People have been recording and extracting knowledge from data since the beginning of time. The cave drawings of Arles, the cuneiform tablets documenting shipboard loading manifests of ancient Babylon, and the Rosetta Stone are examples of the defining human characteristic to make sense of the world through data constructs recorded in symbolic—frequently numeric—form. The cave drawings capture the experience of the day—the life and death dramas of the hunt, the harvest, the feasting, and the fertility; the cuneiform tablets record the minutiae of early trade—counting the weight, cut, and number of precious stones or the number and volume of amphorae filled with olive oil; and the Rosetta Stone provides a key to Egyptian hieroglyphics.

Everywhere and always people reflect and record their reality in data laid down in various recording media. The earliest data miners reconstructed life styles from cave drawings so as to describe and predict human activity in those circumstances. They could describe and predict trading patterns and the effect of variables on the olive tree harvest in the ancient Mediterranean Sea area. Indeed, even today archeologists and anthropologists can infer effects on current-day trading patterns based on early trading models built from examining the data contained in these and other tablets. These tablets, of course, are “little tables”—the precursors of modern database systems.
So data mining has its roots in one of the oldest of human activities: the desire to summarize experience in some numeric or symbolic form so as to describe it better and preserve both meaning and experience. As soon as we describe and preserve experience through data and symbolic traces, we inevitably begin the process of disentangling the meanings through some kind of data mining process. Regardless of the source of the record, it seems inevitable that someone will come along to interpret it so as to make better predictions about the experience that has been recorded. Often the description seems out of idle curiosity, but inevitably the motivation turns toward extracting some kind of knowledge for profit or knowledge that can potentially be translated into another kind of spiritual or material return.

Although data mining, or knowledge discovery as it is sometimes called, seems to be a very recent and novel invention, the origins of data mining and knowledge discovery are as old as the record of civilization. Nowadays, as our ability to record data increases—astronomically it seems—so too does our ingenuity turn to develop more powerful data mining (data disentangling) methods to keep up with the interpretation of the constant, and growing, accumulation of data. This leads us to a definition of data mining: Data mining is a current-day term for the computer implementation of a timeless human activity. It is the process of using automated methods to uncover meaning—in the form of trends, patterns, and relationships—from accumulated electronic traces of data.

It is normal to use data mining for a purpose—typically to gain insight and improvements in business functions. The utility of data is unquestioned. But how does the utility present itself? Utility presents itself in the form of a model. If I can describe the operation of natural phenomena with a few well-chosen data elements, then I can present a simple data summarization—a model built from data—that is easy to grasp and conceptualize. Knowing my minimal monthly average in my savings account provides me with a simple and readily understandable indicator of many aspects of my financial well-being. The average is a conceptual construct—built from data. It is a model of the world and, by manipulating the model in my mind, I can make intelligence guesses, or inferences, about the real world that is reflected in my model. For example, if I can plot an upward or downward trend in my account earnings rate, then I can predict with greater certainty the likely rate of return in the next period. If I double (or halve) my average savings rate, then I can make some important inferences about the state of my earning power that is behind this doubling or halving.

We use symbolic models to reflect real-world events in an ever-broadening range of areas. By manipulating the models we find out more about the
1.1 Something old, something new

As suggested previously, civilization has always tapped data to uncover meaning and to make intellectual, economic, and technical progress. Until the late 1990s, most of the data tapping and disentangling of meaning took place in a specialized research and development–oriented environment and took specialized skills to produce results. Data mining techniques were developed in scientific settings and, originally, had scientific goals and objectives in mind. The computer algorithms to perform data mining tasks were developed by statisticians and artificial intelligence researchers.

However, by the turn of the century, computer technology—and associated computer networks—had become commodity items. Just as a spreadsheet program provides sophisticated business planning functions on the desktop, so too could tools be designed to provide sophisticated statistical and artificial intelligence functions on the desktop. If statistical and numer-
ical algorithms could be harnessed to design buildings, bridges, and even nuclear weapons, so too could these same algorithms be used to build new products, better customer relationships, and, quite possibly, new forms of businesses based on intelligent and automated data mining algorithms.

To mine data you need to have access to data. It is no coincidence that data mining grew at the same time that data warehousing developed. As computer power and database capability grew through the late 1900s, it became increasingly clear that data were not simply passive receptacles, useful in performing billing or order-entry functions, but data could also be used in a more proactive role so as to provide predictive value that would be useful in guiding a business forward. This concept led to the development of computer decision support, or executive information systems (EIS). The idea was to harness growing computing power and improved graphical interfaces in order to slice and dice data in novel ways to blow away old, static reporting concepts. Slicing and dicing data—drilling down into many detailed reports or zooming up to a 10,000-foot “big-picture” view—required special ways of organizing data for decision making. This gave rise to the concept of the data warehouse.

Decision support systems and associated data warehousing created an environment that integrated data from disparate business systems. This extended traditional business reporting to support consolidated reporting across multiple sources of data, usually in an interactive, graphically-enhanced mode.

The term data warehousing was virtually unknown in 1990. Ten years later data warehousing had become a $10+ billion business annually—a business that was devoted to capturing and organizing data so as to provide a proactive analytical (versus operational) environment in which to deploy data in the service of defining and guiding business activity.

Business caught on to the same thing that science caught on to: Data capture experience and, appropriately treated, can provide lots of ammunition to win competitive battles. Data warehousing, by organizing data for analysis, provides the raw material of data organized for analysis and decision making.

The field of decision support and executive information systems continued to evolve in line with the growth of data warehousing. Decision support and executive information systems gave way to the more general concept of business intelligence (coined in 1996 by IT trend watcher, Howard Dresner of the Gartner Group). Dresner’s insight suggested that as data moved from supporting operational purposes to include analytical purposes, the analyti-
The general term business intelligence involved the concept of organizing data along various potential dimensions of analysis so that any one view of data—for example, sales results—could be cross-referenced and displayed from within any number of other potential dimensions—for example, region or product line. The ability to move up and down dimensions involved the concept of drilling down into detail or zooming up for a more general view. The ability to show variations in data along several dimensions involved the concept of slicing and dicing data along various dimensional views. This general approach became known as on-line analytical processing—that is, processing data for analytical purposes (as opposed to operational purposes). The on-line concept referred to the idea of having the analytical data continuously available. This on-line analytical processing, or OLAP, required, therefore, the existence of a data warehouse so that data were continuously available and available in a form that would support analytical, decision support tasks.

While OLAP certainly represented a major step forward, particularly when compared with old-style, static batch reports and difficult to execute ad hoc queries, the approach ran into significant limitations when faced with the multiple fields and dimensions of data that were accumulating on mainframes and servers at the turn of the century. To display four products, in five sales regions, at three discount rates, over four quarters is a $4 \times 5 \times 3 \times 4$ combinatorial problem—that is, 240 combinations of data. Add another dimension—for example, channel (direct, mail order, regional, and national distributor)—and the combinations increase to 960. Modern databases and information collection mechanisms can deliver hundreds of potential dimensions of analysis—enough to provide significant conceptual obstacles to effective OLAP processing.

The OLAP dimensional data representation is constructed so as to contain all relevant dimensions of analysis and associated summaries in easy to retrieve, preprocessed form. This preprocessing of data means that the results of a query can be presented almost instantaneously—if the dimensional combination the end users want has been reflected in the construction of the dimensional representation. What is missing in OLAP is the ability to sort through all the potential dimensions of analysis and quickly identify only the meaningful ones. A dimension is meaningful if it displays data in a fashion that leads to better business decisions—for example, in a
Something old, something new

In a regional sales report, only some regions might be under or over projected sales. So the end user might only want to look at certain regions. In order to understand the deviations of the actual figures compared with planned figures, only a few other factors in the data—for example, discount rate, product line—might turn out to be influential in explaining the sales figure deviations. In an OLAP environment, it is possible to spend a lot of time in a fruitless search for meaningful dimensions of analysis in order to correctly identify the patterns in data that can make or break a good sales analysis. Often, the relevant dimensions are not presented in the dimensional representation, because the relevancy only becomes obvious when presented in combination with other dimensions. If this is not known ahead of time, the combination is not built.

Figure 1.1 illustrates the origins of data mining.

A cohesive approach to looking through many combinations of data dimensions in order to identify useful patterns in data would be very helpful and would complement OLAP styles of analysis very well. Weeding through hundreds of competing and potentially useful dimensions of analysis and associated combinations is a job that is custom-crafted for data mining solutions. All data mining algorithms have built-in mechanisms to examine huge numbers of potential patterns in data in order to reduce the results to a simple summary report.
1.2 Microsoft’s approach to developing the right set of tools

Doing knowledge discovery by browsing through the enterprise data store is a bit like working an archaeological dig. You know that the evidence that could yield a brilliant insight on the significant drivers of the enterprise is there, but it is deeply buried, concealed by a lot of noisy data, and, if you are not careful, are impatient, or simply not on your toes, you are going to miss it.

Welcome to the dig. Given the fortunes to be made in knowledge discovery, the intrepid knowledge discovery agent would be wise to assemble a comprehensive, effective, easy to use, and efficient tool kit to execute the knowledge discovery mission. What are the components of this tool kit?

1.2.1 Discovery of relationships

One of the key features of a knowledge discovery tool is the ability to unearth relationships. These are the dependencies that are captured in the data and that reveal the patterns characterizing the operation of the phenomenon under investigation. Relationships may be subtle and multi-
Microsoft’s approach to developing the right set of tools

1.2.2 Defining the model of the world

There is a Catch-22 to knowledge discovery. The end user, acting as a discovery agent, cannot know in advance whether the relationships to be discovered are simple or complex. Ahead of time it is difficult to assess which dependencies are spurious (i.e., spring from the operation of another dependency in the data) and which dependencies are fundamental patterns that characterize the operation of the question under investigation. To escape from this Catch-22, users of the discovery tool need to employ their knowledge of the area under investigation and need to incorporate that knowledge into the search for data patterns so as to guide the search engine to identify fundamental relationships and ignore the spurious ones. Briefly, this means that the knowledge discovery agent needs to speculate on the form of the data pattern and associated data dependencies by forming a mental model of the operation of the data pattern in terms of how data elements interact to produce the effect that characterizes the question being investigated.

The mental model formed by the knowledge discovery agent will have components that match the contents of the fields of data contained in the
data store. The knowledge discovery agent has to determine how the components of the mental model relate to the contents of the data store. The fields of data in the data store contain measurements that form a trace of the phenomenon that the data store is meant to reflect. Thus, discount rate becomes a measure of purchase practice and, possibly, supply and demand. Purchase practice and supply and demand may be components of the mental model formed by the knowledge discovery agent.

The knowledge discovery tool needs to illuminate the relationship between the data store and the mental model. It needs to facilitate the construction of a mental model the data patterns can fit into. In this manner not only will the data patterns show strong effects, but these effects will support the viewpoint and understanding of the knowledge discovery agent, as revealed by the model of the phenomenon that is being employed. This, in turn, will contribute to a greater understanding of the theory of operation that is used to explain the question under examination. It will ensure that the dependencies that are revealed will be fundamental ones, not spurious ones.

1.2.3 Dealing with data quality

The tools need to be robust to deal with the reality of real-world data. The data store may not be fully suited to capture the essence of the model construct, simple or complex, that eventually is constructed. The data store may contain missing data, so the data discovery tool has to make provisions for this.

Most certainly, the data store will contain data that are measured in various ways— for example, discount rate may be captured as a percentage or it may be captured as a range (e.g., 5 percent to 15 percent). So the data discovery tool has to deal with data, like percentage, that are quantitative (i.e., range from 1 to 100 in continuous increments) or qualitative (i.e., more than 50 percent, less than 25 percent). It needs to deal with data that are very simple, such as the qualitative type of data—M for male and F for female—and data that are very complex, such as the quantitative data refractive index, which is used in the manufacturing of optical components.

The data discovery tool also has to deal with data that are stored in a variety of formats— for example, alphabetic formats (as seen in the M and F, above), numeric formats (like, percentage), and even specialized data formats, such as time and date stamps (e.g., day, month, year codes).
1.2.4 Consolidated data mining tool requirements

This perspective on data mining and knowledge discovery provides a number of criteria on the requirements for a robust, pragmatic, real-world knowledge discovery tool. These requirements can be summarized as follows:

- The tool should be able to dig into the data set to identify effects within effects; it should not automatically assume that the strongest effect is the best effect, since this effect may be spurious—that is, it may be the result of a more fundamental relationship contained within the data.

- To help develop a mental model of the phenomenon under study, the tool should provide features that enable the association of data values, and groups of data values, with corresponding components of the mental model used by the knowledge discovery agent to model and explain the theory of operation under investigation.

- The tool needs to deal with missing values in the data elements of fields of the data store.

- The tool should be able to support both qualitative measurements of concepts as well as quantitative measurements and should be able to describe the relationships between any two or more concepts regardless of whether the measurement is qualitative or quantitative.

1.3 Benefits of data mining

Data mining offers three major advantages to the enterprise:

1. It provides information about business processes, the customer, and market behavior.

2. It takes advantage of data that may already be available in operational data collections, data marts, and the data warehouse.

3. It provides patterns of behavior, reflected in data, that can drive the accumulation of business knowledge and the ability to foresee and shape future events.

Data mining, by providing more information about the market, goes to the heart of the competitive advantage enjoyed by the enterprise. In this view, the benefit that can be delivered by data mining is only constrained by the amount of novel, useful information it can deliver as a result of its analysis of the data store.
Data mining allows for additional leverage of operational data and associated data collections in data marts and data warehouses. There is no shortage of raw data in today's networked, computer-mediated world. Data mining turns this accelerating accumulation of data into an inexhaustible supply of raw materials for the generation of business advantage.

Since the range of data, and associated, revealed trends and patterns, is very large and includes marketing, sales, engineering, finance, and human resource information, the potential application of data mining, and its associated reach into all key areas of business, is broad. According to a model developed by Treacy and Wiersma, operational excellence, customer excellence, and product excellence distinguish businesses (see Figure 1.2). Since data mining can tap into data sources that reflect upon any of these three areas, it can provide business advantage in these areas.

Various data mining groups routinely deploy data mining in these three areas, as shown in the following examples:

- Customer excellence. Bank of America used data mining to identify the characteristics of customers for their high-margin home equity loan product to find likely new prospects for the offering. The associated targeting profile was so effective that it ended up identifying some new equity loan prospects who were already in the preliminary application phases of the loan.

- Operational excellence. American Express uses a worldwide operation data repository to negotiate volume discounts from suppliers and to

---

1.3 Benefits of data mining

pinpoint—and eliminate—high-cost activities (and, conversely, to identify and promote high-margin operational practices).

- Product excellence. Bell Canada is one of the many providers in the area of telecommunications that is using business intelligence and customer relationship management tools to ensure that it gets the right products to the right customers at the right time. “Delivering products to customers becomes more complicated all the time,” says Bill Comeau, senior director of database marketing services for Bell Canada International Inc. in Toronto (as quoted in the March 13, 2000, issue of Information Week). The goal is to create a unified system that can discover patterns in information and support development and marketing of products to customers.

1.3.1 Enterprise applications of data mining

A review of the applications of data mining shows that indeed, if data can be found to capture and record the operation of a given phenomenon, then the meaning of the phenomenon can be described, modeled, and predicted using data mining approaches. This means that data mining can be used to work toward customer, operational, and product excellence in both business-to-business and business-to-customer situations.

Here are just a few examples, drawn from media reports, that describe how data mining is being used today. As shown in the following list, data mining applies to the entire customer life cycle—from product conceptualization through customer acquisition, servicing, retention, and lifetime value optimization—and to many business-to-business life-cycle activities as well. In any area where we can collect data, we can use data mining to extract knowledge for competitive advantage. This clearly involves creating advantage in the three areas of enterprise excellence: products, customers, and operations (according to Treacy and Wiersma). These examples are shown in greater detail in Table 1.1.

- Customer acquisition and customer targeting
- Profitability and risk reduction
- Loyalty management and cross-selling
- Operational analysis and optimization
- Relationship marketing
- Customer attrition and churn reduction
Fraud detection
Campaign management
Business-to-business/channel, inventory, and supply chain management
Market research, product conceptualization
Product development, engineering and quality control
Sales and sales management

Table 1.1  Illustrative Data Mining Best Practices Drawn from Media Reports

| Customer acquisition and customer targeting | Data mining can reveal the trends and patterns that characterize the best customers. Data mining can also show what unique combination of characteristics tends to be associated with customer use of specific products. Knowledge of these characteristics can be used to target specific groups of people for acquisition or for specific product promotions. Firstar Bank of Milwaukee is a typical user of this kind of data mining approach to direct marketing. Firstar uses information gained from data mining to rank order customers into different groups according to propensity to purchase home equity loans, charge cards, CD's, savings, or investments. This ranking is then used to target offers to the most likely prospects. The payoff for Firstar is a fourfold improvement in response to promotions—a substantial improvement in the cost/return ratio of direct marketing campaigns. FirstUnion is the sixth largest banking company in the United States. FirstUnion now also ranks among the country's largest credit card issuers and has climbed to the number six slot of top debit card issuers. In one series of data mining–driven acquisition campaigns, the Card Products Division nearly tripled the size of its managed credit card portfolio to more than $6.5 billion in receivables. FirstUnion also developed techniques for acquisitions and prospect targeting and employs a market test strategy prior to rollout and data mining techniques to target prospects. Capital One Financial Corp., one of the largest credit card issuers in the United States, uses data mining to help sell the most appropriate of its 3,000 financial products—including secured, joint, cobranded, and college student cards—to 150 million potential prospects in its data warehouse. Capital One's data mining techniques, which use actuarial and behavioral principles, not only track the success of various mailings but also the ongoing profitability and other characteristics of the 8.6 million customers who have signed up. These capabilities helped the company pioneer a "balance transfer" strategy—offering prospects a temporarily low interest rate to move balances from competing cards—that is now a common industry feature. Data mining and other information-based strategies not only helped the firm expand from $1 billion to $12.8 billion in managed loans from 1988 to 1996 but also to win the 1996 Excellence in Technology Award from the Gartner Group. |
1.3 Benefits of data mining

Profitability and risk reduction

Profitability and risk reduction use data mining to identify the attributes of the best customers—to characterize customer characteristics through time so as to target the appropriate customer with the appropriate product at the appropriate time. Risk reduction approaches match the discovery of poor risk characteristics against customer loan applications. This may suggest that some risk management procedures are not necessary with certain customers—a profit maximization move. It may also suggest which customers require special processing.

As can be expected, financial companies are heavy users of data mining to improve profitability and reduce risk. Home Savings of America FSB, Irwindale, CA, the nation's largest savings and loan company, analyzes mortgage delinquencies, foreclosures, sales activity, and even geological trends over five years to drive risk pricing.

According to Susan Osterfeldt, senior vice president of strategic technologies at NationsBank Services Co., "We've been able to use a neural network to build models that reduce the time it takes to process loan approvals. The neural networks speed processing. A human has to do almost nothing to approve it once it goes through the model."

Loyalty management and cross-selling

Cross-selling relies on identifying new prospects based on a match of their characteristics with known characteristics of existing customers who have been and still are satisfied with a given product. Reader's Digest does analysis of cross-selling opportunities to see if a promotional activity in one area is likely to respond to needs in another area so as to meet as many customer needs as possible.

This is a cross-sell application that involves assessing the profile of likely purchasers of a product and matching that profile to other products to find similarities in the portfolio. Cross-selling and customer relationship management are treated extensively in Mastering Data Mining (Berry and Linoff, 2000) and Building Data Mining Applications for CRM (Berson, Smith, and Thearling).

Operational analysis and optimization

Operational analysis encompasses the ability to merge corporate purchasing systems to review and manage global expenditures and to detect spending anomalies. It also includes the ability to capture and analyze operational patterns in successful branch locations, so as to compare and apply lessons learned to other branches.

American Express is using a data warehouse and data mining technique to reduce unnecessary spending, leverage its global purchasing power, and standardize equipment and services in its offices worldwide. In the late 1990s, American Express began merging its worldwide purchasing system, corporate purchasing card, and corporate card databases into a single Microsoft SQL Server database. The system allows American Express to pinpoint, for example, employees who purchase computers or other capital equipment with corporate credit cards meant for travel and entertainment. It also eliminates what American Express calls "contract bypass"—purchases from vendors other than those the company has negotiated with for discounts in return for guaranteed purchase levels.

Table 1.1

<table>
<thead>
<tr>
<th>Illustrative Data Mining Best Practices Drawn from Media Reports (continued)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Profitability and risk reduction</strong></td>
</tr>
<tr>
<td><strong>Loyalty management and cross-selling</strong></td>
</tr>
<tr>
<td><strong>Operational analysis and optimization</strong></td>
</tr>
</tbody>
</table>
Chapter 1

Operational analysis and optimization

American Express uses Quest, from New York–based Information Builders, to score the best suppliers according to 24 criteria, allowing managers to perform best-fit analyses and trade-off analyses that balance competing requirements. By monitoring purchases and vendor performance, American Express can address quality, reliability, and other issues with IBM, Eastman Kodak Co., and various worldwide vendors. According to an American Express senior vice president, “Many of the paybacks from data mining, even at this early stage, will result from our increased buying power, fewer uncontrolled expenses, and improved supplier responsiveness.”

Relationship marketing

American Express has invested in a massively parallel processor, which allows it to vastly expand the profile of every customer. The company can now store every transaction. Seventy workstations at the American Express Decision Sciences Center in Phoenix, AZ, look at data about millions of AmEx card members—the stores they shop in, the places they travel to, the restaurants they’ve eaten in, and even economic conditions and weather in the areas where they live. Every month, AmEx uses that information to send out precisely aimed offers. AmEx has seen an increase of 15 percent to 20 percent in year over year card member spending in its test market and attributes much of the increase to this approach.

Customer attrition and churn reduction

Mellon Bank of Pittsburgh is using Intelligent Miner to analyze data on the bank’s existing credit card customers to characterize their behavior and predict, for example, which customers are most likely to take their business elsewhere. “We decided it was important for us to generate and manage our own attrition models,” said Peter Johnson, vice president of the Advanced Technology Group at Mellon Bank.

Fraud detection

Another strategic benefit of Capital One’s data mining capabilities is fraud detection. In 1995, for instance, Visa and MasterCard’s U.S. losses from fraud totaled $702 million. Although Capital One will not discuss its fraud detection efforts specifically, it noted that its losses from fraud declined more than 50 percent last year, in part due to its proprietary data mining tools and San Diego–based HNC Software Inc.’s Falcon, a neural network–based credit card fraud detection system.
Campaign management  IBM's DecisionEdge campaign management module is designed to help businesses personalize marketing messages and pass them to clients through direct mail, telemarketing, and face to face interactions. The product works with IBM's Intelligent Miner for Relationship Marketing.

Among the software's features is a load-management tool, which lets companies give more lucrative campaigns priority status. "If I can only put out so many calls from my call center today, I want to make sure I make the most profitable ones," said David Raab at the analyst firm Raab Associates. "This feature isn't present in many competing products," he said.

IBM's DecisionEdge campaign management module is designed to help businesses personalize marketing messages and pass them to clients through direct mail, telemarketing, and face to face interactions. The product works with IBM's Intelligent Miner for Relationship Marketing.

Among the software's features is a load-management tool, which lets companies give more lucrative campaigns priority status. "If I can only put out so many calls from my call center today, I want to make sure I make the most profitable ones," said David Raab at the analyst firm Raab Associates. "This feature isn't present in many competing products," he said.

| Business-to-business channel, inventory, and supply chain management | The Zurich Insurance Group, a global, Swiss-based insurer, uses data mining to analyze broker performance in order to increase the efficiency and effectiveness of its business-to-business channel. Its primary utility is to look at broker performance relative to past performance and to predict future performance. Supply chains and inventory management are expensive operational overheads. In terms of sales and sales forecasting price is only one differentiator. Others include product range and image, as well as the ability to identify trends and patterns ahead of the competition. A large European retailer, using a data warehouse and data mining tools, spotted an unexpected downturn in sales of computer games. This was before Christmas. The retailer canceled a large order and watched the competition stockpile unsold computer games before Christmas. Superbrugsen, a leading Danish supermarket chain, uses data mining to optimize every single product area, and product managers must therefore have as much relevant information as possible to assist them when negotiating with suppliers to obtain the best prices. Marks and Spencer use customer profiling to determine what messages to send to certain customers. In the financial services area, for example, data mining is used to determine the characteristics of customers who are most likely to respond to a credit offer. |

Table 1.1  Illustrative Data Mining Best Practices Drawn from Media Reports (continued)
Chapter 1

Market research, product conceptualization

Blue Cross/Blue Shield is one of the largest health care providers in the United States. The organization provides analysts financial, enrollment, market penetration, and provider network information. This yields enrollment, new product development, sales, market segment, and group size estimates for marketing and sales support. Located in Dallas, TX, Rapp Collins is the second largest market research organization in the United States. It provides a wide range of marketing-related services. One involves applications that measure the effectiveness of reward incentive programs. Data mining is a core technology used to identify the many factors that influence attraction to incentives.

J. D. Power and Associates, located in Augora Hills, CA, produce a monthly forecast of car and truck sales for about 300 different vehicles. Their specialty is polling the customer after the sale regarding the purchase experience and the product itself. Forecasts are driven by sales data, economic data, and data about the industry. Data mining is used to sort through these various classes of data to produce effective forecasting models.

Product development, engineering and quality control

Quality management is a significant application area for data mining. In the manufacturing area the closer that a defect is detected to the source of the defect the easier—and less costly—it is to fix. So there is a strong emphasis on measuring progress through the various steps of manufacturing in order to find problems sooner rather than later. Of course, this means huge amounts of data are generated on many, many measurement points. This is an ideal area for data mining:

- Hewlett-Packard has used data mining to sort out a perplexing problem with a color printer that periodically produced fuzzy images. It turned out the problem was in the alignment of the lenses that blended the three primary colors to produce the output. The problem was caused by variability in the glue curing process that only affected one of the lenses. Data mining was used to find which lens, under what curing circumstances, produced the fuzzy printing resolution.
- R. R. Donnelley and Sons is the largest printing company in the United States. Their printing presses include rollers that weigh several tons and spit out results at the rate of 1,000 feet per minute. The plant experienced an occasional problem with the print quality, caused by a collection of ink on the rollers called “banding.” A task force was struck to find the cause of the problem. One of the task force members, Bob Evans, used data mining to sort through thousands of fields of data related to press performance in order to find a small subset of variables that, in combination, could be used to predict the banding problem. His work is published in the February 1994 issue of IEEE Expert and the April 1997 issue of Database Programming & Design.

<table>
<thead>
<tr>
<th>Market research, product conceptualization</th>
<th>Blue Cross/Blue Shield is one of the largest health care providers in the United States. The organization provides analysts financial, enrollment, market penetration, and provider network information. This yields enrollment, new product development, sales, market segment, and group size estimates for marketing and sales support. Located in Dallas, TX, Rapp Collins is the second largest market research organization in the United States. It provides a wide range of marketing-related services. One involves applications that measure the effectiveness of reward incentive programs. Data mining is a core technology used to identify the many factors that influence attraction to incentives. J. D. Power and Associates, located in Augora Hills, CA, produce a monthly forecast of car and truck sales for about 300 different vehicles. Their specialty is polling the customer after the sale regarding the purchase experience and the product itself. Forecasts are driven by sales data, economic data, and data about the industry. Data mining is used to sort through these various classes of data to produce effective forecasting models.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Product development, engineering and quality control</td>
<td>Quality management is a significant application area for data mining. In the manufacturing area the closer that a defect is detected to the source of the defect the easier—and less costly—it is to fix. So there is a strong emphasis on measuring progress through the various steps of manufacturing in order to find problems sooner rather than later. Of course, this means huge amounts of data are generated on many, many measurement points. This is an ideal area for data mining: Hewlett-Packard has used data mining to sort out a perplexing problem with a color printer that periodically produced fuzzy images. It turned out the problem was in the alignment of the lenses that blended the three primary colors to produce the output. The problem was caused by variability in the glue curing process that only affected one of the lenses. Data mining was used to find which lens, under what curing circumstances, produced the fuzzy printing resolution. R. R. Donnelley and Sons is the largest printing company in the United States. Their printing presses include rollers that weigh several tons and spit out results at the rate of 1,000 feet per minute. The plant experienced an occasional problem with the print quality, caused by a collection of ink on the rollers called “banding.” A task force was struck to find the cause of the problem. One of the task force members, Bob Evans, used data mining to sort through thousands of fields of data related to press performance in order to find a small subset of variables that, in combination, could be used to predict the banding problem. His work is published in the February 1994 issue of IEEE Expert and the April 1997 issue of Database Programming &amp; Design.</td>
</tr>
<tr>
<td>Table 1.1</td>
<td>Illustrative Data Mining Best Practices Drawn from Media Reports (continued)</td>
</tr>
</tbody>
</table>
1.4 Microsoft’s entry into data mining

Obviously, data mining is not just a back-room, scientific type of activity anymore. Just as document preparation software and row/column-oriented workbooks make publishers and business planners of us all, so too are we sitting on the threshold of a movement that will bring data mining—integrated with OLAP—to the desktop. What is the Microsoft strategy to achieve this?

Microsoft is setting out to solve three perceived problems:

1. Data mining tools are too expensive.
2. Data mining tools are not integrated with the underlying database.
3. Data mining algorithms, in general, reflect their scientific roots and, while they work well with small collections of data, do not scale well with the large gigabyte- and terabyte-size databases of today’s business environment.

Microsoft’s strategy to address these problems revolves around three thrusts:

1. Accessibility. Make complex data operations accessible and available to nonprofessionals, by generalizing the accessibility and lowering the cost.
2. Seamless reporting. Promote access and usability by providing a common data reporting paradigm through simple to complex business queries.
3. Scalability. To ensure access to data operations across increasingly large collections of data, provide an integration layer between the data mining algorithms and the underlying database.

Integration with the database engine occurs in three ways:

1. Preprocessing functionality is done in the database, thus providing native database access to sophisticated and heretofore specialized data cleaning, transforming, and preparation facilities.
2. Provide a core set of data mining algorithms directly in the database and provide a broadly accessible application programming interface (API) to ensure easy integration of external data mining algorithms.
3. Provide a deployment mechanism to ensure that modeling results can be readily built into other applications—both on the server and on the desktop—and to break down business process barriers to effective data mining results utilization.

Figure 1.3 shows the development of the current Microsoft architectural approach to data mining, as Microsoft migrated from the SQL Server 7 release to the SQL Server 2000 release.

One message from this figure is that data mining, as with OLAP and ad hoc reports before it, is just another query function—albeit a rather super query. Whereas in the past an end user might ask for a sales by region report, in the Microsoft world of data mining the query now becomes: Show me the main factors that were driving my sales results last period. In this way, one query can trigger millions—even trillions—of pattern matching and search operations to find the optimal results. Often many results will be produced for the reader to view. However, before long, many reader models of the world will be solicited and presented—all in template style—so that more and more preprocessing will take place to ensure that the appropriate results are presented for display (and to cut down on the amount of pattern searching and time required to respond to a query).

1.5 Concept of operations

As can be seen in Figure 1.3, the data mining component belongs to the DB query engine (DMX expressions). With the growth—depth and breadth—of data sources, it is clear that data mining algorithmic work belongs on the
server (shown in the figure as Commerce Server). We can also see that the core data mining algorithms include segmentation capabilities and associated description and prediction facilities and cross-selling components. This particular thrust has a decidedly e-commerce orientation, since cross-sell, prediction, and segmentation are important e-commerce customer relationship management functions.

Whatever algorithms are not provided on board will be provided through a common API, which extends the OLE DB for data access convention to include data mining extensions.

The Socrates project, formed to develop the Microsoft approach to data mining, is a successor to the Plato Group (the group that built the Microsoft OLAP services SQL Server 7 functionality). Together with the Database Research Group, they are working on data mining concepts for the future. Current projects this group is looking at include the following:

- It is normal to view the database or data warehouse as a data snapshot, frozen in time (the last quarter, last reporting period, and so on). Data change through time, however, and this change requires the mining algorithms to look at sequential data and patterns.

- Most of the world's data are not contained as structured data but as relatively unstructured text. In order to harvest the knowledge contained in this source of data, text mining is required.

- There are many alternative ways of producing segmentations. One of the most popular is K-means clustering. Microsoft is also exploring other methods—based on expectation maximization—that will provide more reliable clusters than the popular K-means algorithms.

- The problem of scaling algorithms to apply data mining to large databases is a continuing effort. One area—sufficiency statistics—seeks to find optimal ways of computing the necessary pattern-matching rules so that the rules that are discovered are reliable across the entire large collection of data.

- Research is underway on a general data mining query language (DMQL). This is to devise general methods within the DBMS query language to form data mining queries. Current development efforts focus on SQL operators Unipivot and DataCube.

- There are continuing efforts regarding OLAP refinements in the direction of data mining to continue integration of OLAP and data mining.
A promising area of data mining is to define methods and procedures to continue to automate more and more of the searching that is undertaken automatically. This area of metarule-guided mining is a continuing effort in the Socrates project.