CHAPTER 1

Wholeness of Business Intelligence and Data Mining

Business is the act of doing something productive to serve someone's needs, and thus earn a living and make the world a better place. Business activities are recorded on paper or using electronic media, and then these records become data. There is more data from customers' responses and on the industry as a whole. All this data can be analyzed and mined using special tools and techniques to generate patterns and intelligence, which reflect how the business is functioning. These ideas can then be fed back into the business so that it can evolve to become more effective and efficient in serving customer needs. And the cycle continues on (Figure 1.1).

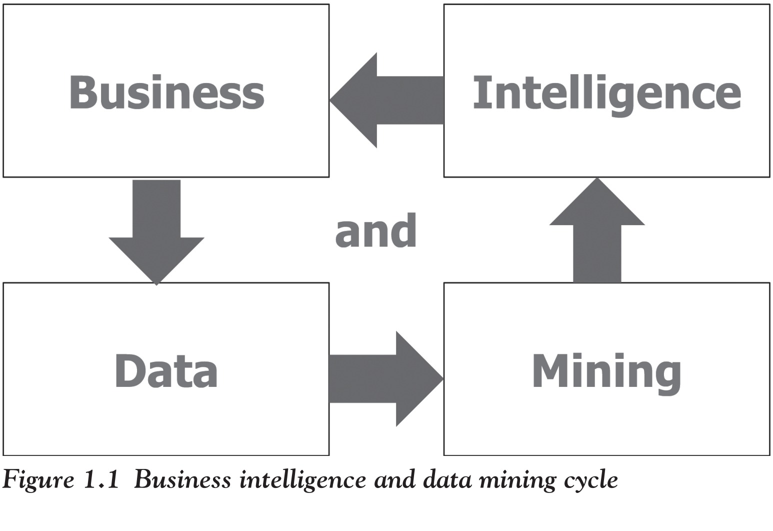


Figure 1.1 Business intelligence and data mining cycle

Business Intelligence

Any business organization needs to continually monitor its business environment and its own performance, and then rapidly adjust its future plans. This includes monitoring the industry, the competitors, the suppliers, and the customers. The organization needs to also develop a balanced scorecard to track its own health and vitality. Executives typically determine what they want to track based on their key performance Indexes (KPIs) or key result areas (KRAs). Customized reports need to be designed to deliver the required information to every executive. These reports can be converted into customized dashboards that deliver the information rapidly and in easy-to-grasp formats.

Caselet: MoneyBall—Data Mining in Sports

Analytics in sports was made popular by the book and movie, Moneyball. Statistician Bill James and Oakland A's General Manager Billy Bean placed emphasis on crunching numbers and data instead of watching an athlete's style and looks. Their goal was to make a team better while using fewer resources. The key action plan was to pick important role players at a lower cost while avoiding the famous players who demand higher salaries but may provide a low return on a team's investment. Rather than relying on the scouts' experience and intuition Bean selected players based almost exclusively on their on-base percentage (OBP). By finding players with a high OBP but, with characteristics that lead scouts to dismiss them, Bean assembled a team of undervalued players with far more potential than the A's hamstrung finances would otherwise allow.

Using this strategy, they proved that even small market teams can be competitive—a case in point, the Oakland A's. In 2004, two years after adopting the same sabermetric model, the Boston Red Sox won their first World Series since 1918. (Source: Moneyball 2004)

Q1. *Could similar techniques apply to the games of soccer, or cricket? If so, how?*

Q2. *What are the general lessons from this story?*

Business intelligence is a broad set of information technology (IT) solutions that includes tools for gathering, analyzing, and reporting information to the users about performance of the organization and its environment. These IT solutions are among the most highly prioritized solutions for investment.

Consider a retail business chain that sells many kinds of goods and services around the world, online and in physical stores. It generates data about sales, purchases, and expenses from multiple locations and time frames. Analyzing this data could help identify fast-selling items, regional-selling items, seasonal items, fast-growing customer segments, and so on. It might also help generate ideas about what products sell together, which people tend to buy which products, and so on. These insights and intelligence can help design better promotion plans, product bundles, and store layouts, which in turn lead to a better-performing business.

The vice president of sales of a retail company would want to track the sales to date against monthly targets, the performance of each store and product category, and the top store managers that month. The vice president of finance would be interested in tracking daily revenue, expense, and cash flows by store; comparing them against plans; measuring cost of capital; and so on.

Pattern Recognition

A pattern is a design or model that helps grasp something. Patterns help connect things that may not appear to be connected. Patterns help cut through complexity and reveal simpler understandable trends. Patterns can be as definitive as hard scientific rules, like the rule that the sun always rises in the east. They can also be simple generalizations, such as the Pareto principle, which states that 80 percent of effects come from 20 percent of the causes.

A perfect pattern or model is one that (a) accurately describes a situation, (b) is broadly applicable, and (c) can be described in a simple manner. $E = MC^2$ would be such a *general*, *accurate*, and *simple* (GAS) model. Very often, all three qualities are not achievable in a single model, and one has to settle for two of three qualities in the model.

Patterns can be temporal, which is something that regularly occurs over time. Patterns can also be spatial, such as things being organized in a certain way. Patterns can be functional, in that doing certain things leads to certain effects. Good patterns are often symmetric. They echo basic structures and patterns that we are already aware of.

A temporal rule would be that "some people are always late," no matter what the occasion or time. Some people may be aware of this pattern and some may not be. Understanding a pattern like this would help dissipate a lot of unnecessary frustration and anger. One can just joke that some people are born "10 minutes late," and laugh it away. Similarly, Parkinson's law states that works expands to fill up all the time available to do it.

A spatial pattern, following the 80–20 rule, could be that the top 20 percent of customers lead to 80 percent of the business. Or 20 percent of products generate 80 percent of the business. Or 80 percent of incoming customer service calls are related to just 20 percent of the products. This last pattern may simply reveal a discrepancy between a product's features and what the customers believe about the product. The business can then decide to invest in educating the customers better so that the customer service calls can be significantly reduced.

A functional pattern may involve test-taking skills. Some students perform well on essay-type questions. Others do well in multiple-choice questions. Yet other students excel in doing hands-on projects, or in oral presentations. An awareness of such a pattern in a class of students can help the teacher design a balanced testing mechanism that is fair to all.

Retaining students is an ongoing challenge for universities. Recent data-based research shows that students leave a school for social reasons more than they do for academic reasons. This pattern/insight can instigate schools to pay closer attention to students engaging in extracurricular activities and developing stronger bonds at school. The school can invest in entertainment activities, sports activities, camping trips, and other activities. The school can also begin to actively gather data about every student's participation in those activities, to predict at-risk students and take corrective action.

However, long-established patterns can also be broken. The past cannot always predict the future. A pattern like "all swans are white" does not mean that there may not be a black swan. Once enough anomalies are discovered, the underlying pattern itself can shift. The economic meltdown in 2008 to 2009 was because of the collapse of the accepted pattern, that is, "housing prices always go up." A deregulated financial environment made markets more volatile and led to greater swings in markets, leading to the eventual collapse of the entire financial system.

Diamond mining is the act of digging into large amounts of unrefined ore to discover precious gems or nuggets. Similarly, data mining is the act of digging into large amounts of raw data to discover unique nontrivial useful patterns. Data is cleaned up, and then special tools and techniques can be applied to search for patterns. Diving into clean and nicely organized data from the right perspectives can increase the chances of making the right discoveries.

A skilled diamond miner knows what a diamond looks like. Similarly, a skilled data miner should know what kinds of patterns to look for. The patterns are essentially about what hangs together and what is separate. Therefore, knowing the business domain well is very important. It takes knowledge and skill to discover the patterns. It is like finding a needle in a haystack. Sometimes the pattern may be hiding in plain sight. At other times, it may take a lot of work, and looking far and wide, to find surprising useful patterns. Thus, a systematic approach to mining data is necessary to efficiently reveal valuable insights.

For instance, the attitude of employees toward their employer may be hypothesized to be determined by a large number of factors, such as level of education, income, tenure in the company, and gender. It may be surprising if the data reveals that the attitudes are determined first and foremost by their age bracket. Such a simple insight could be powerful in designing organizations effectively. The data miner has to be open to any and all possibilities.

When used in clever ways, data mining can lead to interesting insights and be a source of new ideas and initiatives. One can predict the traffic pattern on highways from the movement of cell phone (in the car) locations on the highway. If the locations of cell phones on a highway or roadway are not moving fast enough, it may be a sign of traffic congestion. Telecom companies can thus provide real-time traffic information to the drivers on their cell phones, or on their GPS devices, without the need of any video cameras or traffic reporters.

Similarly, organizations can find out an employee's arrival time at the office by when their cell phone shows up in the parking lot. Observing the record of the swipe of the parking permit card in the company parking garage can inform the organization whether an employee is in the office building or out of the office at any moment in time.

Some patterns may be so sparse that a very large amount of diverse data has to be seen together to notice any connections. For instance, locating the debris of a flight that may have vanished midcourse would require bringing together data from many sources, such as satellites, ships, and navigation systems. The raw data may come with various levels of quality, and may even be conflicting. The data at hand may or may not be adequate for finding good patterns. Additional dimensions of data may need to be added to help solve the problem.

Data Processing Chain

Data is the new natural resource. Implicit in this statement is the recognition of hidden value in data. Data lies at the heart of business intelligence. There is a sequence of steps to be followed to benefit from the data in a systematic way. Data can be modeled and stored in a database. Relevant data can be extracted from the operational data stores according to certain reporting and analyzing purposes, and stored in a data warehouse. The data from the warehouse can be combined with other sources of data, and mined using data mining techniques to generate new insights. The insights need to be visualized and communicated to the right audience in real time for competitive advantage. Figure 1.2 explains the progression of data processing activities. The rest of this chapter will cover these five elements in the data processing chain.

Data

Anything that is recorded is data. Observations and facts are data. Anecdotes and opinions are also data, of a different kind. Data can be numbers, such as the record of daily weather or daily sales. Data can be alphanumeric, such as the names of employees and customers.

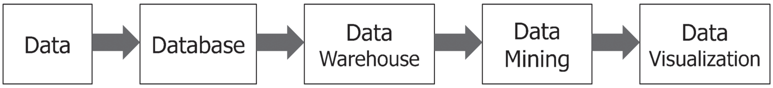


Figure 1.2 Data processing chain

1. Data could come from any number of sources. It could come from operational records inside an organization, and it can come from records compiled by the industry bodies and government agencies. Data could come from individuals telling stories from memory and from people's interaction in social contexts. Data could come from machines reporting their own status or from logs of web usage.
2. Data can come in many ways. It may come as paper reports. It may come as a file stored on a computer. It may be words spoken over the phone. It may be e-mail or chat on the Internet. It may come as movies and songs in DVDs, and so on.
3. There is also data about data. It is called metadata. For example, people regularly upload videos on YouTube. The format of the video file (whether it was a high-def file or lower resolution) is metadata. The information about the time of uploading is metadata. The account from which it was uploaded is also metadata. The record of downloads of the video is also metadata.

Data can be of different types.

1. Data could be an unordered collection of values. For example, a retailer sells shirts of red, blue, and green colors. There is no intrinsic ordering among these color values. One can hardly argue that any one color is higher or lower than the other. This is called nominal (means names) data.
2. Data could be ordered values like small, medium, and large. For example, the sizes of shirts could be extra-small, small, medium, and large. There is clarity that medium is bigger than small, and large is bigger than medium. But the differences may not be equal. This is called ordinal (ordered) data.
3. Another type of data has discrete numeric values defined in a certain range, with the assumption of equal distance between the values. Customer satisfaction score may be ranked on a 10-point scale with 1 being lowest and 10 being highest. This requires the respondent to carefully calibrate the entire range as objectively as possible and place his or her own measurement in that scale. This is called interval (equal intervals) data.

4. The highest level of numeric data is ratio data that can take on any numeric value. The weights and heights of all employees would be exact numeric values. The price of a shirt will also take any numeric value. It is called ratio (any fraction) data.
5. There is another kind of data that does not lend itself to much mathematical analysis, at least not directly. Such data needs to be first structured and then analyzed. This includes data like audio, video, and graphs files, often called BLOBs (Binary Large Objects). These kinds of data lend themselves to different forms of analysis and mining. Songs can be described as happy or sad, fast-paced or slow, and so on. They may contain sentiment and intention, but these are not quantitatively precise.

The precision of analysis increases as data becomes more numeric. Ratio data could be subjected to rigorous mathematical analysis. For example, precise weather data about temperature, pressure, and humidity can be used to create rigorous mathematical models that can accurately predict future weather.

Data may be publicly available and sharable, or it may be marked private. Traditionally, the law allows the right to privacy concerning one's personal data. There is a big debate on whether the personal data shared on social media conversations is private or can be used for commercial purposes.

Datafication is a new term that means that almost every phenomenon is now being observed and stored. More devices are connected to the Internet. More people are constantly connected to "the grid," by their phone network or the Internet, and so on. Every click on the web, and every movement of the mobile devices, is being recorded. Machines are generating data. The "Internet of things" is growing faster than the Internet of people. All of this is generating an exponentially growing volume of data, at high velocity. Kryder's law predicts that the density and capability of hard drive storage media will double every 18 months. As storage costs keep coming down at a rapid rate, there is a greater incentive to record and store more events and activities at a higher resolution. Data is getting stored in more detailed resolution, and many more variables are being captured and stored.

Database

A database is a modeled collection of data that is accessible in many ways. A data model can be designed to integrate the operational data of the organization. The data model abstracts the key entities involved in an action and their relationships. Most databases today follow the relational data model and its variants. Each data modeling technique imposes rigorous rules and constraints to ensure the integrity and consistency of data over time.

Take the example of a sales organization. A data model for managing customer orders will involve data about customers, orders, products, and their interrelationships. The relationship between the customers and orders would be such that one customer can place many orders, but one order will be placed by one and only one customer. It is called a one-to-many relationship. The relationship between orders and products is a little more complex. One order may contain many products. And one product may be contained in many different orders. This is called a many-to-many relationship. Different types of relationships can be modeled in a database.

Databases have grown tremendously over time. They have grown in complexity in terms of number of the objects and their properties being recorded. They have also grown in the quantity of data being stored. A decade ago, a terabyte-sized database was considered big. Today databases are in petabytes and exabytes. Video and other media files have greatly contributed to the growth of databases. E-commerce and other web-based activities also generate huge amounts of data. Data generated through social media has also generated large databases. The e-mail archives, including attached documents of organizations, are in similar large sizes.

Many database management software systems (DBMSs) are available to help store and manage this data. These include commercial systems, such as Oracle and DB2 system. There are also open-source, free DBMS, such as MySQL and Postgres. These DBMSs help process and store millions of transactions worth of data every second.

Here is a simple database of the sales of movies worldwide for a retail organization. It shows sales transactions of movies over three quarters. Using such a file, data can be added, accessed, and updated as needed.

Movies Transaction Database				
Order #	**Date sold**	**Product name**	**Location**	**Total value**
1	April 2013	Monty Python	United States	$9
2	May 2013	Gone With the Wind	United States	$15
3	June 2013	Monty Python	India	$9
4	June 2013	Monty Python	United Kingdom	$12
5	July 2013	Matrix	United States	$12
6	July 2013	Monty Python	United States	$12
7	July 2013	Gone With the Wind	United States	$15
8	Aug 2013	Matrix	United States	$12
9	Sept 2013	Matrix	India	$12
10	Sept 2013	Monty Python	United States	$9
11	Sept 2013	Gone With the Wind	United States	$15
12	Sept 2013	Monty Python	India	$9
13	Nov 2013	Gone With the Wind	United States	$15
14	Dec 2013	Monty Python	United States	$9
15	Dec 2013	Monty Python	United States	$9

Data Warehouse

A data warehouse is an organized store of data from all over the organization, specially designed to help make management decisions. Data can be extracted from operational database to answer a particular set of queries. This data, combined with other data, can be rolled up to a consistent granularity and uploaded to a separate data store called the data warehouse. Therefore, the data warehouse is a simpler version of the operational data base, with the purpose of addressing reporting and decision-making needs only. The data in the warehouse cumulatively grows as more operational data becomes available and is extracted and appended to the data warehouse. Unlike in the operational database, the data values in the warehouse are not updated.

To create a simple data warehouse for the movies sales data, assume a simple objective of tracking sales of movies and making decisions about managing inventory. In creating this data warehouse, all the sales transaction data will be extracted from the operational data files. The data will be rolled up for all combinations of time period and product number. Thus, there will be one row for every combination of time period and product. The resulting data warehouse will look like the table what follows.

Movies Sales Data Warehouse			
Row #	**Qtr Sold**	**Product Name**	**Total Value**
1	Q2	Gone With the Wind	$15
2	Q2	Monty Python	$30
3	Q3	Gone With the Wind	$30
4	Q3	Matrix	$36
5	Q3	Monty Python	$30
6	Q4	Gone With the Wind	$15
7	Q4	Monty Python	$18

The data in the data warehouse is at much less detail than the transaction database. The data warehouse could have been designed at a lower or higher level of detail, or granularity. If the data warehouse were designed on a monthly level, instead of a quarterly level, there would be many more rows of data. When the number of transactions approaches millions and higher, with dozens of attributes in each transaction, the data warehouse can be large and rich with potential insights. One can then mine the data (slice and dice) in many different ways and discover unique meaningful patterns. Aggregating the data helps improve the speed of analysis. A separate data warehouse allows analysis to go on separately in parallel, without burdening the operational database systems (Table 1.1).

Data Mining

Data Mining is the art and science of discovering useful innovative patterns from data. There is a wide variety of patterns that can be found in the data. There are many techniques, simple or complex, that help with finding patterns.

Table 1.1 Comparing database systems with data warehousing systems

Function	Database	Data Warehouse
Purpose	Data stored in databases can be used for many purposes including day-to-day operations	Data in data warehouse is cleansed data, which is useful for reporting and analysis
Granularity	Highly granular data including all activity and transaction details	Lower granularity data; rolled up to certain key dimensions of interest
Complexity	Highly complex with dozens or hundreds of data files, linked through common data fields	Typically organized around a large fact tables, and many lookup tables
Size	Database grows with growing volumes of activity and transactions. Old completed transactions are deleted to reduce size	Grows as data from operational databases is rolled up and appended every day. Data is retained for long-term trend analyses
Architectural choices	Relational, and object-oriented, databases	Star schema or Snowflake schema
Data access mechanisms	Primarily through high-level languages such as SQL. Traditional programming access database through Open Database Connectivity (ODBC) interfaces	Accessed through SQL; SQL output is forwarded to reporting tools and data visualization tools

In this example, a simple data analysis technique can be applied to the data in the data warehouse mentioned earlier. A simple cross-tabulation of results by quarter and products will reveal some easily visible patterns.

Movies Sales by Quarters—Cross-tabulation				
Qtr/Product	Gone With the Wind	Matrix	Monty Python	Total Sales
Q2	$15	0	$30	$45
Q3	$30	$36	$30	$96
Q4	$15	0	$18	$33
Total Sales	$60	$36	$78	$174

Based on this cross-tabulation, one can readily answer some product sales questions, such as:

1. What is the best selling movie by revenue?—***Monty Python***
2. What is the best quarter by revenue this year?—***Q3***
3. Any other patterns?—Matrix movie sells only in ***Q3 (seasonal item).***

These simple insights can help plan marketing promotions and manage inventory of various movies.

If a cross-tabulation was designed to include customer location data, one could answer other questions, such as:

1. What is the best selling geography?—United States
2. What is the worst selling geography?—United Kingdom
3. Any other patterns?—Monty Python sells globally, while Gone with the Wind sells only in the United States.

If the data mining was done at the monthly level of data, it would be easy to miss the seasonality of the movies. However, one would have observed that September is the highest selling month.

The previous example shows that many differences and patterns can be noticed by analyzing data in different ways. However, some insights are more important than others. The value of the insight depends upon the problem being solved. The insight that there are more sales of a product in a certain quarter helps a manager plan what products to focus on. In this case, the store manager should stock up on Matrix in Quarter 3 (Q3). Similarly, knowing which quarter has the highest overall sales allows for different resource decisions in that quarter. In this case, if Q3 is bringing more than half of total sales, this requires greater attention on the e-commerce website in the third quarter.

Data mining should be done to solve high-priority, high-value problems. Much effort is required to gather data, clean and organize it, mine it with many techniques, interpret the results, and find the right insight. It is important that there be a large expected payoff from finding the insight. One should select the right data (and ignore the rest), organize it into a nice and imaginative framework that brings relevant data together, and then apply data mining techniques to deduce the right insight.

A retail company may use data mining techniques to determine which new product categories to add to which of their stores; how to increase sales of existing products; which new locations to open stores in; how to segment the customers for more effective communication; and so on.

Data can be analyzed at multiple levels of granularity and could lead to a large number of interesting combinations of data and interesting patterns. Some of the patterns may be more meaningful than the others. Such highly granular data is often used, especially in finance and high-tech areas, so that one can gain even the slightest edge over the competition.

Following are the brief descriptions of some of the most important data mining techniques used to generate insights from data.

Decision trees: They help classify populations into classes. It is said that 70 percent of all data mining work is about classification solutions; and that 70 percent of all classification work uses decision trees. Thus, decision trees are the most popular and important data mining technique. There are many popular algorithms to make decision trees. They differ in terms of their mechanisms and each technique work well for different situations. It is possible to try multiple algorithms on a data set and compare the predictive accuracy of each tree.

Regression: This is a well-understood technique from the field of statistics. The goal is to find a best fitting curve through the many data points. The best fitting curve is that which minimizes the (error) distance between the actual data points and the values predicted by the curve. Regression models can be projected into the future for prediction and forecasting purposes.

Artificial neural networks (ANNs): Originating in the field of artificial intelligence and machine learning, ANNs are multilayer nonlinear information processing models that learn from past data and predict future values. These models predict well, leading to their popularity. The model's parameters may not be very intuitive. Thus, neural networks are opaque like a black box. These systems also require a large amount of past data to adequately train the system.

Cluster analysis: This is an important data mining technique for dividing and conquering large data sets. The data set is divided into a certain number of clusters, by discerning similarities and dissimilarities within the data. There is no one right answer for the number of clusters in the data. The user needs to make a decision by looking at how well the number of clusters chosen fit the data. This is most commonly used for market segmentation. Unlike decision trees and regression, there is no one right answer for cluster analysis.

Association rule mining: Also called market basket analysis when used in retail industry, these techniques look for associations between data values. An analysis of items frequently found together in a market basket can help cross-sell products and also create product bundles.

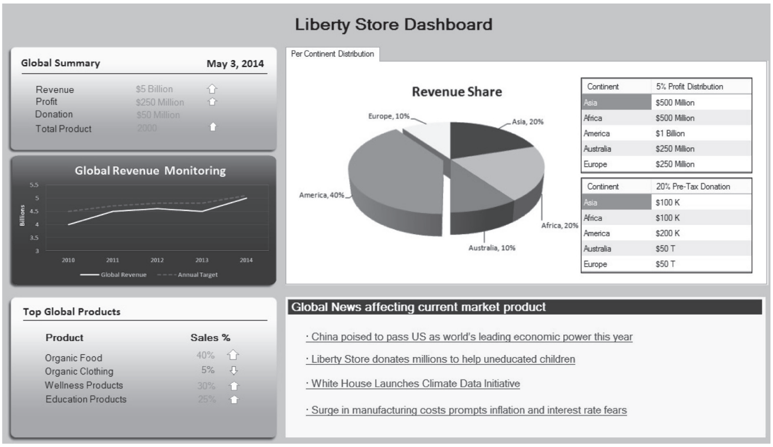
Data Visualization

As data and insights grow in number, a new requirement is the ability of the executives and decision makers to absorb this information in real time. There is a limit to human comprehension and visualization capacity. That is a good reason to prioritize and manage with fewer but key variables that relate directly to the key result areas of a role.

Here are few considerations when presenting data:

1. Present the conclusions and not just report the data.
2. Choose wisely from a palette of graphs to suit the data.
3. Organize the results to make the central point stand out.
4. Ensure that the visuals accurately reflect the numbers. Inappropriate visuals can create misinterpretations and misunderstandings.
5. Make the presentation unique, imaginative, and memorable.

Executive dashboards are designed to provide information on select few variables for every executive. They use graphs, dials, and lists to show the status of important parameters. These dashboards also have a drill-down capability to enable a root-cause analysis of exceptional situations (Figure 1.3).



Liberty Store Dashboard

Global Summary	May 3, 2014	
Revenue	$5 Billion	⇧
Profit	$250 Million	⇧
Donation	$50 Million	
Total Product	2000	⬆

Continent	5% Profit Distribution
Asia	$500 Million
Africa	$500 Million
America	$1 Billion
Australia	$250 Million
Europe	$250 Million

Continent	20% Pre-Tax Donation
Asia	$100 K
Africa	$100 K
America	$200 K
Australia	$50 T
Europe	$50 T

Top Global Products

Product	Sales %	
Organic Food	40%	⇧
Organic Clothing	5%	⇩
Wellness Products	30%	⬆
Education Products	25%	⬆

Global News affecting current market product

· China poised to pass US as world's leading economic power this year

· Liberty Store donates millions to help uneducated children

· White House Launches Climate Data Initiative

· Surge in manufacturing costs prompts inflation and interest rate fears

Figure 1.3 Sample executive dashboard

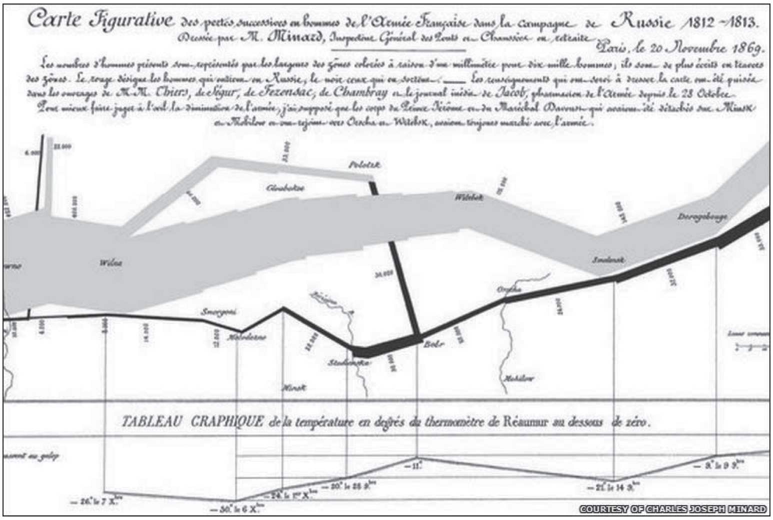

Carte Figurative des pertes successives en hommes de l'Armée Française dans la campagne de Russie 1812-1813.
Dressée par M. Minard, Inspecteur Général des Ponts et Chaussées en retraite.
Paris, le 20 Novembre 1869.
Wilna
Smorgoni
Molodezno
Minsk
Studienska
Bobr
Polotzk
Glubokoe
Witebsk
Orscha
Mohilow
Smolensk
Dorogobouge
TABLEAU GRAPHIQUE de la température en degrés du thermomètre de Réaumur au dessous de zéro.
-26° le 7 X.bre
-30° le 6 X.bre
-24° le 1er X.bre
-20° le 28 9.bre
-11°
-21° le 14 9.bre
-9° le 9 9.bre
COURTESY OF CHARLES JOSEPH MINARD

Figure 1.4 Sample data visualization

Data visualization has been an interesting problem across the disciplines. Many dimensions of data can be effectively displayed on a two-dimensional surface to give a rich and more insightful description of the totality of the story.

The classic presentation of the story of Napoleon's march to Russia in 1812, by French cartographer Joseph Minard, is shown in Figure 1.4. It covers about six dimensions. Time is on horizontal axis. The geographical coordinates and rivers are mapped in. The thickness of the bar shows the number of troops at any point of time that is mapped. One color is used for the onward march and another for the retreat. The weather temperature at each time is shown in the line graph at the bottom.

Organization of the Book

This chapter is designed to provide the wholeness of business intelligence and data mining, to provide the reader with an intuition for this area of knowledge. The rest of the book can be considered in three sections.

Section 1 will cover high-level topics. Chapter 2 will cover the field of business intelligence and its applications across industries and functions. Chapter 3 will briefly explain what data warehousing is and how it helps with data mining. Chapter 4 will then describe data mining in some detail with an overview of its major tools and techniques.

Section 2 is focused on data mining techniques. Every technique will be shown through solving an example in detail. Chapter 5 will show the power and ease of decision trees, which are the most popular data mining technique. Chapter 6 will describe statistical regression modeling techniques. Chapter 7 will provide an overview of ANNs. Chapter 8 will describe how cluster analysis can help with market segmentation. Finally, Chapter 9 will describe the association rule mining technique, also called market basket analysis, which helps find shopping patterns.

Section 3 will cover more advanced new topics. Chapter 10 will introduce the concepts and techniques of text mining, which helps discover insights from text data, including social media data. Chapter 11 will provide an overview of the growing field of web mining, which includes mining the structure, content, and usage of websites. Chapter 12 will provide an overview of the field of Big Data. Chapter 13 has been added as a primer on data modeling, for those who do not have any background in databases, and should be used if necessary.

Review Questions

1. Describe the business intelligence and data mining cycle.
2. Describe the data processing chain.
3. What are the similarities between diamond mining and data mining?
4. What are the different data mining techniques? Which of these would be relevant in your current work?
5. What is a dashboard? How does it help?
6. Create a visual to show the weather pattern in your city. Could you show together temperature, humidity, wind, and rain/snow over a period of time.

SECTION 1

This section covers three important high-level topics.

Chapter 2 will cover business intelligence concepts, and its applications in many industries.

Chapter 3 will describe data warehousing systems, and ways of creating and managing them.

Chapter 4 will describe data mining as a whole, with many do's and don'ts of effective data mining.

CHAPTER 2

Business Intelligence Concepts and Applications

Business intelligence (BI) is an umbrella term that includes a variety of IT applications that are used to analyze an organization's data and communicate the information to relevant users. Its major components are data warehousing, data mining, querying, and reporting (Figure 2.1).

The nature of life and businesses is to grow. Information is the lifeblood of business. Businesses use many techniques for understanding their environment and predicting the future for their own benefit and growth. Decisions are made from facts and feelings. Data-based decisions are more effective than those based on feelings alone. Actions based on accurate data, information, knowledge, experimentation, and testing, using fresh insights, can more likely succeed and lead to sustained growth.

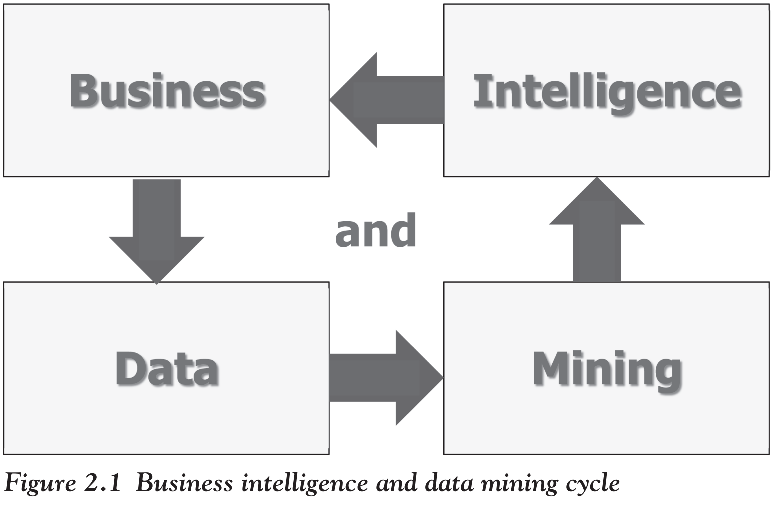


Figure 2.1 Business intelligence and data mining cycle

One's own data can be the most effective teacher. Therefore, organizations should gather data, sift through it, analyze and mine it, find insights, and then embed those insights into their operating procedures.

There is a new sense of importance and urgency around data as it is being viewed as a new natural resource. It can be mined for value, insights, and competitive advantage. In a hyperconnected world, where everything is potentially connected to everything else, with potentially infinite correlations, data represents the impulses of nature in the form of certain events and attributes. A skilled business person is motivated to use this cache of data to harness nature, and to find new niches of unserved opportunities that could become profitable ventures.

Caselet: Khan Academy—BI in Education

Khan Academy is an innovative nonprofit educational organization that is turning the K-12 education system upside down. It provides short YouTube-based video lessons on thousands of topics for free. It shot into prominence when Bill Gates promoted it as a resource that he used to teach his own children. With this kind of a resource, classrooms are being flipped—that is, students do their basic lecture-type learning at home using those videos, while the class time is used for more one-on-one problem solving and coaching. Students can access the lessons at any time to learn at their own pace. The students' progress is recorded, including what videos they watched, how many times they watched, which problems they stumbled on, and what scores they got on online tests.

Khan Academy has developed tools to help teachers get a pulse on what is happening in the classroom. Teachers are provided a set of real-time dashboards to give them information from the macrolevel ("How is my class doing on geometry?") to the micro level ("How is Jane doing on mastering polygons?"). Armed with this information, teachers can place selective focus on the students that need certain help. (Source: KhanAcademy.org)

Q1. *How does a dashboard improve the teaching experience and the student's learning experience?*

Q2. *Design a dashboard for tracking your own career.*

BI for Better Decisions

The future is inherently uncertain. Risk is the result of a probabilistic world where there are no certainties and complexities abound. People use crystal balls, astrology, palmistry, ground hogs, and also mathematics and numbers to mitigate risk in decision-making. The goal is to make effective decisions, while reducing risk. Businesses calculate risks and make decisions based on a broad set of facts and insights. Reliable knowledge about the future can help managers make the right decisions with lower levels of risk.

The speed of action has risen exponentially with the growth of the Internet. In a hypercompetitive world, the speed of a decision and the consequent action can be a key advantage. The Internet and mobile technologies allow decisions to be made anytime, anywhere. Ignoring fast-moving changes can threaten the organization's future. Research has shown that an unfavorable comment about the company and its products on social media should not go unaddressed for long. Banks have had to pay huge penalties to Consumer Financial Protection Bureau (CFPB) in United States in 2013 for complaints made on CFPB's websites. On the other hand, a positive sentiment expressed on social media should also be utilized as a potential sales and promotion opportunity, while the opportunity lasts.

Decision Types

There are two main kinds of decisions: strategic decisions and operational decisions. BI can help make both better. Strategic decisions are those that impact the direction of the company. The decision to reach out to a new customer set would be a strategic decision. Operational decisions are more routine and tactical decisions, focused on developing greater efficiency. Updating an old website with new features will be an operational decision.

In strategic decision-making, the goal itself may or may not be clear, and the same is true for the path to reach the goal. The consequences of the decision would be apparent some time later. Thus, one is constantly scanning for new possibilities and new paths to achieve the goals. BI can help with what-if analysis of many possible scenarios. BI can also help create new ideas based on new patterns found from data mining.

Operational decisions can be made more efficient using an analysis of past data. A classification system can be created and modeled using the data of past instances to develop a good model of the domain. This model can help improve operational decisions in the future. BI can help automate operations level decision-making and improve efficiency by making millions of microlevel operational decisions in a model-driven way. For example, a bank might want to make decisions about making financial loans in a more scientific way using data-based models. A decision-tree-based model could provide a consistently accurate loan decisions. Developing such decision tree models is one of the main applications of data mining techniques.

Effective BI has an evolutionary component, as business models evolve. When people and organizations act, new facts (data) are generated. Current business models can be tested against the new data, and it is possible that those models will not hold up well. In that case, decision models should be revised and new insights should be incorporated. An unending process of generating fresh new insights in real time can help make better decisions, and thus can be a significant competitive advantage.

BI Tools

BI includes a variety of software tools and techniques to provide the managers with the information and insights needed to run the business. Information can be provided about the current state of affairs with the capability to drill down into details, and also insights about emerging patterns which lead to projections into the future. BI tools include data warehousing, online analytical processing, social media analytics, reporting, dashboards, querying, and data mining.

BI tools can range from very simple tools that could be considered end-user tools, to very sophisticated tools that offer a very broad and complex set of functionality. Thus, Even executives can be their own BI experts, or they can rely on BI specialists to set up the BI mechanisms for them. Thus, large organizations invest in expensive sophisticated BI solutions that provide good information in real time.

A spreadsheet tool, such as Microsoft Excel, can act as an easy but effective BI tool by itself. Data can be downloaded and stored in the spreadsheet, then analyzed to produce insights, then presented in the form of graphs and tables. This system offers limited automation using macros and other features. The analytical features include basic statistical and financial functions. Pivot tables help do sophisticated what-if analysis. Add-on modules can be installed to enable moderately sophisticated statistical analysis.

A dashboarding system, such as Tableau, can offer a sophisticated set of tools for gathering, analyzing, and presenting data. At the user end, modular dashboards can be designed and redesigned easily with a graphical user interface. The back-end data analytical capabilities include many statistical functions. The dashboards are linked to data warehouses at the back end to ensure that the tables and graphs and other elements of the dashboard are updated in real time (Figure 2.2).

Data mining systems, such as IBM SPSS Modeler, are industrial strength systems that provide capabilities to apply a wide range of analytical models on large data sets. Open source systems, such as Weka, are popular platforms designed to help mine large amounts of data to discover patterns.

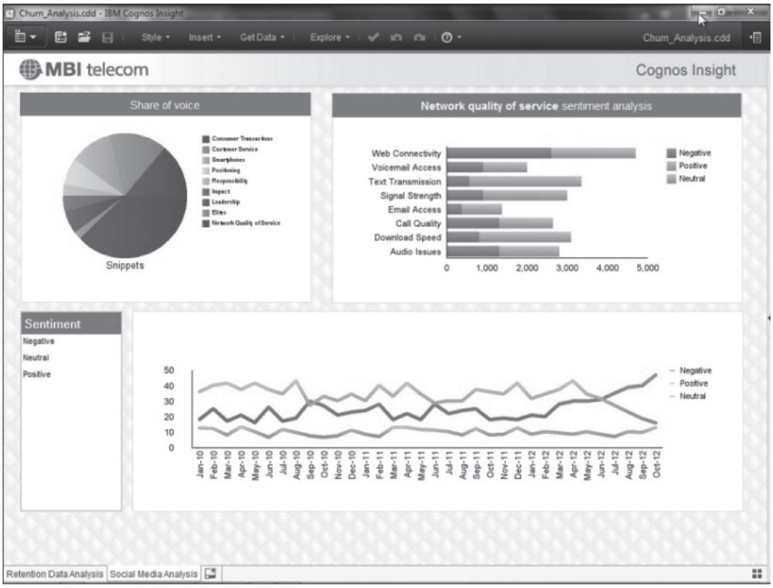


Figure 2.2 Sample executive dashboard

BI Skills

As data grows and exceeds our capacity to make sense of it, the tools need to evolve, and so should the imagination of the BI specialist. "Data Scientist" has been called as the hottest job of this decade.

A skilled and experienced BI specialist should be open enough to go outside the box, open the aperture and see a wider perspective that includes more dimensions and variables, in order to find important patterns and insights. The problem needs to be looked at from a wider perspective to consider many more angles that may not be immediately obvious. An imaginative solution should be proposed for the problem so that interesting and useful results can emerge.

A good data mining project begins with an interesting problem to solve. Selecting the right data mining problem is an important skill. The problem should be valuable enough that solving it would be worth the time and expense. It takes a lot of time and energy to gather, organize, cleanse, and prepare the data for mining and other analysis. The data miner needs to persist with the exploration of patterns in the data. The skill level has to be deep enough to engage with the data and make it yield new useful insights.

BI Applications

BI tools are required in almost all industries and functions. The nature of the information and the speed of action may be different across businesses, but every manager today needs access to BI tools to have up-to-date metrics about business performance. Businesses need to embed new insights into their operating processes to ensure that their activities continue to evolve with more efficient practices. The following are some areas of applications of BI and data mining.

Customer Relationship Management

A business exists to serve a customer. A happy customer becomes a repeat customer. A business should understand the needs and sentiments of the customer, sell more of its offerings to the existing customers, and also, expand the pool of customers it serves. BI applications can impact many aspects of marketing.

1. *Maximize the return on marketing campaigns:* Understanding the customer's pain points from data-based analysis can ensure that the marketing messages are fine-tuned to better resonate with customers.
2. *Improve customer retention (churn analysis):* It is more difficult and expensive to win new customers than it is to retain existing customers. Scoring each customer on their likelihood to quit can help the business design effective interventions, such as discounts or free services, to retain profitable customers in a cost-effective manner.
3. *Maximize customer value (cross-selling, upselling):* Every contact with the customer should be seen as an opportunity to gauge their current needs. Offering a customer new products and solutions based on those imputed needs can help increase revenue per customer. Even a customer complaint can be seen as an opportunity to wow the customer. Using the knowledge of the customer's history and value, the business can choose to sell a premium service to the customer.
4. *Identify and delight highly valued customers:* By segmenting the customers, the best customers can be identified. They can be proactively contacted, and delighted, with greater attention and better service. Loyalty programs can be managed more effectively.
5. *Manage brand image:* A business can create a listening post to listen to social media chatter about itself. It can then do sentiment analysis of the text to understand the nature of comments and respond appropriately to the prospects and customers.

Health Care and Wellness

Health care is one of the biggest sectors in advanced economies. Evidence-based medicine is the newest trend in data-based health care management. BI applications can help apply the most effective diagnoses and prescriptions for various ailments. They can also help manage public health issues, and reduce waste and fraud.

1. *Diagnose disease in patients:* Diagnosing the cause of a medical condition is the critical first step in a medical engagement. Accurately diagnosing cases of cancer or diabetes can be a matter of life and death for the patient. In addition to the patient's own current situation, many

other factors can be considered, including the patient's health history, medication history, family's history, and other environmental factors. This makes diagnosis as much of an art form as it is science. Systems, such as IBM Watson, absorb all the medical research to date and make probabilistic diagnoses in the form of a decision tree, along with a full explanation for their recommendations. These systems take away most of the guess work done by doctors in diagnosing ailments.

2. *Treatment effectiveness:* The prescription of medication and treatment is also a difficult choice out of so many possibilities. For example, there are more than 100 medications for hypertension (high blood pressure) alone. There are also interactions in terms of which drugs work well with others and which drugs do not. Decision trees can help doctors learn about and prescribe more effective treatments. Thus, the patients could recover their health faster with a lower risk of complications and cost.
3. *Wellness management:* This includes keeping track of patient health records, analyzing customer health trends, and proactively advising them to take any needed precautions.
4. *Manage fraud and abuse:* Some medical practitioners have unfortunately been found to conduct unnecessary tests and/or overbill the government and health insurance companies. Exception-reporting systems can identify such providers, and action can be taken against them.
5. *Public health management:* The management of public health is one of the important responsibilities of any government. By using effective forecasting tools and techniques, governments can better predict the onset of disease in certain areas in real time. They can thus be better prepared to fight the diseases. Google has been known to predict the movement of certain diseases by tracking the search terms (like flu, vaccine) used in different parts of the world.

Education

As higher education becomes more expensive and competitive, it is a great user of data-based decision-making. There is a strong need for efficiency, increasing revenue, and improving the quality of student experience at all levels of education.

1. *Student enrolment (recruitment and retention):* Marketing to new potential students requires schools to develop profiles of the students that are most likely to attend. Schools can develop models of what kinds of students are attracted to the school, and then reach out to those students. The students at risk of not returning can be flagged, and corrective measures can be taken in time.
2. *Course offerings:* Schools can use the class enrolment data to develop models of which new courses are likely to be more popular with students. This can help increase class size, reduce costs, and improve student satisfaction.
3. *Alumni pledges:* Schools can develop predictive models of which alumni are most likely to pledge financial support to the school. Schools can create a profile for alumni more likely to pledge donations to the school. This could lead to a reduction in the cost of mailings and other forms of outreach to alumni.

Retail

Retail organizations grow by meeting customer needs with quality products, in a convenient, timely, and cost-effective manner. Understanding emerging customer shopping patterns can help retailers organize their products, inventory, store layout, and web presence in order to delight their customers, which in turn would help increase revenue and profits. Retailers generate a lot of transaction and logistics data that can be used to solve problems.

1. *Optimize inventory levels at different locations:* Retailers need to manage their inventories carefully. Carrying too much inventory imposes carrying costs, while carrying too little inventory can cause stock-outs and lost sales opportunities. Predicting sales trends dynamically can help retailers move inventory to where it is most in demand. Retail organizations can provide their suppliers with real-time information about sales of their items so that the suppliers can deliver their product to the right locations and minimize stock-outs.
2. *Improve store layout and sales promotions:* A market basket analysis can develop predictive models of which products sell together

often. This knowledge of affinities between products can help retailers co-locate those products. Alternatively, those affinity products could be located farther apart to make the customer walk the length and breadth of the store, and thus be exposed to other products. Promotional discounted product bundles can be created to push a nonselling item along with a set of products that sell well together.

3. *Optimize logistics for seasonal effects:* Seasonal products offer tremendously profitable short-term sales opportunities, yet they also offer the risk of unsold inventories at the end of the season. Understanding which products are in season in which market can help retailers dynamically manage prices to ensure their inventory is sold during the season. If it is raining in a certain area, then the inventory of umbrella and ponchos could be rapidly moved there from nonrainy areas to help increase sales.
4. *Minimize losses due to limited shelf life:* Perishable goods offer challenges in terms of disposing off the inventory in time. By tracking sales trends, the perishable products at risk of not selling before the sell-by date can be suitably discounted and promoted.

Banking

Banks make loans and offer credit cards to millions of customers. They are most interested in improving the quality of loans and reducing bad debts. They also want to retain more good customers and sell more services to them.

1. *Automate the loan application process:* Decision models can be generated from past data that predict the likelihood of a loan proving successful. These can be inserted in business processes to automate the financial loan application process.
2. *Detect fraudulent transactions:* Billions of financial transactions happen around the world every day. Exception-seeking models can identify patterns of fraudulent transactions. For example, if money is being transferred to an unrelated account for the first time, it could be a fraudulent transaction.

3. *Maximize customer value (cross-selling, upselling):* Selling more products and services to existing customers is often the easiest way to increase revenue. A checking account customer in good standing could be offered home, auto, or educational loans on more favorable terms than other customers, and thus, the value generated from that customer could be increased.
4. *Optimize cash reserves with forecasting:* Banks have to maintain certain liquidity to meet the needs of depositors who may like to withdraw money. Using past data and trend analysis, banks can forecast how much to keep, and invest the rest to earn interest.

Financial Services

Stock brokerages are an intensive user of BI systems. Fortunes can be made or lost based on access to accurate and timely information.

1. *Predict changes in bond and stock prices:* Forecasting the price of stocks and bonds is a favorite pastime of financial experts as well as lay people. Stock transaction data from the past, along with other variables, can be used to predict future price patterns. This can help traders develop long-term trading strategies.
2. *Assess the effect of events on market movements:* Decision models using decision trees can be created to assess the impact of events on changes in market volume and prices. Monetary policy changes (such as Fed Reserve interest rate change) or geopolitical changes (such as war in a part of the world) can be factored into the predictive model to help take action with greater confidence and less risk.
3. *Identify and prevent fraudulent activities in trading:* There have unfortunately been many cases of insider trading, leading to many prominent financial industry stalwarts going to jail. Fraud detection models can identify and flag fraudulent activity patterns.

Insurance

This industry is a prolific user of prediction models in pricing insurance proposals and managing losses from claims against insured assets.

1. *Forecast claim costs for better business planning:* When natural disasters, such as hurricanes and earthquakes, strike, loss of life and property occurs. By using the best available data to model the likelihood (or risk) of such events happening, the insurer can plan for losses and manage resources and profits effectively.
2. *Determine optimal rate plans:* Pricing an insurance rate plan requires covering the potential losses and making a profit. Insurers use actuary tables to project life spans and disease tables to project mortality rates, and thus price themselves competitively yet profitably.
3. *Optimize marketing to specific customers:* By microsegmenting potential customers, a data-savvy insurer can cherry-pick the best customers and leave the less profitable customers to its competitors. Progressive Insurance is a U.S.-based company that is known to actively use data mining to cherry-pick customers and increase its profitability.
4. *Identify and prevent fraudulent claim activities:* Patterns can be identified as to where and what kinds of fraud are more likely to occur. Decision-tree-based models can be used to identify and flag fraudulent claims.

Manufacturing

Manufacturing operations are complex systems with interrelated subsystems. From machines working right, to workers having the right skills, to the right components arriving with the right quality at the right time, to money to source the components, many things have to go right. Toyota's famous lean manufacturing company works on just-in-time inventory systems to optimize investments in inventory and to improve flexibility in their product mix.

1. *Discover novel patterns to improve product quality:* Quality of a product can also be tracked, and this data can be used to create a predictive model of product quality deteriorating. Many companies, such as automobile companies, have to recall their products if they have found defects that have a public safety implication. Data mining can help with root cause analysis that can be used to identify sources of errors and help improve product quality in the future.

2. *Predict/prevent machinery failures:* Statistically, all equipment is likely to break down at some point in time. Predicting which machine is likely to shut down is a complex process. Decision models to forecast machinery failures could be constructed using past data. Preventive maintenance can be planned, and manufacturing capacity can be adjusted, to account for such maintenance activities.

Telecom

BI in telecom can help with churn management, marketing/customer profiling, network failure, and fraud detection.

1. *Churn management:* Telecom customers have shown a tendency to switch their providers in search for better deals. Telecom companies tend to respond with many incentives and discounts to hold on to customers. However, they need to determine which customers are at a real risk of switching and which others are just negotiating for a better deal. The level of risk should to be factored into the kind of deals and discounts that should be given. Millions of such customer calls happen every month. The telecom companies need to provide a consistent and data-based way to predict the risk of the customer switching, and then make an operational decision in real time while the customer call is taking place. A decision-tree- or a neural network-based system can be used to guide the customer-service call operator to make the right decisions for the company, in a consistent manner.
2. *Marketing and product creation:* In addition to customer data, telecom companies also store call detail records (CDRs), which precisely describe the calling behavior of each customer. This unique data can be used to profile customers and then can be used for creating new products/services bundles for marketing purposes. An American telecom company, MCI, created a program called Friends & Family that allowed calls with one's friends and family on that network to be totally free and thus, effectively locked many people into their network.
3. *Network failure management:* Failure of telecom networks for technical failures or malicious attacks can have devastating impacts on

people, businesses, and society. In telecom infrastructure, some equipment will likely fail with certain mean time between failures. Modeling the failure pattern of various components of the network can help with preventive maintenance and capacity planning.

4. *Fraud management:* There are many kinds of fraud in consumer transactions. Subscription fraud occurs when a customer opens an account with the intention of never paying for the services. Superimposition fraud involves illegitimate activity by a person other than the legitimate account holder. Decision rules can be developed to analyze each CDR in real time to identify chances of fraud and take effective action.

Government

Government gathers a large amount of data by virtue of their regulatory function. That data could be analyzed for developing models of effective functioning.

1. *Law enforcement:* Social behavior is a lot more patterned and predictable than one would imagine. For example, Los Angeles Police Department (LAPD) mined the data from its 13 million crime records over 80 years and developed models of what kind of crime going to happen when and where. By increasing patrolling in those particular areas, LAPD was able to reduce property crime by 27 percent. Internet chatter can be analyzed to learn of and prevent any evil designs.
2. *Scientific research:* Any large collection of research data is amenable to being mined for patterns and insights. Protein folding (microbiology), nuclear reaction analysis (subatomic physics), disease control (public health) are some examples where data mining can yield powerful new insights.

Conclusion

BI is a comprehensive set of IT tools to support decision-making with imaginative solutions for a variety of problems. BI can help improve the performance in nearly all industries and applications.

Review Questions

1. Why should organizations invest in business intelligence solutions? Are these more important than IT security solutions? Why or why not?
2. List three business intelligence applications in the hospitality industry.
3. Describe two business intelligence tools used in your organization.
4. Businesses need a "two-second advantage" to succeed. What does that mean to you?

Liberty Stores Case Exercise: Step 1

Liberty Stores Inc is a specialized global retail chain that sells organic food, organic clothing, wellness products, and education products to enlightened LOHAS (Lifestyles of the Healthy and Sustainable) citizens worldwide. The company is 20 years old and is growing rapidly. It now operates in 5 continents, 50 countries, 150 cities, and has 500 stores. It sells 20,000 products and has 10,000 employees. The company has revenues of over $5 billion and has a profit of about 5 percent of revenue. The company pays special attention to the conditions under which the products are grown and produced. It donates about one-fifth (20 percent) of its pretax profits from global local charitable causes.

1. *Create a comprehensive dashboard for the CEO of the company.*
2. *Create another dashboard for a country head.*

CHAPTER 3

Data Warehousing

A data warehouse (DW) is an organized collection of integrated, subject-oriented databases designed to support decision support functions. DW is organized at the right level of granularity to provide clean enterprise-wide data in a standardized format for reports, queries, and analysis. DW is physically and functionally separate from an operational and transactional database. Creating a DW for analysis and queries represents significant investment in time and effort. It has to be constantly kept up-to-date for it to be useful. DW offers many business and technical benefits.

DW supports business reporting and data mining activities. It can facilitate distributed access to up-to-date business knowledge for departments and functions, thus improving business efficiency and customer service. DW can present a competitive advantage by facilitating decision making and helping reform business processes.

DW enables a consolidated view of corporate data, all cleaned and organized. Thus, the entire organization can see an integrated view of itself. DW thus provides better and timely information. It simplifies data access and allows end users to perform extensive analysis. It enhances overall IT performance by not burdening the operational databases used by Enterprise Resource Planning (ERP) and other systems.

Caselet: University Health System—BI in Health Care

Indiana University Health (IUH), a large academic health care system, decided to build an enterprise data warehouse (EDW) to foster a genuinely data-driven management culture. IUH hired a DW vendor to develop EDW, which also integrates with their electronic health record (EHR) *system. They loaded 14 billion rows of data into EDW—fully 10 years of clinical data from across IUH's network. Clinical events, patient encounters, lab and radiology, and other patient data were included, as were IUH's performance management, revenue cycle, and patient satisfaction data. They soon put in a new interactive dashboard using the EDW that provided IUH's leadership with the daily operational insights they need to solve the quality/cost equation. It offers visibility into key operational metrics and trends to easily track the performance measures critical to controlling costs and maintaining quality. EDW can easily be used across IUH's departments to analyze, track, and measure clinical, financial, and patient experience outcomes. (Source: healthcatalyst.com)*

Q1. *What are the benefits of a single large comprehensive EDW?*
Q1. *What kinds of data would be needed for EDW for an airline company?*

Design Considerations for DW

The objective of DW is to provide business knowledge to support decision making. For DW to serve its objective, it should be aligned around those decisions. It should be comprehensive, easy to access, and up-to-date. Here are some requirements for a good DW:

1. *Subject-oriented:* To be effective, DW should be designed around a subject domain, that is, to help solve a certain category of problems.
2. *Integrated:* DW should include data from many functions that can shed light on a particular subject area. Thus, the organization can benefit from a comprehensive view of the subject area.
3. *Time-variant (time series):* The data in DW should grow at daily or other chosen intervals. That allows latest comparisons over time.
4. *Nonvolatile:* DW should be persistent, that is, it should not be created on the fly from the operations databases. Thus, DW is consistently available for analysis, across the organization and over time.
5. *Summarized:* DW contains rolled-up data at the right level for queries and analysis. The rolling up helps create consistent granularity for effective comparisons. It helps reduces the number of variables or dimensions of the data to make them more meaningful for the decision makers.

6. *Not normalized:* DW often uses a star schema, which is a rectangular central table, surrounded by some lookup tables. The single-table view significantly enhances speed of queries.
7. *Metadata:* Many of the variables in the database are computed from other variables in the operational database. For example, total daily sales may be a computed field. The method of its calculation for each variable should be effectively documented. Every element in DW should be sufficiently well-defined.
8. *Near real-time and/or right-time (active):* DWs should be updated in near real-time in many high-transaction volume industries, such as airlines. The cost of implementing and updating DW in real time could discourage others. Another downside of real-time DW is the possibilities of inconsistencies in reports drawn just a few minutes apart.

DW Development Approaches

There are two fundamentally different approaches to developing DW: top down and bottom up. The top-down approach is to make a comprehensive DW that covers all the reporting needs of the enterprise. The bottom-up approach is to produce small data marts, for the reporting needs of different departments or functions, as needed. The smaller data marts will eventually align to deliver comprehensive EDW capabilities. The top-down approach provides consistency but takes time and resources. The bottom-up approach leads to healthy local ownership and maintainability of data (Table 3.1).

Table 3.1 Comparing data mart and data warehouse

	Functional Data Mart	Enterprise Data Warehouse
Scope	One subject or functional area	Complete enterprise data needs
Value	Functional area reporting and insights	Deeper insights connecting multiple functional areas
Target organization	Decentralized management	Centralized management
Time	Low to medium	High
Cost	Low	High
Size	Small to medium	Medium to large
Approach	Bottom up	Top down
Complexity	Low (fewer data transformations)	High (data standardization)
Technology	Smaller scale servers and databases	Industrial strength

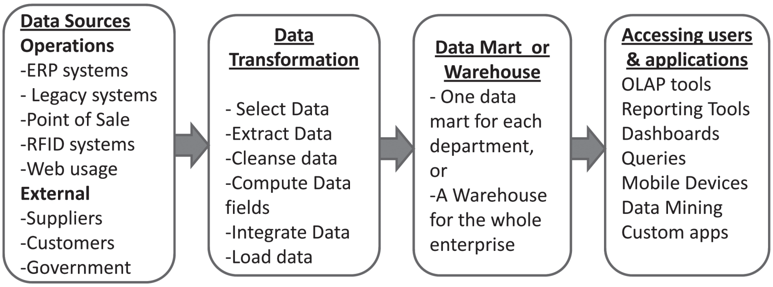


Figure 3.1 Data warehousing architecture

DW Architecture

DW has four key elements (Figure 3.1). The first element is the data sources that provide the raw data. The second element is the process of transforming that data to meet the decision needs. The third element is the methods of regularly and accurately loading of that data into EDW or data marts. The fourth element is the data access and analysis part, where devices and applications use the data from DW to deliver insights and other benefits to users.

Data Sources

DWs are created from structured data sources. Unstructured data, such as text data, would need to be structured before inserted into DW.

1. Operations data include data from all business applications, including from ERPs systems that form the backbone of an organization's IT systems. The data to be extracted will depend upon the subject matter of DW. For example, for a sales/marketing DW, only the data about customers, orders, customer service, and so on would be extracted.
2. Other applications, such as point-of-sale (POS) terminals and e-commerce applications, provide customer-facing data. Supplier data could come from supply chain management systems. Planning and budget data should also be added as needed for making comparisons against targets.
3. External syndicated data, such as weather or economic activity data, could also be added to DW, as needed, to provide good contextual information to decision makers.

Data Transformation Processes

The heart of a useful DW is the processes to populate the DW with good quality data. This is called the extract-transform-load (ETL) cycle.

1. Data should be extracted from many operational (transactional) database sources on a regular basis.
2. Extracted data should be aligned together by key fields. It should be cleansed of any irregularities or missing values. It should be rolled up together to the same level of granularity. Desired fields, such as daily sales totals, should be computed. The entire data should then be brought to the same format as the central table of DW.
3. The transformed data should then be uploaded into DW.

This ETL process should be run at a regular frequency. Daily transaction data can be extracted from ERPs, transformed, and uploaded to the database the same night. Thus, DW is up-to-date next morning. If DW is needed for near-real-time information access, then the ETL processes would need to be executed more frequently. ETL work is usually automated using programing scripts that are written, tested, and then deployed for periodic updating DW.

DW Design

Star schema is the preferred data architecture for most DWs. There is a central fact table that provides most of the information of interest. There are lookup tables that provide detailed values for codes used in the central table. For example, the central table may use digits to represent a sales person. The lookup table will help provide the name for that sales person code. Here is an example of a star schema for a data mart for monitoring sales performance (Figure 3.2).

Other schemas include the snowflake architecture. The difference between a star and snowflake is that in the latter, the lookup tables can have their own further lookup tables.

There are many technology choices for developing DW. This includes selecting the right database management system and the right set of data management tools. There are a few big and reliable providers of DW systems. The provider of the operational DBMS may be chosen for DW also.

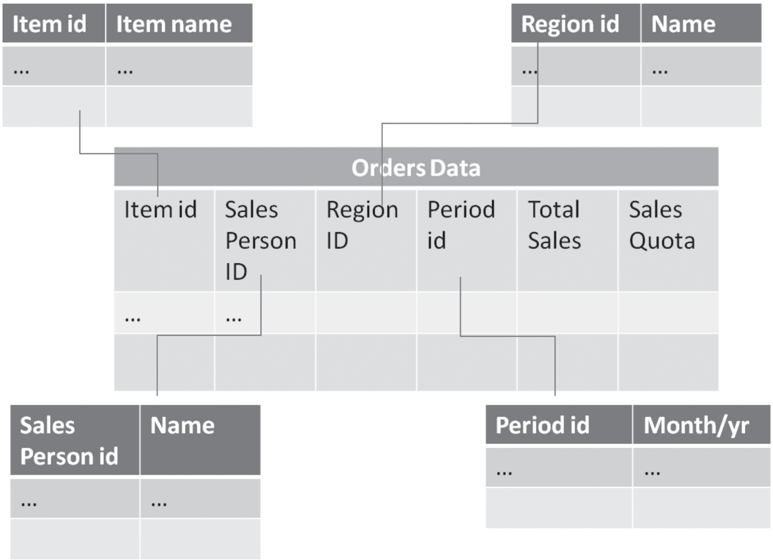


Figure 3.2 Star schema architecture

Alternatively, a best-of-breed DW vendor could be used. There are also a variety of tools out there for data migration, data upload, data retrieval, and data analysis.

DW Access

Data from DW could be accessed for many purposes, through many devices.

1. A primary use of DW is to produce routine management and monitoring reports. For example, a sales performance report would show sales by many dimensions, and compared with plan. A dashboarding system will use data from the warehouse and present analysis to users. The data from DW can be used to populate customized performance dashboards for executives. The dashboard could include drill-down capabilities to analyze the performance data for root cause analysis.
2. The data from the warehouse could be used for ad hoc queries and any other applications that make use of the internal data.

3. Data from DW is used to provide data for mining purposes. Parts of the data would be extracted, and then combined with other relevant data, for data mining.

DW Best Practices

A DW project reflects a significant investment into IT. All of the best practices in implementing any IT project should be followed.

1. The DW project should align with the corporate strategy. Top management should be consulted for setting objectives. Financial viability Return on Investment (ROI) should be established. The project must be managed by both IT and business professionals. The DW design should be carefully tested before beginning development work. It is often much more expensive to redesign after development work has begun.
2. It is important to manage user expectations. DW should be built incrementally. Users should be trained in using the system, and absorb the many features of the system.
3. Quality and adaptability should be built in from the start. Only cleansed and high-quality data should be loaded. The system should be able to adapt to new access tools. As business needs change, new data marts can be created for new needs.

Conclusion

DWs are special data management facilities intended for creating reports and analysis to support managerial decision making. They are designed to make reporting and querying simple and efficient. The sources of data are operational systems and external data sources. DW needs to be updated with new data regularly to keep it useful. Data from DW provides a useful input for data mining activities.

Review Questions

1. What is the purpose of a data warehouse?
2. What are the key elements of a data warehouse? Describe each.

3. What are the sources and types of data for a data warehouse?
4. How will data warehousing evolve in the age of social media?

Liberty Stores Case Exercise: Step 2

The Liberty Stores company wants to be fully informed about its sales of products and take advantage of growth opportunities as they arise. It wants to analyze sales of all its products by all store locations. The newly hired chief knowledge officer has decided to build a data warehouse.

1. *Design a DW structure for the company to monitor its sales performance. (Hint: Design the central table and lookup tables.)*
2. *Design another DW for the company's sustainability and charitable activities.*

CHAPTER 4

Data Mining

Data mining is the art and science of discovering knowledge, insights, and patterns in data. It is the act of extracting useful patterns from an organized collection of data. Patterns must be valid, novel, potentially useful, and understandable. The implicit assumption is that data about the past can reveal patterns of activity that can be projected into the future.

Data mining is a multidisciplinary field that borrows techniques from a variety of fields. It utilizes the knowledge of data quality and data organizing from the databases area. It draws modeling and analytical techniques from statistics and computer science (artificial intelligence) areas. It also draws the knowledge of decision-making from the field of business management.

The field of data mining emerged in the context of pattern recognition in defense, such as identifying a friend-or-foe on a battlefield. Like many other defense-inspired technologies, it has evolved to help gain a competitive advantage in business.

For example, "customers who buy *cheese* and *milk* also buy *bread* 90 percent of the time" would be a useful pattern for a grocery store, which can then stock the products appropriately. Similarly, "people with blood pressure greater than 160 and an age greater than 65 were at a high risk of dying from a heart stroke" is of great diagnostic value for doctors, who can then focus on treating such patients with urgent care and great sensitivity.

Past data can be of predictive value in many complex situations, especially where the pattern may not be so easily visible without the modeling technique. Here is a dramatic case of a data-driven decision-making system that beats the best of human experts. Using past data, a decision tree model was developed to predict votes for Justice Sandra Day O'Connor, who had a swing vote in a 5–4 divided US Supreme Court. All her previous decisions were coded on a few variables. What emerged from data mining was a simple four-step decision tree that was able to accurately predict her votes

71 percent of the time. In contrast, the legal analysts could at best predict correctly 59 percent of the time. *(Source: Martin et al. 2004)*

Caselet: Target Corp—Data Mining in Retail

Target is a large retail chain that crunches data to develop insights that help target marketing and advertising campaigns. Target analysts managed to develop a pregnancy-prediction score based on a customer's purchasing history of 25 products. In a widely publicized story, they figured out that a teenage girl was pregnant before her father did. The targeting can be quite successful and dramatic as this example published in the New York Times illustrates as follows:

About a year after Target created their pregnancy-prediction model, a man walked into a Target store and demanded to see the manager. He was clutching coupons that had been sent to his daughter and he was angry, according to an employee who participated in the conversation. "My daughter got this in the mail!" he said. "She's still in high school, and you're sending her coupons for baby clothes and cribs? Are you trying to encourage her to get pregnant?"

The manager didn't have any idea what the man was talking about. He looked at the mailer. Sure enough, it was addressed to the man's daughter and contained advertisements for maternity clothing, nursery furniture and pictures of smiling infants. The manager apologized and then called a few days later to apologize again.

On the phone, though, the father was somewhat subdued. "I had a talk with my daughter," he said. "It turns out there's been some activities in my house I haven't been completely aware of. I owe you an apology." (Source: New York Times)

Q1. *Do Target and other retailers have full rights to use their acquired data as it sees fit, and to contact desired consumers with all legally admissible means and messages? What are the issues involved here?*

Q2. *FaceBook and Google provide many services for free. In return they mine our email and blogs and send us targeted ads. Is that a fair deal?*

Gathering and Selecting Data

The total amount of data in the world is doubling every 18 months. There is an ever-growing avalanche of data coming with higher velocity, volume, and variety. One has to quickly use it or lose it. Smart data mining requires choosing where to play. One has to make judicious decisions about what to gather and what to ignore, based on the purpose of the data mining exercises. It is like deciding where to fish; not all streams of data will be equally rich in potential insights.

To learn from data, one needs to effectively gather quality data, clean and organize it, and then efficiently process it. One requires the skills and technologies for consolidation and integration of data elements from many sources. Most organizations develop an enterprise data model (EDM), which is a unified, high-level model of all the data stored in an organization's databases. The EDM will be inclusive of the data generated from all internal systems. The EDM provides the basic menu of data to create a data warehouse for a particular decision-making purpose. Data warehouses help organize all this data in a useful manner so that it can be selected and deployed for mining. The EDM can also help imagine what relevant external data should be gathered to develop good predictive relationships with the internal data. In the United States, the governments and their agencies make a vast variety and quantity of data available at data.gov.

Gathering and curating data takes time and effort, particularly when it is unstructured or semistructured. Unstructured data can come in many forms like databases, blogs, images, videos, and chats. There are streams of unstructured social media data from blogs, chats, and tweets. There are also streams of machine-generated data from connected machines, RFID tags, the internet of things, and so on. The data should be put in rectangular data shapes with clear columns and rows before submitting it to data mining.

Knowledge of the business domain helps select the right streams of data for pursuing new insights. Data that suits the nature of the problem being solved should be gathered. The data elements should be relevant, and suitably address the problem being solved. They could directly impact the problem, or they could be a suitable proxy for the effect being measured. Select data will also be gathered from the data warehouse.

Industries and functions will have their own requirements and constraints. The health care industry will provide a different type of data with different data names. The HR function would provide different kinds of data. There would be different issues of quality and privacy for these data.

Data Cleansing and Preparation

The quality of data is critical to the success and value of the data mining project. Otherwise, the situation will be of the kind of garbage in and garbage out (GIGO). The quality of incoming data varies by the source and nature of data. Data from internal operations is likely to be of higher quality, as it will be accurate and consistent. Data from social media and other public sources is less under the control of business, and is less likely to be reliable.

Data almost certainly needs to be cleansed and transformed before it can be used for data mining. There are many ways in what data may need to be cleansed—filling missing values, reigning in the effects of outliers, transforming fields, binning continuous variables, and much more—before it can be ready for analysis. Data cleansing and preparation is a labor-intensive or semiautomated activity that can take up to 60 to 70 percent of the time needed for a data mining project.

1. Duplicate data needs to be removed. The same data may be received from multiple sources. When merging the data sets, data must be de-duped.
2. Missing values need to be filled in, or those rows should be removed from analysis. Missing values can be filled in with average or modal or default values.
3. Data elements may need to be transformed from one unit to another. For example, total costs of health care and the total number of patients may need to be reduced to cost/patient to allow comparability of that value.
4. Continuous values may need to be binned into a few buckets to help with some analyses. For example, work experience could be binned as low, medium, and high.
5. Data elements may need to be adjusted to make them comparable over time. For example, currency values may need to be adjusted

for inflation; they would need to be converted to the same base year for comparability. They may need to be converted to a common currency.

6. Outlier data elements need to be removed after careful review, to avoid the skewing of results. For example, one big donor could skew the analysis of alumni donors in an educational setting.
7. Any biases in the selection of data should be corrected to ensure the data is representative of the phenomena under analysis. If the data includes many more members of one gender than is typical of the population of interest, then adjustments need to be applied to the data.
8. Data should be brought to the same granularity to ensure comparability. Sales data may be available daily, but the sales person compensation data may only be available monthly. To relate these variables, the data must be brought to the lowest common denominator, in this case, monthly.
9. Data may need to be selected to increase information density. Some data may not show much variability, because it was not properly recorded or for any other reasons. This data may dull the effects of other differences in the data and should be removed to improve the information density of the data.

Outputs of Data Mining

Data mining techniques can serve different types of objectives. The outputs of data mining will reflect the objective being served. There are many representations of the outputs of data mining.

One popular form of data mining output is a decision tree. It is a hierarchically branched structure that helps visually follow the steps to make a model-based decision. The tree may have certain attributes, such as probabilities assigned to each branch. A related format is a set of business rules, which are if-then statements that show causality. A decision tree can be mapped to business rules. If the objective function is prediction, then a decision tree or business rules are the most appropriate mode of representing the output.

The output can be in the form of a regression equation or mathematical function that represents the best fitting curve to represent the data. This equation may include linear and nonlinear terms. Regression equations are a good way of representing the output of classification exercises. These are also a good representation of forecasting formulae.

Population "centroid" is a statistical measure for describing central tendencies of a collection of data points. These might be defined in a multidimensional space. For example, a centroid could be "middle-aged, highly educated, high-net worth professionals, married with two children, living in the coastal areas". Or a population of "20-something, ivy-league-educated, tech entrepreneurs based in Silicon Valley". Or a collection of "vehicles more than 20 years old, giving low mileage per gallon, which failed the environmental inspection". These are typical representations of the output of a cluster analysis exercise.

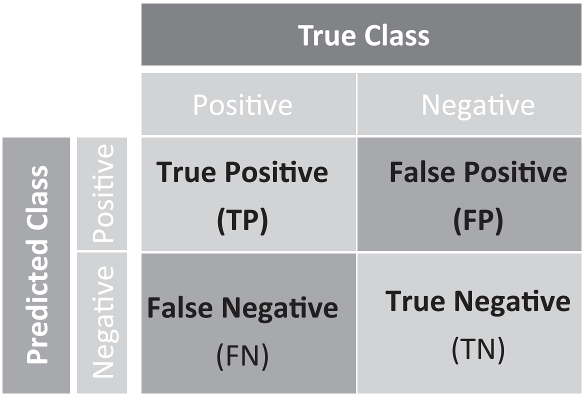
Business rules are an appropriate representation of the output of a market basket analysis exercise. These rules are if-then statements with some probability parameters associated with each rule. For example, those that buy milk and bread will also buy butter (with 80 percent probability).

Evaluating Data Mining Results

There are two primary kinds of data mining processes: supervised learning and unsupervised learning. In supervised learning, a decision model can be created using past data, and the model can then be used to predict the correct answer for future data instances. Classification is the main category of supervised learning activity. There are many techniques for classification, decision trees being the most popular one. Each of these techniques can be implemented with many algorithms. A common metric for all of classification techniques is predictive accuracy.

Predictive Accuracy = (Correct Predictions) / Total Predictions

Suppose a data mining project has been initiated to develop a predictive model for cancer patients using a decision tree. Using a relevant set of variables and data instances, a decision tree model has been created. The model is then used to predict other data instances. When a true positive data point is positive, that is a correct prediction, called a true positive (TP). Similarly, when a true negative data point is classified as negative, that is a true negative (TN). On the other hand, when a true-positive data



		True Class	
		Positive	Negative
Predicted Class	Positive	**True Positive (TP)**	**False Positive (FP)**
	Negative	**False Negative** (FN)	**True Negative** (TN)

Figure 4.1 Confusion matrix

point is classified by the model as negative, that is an incorrect prediction, called a false negative (FN). Similarly, when a true-negative data point is classified as positive, that is classified as a false positive (FP). This is called the confusion matrix (Figure 4.1).

Thus, the predictive accuracy can be specified by the following formula.

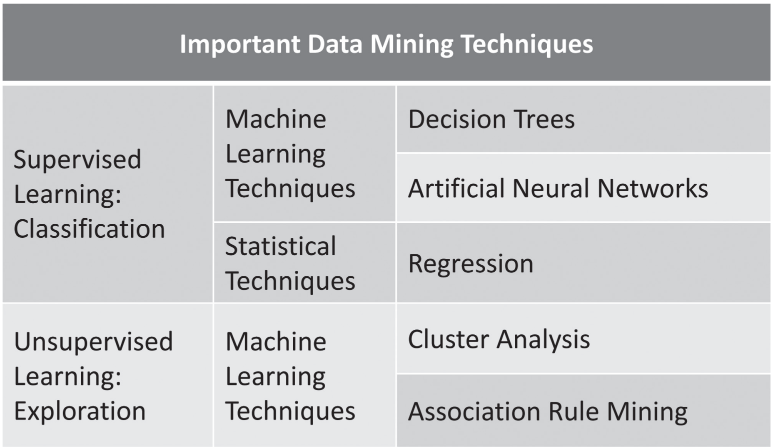
Predictive Accuracy = (TP + TN) / (TP + TN + FP + FN).

All classification techniques have a predictive accuracy associated with a predictive model. The highest value can be 100 percent. In practice, predictive models with more than 70 percent accuracy can be considered usable in business domains, depending upon the nature of the business.

There are no good objective measures to judge the accuracy of unsupervised learning techniques, such as cluster analysis. There is no single right answer for the results of these techniques. The value of the segmentation model depends upon the value the decision maker sees in those results.

Data Mining Techniques

Data may be mined to help make more efficient decisions in the future. Or it may be used to explore the data to find interesting associative patterns. The right technique depends upon the kind of problem being solved (Figure 4.2).



Important Data Mining Techniques		
Supervised Learning: Classification	Machine Learning Techniques	Decision Trees
		Artificial Neural Networks
	Statistical Techniques	Regression
Unsupervised Learning: Exploration	Machine Learning Techniques	Cluster Analysis
		Association Rule Mining

Figure 4.2 Important data mining techniques

The most important class of problems solved using data mining are classification problems. These are problems where data from past decisions is mined to extract the few rules and patterns that would improve the accuracy of the decision-making process in the future. The data of past decisions is organized and mined for decision rules or equations, which are then codified to produce more accurate decisions. Classification techniques are called supervised learning as there is a way to supervise whether the model's prediction is right or wrong.

A decision tree is a hierarchically organized branched, structured to help make decision in an easy and logical manner. *Decision trees* are the most popular data mining technique, for many reasons.

1. Decision trees are easy to understand and easy to use, by analysts as well as executives. They also show a high predictive accuracy.
2. They select the most relevant variables automatically out of all the available variables for decision-making.
3. Decision trees are tolerant of data quality issues and do not require much data preparation from the users.
4. Even nonlinear relationships can be handled well by decision trees.

There are many algorithms to implement decision trees. Some of the popular ones are C5, CART, and CHAID.

Regression is a relatively simple and the most popular statistical data mining technique. The goal is to fit a smooth well-defined curve to the data. Regression analysis techniques, for example, can be used to model and predict the energy consumption as a function of daily temperature. Simply plotting the data shows a nonlinear curve. Applying a nonlinear regression equation will fit the data very well with high accuracy. Thus, the energy consumption on any future day can be predicted using this equation.

Artificial neural network (ANN) is a sophisticated data mining technique from the Artificial Intelligence stream in Computer Science. It mimics the behavior of human neural structure: Neurons receive stimuli, process them, and communicate their results to other neurons successively, and eventually a neuron outputs a decision. A decision task may be processed by just one neuron and the result may be communicated soon. Alternatively, there could be many layers of neurons involved in a decision task, depending upon the complexity of the domain. The neural network can be trained by making a decision over and over again with many data points. It will continue to learn by adjusting its internal computation and communication parameters based on feedback received on its previous decisions. The intermediate values passed within the layers of neurons may not make intuitive sense to an observer. Thus, the neural networks are considered a black-box system.

At some point, the neural network will have learned enough and begin to match the predictive accuracy of a human expert or alternative classification techniques. The predictions of some ANNs that have been trained over a long period of time with a large amount of data have become decisively more accurate than human experts. At that point, the ANNs can begin to be seriously considered for deployment, in real situations in real time.

ANNs are popular because they are eventually able to reach a high predictive accuracy. ANNs are also relatively simple to implement and do not have any issues with data quality. ANNs require a lot of data to train to develop good predictive ability.

Cluster analysis is an exploratory learning technique that helps in identifying a set of similar groups in the data. It is a technique used for automatic identification of natural groupings of things. Data instances that are similar to (or near) each other are categorized into one cluster, while data instances that are very different (or far away) from each other are categorized into separate clusters. There can be any number of clusters that could be produced by the data. The K-means technique is a popular technique and allows the user guidance in selecting the right number (K) of clusters from the data.

Clustering is also known as the segmentation technique. The technique shows the clusters of things from past data. The output is the centroids for each cluster and the allocation of data points to their cluster. The centroid definition is used to assign new data instances that can be assigned to their cluster homes. Clustering is also a part of the artificial intelligence family of techniques.

Association rules are a popular data mining method in business, especially where selling is involved. Also known as market basket analysis, it helps in answering questions about cross-selling opportunities. This is the heart of the personalization engine used by e-commerce sites like Amazon.com and streaming movie sites like Netflix.com. The technique helps find interesting relationships (affinities) between variables (items or events). These are represented as rules of the form $X \Rightarrow Y$, where X and Y are sets of data items. A form of unsupervised learning, it has no dependent variable; and there are no right or wrong answers. There are just stronger and weaker affinities. Thus, each rule has a confidence level assigned to it. A part of the machine-learning family, this technique achieved legendary status when a fascinating relationship was found in the sales of diapers and beers.

Tools and Platforms for Data Mining

Data mining tools have existed for many decades. However, they have recently become more important as the values of data have grown and the field of big data analytics has come into prominence. There are a wide range of data mining platforms available in the market today.

1. There are simple end-user data mining tools, such as MS Excel, and there are more sophisticated tools, such as IBM SPSS Modeler.
2. There are stand-alone tools, and there are tools embedded in an existing transaction processing or data warehousing or ERP system.
3. There are open-source and freely available tools, such as Weka, and there are commercial products.

4. There are text-based tools that require some programing skills, and there are Graphical User Interface (GUI)-based drag-and-drop format tools.
5. There are tools that work only on proprietary data formats, and there are those directly accept data from a host of popular data management tools formats.

Here, we compare three platforms that we have used extensively and effectively for many data mining projects (Table 4.1).

MS Excel is a relatively simple and easy data mining tool. It can get quite versatile once analyst pack and some other add-on products are installed on it.

IBM's SPSS Modeler is an industry-leading data mining platform. It offers a powerful set of tools and algorithms for most popular data mining capabilities. It has colorful GUI format with drag-and-drop capabilities. It can accept data in multiple formats, including reading Excel files directly.

Weka is an open-source GUI-based tool that offers a large number of data mining algorithms.

ERP systems include some data analytic capabilities, too. SAP has its Business Objects BI software. Business Objects is considered one of the leading BI suites in the industry and is often used by organizations that use SAP.

Table 4.1 Comparison of popular data mining platforms

Feature	Excel	IBM SPSS Modeler	Weka
Ownership	Commercial	Commercial, expensive	Open-source, free
Data mining features	Limited, extensible with add-on modules	Extensive features, unlimited data sizes	Extensive, performance issues with large data
Stand-alone	Stand-alone	Embedded in BI software suites	Stand-alone
User skills needed	End users	Skilled BI analysts	Skilled BI analysts
User interface	Select and click, easy	Drag-and-drop use, colorful, beautiful GUI	GUI, mostly b&w text output
Data formats	Industry standard	Variety of data sources accepted	Proprietary

Data Mining Best Practices

Effective and successful use of data mining activity requires both business and technology skills. The business aspects help understand the domain and the key questions. It also helps one imagine possible relationships in the data and create hypotheses to test it. The IT aspects help fetch the data from many sources, clean up the data, assemble it to meet the needs of the business problem, and then run the data mining techniques on the platform.

An important element is to go after the problem iteratively. It is better to divide and conquer the problem with smaller amounts of data, and get closer to the heart of the solution in an iterative sequence of steps. There are several best practices learned from the use of data mining techniques over a long period of time. The data mining industry has proposed a Cross-Industry Standard Process for Data Mining (CRISP-DM). It has six essential steps (Figure 4.3):

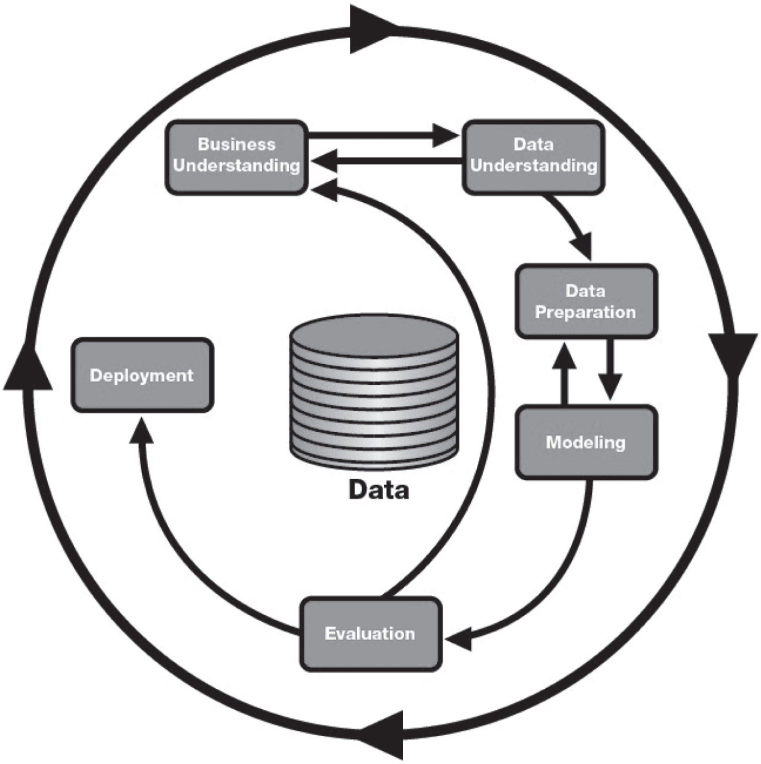


Figure 4.3 CRISP-DM data mining cycle

1. The first and most important step in data mining is business understanding, that is, asking the right business questions. A question is a good one if answering it would lead to large payoffs for the organization, financially and otherwise. In other words, selecting a data mining project is like any other project, in which it should show strong payoffs if the project is successful. There should be strong executive support for the data mining project, which means that the project aligns well with the business strategy.
2. A second important step is to be creative and open in proposing imaginative hypotheses for the solution. Thinking outside the box is important, both in terms of a proposed model as well in the data sets available and required.
3. The data should be clean and of high quality. It is important to assemble a team that has a mix of technical and business skills, who understand the domain and the data. Data cleaning can take 60 to 70 percent of the time in a data mining project. It may be desirable to add new data elements from external sources of data that could help improve predictive accuracy.
4. Patience is required in continuously engaging with the data until the data yields some good insights. A host of modeling tools and algorithms should be used. A tool could be tried with different options, such as running different decision tree algorithms.
5. One should not accept what the data says at first. It is better to triangulate the analysis by applying multiple data mining techniques and conducting many what-if scenarios, to build confidence in the solution. Evaluate the model's predictive accuracy with more test data.
6. The dissemination and rollout of the solution is the key to project success. Otherwise the project will be a waste of time and will be a setback for establishing and supporting a data-based decision-process culture in the organization. The model should be embedded in the organization's business processes.

Myths about Data Mining

There are many myths about this area, scaring away many business executives from using data mining.

Myth #1: Data mining is about algorithms: Data mining is used by business to answer important and practical business questions. Formulating the problem statement correctly and identifying imaginative solutions for testing are far more important before the data mining algorithms get called in.

Myth #2: Data mining is about predictive accuracy: While important, predictive accuracy is a feature of the algorithm. As in myth #1, the quality of output is a strong function of the right problem, right hypothesis, and the right data.

Myth #3: Data mining requires a data warehouse: While the presence of a data warehouse assists in the gathering of information, sometimes the creation of the data warehouse itself can benefit from some exploratory data mining.

Myth #4: Data mining requires large quantities of data: Many interesting data mining exercises are done using small- or medium-sized data sets.

Myth #5: Data mining requires a technology expert: Many interesting data mining exercises are done by end users and executives using simple everyday tools like spreadsheets.

Data Mining Mistakes

Data mining is an exercise in extracting nontrivial useful patterns in the data. It requires a lot of preparation and patience to pursue the many leads that data may provide. Much domain knowledge, tools, and skill are required to find such patterns. Here are some of the more common mistakes in doing data mining, and should be avoided.

Mistake #1: Selecting the wrong problem for data mining: Without the right goals or having no goals, data mining leads to a waste of time. Getting the right answer to an irrelevant question could be interesting, but it would be pointless.

Mistake #2: Buried under mountains of data without clear metadata: It is more important to be engaged with the data, than to have lots of data. The relevant data required may be much less than initially thought. There may be insufficient knowledge about the data or metadata.

Mistake #3: Disorganized data mining: Without clear goals, much time is wasted. Doing the same tests using the same mining algorithms repeatedly and blindly, without thinking about the next stage, without a plan, would lead to wasted time and energy. This can come from being sloppy about keeping track of the data mining procedure and results.

Mistake #4: Insufficient business knowledge: Without a deep understanding of the business domain, the results would be gibberish and meaningless. Do not make erroneous assumptions, courtesy of experts. Do not rule out anything when observing data analysis results. Do not ignore suspicious (good or bad) findings and quickly move on. Be open to surprises. Even when insights emerge at one level, it is important to sliced and dice the data at other levels to see if more powerful insights can be extracted.

Mistake #5: Incompatibility of data mining tools: All the tools from data gathering, preparation, mining, and visualization should work together.

Mistake #6: Locked in the data jailhouse: Use tools that can work with data from multiple sources in multiple industry standard formats.

Mistake #7: Looking only at aggregated results and not at individual records/predictions. It is possible that the right results at the aggregate level provide absurd conclusions at an individual record level.

Mistake #8: Running out of time: Not leaving sufficient time for data acquisition, selection, and preparation can lead to data quality issues and GIGO. Similarly not providing enough time for testing the model, training the users and deploying the system can make the project a failure.

Mistake #9: Measuring your results differently from the way your sponsor measures them: This comes from losing a sense of business objectives and beginning to mine data for its own sake.

Mistake #10: Naively believing everything you are told about the data: Also naively believing everything you are told about your own data mining analysis.

Conclusion

Data mining is like diving into the rough material to discover a valuable finished nugget. While the technique is important, domain knowledge is also important to provide imaginative solutions that can then be tested with data mining. The business objective should be well understood and should always be kept in mind to ensure that the results are beneficial to the sponsor of the exercise.

Review Questions

1. What is data mining? What are supervised and unsupervised learning techniques?
2. Describe the key steps in the data mining process. Why is it important to follow these processes?
3. What is a confusion matrix?
4. Why is data preparation so important and time consuming?
5. What are some of the most popular data mining techniques?
6. What are the major mistakes to be avoided when doing data mining?
7. What are the key requirements for a skilled data analyst?

Liberty Stores Case Exercise: Step 3

Liberty is constantly evaluating opportunities for improving efficiencies in all its operations, including the commercial operations as well its charitable activities.

1. *What data mining techniques would you use to analyze and predict sales patterns?*
2. *What data mining technique would you use to categorize its customers?*