CHAPTER 1

Data Mining in Context

To say that the past century saw rapid change is both a cliché and an understatement. Although the rapid pace of change was felt in nearly every area, it is hard to find examples of anything, anywhere that has changed as fast as the quantity of stored information. This information explosion has created new opportunities and new headaches in every field, from manufacturing to medicine to marketing. To appreciate just how fast the world's store of information has grown in recent years, it is instructive to compare it against some of the standard benchmarks of the twentieth century.

In 1900, the world population (another area where growth has frequently been described as "explosive") was 1.6 billion. A hundred years later, the population topped 6 billion. So, the population explosion caused the number of people living on Earth to increase by a factor of 3.75 over the course of the century.

In 1906, the Stanley twins, Francis and Freelan, established a world land speed record of 122 miles per hour with their Stanley Steamer. The land speed record was, of course, the only one that counted; 15 miles per hour was a pretty good speed for a ship and airplanes had only been in the air for three years, so they were in no position to challenge the speed record. On their journey to the moon 63 years later, the Apollo astronauts traveled at nearly 25,000 miles per hour—223 times as fast.

The trip to the moon provides another yardstick. In 1900, the longest journey one could reasonably make was 25,000 miles—the distance required to circumnavigate the globe. The round trip distance to the moon is 19 times as far.

Impressive numbers all, but nothing compared to the growth in corporate data. At the beginning of the twentieth century, or even at its midpoint, no company had more than a few tens of megabytes of data lying around in company ledgers, order books, and file cabinets. Today, the largest corporate databases are measured in terabytes. For these organizations, corporate data has grown by a factor of 100,000. This comparison (like all the size comparisons in this book) is made on the basis of text and numbers alone. Video and audio recordings take up huge amounts of storage, and are very interesting in their own right, but the data mining techniques that we discuss cannot yet easily be applied to them.

WARNING

Comparing data sizes is a tricky business. This is especially true when we get away from structured data such as the records in a database and into areas like music or photography. How large is a picture? That depends on the method of encoding it, the resolution of the image, the amount of data compression, and so on. Even comparisons of structured, relational databases can be misleading. Should we count the space used for indexes? What about the disk required for scratch space? Vendors sometimes report the total disk capacity of the database server instead of the actual size of the database in order to come up with a more impressive number. The data size comparisons in this chapter are based on the actual space required to store uncompressed text with no illustrations or indexes.

How large are today's databases? A comparison is instructive: From the time, over five thousand years ago, when some Sumerian took a reed to a clay table to scratch out the world's oldest known shopping list, an awful lot has been written. The Library of Congress contains 17 million books (and millions more manuscripts, maps, works of art, etc.). If each of those books were the same size as this one—about a megabyte of text in MS Word format—then the world's largest compendium of human knowledge, if typed into a computer database, would consume about 17 terabytes of disk space. As it happens, 17 terabytes is also the size of the package-level detail database used to track shipments at UPS.

In other words, a single company has as much data on where and when its customers are shipping packages as is contained in all the books in the Library of Congress. It seems incredible, but it is a natural consequence of increasingly automated operations. When every bar-coded package is scanned several times in transit, it doesn't take long to build up a lot of records—especially when you ship millions of packages every day. The same is true of manufacturing processes that are run by statistical controls based on readings from thousands of sensors, or telephone systems that keep track of time, duration, and network routing for every call.

For the most part, this data is not being collected so that it can be analyzed or used for predictive modeling; it is being collected to improve the efficiency of underlying operations. Once collected, however, it represents a wealth of information that can be used to improve business decision-making in every area, including the one that is the primary focus of this book: customer relationship management.

As more and more commerce moves onto the Web, the volume of data collected and the need for analysis are both increasing dramatically. The volume increases because e-commerce Web sites can keep track of much more than orders: every link followed and every item viewed is noted in the log. The need for analysis increases because these slender electronic traces are all the e-business has to go on when trying to form lasting, profitable relationships with customers who may be anywhere on earth and are free to switch to another supplier at the click of a mouse. Turning millions of transaction records from Web log files, call detail databases, or point-of-sale devices into recognizable portraits of consumers or business-to-business customers requires art and science, mathematics and intuition.

Data miners—the people who apply this potent mixture of massive computing power, clever algorithms, business knowledge, and human intuition—do not ply their trade in a vacuum. This chapter attempts to put data mining in its proper context by showing how it relates to business processes, information technology, and the wider world.

What Is Data Mining?

In our earlier book, *Data Mining Techniques for Marketing, Sales and Customer Support* (1997, John Wiley & Sons), we gave the following definition:

> Data mining is the process of exploration and analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns and rules.

Revisiting that definition a few years later, we find that, for the most part, it stands up pretty well. We like the emphasis on large quantities of data, since data volumes continue to increase. We like the notion that the patterns and rules to be found ought to be *meaningful*. If there is anything we regret, it is the phrase "by automatic or semi-automatic means." Not because it is untrue—without automation it would be impossible to mine the huge quantities of data being generated today—but because we feel there has come to be too much focus on the automatic techniques and not enough on the exploration and analysis. This

has misled many people into believing that data mining is a product that can be bought rather than a discipline that must be mastered. In this book, although we discuss algorithms and techniques when necessary, we put the focus back where it belongs: the data mining *process*. But before we can discuss process, we need to establish a common understanding of what data mining is and how it can be used. For readers of our earlier book, this will be review.

What Can Data Mining Do?

The term data mining is often thrown around rather loosely. In this book, we use the term for a specific set of activities, all of which involve extracting meaningful new information from the data. The six activities are:

- Classification
- Estimation
- Prediction
- Affinity grouping or association rules
- Clustering
- Description and visualization

The first three tasks—classification, estimation, and prediction—are all examples of *directed* data mining. In directed data mining, the goal is to use the available data to build a model that describes one particular variable of interest in terms of the rest of the available data. The next three tasks are examples of *undirected* data mining. In undirected data mining, no variable is singled out as the target; the goal is to establish some relationship among all the variables.

Classification

Classification consists of examining the features of a newly presented object and assigning to it a predefined class. For our purposes, the objects to be classified are generally represented by records in a database. The act of classification consists of updating each record by filling in a field with a class code.

The classification task is characterized by a well-defined definition of the classes, and a training set consisting of preclassified examples. The task is to build a model that can be applied to unclassified data in order to classify it. Examples of classification tasks include:

- Assigning keywords to articles as they come in off the news wire.
- Classifying credit applicants as low, medium, or high risk.
- Determining which home telephone lines are used for Internet access.
- Assigning customers to predefined customer segments.

In all of these examples, there are a limited number of already-known classes and we expect to be able to assign any record into one or another of them.

Estimation

Classification deals with discrete outcomes: yes or no, debit card, mortgage, or car loan. Estimation deals with continuously valued outcomes. Given some input data, we use estimation to come up with a value for some unknown continuous variable such as income, height, or credit card balance.

In practice, estimation is often used to perform a classification task. A bank trying to decide to whom they should offer a home equity loan might run all its customers through a model that gives them each a score, such as a number between 0 and 1. This is actually an estimate of the probability that the person will respond positively to an offer. This approach has the great advantage that the individual records may now be rank ordered from most likely to least likely to respond. The classification task now comes down to establishing a threshold score. Anyone with a score greater than or equal to the threshold will receive the offer.

Other classification tasks can also be recast as estimation tasks. For instance, churn modeling—figuring out which customers are likely to stop being customers (*churn*)—may be viewed as classifying people as likely or unlikely to churn, or it may be viewed as estimating the length of time that a customer will stay.

Other examples of estimation tasks include:

- Estimating the number of children in a family.
- Estimating a family's total household income.
- Estimating the value of a piece of real estate.

Often, classification and estimation are used together, as when data mining is used to predict who is likely to respond to a credit card balance transfer offer and also to estimate the size of the balance to be transferred.

Prediction

Arguably, there should not be a separate heading for prediction. Any prediction can be thought of as classification or estimation. The difference is one of emphasis. When data mining is used to classify a phone line as primarily used for Internet access or a credit card transaction as fraudulent, we do not expect to be able to go back later to see if the classification was correct. Our classification may be correct or incorrect, but the uncertainty is due only to incomplete knowledge: out in the real world, the relevant actions have already taken place. The phone is or is not used primarily to dial the local ISP. The credit card transaction is or is not fraudulent. With enough effort, it is possible to check. Predictive tasks feel different because the records are classified according to some predicted future behavior or estimated future value. With prediction, the only way to check the accuracy of the classification is to wait and see.

Examples of prediction tasks include:

- Predicting the size of the balance that will be transferred if a credit card prospect accepts a balance transfer offer.
- Predicting which customers will leave within the next six months.
- Predicting which telephone subscribers will order a value-added service such as three-way calling or voice mail.

Any of the techniques used for classification and estimation can be adapted for use in prediction by using training examples where the value of the variable to be predicted is already known, along with historical data for those examples. The historical data is used to build a model that explains the current observed behavior. When this model is applied to current inputs, the result is a prediction of future behavior.

Affinity Grouping or Association Rules

The task of affinity grouping is to determine which things go together. The prototypical example is determining what things go together in a shopping cart at the supermarket. Retail chains can use affinity grouping to plan arrangement of items on store shelves or in a catalog so that items often purchased together will be seen together. Affinity grouping can also be used to identify cross-selling opportunities and to design attractive packages or groupings of products and services.

Clustering

Clustering is the task of segmenting a diverse group into a number of more similar subgroups or clusters. What distinguishes clustering from classification is that clustering does not rely on predefined classes.

In clustering, there are no predefined classes and no examples. The records are grouped together on the basis of self-similarity. It is up to the miner to determine what meaning, if any, to attach to the resulting clusters. A particular cluster of symptoms might indicate a particular disease. Dissimilar clusters of video and music purchases might indicate membership in different subcultures.

Clustering is often done as a prelude to some other form of data mining or modeling. For example, clustering might be the first step in a market segmentation effort. Instead of trying to come up with a one-size-fits-all rule for "what kind of promotion do customers respond to best," first divide the customer base into clusters or people with similar buying habits, and then ask what kind of promotion works best for each cluster.

Description and Visualization

Sometimes the purpose of data mining is simply to describe what is going on in a complicated database in a way that increases our understanding of the people, products, or processes that produced the data in the first place. A good enough description of a behavior will often suggest an explanation for it as well. At the very least, a good description suggests where to start looking for an explanation. The famous gender gap in American politics is an example of how a simple description, "women support Democrats in greater numbers than do men," can provoke large amounts of interest and further study on the part of journalists, sociologists, economists, and political scientists, not to mention candidates for public office.

Data visualization is one powerful form of descriptive data mining. It is not always easy to come up with meaningful visualizations, but the right picture really can be worth a thousand association rules since human beings are extremely practiced at extracting meaning from visual scenes.

The Business Context for Data Mining

Data mining—extracting meaningful patterns and rules from large quantities of information—is clearly useful in any field where there are large quantities of data and something worth learning. We would not be surprised to learn, for example, that military intelligence organizations use data mining techniques to process large quantities of satellite imagery in an attempt to classify things on the ground as tanks or tractors—targets or public relations disasters in the making.

In the business context, the same rule applies: Data mining is useful wherever there are large quantities of data and something worth learning. In business, there is an explicit definition of what it means for a thing to be worth learning. For a business, something is worth learning if the resulting knowledge is worth more money than it costs to discover. Actually, the definition is even

stricter than that: Something is worth knowing if the return on the investment required to learn it is greater than the return from investing the same funds some other way.

In academia, knowledge is considered to have intrinsic value quite apart from any application. In the business context however, knowledge can be valuable in two ways: It can increase profit by *lowering cost*, or it can increase profit by *raising revenue*. Actually, there is a third way: it can increase the stock price by holding out the promise of *future* increased profit via either of these mechanisms.

Data Mining as a Research Tool

One way that data mining can lower costs is at the very beginning of the product lifecycle, during research and development. The pharmaceuticals industry provides a good example. This industry is characterized by very high R&D costs. Globally, the industry spends $13 billion a year on research and development of prescription drugs. At the risk of oversimplifying a very complex business, the drug development process is like a large funnel with millions of chemical compounds going in the wide end and only a few safe and useful drugs emerging at the narrow spout.

To make it through the funnel a potential drug must first be found to bind with the correct target molecule. Those that do are called *leads*. Next, the lead must be shown to have some desired effect in the test tube—those that show no effect are discarded. Next the lead must be shown to be capable of being absorbed by the body and surviving in the complex and hostile environment of a living organism—those that are not absorbed are discarded. Then the lead must be shown to be nontoxic and to have some useful effect in animal trials—those that are toxic or useless are discarded. Finally, the lead must go through a series of clinical trials to prove that it is safe and effective in humans. Very few candidate molecules make it through this whole process and become drugs—about one in ten thousand—at an average cost of $300 million.

There is certainly plenty of data on which predictions might be based. Modern pharmaceutical laboratories employ a technique called high-throughput screening (HTS) to select candidate drugs. Automated systems perform combinatorial chemistry to create a wide variety of organic molecules from a small set of known reagents. These molecules are then screened by exotic robots that manipulate special plates containing dozens of cavities that hold chemical solutions with the target receptor molecule. These robots can recognize when a molecule binds to the target receptor. Molecules that bind to the target receptor, but not to similar ones, are likely leads.

All this automated testing yields data that is perfect for data mining—many input variables and a simple yes/no output variable. The pharmaceutical

companies use sophisticated prediction techniques to determine which chemicals are likely (or highly unlikely) to produce useful drugs. By focusing the research on the appropriate chemicals (or not doing the research on unlikely chemicals), these companies can save millions and millions of dollars.

An entire discipline of *bioinformatics* has grown up to mine and interpret the data being generated by high throughput screening and other frontiers of biology such as the mapping of the entire human genetic sequence.

Data Mining for Process Improvement

Another area where data mining can be profitably employed to save money is manufacturing. Many modern manufacturing processes from chip fabrication to brewing are controlled using statistical process control. Sensors keep track of pressure, temperature, speed, color, humidity, and whatever other variables are appropriate to a particular process. Computer programs watch over the output from these sensors and order slight adjustments to keep the readings within the proper bounds. But what *are* the proper bounds? There are so many complex interactions between the variables that even expert human operators often cannot be sure.

The consequence of a problem in the manufacturing process is often a ruined batch of product and a very costly shutdown and restart of the process. Once again, the data produced by automated manufacturing systems is perfect for data mining: huge volumes of input, consisting of precisely measured values for scores or hundreds of variables and a few simple outputs like "good" and "bad." In Chapter 14 there are two case studies of data mining in the commercial printing industry where millions of dollars were saved by using data mining techniques to generate process control rules.

Data Mining for Marketing

Many of the most successful applications of data mining are in the marketing arena, especially in the area known as database marketing. In this world, data mining is used in both the cost term and the revenue term of the profit equation. In database marketing, the database refers to a collection of data on prospective targets of a marketing campaign. Depending on the situation, this data may be detailed behavioral data on existing customers culled from operational systems such as order tracking systems, billing systems, point-of-sale systems, etc. Or it may be rudimentary information of the kind that, for a fee, is readily available on every U.S. household and, to a lesser extent, on households in other countries as well.

Data mining can be used to cut marketing costs by eliminating calls and letters to people who are very unlikely to respond to an offer. For example, the AARP

(formerly, the American Association of Retired People, now an allegedly meaningless acronym) saved millions of dollars by excluding the 10 percent of eligible households who were judged least likely to become members. On the revenue side, data mining can be used to spot the most valuable prospects, those likely to buy the highest insurance coverage amounts or the most expensive, high-margin automobiles. Since much of the authors' experience is in this area, this book is full of examples of data mining used to improve the targeting of marketing campaigns.

Data Mining for Customer Relationship Management

The phrase "customer relationship management" seems to be on the lips of every chief executive and management consultant these days. The term has come to embody much of what used to be called one-to-one marketing, along with ideas about sales force automation and customization. Good customer relationship management means presenting a single image of the company across all the many channels a customer may use to interact with the firm, and keeping a single image of the customer that is shared across the enterprise. Good customer relationship management requires understanding who customers are and what they like and don't like. It means anticipating their needs and addressing them proactively. It means recognizing when they are unhappy and doing something about it before they get fed up and go to a competitor.

Data mining plays a leading role in every facet of customer relationship management. Only through the application of data mining techniques can a large enterprise hope to turn the myriad records in its customer databases into some sort of coherent picture of its customers. In Chapter 2, we see how data mining allows a corporation to *learn* from all the observations of customer behavior that are stored in the customer information warehouse that serves as the corporate memory. In Chapter 4 we trace the customer lifecycle from before the prospect becomes a customer until after the customer has left, and show how data mining can be applied to improve customer relationship management at every point along the way.

The Technical Context for Data Mining

In the second part of this book, we will be looking at the technical context of data mining. This context has three main areas:

1. Algorithms and techniques
2. Data
3. Modeling practices

The field that has come to be called *data mining* has grown from several antecedents. On the academic side are machine learning and statistics. Machine learning has contributed important algorithms for recognizing patterns in data. Machine-learning researchers are on the bleeding edge, conjuring ideas about how to make computers think. Statistics is another important area that provides background for data mining. Statisticians offer mathematical rigor; not only do they understand the algorithms, they understand the best practices in modeling and experimental design.

The final thread is decision support. Over the past few decades, people have been gathering data into databases to make better informed decisions. Data mining is a natural extension of this effort.

Data Mining and Machine Learning

The machine learning people come from the computer science and artificial intelligence worlds. They have focused their efforts on getting computers to display intelligence. In particular, the machine learning community is interested in writing computer programs that are capable of learning by example. The first kind of learning manifests itself by a newfound ability to perform some task such as balancing a broom handle or recognizing written characters. In other cases the new learning is expressed as rules that have been induced from the examples. Neural networks have proved to be very successful at the first kind of learning and decision trees have proved to be very successful at the second.

The term data mining, in its present, nonpejorative sense, was first used by people who took the methods of the machine learning and began to apply them to fields outside of computer science and artificial intelligence (AI)—fields such as industrial process control and direct marketing. This search for practical applications was probably encouraged by the collapse of funding for artificial intelligence research in the early 1980s when the over-ambitious claims from the 1960s and 1970s (machine translation, natural language recognition) failed to materialize. The choice of the term *data mining* for the new, business-oriented applications of AI research shows how little overlap there was between this group and the statisticians, actuaries, and economists who had long been doing predictive modeling. For the latter group, the term "data mining" meant searching for data to support a particular point of view rather than letting the facts speak for themselves. The data miners were smart people getting good results, but they were not mathematicians.

Data Mining and Statistics

Statistics has been another important thread that has supported data mining. For centuries, people have used statistical techniques to understand the natural world. These have included predictive algorithms (which statisticians call

regression), sampling methodologies, and experimental design. Now, they are applying these techniques to the business world.

For years, the work done by the machine learning researchers worked in practice but had a very limited foundation theoretically. The machine learning folks preferred anthropomorphic terminology and allusions to the biological sciences to rigorous mathematical proofs. Statisticians, who are by nature much more comfortable with numbers, were not impressed by this approach.

Data Mining and Decision Support

Decision support is a broad term for the entire information technology infrastructure that companies and other organizations use to make informed decisions. The term covers both relational and dimensional databases used for decision support. Decision support systems contrast with online transaction processing systems (OLTP). OLTP databases are designed to process large numbers of transactions very fast. In database terminology, a transaction is a complete action that must either finish successfully or appear not to have happened. For example, transferring money from your savings account to your checking account at an ATM machine is a single transaction. It would not be acceptable for this transaction to be interrupted in the middle after your savings account has been debited but before your checking account has been credited. OLTP databases are designed to ensure the sanctity of such transactions and to allow very efficient access to single records, such as one customer's account balance—or room reservation, or airline seat assignment, or last payment.

Decision support databases have very different requirements. In decision support, it is rarely useful to look at individual records. Decision support databases are designed to support complex queries such as "Which customers spent more than $100 at a restaurant more than 100 miles from home in two of the last three months?" To answer this question, you need to be able to translate each customer's address into geographic coordinates (using the centroid of the zip code is sufficient) and do the same for the restaurants before beginning to aggregate the charges.

It turns out that the design requirements for the two types of database are so incompatible that the same information must often be stored at least twice—once in an operational system that takes care of transactions and once in a decision support system where the historical record can be studied.

Data Warehouses

A special case of a decision support database is the *data warehouse*. A data warehouse is a large decision support database fed by many operational systems. Data warehouses are motivated by the need to view the entire enterprise from a single point of view instead of as a collection of narrowly defined "silos." On its

way into the warehouse, data from the operational systems is cleaned and transformed so that database fields from disparate sources share the same definitions.

A data warehouse sounds like a great resource for data mining, and in some cases it is, but all too often the transformations applied to the incoming data destroy valuable information. Data warehouses are often *normalized*, so they have the property that any given item of information is stored exactly once. This is very efficient for storing large amounts of data. However, it often requires "killer queries" to access data of interest.

Of particular concern is aggregation and summarization. Although it is increasingly common for data warehouses to store atomic data, and indeed, some warehouses are now designed with data mining in mind, we still run across many data warehouses that contain only summaries of historical data. For instance, a data warehouse may store the monthly balances on a credit card, but may not include the individual transactions. In the past, data warehouses have been designed for reporting, not for mining. Data mining requires access to data at a detailed level because that is where the most interesting patterns are discovered. This is not to say that the inputs to data mining models cannot be summarized; they often are. It is just that the aggregations performed by the data miner may not have been anticipated by a warehouse designer.

Data warehouses often focus on the customer level—a useful and appropriate focus. However, the focus on customers often requires summarizing more detailed data, such as call records, line item detail, or individual banking transactions. Even if the detailed data is supposed to be available in the warehouse, accessing a normalized set of tables requires "killer queries" that never get run for practical reasons. When the data is organized in a "star schema" as advocated by Ralph Kimball in his excellent books on data warehouse design, the response time may be better. These star-schema databases are discussed in the following section.

OLAP, Data Marts, and Multidimensional Databases

Another common type of decision support database is the data mart designed for online analytical processing (OLAP). Data marts address a weakness of normalized data warehouses, which is that in their desire to represent a holistic view of an entire enterprise, they insist on centralizing the implementation, which may add months or years to the project.

OLAP databases for decision support offer improved speed and responsiveness by limiting themselves to a single view of the data. Typically, this point of view is that of the department that owns the database. OLAP databases are organized along the different *dimensions* of the business such as time, product type, and geography, allowing analysts to slice and dice data along them. This multidimensional structure is often called a *cube* (even when there are many more than

three dimensions). Each dimension can, and usually does, have many levels of aggregation. In fact, a single dimension may have multiple hierarchies. So the geography dimension might be arranged into stores, cities, states, and countries, or into stores, sales regions, and countries, or both. Or the time dimensions may be organized into days, weeks, and years, or days, months, quarters, and years. An added complication is that the dimensions are not static. For example, many companies redraw their sales regions every time a new VP of sales comes along, if not more frequently. The structure chosen for a multidimensional database limits the kind of analysis we can perform. Often, a dimension of particular interest to us, the *customer dimension*, is entirely absent!

When a multidimensional database is stored in a relational database system, the arrangement of the tables—one central fact table with many dimensional tables surrounding it—is called a *star schema*, or dimensional model. Such an arrangement is especially appropriate for a database that serves the interests of a particular department such as finance or marketing since it is easier to get agreement as to what are the central facts to be tracked and what dimensions would be useful for tracking them within a single department than across a larger enterprise. These specialized decision support databases are often called *data marts*. Sometimes a data warehouse is made up of a collection of data marts.

In this approach to data warehouse design, dimensional models are the basis for a style of incremental design of an enterprise data warehouse that is inherently distributed. The main challenge for the data warehouse teams building these distributed data marts is to establish what are called conformed dimensions and conformed facts so that the separate data marts will work together. For more on this topic, see Ralph Kimball, et al.'s book, *The Data Warehouse Lifecycle Toolkit* (Wiley, 1998). We discuss OLAP and multi-dimensional databases in more detail in Chapter 6.

Decision Support Fusion

To the business user, the distinctions between online analytic processing, data mining, and data visualization seem pointless. To make better decisions, management needs answers to all kinds of questions. Today, some of those questions ("How have widget sales changed quarter over quarter by sales rep and widget type?") are answered using OLAP. Other questions ("How are sales varying by geography and widget type?") are best understood through visualization—in this case, perhaps a map with different countries shown in relief with height representing total sales and color representing product mix. Another family of questions ("Which customers should receive the 96-page holiday catalog, and which should receive the 120-page catalog?") are best addressed through data mining.

The VP of Marketing wants answers of all three types of question, and probably does not understand why the answers have to come from three different

software systems, quite possibly running on different computers and accessing slightly different data. We believe that the future will bring tighter integration of various decision support technologies such as data mining, OLAP, data warehousing, and visualization. Already, there are signs: The OLAP Council, an industry group, has changed its name to the Decision Support Council. Some data mining packages (SGI's MineSet is a notable example) include sophisticated visualization tools. As this book was being written, Oracle, the leading relational database vendor, announced the acquisition of the data mining business unit from Thinking Machines and of its Darwin data mining package. All of these demonstrate convergence in the market.

Data Mining and Computer Technology

Data mining requires complex calculations to be applied to vast quantities of data. Only a few years ago much of the work described in this book would not have been technically feasible within reasonable cost constraints. The incredible advances in computing power, price/performance, and data input and output speeds have made large-scale data mining practical and profitable. For the most part, computing technology remains behind the scenes in this book, but in a few cases where large collections of transactions are involved (as in the analysis of telephone call detail records discussed in Chapter 12) the need for parallel processing becomes explicit.

The Societal Context for Data Mining

Data mining as a technical activity and a business activity takes place against the background of rapidly changing societies that are having to adapt to new circumstances brought about by the information revolution. Authoritarian governments that could once "protect" their citizens from outside points of view are finding it very hard to control the Internet. On a more personal level, the authors can determine the number of volumes in the Library of Congress while sitting at home—or search the world for a favorite out-of-print book or vinyl record album that we might never have found in years of searching second-hand stores, and bid on it in an electronic auction.

On the down side, things that once seemed personal and private such as our taste in magazines, the groceries we buy, and even the drugs we are prescribed, are finding their way into databases. Even information that has long been a matter of public record—our real estate tax assessments, car registrations, marriage licenses, and birth announcements—seem a bit more public when they can readily be retrieved from an electronic database rather than copied by hand from a dusty ledger.

Data Mining and Individual Predictions

We often speak of using data mining to "predict who is likely to respond to an offer" or to "predict who is likely to cancel a subscription." Since people often feel they don't know themselves what they are likely to do next month or next year, this supposed ability to predict individual behavior imbues data mining with an unjustified aura of near-magical power. The fact is that, to take a typical application of data mining to direct marketing, *95 percent of the people picked by data mining to be likely responders to an offer will not respond.* In other words, at the level of individual consumers, *data mining predictions are nearly always wrong.* If this is a crystal ball, it is a pretty cloudy one!

The reason that data mining is valuable, despite being so very inaccurate, is that although only 5 percent of the people predicted to respond actually do so, that may be a significantly higher number than would have responded if no data mining model had been used. The ability of data mining to identify a population within which we can expect a 5 percent response rate, instead of the 2.5 percent response rate we could achieve without data mining, makes it worthwhile from a business point of view.

Unfortunately, it is awkward, when talking or writing about data mining, to lard every sentence with words like "overall response in the population" so, as shorthand, we speak of "finding responders" or "identifying churners." It may be comforting to readers who have no idea what kind of car they will buy when the current one dies to know that we don't either!

Since data mining is often portrayed—even by its practitioners—as a means of using this sort of information to predict what individual consumers will do before they even know themselves, it is not surprising that it is sometimes controversial. Actually, as described in the sidebar, data mining is not very good at targeting individuals for anything, but since we data miners have been so caught up in the rhetoric of individualization and one-to-one marketing, we can hardly blame the public for thinking that it is.

In a world where there is suddenly more information more readily available than ever before, data mining is bound to play an increasing role in helping both consumers and businesses find the signal—whatever it is they are really looking for—in all that noise. Along the way, many issues around privacy, ownership of data, and proper use will have to be addressed much more fully than they are today. We explore some of these issues in the final chapter of this book. In the next chapter, we start our journey toward mastering data mining by asking and addressing the question "why bother?"