
7.2	Dodgers Attendance by Month	104
7.3	Dodgers Weather, Fireworks, and Attendance	106
7.4	Dodgers Attendance by Visiting Team	107
7.5	Regression Model Performance: Bobbleheads and Attendance	108
8.1	Competitive Analysis for an NBA Team: Golden State Warriors	129
9.1	Cost-Volume-Profit Analysis	135
9.2	Higher Profits Through Increased Sales	136
9.3	Higher Profits Through Lower Fixed Costs	137
9.4	Higher Profits Through Increased Efficiency	137
9.5	Decision Analysis: Investing in a Sports Franchise (Or Not)	143
10.1	Game-day Simulation (Offense Only)	152

10.2	Mets' Away and Yankees' Home Data (Offense and Defense)	154
10.3	Balanced Game-day Simulation (Offense and Defense)	155
10.4	Actual and Theoretical Runs-scored Distributions	157
10.5	Poisson Model for Mets vs. Yankees at Yankee Stadium	159
10.6	Negative Binomial Model for Mets vs. Yankees at Yankee Stadium	160
10.7	Probability of Home Team Winning (Negative Binomial Model)	162
10.8	Strategic Modeling Techniques in Sports	164
11.1	Software Stack for a Document Search and Selection System	173
11.2	The Information Supply Chain of Professional Team Sports	174
11.3	Automated Data Acquisition by Crawling, Scraping, and Parsing	177
11.4	Automated Data Acquisition with an API	179
11.5	Gathering and Organizing Data for Analysis	180
A.1	Mathematical Programming Modeling Methods	201
A.2	Evaluating the Predictive Accuracy of a Binary Classifier	212
A.3	Linguistic Foundations of Text Analytics	218
A.4	Creating a Terms-by-Documents Matrix	221
A.5	Data and Plots for the Anscombe Quartet	235
A.6	Visualizing Many Games Across a Season: Differential Runs Plot	236
A.7	Moving Fraction Plot for Basketball	237
A.8	Visualizing Basketball Play-by-Play Data	239
A.9	Data Science: The Eclectic Discipline	241

# Tables

1.1	Sports and Recreation Activities in the United States	3
1.2	MLB Team Valuation and Finances (March 2015)	5
1.3	NBA Team Valuation and Finances (January 2015)	6
1.4	NFL Team Valuation and Finances (August 2014)	7
1.5	World Soccer Team Valuation and Finances (May 2015)	8
2.1	Levels of Measurement	29
3.1	NBA Team Records (2014–2015 Season)	39
5.1	Twenty-five States of a Baseball Half-Inning	63
6.1	Dissimilarity Matrix for Entertainment Events and Activities	71
6.2	Consumer Preference Data for Dodger Stadium Seating	76
7.1	Bobbleheads and Dodger Dogs	103
7.2	Regression of Attendance on Month, Day of Week, and Promotion	110
9.1	Discounted Cash Flow Analysis of a Player Contract	139
9.2	Would you like to buy the Brooklyn Nets?	141
10.1	New York Mets' Early Season Games in 2007	149
10.2	New York Yankees' Early Season Games in 2007	150
A.1	Three Generalized Linear Models	209
A.2	Social Network Data: MLB Player Transactions	233
B.1	Women's National Basketball Association (WNBA)	255
B.2	Major League Baseball (MLB)	256
B.3	Major League Soccer (MLS)	257
B.4	National Basketball Association (NBA)	258
B.5	National Football League (NFL)	259
B.6	National Hockey League (NHL)	260

*This page intentionally left blank*

# Exhibits

1.1	MLB, NBA, and NFL Player Salaries (R)	16
1.2	Payroll and Performance in Major League Baseball (R)	18
1.3	Making a Perceptual Map of Sports (R)	19
3.1	Assessing Team Strength by Unidimensional Scaling (R)	43
6.1	Mapping Entertainment Events and Activities (R)	83
6.2	Mapping Entertainment Events and Activities (Python)	86
6.3	Preferences for Sporting Events—Conjoint Analysis (R)	88
6.4	Preferences for Sporting Events—Conjoint Analysis (Python)	99
7.1	Shaking Our Bobbleheads Yes and No (R)	113
7.2	Shaking Our Bobbleheads Yes and No (Python)	116
10.1	Team Winning Probabilities by Simulation (R)	167
10.2	Team Winning Probabilities by Simulation (Python)	168
11.1	Simple One-Site Web Crawler and Scraper (Python)	186
11.2	Gathering Opinion Data from Twitter: Football Injuries (Python)	189
A.1	Programming the Anscombe Quartet (Python)	242
A.2	Programming the Anscombe Quartet (R)	244
A.3	Making Differential Runs Plots for Baseball (R)	245
A.4	Moving Fraction Plot: A Basketball Example (R)	246
A.5	Visualizing Basketball Games (R)	248
A.6	Seeing Data Science as an Eclectic Discipline (R)	252

*This page intentionally left blank*

# 1

## Understanding Sports Markets

“Those of you on the floor at the end of the game, I’m proud of you. You played your guts out. I’m only going to say this one time. All of you have the weekend. Think about whether or not you want to be on this team under the following condition: What I say when it comes to this basketball team is the law, absolutely and without discussion.”

—GENE HACKMAN AS COACH NORMAN DALE IN *Hoosiers* (1986)

In applying the laws of economics to professional sports, we must consider the nature of sports and the motives of owners. Professional sports are different from other forms of business.

There are sellers and buyers of sports entertainment. The sellers are the players and teams within the leagues of professional sports. The buyers are consumers of sports, many of whom never go to games in person but who watch sports on television, listen to the radio, and buy sports team paraphernalia.

Sports compete with other forms of entertainment for people’s time and money. And various sports compete with one another, especially when their seasons overlap. Sports teams produce entertainment content that is distributed through the media. Sports teams license their brand names and logos to other organizations, including sports apparel manufacturers.



Sports teams are not independent businesses competing with one another. While players and teams compete on the fields and courts of play, they cooperate with one another as members of leagues. The core product of sports is the sporting contest, a joint product of two or more players or two or more teams.

Fifty-four sports and recreation activities, shown in table 1.1, are tracked by the National Sporting Goods Association (2015), which serves the sporting goods industry. In recent years, participation in baseball, basketball, football, and tennis has declined, while participation in soccer has increased. There has been growth in individual recreational sports, such as skateboarding and snowboarding. Of course, levels of participation in sports are not necessarily an indicator of levels of interest in sports as entertainment.

Sports businesses produce entertainment products by cooperating with one another. While it is illegal for businesses in most industries to collude in setting output and prices, sports leagues engage in cooperative output and pricing as a standard part of their business model. The number of games, indeed the entire schedule of games in a sport, is determined by the league. In fact, aspects of professional sports are granted monopoly power by the federal government in the United States.

When developing a model for a typical business or firm, an economist would assume profit maximization as a motive. But for a professional sports team, an owner's motives may not be so easily understood. While one owner may operate his or her team for profit year by year, another may seek to maximize wins or overall utility. Another may look for capital appreciation—buying, then selling after a few years. Lacking knowledge of owners' motives, it is difficult to predict what they will do.

Gaining market share and becoming the dominant player is a goal of firms in many industries. Not so in the business of professional sports. If one team were assured of victory in almost all of its contests, interest in those contests could wane. A team benefits by winning more often than losing, but winning all the time may be less beneficial than winning most of the time. Professional sports leagues claim to be seeking competitive balance, although there are dominant teams in many leagues.

*Table 1.1. Sports and Recreation Activities in the United States*

Aerobic Exercising	Ice/Figure Skating
Archery (Target)	In-Line Roller Skating
Backpack/Wilderness Camping	Kayaking
Baseball	Lacrosse
Basketball	Martial Arts/MMA/Tae Kwon Do
Bicycle Riding	Mountain Biking (Off Road)
Billiards/Pool	Muzzleloading
Boating (Motor/Power)	Paintball Games
Bowling	Running/Jogging
Boxing	Scuba Diving (Open Water)
Camping (Vacation/Overnight)	Skateboarding
Canoeing	Skiing (Alpine)
Cheerleading	Skiing (Cross Country)
Dart Throwing	Snowboarding
Exercise Walking	Soccer
Exercising with Equipment	Softball
Fishing (Fresh Water)	Swimming
Fishing (Salt Water)	Table Tennis/Ping Pong
Football (Flag)	Target Shooting (Airgun)
Football (Tackle)	Target Shooting (Live Ammunition)
Football (Touch)	Tennis
Golf	Volleyball
Gymnastics	Water Skiing
Hiking	Weight Lifting
Hockey (Ice)	Work Out at Club/Gym/Fitness Studio
Hunting with Bow & Arrow	Wrestling
Hunting with Firearms	Yoga

Sports is big business as shown by valuations and finances of the major professional sports in the United States and worldwide. Data from *Forbes* for Major League Baseball (MLB), the National Basketball Association (NBA), the National Football League (NFL), and worldwide soccer teams are shown in tables 1.2 through 1.5.

Professional sports teams most certainly compete with one another in the labor market, and labor in the form of star players is in short supply. Some argue that salary caps are necessary to preserve competitive balance. Salary caps also help teams in limiting expenditures on players.

Most professional sports in the United States have salary caps. The 2015 salary cap for NFL teams, with fifty-three player rosters, is set at \$143.28 million (Patra 2015). Most teams have payrolls at or near the cap, making the average salary of an NFL player about \$2.7 million. One player on an NFL team may be designated as a *franchise player*, restricting that player from entering free agency. The league sets minimum salaries for franchise players. For example, a franchise quarterback has a minimum salary of \$18.544 million in 2015. The highest annual salary among NFL players is \$22 million for Aaron Rodgers, Green Bay Packers quarterback (spotrac 2015c). The minimum annual salary is \$420 thousand.

NBA teams have a \$70 million salary cap for the 2015–16 season, with penalties for teams going over the cap. Maximum player salaries are based on a percentage of cap and years of service. For example, LeBron James, with ten years of experience, would have a maximum salary of \$23 million (Mahoney 2015). New Orleans Pelicans Anthony Davis' average salary of \$29 million is the highest among NBA players (spotrac 2015b). Team rosters include fifteen players under contract, with as many as thirteen available to play in any particular game. The minimum annual salary is \$428,498.

Major League Baseball (MLB) has a "luxury tax" for teams with payrolls in excess of \$189 million. There is a regular-player roster of twenty-five or twenty-six players for double-header days/nights. A forty-man roster includes players under contract and eligible to play. Between September 1 and the end of the regular season the roster is expanded to forty players. The roster drops back to twenty-five players for the playoffs. The minimum MLB annual salary is \$505,700 in 2015. The highest MLB annual salary is \$31 million for Miguel Cabrera of the Detroit Tigers (spotrac 2015a).

*Table 1.2. MLB Team Valuation and Finances (March 2015)*

Team Rank	Team	Current Value (\$ Millions)	One-Year Change in Value (Percentage)	Debt/Value (Percentage)	Revenue (\$ Millions)	Operating Income (\$ Millions)
1	New York Yankees	3,200	28	0	508	8.1
2	Los Angeles Dodgers	2,400	20	17	403	-12.2
3	Boston Red Sox	2,100	40	0	370	49.2
4	San Francisco Giants	2,000	100	4	387	68.4
5	Chicago Cubs	1,800	50	24	302	73.3
6	St Louis Cardinals	1,400	71	21	294	73.6
7	New York Mets	1,350	69	26	263	25.0
8	Los Angeles Angels	1,300	68	0	304	16.7
9	Washington Nationals	1,280	83	27	287	41.4
10	Philadelphia Phillies	1,250	28	8	265	-39.0
11	Texas Rangers	1,220	48	13	266	3.5
12	Atlanta Braves	1,150	58	0	267	33.2
13	Detroit Tigers	1,125	65	15	254	-20.7
14	Seattle Mariners	1,100	55	0	250	26.4
15	Baltimore Orioles	1,000	61	15	245	31.4
16	Chicago White Sox	975	40	5	227	31.9
17	Pittsburgh Pirates	900	57	10	229	43.6
18	Minnesota Twins	895	48	25	223	21.3
19	San Diego Padres	890	45	22	224	35.0
20	Cincinnati Reds	885	48	6	227	2.2
21	Milwaukee Brewers	875	55	6	226	11.3
22	Toronto Blue Jays	870	43	0	227	-17.9
23	Colorado Rockies	855	49	7	214	12.6
24	Arizona Diamondbacks	840	44	17	211	-2.2
25	Cleveland Indians	825	45	9	207	8.9
26	Houston Astros	800	51	34	175	21.6
27	Oakland Athletics	725	46	8	202	20.8
28	Kansas City Royals	700	43	8	231	26.6
29	Miami Marlins	650	30	34	188	15.4
30	Tampa Bay Rays	625	29	22	188	7.9

Source. Badenhausen, Ozanian, and Settini (2015b).

**Table 1.3.** *NBA Team Valuation and Finances (January 2015)*

Team Rank	Team	Current Value (\$ Millions)	One-Year Change in Value (Percentage)	Debt/Value (Percentage)	Revenue (\$ Millions)	Operating Income (\$ Millions)
1	Los Angeles Lakers	2,600	93	2	293	104.1
2	New York Knicks	2,500	79	0	278	53.4
3	Chicago Bulls	2,000	100	3	201	65.3
4	Boston Celtics	1,700	94	9	173	54.9
5	Los Angeles Clippers	1,600	178	0	146	20.1
6	Brooklyn Nets	1,500	92	19	212	-99.4
7	Golden State Warriors	1,300	73	12	168	44.9
8	Houston Rockets	1,250	61	8	175	38.0
9	Miami Heat	1,175	53	8	188	12.6
10	Dallas Mavericks	1,150	50	17	168	30.4
11	San Antonio Spurs	1,000	52	8	172	40.9
12	Portland Trail Blazers	940	60	11	153	11.7
13	Oklahoma City Thunder	930	58	15	152	30.8
14	Toronto Raptors	920	77	16	151	17.9
15	Cleveland Cavaliers	915	78	22	149	20.6
16	Phoenix Suns	910	61	20	145	28.2
17	Washington Wizards	900	86	14	143	10.1
18	Orlando Magic	875	56	17	143	20.9
19	Denver Nuggets	855	73	1	136	14.0
20	Utah Jazz	850	62	6	142	32.7
21	Indiana Pacers	830	75	18	139	25.0
22	Atlanta Hawks	825	94	21	133	14.8
23	Detroit Pistons	810	80	23	144	17.6
24	Sacramento Kings	800	45	29	125	8.9
25	Memphis Grizzlies	750	66	23	135	10.5
26	Charlotte Hornets	725	77	21	130	1.2
27	Philadelphia 76ers	700	49	21	125	24.4
28	New Orleans Pelicans	650	55	19	131	19.0
29	Minnesota Timberwolves	625	45	16	128	6.9
30	Milwaukee Bucks	600	48	29	110	11.5

Source. Badenhausen, Ozanian, and Settini (2015a).

*Table 1.4. NFL Team Valuation and Finances (August 2014)*

Team Rank	Team	Current Value (\$ Millions)	One-Year Change in Value (Percentage)	Debt/Value (Percentage)	Revenue (\$ Millions)	Operating Income (\$ Millions)
1	Dallas Cowboys	3,200	39	6	560	245.7
2	New England Patriots	2,600	44	9	428	147.2
3	Washington Redskins	2,400	41	10	395	143.4
4	New York Giants	2,100	35	25	353	87.3
5	Houston Texans	1,850	28	11	339	102.8
6	New York Jets	1,800	30	33	333	79.5
7	Philadelphia Eagles	1,750	33	11	330	73.2
8	Chicago Bears	1,700	36	6	309	57.1
9	San Francisco 49ers	1,600	31	53	270	24.8
10	Baltimore Ravens	1,500	22	18	304	56.7
11	Denver Broncos	1,450	25	8	301	30.7
12	Indianapolis Colts	1,400	17	4	285	60.7
13	Green Bay Packers	1,375	16	1	299	25.6
14	Pittsburgh Steelers	1,350	21	15	287	52.4
15	Seattle Seahawks	1,330	23	9	288	27.3
16	Miami Dolphins	1,300	21	29	281	8.0
17	Carolina Panthers	1,250	18	5	283	55.6
18	Tampa Bay Buccaneers	1,225	15	15	275	46.4
19	Tennessee Titans	1,160	10	11	278	35.6
20	Minnesota Vikings	1,150	14	43	250	5.3
21	Atlanta Falcons	1,125	21	27	264	13.1
22	Cleveland Browns	1,120	11	18	276	35.0
23	New Orleans Saints	1,110	11	7	278	50.1
24	Kansas City Chiefs	1,100	9	6	260	10.0
25	Arizona Cardinals	1,000	4	15	266	42.8
26	San Diego Chargers	995	5	10	262	39.9
27	Cincinnati Bengals	990	7	10	258	11.9
28	Oakland Raiders	970	18	21	244	42.8
29	Jacksonville Jaguars	965	15	21	263	56.9
30	Detroit Lions	960	7	29	254	-15.9
31	Buffalo Bills	935	7	13	252	38.0
32	St Louis Rams	930	6	12	250	16.2

Source: Badenhausen, Ozanian, and Settini (2014).

*Table 1.5. World Soccer Team Valuation and Finances (May 2015)*

Team Rank	Team	Current Value (\$ Millions)	One-Year Change in Value (Percentage)	Debt/Value (Percentage)	Revenue (\$ Millions)	Operating Income (\$ Millions)
1	Real Madrid	3,263	-5	4	746	170
2	Barcelona	3,163	-1	3	657	174
3	Manchester United	3,104	10	20	703	211
4	Bayern Munich	2,347	27	0	661	78
5	Manchester City	1,375	59	0	562	122
6	Chelsea	1,370	58	0	526	83
7	Arsenal	1,307	-2	30	487	101
8	Liverpool	982	42	10	415	86
9	Juventus	837	-2	9	379	50
10	AC Milan	775	-10	44	339	54
11	Borussia Dortmund	700	17	6	355	55
12	Paris Saint-Germain	634	53	0	643	-1
13	Tottenham Hotspur	600	17	9	293	63
14	Schalke 04	572	-1	0	290	57
15	Inter Milan	439	-9	56	222	-41
16	Atletico de Madrid	436	33	53	231	47
17	Napoli	353	19	0	224	43
18	Newcastle United	349	33	0	210	44
19	West Ham United	309	33	12	186	54
20	Galatasaray	294	-15	17	220	-37

Source. Ozanian (2015).

Figure 1.1, a histogram lattice, shows how player salaries compare across the MLB, NBA, and NFL in August 2015. Player salary distributions are positively skewed. The mean salary across NFL players is around \$1.7 million, but the median is \$630 thousand. The mean salary across NBA players is around \$5.1 million, with median salary \$2.8 million. The mean salary across MLB players is around \$4.1 million, with the median \$1.1 million.

Do team expenditures on players buy success? This is a meaningful question to ask for leagues that have no salary caps. Szymanski (2015) reports studies showing that between 60 and 90 percent of the variability in U.K. soccer team positions may be explained by wages paid to players. Major League Baseball has a luxury tax in place of a salary cap, and team payrolls vary widely in size. The New York Yankees have been known for having the highest payrolls in baseball. Recently, the Los Angeles Dodgers have surpassed the Yankees with the highest player payroll—more than \$257 million at the end of the 2014 season (Woody 2014).

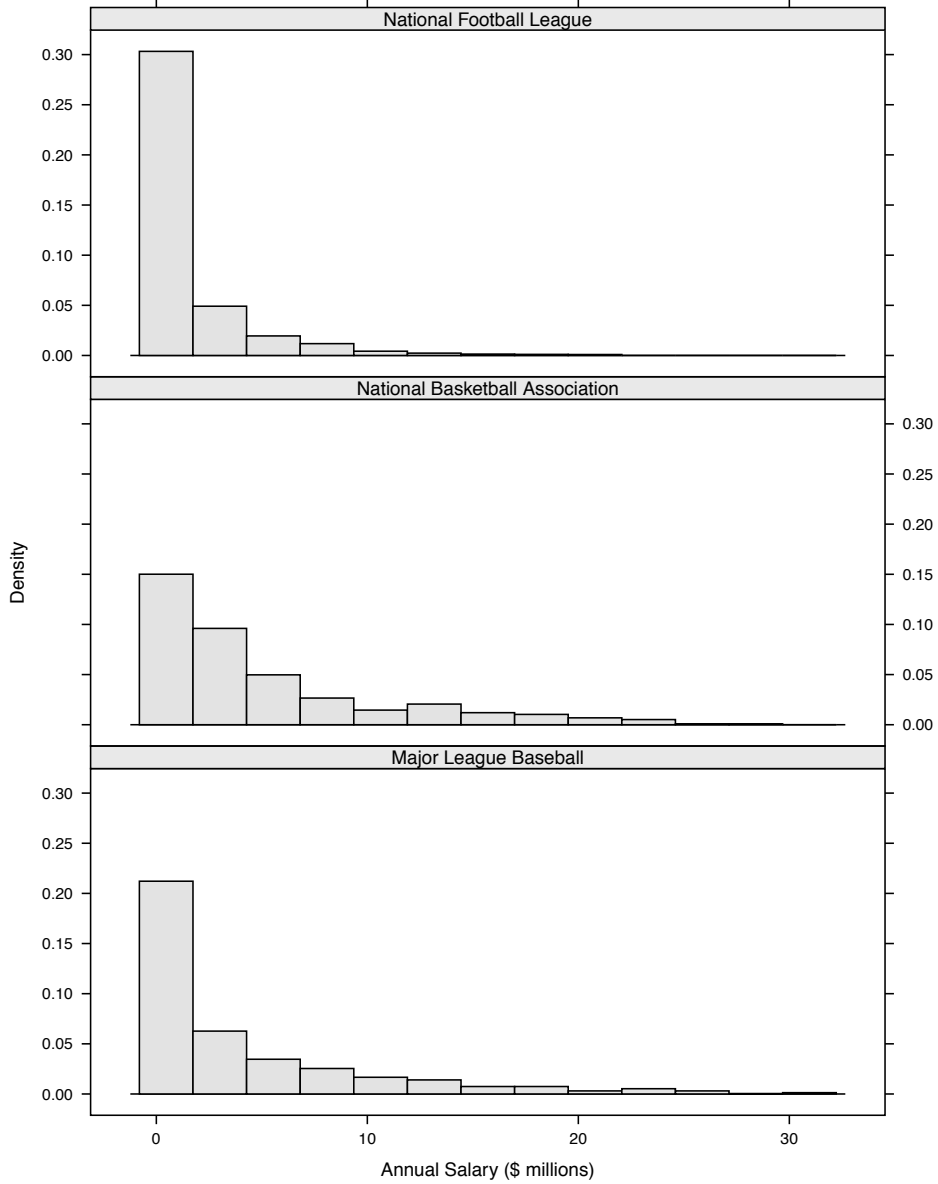
Figure 1.2 shows baseball team salaries at the beginning of the 2014 season plotted against the percentage of games won across the regular season. Notice how teams that made the playoffs in 2014, labeled with team abbreviations, have a wide range of payrolls. While the biggest spenders in baseball are often among the set of teams going to the playoffs, the relationship between team payrolls and team performance is weak at best—less than 7 percent of the variability in win/loss percentages is explained by player payrolls.

The thesis of Michael Lewis' *Moneyball* (2003) and what has become the ethos of sports analytics is that small-market baseball teams can win by spending their money wisely. Star players demand top salaries due as much to their celebrity status as to their skills. Players with high on-base percentages, overlooked by major-market teams, can be hired at much lower salaries than star players.

Teams, although associated with particular cities, can be known nationwide or worldwide. The media of television and the Internet provide opportunities for reaching consumers across the globe. A Super Bowl at the Rose Bowl in Pasadena, California or AT&T Stadium in Arlington, Texas may be attended by around 100 thousand fans (Alder 2015), while U.S. television audiences have grown to over 100 million (statista 2015).

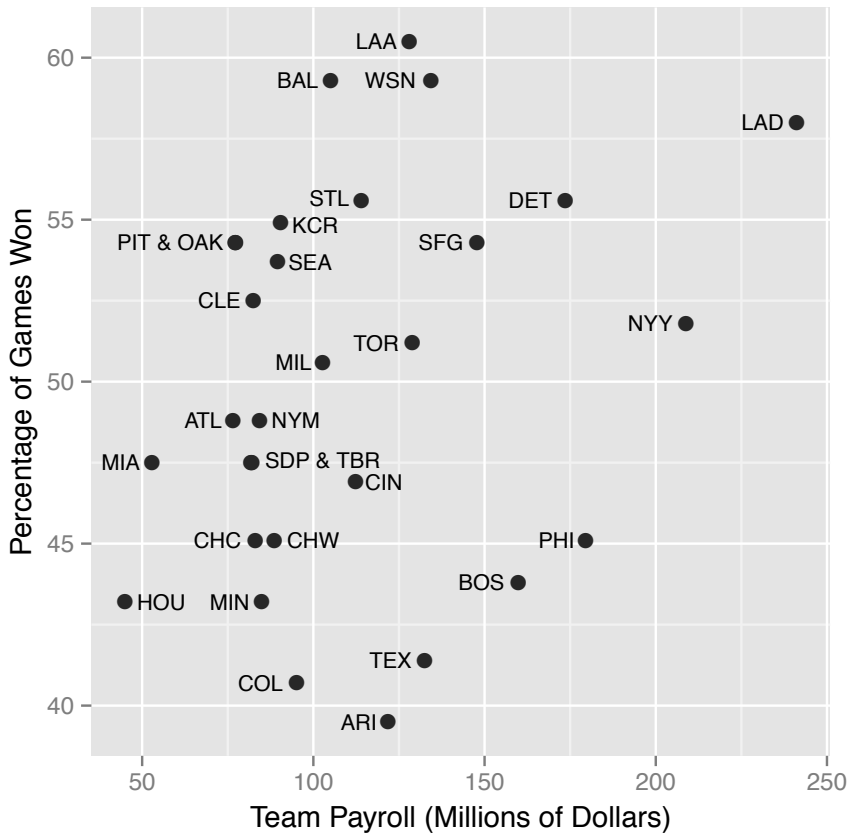


Figure 1.1. MLB, NBA, and NFL Average Annual Salaries



Sources. spotrac (2015a, 2015b, 2015c).

Figure 1.2. MLB Team Payrolls and Win/Loss Performance (2014 Season)



Sources. Sports Reference LLC (2015b) and USA Today (2015).

See Appendix B, page 255, for team abbreviations and names.

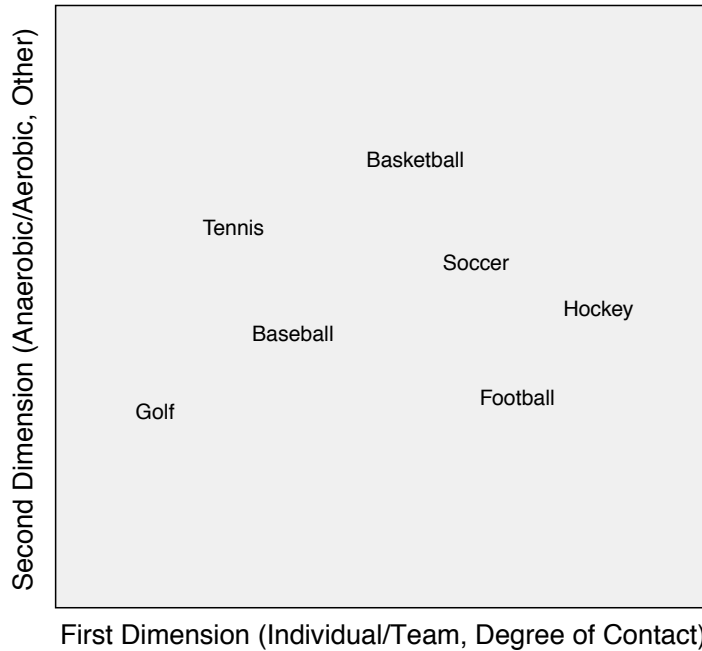
Media revenues are important to successful sports teams. Other revenues come from business partnerships, sponsorships, advertising, and stadium naming rights. City governments understand well the power of sports to promote business. Locating sports arenas in cities can help to revitalize downtown areas, as demonstrated by the experience of the Oklahoma City Thunder. Indianapolis, Indiana promotes itself as a sports capital with the Colts and Pacers (Rein, Shields, and Grossman 2015).

Teams seek to build their brands, developing a positive reputation in the minds of consumers. Players, like fans, are attracted to teams with a reputation for hard work, courage, fair play, honesty, teamwork, and community service. The character of a team is often as important as its likelihood of winning. The Cubs are associated with Chicago, but Cub fans may be found from Maine to California. This is despite the fact that the Cubs have not won the World Series since 1908. Teams in U.S. professional sports vie to become “America’s team,” with fans across the land wearing their logo-embossed hats and jerseys.

The demand for sports and the feelings of sports consumers are not so easily understood. Fans can be fickle and fandom fleeting. Fans can be loyal to a sport, to a team, or to individual players. Multivariate methods can help us understand how sports consumers think by revealing relationships among products or brands.

Figure 1.3 provides an example, a *perceptual map* of seven sports. Along the horizontal dimension, we move from individual, non-contact sports on the left-hand side, to team sports with little contact, to team sports with contact on the right-hand side. The vertical dimension, less easily described, may be thought of as relating to the aerobic versus anaerobic nature of sports and to other characteristics such as physicality and skill. Sports such as tennis, soccer, and basketball entail aerobic exercise. These are endurance sports, while football is an example of a sport that involves both aerobic and anaerobic exercise, including intense exercise for short durations. Sports close together on the map have similarities. Baseball and golf, for example, involve special skills, such as precision in hitting a ball. Soccer and hockey involve almost continuous movement and getting a ball through the goal. Football and hockey have high physicality or player contact.

Figure 1.3. A Perceptual Map of Seven Sports



In many respects, professional sports teams are decidedly different from other businesses. They are in the public eye. They live and die in the media. And a substantial portion of their revenues come from media.

Késenne (2007), Szymanski (2009), Fort (2011), Fort and Winfree (2013), Leeds and von Allmen (2014), and the edited volumes of Humphreys and Howard (2008a, 2008b, 2008c) review sports economics and business issues.

Gorman and Calhoun (1994) and Rein, Shields, and Grossman (2015) focus on alternative sources of revenue for sports teams and how these relate to business strategy. The business of baseball has been the subject of numerous volumes (Miller 1990; Zimbalist 1992; Powers 2003; Bradbury 2007; Pessah 2015). And Jozsa (2010) reviews the history of the National Basketball Association.

An overview of sports marketing is provided by Mullin, Hardy, and Sutton (2014). Rein, Kotler, and Shields (2006) and Carter (2011) discuss the convergence of entertainment and sports. Miller (2015a) reviews methods in marketing data science, including product positioning maps, market segmentation, target marketing, customer relationship management, and competitive analysis.

Sports also represents a laboratory for labor market research. Sports is one of the few industries in which job performance and compensation are public knowledge. Economic studies examine player performance measures and value of individual players to teams (Kahn 2000; Bradbury 2007). Miller (1991), Abrams (2010), and Lowenfish (2010) review baseball labor relations. And Early (2011) provides insight into labor and racial discrimination in professional sports.

Sports wagering markets have been studied extensively by economists because they provide public information about price, volume, and rates of return. Furthermore, sports betting opportunities have fixed beginning and ending times and published odds or point spreads, making them easier to study than many financial investment opportunities. As a result, sports wagering markets have become a virtual field laboratory for the study of market efficiency. Sauer (1998) provides a comprehensive review of the economics of wagering markets.

When management objectives can be defined clearly in mathematical terms, teams use mathematical programming methods—constrained optimization. Teams attempt to maximize revenue or minimize costs subject to known situational factors. There has been extensive work on league schedules, for which the league objective may be to have teams playing one another an equal number of times while minimizing total distance traveled between cities. Alternatively, league officials may seek home/away schedules, revenue sharing formulas, or draft lottery rules that maximize competitive balance. Briskorn (2008) reviews methods for scheduling sports competition, drawing on integer programming, combinatorics, and graph theory. Wright (2009) provides an overview of operations research in sport.

Extensive data about sports are in the public domain, readily available in newspapers and online sources. These data offer opportunities for predictive modeling and research. Throughout the book we also identify places to apply methods of operations research, including mathematical programming and simulation.

Exhibit 1.1 shows an R program for exploring distributions of player salaries across the MLB, NBA, and NFL. The program draws on software for statistical graphics from Sarkar (2008).

Exhibit 1.2 (page 18) shows an R program for examining the relationship between MLB payrolls and win-loss performance. The program draws on software for statistical graphics from Wickham and Chang (2014).

Exhibit 1.3 (page 19) shows an R program to obtain a perceptual map of seven sports, showing their relationships with one another. The program draws on modeling software for multidimensional scaling.

*Exhibit 1.1. MLB, NBA, and NFL Player Salaries (R)*

```

# MLB, NBA, and NFL Player Salaries (R)

library(lattice) # statistical graphics

# variables in contract data from spotrac.com (August 2015)
# player: player name (contract years)
# position: position on team
# team: team abbreviation
# teamsignedwith: team that signed the original contract
# age: age in years as of August 2015
# years: years as player in league
# contract: dollars in contract
# guaranteed: guaranteed dollars in contract
# guaranteedpct: percentage of contract dollars guaranteed
# salary: annual salary in dollares
# yearfreeagent: year player becomes free agent
#
# additional created variables
# salarymm: salary in millions
# leaguename: full league name
# league: league abbreviation

# read data for Major League Baseball
mlb_contract_data <- read.csv("mlb_player_salaries_2015.csv")
mlb_contract_data$leaguename <- rep("Major League Baseball",
  length = nrow(mlb_contract_data))
for (i in seq(along = mlb_contract_data$yearfreeagent))
  if (mlb_contract_data$yearfreeagent[i] == 0)
    mlb_contract_data$yearfreeagent[i] <- NA
for (i in seq(along = mlb_contract_data$age))
  if (mlb_contract_data$age[i] == 0)
    mlb_contract_data$age[i] <- NA
mlb_contract_data$salarymm <- mlb_contract_data$salary/1000000
mlb_contract_data$league <- rep("MLB", length = nrow(mlb_contract_data))
print(summary(mlb_contract_data))
# variables for plotting
mlb_data_plot <- mlb_contract_data[, c("salarymm", "leaguename")]

nba_contract_data <- read.csv("nba_player_salaries_2015.csv")
nba_contract_data$leaguename <- rep("National Basketball Association",
  length = nrow(nba_contract_data))
for (i in seq(along = nba_contract_data$yearfreeagent))
  if (nba_contract_data$yearfreeagent[i] == 0)
    nba_contract_data$yearfreeagent[i] <- NA
for (i in seq(along = nba_contract_data$age))
  if (nba_contract_data$age[i] == 0)
    nba_contract_data$age[i] <- NA
nba_contract_data$salarymm <- nba_contract_data$salary/1000000
nba_contract_data$league <- rep("NBA", length = nrow(nba_contract_data))
print(summary(nba_contract_data))

```

```
# variables for plotting
nba_data_plot <- nba_contract_data[, c("salarymm", "leaguename")]

nfl_contract_data <- read.csv("nfl_player_salaries_2015.csv")
nfl_contract_data$leaguename <- rep("National Football League",
  length = nrow(nfl_contract_data))
for (i in seq(along = nfl_contract_data$yearfreeagent))
  if (nfl_contract_data$yearfreeagent[i] == 0)
    nfl_contract_data$yearfreeagent[i] <- NA
for (i in seq(along = nfl_contract_data$age))
  if (nfl_contract_data$age[i] == 0)
    nfl_contract_data$age[i] <- NA
nfl_contract_data$salarymm <- nfl_contract_data$salary/1000000
nfl_contract_data$league <- rep("NFL", length = nrow(nfl_contract_data))
print(summary(nfl_contract_data))
# variables for plotting
nfl_data_plot <- nfl_contract_data[, c("salarymm", "leaguename")]

# merge contract data with variables for plotting
plotting_data_frame <- rbind(mlb_data_plot, nba_data_plot, nfl_data_plot)

# generate the histogram lattice for comparing player salaries
# across the three leagues in this study
lattice_object <- histogram(~salarymm | leaguename, plotting_data_frame,
  type = "density", xlab = "Annual Salary ($ millions)", layout = c(1,3))

# print to file
pdf(file = "fig_understanding_markets_player_salaries.pdf",
  width = 8.5, height = 11)
print(lattice_object)
dev.off()
```



*Exhibit 1.2. Payroll and Performance in Major League Baseball (R)*

```

# Payroll and Performance in Major League Baseball (R)

library(ggplot2) # statistical graphics

# functions used with grid graphics to split the plotting region
# to set margins and to plot more than one ggplot object on one page/screen
vplayout <- function(x, y)
viewport(layout.pos.row=x, layout.pos.col=y)

# user-defined function to plot a ggplot object with margins
ggplot.print.with.margins <- function(ggplot.object.name,
  left.margin.pct=10,
  right.margin.pct=10,top.margin.pct=10,bottom.margin.pct=10)
{ # begin function for printing ggplot objects with margins
  # margins expressed as percentages of total... use integers
  grid.newpage()
  pushViewport(viewport(layout=grid.layout(100,100)))
  print(ggplot.object.name,
    vp=vplayout((0 + top.margin.pct):(100 - bottom.margin.pct),
      (0 + left.margin.pct):(100 - right.margin.pct)))
} # end function for printing ggplot objects with margins

# read in payroll and performance data
# including annotation text for team abbreviations
mlb_data <- read.csv("mlb_payroll_performance_2014.csv")
mlb_data$millions <- mlb_data$payroll/1000000
mlb_data$winpercent <- mlb_data$wlpct * 100

cat("\nCorrelation between Payroll and Performance:\n")
with(mlb_data, print(cor(millions, winpercent)))

cat("\nProportion of win/loss percentage explained by payrolls:\n")
with(mlb_data, print(cor(millions, winpercent)^2))

pdf(file = "fig_understanding_markets_payroll_performance.pdf",
  width = 5.5, height = 5.5)
ggplot_object <- ggplot(data = mlb_data,
  aes(x = millions, y = winpercent)) +
  geom_point(colour = "darkblue", size = 3) +
  xlab("Team Payroll (Millions of Dollars)") +
  ylab("Percentage of Games Won") +
  geom_text(aes(label = textleft), size = 3, hjust = 1.3) +
  geom_text(aes(label = textright), size = 3, hjust = -0.25)

ggplot.print.with.margins(ggplot_object, left.margin.pct = 5,
  right.margin.pct = 5, top.margin.pct = 5, bottom.margin.pct = 5)

dev.off()

```

*Exhibit 1.3. Making a Perceptual Map of Sports (R)*

```
# Making a Perceptual Map of Sports (R)

library(MASS) # includes functions for multidimensional scaling
library(wordcloud) # textplot utility to avoid overlapping text

USE_METRIC_MDS <- FALSE # metric versus non-metric toggle

# utility function for converting a distance structure
# to a distance matrix as required for some routines and
# for printing of the complete matrix for visual inspection.
make.distance.matrix <- function(distance_structure)
  { n <- attr(distance_structure, "Size")
    full <- matrix(0,n,n)
    full[lower.tri(full)] <- distance_structure
    full+t(full)
  }

# enter data into a distance structure as required for various
# distance-based routines. That is, we enter the upper triangle
# of the distance matrix as a single vector of distances
distance_structure <-
  as.single(c(9,11,10,5,14,4,15,6,12,13,16,1,18,2,20,7,3,19,17,8,21))

# provide a character vector of sports names
sport_names <- c("Baseball", "Basketball", "Football",
  "Soccer", "Tennis", "Hockey", "Golf")

attr(distance_structure, "Size") <- length(sport_names) # set size attribute

# check to see that the distance structure has been entered correctly
# by converting the distance structure to a distance matrix
# using the utility function make.distance.matrix, which we had defined
distance_matrix <- unlist(make.distance.matrix(distance_structure))
cat("\n", "Distance Matrix of Seven Sports", "\n")
print(distance_matrix)

if (USE_METRIC_MDS)
  {
    # apply the metric multidimensional scaling algorithm and plot the map
    mds_solution <- cmdscale(distance_structure, k=2, eig=T)
  }

# apply the non-metric multidimensional scaling algorithm
# this is more appropriate for rank-order data
# and provides a more satisfactory solution here

if (!USE_METRIC_MDS)
  {
    mds_solution <- isoMDS(distance_matrix, k = 2, trace = FALSE)
  }
}
```

```

pdf(file = "plot_nonmetric_mds_seven_sports.pdf",
     width=8.5, height=8.5) # opens pdf plotting device
# use par(mar = c(bottom, left, top, right)) to set up margins on the plot
par(mar=c(7.5, 7.5, 7.5, 5))

# original solution
First_Dimension <- mds_solution$points[,1]
Second_Dimension <- mds_solution$points[,2]

# set up the plot but do not plot points... use names for points
plot(First_Dimension, Second_Dimension, type = "n", cex = 1.5,
     xlim = c(-15, 15), ylim = c(-15, 15)) # first page of pdf plots
# We plot the sport names in the locations where points normally go.
text(First_Dimension, Second_Dimension, labels = sport_names,
     offset = 0.0, cex = 1.5)
title("Seven Sports (initial solution)")

# reflect the horizontal dimension
# multiply the first dimension by -1 to get reflected image
First_Dimension <- mds_solution$points[,1] * -1
Second_Dimension <- mds_solution$points[,2]
plot(First_Dimension, Second_Dimension, type = "n", cex = 1.5,
     xlim = c(-15, 15), ylim = c(-15, 15)) # second page of pdf plots
text(First_Dimension, Second_Dimension, labels = sport_names,
     offset = 0.0, cex = 1.5)
title("Seven Sports (horizontal reflection)")

# reflect the vertical dimension
# multiply the second dimension by -1 to get reflected image
First_Dimension <- mds_solution$points[,1]
Second_Dimension <- mds_solution$points[,2] * -1
plot(First_Dimension, Second_Dimension, type = "n", cex = 1.5,
     xlim = c(-15, 15), ylim = c(-15, 15)) # third page of pdf plots
text(First_Dimension, Second_Dimension, labels = sport_names,
     offset = 0.0, cex = 1.5)
title("Seven Sports (vertical reflection)")

# multiply the first and second dimensions by -1
# for reflection in both horizontal and vertical directions
First_Dimension <- mds_solution$points[,1] * -1
Second_Dimension <- mds_solution$points[,2] * -1
plot(First_Dimension, Second_Dimension, type = "n", cex = 1.5,
     xlim = c(-15, 15), ylim = c(-15, 15)) # fourth page of pdf plots
text(First_Dimension, Second_Dimension, labels = sport_names,
     offset = 0.0, cex = 1.5)
title("Seven Sports (horizontal and vertical reflection)")
dev.off() # closes the pdf plotting device

pdf(file = "plot_pretty_original_mds_seven_sports.pdf",
     width=8.5, height=8.5) # opens pdf plotting device
# use par(mar = c(bottom, left, top, right)) to set up margins on the plot
par(mar=c(7.5, 7.5, 7.5, 5))

```

```
First_Dimension <- mds_solution$points[,1] # no reflection
Second_Dimension <- mds_solution$points[,2] # no reflection
# wordcloud utility for plotting with no overlapping text
textplot(x = First_Dimension,
         y = Second_Dimension,
         words = sport_names,
         show.lines = FALSE,
         xlim = c(-15, 15), # extent of horizontal axis range
         ylim = c(-15, 15), # extent of vertical axis range
         xaxt = "n", # suppress tick marks
         yaxt = "n", # suppress tick marks
         cex = 1.15, # size of text points
         mgp = c(0.85, 1, 0.85), # position of axis labels
         cex.lab = 1.5, # magnification of axis label text
         xlab = "",
         ylab = "")
dev.off() # closes the pdf plotting device

pdf(file = "fig_sports_perceptual_map.pdf",
     width=8.5, height=8.5) # opens pdf plotting device
# use par(mar = c(bottom, left, top, right)) to set up margins on the plot
par(mar=c(7.5, 7.5, 7.5, 5))
First_Dimension <- mds_solution$points[,1] * -1 # reflect horizontal
Second_Dimension <- mds_solution$points[,2]
# wordcloud utility for plotting with no overlapping text
textplot(x = First_Dimension,
         y = Second_Dimension,
         words = sport_names,
         show.lines = FALSE,
         xlim = c(-15, 15), # extent of horizontal axis range
         ylim = c(-15, 15), # extent of vertical axis range
         xaxt = "n", # suppress tick marks
         yaxt = "n", # suppress tick marks
         cex = 1.15, # size of text points
         mgp = c(0.85, 1, 0.85), # position of axis labels
         cex.lab = 1.5, # magnification of axis label text
         xlab = "First Dimension (Individual/Team, Degree of Contact)",
         ylab = "Second Dimension (Anaerobic/Aerobic, Other)")
dev.off() # closes the pdf plotting device
```

*This page intentionally left blank*

# Index

## A

adjacency matrix, 261  
advertising research, 229  
agent, 261  
agent-based modeling, 261  
Akaike information criterion (AIC), 54  
algebraic modeling system, 145, 261  
Apache Hadoop, *see* Hadoop  
Apache Lucene, *see* Lucene  
Apache Software Foundation, 261  
Apache Spark, *see* Spark  
API, *see* application programming interface  
application programming interface, 175, 178, 179, 261  
ARPANET, 175, 262  
ASP, 262  
association rules, 262  
asynchronous focus group, *see* blog

## B

bandwidth, 262  
baseball term  
    “out”, 289  
    “play ball”, 291  
    “safe”, 293  
    “time”, 296  
ahead in the count, 279  
All Star Game, 279  
American League, 279  
around the horn, 279  
at bats (AB), 279  
away team, 279  
bailing out, 279  
balk, 279  
ball, 280  
ban on women, 280  
base, 280  
base coach, 280

base hit, 280  
base on balls (BB), 280  
base runner, 280  
baserunner, 280  
baserunning error, 280  
bases loaded, 280  
batter, 280  
batter in the hole, 280  
batter on deck, 280  
batter’s box, 280  
battery, 280  
batting average (BA, AVG), 280  
batting stance, 281  
batting team, 281  
behind in the count, 281  
bench, 281  
big leagues, 281  
BIP (BPIP), 281  
bloop single, 281  
bunt, 281  
call, 281  
called game, 281  
catcher, 281  
caught looking, 281  
caught off base, 281  
caught stealing (CS), 281  
center fielder, 282  
Championship Series, 282  
changeup, 282  
check swing, 282  
choking up, 282  
chop single, 282  
clean-up hitter, 282  
closed batting stance, 282  
closer, 282  
clutch hitter, 282  
coach, 282  
command, 282  
control, 282

- cover the bases, 282
- crowd the plate, 283
- curveball (curve), 283
- cut fastball (cutter), 283
- cut-off position, 283
- dead ball, 283
- dead ball era, 283
- dead red hitter, 283
- defense, 283
- defensive indifference, 283
- designated hitter (DH), 283
- diamond, 283
- dig in, 284
- Division Series, 284
- double (2B), 284
- double play, 284
- double-header, 284
- double-switch, 284
- dugout, 284
- earned run average (ERA), 284
- expected runs, 62–65, 284
- extra-base hit, 284
- fair ball, 284
- fair territory, 284
- fan, 284
- fantasy baseball, 284
- fastball, 284
- fielder, 285
- fielder's choice, 285
- fielding error, 285
- first base, 285
- first baseman, 285
- five-tool player, 285
- fly ball, 285
- fly out, 285
- force out, 285
- forfeited game, 285
- foul ball, 285
- foul territory, 285
- foul tip, 285
- frame (a pitch), 285
- free agent, 285
- full count, 285
- game (G), 285
- grand slam, 285
- ground ball, 285
- ground out, 286
- ground-rule double, 286
- hit (H), 286
- hit batsman (hit by pitch, HBP), 286
- hit for the cycle, 286
- hit-and-run, 286
- hitter, 286
- hitter's park, 286
- hitting for power, 286
- hitting slump, 286
- hitting streak, 286
- holding runner on base, 286
- home plate, 286
- home run (HR), 286
- home team, 287
- illegal pitch, 287
- in the hole, 287
- infielder, 287
- inning, 287
- intentional base on balls, 287
- interference, 287
- keystone sack, 287
- knuckleball, 287
- lead-off hitter, 287
- leave the yard (go yard), 287
- left fielder (LF), 287
- left on base (LOB), 287
- lefty, 287
- line drive, 288
- lineup, 288
- live ball, 288
- live ball era, 288
- making the turn, 288
- manager, 288
- manufactured run, 193, 288
- men on base, 288
- Mendoza Line, 288
- middle infielder, 288
- middle reliever, 288
- National League, 288
- neighborhood play, 288
- no hitter, 289
- no-no, 289
- obstruction, 289
- offense, 289
- official scorer, 289
- on the field (team), 289
- on-base percentage (OBP), 289
- open batting stance, 289
- out, 289
- outfielder, 289
- overslide, 290
- pace of play, 290
- passed ball, 290
- PECOTA, 33, 34, 290
- perfect game, 290
- pick off assignment, 290
- pick off play, 290
- pinch hitter, 290
- pinch runner, 290
- pitch, 290
- pitch count, 290

- pitcher (P), 291
- pitcher's duel, 291
- pitcher's park, 291
- pitcher's plate, 291
- pitching depth, 291
- pitching from the stretch, 291
- pitching mound, 291
- pitching rotation, 291
- pivot foot, 291
- place hitter, 291
- plate, 291
- plate appearance, 291
- platooning, 291
- pop-up, 292
- position number, 292
- position player, 292
- power hitter, 292
- productive at bat, 292
- pull hitter, 292
- pull the string, 292
- quick pitch, 292
- reaching for the fences, 292
- regulation game, 292
- relief pitcher, 292
- replay review, 292
- retouch, 292
- reverse curve, 292
- right fielder (RF), 292
- RISP, 293
- rounding the bases, 293
- run, 293
- run batted in (RBI), 293
- run down, 293
- runner, 293
- sacrifice bunt, 293
- sacrifice fly, 293
- scoring position, 293
- screwball, 293
- season, 293
- second base, 293
- second baseman, 293
- secondary lead, 293
- semi-intentional walk, 293
- set position, 293
- shadow ball, 294
- shift, 294
- shine ball, 294
- shortstop, 294
- shutout, 294
- side-arm delivery, 294
- single (1B), 294
- slider, 294
- slugging percentage (SLG), 294
- small ball, 294
- spin rate, 294
- spitball, 295
- squeeze play, 295
- starting pitcher, 295
- steal (stolen base, SB), 295
- stepping in the bucket, 295
- strike, 295
- strike zone, 295
- strikeout (K), 295
- suspended game, 296
- sweep, 296
- switch hitter, 296
- switch pitcher, 296
- tag out, 296
- tagging up, 296
- take a lead (off base), 296
- take a pitcher deep, 296
- Texas Leaguer, 296
- third base, 296
- third baseman, 296
- three-bagger, 296
- throw, 296
- throwing error, 296
- tie game, 296
- tip a pitch, 296
- total bases (TB), 297
- triple (3B), 297
- triple crown, 297
- triple-play, 297
- two-bagger, 297
- umpire, 297
- umpire-in-chief, 297
- up the middle, 297
- up to bat (team), 297
- visiting team, 297
- VORP, 297
- walk, 297
- walk-off balk, 297
- walk-off hit, 297
- walk-off home run, 298
- WAR (WARP), 298
- WHIP, 298
- Wild Card Game, 298
- wild pitch, 298
- windup position, 298
- World Series, 298
- Bayesian statistics, 54, 144, 198, 199, 204, 205, 262
  - Bayes information criterion (BIC), 54
  - Bayes' theorem, 204
  - hierarchical modeling, 79, 267
- best-case/worst-case approach, 144, 262
- betweenness centrality, 232, 262
- big data, 262



binary variable, 262  
 biologically-inspired methods, 216  
 black box model, 215  
 bootstrap method, 55, 57  
 bootstrap sampling, 262  
 bot, *see* crawler (web crawler)  
 boundary (of a network), 263  
 bps, 263  
 brand positioning, *see* marketing, brand positioning  
 breakeven analysis, *see* cost-volume-profit analysis  
 brute force approach, 263  
 bulletin board, *see* blog

## C

C, C++, C#, 172, 173, 263  
 CART, 263  
 cascading style sheet (CSS), 263  
 case study  
   Return of the Bobbleheads, 112  
 censoring, 112, 206, 207  
 chat room, 263  
 choice uncertainty, *see* decision analysis  
 classical statistics, 199, 203, 205, 263  
   null hypothesis, 203  
   power, 204  
   statistical significance, 203, 204  
 classification, 198, 211–213, 215, 263  
   predictive accuracy, 211–213  
 client, 263  
 client-server application, 263  
 closeness centrality, 232, 263  
 cloud computing, 183  
 cluster analysis, 71, 73, 122, 216, 263  
 coefficient of determination, 211  
 comma-delimited text (csv), 263  
 compile cycle, 264  
 complexity, of model, 213  
 conjoint analysis, 78, 79, 264  
 consumer heterogeneity, *see* market segmentation, consumer heterogeneity  
 consumer surplus, *see* pricing research, consumer surplus  
 content analysis, 264  
 continuous random variable, *see* random variable, continuous  
 cookie, 264  
 corpus, 264  
 cost-volume-profit analysis, 135–138  
 crawler, 175–178, 183, 264  
 cross-sectional data, 264  
 cross-validation, 55, 56, 214, 264  
 CSS, *see* cascading style sheet (CSS)  
 csv, *see* comma-delimited text (csv)  
 cumulative frequency distribution, *see* data visualization, cumulative frequency distribution  
 customer lifetime value, 125  
 CVP analysis, *see* cost-volume-profit analysis  
 Cython, 172

## D

data mining, *see* machine learning  
 data organization, 227  
 data partitioning, 55  
 data preparation, 181  
   missing data, 182  
 data science, 197, 198  
 data visualization, 264  
   biplot, 81  
   box plot, 102, 104  
   cumulative frequency distribution, 239, 248  
   diagnostics, 109, 213  
   differential runs plot, 236  
   heat map, 161, 162  
   histogram, 10, 154, 157, 159, 160  
   lattice plot, 10, 105–108, 238  
   moving fraction plot, 237, 245, 246  
   Sankey diagram, 80  
   scatter plot, 11  
   spine chart, 72, 77, 78, 82, 88  
   strip plot, 105, 107  
 data-adaptive research, 52, 53  
 database system, 180, 181  
   MongoDB, 183, 270  
   MySQL, 270  
   non-relational, 183  
   NoSQL, 271  
   PostgreSQL, 183, 272  
   relational, 181, 183  
 decision analysis, 142, 143  
 decision tree, *see* decision analysis  
 decision uncertainty, *see* decision analysis  
 declarative language, 264  
 degree (of a network), 264  
 degree centrality, 265  
 degree distribution, 265  
 density (of a network), 265  
 descriptive statistics, 265  
 discounted cash flow analysis, 139–141  
   discount rate, 139  
   net present value (NPV), 139  
   payback period, 140  
   return on investment (ROI), 140  
 discrete random variable, *see* random variable, discrete  
 Document Object Model (DOM), 176, 265, 267

document store, 173, 177, 179, 180  
DOM, *see* Document Object Model (DOM)

## E

e-mail, 265  
eigenvector centrality, 232, 265  
Elasticsearch, 173, 182, 183, 225, 265  
Elo rating/ranking system, 40  
emoji, *see* emoticon  
emoticon, 265  
ethnography, 265  
expected value, 265  
experimental research, 265  
explanatory model, 198, 266  
explanatory variable, 51, 52, 124, 207, 266  
exploratory data analysis, 102

## F

factor analysis, 81, 266  
fantasy sports, 35, 129, 130  
forecasting, 229  
Fortran, 172  
frame, 266  
ftp, 266  
functional language, 266

## G

game theory, 266  
game-day strategy, *see* strategy, playing  
General Inquirer, 223  
generalized linear model, 214, 266  
generative grammar, 266  
genetic algorithm, 266  
genetic algorithms, 216  
Go (Golang), 172, 182, 266  
graph theory, 267  
grounded theory, 267

## H

Hadoop, 267  
heuristic, 216, 267  
hierarchical modeling, *see* Bayesian statistics,  
hierarchical modeling  
HTML, 175, 267  
HTTP, 267

## I

imputation, *see* multiple imputation  
in-game strategy, *see* strategy, playing  
indexing, 267  
inferential statistics, 267  
information retrieval, 182, 183, 220, 225, 267

injuries, 185, 189  
integer knapsack problem, *see* mathematical  
programming, knapsack problem  
integer programming, 267, *see* mathematical  
programming, integer programming  
integrated development environment (IDE),  
267  
interaction effect, 213  
Internet, 175, 267  
Internet of Things, 170, 268  
Internet Services Provider, 268  
intranet, 268  
investment analysis, *see* discounted cash flow  
analysis  
IoT, *see* Internet of Things  
IRC, 268  
IT, 268

## J

Java, 172, 173, 268  
Java Virtual Machine, *see* Java  
JavaScript, 268  
JavaScript Object Notation (JSON), 179, 268  
JPEG, 268  
JSON, *see* JavaScript Object Notation (JSON)  
JVM, *see* Java

## K

kbps, 268  
keyword, 268

## L

LAMP, 268  
leading indicator, 229  
leave-one-out cross-validation, *see* cross-validation  
levels of measurement, 269  
linear model, 207, 214  
linear predictor, 207  
linear programming, 269, *see* mathematical  
programming, linear programming  
linear regression, 269  
Linux, 173, 183  
listserv, 269  
logging, *see* system logging  
logistic regression, 42, 52, 211, 269  
longitudinal data, 269  
Lucene, 173, 225, 269

**M**

machine learning, 215, 216, 269  
 Major League Baseball, *see* MLB  
 Major League Soccer, *see* MLS  
 market basket analysis, 73  
 market response model, 111, 226–229  
 market segmentation, 79, 120–122  
   consumer heterogeneity, 127, 131  
 marketing  
   competitive analysis, 128, 129  
   mass marketing, 123  
   one-to-one marketing, 123  
   product positioning, 71, 72, 74  
   strategy, 128, 129  
   substitute products, 74  
   target marketing, 120  
 marketing mix model, 111, 228  
 Markov chain, 62, 269  
   transition probability, 62, 269  
 mass marketing, *see* marketing, mass marketing  
 mathematical programming, 127, 200, 269  
   integer programming, 200, 201, 267  
   knapsack problem, 41, 200, 201, 268  
   linear programming, 200, 269  
   mixed integer programming, 200, 269  
   stochastic programming, 134, 145, 275  
 measurement, 23–35, 269  
   accessible measure, 31  
   comprehensible measure, 31, 32  
   construct validity, 24, 27  
   content validity, 24, 224  
   convergent validity, 27  
   discriminant validity, 27  
   explicit measure, 31  
   face validity, 24, 224  
   levels of measurement, 28–29  
   multitrait-multimethod matrix, 24–26  
   predictive validity, 24  
   reliability, 24–26, 30, 274  
   tractable measure, 31  
   transparent measure, 31, 33  
   validity, 24, 26, 27, 31, 32, 277  
 MEG, *see* listserv  
 mixed integer programming, 269, *see* mathematical programming, mixed integer programming  
 MKP (multidimensional knapsack problem), *see* mathematical programming, knapsack problem  
 MLB, 4, 5, 9–11, 64, 65, 170, 201, 232, 233, 256, 288  
 MLS, 257  
 model, 270

model-dependent research, 52, 53  
 modem, 270  
 Moneyball, 9  
 MongoDB, *see* database system, MongoDB  
 morphology, 270  
 multi-fold cross-validation, *see* cross-validation  
 multi-level categorical variable, 270  
 multi-level modeling, *see* Bayesian statistics, hierarchical modeling  
 multidimensional knapsack problem, *see* mathematical programming, knapsack problem  
 multidimensional scaling, 71, 72, 81  
   Jaccard index, 73  
   perceptual map, 12, 13, 70–72, 241  
   similarity judgment, 71–73  
 multinomial variable, *see* multi-level categorical variable  
 multiple imputation, 270  
 multivariate methods, 81  
 MySQL, *see* database system, MySQL

**N**

naïve Bayes model, 270  
 National Basketball Association, *see* NBA  
 National Football League, *see* NFL  
 National Hockey League, *see* NHL  
 natural language processing, *see* text analytics, natural language processing  
 NBA, 4, 6, 9, 10, 12, 38–40, 65–67, 170, 238, 239, 258  
 nearest-neighbor analysis, 73  
 net present value, *see* discounted cash flow analysis, net present value (NPV)  
 netiquette, 270  
 network, 270  
 network science, 270  
 neural network, 270  
 NFL, 4, 7, 9, 10, 170, 259  
 NHL, 260  
 NLP, *see* text analytics, natural language processing  
 NoSQL, *see* database system, NoSQL  
 NPV, *see* discounted cash flow analysis, net present value (NPV)

**O**

object-oriented language, 271  
 observational research, 271  
 on-base percentage plus slugging (OPS), 289  
 one-to-one marketing, *see* marketing, one-to-one marketing  
 online community, 271, 277

operations research, 271  
 optimization, 216  
 organic search, 271  
 Orme, B. K., 131  
 outlier, 271  
 over-fitting, 213

## P

PageRank, 271  
 paid search, 271  
 paired comparisons, 42  
 panel, 271  
 panel data, 271, *see* longitudinal data  
 parameter, 272  
 parametric models, 213  
 parser, 176, 177, 272  
 payback period, *see* discounted cash flow analysis  
 PECOTA, *see* baseball term, PECOTA  
 perceptual map, *see* multidimensional scaling, perceptual map  
 Perl, 272  
 PHP, 272  
 Poisson distribution, 272  
 Poisson regression, 206, 207, 210, 214, 272  
 population, 272  
 population distribution, 272  
 Porter five-forces model, 128, 129  
 post, 272  
 posterior distribution, 272  
 PostgreSQL, *see* database system, PostgreSQL  
 predictive analytics  
   definition, 51  
 predictive model, 198, 272  
 preference scaling, 81  
 price elasticity, *see* pricing research, price elasticity  
 price sensitivity, *see* pricing research, price sensitivity  
 pricing research, 126  
   consumer surplus, 128  
   price elasticity, 126  
   price sensitivity, 127  
   reference price, 127  
   three Cs of pricing, 126  
   willingness to pay, 75, 78, 79, 126, 127  
 principal component analysis, 81, 216, 272  
 prior distribution, 273  
 probability, 64  
   binomial distribution, 158  
   negative binomial distribution, 156, 158, 161  
   Poisson distribution, 156, 158, 161  
 procedural language, 273

product positioning, 81, *see* marketing, product positioning  
 psychographics, 273  
 Python, 171–173, 182, 273  
 Python package  
   json, 189  
   matplotlib, 86, 116, 242  
   numpy, 86, 99, 116, 168, 242  
   os, 186  
   pandas, 99, 116, 242  
   patsy, 99  
   scipy, 116, 168  
   scrapy, 186  
   sklearn, 86  
   statsmodels, 99, 116, 242  
   twitter, 189

## Q

queuing theory, 273

## R

R, 171–173, 182, 273  
 R package  
   car, 113  
   ggplot2, 18, 245, 246, 248  
   grid, 18, 245, 246, 248  
   lattice, 16, 113, 167  
   lubridate, 248  
   MASS, 83  
   plyr, 248  
   support.CEs, 88  
 R-squared, 211  
 random forests, 273  
 random network (random graph), 231, 273  
 random variable, 273  
   continuous, 273  
   discrete, 273  
 real-time analytics, 171, 172, 182  
 real-time focus group, *see* chat room  
 recency-frequency-monetary value (RFM) model, 126  
 reference price, *see* pricing research, reference price  
 regression, 52, 108–111, 197, 206, 213, 273  
   nonlinear regression, 214  
   robust methods, 214  
   time series regression, 228  
 regular expressions, 178, 273  
 regularized regression, 213  
 relational database, 274  
 reliability, *see* measurement, reliability  
 resampling, 274  
 response, 51  
 response variable, 124, 206

return on investment, *see* discounted cash flow analysis  
 robot, *see* crawler (web crawler)  
 ROI, *see* discounted cash flow analysis  
 root-mean-square error (RMSE), 211

## S

sales forecasting, 226, 228  
 sampling, 274  
   sampling variability, 204  
 sampling distribution, 274  
 sampling frame, 274  
 Scala, 172, 274  
 scatter plot, 274  
 scheduling, 216  
 scraper, 176, 177, 274  
 search, *see* information retrieval  
 semantic web, 274  
 semantics, 274  
 semi-supervised learning, 216, 274  
 sensitivity analysis, 145, 274  
 sentiment analysis, 222–225  
 shrinkage estimators, 213  
 similarity judgment, *see* multidimensional scaling, similarity judgment  
 simulated annealing, 274  
 simulation, 148, 153, 214, 275  
   benchmark study, 214, 215  
   game-day, 59, 151, 152  
   what-if analysis, 198  
 small-world network, 231  
 smoothing methods, 213  
   splines, 214  
 social network analysis, 230, 275  
 Solr, 225, 275  
 Spark, 172, 182, 275  
 spatial data, 275  
 spider, *see* crawler (web crawler)  
 SQL, 181, 275  
 statistic  
   interval estimate, 203  
   p-value, 203  
   point estimate, 203  
   test statistic, 203  
 statistical learning, *see* machine learning  
 stemming (word stemming), 275  
 stochastic process, 275  
 stochastic programming, *see* mathematical programming, stochastic programming  
 strategy  
   playing, 61–68  
 strength of schedule, 38  
 structured query language, *see* SQL

substitute product, *see* marketing, substitute product  
 supervised learning, 206, 216, 222, 275  
 support vector machine, 275  
 survival analysis, 206, 207  
 synchronous focus group, *see* chat room  
 syntax, 275  
 system logging, 275

## T

target marketing, *see* marketing, target marketing  
 TCP/IP, 275  
 telnet, 276  
 term frequency-inverse document frequency, *see* TF-IDF  
 text analytics, 217–225  
   content analysis, 223  
   document annotation, 35  
   generative grammar, 217, 218  
   latent Dirichlet allocation, 216  
   latent semantic analysis, 216  
   morphology, 218  
   natural language processing, 217, 224, 270  
   semantics, 218  
   stemming, 219  
   syntax, 218  
   terms-by-documents matrix, 219, 221  
   text analysis, 276  
   text feature, 35  
   text measure, 35, 223–225  
   text summarization, 222  
   thematic analysis, 216, 223  
 text measure, 276  
 text mining, 276  
 text parser, *see* parser  
 TF-IDF, 220, 276  
 thread, of discussion, 276  
 time series, 276  
 time series analysis, 226–229  
   ARIMA model, 227  
   state space model, 228  
 time-value of money, 142  
 traditional research, 52  
 traditional statistics, 276  
 training-and-test regimen, 54, 55, 105, 108, 198, 199  
 transition probability, *see* Markov chain, transition probability  
 transitivity, 276  
 tree-structured model, 276

**U**

unidimensional scaling, 38, 40, 42, 43  
unsupervised learning, 216, 220, 276  
URL, 175, 277  
Usenet, 277

**V**

validity, *see* measurement, validity  
variable transformation, 213  
virtual facility, 277  
virtual private network (VPN), 277  
VPN, *see* virtual private network

**W**

web, *see* World Wide Web

web board, *see* blog  
web presence testing, 277  
web scraper, *see* scraper  
weblog, *see* blog  
Wiki, 277  
willingness to pay, *see* pricing research, willingness to pay  
WNBA, 255  
Women's National Basketball Association, *see* WNBA  
World Wide Web, 174, 175, 183, 277  
WWW, *see* World Wide Web

**X**

XML, 277  
XPath, 176, 277