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).

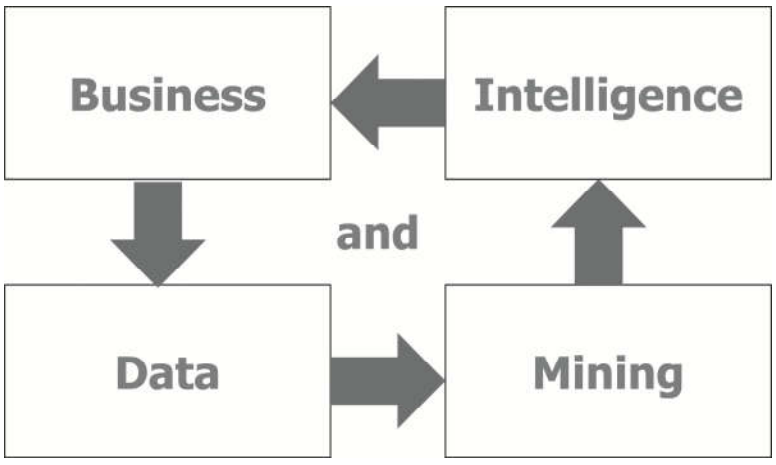

Figure 1.1 Business intelligence and data mining cycle

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).

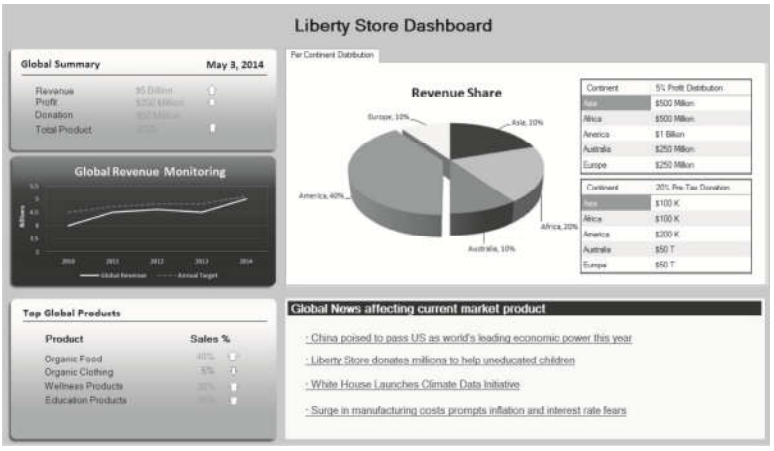

Figure 1.3 Sample executive dashboard

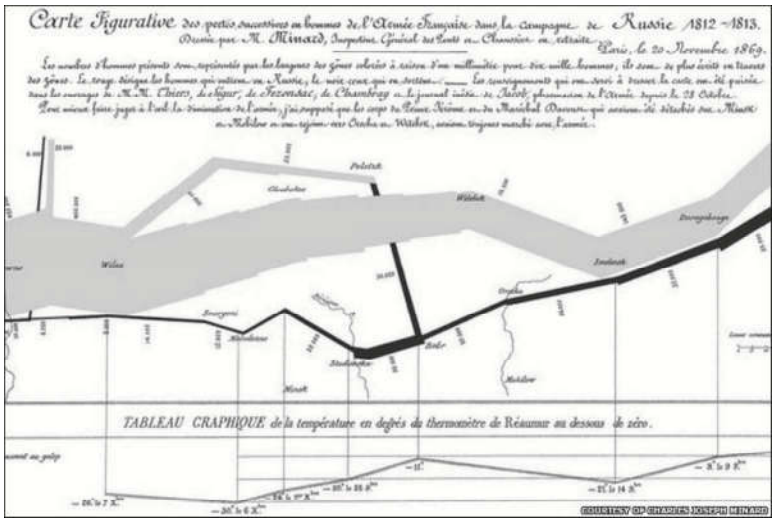

Figure 1.4 Sample data visualization

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.