**An Introduction to Data Mining**

**Overview**

Data mining, *the extraction of hidden predictive information from large databases*, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems. Data mining tools can answer business questions that traditionally were too time consuming to resolve. They scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations.

Most companies already collect and refine massive quantities of data. Data mining techniques can be implemented rapidly on existing software and hardware platforms to enhance the value of existing information resources, and can be integrated with new products and systems as they are brought on-line. When implemented on high performance client/server or parallel processing computers, data mining tools can analyze massive databases to deliver answers to questions such as, "Which clients are most likely to respond to my next promotional mailing, and why?"

This white paper provides an introduction to the basic technologies of data mining. Examples of profitable applications illustrate its relevance to today’s business environment as well as a basic description of how data warehouse architectures can evolve to deliver the value of data mining to end users.

**The Foundations of Data Mining**

Data mining techniques are the result of a long process of research and product development. This evolution began when business data was first stored on computers, continued with improvements in data access, and more recently, generated technologies that allow users to navigate through their data in real time. Data mining takes this evolutionary process beyond retrospective data access and navigation to prospective and proactive information delivery. Data mining is ready for application in the business community because it is supported by three technologies that are now sufficiently mature:

* Massive data collection
* Powerful multiprocessor computers
* Data mining algorithms

Commercial databases are growing at unprecedented rates. A recent META Group survey of data warehouse projects found that 19% of respondents are beyond the 50 gigabyte level, while 59% expect to be there by second quarter of 1996.1 In some industries, such as retail, these numbers can be much larger. The accompanying need for improved computational engines can now be met in a cost-effective manner with parallel multiprocessor computer technology. Data mining algorithms embody techniques that have existed for at least 10 years, but have only recently been implemented as mature, reliable, understandable tools that consistently outperform older statistical methods.

In the evolution from business data to business information, each new step has built upon the previous one. For example, dynamic data access is critical for drill-through in data navigation applications, and the ability to store large databases is critical to

data mining. From the user’s point of view, the four steps listed in Table 1 were revolutionary because they allowed new business questions to be answered accurately and quickly.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Evolutionary Step** | **Business Question** | **Enabling Technologies** | **Product Providers** | **Characteristics** |
| Data Collection  (1960s) | "What was my total revenue in the last five years?" | Computers, tapes, disks | IBM, CDC | Retrospective, static data delivery |
| Data Access  (1980s) | "What were unit sales in New England last March?" | Relational databases (RDBMS), Structured Query Language (SQL), ODBC | Oracle, Sybase, Informix, IBM, Microsoft | Retrospective, dynamic data delivery at record level |
| Data Warehousing &  Decision Support  (1990s) | "What were unit sales in New England last March? Drill down to Boston." | On-line analytic processing (OLAP), multidimensional databases, data warehouses | Pilot, Comshare, Arbor, Cognos, Microstrategy | Retrospective, dynamic data delivery at multiple levels |
| Data Mining  (Emerging Today) | "What’s likely to happen to Boston unit sales next month? Why?" | Advanced algorithms, multiprocessor computers, massive databases | Pilot, Lockheed, IBM, SGI, numerous startups (nascent industry) | Prospective, proactive information delivery |

Table 1. Steps in the Evolution of Data Mining.

The core components of data mining technology have been under development for decades, in research areas such as statistics, artificial intelligence, and machine learning. Today, the maturity of these techniques, coupled with high-performance relational database engines and broad data integration efforts, make these technologies practical for current data warehouse environments.

**The Scope of Data Mining**

Data mining derives its name from the similarities between searching for valuable business information in a large database — for example, finding linked products in gigabytes of store scanner data — and mining a mountain for a vein of valuable ore. Both processes require either sifting through an immense amount of material, or intelligently probing it to find exactly where the value resides. Given databases of sufficient size and quality, data mining technology can generate new business opportunities by providing these capabilities:

Automated prediction of trends and behaviors. Data mining automates the process of finding predictive information in large databases. Questions that traditionally required extensive hands-on analysis can now be answered directly from the data — quickly. A typical example of a predictive problem is targeted marketing. Data mining uses data on past promotional mailings to identify the targets most likely to maximize return on

investment in future mailings. Other predictive problems include forecasting bankruptcy and other forms of default, and identifying segments of a population likely to respond similarly to given events.

Automated discovery of previously unknown patterns. Data mining tools sweep through databases and identify previously hidden patterns in one step. An example of pattern discovery is the analysis of retail sales data to identify seemingly unrelated products that are often purchased together. Other pattern discovery problems include detecting fraudulent credit card transactions and identifying anomalous data that could represent data entry keying errors.

Data mining techniques can yield the benefits of automation on existing software and

hardware platforms, and can be implemented on new systems as existing platforms are upgraded and new products developed. When data mining tools are implemented on high performance parallel processing systems, they can analyze massive databases in minutes. Faster processing means that users can automatically experiment with more models to understand complex data. High speed makes it practical for users to analyze huge quantities of data. Larger databases, in turn, yield improved predictions.

Databases can be larger in both depth and breadth:

* **More columns**. Analysts must often limit the number of variables they examine when doing hands-on analysis due to time constraints. Yet variables that are discarded because they seem unimportant may carry information about unknown patterns. High performance data mining allows users to explore the full depth of a database, without preselecting a subset of variables.
* **More rows**. Larger samples yield lower estimation errors and variance, and allow users to make inferences about small but important segments of a population.

A recent Gartner Group Advanced Technology Research Note listed data mining and artificial intelligence at the top of the five key technology areas that "will clearly have a major impact across a wide range of industries within the next 3 to 5 years."2 Gartner also listed parallel architectures and data mining as two of the top 10 new technologies in which companies will invest during the next 5 years. According to a recent Gartner HPC Research Note, "With the rapid advance in data capture, transmission and storage, large-systems users will increasingly need to implement new and innovative ways to mine the after-market value of their vast stores of detail data, employing MPP [massively parallel processing] systems to create new sources of business advantage (0.9 probability)."3

The most commonly used techniques in data mining are:

**Artificial neural networks**: Non-linear predictive models that learn through training and resemble biological neural networks in structure.

Decision trees: Tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset. Specific decision tree methods include Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID) .

Genetic algorithms: Optimization techniques that use processes such as genetic combination, mutation, and natural selection in a design based on the concepts of evolution.

* **Nearest neighbor method**: A technique that classifies each record in a dataset based on a combination of the classes of the k record(s) most similar to it in a historical dataset (where k ³ 1). Sometimes called the k-nearest neighbor technique.
* **Rule induction**: The extraction of useful if-then rules from data based on statistical significance.

Many of these technologies have been in use for more than a decade in specialized analysis tools that work with relatively small volumes of data. These capabilities are now evolving to integrate directly with industry-standard data warehouse and OLAP platforms. The appendix to this white paper provides a glossary of data mining terms.

**How Data Mining Works**

How exactly is data mining able to tell you important things that you didn't know or what is going to happen next? The technique that is used to perform these feats in data mining is called modeling. Modeling is simply the act of building a model in one situation where you know the answer and then applying it to another situation that you don't. For instance, if you were looking for a sunken Spanish galleon on the high seas the first thing you might do is to research the times when Spanish treasure had been found by others in the past. You might note that these ships often tend to be found off the coast of Bermuda and that there are certain characteristics to the ocean currents,

and certain routes that have likely been taken by the ship’s captains in that era. You note these similarities and build a model that includes the characteristics that are common to the locations of these sunken treasures. With these models in hand you sail off looking for treasure where your model indicates it most likely might be given a similar situation in the past. Hopefully, if you've got a good model, you find your treasure.

This act of model building is thus something that people have been doing for a long time, certainly before the advent of computers or data mining technology. What happens on computers, however, is not much different than the way people build models. Computers are loaded up with lots of information about a variety of situations where an answer is known and then the data mining software on the computer must run through that data and distill the characteristics of the data that should go into the model. Once the model is built it can then be used in similar situations where you don't know the answer. For example, say that you are the director of marketing for a telecommunications company and you'd like to acquire some new long distance phone customers. You could just randomly go out and mail coupons to the general population - just as you could randomly sail the seas looking for sunken treasure. In neither case would you achieve the results you desired and of course you have the opportunity to do much better than random - you could use your business experience stored in your database to build a model.

As the marketing director you have access to a lot of information about all of your customers: their age, sex, credit history and long distance calling usage. The good news is that you also have a lot of information about your prospective customers: their age, sex, credit history etc. Your problem is that you don't know the long distance calling usage of these prospects (since they are most likely now customers of your competition). You'd like to concentrate on those prospects who have large amounts of long distance usage. You can accomplish this by building a model. Table 2 illustrates the data used for building a model for new customer prospecting in a data warehouse.

|  |  |  |
| --- | --- | --- |
|  | Customers | Prospects |
| General information (e.g. demographic data) | Known | Known |
| Proprietary information (e.g. customer transactions) | Known | Target |

Table 2 - Data Mining for Prospecting

The goal in prospecting is to make some calculated guesses about the information in the lower right hand quadrant based on the model that we build going from Customer General Information to Customer Proprietary Information. For instance, a simple model for a telecommunications company might be:

98% of my customers who make more than $60,000/year spend more than $80/month on long distance

This model could then be applied to the prospect data to try to tell something about the proprietary information that this telecommunications company does not currently have access to. With this model in hand new customers can be selectively targeted.

Test marketing is an excellent source of data for this kind of modeling. Mining the results of a test market representing a broad but relatively small sample of prospects can provide a foundation for identifying good prospects in the overall market. Table 3 shows another common scenario for building models: predict what is going to happen in the future.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Yesterday | Today | Tomorrow |
| Static information and current plans (e.g. demographic data, marketing plans) | Known | Known | Known |
| Dynamic information (e.g. customer transactions) | Known | Known | Target |

Table 3 - Data Mining for Predictions

If someone told you that he had a model that could predict customer usage how would you know if he really had a good model? The first thing you might try would be to ask him to apply his model to your customer base - where you already knew the answer. With data mining, the best way to accomplish this is by setting aside some of your data in a vault to isolate it from the mining process. Once the mining is complete, the results can be tested against the data held in the vault to confirm the model’s validity. If the model works, its observations should hold for the vaulted data.

**An Architecture for Data Mining**

To best apply these advanced techniques, they must be fully integrated with a data warehouse as well as flexible interactive business analysis tools. Many data mining tools currently operate outside of the warehouse, requiring extra steps for extracting, importing, and analyzing the data. Furthermore, when new insights require operational implementation, integration with the warehouse simplifies the application of results from data mining. The resulting analytic data warehouse can be applied to improve business processes throughout the organization, in areas such as promotional campaign management, fraud detection, new product rollout, and so on. Figure 1 illustrates an architecture for advanced analysis in a large data warehouse. 

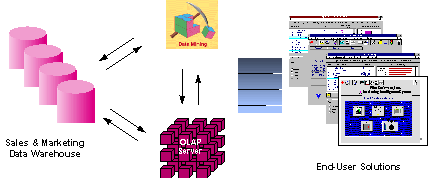


Figure 1 - Integrated Data Mining Architecture

The ideal starting point is a data warehouse containing a combination of internal data tracking all customer contact coupled with external market data about competitor activity. Background information on potential customers also provides an excellent basis for prospecting. This warehouse can be implemented in a variety of relational database systems: Sybase, Oracle, Redbrick, and so on, and should be optimized for flexible and fast data access.

An OLAP (On-Line Analytical Processing) server enables a more sophisticated end-user business model to be applied when navigating the data warehouse. The multidimensional structures allow the user to analyze the data as they want to view their business – summarizing by product line, region, and other key perspectives of their business. The Data Mining Server must be integrated with the data warehouse and the OLAP server to embed ROI-focused business analysis directly into this infrastructure. An advanced, process-centric metadata template defines the data mining objectives for specific business issues like campaign management, prospecting, and promotion optimization. Integration with the data warehouse enables

operational decisions to be directly implemented and tracked. As the warehouse grows with new decisions and results, the organization can continually mine the best practices and apply them to future decisions.

This design represents a fundamental shift from conventional decision support systems. Rather than simply delivering data to the end user through query and reporting software, the Advanced Analysis Server applies users’ business models directly to the warehouse and returns a proactive analysis of the most relevant information. These results enhance the metadata in the OLAP Server by providing a dynamic metadata layer that represents a distilled view of the data. Reporting, visualization, and other analysis tools can then be applied to plan future actions and confirm the impact of those plans.

**Profitable Applications**

A wide range of companies have deployed successful applications of data mining. While early adopters of this technology have tended to be in information-intensive industries such as financial services and direct mail marketing, the technology is applicable to any company looking to leverage a large data warehouse to better manage their customer relationships. Two critical factors for success with data mining are: a large, well-integrated data warehouse and a well-defined understanding of the business process within which data mining is to be applied (such as customer prospecting, retention, campaign management, and so on).

Some successful application areas include:

A pharmaceutical company can analyze its recent sales force activity and their results to improve targeting of high-value physicians and determine which marketing activities will have the greatest impact in the next few months. The data needs to include competitor market activity as well as information about the local health care systems. The results can be distributed to the sales force via a wide-area network that enables the representatives to review the recommendations from the perspective of the key attributes in the decision process. The ongoing, dynamic analysis of the data warehouse allows best practices from throughout the organization to be applied in specific sales situations.

A credit card company can leverage its vast warehouse of customer transaction data to identify customers most likely to be interested in a new credit product. Using a small test mailing, the attributes of customers with an affinity for the product can be identified. Recent projects have indicated more than a 20-fold decrease in costs for targeted mailing campaigns over conventional approaches.

A diversified transportation company with a large direct sales force can apply data mining to identify the best prospects for its services. Using data mining to analyze its own customer experience, this company can build a unique segmentation identifying the attributes of high-value prospects. Applying this segmentation to a general business database such as those provided by Dun & Bradstreet can yield a prioritized list of prospects by region.

A large consumer package goods company can apply data mining to improve its sales process to retailers. Data from consumer panels, shipments, and competitor activity can be applied to understand the reasons for brand and store switching. Through this analysis, the manufacturer can select promotional strategies that best reach their target customer segments.

Each of these examples have a clear common ground. They leverage the knowledge about customers implicit in a data warehouse to reduce costs and improve the value of customer relationships. These organizations can now focus their efforts on the most important (profitable) customers and prospects, and design targeted marketing strategies to best reach them.

**Conclusion**

Comprehensive data warehouses that integrate operational data with customer, supplier, and market information have resulted in an explosion of information. Competition requires timely and sophisticated analysis on an integrated view of the data. However, there is a growing gap between more powerful storage and retrieval systems and the users’ ability to effectively analyze and act on the information they contain. Both relational and OLAP technologies have tremendous capabilities for navigating massive data warehouses, but brute force navigation of data is not enough. A new technological leap is needed to structure and prioritize information for specific end-user problems. The data mining tools can make this leap. Quantifiable business benefits have been proven through the integration of data mining with current information systems, and new products are on the horizon that will bring this integration to an even wider audience of users.

**Glossary of Data Mining Terms**

|  |  |
| --- | --- |
| analytical model | A structure and process for analyzing a dataset. For example, a decision tree is a model for the classification of a dataset. |
| anomalous data | Data that result from errors (for example, data entry keying errors) or that represent unusual events. Anomalous data should be examined carefully because it may carry important information. |
| artificial neural networks | Non-linear predictive models that learn through training and resemble biological neural networks in structure. |
| CART | Classification and Regression Trees. A decision tree technique used for classification of a dataset. Provides a set of rules that you can apply to a new (unclassified) dataset to predict which records will have a given outcome. Segments a dataset by creating 2-way splits. Requires less data preparation than CHAID. |
| CHAID | Chi Square Automatic Interaction Detection. A decision tree technique used for classification of a dataset. Provides a set of rules that you can apply to a new (unclassified) dataset to predict which records will have a given outcome. Segments a dataset by using chi square tests to create multi-way splits. Preceded, and requires more data preparation than, CART. |
| classification | The process of dividing a dataset into mutually exclusive groups such that the members of each group are as "close" as possible to one another, and different groups are as "far" as possible from one another, where distance is measured with respect to specific variable(s) you are trying to predict. For example, a typical classification problem is to divide a database of companies into groups that are as homogeneous as possible with respect to a creditworthiness variable with values "Good" and "Bad." |
| clustering | The process of dividing a dataset into mutually exclusive groups such that the members of each group are as "close" as possible to one another, and different groups are as "far" as possible from one another, where distance is measured with respect to all available variables. |
| data cleansing | The process of ensuring that all values in a dataset are consistent and correctly recorded. |
| data mining | The extraction of hidden predictive information from large databases. |
| data navigation | The process of viewing different dimensions, slices, and levels of detail of a multidimensional database. See OLAP. |
| data visualization | The visual interpretation of complex relationships in multidimensional data. |
| data warehouse | A system for storing and delivering massive quantities of data. |
| decision tree | A tree-shaped structure that represents a set of decisions. These decisions generate rules for the classification of a dataset. See CART and CHAID. |
| dimension | In a flat or relational database, each field in a record represents a dimension. In a multidimensional database, a dimension is a set of similar entities; for example, a multidimensional sales database might include the dimensions Product, Time, and City. |
| exploratory data analysis | The use of graphical and descriptive statistical techniques to learn about the structure of a dataset. |
| genetic algorithms | Optimization techniques that use processes such as genetic combination, mutation, and natural selection in a design based on the concepts of natural evolution. |
| linear model | An analytical model that assumes linear relationships in the coefficients of the variables being studied. |
| linear regression | A statistical technique used to find the best-fitting linear relationship between a target (dependent) variable and its predictors (independent variables). |
| logistic regression | A linear regression that predicts the proportions of a categorical target variable, such as type of customer, in a population. |
| multidimensional database | A database designed for on-line analytical processing. Structured as a multidimensional hypercube with one axis per dimension. |
| multiprocessor computer | A computer that includes multiple processors connected by a network. See parallel processing. |
| nearest neighbor | A technique that classifies each record in a dataset based on a combination of the classes of the k record(s) most similar to it in a historical dataset (where k ³ 1). Sometimes called a k-nearest neighbor technique. |
| non-linear model | An analytical model that does not assume linear relationships in the coefficients of the variables being studied. |
| OLAP | On-line analytical processing. Refers to array-oriented database applications that allow users to view, navigate through, manipulate, and analyze multidimensional databases. |
| outlier | A data item whose value falls outside the bounds enclosing most of the other corresponding values in the sample. May indicate anomalous data. Should be examined carefully; may carry important information. |
| parallel processing | The coordinated use of multiple processors to perform computational tasks. Parallel processing can occur on a multiprocessor computer or on a network of workstations or PCs. |
| predictive model | A structure and process for predicting the values of specified variables in a dataset. |
| prospective data analysis | Data analysis that predicts future trends, behaviors, or events based on historical data. |
| RAID | Redundant Array of Inexpensive Disks. A technology for the efficient parallel storage of data for high-performance computer systems. |
| retrospective data analysis | Data analysis that provides insights into trends, behaviors, or events that have already occurred. |
| rule induction | The extraction of useful if-then rules from data based on statistical significance. |
| SMP | Symmetric multiprocessor. A type of multiprocessor computer in which memory is shared among the processors. |
| terabyte | One trillion bytes. |
| time series analysis | The analysis of a sequence of measurements made at specified time intervals. Time is usually the dominating dimension of the data. |

**What Is Data Mining?**

Data mining is the practice of automatically searching large stores of data to discover patterns and trends that go beyond simple analysis. Data mining uses sophisticated mathematical algorithms to segment the data and evaluate the probability of future events. Data mining is also known as Knowledge Discovery in Data (KDD).

The key properties of data mining are:

* Automatic discovery of patterns
* Prediction of likely outcomes
* Creation of actionable information
* Focus on large data sets and databases

Data mining can answer questions that cannot be addressed through simple query and reporting techniques.

**Automatic Discovery**

Data mining is accomplished by building models. A model uses an algorithm to act on a set of data. The notion of automatic discovery refers to the execution of data mining models.

Data mining models can be used to mine the data on which they are built, but most types of models are generalizable to new data. The process of applying a model to new data is known as scoring.

Oracle Data Mining Application Developer's Guide for a discussion of scoring and deployment in Oracle Data Mining

Prediction

Many forms of data mining are predictive. For example, a model might predict income based on education and other demographic factors. Predictions have an associated probability (How likely is this prediction to be true?). Prediction probabilities are also known as confidence (How confident can I be of this prediction?).

Some forms of predictive data mining generate rules, which are conditions that imply a given outcome. For example, a rule might specify that a person who has a bachelor's degree and lives in a certain neighborhood is likely to have an income greater than the regional average. Rules have an associated support (What percentage of the population satisfies the rule?).

**Grouping**

Other forms of data mining identify natural groupings in the data. For example, a model might identify the segment of the population that has an income within a specified range, that has a good driving record, and that leases a new car on a yearly basis.

**Actionable Information**

Data mining can derive actionable information from large volumes of data. For example, a town planner might use a model that predicts income based on demographics to develop a plan for low-income housing. A car leasing agency might a use model that identifies customer segments to design a promotion targeting high-value customers.

"Data Mining Functions" for an overview of predictive and descriptive data mining. A general introduction to algorithms is provided in "Data Mining Algorithms".

Data Mining and Statistics

There is a great deal of overlap between data mining and statistics. In fact most of the techniques used in data mining can be placed in a statistical framework. However, data mining techniques are not the same as traditional statistical techniques.

Traditional statistical methods, in general, require a great deal of user interaction in order to validate the correctness of a model. As a result, statistical methods can be difficult to automate. Moreover, statistical methods typically do not scale well to very large data sets. Statistical methods rely on testing hypotheses or finding correlations based on smaller, representative samples of a larger population.

Data mining methods are suitable for large data sets and can be more readily automated. In fact, data mining algorithms often require large data sets for the creation of quality models.

**Data Mining and OLAP**

On-Line Analytical Processing (OLAP) can been defined as fast analysis of shared multidimensional data. OLAP and data mining are different but complementary activities.

**OLAP** supports activities such as data summarization, cost allocation, time series analysis, and what-if analysis. However, most OLAP systems do not have inductive inference capabilities beyond the support for time-series forecast. Inductive inference, the process of reaching a general conclusion from specific examples, is a characteristic of data mining. Inductive inference is also known as computational learning.

OLAP systems provide a multidimensional view of the data, including full support for hierarchies. This view of the data is a natural way to analyze businesses and organizations. Data mining, on the other hand, usually does not have a concept of dimensions and hierarchies.

Data mining and OLAP can be integrated in a number of ways. For example, data mining can be used to select the dimensions for a cube, create new values for a dimension, or create new measures for a cube. OLAP can be used to analyze data mining results at different levels of granularity.

Data Mining can help you construct more interesting and useful cubes. For example, the results of predictive data mining could be added as custom measures to a cube. Such measures might provide information such as "likely to default" or "likely to buy" for each customer. OLAP processing could then aggregate and summarize the probabilities.

**Data Mining and Data Warehousing**

Data can be mined whether it is stored in flat files, spreadsheets, database tables, or some other storage format. The important criteria for the data is not the storage format, but its applicability to the problem to be solved.

Proper data cleansing and preparation are very important for data mining, and a data warehouse can facilitate these activities. However, a data warehouse will be of no use if it does not contain the data you need to solve your problem.

Oracle Data Mining requires that the data be presented as a case table in single-record case format. All the data for each record (case) must be contained within a row. Most typically, the case table is a view that presents the data in the required format for mining.

**What Can Data Mining Do and Not Do?**

Data mining is a powerful tool that can help you find patterns and relationships within your data. But data mining does not work by itself. It does not eliminate the need to know your business, to understand your data, or to understand analytical methods. Data mining discovers hidden information in your data, but it cannot tell you the value of the information to your organization.

You might already be aware of important patterns as a result of working with your data over time. Data mining can confirm or qualify such empirical observations in addition to finding new patterns that may not be immediately discernible through simple observation.

It is important to remember that the predictive relationships discovered through data mining are not necessarily causes of an action or behavior. For example, data mining might determine that males with incomes between $50,000 and $65,000 who subscribe to certain magazines are likely to buy a given product. You can use this information to help you develop a marketing strategy. However, you should not assume that the population identified through data mining will buy the product because they belong to this population.

**Asking the Right Questions**

Data mining does not automatically discover solutions without guidance. The patterns you find through data mining will be very different depending on how you formulate the problem.

To obtain meaningful results, you must learn how to ask the right questions. For example, rather than trying to learn how to "improve the response to a direct mail solicitation," you might try to find the characteristics of people who have responded to your solicitations in the past.

**Understanding Your Data**

To ensure meaningful data mining results, you must understand your data. Data mining algorithms are often sensitive to specific characteristics of the data: outliers (data values that are very different from the typical values in your database), irrelevant columns, columns that vary together (such as age and date of birth), data coding, and data that you choose to include or exclude. Oracle Data Mining can automatically perform much of the data preparation required by the algorithm. But some of the data preparation is typically specific to the domain or the data mining problem. At any rate, you need to understand the data that was used to build the model in order to properly interpret the results when the model is applied.

**The Data Mining Process**

Figure 1-1 illustrates the phases, and the iterative nature, of a data mining project. The process flow shows that a data mining project does not stop when a particular solution is deployed. The results of data mining trigger new business questions, which in turn can be used to develop more focused models.

**Problem Definition**

This initial phase of a data mining project focuses on understanding the project objectives and requirements. Once you have specified the project from a business perspective, you can formulate it as a data mining problem and develop a preliminary implementation plan.

For example, your business problem might be: "How can I sell more of my product to customers?" You might translate this into a data mining problem such as: "Which customers are most likely to purchase the product?" A model that predicts who is most likely to purchase the product must be built on data that describes the customers who have purchased the product in the past. Before building the model, you must assemble the data that is likely to contain relationships between customers who have purchased the product and customers who have not purchased the product. Customer attributes might include age, number of children, years of residence, owners/renters, and so on.

**Data Gathering and Preparation**

The data understanding phase involves data collection and exploration. As you take a closer look at the data, you can determine how well it addresses the business problem. You might decide to remove some of the data or add additional data. This is also the time to identify data quality problems and to scan for patterns in the data.

The data preparation phase covers all the tasks involved in creating the case table you will use to build the model. Data preparation tasks are likely to be performed multiple times, and not in any prescribed order. Tasks include table, case, and attribute selection as well as data cleansing and transformation. For example, you might transform a DATE\_OF\_BIRTH column to AGE; you might insert the average income in cases where the INCOME column is null.

Additionally you might add new computed attributes in an effort to tease information closer to the surface of the data. For example, rather than using the purchase amount, you might create a new attribute: "Number of Times Amount Purchase Exceeds $500 in a 12 month time period." Customers who frequently make large purchases may also be related to customers who respond or don't respond to an offer.

Thoughtful data preparation can significantly improve the information that can be discovered through data mining.

**Model Building and Evaluation**

In this phase, you select and apply various modeling techniques and calibrate the parameters to optimal values. If the algorithm requires data transformations, you will need to step back to the previous phase to implement them (unless you are using Oracle Automatic Data Preparation, as described in Chapter 19).

In preliminary model building, it often makes sense to work with a reduced set of data (fewer rows in the case table), since the final case table might contain thousands or millions of cases.

At this stage of the project, it is time to evaluate how well the model satisfies the originally-stated business goal (phase 1). If the model is supposed to predict customers who are likely to purchase a product, does it sufficiently differentiate between the two classes? Is there sufficient lift? Are the trade-offs shown in the confusion matrix acceptable? Would the model be improved by adding text data? Should transactional data such as purchases (market-basket data) be included? Should

costs associated with false positives or false negatives be incorporated into the model?

**Knowledge Deployment**

Knowledge deployment is the use of data mining within a target environment. In the deployment phase, insight and actionable information can be derived from data.

Deployment can involve scoring (the application of models to new data), the extraction of model details (for example the rules of a decision tree), or the integration of data mining models within applications, data warehouse infrastructure, or query and reporting tools.

Because Oracle Data Mining builds and applies data mining models inside Oracle Database, the results are immediately available. BI reporting tools and dashboards can easily display the results of data mining. Additionally, Oracle Data Mining supports scoring in real time: Data can be mined and the results returned within a single database transaction. For example, a sales representative could run a model that predicts the likelihood of fraud within the context of an online sales transaction.

**Data Mining Process Steps:**

Below mentioned are steps to identify hidden pattern in data in data mining process-data-mining-process

**1. Data Integration:** At the very beginning we collects data from variety of sources (CRM, OLTP, Excel,MS Access, Oracle, SQL Server, csv etc) into a single data source called target data/ database using some technology.

**2. Data Selection:** In this step focus on only those data set which are required to fulfill our hypothesis/ research/ assumption it means only meaningful data is selected to proceed.

**3. Data Cleaning:** Data imported from data sources may be in different format than target database then we need to clean the data using some data cleaning algorithm.

Data Transformation: Data is then prepossessed and transformed into standard format.

**4. Data Mining:** Then we analyse and identify the type of data mining algorithm which will be suitable for current data and then applying algorithm(s) to identify the hidden patterns.

**5. Pattern Evaluation:** The pattern we identified from the data is then interpreted and evaluated to gain knowledge out of it.

Knowledge Presentation: This is the goal of data mining technique where knowledge collected from data mining process is then taken into consideration to make crucial business decision for the benefit of organization. Data Mining Process Steps:

**6. Data Integration:** At the very beginning we collects data from variety of sources (CRM, OLTP, Excel,MS Access, Oracle, SQL Server, csv etc) into a single data source called target data/ database using some technology.

**7. Pattern Evaluation:** The pattern we identified from the data is then interpreted and evaluated to gain knowledge out of it.

**8. Knowledge Presentation:**This is the goal of data mining technique where knowledge collected from data mining process is then taken into consideration to make crucial business decision for the benefit of organization.

