
Introduction

Chapter Outline

Introduction, 3	<i>Statistical Tools, 22</i>
Business Situation, 4	<i>Production Systems and Scaling, 22</i>
<i>Finance, Risk, and Costs, 4</i>	Potential Dangers, 24
<i>Marketing, 6</i>	<i>Human Errors , 24</i>
<i>Production and Supply Chain Management, 11</i>	<i>Insufficient Data, 25</i>
<i>Human Resources Management, 12</i>	<i>Bad Data , 25</i>
Perspectives, 14	<i>Over Fitting, 26</i>
<i>Probability and Statistics, 15</i>	<i>Random Chance, 27</i>
<i>Machine Learning, 16</i>	<i>Estimation Instability, 28</i>
<i>Computer Science: Challenges of Large Data Sets, 16</i>	<i>Model Instability, 29</i>
<i>Management Applications, 16</i>	Introduction to Cases, 30
Database, 17	<i>Rolling Thunder Bicycle Company, 30</i>
<i>Traditional Transactions Processing, 18</i>	<i>Diner, 31</i>
<i>Data Warehouse and Analytical Processing, 19</i>	<i>Corner Med, 31</i>
<i>Data Sources, 19</i>	<i>Basketball, 32</i>
<i>Data Extraction, Transformation, and Loading, 19</i>	<i>Bakery, 32</i>
Software Tools, 20	<i>Cars, 32</i>
<i>Data Mining Techniques, 20</i>	Summary, 33
<i>Data Mining Tools, 21</i>	Key Words, 34
	Review Questions, 34
	Exercises, 34
	Additional Reading, 37

What You Will Learn in This Chapter

- What is data mining and what do managers need to know about it?
- How is data mining used in business?
- What do managers need to know about data mining?
- What statistical and data mining tools are used in this book?
- What can go wrong?
- What are the cases used in the book?

Wal-Mart

The amount of digital information increases ten times every five years, and Cisco estimates that in the year 2013, 667 exabytes of data will travel over the Internet. [Economist 2010]

In 2004, Wal-Mart had 460 terabytes of sales and customer data stored at its Bentonville headquarters. Wal-Mart's CIO Linda M. Dillman challenged her staff to help Wal-Mart's Florida stores before the onset of Hurricane Frances. Specifically, what products would residents want to buy before the storm hit? The analysts turned to the historical sales data for stores in the paths of prior hurricanes and applied some early analytical tools. While some might expect that survival items such as flashlights would be popular, it turns out the two biggest-selling items before a hurricane are: strawberry Pop-Tarts and beer. Ms. Dillman noted that "We didn't know in the past that strawberry Pop-Tarts increase in sales, like seven times their normal sales rate, ahead of a hurricane. And the pre-hurricane top-selling item was beer." So, the company loaded the trucks and shipped beer and Pop-Tarts to the stores in the path of the new storm—quickly selling most of the products stocked for the storm. [Hays 2004] By 2011, Wal-Mart's database was estimated to contain 2.5 petabytes of data and the company was expanding its use of business intelligence and data mining tools. [Rogers 2011] Wal-Mart makes some of its sales data directly visible to vendors—leaning towards a pass-through system where vendors monitor sales and build and ship products to match forecasts based on end-point sales.

Even with years of experience, managers and analysts can miss details and changes. Asking questions and analyzing data are critical to making better decisions.

The Economist, "Data, Data Everywhere," February 25, 2010. <http://www.economist.com/node/15557443>

Constance L. Hays, "What Wal-Mart Knows About Customers' Habits," *The New York Times*, November 14, 2004. <http://www.nytimes.com/2004/11/14/business/yourmoney/14wal.html>

Shawn Rogers, "Big Data is Scaling BI and Analytics," *Information Management*, September 1, 2011. http://www.information-management.com/issues/21_5/big-data-is-scaling-bi-and-analytics-10021093-1.html

Introduction

What is data mining and what do managers need to know about it? **Data mining** is the process of using analytical tools to scan large data sets for patterns and provide insight to analysts and managers. Data mining can also be considered as machine or statistical learning, where data is used to train a system to identify categories and patterns so these can be used to make future decisions. Different names are sometimes used, including business intelligence and analytics. Some people categorize data mining and business intelligence as different tools. Data mining tends to focus on the database and algorithm topics while business intelligence emphasizes the business applications and evaluation of the tool results. This book serves as an introduction to all of the related topics, so it covers the data tools, some useful aspects of the algorithms, and the business applications. The goal of the book is to provide enough background to get the reader started in using the common tools with a basic ability to read and apply the results to business problems.

It is critical to recognize that data mining is different from statistical research. Statistical theory is used in many data mining techniques, but the goals of data mining and statistical research are radically different. The goal of statistical research is to prove (or disprove) hypotheses. Specific processes must be followed when evaluating results statistically in order to apply the fundamental theorems. The goal of data mining is much looser—it is to help analysts and managers explore the data and spot potential patterns. These patterns might be small and below the thresholds required for statistical testing. Applying multiple analyses to the same set of data, searching for small patterns, usually violates key assumptions required in statistical research. Data used for data mining can help analysts and managers identify potential patterns and hypotheses to test later, but the same data should not be used for formal research.

Why are data mining tools necessary? Can managers simply examine the data and find patterns on their own? Yes, some patterns are large enough to see simply by examining the data and basic charts. Chapter 3 shows how **hyper cubes** are used to explore data subtotals and enable managers to browse the data and compare various subgroups. But some patterns are complex and difficult or time-consuming to identify. Statistical data mining tools are much more efficient at examining many factors and finding patterns and important relationships. Most statistical tools provide measures of the strength of relationships, making it possible to focus on key properties in the data. Statistical tools are also useful for **forecasting**—extending existing data to predict potential outcomes as the data change.

As the cost of computers and data storage has declined over the years, the amount of data collected by organizations has dramatically increased. Managers can access terabytes and petabytes of **data** on every aspect of business including production, costs, marketing, and human resources management. But do you know which data is important? Max Hopper, CIO of American Airlines, developed a hierarchy that describes how managers can collect data, use it to create information and develop knowledge, which ultimately should lead to wisdom.

Can the data be converted into **information**, where data provides answers to questions and is used to make decisions? A **database management system (DBMS)** is often used to retrieve data to answer questions. Chapter 2 describes how queries are created to answer business questions. But queries require managers to know which questions to ask. The next step is to use information to develop

knowledge, or a deeper understanding of the data including rules and patterns. This level is the main target of data mining—using semi-automated tools to find patterns and rules in the data. How this knowledge is used and applied can lead to **wisdom**—finding new knowledge and making decisions in the presence of a changing environment. Data mining goes beyond basic database concepts and focuses on finding useful patterns in the data, as opposed to simply looking up values. Many of the data mining tools are intended to be interactively used by managers.

Business Situation

How is data mining used in business? This book is largely written for business managers and analysts. The sample datasets, applications, and exercises are all based on common business problems. Yet, the examples and discussion cover only a small sampling of the possible business applications. There is almost no limit to the type of business problem that can be handled by data mining tools. As long as sufficient data points are available, it should be possible to find a data mining tool that can help identify patterns and rules. A key aspect of data mining is to understand each of the main tools so you know which tool to apply to specific types of data.

Many companies use data mining in multiple areas of business—some examples are publicly reported, others are carefully guarded secrets. It is helpful to consider a few examples of how data mining can be used in actual companies, so this section describes a few cases where information is publicly available.

Data analysis is critical to business decisions. Without data and critical analysis, people make mistakes. Relying on intuition in business decisions is usually a recipe for disaster. No matter how much experience a person has, it is critical to “do the math” and evaluate all of the data.

Finance, Risk, and Costs

Finance, particularly investment managers, has made heavy use of analytical tools to identify patterns and rules. Predicting future changes is critical to investments. Some systems use highly complex models to determine how financial and economic variables move together and how patterns might move into the future. Some systems are simpler but are completely automated. They monitor millions of transactions around the world, looking for small differentials in markets and making trades faster than humans. Other issues are more complex—because they involve people. Predicting how people will perform in the future is one of the classic questions in finance. For example, determining how much money to lend to people and what interest rate to charge are key questions for bank loans, mortgages, and credit cards.

Investments are tricky and you can find some of the reasons in the theory of finance and economics. But, to illustrate a specific problem related to data mining, consider a classic scam. Do not attempt the scam; just think about it for a minute. A con-artist chooses a stock. He sends e-mails to 1,000 people stating that he can predict the future and that the stock is going to increase in price. He sends messages to another 1,000 people that the stock is going to drop in price. The next week, if the stock has increased in price, he sends a new message to 500 of the first group with a message that a second stock will increase in price and the reverse message to the second half of the group. If the stock price had fallen, he would switch to the second group and do the same thing. After three weeks, the

swindler has a group of 125 people convinced that he can accurately predict stock prices. Who else can predict three events in a row? At that point, the swindler asks the people for money. Does the story seem absurd? What does it have to do with data mining? On a broader scale, the same problem exists if you examine stock-picking data for brokers or mutual funds. Given a large enough sample, some company or person is likely to have a successful string of selections simply by random chance. Out of the millions of investors and thousands of firms, this one lucky firm stands out and receives all of the attention from the press. Does this firm or person truly have the “formula for success,” or is it merely chance? The answer to the relationship with data mining is that with enough data, a data mining approach will eventually find that one firm or one person who has been successful more times than anyone else. But, did the data mining find a useful set of rules or merely the result of chance applied to the huge amount of data?

Washington Mutual

Washington Mutual, or WaMu as it called itself, was a large savings and loan bank with hundreds of branches on the West coast. The bank largely focused on lower-income customers with an emphasis on free checking, low fees, and liberal lending policies. As the real estate market heated up from 2003-2006, WaMu was one of the leaders in providing mortgages to thousands of borrowers. As Goodman and Morgenson (2008) reported, one interesting loan involved a mariachi singer claiming a six-figure income. John D. Parsons, a WaMu supervisor could not verify the income so he asked for a photograph of the singer in his mariachi costume standing in front of his house—and then approved the loan. With the “Power of Yes” campaign, WaMu’s home-lending unit generated almost \$2 billion in revenue in 2003. In 2005, WaMu purchased Provident, one of the leaders in selling credit cards in the subprime market. Ultimately, the goal of the combined company was to obtain higher interest rates on riskier loans. Like many banks, the company used financial data on customers to identify potentially risky loans; but unlike many banks, Provident and WaMu managers believed they had found the “best of the bad,” (Koudsi 2002). Loan applications were entered into the bank’s computer system for initial approval. Loans not approved automatically were overridden by humans. Many of the loans were repackaged and sold to investors. Yet, midway through 2008, the value of WaMu’s bad loans had reached \$11.5 billion—triple the level from the year before (Goodman 2008). In September 2008, WaMu was sold to JPMorgan Chase for the fire sale amount of \$1.9 billion.

The WaMu situation was not unique. The resulting crash in the housing industry drove the U.S. into the start of the Great Recession as it was called. From a data mining perspective, the key elements are that (1) the financial data mining models used by the company did not adequately estimate the risks involved, and (2) the models were apparently ignored and overruled by human managers in many cases. Other financial institutions also struggled, but many survived by having better models and better understanding of the data and patterns.

High-Frequency Trading

Automated trading systems have existed for several years. Over time, the systems have become more sophisticated and faster. The early systems concentrated on arbitrage—identifying risk-free transactions to take advantage of small differences

in price. For example, if the price of a currency in the Tokyo market differs slightly from its price in the London market, a computer system can quickly identify the difference, and buy a currency on one market, resell it on the other and gain the difference as a profit. As the markets themselves moved to automated processing, several trading firms wrote data analysis algorithms to monitor trades and spot patterns instantly. The tools rely on the ability to monitor data in real time and submit trades before anyone else can react. In most cases, these trading systems are hosted on high-speed networks in the same building as the trade computers. According to (Duhigg 2009), stock exchanges suggest that by 2009, a handful of these trading systems accounted for more than half of all trades. The systems monitor all trades looking for patterns—particularly rising demand for a specific stock. Within milliseconds, the computer systems buy up shares of the desired stock and turn around and resell it to slower-ordering humans, making a profit on the difference as the price rises. Some of the systems issue and cancel small orders just to gauge the extent of the market to determine the maximum price slow traders are willing to pay. Even if the difference is only pennies a share, multiply the difference by thousands of shares across multiple stocks every day, and the profits multiply—perhaps to as much as \$21 billion in 2008 (Duhigg 2009 citing Tabb Group research).

Autonomous systems that can analyze data automatically are relatively rare, but they can be important. As the data mining tools improve, more autonomous systems will be added to financial analyses and other business problems.

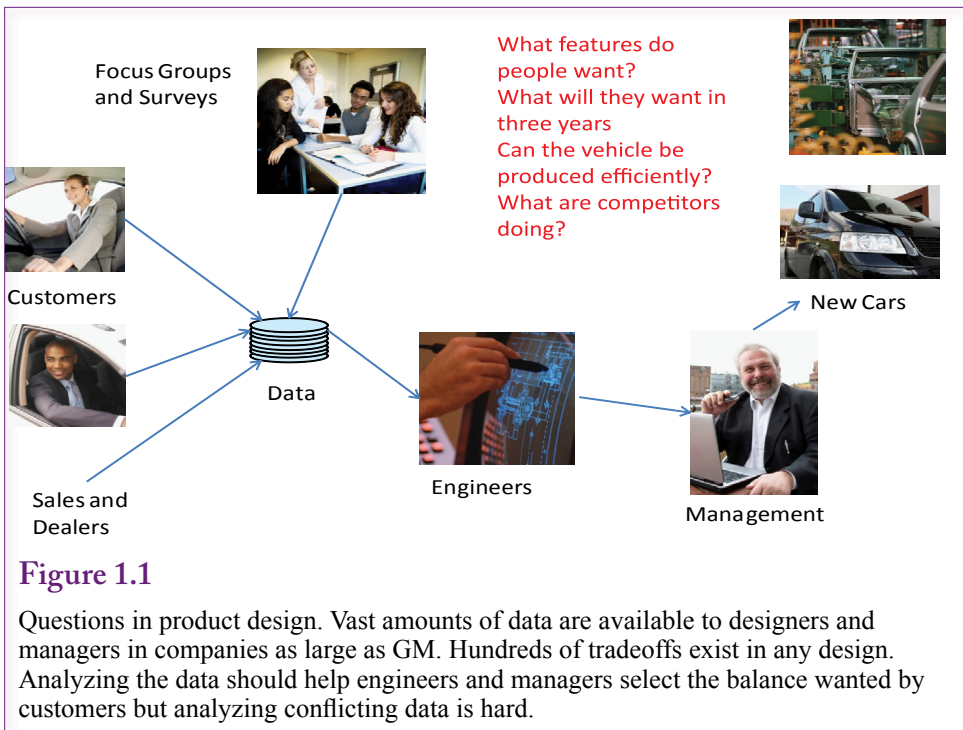
Marketing

Many data mining tools are useful in marketing decisions. Predicting consumer demand is a key element of almost any organization. Designing, building, and shipping products often require time, and companies need some idea of the demand before production begins. But forecasting exactly how customers are going to behave is difficult. Any pattern or forecast can give an edge to the company that understands the data and implications.

Data mining is also used to assist customers and cross sell products. The recommendation tools used by bookstores, music systems, and movie rental companies such as Netflix are useful for encouraging customers to purchase related items. These systems rely on data mining to identify similar products that might appeal to each customer.

General Motors

The issues of determining customer demand are critical in the automobile industry. It takes about three years to design and develop a new automobile. It takes several days and advance planning to physically produce a car. The cost of a car is driven down through mass production—making thousands of almost identical cars at the same time. Minor changes can be customized as a car is built—interior, the choice of radios, tires and so on are relatively easy to make as the car is built. Major changes—particularly colors are difficult to customize and retain low costs. Consequently, cars are designed and built long before they are sold. Many economic changes can occur before the cars make it to dealer showrooms. Selecting design features and forecasting demand for each model is critical in this environment. Figure 1.1 shows some of the types of data available and the basic questions faced by designers and managers. Designing complex products is difficult because of the hundreds of tradeoffs. Yes, customers want powerful engines, but they also



want high gas mileage and low prices. Obviously, one car does not fit everyone. But exactly what features should be included on each type of car and how will the collection of attributes be perceived?

A precursor of GM's design evolved in the early 1980s. The 1970s were hard on GM—a huge increase in the price of gas resulted in a dramatic drop in sales of their over-powered gas-guzzling cars. GM's response was to build an entirely new set of cars for the 1980s. Interestingly, all four major divisions at that time (Chevrolet, Pontiac, Buick, and Oldsmobile) designed almost exactly the same car. Why did each division build the same car when different people want different vehicles? Jump forward to the early 2000s and almost the same problem emerged—each division producing identical vehicles. Throw in the gas price increases of 2007 and GM's gas guzzling trucks and it starts to become obvious why GM was forced into bankruptcy and bailouts by the U.S. and Canadian governments in 2009. GM's designers have tons of data and computing power. How is it that GM was unable to identify customer needs when Toyota and Honda were so much better? See (Taylor 2008) for more details. As a result of the initial questions raised in the 1980s, (Barabba and Zaltman 1991) examined issues in collecting data and evaluating data. The chapter on human biases is particularly useful. Just collecting data is not enough. Even if good data mining tools are used, managers must still understand and believe the results. The GM case clearly indicates the importance of analyzing data, forecasting demand, and overcoming personal biases when designing products.

Pfizer

Another interesting marketing design case involves Pfizer, one of the largest pharmaceutical companies in the world. See (Johnson 2007) for details about how



Pfizer wrote off \$2.8 billion in a project to create an insulin inhaler. Pfizer spent 11 years trying to develop an inhaler that could be used by diabetics to put insulin into the blood stream instead of using injections. After passing all medical tests, the project made it to the market and Pfizer was pushing the product through physicians who specialized in the treatment of diabetes. After lackluster sales of a few million—compared to the projected billions—of dollars, Pfizer pulled the plug on the new product in 2007. In the end, it appears that patients rejected the device because of potential complications with inhaled insulin, because it cost twice as much as injections, and because the device resembled a bong for smoking marijuana. Also, over the course of the 11-year development, insulin pens became available that are substantially easier to use than old-style syringes. As noted in Figure 1.2, pharmaceutical companies often focus on clinical drug trials to generate data and analyses for the Food and Drug Administration. These trials rely on research methods to prove the efficacy and safety of potential new treatments. But, increasingly, pharmaceutical manufacturers such as Pfizer also need to consider the business aspects of treatments, including the ability to manufacture quality products at low prices and to produce items that customers and physicians will prefer over existing treatments. The point is that marketers and managers did not correctly identify the attributes that were important to customers. Identifying attributes and relationships are key aspects in data mining.

Retail Stores

Think about some of the critical decisions for retailers. One of the most challenging problems is identifying which products to stock on shelves and the quantity to carry of each item. As a retailer, it is difficult or impossible to sell products you do not have. Running out of stock or not carrying a popular product is going to result in customers going to other stores and not coming back. Carrying too many

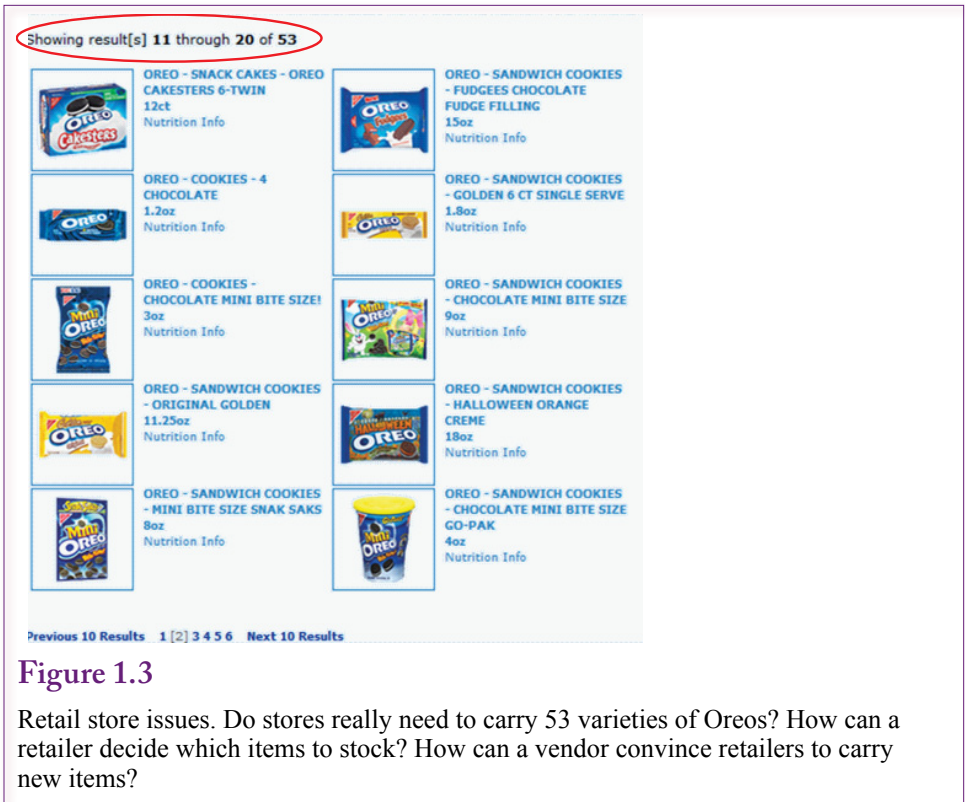


Figure 1.3


Retail store issues. Do stores really need to carry 53 varieties of Oreos? How can a retailer decide which items to stock? How can a vendor convince retailers to carry new items?

products costs shelf space and stocking seldom-sold items takes up space on the shelf and warehouse that could be used for popular products. The challenge lies in identifying exactly which products fall into each category, and predicting the demand for thousands of varieties. For many large retailers, the decision for years was to lean towards multiple products. Figure 1.3 from Nabisco's Web site shows an extreme case: 53 varieties of Oreo cookies. By 2008, a typical food retailer carried 47,000 different products and 47,113 new product variations or sizes were introduced in 2008. By 2009 (Brat et al. 2009), supermarkets, pharmacies, and other retailers began culling their product lists. For example, Walgreen decided that 25 versions of superglue was overkill and reduced the list to 11. Wal-Mart reduced the number of tape measures from 24 down to 20.

Huge amounts of data exist on retail sales. Bar-code scanners and loyalty cards provide data on every transaction by basket, time of day, store, and type of customer. But data mining is needed to find associations, trends, and forecast future demands.

Netflix

The rise of e-commerce has increased the interaction with customers. As customers browse products, recommendation systems can evaluate each selection and compare it to products that other customers have selected. Particularly for subjective items such as books, music, and video, customers often enjoy the assistance provided by seeing what other customers purchase. Perhaps the author or artist has a new release, or a related product might often work well with a specific item. Netflix has a library of tens of thousands of movies. As shown in Figure 1.4,



The image shows a screenshot of the Netflix website interface. At the top, there is a red banner with the Netflix logo and two buttons: "Start Your 1 Month Free Trial" and "Browse Selection". Below the banner is a sidebar menu on the left with the following categories: "Recently Added", "Popular Picks" (highlighted in red), "TV", "Action", "Anime", "Children's Movies", "Classic Movies", "Comedies", "Cult Movies", "Docs", "Dramas", "Faith and Spirituality", "Foreign Movies", "Gay & Lesbian Movies", "Horror", "Indie Movies", "Musicals", and "Music". The main content area displays a grid of movie posters. The top row includes "TARZAN AMERICA", "WARRIOR", and "THE SCORPION". The middle row includes "HUNTER", "NATIONAL SECURITY", and another poster. The bottom row includes "TRANSFORMERS", "THOR", and "ALPHAS".

Data
Customer rentals.
Customer ratings.
Movie ratings/sales.

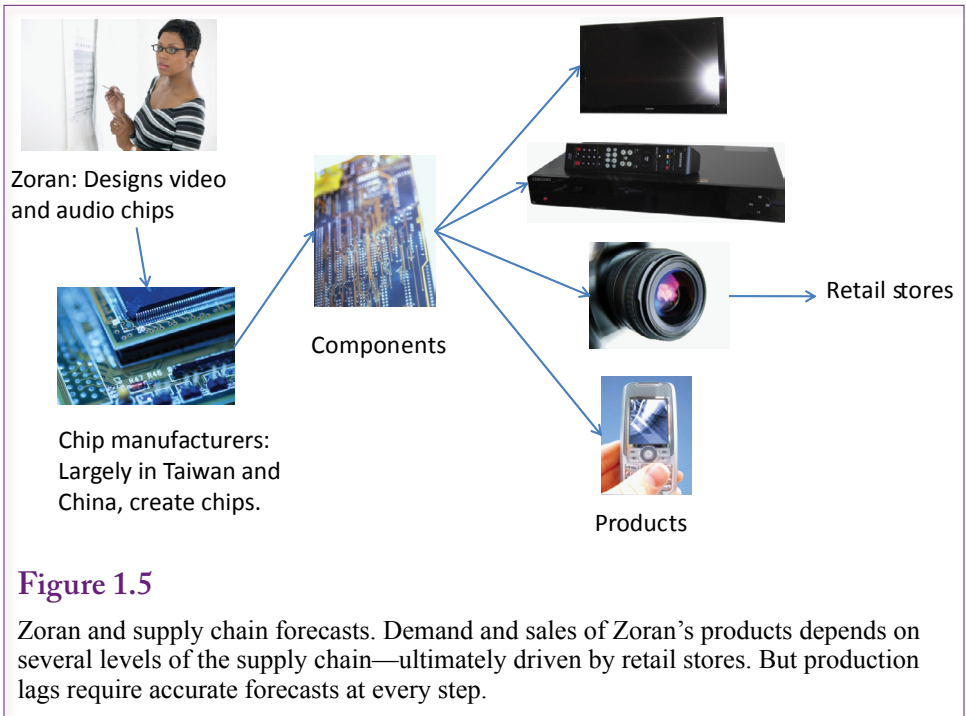
Goal
Recommend movies that customers will enjoy.
Convince customers to rent older movies.

Figure 1.4

Netflix recommendations. Using customer ratings of existing movies, the system must automatically find similar movies that the customer might enjoy. The ultimate goal is to convince customers to rent movies beyond the current releases.

Netflix needs to convince customers to rent movies beyond the handful of latest releases. The Netflix recommendation system is a critical element in the business model of the company. In 2006, Netflix initiated a \$1 million competition—open to anyone who could create a system that could improve recommendation results by at least 10 percent. Any tools could be used but the resulting tools had to be published and licensed to Netflix. A sample data set scrubbed of personal information was provided to contestant teams and results were measured against an internal set of data using a traditional measure (root-mean-square-error). In early results, individual teams made minor progress. In 2009, one team took the step of combining multiple data mining approaches and averaging the results. Based on public results, it accomplished the 10 percent gain, but other teams then combined their approaches in an attempt to duplicate or surpass the results. See (Lohr 2009) for details.

Other than the potential for winning a million dollars, the interesting aspect of the Netflix competition is that a combination of data mining techniques was the winning strategy. Often, analysts and researchers get fixated on a single approach. Observe that it took almost three years for someone to test multiple approaches. As you learn the different techniques, particularly the evaluation methods in Chapter 7, you might wonder why so many different methods exist. Keep in mind that some problems might require multiple approaches.



Production and Supply Chain Management

Adam Smith, one of the early economists, in 1776 wrote about the value of specialization and exchange. Individuals and firms could increase production by specializing. Each person (firm or nation) could produce more of a single item or single step in production by focusing on that one item. One challenge with specialization is the cost of coordinating and paying each specialist. By 2000, information technology reduced transaction costs far enough so that the tasks of building complex products was scattered across many companies and multiple nations. Consider the case of Zoran, a company that designs video and audio processing chips that are used to control digital televisions, cameras, and other gadgets. With the 2009 switch to digital television in the U.S., demand for digital products that use Zoran's chips was booming. As shown in Figure 1.5, demand for Zoran's products depends on several levels of the supply chain for digital products. In a few short weeks in 2008, demand for consumer products evaporated to nothing. As the recession loomed, manufacturers predicted declines in sales, so at every stage of the process, vendors and producers slammed the brakes on production. Rick Tsai, CEO of Taiwan Semiconductor Manufacturing (TSMC) noted that consumer purchases of electronic products in the U.S. fell 8 percent in the last quarter of 2008, but shipments of chips dropped 20 percent (Dvorak 2009). The extent of the drop surprised most producers, but it also meant that as product demand increased, demand for chips also increased early. However, David Pederson, vice president of marketing at Zoran noted that forecasting demand for chips was difficult—sometimes sales are redirected to other uses, such as chips purchased for televisions being installed in digital photo frames. In other cases, growth forecasts are difficult because multiple customers of Zoran's chips might be competing for the same end contract, and only one or two might win. Consider the case of DVD players. Best

Buy, one of the largest U.S. retailers places orders each week along with forecasts for future needs. In general, the company orders DVD players about six weeks in advance. But, the components cannot be produced that quickly, so suppliers need to predict demand and build parts in advance of the orders. Profit margins are extremely thin at each step, and no firm wants to get caught with excess inventory. When the financial crisis hit in the U.S. in 2008 and consumers disappeared, Michael Vitelli, Best Buy's merchandizing chief slashed orders, noting that "you actually had to pick a number with no knowledge whatsoever, because nobody knows anything," (Dvorak 2009). Shipments of audio and visual equipment fell 19 percent in November, 21 percent in December, and 58 percent in January. Producers such as Zoran responded by cutting even deeper. Mr. Pederson noted that "everybody under-cut a certain extent." TSMC, the company that manufactures chips for Zoran and others, slashed production to 35 percent of capacity.

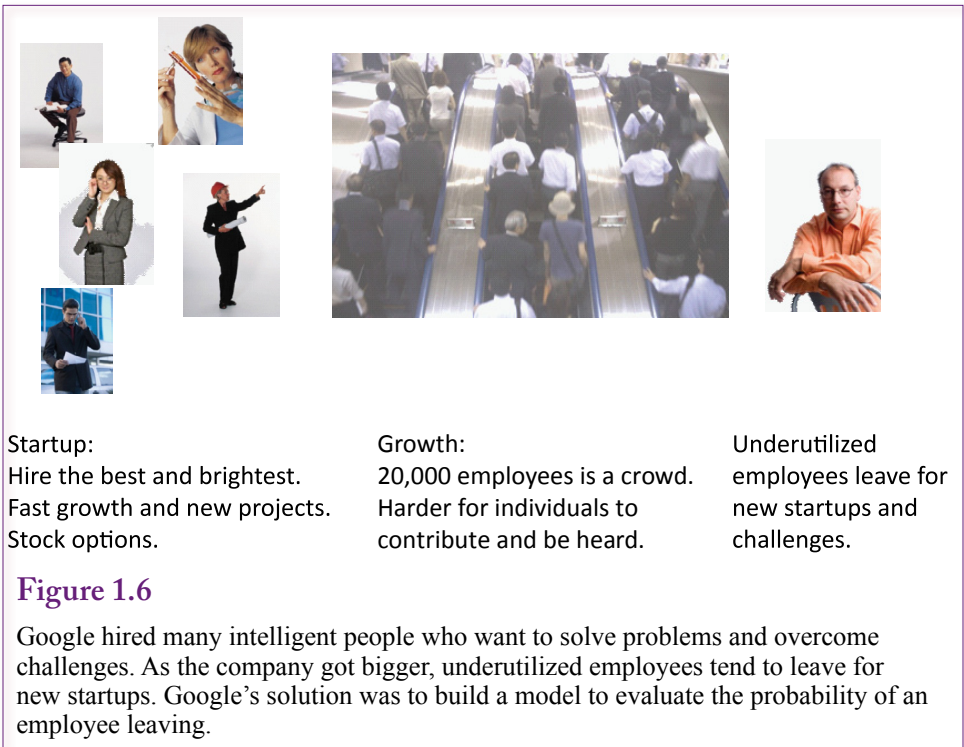
In terms of data mining, the case of Zoran and similar firms presents immense challenges. First, prediction is critical for all levels of production. **Just-in-time** systems require planning and building items in advance of when they will be needed. Second, most models were estimated in fairly stable economic environments. When data shifts far beyond normal ranges, model estimates may no longer work. Third, building models of complex supply chains with multiple interactions is extremely difficult. It might not be possible to obtain reliable data, and constantly changing firms can make it difficult to identify interactions, much less predict them. Yet, extreme changes can often provide useful data for estimating complex reactions. If the only data you ever see comes from the same constant environment, there will never be enough information to estimate radical changes.

Human Resources Management

Employees are the most important part of many businesses. For service companies, employees interact with customers and hold the knowledge and skill to advance the company. Identifying and rewarding the best employees, supporting communication among employees and improving teamwork are often critical components to any business. In some cases, employees are a major component of costs. Determining efficient schedules requires predicting the number of workers needed at each point in time. In more extreme situations, monitoring employees for patterns of fraud or theft is also important. All of these activities use data mining to evaluate patterns.

Google

At heart, Google is a knowledge company that depends heavily on its 20,000 employees. The company hires creative people and relies on employees to develop new products and new ideas. Most employees are given time to work on experimental ideas and new applications. But hiring and managing 20,000 employees is difficult—some employees are bound to slip through the cracks. As indicated in Figure 1.6, as the company gets bigger, the startup atmosphere and stock option growth rates disappear. In 2008 and 2009, several high-level and mid-level managers left to work on smaller companies such as Facebook and Twitter. Those workers who feel underused, or unable to contribute, have a tendency to leave. Finding new employees with the talent and creativity to help grow the company is difficult, time consuming, and expensive. So Google developed a formula based on employee surveys and pay raises among other data that identifies workers most likely to leave. Laszlo Bock, head of HRM for the company notes that the system

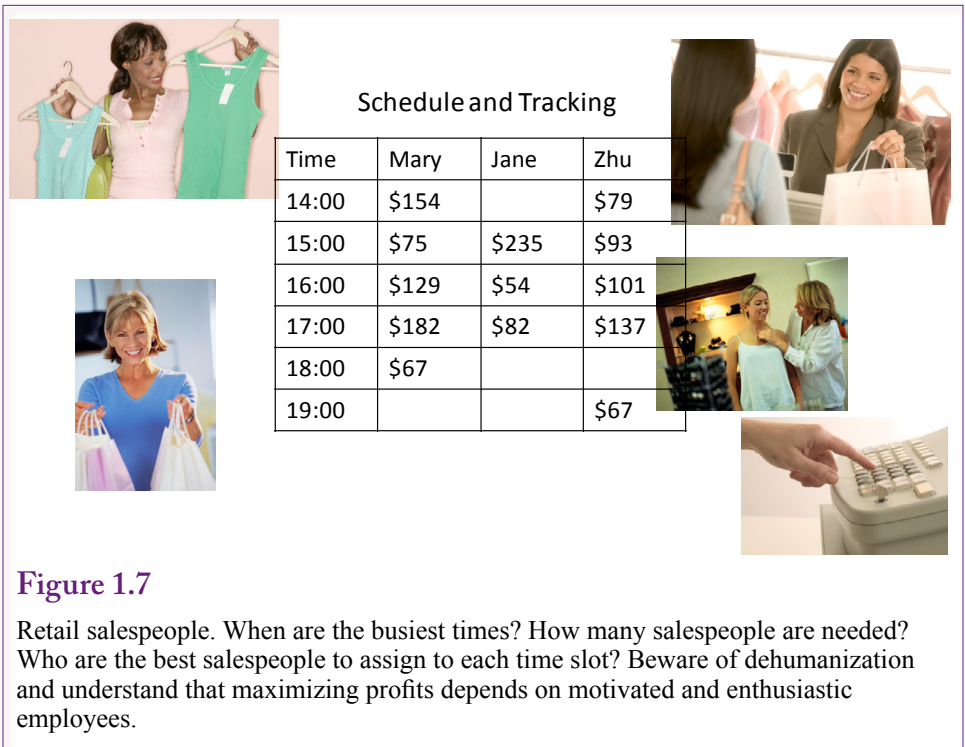


enables the company to “get inside people’s heads even before they know they might leave,” (Morrison 2009).

Applying data mining tools to human capital data is a relatively new experience. Google’s application is ahead of most other companies, and it does face somewhat unique circumstances. Most smaller companies would simply rely on human managers who work with employees on a daily basis to monitor attitudes and find useful tasks that keep workers engaged and creative. There is also the issue of how employees will perceive the use of data mining tools to monitor behavior. If a company does develop a powerful model to predict employee actions, it would probably be wise not to brag to the press.

AnnTaylor Stores Corp.

More traditional businesses have simpler problems—scheduling employees efficiently. Retail stores are classic examples. Having too few salespeople causes customers to walk away. Having too many workers increases costs and reduces profits. Knowing the proper number depends on predicting the number of customers at each point in time. It also depends on the productivity of each worker. As shown in Figure 1.7, AnnTaylor Stores Corp. installed a workforce management system in 2007 that monitors the performance of salespeople in terms of average sales per hour, dollars per transaction, and units sold. The system forecasts the busiest hours and schedules the most productive employees during those times (O’Connell 2008). Before the system was implemented, most store managers scheduled workers according to their preferences, and staffing rarely matched the peak time loads. The new system encourages workers to increase sales. However, this and similar systems often have the effect of shuffling worker schedules each



week—leading to unpredictability of hours for the workers and discontent over the loss of weekly wages and the dehumanization.

Scheduling employees is a question of prediction and of matching employee skills to the needs of the company. Data mining can make the process more efficient. But, any actions involving employees need to be carefully considered because people can be offended by mechanical decisions made by computers. In the end, companies need to evaluate the labor market as well as relationships with workers. Improving efficiency can be important, but in tight labor markets, keeping and improving existing employees is also critical to the success of the company. One of the challenges of data mining and optimization is identifying the true objectives to be maximized. Is it really best to maximize revenue per employee, or is it necessary to include employee hiring and retraining costs to maximize profits? Data mining and optimization can find patterns and improve specified goals, but it is vital to define the correct goals.

Perspectives

What do managers need to know about data mining? The answer to this question is elusive because the technologies are still relatively young. Figure 1.8 outlines the basic perspectives on data mining. The theoretical foundations come from probability and statistics, so understanding the tools and results requires at least a basic knowledge of the mathematics of probability. Chapter 2 summarizes many of these concepts. Machine and statistical learning research led to the development of some powerful tools and concepts of identifying patterns. Computer science research has led to the development of efficient algorithms for searching large data sets. Many data mining tools seem straightforward mathe-

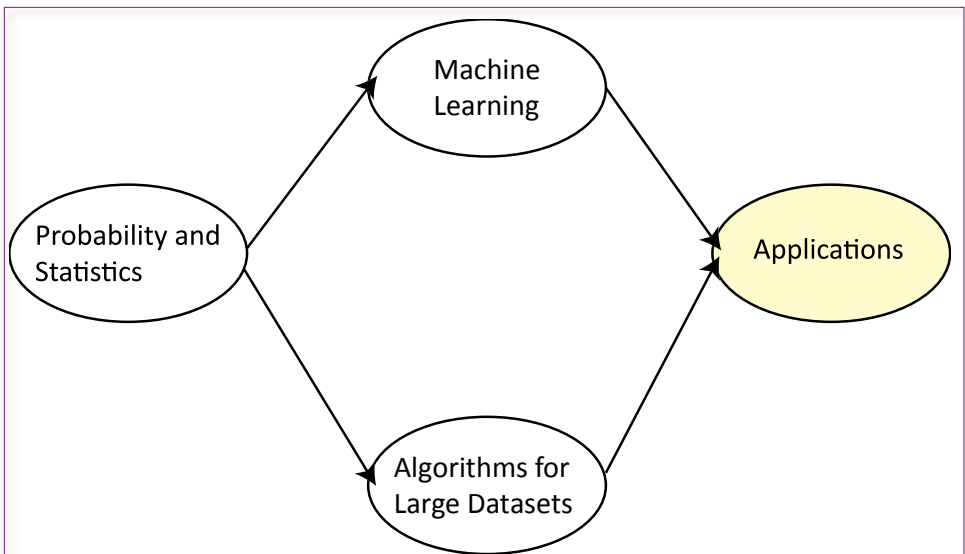


Figure 1.8

Perspectives on data mining. Data mining requires the integration of several complex topics. Probability and statistics form the foundation of the theory. Machine (and statistical) learning led to the development of important tools. Computer science research led to efficient methods to analyze large data sets. But the focus of this book is on understanding the tools to solve business problems.

matically, but with large data sets, they can be impossible to compute without some clever programming.

Several textbooks and courses exist in all three of these areas. In many programs, the emphasis lies in these areas because the concepts and tools are relatively new. Considerable work remains to be done in developing and improving the concepts and tools. Consequently, less work and even fewer books exist in the area of applying the tools to solving business problems. This last area, business applications, is the focus of this book. Many tools now exist and are easy to use without requiring detailed knowledge of statistics and computer programming. It is still important to understand some of the basic concepts of these underlying disciplines but it is not longer necessary to be an expert in those disciplines to participate in data mining projects.

Probability and Statistics

In many ways, probability and statistics are the true fields of data mining. Statistics is concerned with identifying relationships—particularly with proving which relationships are strong enough to exceed random chance. By creating solid mathematical foundations and definitions, probability and statistical theory make it possible to discuss and explore complex relationships in precise terms. The drawback is that it takes several courses and months or years of work to understand the underlying concepts used by the high-level data mining techniques. If you have the mathematical background, check out (Hastie et al. 2009), which explains the statistical foundations of most of the common data mining tools.

Machine Learning

Machine learning is an important area of research that developed computer algorithms for identifying patterns. The research led to the development of **neural networks** which is a way of modeling data that is designed to mimic the way the human brain works. The technique is explained in more detail in Chapter 7. It is a version of machine learning because the network has to be trained based on data with known outcomes. Essentially, a set of nonlinear relationships are estimated using input data with known outputs. The network is then useful at identifying patterns from new data—it has learned the relationships. The method has proven useful at solving complex problems that were difficult or impossible to solve with traditional programming techniques. In particular, it is commonly used for handwriting and speech recognition tasks. It can also be used to analyze business data. Much of the early work is described in the collected work of (Rumelhart and McClelland 1986). These volumes also contain interesting discussions on the nature of machine learning. The basic question is how to design algorithms that can use new data to improve the underlying models to make better decisions. Much of the work is related to statistics, but the machine part of the problem often focuses on the development of new algorithms or new ways to examine the problem.

Computer Science: Challenges of Large Data Sets

Understanding the mathematics and statistics is only one aspect to developing data mining tools. Remember that data mining is often used for huge data sets—containing complex interactions. As fast as computers are, problems still exist that cannot be solved with brute force. It is just not possible to ask a computer to examine trillions of rows of data with thousands of attributes and come up with all possible relationships—unless you are willing to wait a few thousand years for an answer. This problem is the focus of computer science—finding ways to examine huge data sets efficiently. One of the classic data mining technologies (market basket or association rules covered in Chapter 6) gained prominence when Agrawal created the a priori algorithm in 1995 that could efficiently examine millions of combinations of attributes. Solving these types of problems is still an important aspect of computer science research. Even a few years ago, people working in data mining needed to understand some of the algorithms and tricks because they often had to write or customize programs to analyze the data.

Today, most of the existing basic data mining methods are well understood and common programs exist to handle even relatively large problems. A few examples in this book have millions of rows of data and they are solvable in a few seconds. However, be careful. Super large problems still require the assistance of experts trained to optimize the hardware and software. Experts have the hardware and knowledge to reduce computation times from days or weeks to a few minutes or hours. This book contains a few hints about where to expect problems, and the occasional hint of how some tools can be tweaked to improve performance. But, these are merely warnings. This book does not cover the optimization of the tools or applying them to truly gigantic data sets. In general, if you have that much data available, you will need huge servers, and you can afford to hire experts.

Management Applications

The focus of this book is on how to use data mining tools in business applications. It explores the basic tools in terms of business tasks. It contains several examples

<u>SaleID</u>	SaleDate	CustomerID	EmployeeID
1001	9/12/....	15	7
1002	9/12/....	61	3
1003	9/13/....	83	7
1004	9/13/....	15	2

<u>SaleID</u>	<u>ItemID</u>	Quantity	SalePrice
1001	9301	2	5.95
1001	2932	4	12.39
1002	9301	6	5.75
1002	3351	1	12.15
1002	8371	2	16.39

Figure 1.9

Relational tables. Sales data is broken into small pieces stored separately in each table. Updates are fast and can be protected from common problems such as power failures. But retrieving data requires linking the tables.

of realistic data and applies the tools to explain the configuration of the data. It also explains how to understand and analyze the results from various tools.

Using data mining tools requires (1) understanding the purpose of the tool, (2) configuring the data to match the layout required for the tool, and (3) understanding and evaluating the results to make decisions. The main chapters in this book focus on these three steps, and the chapters are organized by the purpose of the tool. In each chapter, study the examples, set up the sample data, and run the same problems. Configuring the data and the tool options is often an important step in the process. You will learn the methods faster by performing the basic steps.

Database

Most companies store data in a DBMS, typically one that relies on the relational model. The DBMS is typically optimized for transactions processing—to store incoming data as quickly and efficiently as possible. These systems are capable of handling large numbers of transactions simultaneously and of protecting the data in case of hardware or even power failures. The details of designing and building database systems are covered in database textbooks (such as Post 2011).

Databases also have query systems to retrieve individual sets of data. **SQL** is the most common query language in use and it has many powerful capabilities. SQL is useful for programmers developing applications, and it can be used to answer ad hoc questions that involve simple subtotals or details. Chapter 3 covers the basics of SQL because it is a useful way to obtain answers to detailed questions. Questions such as: Which customers bought the most items last year, and which salespeople worked with those customers?

ID	LastName	FirstName	Phone	ID	LastName	FirstName	Email
32	Jones	Martha	111-3333	1009	Smith	Jon	J_Smith@gmail.com
35	Brown	Jack	222-3555	1010	Sanchez	Emir	E_San332@gmail.com
36	Smith	Jonathon	777-0222	1011	James	Jeff	Jeff_J009@live.com
39	Masters	Penny	333-4444	1013	Monday	Mary	MandM19@gmail.com
				1014	Stiles	Donna	Dstiles@live.com

NewID	LastName	FirstName	Phone	Email
1	Jones	Martha	111-3333	
2	Brown	Jack	222-3555	
3	Smith	Jonathon	777-0222	J_Smith@gmail.com
4	Masters	Penny	333-4444	
5	Sanchez	Emir		E_San332@gmail.com
6	James	Jeff		Jeff_J009@live.com
7	Monday	Mary		MandM19@gmail.com
8	Stiles	Donna		Dstiles@live.com

Figure 1.10
Extraction, transformation, and loading. Matching data is a common problem. Once it is matched, ID values have to be assigned to enable automated matching for future updates. Numbers might have to be transformed (such as millions to billions). The entire process must be automated.

Traditional Transactions Processing

Online transaction processing (OLTP) represents traditional database applications and is characterized by the need to add small amounts of data from multiple sources. The data has to be consistent and added quickly so actions from one transaction do not interfere with another. Also, the updates need to be protected so that if anything happens during the middle of the update, such as a power failure, the changes can be recovered and the data always remain consistent. One of the most common solutions to these problems is to use a relational database system that breaks data into small pieces stored in separate tables. For example, as shown in Figure 1.9, a typical Sale table might contain columns for an identifier (SaleID), a reference to the customer (CustomerID), the date (SaleDate), and possibly a salesperson (EmployeeID). Creating a new sale simply requires inserting one row of data with values for those attributes. A list of items sold will be stored in a second table (SaleItem) with columns such as SaleID, ItemID, Quantity, and SalePrice. Updating the tables with new data is fast and straightforward to protect.

The challenge lies in retrieving the data. Answering questions involving sales involves both of the main tables, plus probably a Products table and a Customer table. Each row from the tables has to be joined to data in another table. Query languages such as SQL make it easy to define the connections. However, the query engine has to retrieve data by matching values in one table with data in other tables. This process requires many lookups. Most systems seek to improve performance by creating **indexes** on the columns—almost always on the primary keys. But indexes need to be updated when rows are added to a table, so adding several indexes to the database means each update now requires many changes to indexes as well. What began as a single row insert suddenly becomes complicated and can potentially slow down all updates.

Data Warehouse and Analytical Processing

In essence, storing data and performing intensive searches are conflicting uses of a database. One solution is to create a second database or **data warehouse** which is a copy of the transaction data. This copy is bulk-updated on a schedule—not for every new transaction. Consequently, the data warehouse can be stored in new formats that include multiple search indexes and even duplicate data to increase performance for data retrieval. New methods of retrieving data that do not require SQL queries are used to make it easier to retrieve data for exploration and analysis. These **online analytical processing (OLAP)** systems are explained in Chapter 4.

Data Sources

One function of a data warehouse is to combine data from multiple sources. Many companies today use enterprise resource planning (ERP) systems to handle accounting and other data-intensive tasks. This data is usually stored in a relational DBMS and it is internally consistent. Many of these systems have the ability to transfer specific data to data warehouses. These systems are relatively easy to configure and use with OLAP and data mining systems.

On the other hand, companies also tend to have other sources of data that might be stored in separate databases or even individual files. Engineering, research, marketing, finance, HRM, and other areas of the company might have created unique systems to hold data. If this data is needed for analysis, it has to be located and identified. One challenge with finding data is that it is often unclear exactly how it is defined or what it means. It is even more confusing when departments use different terms—such as Client instead of Customer. A key aspect of data warehouses is the ability to retrieve data from diverse sources, including PC databases and spreadsheets. As you locate each of these sources, it is critical to maintain detailed records of the data definitions, formats, locations, ownership, and security constraints. Eventually, a process has to be created to extract the data, transform it, and load it into the warehouse—and all steps have to be handled automatically with minimal errors or human involvement.

Data Extraction, Transformation, and Loading

With luck, most important data will come from an ERP system. A primary benefit of ERP systems is that the major work of integrating the data and ensuring consistency has already been performed. Data taken from other sources typically needs considerable work to ensure all of the pieces will match. Seemingly simple things such as geographic locations (City, State, and Country) can cause problems. If people are allowed to enter data into simple text boxes—the data is guaranteed to be inconsistent. People will always abbreviate and misspell names and never do it consistently. Similarly, if multiple sources exist for customer or employee data, matching the data is going to be painful. Consider two data files with contact information created by two executives that contain basic employee information (Last Name, First Name, and Address). If each file has an entry for John Smith, do they both refer to the same person? Perhaps you can compare the address or even a phone number or e-mail address; but what if one of the lists is older than the other and the person has moved? Figure 1.10 shows an example of a simple merge of contact data. What if one of the lists has “John Smith,” and the other “Jonathon Smith,” are they the same people? Similar problems exist with almost any type of list, but the contact problem is common today when managers create their own

personal lists. If the lists contain only a few dozen entries, the problem is reasonable. When the lists contain hundreds of entries, they become difficult to integrate. Carefully crafted database queries can help, but some names will still need to be checked by humans. Moving forward, if the separate lists will be maintained in the future, the matching process has to be automated. Typically, a master list is created with unique ID values, and the individual data sets are modified to include the ID number from the master list. Then the data can be merged in the future by matching the ID values.

The process of retrieving data, cleaning it to verify consistency, and loading it into the data warehouse is often referred to as **extraction, transformation, and loading (ETL)**. The examples given here provide only a glimpse into the challenges faced at this stage of designing a data warehouse. In most OLAP projects, developing the automated ETL processes takes 60 to 80 percent of the time and effort of the entire project. This process often requires programmers with advanced knowledge of databases and queries. But it also requires business analysts who talk to the managers and workers to find the data, determine its precise definition, and write the transformation steps.

Software Tools

What statistical and data mining tools are used in this book?

The tools available for data mining and statistical analysis have expanded considerably in the past few years. The big database vendors now provide tools—typically integrated with their database systems. Oracle has a data mining add-on for the enterprise suite. Microsoft has SQL Server Analysis Services (SSAS), and IBM purchased the statistical tool vendor SPSS in 2009. Standalone tools also exist, and a couple of them are used in this book. In general, the tools integrated with a DBMS are easier to use—particularly in a production environment. Although they are often more expensive than standalone tools, the integrated environment can save time and money during development. Still, the tradeoffs are difficult decisions that need to be made when designing a data mining strategy.

Data Mining Techniques

Many techniques can be classified as data mining tools. This book concentrates on the core group of methods readily available in common tools. Rather than focus a chapter on each specific technique, this book is organized by the application goals. It begins with techniques that can run largely unsupervised—where the tools learn from the data and make decisions without requiring the analyst to build models and carefully guide the tool. Chapter 5 examines the issues of clusters. How close are data points to each other? What groups can be formed from the data where items within the group have similar attributes compared to items in other groups. Chapter 6 examines the classical data mining technique of association or market basket analysis. It addresses questions such as which items are commonly purchased together? It also applies to any events that might happen at the same time. Chapter 7 focuses on the evaluation of dimensions or attributes. It examines several techniques that require interaction or supervision with the analyst, including regression analysis, Naïve Bayes, Decision trees, and neural networks. All of these tools are defined and described in that chapter. Chapter 8 examines common techniques for evaluating and predicting time series data. Time series data consists of observations over time. Patterns often exist in terms of seasons, months, or daily changes. Chapter 9 introduces geographical analysis—which focuses on the visual display of data related to location.

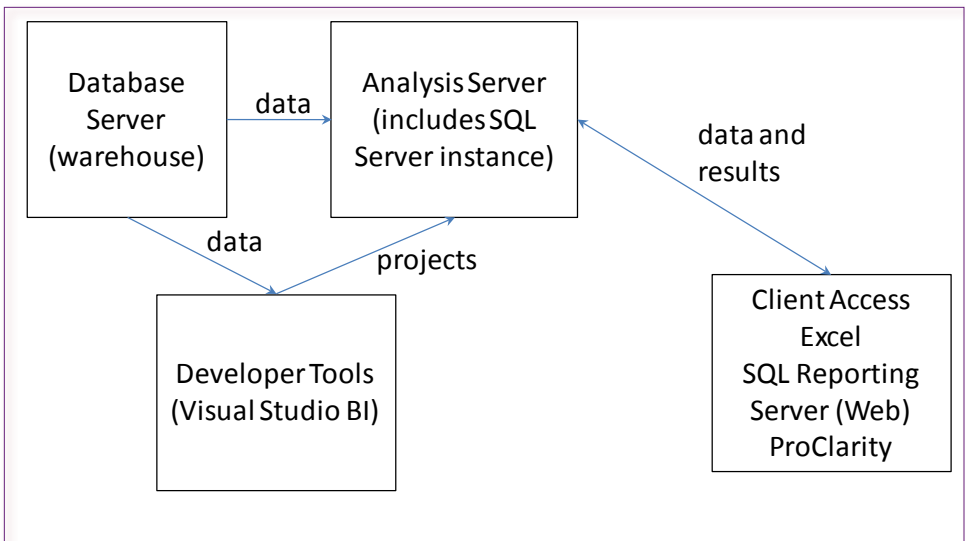


Figure 1.11

SQL Server Analysis Services components. In a production environment, these pieces are usually installed on separate machines. For development, it is easiest to install everything on a single computer.

Data Mining Tools

The primary data mining tool used in this book is Microsoft's **SQL Server Analysis Services (SSAS)** data warehouse and data mining tool. Through Microsoft's MSDN Academic Alliance program, the tools are relatively inexpensive for educational purposes. The database and other tools are easy to install, can run on simple hardware including laptops, and are comparatively easy to use. (Although the existing documentation is weak and a few tricks are needed, but they are explained in this book.) The tools and projects can be scaled up to production environments, and Microsoft's developers have reported the ability to handle huge data sets—with the appropriate hardware (server farms), and expert tuning. Other tools exist and can also be obtained on educational licenses (e.g., Oracle and IBM), but their tools tend to be more difficult to install and administer. It is also straightforward to get demonstration copies of the software for free use for several months, so students can load copies onto their own computers..so straightforward to get demonstration copies of the software for free use for several months.

Figure 1.11 shows the basic layout of SSAS. Each of the components is necessary to develop any projects—except the separate client tools are not needed for initial development and testing. In a development system, it is easiest to install the database server, analysis server, and the development tools (Visual Studio) all on the same computer. Basically, use the SQL Server installation process and choose the option to install everything. Putting everything on one machine simplifies the security permission issues. Be sure to add the appropriate users as administrators for the servers. If the pieces need to be split, moving the database server to a separate machine is the easiest—particularly if mixed authentication is used so that users can log in with a SQL Server username and password. Creating a separate

Analysis Services Server is more complex because developers need administrator access to the service, which means establishing user logins across the network. The process is straightforward if all of the machines and users are logged into Active Directory, but it still requires time to set up the permissions. Also, every project created requires modification of the project's deployment property to specify the new server—the default server is always “localhost.”

Several standalone tools exist for data mining. Some are expensive, others are free. The two leading free tools are the R System (<http://www.r-project.org/>) and Weka (<http://www.cs.waikato.ac.nz/ml/weka/>). The R System is actually more of a programming language and requires some time to learn—although a few books and sample code exist on the Internet. The Weka system is relatively easy to use and is demonstrated in a few cases in this book. If you want the Windows version, check carefully to find the current version.

Statistical Tools

Several standalone statistical packages can be used for traditional statistical analyses as well as data mining. The specific features vary by tool and the prices vary enormously. The high-end packages SPSS, SAS, and Stata are commonly used in research disciplines and are often available at universities. Their usability, configuration, and results tend to be specific to the individual packages. Their main drawback is the price and the fact that data has to be specially configured to work with the tools. These packages are not covered in this book, although SAS does sell one of the leading data mining tools.

A few open-source statistical tools are available now. They also require data to be stored in specific formats, but they are low cost and relatively easy to install and use. One of the interesting tools is gretl (<http://gretl.sourceforge.net/>) which is useful for econometrics (logistic) and time series analysis. It is available free from source forge and is easy to install and run.

Most standalone tools can read data that is stored in comma-separated-values (CSV) formats. These files are text based and are relatively easy to create for small datasets. They are more difficult to deal with when the number of data rows gets into the millions. Most of the tools have some ability to use Microsoft's ODBC database connection methods, so it is possible to retrieve data directly from the database. However, for the relatively small examples in this book, the CSV format is simpler because it requires less configuration effort and fewer problems with security issues.

The standalone tools are useful in this book for a few reasons. First, to show that data mining can actually be performed at relatively low cost. Many people use MySQL or PostgreSQL as the database systems to reduce costs even further. Second, Microsoft's Analysis Services has some slightly unusual approaches to many tools. The methods work, and they are designed to run relatively unsupervised. However, the quirks lead to output that is somewhat different from traditional data mining tools. Hence, it is useful to compare the Microsoft results to those from the traditional or theoretical models. To follow along, you should obtain and install both the Weka and gretl packages.

Production Systems and Scaling

Data mining and OLAP tools are designed to handle huge data sets—into the terabytes of storage. The tools used here (particularly Microsoft SQL Server Analysis Services and Weka) have all been used on giant problems. However, these huge

Acquisition/Input		
Bias	Description	Example
Data availability	Ease with which specific instances can be recalled affects judgments of frequency.	People overestimate the risk of dying due to homicides compared to heart disease.
Illusory correlation	Belief that two variables are related when they are not.	Ask any conspiracy buff about the death of JFK.
Data presentation	Order effects.	First (or last) items in a list are given more importance.
Processing		
Inconsistency	Difficulty in being consistent for similar decisions.	Judgments involving selection, such as personnel.
Conservatism	Failure to completely use new information.	Resistance to change.
Stress	Stress causes people to make hasty decisions.	Panic judgments and quick fixes.
Social pressure	Social pressures cause people to alter their decisions and decision-making processes.	Majority opinion can unduly influence everyone else: mob rule.
Output		
Scale effects	The scale on which responses are recorded can affect responses.	Ask a group of people to rate how they feel on a scale from 1 to 10. Ask a similar group to use a scale from 1 to 1,000.
Wishful thinking	Preference for an outcome affects the assessment.	People sometimes place a higher probability on events that they want to happen.
Feedback		
Learning from irrelevant outcomes	People gain unrealistic expectations when they see incomplete or inaccurate data.	In personnel selection you see how good your selection is for candidates you accepted. You do not receive data on candidates you rejected.
Success/failure attributions	Tendency to attribute success to one's skill and failure to chance.	Only taking credit for the successes in your job.

Figure 1.12

Biases in decision making. Without models, people tend to rely on simplistic “rules of thumb” and fall prey to a variety of common mistakes. These errors can be minimized with training and experience in a discipline. They can also be minimized by having computer systems perform much of the initial analysis.

problems require specialized hardware and careful configuration to function in realistic times. Although, some bloggers on the Web have reported running Weka for several months on a single problem, so it depends on how long you are willing to wait. The point is that the high-end tools such as SSAS can scale to handle huge problems. This scaling is typically accomplished by running the SQL Server database on a specialized server farm with multiple computers and high-speed

networked drives. The Analysis Services are run on a separate server farm with multiple high-speed processors. The enterprise versions of Windows Server and SQL Server are required to distribute the processing load across multiple computers in the server farm.

The high-end tools often utilize parallel processing to split computations into pieces or threads that can be run on different computers. Open source tools sometimes lag behind in support for parallel processing. However, with the source code available, you could always find a programmer to improve the performance.

In most cases, it is easiest to develop and test data mining models for SSAS on a single developer computer. If the datasets are huge, the initial versions can be built and tested on reasonably-sized samples pulled from the main data. Once the model is developed, the project can be deployed to the main analysis server by changing the project's deployment property. The database connections can also be changed simply by editing a connection string to point to a different server.

Potential Dangers

What can go wrong? With any project that is used to guide decisions, it is useful to be cautious and recognize the potential dangers and errors that can arise. Obviously, many things can go wrong, but this section focuses on a few of the problems that are specific to data mining. Recognizing problems means they can be monitored. Even if some cannot be completely avoided, at least the level can be assessed and the results can be interpreted in terms of their usefulness and reliability. For example, decision makers need to know if results are highly variable and sensitive to small changes in assumptions or data.

Human Errors

Humans can make many errors at every stage in the analysis process. Basic mechanical errors such as recording data incorrectly, converting numbers with the wrong values, recording results with errors, and even typographical errors in the analysis can have serious consequences on the analysis and interpretation of the data. These errors are best avoided through careful procedures. In many cases, it is helpful to use pairs or teams of people to cross-check the work. Also, experience is useful because analysts learn to evaluate data and results in terms of "reasonableness." If extreme variations or outliers exist in the data, they should be verified. When results are highly unusual or surprising, they should be validated. Experience with the tools and with the specific data helps analysts recognize what is "unusual" and overly "surprising."

More subtle human errors arise because the results are interpreted and applied through the human lens of perception. Particularly with complex problems, people are often selective and see the results that are familiar or that they want to see. Remaining neutral, evaluating all possibilities, and considering new conclusions is difficult. Barraba and Zaltman (1991) contains an excellent chapter on the perils of human decision making. Even if data is readily available, and even if analysts use comprehensive data mining tools, ultimately the results are evaluated by humans. And humans make many mistakes in evaluating data and results. Figure 1.12 shows a partial list of the more than 100 biases explored by Barabba and Zaltman. Data and careful analysis can help reduce some of the biases. Using multiple, independent people to evaluate data can help; but it is difficult to find independent people, and beware of "group think." Other studies have shown that

groups tend to make suboptimal decisions—with members often agreeing to go along with weak choices simply to avoid causing conflicts.

A big problem in data mining is applying the appropriate tool for the data and the type of decision. Some errors are blatant; others are subtle and might require a statistician or econometrician to explain the problem (for instance, using simple regression with limited dependent variables). It is also easy to misinterpret results or to extend results to decisions that no longer meet the assumptions. Classic examples include trying to forecast data too far into the future or to time intervals that do not match the underlying data. No simple rules exist to avoid these problems. Only that if the problem is critical, it would be wise to consult a statistician to verify the approach and overall tools.

Insufficient Data

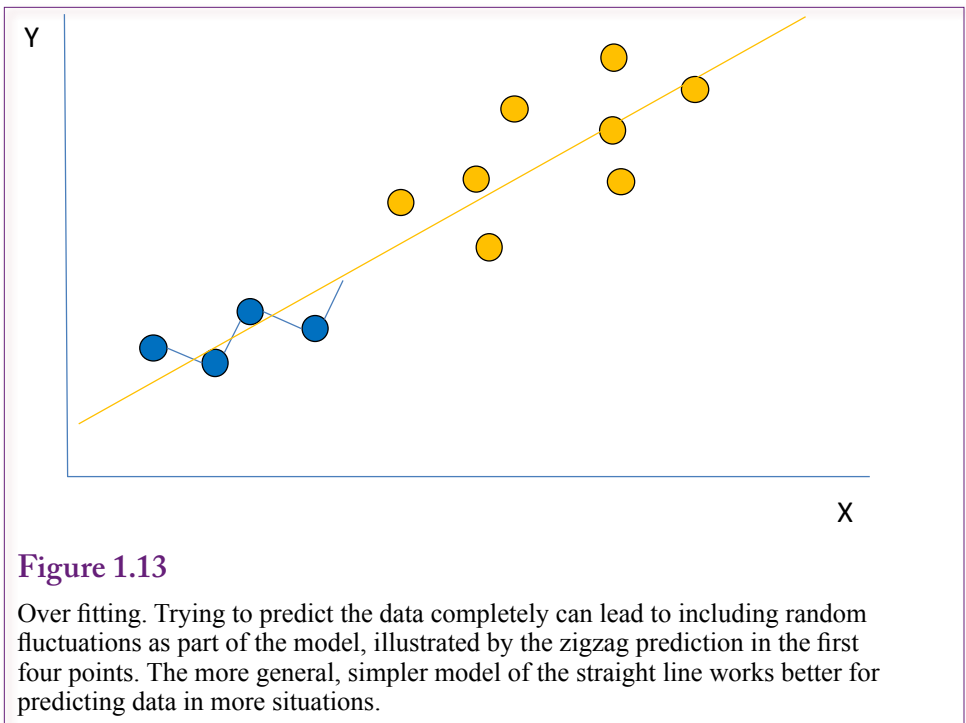
A problem that can arise in research is the issue of insufficient data. In business data mining situations, the opposite problem is more common—too much data. If a company has been operating for a while, usually a large amount of data exists. However, in some cases, it is possible to summarize the data down so far that it reduces the value. For example, it would be best to avoid evaluating data at the annual level. Even if a company has been operating for dozens of years, you would be throwing away detailed quarterly or monthly details that would be useful. Operating on annual data might not provide enough observations to generate robust results—particularly because the underlying economic and structural foundations can change considerably over a couple of decades.

The amount of data needed depends heavily on the type of problem being analyzed. In some cases, it is possible to estimate the minimum number of observations needed, but this approach is rarely needed for data mining. Typically, regression and time series problems are relatively robust with a hundred observations or more, and these levels are easy to achieve in most business problems. But, research involving customers or employees might have more limited data. With smaller data sets the choice of methodologies becomes more critical. Avoid making conclusions based on small sets of data. When in doubt, consult a statistician with knowledge and experience in research techniques. Remember that no matter what techniques might be used to compensate for small data sets, the results ultimately are based on a limited amount of information and can easily be affected by random peculiarities within that set.

A common technique used for 25 years in research is bootstrapping, which draws thousands of random samples from the existing data and generates new data sets. In theory, these new datasets should have the same distribution as the original data, and the expanded amount of data makes it easier to use some data mining techniques. However, a study in 2008 indicated that bootstrapping (and cross validation explained in another section) are ineffective and misleading in small samples (Isaksson et al. 2008).

Bad Data

Most corporate data is probably sound. Sales and financial data tend to be critical to the company and investors, with accountants and auditors examining and verifying the accuracy of much of the data. However, bear in mind that some publicly-released accounting data is not audited—such as quarterly or monthly accounting statements. Even if the base data is accurate, some unaudited data might contain errors.



Still, financial data is going to be better than many other kinds. Asking customers, and even employees, to respond to surveys can lead to questionable data. Although few studies exist, it is likely that as much as 60 or 70 percent of the personal data collected via Web sites is inaccurate. Asking customers for age, income, and other demographic data might irritate customers. In attempts to protect their privacy, people will randomly click entries. When creating surveys, try to avoid asking questions that people are unwilling to answer. Try to provide incentives for people to be accurate.

Integrating data is also a difficult problem in some companies. When data is stored in diverse systems across the company, it is challenging to clean the data and ensure that everything matches. Mismatched data can lead to serious errors in the analysis and results. It is one of the reasons the ETL stage takes so much time—it is crucial that only correct data be loaded into the data warehouse. In almost every case, it is better to leave out data than to include data that can be wrong.

Over Fitting

Over fitting is one of the classic issues with data mining. In fact, it is the reason that the term data mining was originally pejorative. Research statisticians are warned against data mining because it can violate the assumptions required for statistical testing. Think about a small problem with sales data from a couple hundred customers. With data mining, it is tempting to throw every possible tool on the data and see what shows up. With some tools, it is even possible to tweak the parameters to more closely fit the specific data. In the end, this process can generate a model that almost perfectly describes the specific set of data. But, what if this specific set of data is not completely representative of all of the customers or every situation? The model is over fitted to the specific data.

Another way to describe the problem is to note that most situations include three major elements: (1) A dependent variable to be forecast or modeled, (2) A set of independent variables that represent changes in underlying factors, and (3) Random error. The goal of statistics is to find the relationship between the dependent variables and independent variables while measuring (and ignoring) the random error. Over fitting a problem results in a model that incorporates the random error into the factor relationships. These random effects are likely to be different in other times and other cases, but the over-fit model treats them as if they will always be the same. Figure 1.13 illustrates the problem. Over fitting the first four points leads to a zigzag model that is unlikely to fit other points and can lead to strange predictions. The simpler, more general model of a straight line does not perfectly fit the four points, but it does a much better job of applying to additional data.

In any case, a model that is over-fit will work well on the specific training data, but perform poorly on any other version of the data. In other words, it is dangerous, because the results are not applicable to any other situation. And what is the point in describing and predicting something that already happened? The goal is to create a model that describes the overall situation and can be applied to other cases (subject to random error).

One method commonly used to test for over fitting is to withhold a random sample of data from the training set. By default, SSAS reserves 30 percent of the observations for this testing sample. The model is estimated on the main 70 percent of the data. It can then be tested on the withheld 30 percent to see if the results are close to the same. If they are radically different, the model is probably over fit and needs to be discarded.

A related approach is to split the data into multiple sets (typically 10 sets) and estimate models on combinations that leave out one of the sets at a time—resulting in ten different model estimates, each based on a different 90 percent of the data. This **cross validation** approach is a useful way to test for over fitting. However, it still applies to only the data in the entire sample. If the data is not representative of the entire set of possible data, the model can still be over fit to that specific time period or set of data.

Random Chance

All events are affected by random chance, so real-world data contains random errors. As shown in Chapter 2, most random errors fall within a fairly tight range, but extraordinary events can always arise—particularly with thousands or millions of observations. Consider simple events such as lotteries. The vast majority of people lose every week, but eventually someone wins. This winning selection gets all of the press, which distorts perceptions, but the point is that even rare events happen. Do not mistake random events for causation. Consider the slightly more complex case of financial investments. Check any financial newspaper or Web site. You will see rankings and stories on investment firms that have successful track records. But, do these firms, or their data mining models, truly have a method that predicts in all cases, or is it the same as the lottery winners? A random winner gets all of the press attention. At any point in time, some person or model can have amazing results—simply by random chance.

Zweig (2009) makes the point by being critical of data mining tools in the investment community. He mentions examples from David Leinweber's book *Nerds on Wall Street*. To prove his point, Mr. Leinweber built a model that

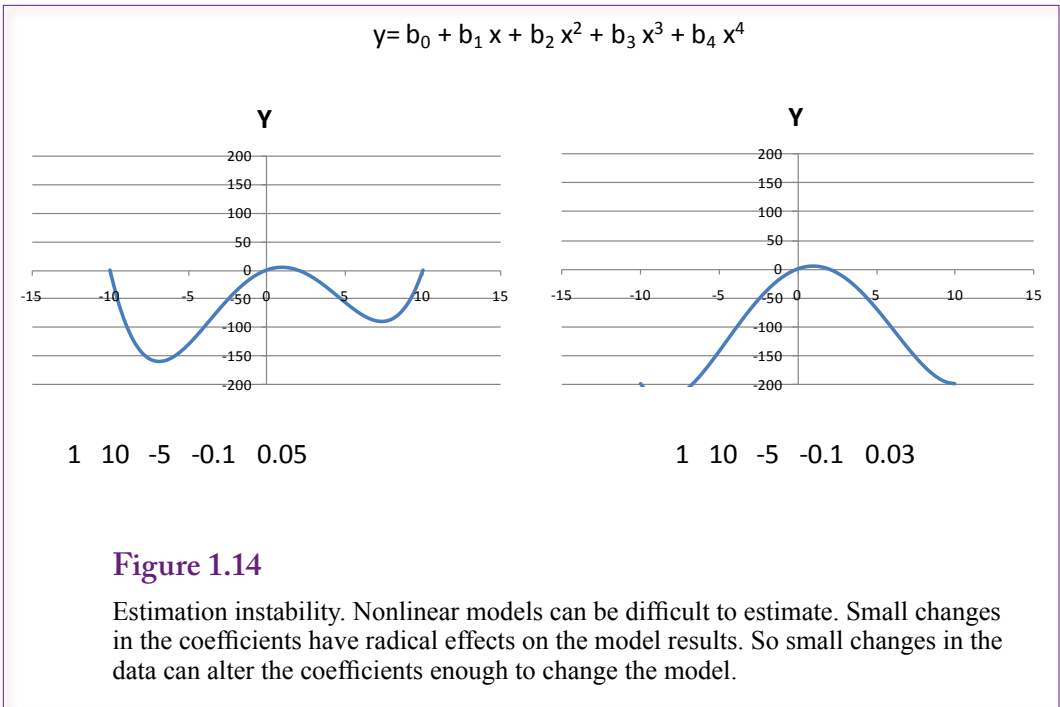


Figure 1.14

Estimation instability. Nonlinear models can be difficult to estimate. Small changes in the coefficients have radical effects on the model results. So small changes in the data can alter the coefficients enough to change the model.

showed over a 13-year period 75 percent of the variation in U.S. stock prices was explained by the annual production of butter in Bangladesh. He increased the percentage to 99 percent by including U.S. cheese production and the total population of sheep in the U.S. and Bangladesh. The model is clearly nonsensical. Yet, Mr. Leinweber notes that he still gets calls from would-be investment managers who ask for details on the model to create a new investment fund. The point is that random events do arise in practice, and random correlations are going to show up if enough data is examined in a large number of combinations. All results need to be examined for reasonableness and tested against other sets of data.

Estimation Instability

A difficult problem that arises in complex data is that the results can be unstable. Small changes in the data result in radical shifts in the model coefficients and results. If the estimation is run only one time, this instability can be hard to spot. Changing the data set by removing observations or estimating the model on new datasets will reveal these instabilities—simply compare the coefficients and results from different data sets.

Instability occurs frequently with non-linear models. Figure 1.14 shows what happens with a relatively simple quartic function (fourth-degree polynomial). Notice the difference in the single coefficient from 0.05 to 0.03. The resulting model is quite different. Given random changes in data, it is possible that either value could be estimated with slightly different versions of the data. Some tools are even more nonlinear and are sensitive to small changes in the data. Even if it is possible to estimate the model reliably, it would be risky to apply the model to new situations where it might predict radically different results.







	Company	ZIP size	DB size
	Rolling Thunder Bicycle Company	11 MB	78.8 MB
	Diner	0.6	11.2
	Corner Med	1.5	55.3
	Basketball	1.3	29.7
	Bakery	47	349
	Cars	0	3

Figure 1.15

Cases and database sizes. ZIP files are the size of the file to download. DB Size is the typical size of the MDF database file. It is best to configure the database with the option for Recovery Mode set to Simple to eliminate the transaction log.

A second common situation that results in unstable coefficients is created when multicollinearity is high among the attributes. **Multicollinearity** means that the data from multiple variables can be expressed as a linear combination. For example, $X_1 = 5X_2 + 4X_3$. Using all three X variables in a model will fail in most cases because there are only two pieces of information instead of three. If the data are not quite perfectly linear, the values can be estimated, but they tend to be unstable. Small changes in the data can result in highly different values for the coefficients.

Some techniques exist for finding stable estimates (such as ridge regression), but they are not covered in this book. The most important point is to test for instability. The simple solutions are to reduce the number of attributes in the model and move closer to linear models.

Model Instability

Model instability is another potential problem, and it is different from estimation instability. It is possible to estimate coefficients that are stable and consistent, yet lead to a model that is unstable. For example, models that include observations over time often use difference equations, such as $Y_t = aY_{t-1} + bX$. Some values of the coefficients will cause the model to be unstable—for example, increasing values of Y will lead to ever larger values and the forecasts will quickly lead to huge numbers that are out of control.

These models can work for limited ranges of data but extrapolating them into the future results in unrealistic projections. It is important to examine every model over a wide range of data. Projections that are unrealistic will help determine the

valid range of data for the model. Some models will work over a wide range of data; others will work for only a small set—such as one or two years. The key is that you need to know the limits before trying to use the model in real predictions.

Introduction to Cases

What are the cases used in the book? As with many other mathematical and business concepts, data mining is easiest to understand by working with examples. Six different cases are used through this book to illustrate various techniques in data mining. Some of the cases are relatively general; others were chosen to illustrate specific concepts. Four of the cases use data that was generated based on realistic data, but represent fictional companies. It is difficult to obtain actual detailed data from operating companies. Two of the cases do use actual data, but the amount of data is limited as a result.

The databases can be downloaded from the book's Web site. The files are stored in CSV text files that have scripts to build the database in SQL Server and load the table data. Some security permissions are needed to create the tables and bulk load the data. In general, it is easiest to create the databases with DBA permissions. Databases are built and loaded one at a time. Load the main script file into the SQL editor. As indicated in the file, change the database name and the path that points to the data files. Then run the script to create the tables and load the data.

Figure 1.15 summarizes the cases in the book. Note the database sizes. The ZIP size is the size of the downloadable file—it will affect the time it takes to transfer the data. The DB size is the base size of the SQL Server MDF file. Note that the Bakery case contains a table with over seven million rows. When you create these databases, you should be careful to set the Recovery Mode to Simple—it is an optional parameter when creating the database and it can be changed after the database is created. If the mode is left on the default Full method, every data change will be written to a transaction log file—which quickly becomes substantially bigger than the original data file. This file is as much as seven gigabytes for the Bakery database. With most data mining projects, either the Simple or Bulk Load option is preferred because it holds down the size of the log file and improves performance. It is also unnecessary because the data can always be reloaded from the original transaction sources.

Rolling Thunder Bicycle Company

The Rolling Thunder Bicycle Company is a fictional company that builds and sells custom bicycles. Most of the data is realistic, such as prices and weights for components. Sales data is configured to match some economic data. Bicycle prices might seem high to people unfamiliar with the industry, but they are relatively accurate for custom bicycles—possibly even too low. The focus of the data is on customization, sales, and production. Sales start in 1994. Some additional description of the company is available on the base Web site: www.JerryPost.com, under the Rolling Thunder links. For people unfamiliar with the bicycle industry, it might be helpful to download the Microsoft Access copy of the database on the Web site. It contains the transaction forms for the company, so it is possible to work through sample cases of configuring and building bicycles, as well as placing and receiving orders from suppliers. The SQL Server files contain only the data—not the forms. The slides for this chapter contain the relationship diagram of the tables used for the Rolling Thunder Bicycle database. The Bicycle and Customer tables are the most important, but several tables are used as lookups. Note

Code	Name
FG	Field goals made
FGA	Field goals attempted
TP	Three pointers made
TPA	Three pointers attempted
FT	Free throws made
FTA	Free throws attempted
ORB	Offensive rebounds
DRB	Defensive rebounds
TRB	Total rebounds
AST	Assists
STL	Steals
BLK	Blocked shots
TOV	Turnovers
PF	Personal fouls
PTS	Points (total)

Figure 1.16

Player statistic abbreviations. Three point shots TP and TPA are sometimes abbreviated 3P and 3PA, but SQL server requires brackets if columns begin with numbers.

that the Access database also contains forms to generate sales of new bicycles, so it is possible to create new patterns. However, the process requires some thought to get reasonable patterns because of the complexity of the options.

Diner

The Diner database is fictional data that contains a single table. This table (Diners) lists the date, day of week, meal time, gender of the group, the number of people at the table, the total bill, and whether or not dessert was ordered. Treat the data as if it came from a relatively high-end restaurant. The specific meals ordered are not critical because the chef changes the menu depending on the availability of some items, and on his preferences. Dessert is important because desserts have higher profit margins, and because people who stay for dessert often spend more money on drinks, with even higher profit margins.

Corner Med

Corner Med is a small healthcare database for a fictional company that has a store-front treatment office. Several physicians, nurses, and assistants staff the site. Patients arrive with common conditions and are treated. The data is fictional but it includes the ICD10 codes for diseases and treatment. The standard drug table is also included. More importantly, the incidence of diseases and treatments are drawn from the U.S. physician's survey so the combinations of items are accurate representations of real cases. As with all of the databases, the names, addresses, and phone numbers are false. The slides for this chapter contain the relationship diagram for the Corner Med database. The primary tables are the Patient, Visit, Employee, VisitDiagnoses, VisitProcedures, and VisitMedications. The goal of the case is to examine data as a business manager—not as a medical expert. The diagnoses codes are fun to read, but not particularly relevant to the business issues

in the case. Medical researchers would use similar data to examine public health trends and explore the effects of treatment options. Business managers are more interested in costs, revenue, and employee management.

Basketball

The basketball database contains actual data for three seasons of the NBA including playoffs: 2008 through 2011. It lists game statistics for all of the players for every game. The tables are: Teams, Players, Games, GameResults, and PlayerGameStats. From a database perspective, the Games table has redundancies because it lists every game twice—once for the home team and once for the visiting team. Technically, the GameResults table contains the same data, but to track player statistics, it was easier to use the Games table to create the shared GameID. Be careful to avoid double counting when examining team statistics. The included team View is designed to minimize this problem, so use it instead of the base tables. The player statistics are straightforward, but the abbreviations might be unfamiliar to people with limited basketball experience. Figure 1.16 shows the list of abbreviations and their names.

If it seems strange to use sports data in a business case, consider that professional sports is a huge business. No, the case does not focus on ticket sales—although those are likely related to the won/loss record. Instead, think of the management questions involving worker performance and salaries. Sports agents make money by analyzing performance data for their clients and presenting the players (workers) in the best possible light to justify raises. Given the large amount of sports statistics available, data mining is a useful tool for identifying patterns and understanding relationships.

Bakery

The Bakery case is a straightforward example of business sales. The data could almost come from any type of business. The tables consist of: Sale, SaleItem, Product, and ProductCategory. The company does not track employees or customers. Instead, the focus is on the items and categories sold. The SalesDate in this case is also tracked down to the specific time that a sale was made. Sales in the case run from January 1, 1995 through December 31, 2012. The data is fictional, but a relatively sophisticated mechanism was used in the data generator to ensure that interesting results and patterns arise.

The key feature of this database is its size. It contains almost 2 million sales with almost 8 million rows in the SaleItem table. Although the size is still relatively small in terms of data mining problems; it would be difficult to distribute anything larger. If you are interested in larger and more realistic datasets, search for the Wal-Mart data. Wal-Mart is trying to encourage the teaching of data mining by providing a cleaned set of sales data through a university program. There are restrictions, and some complications in its use, but it is a large set of data. Also, Netflix provided scrubbed rental data for its recommendation contest. Again, the data has restrictions on its use and distribution, but it is a large set of realistic data.

Cars

The Cars database contains real data. It is a smaller database than the others, but most people are familiar with attributes for cars and the real data make it interesting and easier to comprehend for certain types of problems. The primary Cars database consists of features for over 300 automobiles in the 2012 model year. Basic

data includes weight, number of cylinders, miles per gallon, number of seats, and list price. In most cases, the data represents values for the base model of the car, so it is not meant as a tool to compare all possible versions. Many vehicles have dozens of variations, and students are encouraged to add more data if they want to compare specific models. If you want to use more data, an older copy of the file contains data for 2009 models.

The data file includes a secondary file (*CarSalesMonthly.csv*) that does not load automatically into the SQL Server database, although it is straightforward to import it. It contains data from U.S. government Web sites that lists total vehicle sales by month from 1967-01 through 2012-01. Total sales are broken down into domestic and foreign vehicles for each month. This set of data is quite different from the detailed attributes for 2012 models, but it provides a useful set of data for time series analysis. The sales are a count of the number of vehicles sold in that month (in thousands).

Summary

Data mining is different from statistical analysis of research data. DM is designed for the exploration of data—providing insights, visualization, and early statistical comparisons of data. It has been successfully used in all areas of business—to analyze trends, identify patterns, understand relationships among attributes, and predict results..

Data mining requires a basic understanding of probability and statistics as well as some knowledge of database concepts and tools—particularly queries. Data warehouses are often created to hold data for analysis and exploration. The designs for retrieving and analyzing large amounts of data are different from the designs for storing large amounts of transaction data. Consequently, most organizations build processes that extract data from the transaction processing systems, clean and format the data, and store it in a specialized data warehouse.

Many different software tools exist to analyze data, from traditional statistical packages to dedicated data mining and business intelligence tools. This book focuses on the readily available data analysis tools in SQL Server 2008, but a few more traditional tools are also used to compare the results.

Many things can go wrong in data mining—from human errors, to bad or insufficient data, failing to account for random chance, estimation and model instability, and over fitting a model to a small set of data. It is important to understand the common types of errors and recognize when one of them appears. Knowing what to watch for helps ensure that the problems are caught early and the risks minimized.

This book contains sample cases for six different types of organizations. The size of the databases varies in terms of the number of rows and the number of attributes. The cases are useful for illustrating the application of the tools. It is also important to work on the chapter exercises to gain experience with configuring and running the tools and to understand the results. Cases are important for practice. One of the complications of learning data mining is trying to understand and interpret results. Many situations require some understanding of the underlying organization. For the most part, the cases used in this book focus on business problems, so a basic background in business is a good starting point. For more detailed interpretations and for making decisions, you should do some background research on the specific industries to understand the terms and relationships.

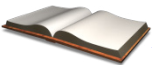
Key Words

cross validation	just-in-time
data	knowledge
data mining	multicollinearity
data warehouse	neural networks
database management system (DBMS)	online analytical processing (OLAP)
extraction, transformation, and loading (ETL)	online transaction processing (OLTP)
forecasting	SQL
hyper cubes	SQL Server Analysis Services (SSAS)
indexes	wisdom
information	

Review Questions

1. What is the purpose of data mining?
2. Why are data, statistics, and data mining important in the decision making process?
3. What are the background disciplines used to develop data mining technologies?
4. What is ETL, why is it so important, and why does it take so much time to configure it?
5. What strengths does SQL Server have as a data mining tool and platform?
6. What are the main dangers in any data mining project?

Exercises



Book

1. If necessary, install the SQL Server database, SQL Server Analysis Studio, and the Business Intelligence Visual Studio components.
2. Identify a decision you have made in the past year or two and describe which human errors and biases you faced while making the decision.
3. Watch a TV game show where contestants must make decisions (not just answer questions), and identify human biases that exist and that are pressed by the host.
4. Choose an industry or specific firm and identify a common decision that must be made. Specify the type of data that should be collected and what types of patterns might be helpful in analyzing the data?

5. Identify the current version of a commercial data mining/business intelligence tool. Summarize its features and estimate the price of the software.
6. Find a business example of a problem that could benefit from the use of data mining. Identify the data available and the specific decisions that need to be made. If possible, interview a manager in that area or find a business case.
7. Use a Netflix account or Amazon search to evaluate the accuracy of the recommendation engine. Do you agree with the proposed matches?



Rolling Thunder Database

8. Create the database and install the data. If the database and tables were already created for you on a central server, test your log in and display the data in the Employee table.
9. Identify at least one major decision that must be made by the managers of this company. Looking at the table definitions, what data could be used to help make this decision?
10. List each table in the database and briefly describe the purpose or data in the table. Keep the description general; it is not necessary to explain every column.
11. If the company emphasized online ordering, what additional data would be available to use for data mining?



Diner

12. Create the database and install the data. If the database and tables were already created for you on a central server, test your log in and display the data in the first 10 rows of the Diners table.
13. Identify at least one major decision that must be made by the managers of this company. Looking at the table definitions, what data could be used to help make this decision?
14. List each table in the database and briefly describe the purpose or data in the table. Keep the description general; it is not necessary to explain every column.



Corner Med

15. Create the database and install the data. If the database and tables were already created for you on a central server, test your log in and display the data in the Employee table.
16. Identify at least one major decision that must be made by the managers of this company. Looking at the table definitions, what data could be used to help make this decision?
17. List each table in the database and briefly describe the purpose or data in the table. Keep the description general; it is not necessary to explain every column.



Basketball

18. Create the database and install the data. If the database and tables were already created for you on a central server, test your log in and display the data in the Teams table.
19. Identify at least one major decision that must be made by the managers of a team. Looking at the table definitions, what data could be used to help make this decision?
20. List each table in the database and briefly describe the purpose or data in the table. Keep the description general; it is not necessary to explain every column.



Bakery

21. Create the database and install the data. If the database and tables were already created for you on a central server, test your log in and display the data in the Cars table.
22. Identify at least one major decision that must be made by the managers of this company. Looking at the table definitions, what data could be used to help make this decision?
23. List each table in the database and briefly describe the purpose or data in the table. Keep the description general; it is not necessary to explain every column.



Cars

24. Create the database and install the data. If the database and tables were already created for you on a central server, test your log in and display the data in the Employee table.
25. Identify at least one major decision that must be made by the managers of an automobile manufacturer. Looking at the table definitions, what data could be used to help make this decision?
26. List each table in the database and briefly describe the purpose or data in the table. Keep the description general; it is not necessary to explain every column.



Teamwork

27. Form teams for future assignments. Obtain contact information and determine a method of communication. If available, find a way to work on files together, such as via SharePoint, Groove, or a Web site such as Google Docs.
28. Each person should select a Web business and identify a major decision that needs to be made by that company and the data available. Share the information and then vote on which decision would be the most difficult to answer.

Additional Reading

Barabba, Vincent P. and Gerald Zaltman, *Hearing the Voice of the Market*, 1991, Harvard Business School Press: Boston. [Marketing perspective on collecting and understanding data—triggered from GM’s mistakes.]

Brat, Ilan, Ellen Bryon, and Ann Zimmerman, “Retailers Cut Back on Variety, Once the Spice of Marketing,” *The Wall Street Journal*, June 26, 2009. [Identifying products to carry.]

Duhigg, Charles “Stock Traders Find Speed Pays, in Milliseconds,” *The New York Times*, July 24, 2009. [High frequency trading data.]

Dvorak, Phred, “Clarity is Missing Link in Supply Chain,” *The Wall Street Journal*, May 18, 2009. [Supply chain challenges with Zoran.]

Goodman, Peter S. and Gretchen Morgenson, “By Saying Yes, WaMu Built Empire on Shaky Loans,” *The New York Times*, December 28, 2008. [WaMu mortgage lending collapse.]

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman, 2001, *The Elements of Statistical Learning*, Springer: New York. [An outstanding book on data mining, with an emphasis on statistical theory. A graduate-level book that requires a strong mathematics background.]

Isaksson, A., M. Wallman, H. Göransson, M.G. Gustafsson, “Cross-Validation and Bootstrapping are Unreliable in Small Sample Classification,” *Pattern Recognition Letters* Volume 29, Issue 14, 1960-1965, October 15, 2008. [Analysis of the performance of bootstrapping and cross-validation in small samples, showed that the results are weak and the recommend a Bayesian approach instead.]

Johnson, Avery, “Insulin Flop Costs Pfizer \$2.8 Billion,” *The Wall Street Journal*, October 19, 2007. [A good summary of the issues and costs faced by Pfizer and the inhaled insulin project.]

Koudsi, Suzanne, “Sleazy Credit,” *Fortune*, February 19, 2002. [WaMu early warnings.]

Lohr, Steve, “Netflix Challenge Ends, but Winner Is in Doubt,” *The New York Times*, July 27, 2009. [Initial results of the Netflix competition.]

Lohr, Steve, “Netflix Competitors Learn the Power of Teamwork,” *The New York Times*, July 28, 2009. [A more detailed examination of the NetFlix competition and combining algorithms.]

Morrison, Scott, “Google Searches for Staffing Answers,” *The Wall Street Journal*, May 19, 2009. [Summary of Google’s use of data to evaluate employee turnover.]

O’Connell, Vanessa, “Retailers Reprogram Workers In Efficiency Push,” *The Wall Street Journal*, September 10, 2008.

Rumelhart, David E. and James L. McClelland, 1986, *Parallel Distributed Processing*, Vol. 1 and 2, MIT Press: Cambridge, Massachusetts. [A classic collection of work on the foundations and start of machine learning (neural networks). Includes descriptive and technical content.]

Taylor III, Alex, “GM: Death of an American Dream,” *Fortune* November 25, 2008. [A good but optimistic business history of GM’s mistakes.]

Zweig, Jason, “Data Mining Isn’t a Good Bet for Stock-Market Predictions,” *The Wall Street Journal*, August 8, 2009. [Comments on problems with data mining in investments.]