CHAPTER ONE:
INTRODUCTION TO DATA MINING AND CRISP-DM

INTRODUCTION

Data mining as a discipline is largely transparent to the world. Most of the time, we never even notice that it's happening. But whenever we sign up for a grocery store shopping card, place a purchase using a credit card, or surf the Web, we are creating data. These **data** are stored in large sets on powerful computers owned by the companies we deal with every day. Lying within those data sets are patterns—indicators of our interests, our habits, and our behaviors. Data mining allows people to locate and interpret those patterns, helping them make better informed decisions and better serve their customers. That being said, there are also concerns about the practice of data mining. Privacy watchdog groups in particular are vocal about organizations that amass vast quantities of data, some of which can be very personal in nature.

The intent of this book is to introduce you to concepts and practices common in data mining. It is intended primarily for undergraduate college students and for business professionals who may be interested in using information systems and technologies to solve business problems by mining data, but who likely do not have a formal background or education in computer science. Although data mining is the fusion of applied statistics, logic, artificial intelligence, machine learning and data management systems, you are not required to have a strong background in these fields to use this book. While having taken introductory college-level courses in statistics and databases will be helpful, care has been taken to explain within this book, the necessary concepts and techniques required to successfully learn how to mine data.

Each chapter in this book will explain a data mining concept or technique. You should understand that the book is not designed to be an instruction manual or tutorial for the tools we will use (RapidMiner and OpenOffice Base and Calc). These software packages are capable of many types of **data analysis**, and this text is not intended to cover all of their capabilities, but rather, to illustrate how these software tools can be used to perform certain kinds of data mining. The book is also not exhaustive; it includes a variety of common data mining techniques, but RapidMiner in particular is capable of many, many data mining tasks that are not covered in the book.

The chapters will all follow a common format. First, chapters will present a scenario referred to as *Context and Perspective*. This section will help you to gain a real-world idea about a certain kind of problem that data mining can help solve. It is intended to help you think of ways that the data mining technique in that given chapter can be applied to organizational problems you might face. Following *Context and Perspective,* a set of *Learning Objectives* is offered. The idea behind this section is that each chapter is designed to teach you something new about data mining. By listing the objectives at the beginning of the chapter, you will have a better idea of what you should expect to learn by reading it. The chapter will follow with several sections addressing the chapter's topic. In these sections, step-by-step examples will frequently be given to enable you to work alongside an actual data mining task. Finally, after the main concepts of the chapter have been delivered, each chapter will conclude with a *Chapter Summary*, a set of *Review Questions* to help reinforce the main points of the chapter, and one or more *Exercise* to allow you to try your hand at applying what was taught in the chapter.

A NOTE ABOUT TOOLS

There are many software tools designed to facilitate data mining, however many of these are often expensive and complicated to install, configure and use. Simply put, they're not a good fit for learning the basics of data mining. This book will use OpenOffice Calc and Base in conjunction with an open source software product called RapidMiner, developed by Rapid-I, GmbH of Dortmund, Germany. Because OpenOffice is widely available and very intuitive, it is a logical place to begin teaching introductory level data mining concepts. However, it lacks some of the tools data miners like to use. RapidMiner is an ideal complement to OpenOffice, and was selected for this book for several reasons:

- RapidMiner provides specific data mining functions not currently found in OpenOffice, such as decision trees and association rules, which you will learn to use later in this book.
- RapidMiner is easy to install and will run on just about any computer.
- RapidMiner's maker provides a Community Edition of its software, making it free for readers to obtain and use.

- Both RapidMiner and OpenOffice provide intuitive graphical user interface environments which make it easier for general computer-using audiences to the experience the power of data mining.

All examples using OpenOffice or RapidMiner in this book will be illustrated in a Microsoft Windows environment, although it should be noted that these software packages will work on a variety of computing platforms. It is recommended that you download and install these two software packages on your computer now, so that you can work along with the examples in the book if you would like.

- OpenOffice can be downloaded from: http://www.openoffice.org/
- RapidMiner Community Edition can be downloaded from: http://rapid-i.com/content/view/26/84/

THE DATA MINING PROCESS

Although data mining's roots can be traced back to the late 1980s, for most of the 1990s the field was still in its infancy. Data mining was still being defined, and refined. It was largely a loose conglomeration of data models, analysis algorithms, and ad hoc outputs. In 1999, several sizeable companies including auto maker Daimler-Benz, insurance provider OHRA, hardware and software manufacturer NCR Corp. and statistical software maker SPSS, Inc. began working together to formalize and standardize an approach to data mining. The result of their work was **CRISP-DM**, the CRoss-Industry Standard Process for Data Mining. Although

the participants in the creation of CRISP-DM certainly had vested interests in certain software and hardware tools, the process was designed independent of any specific tool. It was written in such a way as to be conceptual in nature—something that could be applied independent of any certain tool or kind of data. The process consists of six steps or phases, as illustrated in Figure 1-1.

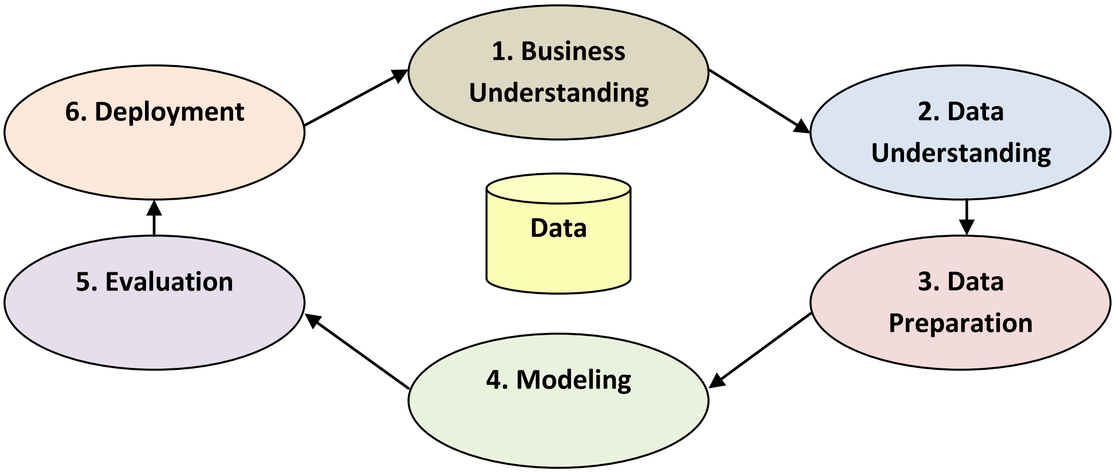

Figure 1-1: CRISP-DM Conceptual Model.

CRISP-DM Step 1: Business (Organizational) Understanding

The first step in CRISP-DM is **Business Understanding**, or what will be referred to in this text as **Organizational Understanding**, since organizations of all kinds, not just businesses, can use data mining to answer questions and solve problems. This step is crucial to a successful data mining outcome, yet is often overlooked as folks try to dive right into mining their data. This is natural of course—we are often anxious to generate some interesting output; we want to find answers. But you wouldn't begin building a car without first defining what you want the vehicle to do, and without first *designing* what you are going to *build*. Consider these oft-quoted lines from Lewis Carroll's *Alice's Adventures in Wonderland*:

> "Would you tell me, please, which way I ought to go from here?"
> "That depends a good deal on where you want to get to," said the Cat.
> "I don't much care where--" said Alice.
> "Then it doesn't matter which way you go," said the Cat.
> "--so long as I get SOMEWHERE," Alice added as an explanation.
> "Oh, you're sure to do that," said the Cat, "if you only walk long enough."

Indeed. You can mine data all day long and into the night, but if you don't know what you want to know, if you haven't defined any questions to answer, then the efforts of your data mining are less likely to be fruitful. Start with high level ideas: What is making my customers complain so much?

How can I increase my per-unit profit margin? How can I anticipate and fix manufacturing flaws and thus avoid shipping a defective product? From there, you can begin to develop the more specific questions you want to answer, and this will enable you to proceed to …

CRISP-DM Step 2: Data Understanding

As with Organizational Understanding, **Data Understanding** is a preparatory activity, and sometimes, its value is lost on people. Don't let its value be lost on you! Years ago when workers did not have their own computer (or multiple computers) sitting on their desk (or lap, or in their pocket), data were centralized. If you needed information from a company's data store, you could request a report from someone who could query that information from a central database (or fetch it from a company filing cabinet) and provide the results to you. The inventions of the personal computer, workstation, laptop, tablet computer and even smartphone have each triggered moves away from data centralization. As hard drives became simultaneously larger *and* cheaper, and as software like Microsoft Excel and Access became increasingly more accessible and easier to use, data began to disperse across the enterprise. Over time, valuable data stores became strewn across hundred and even thousands of devices, sequestered in marketing managers' spreadsheets, customer support databases, and human resources file systems.

As you can imagine, this has created a multi-faceted data problem. Marketing may have wonderful data that could be a valuable asset to senior management, but senior management may not be aware of the data's existence—either because of territorialism on the part of the marketing department, or because the marketing folks simply haven't thought to tell the executives about the data they've gathered. The same could be said of the information sharing, or lack thereof, between almost any two business units in an organization. In Corporate America lingo, the term 'silos' is often invoked to describe the separation of units to the point where interdepartmental sharing and communication is almost non-existent. It is unlikely that effective organizational data mining can occur when employees do not know *what* data they have (or could have) at their disposal or *where* those data are currently located. In chapter two we will take a closer look at some mechanisms that organizations are using to try bring all their data into a common location. These include databases, data marts and data warehouses.

Simply centralizing data is not enough however. There are plenty of question that arise once an organization's data have been corralled. Where did the data come from? Who collected them and was there a standard method of collection? What do the various columns and rows of data mean? Are there acronyms or abbreviations that are unknown or unclear? You may need to do some research in the Data Preparation phase of your data mining activities. Sometimes you will need to meet with subject matter experts in various departments to unravel where certain data came from, how they were collected, and how they have been coded and stored. It is critically important that you verify the accuracy and reliability of the data as well. The old adage "It's better than nothing" does not apply in data mining. Inaccurate or incomplete data could be worse than nothing in a data mining activity, because decisions based upon partial or wrong data are likely to be partial or wrong decisions. Once you have gathered, identified and understood your data assets, then you may engage in…

CRISP-DM Step 3: Data Preparation

Data come in many shapes and formats. Some data are numeric, some are in paragraphs of text, and others are in picture form such as charts, graphs and maps. Some data are anecdotal or narrative, such as comments on a customer satisfaction survey or the transcript of a witness's testimony. Data that aren't in rows or columns of numbers shouldn't be dismissed though—sometimes non-traditional data formats can be the most information rich. We'll talk in this book about approaches to formatting data, beginning in Chapter 2. Although rows and columns will be one of our most common layouts, we'll also get into text mining where paragraphs can be fed into RapidMiner and analyzed for patterns as well.

Data Preparation involves a number of activities. These may include joining two or more data sets together, reducing data sets to only those variables that are interesting in a given data mining exercise, scrubbing data clean of anomalies such as outlier observations or missing data, or re-formatting data for consistency purposes. For example, you may have seen a spreadsheet or database that held phone numbers in many different formats:

(555) 555-5555	555/555-5555
555-555-5555	555.555.5555
555 555 5555	5555555555

Each of these offers the same phone number, but stored in different formats. The results of a data mining exercise are most likely to yield good, useful results when the underlying data are as consistent as possible. Data preparation can help to ensure that you improve your chances of a successful outcome when you begin…

CRISP-DM Step 4: Modeling

A **model**, in data mining at least, is a computerized representation of real-world observations. Models are the application of algorithms to seek out, identify, and display any patterns or messages in your data. There are two basic kinds or types of models in data mining: those that **classify** and those that **predict**.

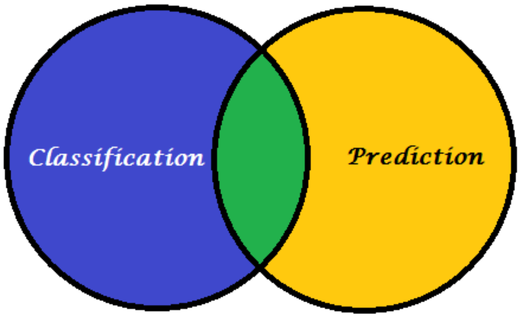

Figure 1-2: Types of Data Mining Models.

As you can see in Figure 1-2, there is some overlap between the types of models data mining uses. For example, this book will teaching you about **decision trees**. Decision Trees are a predictive model used to determine which attributes of a given data set are the strongest indicators of a given outcome. The outcome is usually expressed as the likelihood that an observation will fall into a certain category. Thus, Decision Trees are predictive in nature, but they also help us to classify our data. This will probably make more sense when we get to the chapter on Decision Trees, but for now, it's important just to understand that models help us to classify and predict based on patterns the models find in our data.

Models may be simple or complex. They may contain only a single process, or stream, or they may contain sub-processes. Regardless of their layout, models are where data mining moves from preparation and understanding to development and interpretation. We will build a number of example models in this text. Once a model has been built, it is time for…

CRISP-DM Step 5: Evaluation

All analyses of data have the potential for false positives. Even if a model doesn't yield false positives however, the model may not find any interesting patterns in your data. This may be because the model isn't set up well to find the patterns, you could be using the wrong technique, or there simply may not be anything interesting in your data for the model to find. The Evaluation phase of CRISP-DM is there specifically to help you determine how valuable your model is, and what you might want to do with it.

Evaluation can be accomplished using a number of techniques, both mathematical and logical in nature. This book will examine techniques for cross-validation and testing for false positives using RapidMiner. For some models, the power or strength indicated by certain test statistics will also be discussed. Beyond these measures however, model evaluation must also include a human aspect. As individuals gain experience and expertise in their field, they will have operational knowledge which may not be measurable in a mathematical sense, but is nonetheless indispensable in determining the value of a data mining model. This human element will also be discussed throughout the book. Using both data-driven and instinctive evaluation techniques to determine a model's usefulness, we can then decide how to move on to…

CRISP-DM Step 6: Deployment

If you have successfully identified your questions, prepared data that can answer those questions, and created a model that passes the test of being interesting and useful, then you have arrived at the point of *actually using your results*. This is **deployment**, and it is a happy and busy time for a data miner. Activities in this phase include setting up automating your model, meeting with consumers of your model's outputs, integrating with existing management or operational information systems, feeding new learning from model use back into the model to improve its accuracy and performance, and monitoring and measuring the outcomes of model use. Be prepared for a bit of distrust of your model at first—you may even face pushback from groups who may feel their jobs are threatened by this new tool, or who may not trust the reliability or accuracy of the outputs. But don't let this discourage you! Remember that CBS did not trust the initial predictions of the UNIVAC, one of the first commercial computer systems, when the network used it to predict the eventual outcome of the 1952 presidential election on election night. With only 5% of the votes counted, UNIVAC predicted Dwight D. Eisenhower would defeat Adlai Stevenson in a landslide;

something no pollster or election insider consider likely, or even possible. In fact, most 'experts' expected Stevenson to win by a narrow margin, with some acknowledging that because they expected it to be close, Eisenhower might also prevail in a tight vote. It was only late that night, when human vote counts confirmed that Eisenhower was running away with the election, that CBS went on the air to acknowledge first that Eisenhower had won, and second, that UNIVAC had predicted this very outcome hours earlier, but network brass had refused to trust the computer's prediction. UNIVAC was further vindicated later, when it's prediction was found to be within 1% of what the eventually tally showed. New **technology** is often unsettling to people, and it is hard sometimes to trust what computers show. Be patient and specific as you explain how a new data mining model works, what the results mean, and how they can be used.

While the UNIVAC example illustrates the power and utility of predictive computer modeling (despite inherent mistrust), it should not construed as a reason for blind trust either. In the days of UNIVAC, the biggest problem was the newness of the technology. It was doing something no one really expected or could explain, and because few people understood how the computer worked, it was hard to trust it. Today we face a different but equally troubling problem: computers have become ubiquitous, and too often, we don't question enough whether or not the results are accurate and meaningful. In order for data mining models to be effectively deployed, balance must be struck. By clearly communicating a model's function and utility to stake holders, thoroughly testing and proving the model, then planning for and monitoring its implementation, data mining models can be effectively introduced into the organizational flow. Failure to carefully and effectively manage deployment however can sink even the best and most effective models.

DATA MINING AND YOU

Because data mining can be applied to such a wide array of professional fields, this book has been written with the intent of explaining data mining in plain English, using software tools that are accessible and intuitive to everyone. You may not have studied algorithms, data structures, or programming, but you may have questions that can be answered through data mining. It is our hope that by writing in an informal tone and by illustrating data mining concepts with accessible, logical examples, data mining can become a useful tool for you regardless of your previous level of data analysis or computing expertise. Let's start digging!