REAL-WORLD DATA MINING

DURSUN DELEN
Spears and Patterson Endowed Chairs in Business Analytics
Professor of Management Science and Information Systems
Spears School of Business, Oklahoma State University

APPLIED BUSINESS ANALYTICS
AND DECISION MAKING
Real-World Data Mining
# Books in the FT Press Analytics Series

*Enterprise Analytics* by Thomas Davenport and the International Institute for Analytics (ISBN: 0133039439)

*People Analytics* by Ben Waber (ISBN: 0133158314)


*Applying Advanced Analytics to HR Management Decisions* by James Sesil (ISBN: 0133064603)


*Managerial Analytics* by Michael Watson and Derek Nelson (ISBN: 013340742X)


*Big Data Analytics Beyond Hadoop* by Vijay Agneeswaran (ISBN: 0133837947)

*Computational Intelligence in Business Analytics* by Les Sztaendera (ISBN: 013355208X)

*Big Data Driven Supply Chain Management* by Nada R. Sanders (ISBN: 0133801284)

*Marketing and Sales Analytics* by Cesar Brea (ISBN: 0133592928)

*Cutting-Edge Marketing Analytics* by Rajkumar Venkatesan, Paul Farris, and Ronald T. Wilcox (ISBN: 0133552527)


*Advanced Analytics Methodologies* by Michele Chambers and Thomas W. Dinsmore (ISBN: 0133498603)


*Profiting from the Data Economy* by David A. Schweidel (ISBN: 0133819779)

*Business Analytics with Management Science Models and Methods* by Arben Asllani (ISBN: 0133760359)


*Web and Network Data Science* by Thomas W. Miller (ISBN: 0133886441)


*Trends and Research in the Decision Sciences* by Decision Sciences Institute and Merrill Warkentin (ISBN: 0133925374)
Real-World Data Mining

Applied Business Analytics
and Decision Making

Dursun Delen, Ph.D.
Professor of Management Science and Information Systems
Spears and Patterson Endowed Chairs in Business Analytics
Spears School of Business, Oklahoma State University
To my mother;
without her love and compassion,
I wouldn’t be here.
Contents

Foreword ................................................................. viii

Chapter 1 Introduction to Analytics ................................. 1
   Is There a Difference Between Analytics and Analysis? ........ 3
   Where Does Data Mining Fit In? .............................. 4
   Why the Sudden Popularity of Analytics? ................. 4
   The Application Areas of Analytics ....................... 6
   The Main Challenges of Analytics ......................... 7
   A Longitudinal View of Analytics .......................... 10
   A Simple Taxonomy for Analytics ............................ 15
   The Cutting Edge of Analytics: IBM Watson ............. 20
   References .................................................... 28

Chapter 2 Introduction to Data Mining .......................... 31
   What Is Data Mining? ........................................... 35
   What Data Mining Is Not .................................... 38
   The Most Common Data Mining Applications ............. 39
   What Kinds of Patterns Can Data Mining Discover? ...... 45
   Popular Data Mining Tools .................................. 51
   The Dark Side of Data Mining: Privacy Concerns ......... 57
   References .................................................... 65

Chapter 3 The Data Mining Process ............................... 67
   The Knowledge Discovery in Databases (KDD) Process .... 67
   Cross-Industry Standard Process for Data Mining (CRISP-DM) ..... 69
   SEMMA ..................................................... 78
   SEMMA Versus CRISP-DM .................................. 82
   Six Sigma for Data Mining ................................. 83
   Which Methodology Is Best? ................................ 86
   References .................................................... 91

Chapter 4 Data and Methods in Data Mining ..................... 93
   The Nature of Data in Data Mining ....................... 93
   Preprocessing of Data for Data Mining ................... 97
   Data Mining Methods ....................................... 103
Prediction .......................... 105
Classification .......................... 105
Decision Trees .......................... 114
Cluster Analysis for Data Mining .......................... 117
k-Means Clustering Algorithm .......................... 122
Association .......................... 123
Apriori Algorithm .......................... 127
Data Mining Misconceptions and Realities .......................... 129
References .......................... 139

Chapter 5  Data Mining Algorithms .......................... 141
Nearest Neighbor .......................... 142
Similarity Measure: The Distance Metric .......................... 144
Artificial Neural Networks .......................... 147
Support Vector Machines .......................... 159
Linear Regression .......................... 165
Logistic Regression .......................... 173
Time-Series Forecasting .......................... 175
References .......................... 181

Chapter 6  Text Analytics and Sentiment Analysis .......................... 183
Natural Language Processing .......................... 189
Text Mining Applications .......................... 194
The Text Mining Process .......................... 200
Text Mining Tools .......................... 213
Sentiment Analysis .......................... 215
References .......................... 228

Chapter 7  Big Data Analytics .......................... 231
Where Does Big Data Come From? .......................... 232
The Vs That Define Big Data .......................... 233
Fundamental Concepts of Big Data .......................... 238
The Business Problems That Big Data Analytics Addresses .......................... 244
Big Data Technologies .......................... 244
Data Scientists .......................... 254
Big Data and Stream Analytics .......................... 257
Data Stream Mining .......................... 260
References .......................... 264

Index .......................... 265
Dr. Dursun Delen has written a concise, information-rich book that effectively provides an excellent learning tool for someone who wants to understand what analytics, data mining, and “Big Data” are all about. As business becomes increasingly complex and global, decision-makers must act more rapidly and accurately, based on the best available evidence. Modern data mining and analytics are indispensable for doing this. This reference makes clear the current best practices, demonstrates to readers—students and practitioners—how to use data mining and analytics to uncover hidden patterns and correlations, and explains how to leverage these to improve decision-making.

The author delivers the right amount of concept, technique, and cases to help readers truly understand how data mining technologies work. Coverage includes data mining processes, methods, and techniques; the role and management of data; tools and metrics; text and web mining; sentiment analysis; and integration with cutting-edge Big Data approaches, as presented as follows.

In Chapter 1, he commendably traces the roots of analytics from World War II times to the present, illustrated by Figure 1.2, where he takes the reader from Decision Support Systems in the 1970s, to the Enterprise/Executive IS Systems in the 1980s, to the Business Intelligence (BI) that we all heard about in the 1990s and early 2000s, and finally to our modern day uses of analytics (2000s) and Big Data (2010s). That was all in Chapter 1, creating a preamble for what is to come in the rest of the book: data mining.

Chapter 2 provides a very easy-to-understand description and an excellent taxonomy for data mining. In this chapter, the author differentiates data mining from several other related terminologies, making a strong case for what it really stands for: discovery of knowledge. Identifying data mining as a problem-solving and
decision-making philosophy that sits in the midst of many disciplines is quite refreshing; many people think of data mining as a new discipline of its own. With a number of real-world examples, intuitive graphics, and down-to-earth discussion, this chapter demystifies data mining for the masses. In my opinion, this is an excellent way to portray seemingly complex and highly technical topics like data mining to a wider audience.

In Chapter 3, Dr. Delen provides a rich collection of different approaches to standardized data mining processes in a manner that any reader can understand. KDD (knowledge discovery in databases) is the first standardized process that the chapter talks about, which was developed by Usama Fayyad, an early pioneer in the field. Dr. Delen presents KDD in an engaging discussion enhanced with a diagram (Figure 3.1), which illustrates the flow of the KKD data mining process. Additionally, other data mining “schemas” proposed by various groups and individuals are examined to show the development of the fundamental thinking in this field. To illustrate the usefulness of these schemas, Dr. Delen presents a data mining case study at the end of this chapter: “Mining Cancer Data for New Knowledge.”

Chapter 4 considers the types of data used in data mining, including the ever-increasing use of text data (that is, unstructured non-numerical data, which is probably 90% of the data available to the world today). Data preparation is the most important part of data mining: Data must be clean and good in order to develop useful models (“garbage in, garbage out”); thus, up to 90% of the time involved in data mining can be taken up by the data preparation stage. Dr. Delen goes into all the ways of looking at data to get it clean and ready for data analytics, including developing the train and test data sets, giving one of the most learner-friendly visuals of $k$-fold cross-validation in Figure 4.6.

In Chapter 5, Dr. Delen describes the most common data mining algorithms in a way that the layperson can understand. Among others, neural networks and SVM (support vector machines) are described
thoroughly, with illustrations that help the reader really understand these complicated mathematical processes. Dr. Delen makes his own original illustrations, and they are well worth the price of the book!

Text mining (text analytics) is described thoroughly in Chapter 6, with Dr. Delen starting out with a diagram he originally made for our 2012 book, *Practical Data Mining* (on which I am the lead author: Miner, G.D.; Delen, D.; Elder, J.; Fast, A.; Hill, T.; and Nisbet, B. Elsevier/Academic Press: 2012). Dr. Delen effectively distills our large 1,100 page book into one chapter that tells it all—in other words, is very useful for the new learner. Well done!

In the last chapter, Chapter 7, Dr. Delen goes into the new buzz word in the analytics field: Big Data analytics. Big Data is heard in the news almost daily. What does it mean? It means different things to different people. But I can tell you that, working in the data mining field for more than 15 years now, I have been dealing with Big Data all that time. But the ever-decreasing cost of storage space for data and the availability of cloud storage, plus the availability of faster and faster computers, mean that even a small laptop can do both distributed processing and multi-threading in data analysis. This has made even the small tablet more powerful than the warehouse of air-conditioned mainframe servers of decades ago. One can even run a bank of servers and cloud storage from one’s smartphone these days. So as the data becomes “bigger,” the physical needs to process it become “smaller.”

But Big Data is misunderstood by most, at least it seems that way to me. Many think that data mining requires Big Data. But I have worked with medical residents for 10 years who want to look at lots of variables in their one-year research project but usually can only get a fraction of the cases they need for that many variables in their limited time; thus, traditional statistics are of almost no use to these paltry data sets by traditional statistics standards; yet by using machine-learning, modern data mining methods, I have found that one can usually generate useful hypotheses from these small data sets
and find knowledge that was previously impossible to obtain with only traditional $p$-value Fischerian statistics. Traditional statistics were an anomaly of the twentieth century; prior to 1900, Bayesian statistics had predominated in data analysis for centuries; with the advent of the year 2000, the new modern versions of Bayesian statistics—including the SVM, NN, and other machine-learning modalities—had come of age, and we are now back into the Bayesian age in this twenty-first century. Unfortunately, it is taking a while for the “traditionally statistical training” cadre to understand and catch up…but the cutting edge is with Bayesian, data mining, and Big Data.

Anyone wanting to learn about data mining and have a technical understanding of the topic should get this book. By the end of the read, you will understand the field!

—Gary D. Miner, Ph.D.
Author of two PROSE Award–winning analytics books
Senior Analyst, Healthcare Applications Specialist
Dell, Information Management Group, Dell Software
Without a doubt, we are living in the age of data mining and Big Data analytics. Because of their popularity (and perhaps a little bit of hype), everybody is talking about data mining and Big Data analytics, often in different scope and contexts, using diverse terminology. The main goal of this book is to explain the language of analytics and data mining in a comprehensive yet not-too-technical way. If I have, at least partially, succeeded in achieving this goal, it is because of the direct and indirect contributions of a number of people.

I want to thank my colleagues and my students for providing me with the broader perspective toward analytics that I needed to write this book in a holistic manner. As is the case for most academics, I also have my own opinions and biases toward “what is what” in data mining and in analytics, and thanks to my academic friends, I think I managed to rise above them to make this book inclusive and comprehensive.

We academics tend to be focused on rigor and theory, sometimes in the process moving away from relevance and reality. Thanks to my clients and corporate partners who continuously provide me with the realities of the real world that I need to stay balanced between rigor and relevance. Writing a book titled *Real-World Data Mining* requires such connection to reality, without compromising technical accuracy, and I have to thank to my corporate friends for helping me achieve that in this book.

I want to thank to my publisher, Ms. Jeanne Levine, and Pearson for presenting me with the opportunity to write this book and for being patient and resourceful for me throughout the journey of actually writing it.
Dr. Dursun Delen is an internationally known expert in business analytics and data mining. He is often invited to national and international conferences to deliver keynote presentations on topics related to data/text mining, business intelligence, decision support systems, business analytics, and knowledge management. Prior to his appointment as professor at Oklahoma State University in 2001, Dr. Delen worked for industry for more than 10 years, developing and delivering business analytics solutions to companies. Most recently he worked for a privately owned research and consulting company, Knowledge Based Systems, Inc., in College Station, Texas, as a research scientist. During his five years there, he led a number of projects related to decision support, information systems, and advanced analytics that were funded by federal agencies, including the DoD, NASA, NIST, and the DOE. Today, in addition to his academic endeavors, Dr. Delen provides consulting services to businesses in assessing their information systems needs and developing state-of-the-art business analytics capabilities.

Dr. Delen holds William S. Spears and Neal Patterson Endowed Chairs in Business Analytics, and he is Director of Research for the Center for Health Systems Innovation and Professor of Management Science and Information Systems in the Spears School of Business at Oklahoma State University. His research has appeared in major journals, including Decision Sciences, Decision Support Systems, Communications of the ACM, Computers and Operations Research, Computers in Industry, Journal of Production Operations Management, Artificial Intelligence in Medicine, and Expert Systems with Applications, among others. He has recently published six books: Advanced Data Mining Techniques (Springer, 2008), Decision Support and Business Intelligence Systems (Prentice Hall, 2010), Business
Intelligence: A Managerial Approach (Prentice Hall, 2010), Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications (Elsevier, 2012), Business Intelligence: A Managerial Perspective on Analytics, 3rd edition (Prentice Hall, 2013), and Business Intelligence and Analytics: Systems for Decision Support, 10th edition (Prentice Hall, 2014). He served as the general co-chair for the fourth International Conference on Network Computing and Advanced Information Management, and he regularly chairs tracks and mini-tracks at various information systems conferences. He also serves as associate editor-in-chief, senior editor, associate editor, and editorial board member on a dozen academic and technical journals.
Introduction to Analytics

Business analytics is a relatively new term that is gaining popularity in the business world like nothing else in recent history. In general terms, analytics is the art and science of discovering insight—by using sophisticated mathematical models along with a variety of data and expert knowledge—to support solid, timely decision making. In a sense, analytics is all about decision making and problem solving. These days, analytics can be defined as simply as “the discovery of meaningful patterns in data.” In this era of abundant data, analytics tends to be used on large quantities and varieties of data. Although analytics tends to be data focused, many applications of analytics involve very little or no data; instead, those analytics projects use mathematical models that rely on process description and expert knowledge (e.g., optimization and simulation models).

Business analytics is the application of the tools, techniques, and principles of analytics to complex business problems. Firms commonly apply analytics to business data to describe, predict, and improve business performance. Firms have used analytics in many ways, including the following:

- To improve their relationships with their customers (encompassing all phases of customer relationship management—acquisition, retention, and enrichment), employees, and other stakeholders
• To identify fraudulent transactions and odd behaviors—and, in doing so, saving money
• To enhance product and service features and their pricing, which would lead to better customer satisfaction and profitability
• To optimize marketing and advertising campaigns so they can reach more customers with the right kind of message and promotions with the least amount of expense
• To minimize operational costs by optimally managing inventories and allocating resources wherever and whenever they are needed by using optimization and simulation modeling
• To empower employees with the information and insight they need to make faster and better decisions while they are working with customers or customer-related issues

The term analytics, perhaps because of its rapidly increasing popularity as a buzzword, is being used to replace several previously popular terms, such as intelligence, mining, and discovery. For example, the term business intelligence has now become business analytics; customer intelligence has become customer analytics, Web mining has become Web analytics, knowledge discovery has become data analytics, etc. Modern-day analytics can require extensive computation because of the volume, variety, and velocity of data (which we call Big Data). Therefore, the tools, techniques, and algorithms used for analytics projects leverage the most current, state-of-the-art methods developed in a variety of fields, including management science, computer science, statistics, data science, and mathematics. Figure 1.1 shows a word cloud that includes concepts related to analytics and Big Data.
Is There a Difference Between Analytics and Analysis?

Even though the two terms analytics and analysis are often used interchangeably, they are not the same.

Basically, analysis refers to the process of separating a whole problem into its parts so that the parts can be critically examined at the granular level. It is often used when the investigation of a complete system is not feasible or practical, and the system needs to be simplified by being decomposed into more basic components. Once the improvements at the granular level are realized and the examination of the parts is complete, the whole system (either a conceptual or physical system) can then be put together using a process called synthesis.

Analytics, on the other hand, is a variety of methods, technologies, and associated tools for creating new knowledge/insight to solve complex problems and make better and faster decisions. In essence, analytics is a multifaceted and multidisciplinary approach to addressing complex situations. Analytics take advantage of data and mathematical models to make sense of the complicated world we are living
in. Even though analytics includes the act of analysis at different stages of the discovery process, it is not just analysis but also includes synthesis and other complementing tasks and processes.

**Where Does Data Mining Fit In?**

Data mining is the process of discovering new knowledge in the forms of patterns and relationships in large data sets. The goal of analytics is to convert data/facts into actionable insight, and data mining is the key enabler of that goal. Data mining has been around much longer than analytics, at least in the context of analytics today. As analytics became an overarching term for all decision support and problem-solving techniques and technologies, data mining found itself a rather large space within that arc, ranging from descriptive exploration of identifying relationships and affinities among variables (e.g., market-basket analysis) to developing models to estimate future values of interesting variables. As we will see later in this chapter, within the taxonomy of analytics, data mining plays a key role at every level, from the most simple to the most sophisticated.

**Why the Sudden Popularity of Analytics?**

Analytics is a buzzword of business circles today. No matter what business journal or magazine you look at, it is very likely that you will see articles about analytics and how analytics is changing the way managerial decisions are being made. It has become a new label for evidence-based management (i.e., evidence/data-driven decision making). But why has analytics become so popular? And why now? The reasons (or forces) behind this popularity can be grouped into three categories: need, availability and affordability, and culture change.
**Need**

As we all know, business is anything but “as usual” today. Competition has been characterized progressively as local, then regional, then national, but it is now global. Large to medium to small, every business is under the pressure of global competition. The tariff and transportation cost barriers that sheltered companies in their respective geographic locations are no longer as protective as they once were. In addition to (and perhaps because of) the global competition, customers have become more demanding. They want the highest quality of products and/or services with the lowest prices in the shortest possible time. Success or mere survival depends on businesses being agile and their managers making the best possible decisions in a timely manner to respond to market-driven forces (i.e., rapidly identifying and addressing problems and taking advantage of the opportunities). Therefore, the need for fact-based, better, and faster decisions is more critical now than ever before. In the midst of these unforgiving market conditions, analytics is promising to provide managers the insights they need to make better and faster decisions, which help improve their competitive posture in the marketplace. Analytics today is widely perceived as saving business managers from the complexities of global business practices.

**Availability and Affordability**

Thanks to recent technological advances and the affordability of software and hardware, organizations are collecting tremendous amounts of data. Automated data collections systems—based on a variety of sensors and RFID—have significantly increased the quantity and quality of organizational data. Coupled with the content-rich data collected from Internet-based technologies such as social media, businesses now tend to have more data than they can handle. As the saying goes, “They are drowning in data but starving for knowledge.”
Along with data collection technologies, data processing technologies have also improved significantly. Today’s machines have numerous processors and very large memory capacities, so they are able to process very large and complex data in a reasonable time frame—often in real time. The advances in both hardware and software technology are also reflected in the pricing, continuously reducing the cost of ownership for such systems. In addition to the ownership model, along came the software- (or hardware-) as-a-service business model, which allows businesses (especially small to medium-size businesses with limited financial power) to rent analytics capabilities and pay only for what they use.

**Culture Change**

At the organizational level, there has been a shift from old-fashioned intuition-driven decision making to new-age fact-/evidence-based decision making. Most successful organizations have made a conscious effort to shift to data-/evidence-driven business practices. Because of the availability of data and supporting IT infrastructure, such a paradigm shift is taking place faster than many thought it would. As the new generation of quantitatively savvy managers replaces the baby boomers, this evidence-based managerial paradigm shift will intensify.

**The Application Areas of Analytics**

Even though the business analytics wave is somewhat new, there are numerous applications of analytics, covering almost every aspect of business practice. For instance, in customer relationship management, a wealth of success stories tell of sophisticated models developed to identify new customers, look for up-sell/cross-sell opportunities, and find customers’ with a high propensity toward attrition. Using social media analytics and sentiment analysis, businesses are trying
to stay on top of what people are saying about their products/services and brands. Fraud detection, risk mitigation, product pricing, marketing campaign optimization, financial planning, employee retention, talent recruiting, and actuarial estimation are among the many business applications of analytics. It would be very hard to find a business issue where a number of analytics application could not be found. From business reporting to data warehousing, from data mining to optimization analytics, techniques are used widely in almost every facet of business.

The Main Challenges of Analytics

Even though the advantages of and reasons for analytics are evident, many businesses are still hesitant to jump on the analytics bandwagon. These are the main roadblocks to adoption of analytics:

- **Analytics talent.** Data scientists, the quantitative geniuses who can convert data into actionable insight, are scarce in the market; the really good ones are very hard to find. Because analytics is relatively new, the talent for analytics is still being developed. Many colleges have started undergraduate and graduate programs to address the analytics talent gap. As the popularity of analytics increases, so will the need for people who have the knowledge and skills to convert Big Data into information and knowledge that managers and other decision makers need to tackle real-world complexities.

- **Culture.** As the saying goes, “Old habits die hard.” Changing from a traditional management style (often characterized by intuition as the basis of making decision) to a contemporary management style (based on data and scientific models for managerial decisions and collective organizational knowledge) is not an easy process to undertake for any organization. People do not like to change. Change means losing what you have
learned or mastered in the past and now needing to learn how to do what you do all over again. It suggests that the knowledge (which is also characterized as power) you’ve accumulated over the years will disappear or be partially lost. The culture shift may be the most difficult part of adopting analytics as the new management paradigm.

- **Return on investment.** Another barrier to adoption of analytics is the difficulty in clearly justifying its return on investment (ROI). Analytics projects are complex and costly endeavors, and their return is not immediately clear, many executives are having a hard time investing in analytics, especially on large scales. Will the value gained from analytics outweigh the investment? If so, when? It is very hard to convert the value of analytics into justifiable numbers. Most of the value gained from analytics is somewhat intangible and holistic. If done properly, analytics could transform an organization, putting it on a new and improved level. A combination of tangible and intangible factors needs to be brought to bear to numerically rationalize investment and movement toward analytics and analytically savvy management practice.

- **Data.** The media is taking about Big Data in a very positive way, characterizing it as an invaluable asset for better business practices. This is mostly true, especially if the business understands and knows what to do with it. For those who have no clue, Big Data is a big challenge. Big Data is not just big; it is unstructured, and it is arriving at a speed that prohibits traditional collection and processing means. And it is usually messy and dirty. For an organization to succeed in analytics, it needs to have a well-thought-out strategy for handling Big Data so that it can be converted to actionable insight.
• **Technology.** Even though technology is capable, available, and, to some extent, affordable, technology adoption poses another challenge for traditionally less technical businesses. Although establishing an analytics infrastructure is affordable, it still costs a significant amount of money. Without financial means and/or a clear return on investment, management of some businesses may not be willing to invest in needed technology. For some businesses, an analytics-as-a-service model (which includes both software and the infrastructure/hardware needed to implement analytics) may be less costly and easier to implement.

• **Security and privacy.** One of the most common criticisms of data and analytics is the security. We often hear about data breaches of sensitive information, and indeed, the only completely secured data infrastructure is isolated and disconnected from all other networks (which goes against the very reason for having data and analytics). The importance of data security has made information assurance one of the most popular concentration areas in information systems departments around the world. At the same time that increasingly sophisticated techniques are being used to protect the information infrastructure, increasingly sophisticated attacks are becoming common. There are also concerns about personal privacy. Use of personal data about customers (existing or prospective), even if it is within legal boundaries, should be avoided or carefully scrutinized to protect an organization against bad publicity and public outcry.

Despite the hurdles in the way, analytics adoption is growing, and analytics is inevitable for today’s enterprises, regardless of size or industry segment. As the complexity in conducting business increases, enterprises are trying to find order in the midst of the chaotic behaviors. The ones that succeed will be the ones fully leveraging the capabilities of analytics.
A Longitudinal View of Analytics

Although the buzz about it is relatively recent, analytics isn’t new. It’s possible to find references to corporate analytics as far back as the 1940s, during the World War II era, when more effective methods were needed to maximize output with limited resources. Many optimization and simulation techniques were developed then. Analytical techniques have been used in business for a very long time. One example is the time and motion studies initiated by Frederick Winslow Taylor in the late 19th century. Then Henry Ford measured pacing of assembly lines, which led to mass-production initiatives. Analytics began to command more attention in the late 1960s, when computers were used in decision support systems. Since then, analytics has evolved with the development of enterprise resource planning (ERP) systems, data warehouses, and a wide variety of other hardware and software tools and applications.

The timeline in Figure 1.2 shows the terminology used to describe analytics since the 1970s. During the early days of analytics, prior to the 1970s, data was often obtained from the domain experts using manual processes (i.e., interviews and surveys) to build mathematical or knowledge-based models to solve constraint optimization problems. The idea was to do the best with limited resources. Such decision support models were typically called operations research (OR). The problems that were too complex to solve optimally (using linear or non-linear mathematical programming techniques) were tackled using heuristic methods such as simulation models.
In the 1970s, in addition to the mature OR models that were being used in many industries and government systems, a new and exciting line of models had emerged: rule-based expert systems (ESs). These systems promised to capture experts’ knowledge in a format that computers could process (via a collection of if–then rules) so that they could be used for consultation much the same way that one would use domain experts to identify a structured problem and to prescribe the most probable solution. ESs allowed scarce expertise to be made available where and when needed, using an “intelligent” decision support system. During the 1970s, businesses also began to create routine reports to inform decision makers (managers) about what had happened in the previous period (e.g., day, week, month, quarter). Although it was useful to know what had happened in the past, managers needed more than this: They needed a variety of reports at different levels of granularity to better understand and address changing needs and challenges of the business.

The 1980s saw a significant change in the way organizations captured business-related data. The old practice had been to have multiple disjointed information systems tailored to capture transactional data of different organizational units or functions (e.g., accounting, marketing and sales, finance, manufacturing). In the 1980s, these systems were integrated as enterprise-level information systems that we now commonly call ERP systems. The old mostly sequential and
nonstandardized data representation schemas were replaced by relational database management (RDBM) systems. These systems made it possible to improve the capture and storage of data, as well as the relationships between organizational data fields while significantly reducing the replication of information. The need for RDBM and ERP system emerged when data integrity and consistency became an issue, significantly hindering the effectiveness of business practices. With ERP, all the data from every corner of the enterprise is collected and integrated into a consistent schema so that every part of the organization has access to the single version of the truth when and where needed. In addition to the emergence of ERP systems—or perhaps because of these systems—business reporting became an on-demand, as-needed business practice. Decision makers could decide when they needed to or wanted to create specialized reports to investigate organizational problems and opportunities.

In the 1990s, the need for more versatile reporting led to the development of executive information systems (decision support systems designed and developed specifically for executives and their decision-making needs). These systems were designed as graphical dashboards and scorecards so that they could serve as visually appealing displays while focusing on the most important factors for decision makers to keep track of—the key performance indicators. In order to make this highly versatile reporting possible while keeping the transactional integrity of the business information systems intact, it was necessary to create a middle data tier—known as a data warehouse (DW)—as a repository to specifically support business reporting and decision making. In a very short time, most large to medium-size businesses adopted data warehousing as their platform for enterprise-wide decision making. The dashboards and scorecards got their data from a data warehouse, and by doing so, they were not hindering the efficiency of the business transaction systems—mostly referred to as enterprise resource planning (ERP) systems.
In the 2000s the DW-driven decision support systems began to be called *business intelligence systems*. As the amount of longitudinal data accumulated in the DWs increased, so did the capabilities of hardware and software to keep up with the rapidly changing and evolving needs of the decision makers. Because of the globalized competitive marketplace, decision makers needed current information in a very digestible format to address business problems and to take advantage of market opportunities in a timely manner. Because the data in a DW is updated periodically, it does not reflect the latest information. In order to elevate this information latency problem, DW vendors developed a system to update the data more frequently, which led to the terms *real-time data warehousing* and, more realistically, *right-time data warehousing*, which differs from the former by adopting a data refreshing policy based on the needed freshness of the data items (i.e., not all data items need to be refreshed in real time). Data warehouses are very large and feature rich, and it became necessary to “mine” the corporate data to “discover” new and useful knowledge nuggets to improve business processes and practices—hence the terms *data mining* and *text mining*. With the increasing volumes and varieties of data, the needs for more storage and more processing power emerged. While large corporations had the means to tackle this problem, small to medium-size companies needed financially more manageable business models. This need led to service-oriented architecture and software and infrastructure-as-a-service analytics business models. Smaller companies therefore gained access to analytics capabilities on an as-needed basis and paid only for what they used, as opposed to investing in financially prohibitive hardware and software resources.

In the 2010s we are seeing yet another paradigm shift in the way that data is captured and used. Largely because of the widespread use of the Internet, new data-generation mediums have emerged. Of all the new data sources (e.g., RFID tags, digital energy meters, click-stream Web logs, smart home devices, wearable health monitoring
equipment), perhaps the most interesting and challenging is social networking/social media. This unstructured data is rich in information content, but analysis of such data sources poses significant challenges to computational systems, from both software and hardware perspectives. Recently, the term *Big Data* has been coined to highlight the challenges that these new data streams have brought upon us. Many advancements in both hardware (e.g., massively parallel processing with very large computational memory and highly parallel multiprocessor computing systems) and software/algorithms (e.g., Hadoop with MapReduce and NoSQL) have been developed to address the challenges of Big Data.

It’s hard to predict what the next decade will bring and what the new analytics-related terms will be. The time between new paradigm shifts in information systems and particularly in analytics has been shrinking, and this trend will continue for the foreseeable future. Even though analytics is not new, the explosion in its popularity is very new. Thanks to the recent explosion in Big Data, ways to collect and store this data, and intuitive software tools, data and data-driven insight are more accessible to business professionals than ever before. Therefore, in the midst of global competition, there is a huge opportunity to make better managerial decisions by using data and analytics to increase revenue while decreasing costs by building better products, improving customer experience, and catching fraud before it happens, improving customer engagement through targeting and customization—all with the power of analytics and data. More and more companies are now preparing their employees with the know-how of business analytics to drive effectiveness and efficiency in their day-to-day decision-making processes.
A Simple Taxonomy for Analytics

Because of the multitude of factors related to both the need to make better and faster decisions and the availability and affordability of hardware and software technologies, analytics is gaining popularity faster than any other trends we have seen in recent history. Will this upward exponential trend continue? Many industry experts think it will, at least for the foreseeable future. Some of the most respected consulting companies are projecting that analytics will grow at three times the rate of other business segments in upcoming years; they have also named analytics as one of the top business trends of this decade (Robinson et al., 2010). As interest in and adoption of analytics have grown rapidly, a need to characterize analytics into a simple taxonomy has emerged. The top consulting companies (e.g., Accenture, Gartner, and IDT) and several technologically oriented academic institutions have embarked on a mission to create a simple taxonomy for analytics. Such a taxonomy, if developed properly and adopted universally, could create a contextual description of analytics, thereby facilitating a common understanding of what analytics is, including what is included in analytics and how analytics-related terms (e.g., business intelligence, predictive modeling, data mining) relate to each other. One of the academic institutions involved in this challenge is INFORMS (Institute for Operations Research and Management Science). In order to reach a wide audience, INFORMS hired Capgemini, a strategic management consulting firm, to carry out a study and characterize analytics.

The Capgemini study produced a concise definition of analytics: “Analytics facilitates realization of business objectives through reporting of data to analyze trends, creating predictive models for forecasting and optimizing business processes for enhanced performance.” As this definition implies, one of the key findings from the study is that executives see analytics as a core function of businesses that use it. It spans many departments and functions within organizations, and in mature organizations, it spans the entire business. The
study identified three hierarchical but sometimes overlapping groupings for analytics categories: descriptive, predictive, and prescriptive analytics. These three groups are hierarchical in terms of the level of analytics maturity of the organization. Most organizations start with descriptive analytics, then move into predictive analytics, and finally reach prescriptive analytics, the top level in the analytics hierarchy. Even though these three groupings of analytics are hierarchical in complexity and sophistication, moving from a lower level to a higher level is not clearly separable. That is, a business can be in the descriptive analytics level while at the same time using predictive and even prescriptive analytics capabilities, in a somewhat piecemeal fashion. Therefore, moving from one level to the next essentially means that the maturity at one level is completed and the next level is being widely exploited. Figure 1.3 shows a graphical depiction of the simple taxonomy developed by INFORMS and widely adopted by most industry leaders as well as academic institutions.

![Figure 1.3 A Simple Taxonomy for Analytics](image-url)
Descriptive analytics is the entry level in analytics taxonomy. It is often called business reporting because of the fact that most of the analytics activities at this level deal with creating reports to summarize business activities in order to answer questions such as “What happened?” and “What is happening?” The spectrum of these reports includes static snapshots of business transactions delivered to knowledge workers (i.e., decision makers) on a fixed schedule (e.g., daily, weekly, quarterly); dynamic views of business performance indicators delivered to managers and executives in an easily digestible form—often in a dashboard-looking graphical interface—on a continuous manner; and ad hoc reporting where the decision maker is given the capability of creating his or her own specific report (using an intuitive drag-and-drop graphical user interface) to address a specific or unique decision situation.

Descriptive analytics is also called business intelligence (BI), and predictive and prescriptive analytics are collectively called advanced analytics. The logic here is that moving from descriptive to predictive and/or prescriptive analytics is a significant shift in the level of sophistication and therefore warrants the label advanced. BI has been one of the most popular technology trends for information systems designed to support managerial decision making since the start of the century. It was popular (to some extent, it still is in some business circles) until the arrival of the analytics wave. BI is the entrance to the world of analytics, setting the stage and paving the way toward more sophisticated decision analysis. Descriptive analytics systems usually work off a data warehouse, which is a large database specifically designed and developed to support BI functions and tools.

Predictive analytics comes right after descriptive analytics in the three-level analytics hierarchy. Organizations that are mature in descriptive analytics move to this level, where they look beyond what happened and try to answer the question “What will happen?” In the following chapters, we will cover the predictive capabilities of these analytics techniques in depth as part of data mining; here we provide
only a very short description of the main predictive analytics classes. Prediction essentially is the process of making intelligent/scientific estimates about the future values of some variables, like customer demand, interest rates, stock market movements, etc. If what is being predicted is a categorical variable, the act of prediction is called classification; otherwise, it is called regression. If the predicted variable is time dependent, the prediction process is often called time-series forecasting.

Prescriptive analytics is the highest echelon in analytics hierarchy. It is where the best alternative among many courses of action—that are usually created/identified by predictive and/or descriptive analytics—is determined using sophisticated mathematical models. Therefore, in a sense, this type of analytics tries to answer the question “What should I do?” Prescriptive analytics uses optimization-, simulation-, and heuristics-based decision-modeling techniques. Even though prescriptive analytics is at the top of the analytics hierarchy, the methods behind it are not new. Most of the optimization and simulation models that constitute prescriptive analytics were developed during and right after World War II, when there was a dire need for a lot with limited resources. Since then, some businesses have used these models for some very specific problem types, including yield/revenue management, transportation modeling, scheduling, etc. The new taxonomy of analytics has made them popular again, opening their use to a wide array of business problems and situations.

Figure 1.4 shows a tabular representation of the three hierarchical levels of analytics, along with the questions answered and techniques used at each level. As can be seen data mining is the key enabler of predictive analytics.
### Figure 1.4 Three Levels of Analytics and Their Enabling Techniques

Business analytics is gaining popularity because it promises to provide decision makers with information and knowledge that they need to succeed. Effectiveness of business analytics systems, no matter the level in the analytics hierarchy, depends largely on the quality and quantity of the data (volume and representational richness); the accuracy, integrity, and timeliness of the data management system; and the capabilities and sophistication of the analytical tools and procedures used in the process. Understanding the analytics taxonomy helps organizations to be smart about selecting and implementing analytics capabilities to efficiently navigate through the maturity continuum.
IBM Watson is perhaps the smartest computer system built to date. Since the emergence of computers and subsequently artificial intelligence in the late 1940s, scientists have compared the performance of these “smart” machines with human minds. Accordingly, in the mid- to late 1990s, IBM researchers built a smart machine and used the game of chess (generally credited as the game of smart humans) to test their ability against the best of human players. On May 11, 1997, an IBM computer called Deep Blue beat the world chess grandmaster after a six-game match series: two wins for Deep Blue, one for the champion, and three draws. The match lasted several days and received massive media coverage around the world. It was the classic plot line of human versus machine. Beyond the chess contest, the intention of developing this kind of computer intelligence was to make computers able to handle the kinds of complex calculations needed to help discover new medical drugs, do the broad financial modeling needed to identify trends and do risk analysis, handle large database searches, and perform massive calculations needed in advanced fields of science.

After a couple decades, IBM researchers came up with another idea that was perhaps more challenging: a machine that could not only play Jeopardy! but beat the best of the best. Compared to chess, Jeopardy! is much more challenging. While chess is well structured and has very simple rules, and therefore is very good match for computer processing, Jeopardy! is neither simple nor structured. Jeopardy! is a game designed for human intelligence and creativity, and therefore a computer designed to play it needed to be a cognitive computing system that can work and think like a human. Making sense of imprecision inherent in human language was the key to success.

In 2010 an IBM research team developed Watson, an extraordinary computer system—a novel combination of advanced hardware
and software—designed to answer questions posed in natural human language. The team built Watson as part of the DeepQA project and named it after IBM’s first president, Thomas J. Watson. The team that built Watson was looking for a major research challenge: one that could rival the scientific and popular interest of Deep Blue and would also have clear relevance to IBM’s business interests. The goal was to advance computational science by exploring new ways for computer technology to affect science, business, and society at large. Accordingly, IBM Research undertook a challenge to build Watson as a computer system that could compete at the human champion level in real time on the American TV quiz show *Jeopardy!* The team wanted to create a real-time automatic contestant on the show, capable of listening, understanding, and responding—not merely a laboratory exercise.

**Competing Against the Best at Jeopardy!**

In 2011, as a test of its abilities, Watson competed on the quiz show *Jeopardy!* in the first-ever human-versus-machine matchup for the show. In a two-game, combined-point match (broadcast in three *Jeopardy!* episodes during February 14–16), Watson beat Brad Rutter, the biggest all-time money winner on *Jeopardy!* and Ken Jennings, the record holder for the longest championship streak (75 days). In these episodes, Watson consistently outperformed its human opponents on the game’s signaling device, but it had trouble responding to a few categories, notably those having short clues containing only a few words. Watson had access to 200 million pages of structured and unstructured content, consuming 4 terabytes of disk storage. During the game, Watson was not connected to the Internet.

Meeting the *Jeopardy!* challenge required advancing and incorporating a variety of text mining and natural language processing technologies, including parsing, question classification, question decomposition, automatic source acquisition and evaluation, entity
and relationship detection, logical form generation, and knowledge representation and reasoning. Winning at Jeopardy! requires accurately computing confidence in answers. The questions and content are ambiguous and noisy, and none of the individual algorithms are perfect. Therefore, each component must produce a confidence in its output, and individual component confidences must be combined to compute the overall confidence of the final answer. The final confidence is used to determine whether the computer system should risk choosing to answer at all. In Jeopardy! parlance, this confidence is used to determine whether the computer will “ring in” or “buzz in” for a question. The confidence must be computed during the time the question is read and before the opportunity to buzz in. This is roughly between one and six seconds, with an average around three seconds.

**How Does Watson Do It?**

The system behind Watson, which is called DeepQA, is a massively parallel, text mining–focused, probabilistic evidence-based computational architecture. For the Jeopardy! challenge, Watson used more than 100 different techniques for analyzing natural language, identifying sources, finding and generating hypotheses, finding and scoring evidence, and merging and ranking hypotheses. What is far more important than any particular technique the IBM team used was how it combined them in DeepQA such that overlapping approaches could bring their strengths to bear and contribute to improvements in accuracy, confidence, and speed.

DeepQA is an architecture with an accompanying methodology that is not specific to the Jeopardy! challenge. These are the overarching principles in DeepQA:

- **Massive parallelism.** Watson needed to exploit massive parallelism in the consideration of multiple interpretations and hypotheses.
• **Many experts.** Watson needed to be able to integrate, apply, and contextually evaluate a wide range of loosely coupled probabilistic question and content analytics.

• **Pervasive confidence estimation.** No component of Watson commits to an answer; all components produce features and associated confidences, scoring different question and content interpretations. An underlying confidence-processing substrate learns how to stack and combine the scores.

• **Integration of shallow and deep knowledge.** Watson needed to balance the use of strict semantics and shallow semantics, leveraging many loosely formed ontologies.

Figure 1.5 illustrates the DeepQA architecture at a very high level. More technical details about the various architectural components and their specific roles and capabilities can be found in Ferrucci et al. (2010).

![Figure 1.5 A High-Level Depiction of DeepQA Architecture](image)

**What Is the Future for Watson?**

The *Jeopardy!* challenge helped IBM address requirements that led to the design of the DeepQA architecture and the implementation of Watson. After three years of intense research and development
by a core team of about 20 researchers, as well as a significant R&D budget, Watson managed to perform at human expert levels in terms of precision, confidence, and speed at the Jeopardy! quiz show.

After the show, the big question was “So what now?” Was developing Watson all for a quiz show? Absolutely not! Showing the rest of the world what Watson (and the cognitive system behind it) could do became an inspiration for the next generation of intelligent information systems. For IBM, it was a demonstration of what is possible with cutting-edge analytics and computational sciences. The message is clear: If a smart machine can beat the best of the best in humans at what they are the best at, think about what it can do for your organizational problems. The first industry that utilized Watson was health care, followed by security, finance, retail, education, public services, and research. The following sections provide short descriptions of what Watson can do (and, in many cases, is doing) for these industries.

Health Care

The challenges that health care is facing today are rather big and multifaceted. With the aging U.S. population, which may be partially attributed to better living conditions and advanced medical discoveries fueled by a variety of technological innovations, demand for health care services is increasing faster than the supply of resources. As we all know, when there is an imbalance between demand and supply, the prices go up and the quality suffers. Therefore, we need cognitive systems like Watson to help decision makers optimize the use of their resources, both in clinical and managerial settings.

According to health care experts, only 20% of the knowledge physicians use to diagnose and treat patients is evidence based. Considering that the amount of medical information available is doubling every five years and that much of this data is unstructured, physicians simply don’t have time to read every journal that can help them keep up-to-date with the latest advances. Given the growing demand for services
and the complexity of medical decision making, how can health care providers address these problems? The answer could be to use Watson, or some other cognitive systems like Watson that has the ability to help physicians in diagnosing and treating patients by analyzing large amounts of data—both structured data coming from electronic medical record databases and unstructured text coming from physician notes and published literature—to provide evidence for faster and better decision making. First, the physician and the patient can describe symptoms and other related factors to the system in natural language. Watson can then identify the key pieces of information and mine the patient’s data to find relevant facts about family history, current medications, and other existing conditions. It can then combine that information with current findings from tests, and then it can form and test hypotheses for potential diagnoses by examining a variety of data sources—treatment guidelines, electronic medical record data and doctors’ and nurses’ notes, and peer-reviewed research and clinical studies. Next, Watson can suggest potential diagnostics and treatment options, with a confidence rating for each suggestion.

Watson also has the potential to transform health care by intelligently synthesizing fragmented research findings published in a variety of outlets. It can dramatically change the way medical students learn. It can help healthcare managers to be proactive about the upcoming demand patterns, optimally allocate resources, and improve processing of payments. Early examples of leading health care providers that use Watson-like cognitive systems include MD Anderson, Cleveland Clinic, and Memorial Sloan Kettering.

Security

As the Internet expands into every facet of our lives—ecommerce, ebusiness, smart grids for energy, smart homes for remote control of residential gadgets and appliances—to make things easier to manage, it also opens up the potential for ill-intended people to intrude in our lives. We need smart systems like Watson that are capable of
constantly monitoring for abnormal behavior and, when it is identified, preventing people from accessing our lives and harming us. This could be at the corporate or even national security system level; it could also be at the personal level. Such a smart system could learn who we are and become a digital guardian that could make inferences about activities related to our life and alert us whenever abnormal things happen.

**Finance**

The financial services industry faces complex challenges. Regulatory measures, as well as social and governmental pressures for financial institutions to be more inclusive, have increased. And the customers the industry serves are more empowered, demanding, and sophisticated than ever before. With so much financial information generated each day, it is difficult to properly harness the right information to act on. Perhaps the solution is to create smarter client engagement by better understanding risk profiles and the operating environment. Major financial institutions are already working with Watson to infuse intelligence into their business processes. Watson is tackling data-intensive challenges across the financial services sector, including banking, financial planning, and investing.

**Retail**

Retail industry is rapidly changing with the changing needs and wants of customers. Customers, empowered by mobile devices and social networks that give them easier access to more information faster than ever before, have high expectations for products and services. While retailers are using analytics to keep up with those expectations, their bigger challenge is efficiently and effectively analyzing the growing mountain of real-time insights that could give them the competitive advantage. Watson’s cognitive computing capabilities related to analyzing massive amounts of unstructured data can help
retailers reinvent their decision-making processes around pricing, purchasing, distribution, and staffing. Because of Watson’s ability to understand and answer questions in natural language, it is an effective and scalable solution for analyzing and responding to social sentiment based on data obtained from social interactions, blogs, and customer reviews.

**Education**

With the rapidly changing characteristics of students—more visually oriented/stimulated, constantly connected to social media and social networks, increasingly shorter attention spans—what should the future of education and the classroom look like? The next generation of educational system should be tailored to fit the needs of the new generation, with customized learning plans, personalized textbooks (digital ones with integrated multimedia—audio, video, animated graphs/charts, etc.), dynamically adjusted curriculum, and perhaps smart digital tutors and 24/7 personal advisors. Watson seems to have what it takes to make all this happen. With its natural language processing capability, students can converse with it just as they do with their teachers, advisors, and friends. This smart assistant can answer students’ questions, satisfy their curiosity, and help them keep up with the endeavors of the educational journey.

**Government**

For local, regional, and national governments, the exponential rise of Big Data presents an enormous dilemma. Today’s citizens are more informed and empowered than ever before, and that means they have high expectations for the value of the public sector serving them. And government organizations can now gather enormous volumes of unstructured, unverified data that could serve their citizens—but only if that data can be analyzed efficiently and effectively. IBM Watson’s cognitive computing may help make sense of
this data deluge, speeding governments’ decision-making processes and helping public employees focus on innovation and discovery.

Research

Every year, hundreds of billions of dollars are spent on research and development, most of it documented in patents and publications, creating enormous amount of unstructured data. To contribute to the extant knowledgebase, one needs to sift through these data sources to find the outer boundaries of research in a particular field. This is very difficult, if not impossible work, if it is done with traditional means, but Watson can act as a research assistant to help collect and synthesize information to keep people updated on recent findings and insights. For instance, New York Genome Center is using the IBM Watson cognitive computing system to analyze the genomic data of patients diagnosed with a highly aggressive and malignant brain cancer and to more rapidly deliver personalized, life-saving treatment to patients with this disease (Royyuru, 2014).

References


This page intentionally left blank
Index

Numerics

1-of-N pseudo variables, 97

A

accuracy
of classification models, 107-112
of time-series forecasting methods, assessing, 176-177
AcaXiom, 58
advanced analytics, 17
affinity analysis. See association rule mining
affordability of analytics, 5-6
aggregating data, 74
AIC (Akaike information criterion), 118
algorithms, 104, 141
Apriori algorithm, 127-128
for association rule mining, 127
back-propagation, 155-157
for decision tree creation, 116
genetic algorithms, 114
k-means clustering algorithm, 122-123
k-nearest neighbor algorithm, 113, 142-143
cross-validation, 145-147
Minkowski distance, finding, 144-145
parameter selection, 145
similarity measure, 144-147
learning algorithms, 156-157
logistic regression, 173-175
variables, 97
analytics
advanced analytics, 17
versus analysis, 3-4
Big Data, 14
business analytics, 1-2
business applications, 6-7
cancer survivability studies, 87-90
data mining, 4
descriptive analytics, 17
hierarchy of, 16
history of, 10-14
data warehouses, 12-13
ERP, 10
ERP systems, 12
executive information systems, 12
rule-based ESs, 11
IBM Watson, 20-28
DeepQA, 22-23
future of, 23-28
Jeopardy! challenge, 21-23
in-database analytics, 242
in-memory analytics, 242
predictive analytics, 18, 105
privacy issues, 58-59
prescriptive analytics, 18
reasons for popularity
affordability, 5-6
availability, 5-6
culture change, 6
need, 5
roadblocks to adoption, 7-9
culture, 7-8
data, 8
privacy, 9
return on investment, 8
security, 9
talent, 7
technology, 9
stream analytics, 257-259
taxonomy for, 15-19
terminology, 2
text analytics, 183-184
ANNs (artificial neural networks), 147-158.
See also neural networks
back-propagation algorithm, 155-157
and biological neural networks, 149-152
feed-forward neural networks, 154
MLP architecture, 154-155
network structure, 153-154
neurons, 152
processing elements, 153
sensitivity analysis, 157-158
versus SVMs, 164-165
appliances, 242
application examples
Big Data for political campaigns, 261-263
data mining for complex medical procedures, 177-180
data mining for Hollywood managers, 60-65
methodology, 62
results, 63
sample data, 60-61
predicting NCAA bowl game results, 130-139
evaluation, 136-137
methodology, 131-132
results, 137-139
sample data, 132-136
text mining of research literature, 209-213
text-based deception detection, 225-227
Apriori algorithm, 127-128
area under the ROC curve, 112
association rule mining, 123-127
applications, 124-125
Apriori algorithm, 127-128
associations, 45, 49
assumptions in linear regression, 171-173
attributes, 114
automated data collection, 5-6
availability of analytics, 5-6

B
back-propagation algorithm, 155-157
bag-of-words, 189-190
banking, data mining applications, 40
Bayesian classifiers, 114
benefits of Hadoop, 250-251
BI (business intelligence), 17
BIC (Bayesian information criterion), 118
Big Data, 14, 231, 238-243
challenges to Big Data analytics implementation, 242-243
characteristics of
value proposition, 238
variability, 237-238
variety, 236
velocity, 236-237
veracity, 237
volume, 234-236
critical success factors, 240-241
data scientists, 254-257
educational requirements, 256
skill requirements, 257
high-performance computing, 242
in politics, 261-263
problems addressed by, 244
sources of, 232-233
stream analytics, 257-259
technologies, 244-254
Hadoop, 247-253
MapReduce, 245-247
NoSQL, 254
terminology, 2
binary frequencies, 205
biological neural networks, 148-152
BioText, 199
bootstrapping, 111
branches, 114
brokerages, data mining applications, 41
building
decision trees, 115-116
algorithms, 116
splitting indices, 116
linear regression models, 167-171
SVM models, 161-164
data preprocessing, 162-163
model deployment, 164
model development, 163
business analytics, 1-2, 19
business applications for analytics, 6-7
business intelligence, 2
OLAP, 38

C
cancer survivability, analytic studies in, 87-90
Capgemini, 15-16
case-based reasoning, 113
categorical data, 94-95
challenges of analytics adoption, 7-9
Big Data analytics adoption, 242-243
culture, 7-8
data, 8
privacy, 9
return on investment, 8
security, 9
talent, 7
technology, 9
characteristics of Big Data
value proposition, 238
variability, 237-238
variety, 236
velocity, 236-237
veracity, 237
volume, 234-236
classification, 47-49, 105-114. See also
cluster analysis
Bayesian classifiers, 114
case-based reasoning, 113
decision trees, 49, 113-116
  attributes, 114
  branches, 114
  building, 115-116
  nodes, 114
genetic algorithms, 114
  models, 107
  accuracy, 107
  estimating accuracy of, 107-112
  interpretability, 107
  robustness, 107
  scalability, 107
  speed, 107
nearest-neighbor algorithm, 113
neural networks, 48, 113
N-P polarity classification, 222-223
rough sets, 114
statistical analysis, 113
SVMs, 113
ClearForest, 214
cluster analysis, 50, 117-122
  applications, 117-118
  determining number of clusters, 118
  distance measures, 119
  examples of, 117
  external assessment methods, 121-122
  hierarchical clustering methods, 120
  internal assessment methods, 121
  k-means clustering algorithm, 122-123
  partitive clustering methods, 120
clustering, 46, 50
coefficients (logistic regression), 175
commercial text mining software, 213-214
comparing
  analytics and analysis, 3-4
  ANNs and SVMs, 164-165
  commercial and free data mining tools, 52
  correlation and regression, 166-167
  data mining and statistics, 39
  data mining methodologies, 96-99
  SEMMA and CRISP-DM, 82
concepts, 187
confidence, 126-127
confusion matrices, 108
contingency tables, 108
corpora, 187
correlation versus regression, 166-167
CRISP-DM (Cross-Industry Standard Process for Data Mining), 69-77
  business understanding, 70-71
  comparing with SEMMA, 82
  data preparation, 73-74
  data understanding, 71-73
  deployment, 77
  model building, 74-75
  testing and evaluation, 76
critical success factors for
  Big Data, 240-241
CRM (customer relationship management), 40
cross-validation methodologies, 132
  k-nearest neighbor algorithm, 145-147
cubes, 38
Cutting, Doug, 248

D

data, 32, 35
categorical data, 94-95
discrete data, 95
interval data, 96
nominal data, 95
numeric data, 95
ordinal data, 95
ratio data, 96
structured data, 232
traditional data, 94
unstructured data, 94
Data and Text Analytics Toolkit, 214
data mining, 4, 31-39
  algorithms, 141
  ANNs, 147-158
  nearest-neighbor algorithm, 142-143
  applications
    banking, 40
    brokerages, 41
    CRM, 40
    entertainment industry, 43
    finance, 40
    government, 42
    health care industry, 43
    insurance, 41
    law enforcement, 44
    manufacturing, 41
    marketing, 40
    retailing and logistics, 41
    sports, 44-45
  association rule mining, 49, 123-127
    algorithms, 127
cancer survivability, analytic studies in, 87-90
classification, 47-49, 105-114
  Bayesian classifiers, 114
  case-based reasoning, 113
  decision trees, 49, 113-116
  genetic algorithms, 114
  models, 107
  nearest-neighbor algorithm, 113
  neural networks, 113
  rough sets, 114
  statistical analysis, 113
  SVMs, 113
cluster analysis, 50, 117-122
  applications, 117-118
determining number of clusters, 118
distance measures, 119
topics of, 117
evaluation methods, 121-122
hierarchical clustering methods, 120
internal assessment methods, 121
k-means clustering algorithm, 122-123
partitive clustering methods, 120
data, 35, 93-97
  unstructured data, 94
data preprocessing
data consolidation phase, 99
data reduction phase, 100-102
data scrubbing phase, 99
data transformation phase, 100
data stream mining, 260-261
defining, 36
GIGO rule, 97-98
for Hollywood managers, 60-64
data, 60-61
  methodology, 62
knowledge, 32-33
link analysis, 49
methodologies, comparing, 86-89
misconceptions of, 129-130
neural networks, 48
OLAP, 38
patterns, identifying, 45-51
  associations, 45
  clusters, 46
  predictions, 45, 47
predictive analytics, 105
privacy issues, 57-65
  selling customer data, 58
reasons for popularity, 33-34
rule induction, 49
sequence mining, 49
standardized processes, 67
  CRISP-DM, 69-77
  KDD process, 67-68
  SEMMA, 78-81
  Six Sigma, 83-86
and statistics, 39
structured data, 94
text mining, 185-189
  applications, 186, 195-199
  bag-of-words, 189-190
  initiatives, 199
  NLP, 189-194
tools, 52
  KNIME, 52
  Microsoft SQL Server, 53-54
  RapidMiner, 52
top 10 tools, 55
  vendors, 51
  Weka, 52
visualization, 51
data preprocessing, 73-74
data consolidation phase, 99
data reduction phase, 100-102
data scrubbing phase, 99
data transformation phase, 100
SVM model building, 162-163
in text mining process, 202-204
data scientists, 254-257
  educational requirements, 256
  experimental physicists, 256
  skill requirements, 257
data scrubbing, 99
data stream mining, 260-261
data warehouses, 12-13, 68
databases
  HBase, 254
  in-database analytics, 242
  OLAP, 38
Davenport, Thomas, 32, 255
deception detection, 197
text-based deception detection, 225-227
decision trees, 49, 113-116
  attributes, 114
  branches, 114
  building, 115-116
    algorithms, 116
    splitting indices, 116
nodes, 114
Deep Blue, 20
DeepQA, 22-23
defining data mining, 36
de-identified customer records, 57
descriptive analytics, 17
detecting subjectivity, 222
developing SVM models, 163
dimensional reduction, 101
discrete data, 95
distance measures, 119
DMAIC (Define, Measure, Analyze, Improve, and Control) methodology, 83-86
E
EB (exabyte), 235
ECHELON, 196
educational requirements for data scientists, 256
entertainment industry, data mining applications, 43

ERP (enterprise resource planning), 10, 12
ESs (expert systems), rule-based, 11
estimating accuracy of classification models, 107-112
  k-fold cross-validation, 110-111
  simple split methodology, 109-110
Euclidian distance, 119
evolution of analytics, 10-14
  Big Data, 14
  data warehouses, 12-13
  ERP, 10
  ERP systems, 12
  executive information systems, 12
  rule-based ESs, 11
examples
  application examples
    Big Data for political campaigns, 261-263
data mining for complex medical procedures, 177-180
data mining for Hollywood managers, 60-65
predicting NCAA bowl game results, 130-139
text mining of research literature, 209-213
  of cluster analysis, 117
executive information systems, 12
experimental physicists, 256
explanatory variable, relationship to response variable, 169
explicit sentiment, 217
external assessment methods, 121-122
extracting knowledge, 206-209

F
feed-forward neural networks, 154
finance
  data mining applications, 40
  sentiment analysis applications, 219-220
Ford, Henry, 10
free software tools
  for data mining, 52
  for text mining, 214-215
future of Watson, 23-28
  in education, 27
  in finance, 26
  in government, 27-28
  in health care, 24-25
  in research, 28
  security systems, 25-26

G
GATE, 215
GeB (gegobyte), 235
genetic algorithms, 114
GIGO (garbage in, garbage out) rule, 97-98
Gini index, 116
government
  data mining applications, 42
  sentiment analysis applications, 220
  grid computing, 242

H
Hadoop, 247-253
  benefits of, 250-251
  misconceptions, 251-253
  technical components, 249-250
HBase, 254
HDFS (Hadoop Distributed File System), 248
health care industry, data mining applications, 43
hierarchical clustering methods, 120
hierarchy of analytics, 16
  descriptive analytics, 17
  predictive analytics, 18
  prescriptive analytics, 18
high-performance computing, 242
history of analytics, 10-14
  Big Data, 14
  data warehouses, 12-13
  ERP, 10
  ERP systems, 12
  executive information systems, 12
  rule-based ESs, 11
Hollywood, data mining in motion picture industry, 60-64
homeland security, data mining applications, 44
homonyms, 188
Human Genome Project, 198
human-generated data, 232
hyperplanes, 160

I
IBM Watson, 20-28
  DeepQA, 22-23
  future of, 23-28
    in education, 27
    in finance, 26
    in government, 27-28
    in health care, 24-25
    in research, 28
  security systems, 25-26
Jeopardy! challenge, 21-23
identifying
   patterns in data sets, 45-51
   associations, 45
   clusters, 46
   predictions, 45
   polarity, 224-225
   targets of expressed sentiment, 223-224
in-database analytics, 242
indices, representing, 204-205
information, 32
information gain, 116
INFORMS (Institute for Operations Research and Management Science), 15-16
initiatives in text mining, 199
in-memory analytics, 242
insurance industry, data mining applications, 41
internal assessment methods, 121
interpretability of classification models, 107
interval data, 96
inverse document frequencies, 205

J
jackknifing, 112
Jackman, Simon, 263
Jennings, Ken, 21
Jeopardy! challenge, 21-23
job tracker nodes, 250

K
KDD (knowledge discovery in databases) process, 67-68
KDnuggets.com, 54
k-fold cross-validation, 109-111
k-means clustering algorithm, 122-123
k-nearest neighbor algorithm, 142-147
KNIME, 52
knowledge, 32-33
   extracting, 206-209
   KXEN Text Coder, 214

L
law enforcement, data mining applications, 44
learning algorithms, 156-157
leave-one-out methodology, 111
lift, 126-127
linear regression, 165-173
   assumptions, 171-173
   model building, 167-171
   numeric assessment of model, 169-171
   OLS method, 168
LingPipe, 215
link analysis, 49
location-prediction systems, 198
log frequencies, 205
logistic regression, 173-175
   coefficients, 175
   logistic function, 174
   models, 174
logistics, data mining applications, 41

M
machine-learning techniques, 141
   nearest-neighbor algorithm, 142-143
   SVMs, 159-165
   versus ANNs, 164-165
   hyperplanes, 160
machine-generated data, 232
Manhattan distance, 119
manufacturing, data mining applications, 41
MapReduce, 245-247
market-basket analysis, 123-127
   applications, 124-125
marketing data mining applications, 40
text mining applications, 195
medicine
data mining applications, 43
text mining applications, 197-199
Megaputer Text Analyst, 214
Microsoft Enterprise Consortium, 53-54
Microsoft SQL Server, 53-54
Minkowski distance, finding, 144-145
misconceptions about data mining, 129-130
MLP (multilayered perceptron) architecture, 154-155
models, 45
   classification models, 107
   accuracy, 107
   estimating accuracy of, 107-112
   interpretability, 107
   robustness, 107
   scalability, 107
   speed, 107
   linear regression models, building, 167-171
   logistic regression models, 174
SVM models, building, 161-164
data preprocessing, 162-163
model deployment, 164
model development, 163

morphology, 188
motion picture industry, data mining in, 60-64
data, 60-61
methodology, 62
Movie Forecast Guru, 64
multicollinearity, 172
multiple regression, 167

name nodes, 249
National Centre for Text Mining, 199
nearest-neighbor algorithm, 113, 142-143
cross-validation, 145-147
Minkowski distance, finding, 144-145
parameter selection, 145
similarity measure, 144-147
NER (named entity recognition), 198
network structure of ANNs, 153-154
neural networks, 48, 113, 147-158
ANNs
  back-propagation algorithm, 155-157
  MLP architecture, 154-155
  network structure, 153-154
  processing elements, 153
  sensitivity analysis, 157-158
  versus SVMs, 164-165
  biological neural networks, 148
  feed-forward neural networks, 154
  neurons, 152

neurons, 152
NLP (natural language processing), 189-194. See also text mining
applications, 193-194
challenges associated with, 191-192
WordNet, 192-193

nodes, 114
nominal data, 73, 95
normalization methods, 205
NoSQL, 254
N-P polarity classification, 222-223
numeric data, 95

O
OASIS (Overall Analysis System for
Intelligence Support), 196
objectivity, detecting, 222
OLAP (online analytical processing), 38
OLS (ordinary least squares) method, 168
Open Calais, 215
OR (operations research), 10
ordinal data, 73, 95
OTMI (Open Text Mining Interface), 199

P
partitive clustering methods, 120
part-of-speech tagging, 188
Patil, D. J., 255
patterns, identifying, 45-51
  associations, 45
  clusters, 46
  predictions, 45
phases of data preprocessing, 102-103
data consolidation phase, 99
data reduction phase, 100-102
data scrubbing phase, 99
data transformation phase, 100
polarity, identifying, 224-225
politics
  Big Data, 261-263
  sentiment analysis applications, 220
polysemes, 188
popularity of analytics, reasons for
  affordability, 5-6
  availability, 5-6
  culture change, 6
  need, 5
prediction NCAA bowl game results, 130-139
evaluation, 136-137
methodology, 131-132
results, 137-139
sample data, 132-136
prediction, 45, 47
predictive analytics, 18, 105
  in motion picture industry, 60-64
data, 60-61
  methodology, 62
privacy issues, 58-59
time-series forecasting, 175-180
  accuracy of methods, assessing, 176-177
  averaging methods, 176
prescriptive analytics, 18
privacy, 57-65
  and predictive analytics, 58-59
  as roadblock to analytics adoption, 9
problems addressed by Big Data, 244
processing elements of ANNs, 153
INDEX

Q-R

qualitative data, 73
quantitative data, 73
Quinlan, Ross, 116
RapidMiner, 52, 214
ratio data, 96
regression analysis, 165-167
versus correlation, 166-167
linear regression
assumptions, 171-173
OLS method, 168
logistic regression, 173-175
coefficients, 175
logistic function, 174
models, 174
multiple regression, 167
simple regression, 167
representing indices, 204-205
response variable, relationship to
explanatory variable, 169
retail, data mining applications, 41
RMSE (root mean square error), 169
roadblocks to analytics adoption, 7-9
culture, 7-8
data, 8
privacy, 9
return on investment, 8
security, 9
talent, 7
technology, 9
robustness of classification models, 107
rough sets, 114
rule induction, 49
rule-based ESs, 11
Rutter, Brad, 21

S

SAS Text Miner, 214
scalability of classification models, 107
scatter plots, 167
secondary nodes, 250
securities trading, data mining
applications, 41
security
as roadblock to analytics adoption, 9
text mining applications, 196-197
selling customer data, privacy issues, 58
SEMMA (sample, explore, modify, model, assess), 78-82
sensitivity analysis in ANNs, 157-158
sentiment analysis, 215-227
applications
finance, 219-220
government intelligence, 220
politics, 220

VOC, 218
VOE, 219
VOM, 218-219
explicit sentiment, 217
identifying polarity, 224-225
implicit sentiment, 217
multistep process
collection and aggregation, 224
N-P polarity classification, 222-223
sentiment detection, 222
target identification, 223-224
polarity, 217
sequence mining, 49
Silver, Nate, 263
similarity measure, 144-147
simple regression, 167
simple split methodology, 109-110
singular-value decomposition, 189
Six Sigma, 83-86
skills required for data scientists, 257
slave nodes, 250
sources of Big Data, 232-233
speed of classification models, 107
splitting indices, 116
sports, data mining applications, 44-45
SPSS Modeler, 214
Spy-EM, 215
standardized data mining processes, 67
CRISP-DM, 69-77
business understanding, 70-71
comparing with SEMMA, 82
data preparation, 73-74
data understanding, 71-73
deployment, 77
model building, 74-75
testing and evaluation, 76
KDD process, 67-68
SEMMA, 78-81
Six Sigma, 83-86
Statistica Text Mining engine, 214
statistical analysis, 113
linear regression, 165-173
statistics, 39
stemming, 187
stop words, 187
stratified k-fold cross validation, 137
stream analytics, 257-259
structured data, 94
support, 126-127
survivability of cancer, analytic studies
in, 87-90
SVMs (support vector machines), 113, 159-165
versus ANNs, 164-165
hyperplanes, 160
model building, 161-164
  data preprocessing, 162-163
  model deployment, 164
  model development, 163
synonyms, 188
synthesis, 3

T
Target, use of predictive analytics, 58-59
target of expressed sentiment, identifying, 223-224
taxonomy for analytics, 15-19
  descriptive analytics, 17
  predictive analytics, 18
  prescriptive analytics, 18
Taylor, Frederick Winslow, 10
TDM (term-document matrix), establishing, 202-204
  reducing dimensionality, 206
  representing indices, 204-205
technical components in Hadoop, 249-250
term dictionaries, 188
terms, 187
test sets, 110
text analytics, 183-184
text mining, 185-189
  applications, 186
    marketing, 195
    medicine, 197-199
    security, 196-197
  bag-of-words, 189-190
  initiatives, 199
  NLP, 189-194
    applications, 193-194
    challenges associated with, 191-192
    WordNet, 192-193
  representing indices, 204-205
  three-task process
    data preprocessing, 202-204
    establishing the corpus, 202
    extracting knowledge, 206-209
tools
  commercial software tools, 213-214
  free text mining tools, 214-215
text-based deception detection, 225-227
  three-task text mining process
    data preprocessing, 202-204
    establishing the corpus, 202
    extracting knowledge, 206-209
time-series forecasting, 51, 175-180
  accuracy of methods, assessing, 176-177
  averaging methods, 176
tokenizing, 188
top 10 data mining tools, 55
Torch Concepts, 58
traditional data, 232
training sets, 110
travel industry, data mining
  applications, 42
trend analysis, 208-209
tuples, 259

U-V
unstructured data, 14, 94
text mining, 185-189
VantagePoint, 214
variables, 97
  1-of-N pseudo variables, 97
  relationship between response and explanatory variables, 169
  scatter plots, 167
vendors of data mining tools, 51
visual analytics, 51
visualization, 51
Vivisimo/Clusty, 215
VOC (voice of the customer), 218
VOE (voice of the employee), 219
VOM (voice of the market), 218-219

W
Watson, 20-28
  DeepQA, 22-23
  future of, 23-28
    in education, 27
    in finance, 26
    in government, 27-28
    in health care, 24-25
    in research, 28
    security systems, 25-26
    Jeopardy! challenge, 21-23
Websites, KDnuggets.com, 54
Weka, 52
word counting, 191
word frequency, 188
WordNet, 192-193
WordStat analysis module, 214

X-Y-Z
ZB (zettabyte), 235