CHAPTER TWO:
ORGANIZATIONAL UNDERSTANDING AND DATA UNDERSTANDING

CONTEXT AND PERSPECTIVE

Consider some of the activities you've been involved with in the past three or four days. Have you purchased groceries or gasoline? Attended a concert, movie or other public event? Perhaps you went out to eat at a restaurant, stopped by your local post office to mail a package, made a purchase online, or placed a phone call to a utility company. Every day, our lives are filled with interactions – encounters with companies, other individuals, the government, and various other organizations.

In today's technology-driven society, many of those encounters involve the transfer of information electronically. That information is recorded and passed across networks in order to complete financial transactions, reassign ownership or responsibility, and enable delivery of goods and services. Think about the amount of data collected each time even one of these activities occurs.

Take the grocery store for example. If you take items off the shelf, those items will have to be replenished for future shoppers – perhaps even for yourself – after all you'll need to make similar purchases again when that case of cereal runs out in a few weeks. The grocery store must constantly replenish its supply of inventory, keeping the items people want in stock while maintaining freshness in the products they sell. It makes sense that large databases are running behind the scenes, recording data about what you bought and how much of it, as you check out and pay your grocery bill. All of that data must be recorded and then reported to someone whose job it is to reorder items for the store's inventory.

However, in the world of data mining, simply keeping inventory up-to-date is only the beginning. Does your grocery store require you to carry a frequent shopper card or similar device which, when scanned at checkout time, gives you the best price on each item you're buying? If so, they

can now begin not only keep track of store-wide purchasing trends, but individual purchasing trends as well. The store can target market to you by sending mailers with coupons for products you tend to purchase most frequently.

Now let's take it one step further. Remember, if you can, what types of information you provided when you filled out the form to receive your frequent shopper card. You probably indicated your address, date of birth (or at least birth year), whether you're male or female, and perhaps the size of your family, annual household income range, or other such information. Think about the range of possibilities now open to your grocery store as they analyze that vast amount of data they collect at the cash register each day:

- Using ZIP codes, the store can locate the areas of greatest customer density, perhaps aiding their decision about the construction location for their next store.
- Using information regarding customer gender, the store may be able to tailor marketing displays or promotions to the preferences of male or female customers.
- With age information, the store can avoid mailing coupons for baby food to elderly customers, or promotions for feminine hygiene products to households with a single male occupant.

These are only a few the many examples of potential uses for data mining. Perhaps as you read through this introduction, some other potential uses for data mining came to your mind. You may have also wondered how ethical some of these applications might be. This text has been designed to help you understand not only the possibilities brought about through data mining, but also the techniques involved in making those possibilities a reality while accepting the responsibility that accompanies the collection and use of such vast amounts of personal information.

LEARNING OBJECTIVES

After completing the reading and exercises in this chapter, you should be able to:

- Define the discipline of Data Mining
- List and define various types of data
- List and define various sources of data
- Explain the fundamental differences between databases, data warehouses and data sets

- Explain some of the ethical dilemmas associated with data mining and outline possible solutions

PURPOSES, INTENTS AND LIMITATIONS OF DATA MINING

Data mining, as explained in Chapter 1 of this text, applies statistical and logical methods to large data sets. These methods can be used to *categorize* the data, or they can be used to create *predictive models*. Categorizations of large sets may include grouping people into similar types of classifications, or in identifying similar characteristics across a large number of observations.

Predictive models however, transform these descriptions into expectations upon which we can base decisions. For example, the owner of a book-selling Web site could project how frequently she may need to restock her supply of a given title, or the owner of a ski resort may attempt to predict the earliest possible opening date based on projected snow arrivals and accumulations.

It is important to recognize that data mining cannot provide answers to every question, nor can we expect that predictive models will always yield results which will in fact turn out to be the reality. Data mining is limited to the data that has been collected. And those limitations may be many. We must remember that the data may not be completely representative of the group of individuals to which we would like to apply our results. The data may have been collected incorrectly, or it may be out-of-date. There is an expression which can adequately be applied to data mining, among many other things: *GIGO,* or *Garbage In, Garbage Out.* The quality of our data mining results will directly depend upon the quality of our data collection and organization. Even after doing our very best to collect high quality data, we must still remember to base decisions not only on data mining results, but also on available resources, acceptable amounts of risk, and plain old common sense.

DATABASE, DATA WAREHOUSE, DATA MART, DATA SET…?

In order to understand data mining, it is important to understand the nature of databases, data collection and data organization. This is fundamental to the discipline of Data Mining, and will directly impact the quality and reliability of all data mining activities. In this section, we will

examine the differences between **databases**, **data warehouses**, and **data sets**. We will also examine some of the variations in terminology used to describe data attributes.

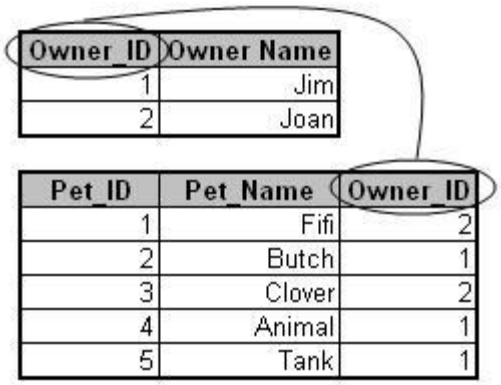
Although we will be examining the differences between databases, data warehouses and data sets, we will begin by discussing what they have in common. In Figure 2-1, we see some data organized into **rows** (shown here as A, B, etc.) and **columns** (shown here as 1, 2, etc.). In varying data environments, these may be referred to by differing names. In a database, rows would be referred to as **tuples** or **records**, while the columns would be referred to as **fields**.

	A	B	C	D
1	3989.408	3989.408	140.4029	2654.278
2	140.4029	4125.044	4125.044	1335.467
3	2654.278	1335.467	2789.76	2789.76
4	5777.168	1788.068	5912.553	3123.153
5	2050.529	6039.689	1915.155	4704.363
6	1435.265	2554.287	1571.295	1219.56
7	4006.104	7994.156	3872.258	6659.535
8	671.2763	3318.277	807.9208	1983.314
9	2622.699	1367.091	2758.56	43.64889
10	8364.031	12353.06	8229.223	11018.06

Figure 2-1: Data arranged in columns and rows.

In data warehouses and data sets, rows are sometimes referred to as **observations**, **examples** or **cases**, and columns are sometimes called **variables** or **attributes.** For purposes of consistency in this book, we will use the terminology of **observations** for rows and **attributes** for columns. It is important to note that RapidMiner will use the term ***examples*** for rows of data, so keep this in mind throughout the rest of the text.

A **database** is an organized grouping of information within a specific structure. Database containers, such as the one pictured in Figure 2-2, are called **tables** in a database environment. Most databases in use today are **relational databases**—they are designed using many tables which relate to one another in a logical fashion. Relational databases generally contain dozens or even hundreds of tables, depending upon the size of the organization.



Owner_ID	Owner Name
1	Jim
2	Joan

Pet_ID	Pet_Name	Owner_ID
1	Fifi	2
2	Butch	1
3	Clover	2
4	Animal	1
5	Tank	1

Figure 2-2: A simple database with a relation between two tables.

Figure 2-2 depicts a relational database environment with two tables. The first table contains information about pet owners; the second, information about pets. The tables are related by the single column they have in common: Owner_ID. By relating tables to one another, we can reduce redundancy of data and improve database performance. The process of breaking tables apart and thereby reducing data redundancy is called **normalization**.

Most relational databases which are designed to handle a high number of reads and writes (updates and retrievals of information) are referred to as **OLTP (online transaction processing)** systems. OLTP systems are very efficient for high volume activities such as cashiering, where many items are being recorded via bar code scanners in a very short period of time. However, using OLTP databases for analysis is generally not very efficient, because in order to retrieve data from multiple tables at the same time, a query containing joins must be written. A **query** is simple a method of retrieving data from database tables for viewing. Queries are usually written in a language called **SQL (Structured Query Language; pronounced '*sequel*').** Because it is not very useful to only query pet names or owner names, for example, we must **join** two or more tables together in order to retrieve both pets and owners at the same time. Joining requires that the computer match the Owner_ID column in the Owners table to the Owner_ID column in the Pets table. When tables contain thousands or even millions of rows of data, this matching process can be very intensive and time consuming on even the most robust computers.

For much more on database design and management, check out geekgirls.com: (http://www.geekgirls.com/_menu_databases.htm).

In order to keep our transactional databases running quickly and smoothly, we may wish to create a data warehouse. A **data warehouse** is a type of large database that has been denormalized and archived. **Denormalization** is the process of intentionally combining some tables into a single table in spite of the fact that this may introduce duplicate data in some columns (or in other words, attributes).

Pet_ID	Pet_Name	Owner_Name
1	Fifi	Joan
2	Butch	Jim
3	Clover	Joan
4	Animal	Jim
5	Tank	Jim

Figure 2-3: A combination of the tables into a single data set.

Figure 2-3 depicts what our simple example data might look like if it were in a data warehouse. When we design databases in this way, we reduce the number of joins necessary to query related data, thereby speeding up the process of analyzing our data. Databases designed in this manner are called **OLAP (online analytical processing)** systems.

Transactional systems and analytical systems have conflicting purposes when it comes to database speed and performance. For this reason, it is difficult to design a single system which will serve both purposes. This is why data warehouses generally contain archived data. **Archived data** are data that have been copied out of a transactional database. Denormalization typically takes place at the time data are copied out of the transactional system. It is important to keep in mind that if a *copy* of the data is made in the data warehouse, the data may become out-of-synch. This happens when a copy is made in the data warehouse and then later, a change to the original record (observation) is made in the source database. Data mining activities performed on out-of-synch observations may be useless, or worse, misleading. An alternative archiving method would be to *move* the data out of the transactional system. This ensures that data won't get out-of-synch, however, it also makes the data unavailable should a user of the transactional system need to view or update it.

A **data set** is a subset of a database or a data warehouse. It is usually denormalized so that only one table is used. The creation of a data set may contain several steps, including appending or combining tables from source database tables, or simplifying some data expressions. One example of this may be changing a date/time format from '10-DEC-2002 12:21:56' to '12/10/02'. If this

latter date format is adequate for the type of data mining being performed, it would make sense to simplify the attribute containing dates and times when we create our data set. Data sets may be made up of a representative sample of a larger set of data, or they may contain all observations relevant to a specific group. We will discuss sampling methods and practices in Chapter 3.

TYPES OF DATA

Thus far in this text, you've read about some fundamental aspects of data which are critical to the discipline of data mining. But we haven't spent much time discussing where that data are going to come from. In essence, there are really two types of data that can be mined: **operational** and **organizational**.

The most elemental type of data, operational data, comes from transactional systems which record everyday activities. Simple encounters like buying gasoline, making an online purchase, or checking in for a flight at the airport all result in the creation of **operational data**. The times, prices and descriptions of the goods or services we have purchased are all recorded. This information can be combined in a data warehouse or may be extracted directly into a data set from the OLTP system.

Often times, transactional data is too detailed to be of much use, or the detail may compromise individuals' privacy. In many instances, government, academic or not-for-profit organizations may create data sets and then make them available to the public. For example, if we wanted to identify regions of the United States which are historically at high risk for influenza, it would be difficult to obtain permission and to collect doctor visit records nationwide and compile this information into a meaningful data set. However, the U.S. Centers for Disease Control and Prevention (CDCP), do exactly that every year. Government agencies do not always make this information immediately available to the general public, but it often can be requested. Other organizations create such summary data as well. The grocery store mentioned at the beginning of this chapter wouldn't necessarily want to analyze records of individual cans of greens beans sold, but they may want to watch trends for daily, weekly or perhaps monthly totals. **Organizational data** sets can help to protect peoples' **privacy**, while still proving useful to data miners watching for trends in a given population.

Another type of data often overlooked within organizations is something called a data mart. A **data mart** is an organizational data store, similar to a data warehouse, but often created in conjunction with business units' needs in mind, such as Marketing or Customer Service, for reporting and management purposes. Data marts are usually intentionally created by an organization to be a type of one-stop shop for employees throughout the organization to find data they might be looking for. Data marts may contain wonderful data, prime for data mining activities, but they must be known, current, and accurate to be useful. They should also be well-managed in terms of privacy and security.

All of these types of organizational data carry with them some concern. Because they are secondary, meaning they have been derived from other more detailed primary data sources, they may lack adequate documentation, and the rigor with which they were created can be highly variable. Such data sources may also not be intended for general distribution, and it is always wise to ensure proper permission is obtained before engaging in data mining activities on any data set. Remember, simply because a data set may have been acquired from the Internet does not mean it is in the public domain; and simply because a data set may exist within your organization does not mean it can be freely mined. Checking with relevant managers, authors and stakeholders is critical before beginning data mining activities.

A NOTE ABOUT PRIVACY AND SECURITY

In 2003, JetBlue Airlines supplied more than one million passenger records to a U.S. government contractor, Torch Concepts. Torch then subsequently augmented the passenger data with additional information such as family sizes and social security numbers—information purchased from a data broker called Acxiom. The data were intended for a data mining project in order to develop potential terrorist profiles. All of this was done without notification or consent of passengers. When news of the activities got out however, dozens of privacy lawsuits were filed against JetBlue, Torch and Acxiom, and several U.S. senators called for an investigation into the incident.

This incident serves several valuable purposes for this book. First, we should be aware that as we gather, organize and analyze data, there are real people behind the figures. These people have certain rights to privacy and protection against crimes such as identity theft. We as data miners

have an ethical obligation to protect these individuals' rights. This requires the utmost care in terms of information security. Simply because a government representative or contractor asks for data does not mean it should be given.

Beyond technological security however, we must also consider our moral obligation to those individuals behind the numbers. Recall the grocery store shopping card example given at the beginning of this chapter. In order to encourage use of frequent shopper cards, grocery stores frequently list two prices for items, one with use of the card and one without. For each individual, the answer to this question may vary, however, answer it for yourself: At what price mark-up has the grocery store crossed an ethical line between encouraging consumers to participate in frequent shopper programs, and forcing them to participate in order to afford to buy groceries? Again, your answer will be unique from others', however it is important to keep such moral obligations in mind when gathering, storing and mining data.

The objectives hoped for through data mining activities should never justify unethical means of achievement. Data mining can be a powerful tool for customer relationship management, marketing, operations management, and production, however in all cases the human element must be kept sharply in focus. When working long hours at a data mining task, interacting primarily with hardware, software, and numbers, it can be easy to forget about the people, and therefore it is so emphasized here.

CHAPTER SUMMARY

This chapter has introduced you to the discipline of data mining. Data mining brings statistical and logical methods of analysis to large data sets for the purposes of describing them and using them to create predictive models. Databases, data warehouses and data sets are all unique kinds of digital record keeping systems, however, they do share many similarities. Data mining is generally most effectively executed on data data sets, extracted from OLAP, rather than OLTP systems. Both operational data and organizational data provide good starting points for data mining activities, however both come with their own issues that may inhibit quality data mining activities. These should be mitigated before beginning to mine the data. Finally, when mining data, it is critical to remember the human factor behind manipulation of numbers and figures. Data miners have an ethical responsibility to the individuals whose lives may be affected by the decisions that are made as a result of data mining activities.

REVIEW QUESTIONS

1) What is data mining in general terms?

2) What is the difference between a database, a data warehouse and a data set?

3) What are some of the limitations of data mining? How can we address those limitations?

4) What is the difference between operational and organizational data? What are the pros and cons of each?

5) What are some of the ethical issues we face in data mining? How can they be addressed?

6) What is meant by out-of-synch data? How can this situation be remedied?

7) What is normalization? What are some reasons why it is a good thing in OLTP systems, but not so good in OLAP systems?

EXERCISES

1) Design a relational database with at least three tables. Be sure to create the columns necessary within each table to relate the tables to one another.

2) Design a data warehouse table with some columns which would usually be normalized. Explain why it makes sense to denormalize in a data warehouse.

3) Perform an Internet search to find information about data security and privacy. List three web sites that you found that provided information that could be applied to data mining. Explain how it might be applied.

4) Find a newspaper, magazine or Internet news article related to information privacy or security. Summarize the article and explain how it might be related to data mining.

5) Using the Internet, locate a data set which is available for download. Describe the data set (contents, purpose, size, age, etc.). Classify the data set as operational or organizational. Summarize any requirements placed on individuals who may wish to use the data set.

6) Obtain a copy of an application for a grocery store shopping card. Summarize the type of data requested when filling out the application. Give an example of how that data may aid in a data mining activity. What privacy concerns arise regarding the data being collected?