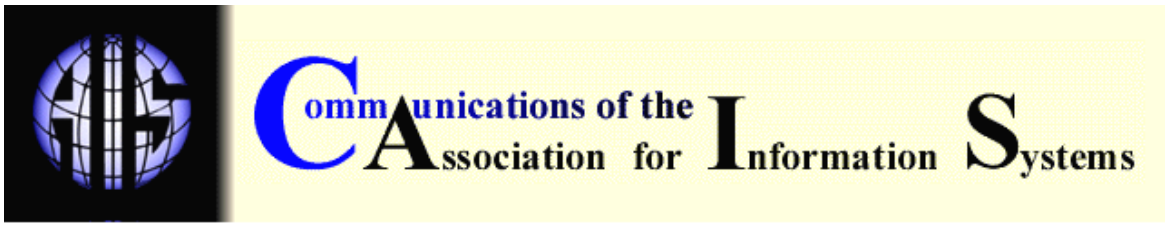


DATA MINING: A CONCEPTUAL OVERVIEW

Joyce Jackson
Management Science Department
University of South Carolina
joyce.jackson@sc.edu

ABSTRACT

This tutorial provides an overview of the data mining process. The tutorial also provides a basic understanding of how to plan, evaluate and successfully refine a data mining project, particularly in terms of model building and model evaluation. Methodological considerations are discussed and illustrated. After explaining the nature of data mining and its importance in business, the tutorial describes the underlying machine learning and statistical techniques involved. It describes the CRISP-DM standard now being used in industry as the standard for a technology-neutral data mining process model. The paper concludes with a major illustration of the data mining process methodology and the unsolved problems that offer opportunities for research. The approach is both practical and conceptually sound in order to be useful to both academics and practitioners.

Keywords: data mining, machine learning, statistics, process methodology

I. INTRODUCTION

DATA MINING

The objective of data mining is to identify valid novel, potentially useful, and understandable correlations and patterns in existing data [Chung and Gray 1999]. Finding useful patterns in data is known by different names (including data mining) in different communities (e.g., knowledge extraction, information discovery, information harvesting, data archeology, and data pattern processing) [Fayyad, et al, 1996]. The term “data mining” is primarily used by statisticians, database researchers, and the MIS and business communities. The term Knowledge Discovery in Databases (KDD) is generally used to refer to the overall process of discovering useful knowledge from data, where data mining is a particular step in this process. [Fayyad, et al, 1996; Peacock, 1998a; Han and Kamber, 2000] The additional steps in the KDD process, such as data preparation, data selection, data cleaning, and proper interpretation of the results of the data mining process, ensure that useful knowledge is derived from the data.

Data mining is an extension of traditional data analysis and statistical approaches in that it incorporates analytical techniques drawn from a range of disciplines including, but not limited to,

- numerical analysis,
- pattern matching and areas of artificial intelligence such as machine learning,
- neural networks and genetic algorithms.

While many data mining tasks follow a traditional, hypothesis-driven data analysis approach, it is commonplace to employ an opportunistic, data driven approach that encourages the pattern detection algorithms to find useful trends, patterns, and relationships.

Essentially, the two types of data mining approaches differ in whether they seek to build models or to find patterns. The first approach, concerned with building models is, apart from the problems inherent from the large sizes of the data sets, similar to conventional exploratory statistical methods. The objective is to produce an overall summary of a set of data to identify and describe the main features of the shape of the distribution [Hand 1998]. Examples of such models include a cluster analysis partition of a set of data, a regression model for prediction, and a tree-based classification rule. In model building, a distinction is sometimes made between empirical and mechanistic models [Box and Hunter 1965; Cox 1990; Hand 1995]. The former (also sometimes called operational) seeks to model relationships without basing them on any underlying theory. The latter (sometimes called substantive or phenomenological) are based on some theory or mechanism for the underlying data generating process. Data mining, almost by definition, is primarily concerned with the operational.

The second type of data mining approach, pattern detection, seeks to identify small (but nonetheless possibly important) departures from the norm, to detect unusual patterns of behavior. Examples include unusual spending patterns in credit card usage (for fraud detection), sporadic waveforms in EEG traces, and objects with patterns of characteristics unlike others. It is this class of strategies that led to the notion of data mining as seeking “nuggets” of information among the mass of data. In general, business databases pose a unique problem for pattern extraction because of their complexity. Complexity arises from anomalies such as discontinuity, noise, ambiguity, and incompleteness [Fayyad, Piatetsky-Shapiro, and Smyth, 1996]. And while most data mining algorithms are able to separate the effects of such irrelevant attributes in determining the actual pattern, the predictive power of the mining algorithms may decrease as the number of these anomalies increase [Rajagopalan and Krovi, 2002].

DATA MINING AND DATA WAREHOUSING

The construction of a data warehouse, which involves data cleaning and data integration, can be viewed as an important pre-processing step for data mining. However, a data warehouse is not a requirement for data mining. Building a large data warehouse that consolidates data from multiple sources, resolves data integrity problems, and loads the data into a database, can be an enormous task, sometimes taking years and costing millions of dollars [Gray and Watson, 1998a]. If a data warehouse is not available, the data to be mined can be extracted from one or more operational or transactional databases, or data marts. Alternatively, the data mining database could be a logical or a physical subset of a data warehouse.

Data mining uses the data warehouse as the source of information for knowledge data discovery (KDD) systems through an amalgam of artificial intelligence and statistics-related techniques to find associations, sequences, classifications, clusters, and forecasts [Gray and Watson, 1998b]. Figures 1 and 2 illustrate this process.

As shown in Figure 1, almost all data enter the warehouse from the operational environment. The data are then "cleaned" and moved into the warehouse.

The data continue to reside in the warehouse until they reach an age where one of three actions is taken: the data are purged; the data, together with other information, are summarized; or the data are archived. An aging process inside the warehouse moves current data into old detail data.

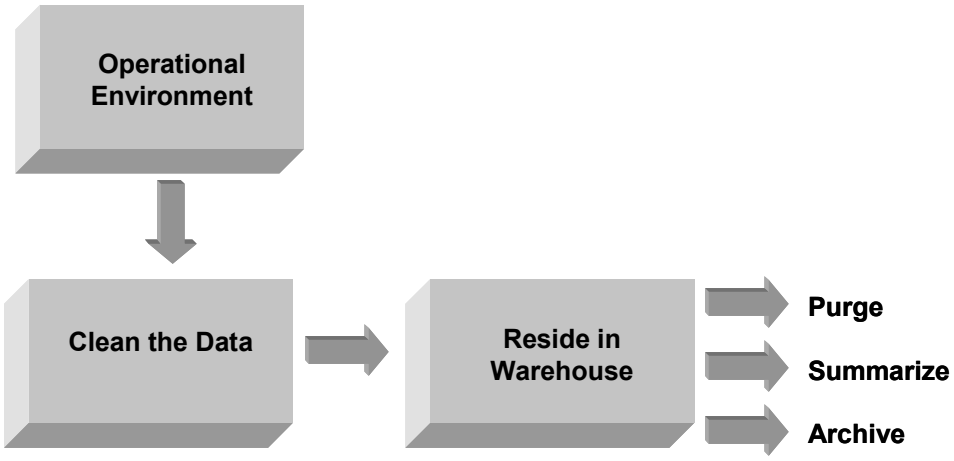


Figure 1. Data Flow
(Adapted from Gray and Watson 1998b)

Figure 2. Data Warehouse Architecture
(Adapted from Gray and Watson 1998b)

Typically the data warehouse architecture has three components:

- Data acquisition software (back-end) which extracts data from legacy systems and external sources, consolidates and summarizes the data, and loads them into the data warehouse.
- The data warehouse itself contains the data and associated database software. It is often referred to as the "target database."
- The client (front-end) software, which allows users and applications (such as DSS and EIS) to access and analyze data in the warehouse.

These three components may reside on different platforms, or two or three of them may be on the same platform. Regardless of the platform combination, all three components are required.

DATA MINING AND OLAP

The question of how data warehousing and OLAP relate to data mining is a question that often arises. The relationship can be succinctly captured as follows: " The capability of OLAP to provide multiple and dynamic views of summarized data in a data warehouse sets a solid foundation for successful data mining." [Han and Kamber 2001] Therefore, data mining and OLAP can be seen as tools than can be used to complement one another. The term OLAP, standing for Online Analytical Processing, is often used to describe the various types of query-

driven analysis that are undertaken when analyzing the data in a database or a data warehouse [Berry and Linoff 2000]. OLAP provides for the selective extraction and viewing of data from different points of view; these views are generally referred to as dimensions [Fayyad 2001]. Each dimension can and generally has many levels of aggregation, i.e. a time dimension can be organized into days, weeks, and years.

The essential distinction between OLAP and data mining is that OLAP is a data summarization/aggregation tool, while data mining thrives on detail. Data mining allows the automated discovery of implicit patterns and interesting knowledge that's hiding in large amounts of data [Han and Kamber 2001]. Prior to acting on the pattern uncovered by data mining, an analyst may use OLAP in order to determine the implications of using the discovered pattern in governing a decision. And while OLAP is considered part of the spectrum of decision support tools, it goes a step further than the traditional query and reporting tools. More specifically, the traditional query and reporting tools describe "what" is in a database, while OLAP is used to answer "why" certain things are true in that the user forms a hypothesis about a relationship and verifies it with a series of queries against the data. For example, an analyst might want to determine the factors that lead to loan defaults. She might initially hypothesize that people with low incomes are bad credits risks and analyze the database with OLAP to verify or disprove this assumption.

Expressions used in OLAP that describe the various functions include:

- *rolling up* (producing marginals),
- *drilling* (going down levels of aggregation—the opposite of rolling up),
- *slicing* (conditioning on one variable),
- *dicing* (conditioning on many variables) and
- *pivoting* (rotating the data axes to provide an alternative presentation of the data [Hand 1998; Han and Kamber 2001].

A powerful paradigm that integrates OLAP with data mining technology is OLAM (Online Analytical Mining) which is sometimes referred to as OLAP mining [Han and Kamber 2001]. OLAM systems are particularly important because most data mining tools need to work on integrated, consistent, and cleaned data, which again, requires costly data cleaning, data transformation, and data integration as pre-processing steps. A data warehouse constructed by such pre-processing serves as a valuable source of high-quality data for OLAP as well as for OLAM. OLAM provides a multi-dimensional view of its data and creates an interactive data mining environment whereby users can dynamically select data mining and OLAP functions, perform OLAP operations (such as drilling, slicing, dicing and pivoting on the data mining results), as well as perform mining operations on OLAP results, that is, mining different portions of data at multiple levels of abstraction [Han and Kamber 2001].

DATA MINING IN PERSPECTIVE

While the term data mining is often used rather loosely, it is generally a term that's used for a specific set of activities, all of which involve extracting meaningful new information from data. However, the term data mining is not new to statisticians. It is a term synonymous with *data dredging* or *data snooping* and has been used to describe the process of trawling through data in the hope of identifying patterns. Data snooping occurs when a given dataset is used more than once for inference or model selection [White 2000]. The connotation is derogatory because a sufficiently exhaustive search will certainly throw up patterns of some kind—by definition, data that are not simply uniform contain differences that can be interpreted as patterns. The trouble is that many of these "patterns" will simply be a product of random fluctuations, and will not represent any underlying structure in the data. The objective of data analysis is not to model the fleeting random patterns of the moment, but to model the underlying structures that give rise to consistent and replicable patterns.

In summary, data mining helps organizations focus on the most important information available in their existing databases. But data mining is only tool; it does not eliminate the need to know the business, to understand the data, or to understand the analytical methods involved. It

must be remembered that the predictive relationships found via data mining are not necessarily causes of an action or a behavior. Causal inference from uncontrolled convenience samples,

SIDEBAR 1

ACTORS IN DATA MINING

Data mining is performed by people, many of whom will be discussed in this tutorial. They include:

The project leader, who has the overall responsibility for planning, coordinating, executing, and deploying the data mining project.

The data mining client, who is the business domain expert that requests the project and utilizes the results, but generally does not possess the technical skills needed to participate in the execution of the more technical phases of the data mining project such as data preparation and modeling.

The data mining analyst, who thoroughly understands, from a business perspective, what the client wants to accomplish and assists in translating those business objectives into technical requirements to be used in the subsequent development of the data mining model(s).

The data mining engineer, who develops, interprets and evaluates the data mining model(s) in light of the business objectives and business success criteria. Data mining engineering is performed in consultation with the data mining client and the data mining analyst in order to assist in achieving business ends.

The IT analyst, who provides access to the hardware, software and data needed to complete the data mining project successfully. It is important to note that data mining is a technology that needs to co-exist harmoniously with other technologies in the organization. In addition, the data to be mined could be coming from virtually any existing system, database, or data warehouse in the organization.

_____

Depending on the scale and scope of the project, multiple individuals may assume each of the various roles. For example, a large project would likely need several data mining analysts and data mining engineers.

such as those used in data mining, are subject to several sources of error such as latent variables, sample selection bias, model equivalence and non-stationarity of the population being studied, or population drift [Glymour and Madigan 1996; Hand 1998]. Further, data mining assists analysts with finding patterns and relationships in the data – it does not indicate the value of the patterns to the organization. The patterns uncovered by data mining must be verified and validated in an appropriate context.

II. THE BUSINESS IMPERATIVE

Data mining offers value across a broad spectrum of industries and can be used as a vehicle to increase profits by reducing costs and/or raising revenue. A few of the common ways in which data mining can accomplish those objectives are

- lowering costs at the beginning of the product life cycle during research and development;
- determining the proper bounds for statistical process control methods in automated manufacturing processes;

- eliminating expensive mailings to customers who are unlikely to respond to an offer during a marketing campaign;
- facilitating one-to-one marketing and mass customization opportunities in customer relationship management.

Many organizations use data mining to help manage all phases of the customer life cycle, including acquiring new customers, increasing revenue from existing customers, and retaining good customers. By determining characteristics of good customers (profiling), a company can target prospects with similar characteristics. By profiling customers who bought a particular product a firm can focus attention on similar customers who have not bought that product (cross-selling). Profiling also enables a company to act to retain customers who are at risk for leaving (reducing churn or attrition), because it is usually far less expensive to retain a customer than acquire a new one [Berry and Linoff 2000]. However, profiling introduces issues of privacy (Section VII).

Examples of other industries where data mining can make a contribution include:

- *Telecommunications and credit card companies* are two of the leaders in applying data mining to detect fraudulent use of their services.
- *Insurance companies and stock exchanges* are interested in applying data mining to reduce fraud.
- *Medical applications* use data mining to predict the effectiveness of surgical procedures, medical tests, or medications.
- *Financial firms* use data mining to determine market and industry characteristics as well as to predict individual company and stock performance.
- *Retailers* make use of data mining to decide which products to stock in particular stores (and even how to place them within a store), as well as to assess the effectiveness of promotions and coupons.
- *Pharmaceutical firms* mine large databases for chemical compounds and genetic material to discover substances that might be candidates for development as agents for the treatments of disease.

III. THE TECHNICAL IMPERATIVE

Data mining uses

- the classical statistical procedures such as logistic regression, discriminant analysis, and cluster analysis,
- machine learning techniques such as neural networks, decision trees, and genetic algorithms.

In the continuum of data analysis techniques, the disciplines of statistics and of machine learning often overlap.

DATA MINING AND MACHINE LEARNING

Machine learning is the study of computational methods for improving performance by mechanizing the acquisition of knowledge from experience [Langley and Simon 1995]. Machine learning aims to provide increasing levels of automation in the knowledge engineering process, replacing much time-consuming human activity with automatic techniques that improve accuracy or efficiency by discovering and exploiting regularities in training data[1]. Although machine learning algorithms are central to the data mining process, it is important to note that the process also involves other important steps, including:

[1] Training data is data that is used to estimate or train a model. Training data are existing data that's pre-classified in that the outcomes are already known.

- building and maintaining the database,
- data formatting and cleansing,
- data visualization and summarization,
- the use of human expert knowledge to formulate the inputs to the learning algorithm and to evaluate the empirical regularities it discovers, and
- determining how to deploy the results.

Following are the basic learning algorithms [Peacock 1998b; Goebel and Gruenwald, 1999]. For an exhaustive review of machine learning algorithms, see Kennedy et al. [1997].

Neural Networks (NN) are a class of systems modeled after the human brain. As the human brain consists of millions of neurons that are inter-connected by synapses, NN are formed from large numbers of simulated neurons, connected to each other in a manner similar to brain neurons. As in the human brain, the strength of neuron inter-connections may change (or be changed by the learning algorithm) in response to a presented stimulus or an obtained output, which enables the network to "learn".

A disadvantage of NN is that building the initial neural network model can be especially time-intensive because input processing almost always means that raw data must be transformed. Variable screening and selection requires large amounts of the analysts' time and skill. Also, for the user without a technical background, figuring out how neural networks operate is far from obvious.

Case-Based Reasoning (CBR) is a technology that tries to solve a given problem by making direct use of past experiences and solutions. A case is usually a specific problem that was encountered and solved previously. Given a particular new problem, CBR examines the set of stored cases and finds similar ones. If similar cases exist, their solution is applied to the new problem, and the problem is added to the case base for future reference.

A disadvantage of CBR is that the solutions included in the case database may not be optimal in any sense because they are limited to what was actually done in the past, not necessarily what should have been done under similar circumstances. Therefore, using them may simply perpetuate earlier mistakes.

Genetic Algorithms (GA) operate through procedures modeled upon the evolutionary biological processes of selection, reproduction, mutation, and survival of the fittest to search for very good solutions to prediction and classification problems. GA are used in data mining to formulate hypotheses about dependencies between variables in the form of association rules or some other internal formalism.

A disadvantage of GA is that the solutions are difficult to explain. Also, they do not provide interpretive statistical measures that enable the user to understand why the procedure arrived at a particular solution.

Decision Trees (DT) are like those used in decision analysis where each non-terminal node represents a test or decision on the data item considered. Depending on the outcome of the test, one chooses a certain branch. To classify a particular data item, one would start at the root node and follow the assertions down until a terminal node (or leaf) is reached; at that point, a decision is made. DT can also be interpreted as a special form of a rule set, characterized by their hierarchical organization of rules.

A disadvantage of DT is that trees use up data very rapidly in the training[2] process. They should never be used with small data sets. They are also highly sensitive to noise in the data, and they try to fit the data exactly, which is referred to as "overfitting. Overfitting[3], discussed further in Section V on modeling, means that the model depends too strongly on the details of the particular dataset used to create it. When a model suffers from overfitting, it is unlikely to be externally valid (i.e., it won't hold up when applied to a new data set).

Association Rules (AR) are statements about relationships between the attributes of a known group of entities and one or more aspects of those entities that enable predictions to be

[2] The training process refers to the process of estimating a model's parameters whereby the machine learning techniques learn or are trained on existing pre-classified data.

[3] An overfitted model occurs when the specification of the model is in large part an artifact of the idiosyncrasies of the data set used to build it and can reduce external validity.

made about aspects of other entities who are not in the group, but who possess the same attributes. More generally, AR state a statistical correlation between the occurrences of certain attributes in a data item, or between certain data items in a data set. The general form of an AR is $X_1 \ldots X_n \Rightarrow Y[C,S]$ which means that the attributes $X_1, \ldots, X_n$ predict Y with a confidence C and a significance S.

While these so-called first-generation algorithms are widely used, they have significant limitations. They typically assume the data contains only numeric and textual symbols and do not contain images. They assume the data was carefully collected into a single database with a specific data mining task in mind. Furthermore, these algorithms tend to be fully automatic and therefore fail to allow guidance from knowledgeable users at key stages in the search for data regularities.

DATA MINING AND STATISTICS

The disciplines of statistics and data mining both aim to discover structure in data. So much do their aims overlap, that some people regard data mining as a subset of statistics. But that is not a realistic assessment as data mining also makes use of ideas, tools, and methods from other areas – particularly database technology and machine learning, and is not heavily concerned with some areas in which statisticians are interested [Hand 1999]. Statistical procedures do, however, play a major role in data mining, particularly in the processes of developing and assessing models. Most of the learning algorithms use statistical tests when constructing rules or trees and also for correcting models that are overfitted. Statistical tests are also used to validate machine learning models and to evaluate machine learning algorithms.

Some of the commonly used statistical analysis techniques are discussed below. For an extensive review of classical statistical algorithms see Johnson and Wicheren [1998].

Descriptive and Visualization Techniques include simple descriptive statistics such as:

- averages and measures of variation,
- counts and percentages, and
- cross-tabs and simple correlations

They are useful for understanding the structure of the data. Visualization is primarily a discovery technique and is useful for interpreting large amounts of data; visualization tools include histograms, box plots, scatter diagrams, and multi-dimensional surface plots [Tegarden 1999].

Cluster Analysis seeks to organize information about variables so that relatively homogeneous groups, or "clusters," can be formed. The clusters formed with this family of methods should be highly internally homogenous (members are similar to one another) and highly externally heterogeneous (members are *not* like members of other clusters).

Correlation Analysis measures the relationship between two variables. The resulting correlation coefficient shows if changes in one variable will result in changes in the other. When comparing the correlation between two variables, the goal is to see if a change in the independent variable will result in a change in the dependent variable. This information helps in understanding an independent variable's predictive abilities. Correlation findings, just as regression findings, can be useful in analyzing causal relationships, but they do not by themselves establish causal patterns.

Discriminant Analysis is used to predict membership in two or more mutually exclusive groups from a set of predictors, when there is no natural ordering on the groups. Discriminant analysis can be seen as the inverse of a one-way multivariate analysis of variance (MANOVA) in that the levels of the independent variable (or factor) for MANOVA become the categories of the dependent variable for discriminant analysis, and the dependent variables of the MANOVA become the predictors for discriminant analysis.

Factor Analysis is useful for understanding the underlying reasons for the correlations among a group of variables. The main applications of factor analytic techniques are to reduce the number of variables and to detect structure in the relationships among variables; that is to classify variables. Therefore, factor analysis can be applied as a data reduction or structure detection method. In an exploratory factor analysis, the goal is to explore or search for a factor structure. Confirmatory factor analysis, on the other hand, assumes the factor structure is known a priori and the objective is to empirically verify or confirm that the assumed factor structure is correct.

Regression Analysis is a statistical tool that uses the relation between two or more quantitative variables so that one variable (dependent variable) can be predicted from the other(s) (independent variables). But no matter how strong the statistical relations are between the variables, no cause-and-effect pattern is necessarily implied by the regression model. Regression analysis comes in many flavors, including simple linear, multiple linear, curvilinear, and multiple curvilinear regression models, as well as logistic regression, which is discussed next.

Logistic Regression is used when the response variable is a binary or qualitative outcome. Although logistic regression finds a "best fitting" equation just as linear regression does, the principles on which it does so are rather different. Instead of using a least-squared deviations criterion[4] for the best fit, it uses a maximum likelihood method, that is, it maximizes the probability of obtaining the observed results given the fitted regression coefficients. Because logistic regression does not make any assumptions about the distribution for the independent variables, it is more robust to violations of the normality assumption. Some of the more common flavors that logistic regression comes in include simple, multiple, polytomous and Poisson logistic regression models.

DATA ANALYSIS TASKS AND TECHNIQUES

Several data mining problem types, or analysis tasks are typically encountered during a data mining project. Depending on the desired outcome, several data analysis techniques with different goals may be applied successively to achieve a desired result. For example, to determine which customers are likely to buy a new product, a business analyst may need first to use cluster analysis to segment the customer database, then apply regression analysis to predict buying behavior for each cluster. The data mining analysis tasks typically fall into the general categories listed below. For each data analysis task, an example of a useful data analysis technique is presented.

Again, there is a continuum of data analysis techniques and the two disciplines of statistics and machine learning often overlap. Table 1 is a matrix that summarizes the data mining analysis tasks and the techniques useful for performing these tasks. The table is representative of the many possibilities since the permutations and combinations of data analysis tasks and techniques are numerous.

Data Summarization gives the user an overview of the structure of the data and is generally carried out in the early stages of a project. This type of initial exploratory data analysis can help to understand the nature of the data and to find potential hypotheses for hidden information. Simple descriptive statistical and visualization techniques generally apply.

Segmentation separates the data into interesting and meaningful sub-groups or classes. In this case, the analyst can hypothesize certain subgroups as relevant for the business question based on prior knowledge or based on the outcome of data description and summarization. Automatic clustering techniques can detect previously unsuspected and hidden structures in data that allow segmentation. Clustering techniques, visualization and neural nets generally apply.

Classification assumes that a set of objects—characterized by some attributes or features—belong to different classes. The class label is a discrete qualitative identifier; for example, large, medium, or small. The objective is to build classification models that assign the correct class to previously unseen and unlabeled objects. Classification models are mostly used for predictive modeling. Discriminant analysis, decision tree, rule induction methods, and genetic algorithms generally apply.

Prediction is very similar to classification. The difference is that in prediction, the class is not a qualitative discrete attribute but a continuous one. The goal of prediction is to find the numerical value of the target attribute for unseen objects; this problem type is also known as regression, and if the prediction deals with time series data, then it is often called forecasting. Regression analysis, decision trees, and neural nets generally apply.

---

[4] The least-squares criterion is a common method used in regression analysis, which finds the regression coefficients that minimize the sum of the squared deviation of the predicted values of the model from the observed values of the data.

Table 1. Data Analysis Tasks and Techniques

DATA ANALYSIS TECHNIQUES	Data Summarization	Segmen-tation	Classifi-cation	Prediction	Dependency Analysis
Descriptive and Visualization	♦	♦			♦
Correlation Analysis					♦
Cluster Analysis		♦			
Discriminant Analysis			♦		
Regression Analysis				♦	♦
Neural Networks		♦	♦	♦	
Case-Based Reasoning					♦
Decision Trees			♦	♦	
Association Rules					♦

Dependency analysis deals with finding a model that describes significant dependencies (or associations) between data items or events. Dependencies can be used to predict the value of an item given information on other data items. Dependency analysis has close connections with classification and prediction because the dependencies are implicitly used for the formulation of predictive models. Correlation analysis, regression analysis, association rules, case-based reasoning and visualization techniques generally apply.

IV. DATA MINING AND THE WEB

With the large amount of information available online, the Web is a fertile area for data mining and knowledge discovery. In Web mining, data can be collected at the

- server-side,
- client-side,
- proxy servers, or
- obtained from an organization's database (which may contain business data or consolidated web data).

Each type of data collection differs not only in terms of the location of the data source, but also

- the kinds of data available,
- the segment of population from which the data was collected, and its
- method of implementation.

A meta-analysis of the web mining literature, categorized web mining into three areas of interest based on which part of the web is to be mined [Kosala and Blockeel, 2000; Srivastava, et al, 2000]:

- Web Content Mining,
- Web Structure Mining and
- Web Usage Mining.

The three areas are described next , although the distinctions among them are not clear-cut.

Web content mining describes the discovery of useful information from the web content/data/documents. Essentially, the web content data consists of the data the web page was designed to convey to the users, including text, image, audio, video, metadata, and hyperlinks.

Web structure mining tries to discover the model underlying the link structures of the Web. *Intra-page* structure information includes the arrangement of various HTML or XML tags within a given page, while *inter-page* structure information is hyper-links connecting one page to another. This model can be used to categorize web pages and is useful to generate information such as the similarity and relationship among Web sites.

Web usage mining (also referred to as click-stream analysis [Edelstein 2001]) is the process of applying data mining techniques to the discovery of usage patterns from Web data, and is targeted towards applications [Srivastava, et al. 2000]. It tries to make sense of the data generated by the Web surfer's sessions or behaviors. While the web content and structure mining use the real or primary data on the web, web usage mining mines the secondary data derived from the interactions of the users during Web sessions. Web usage data includes the data from web server access logs, browser logs, user profiles, registration data, user sessions or transactions, cookies, user queries, mouse clicks, and any other data as the result of interaction with the Web.

Given its application potential, particularly in terms of electronic commerce, interest in web usage mining, increased rapidly in both the research and practice communities. Section V will provide a high level overview of the web usage mining process [Srivastava, et al 2000].

As shown in Figure 3, three main tasks are performed in web usage mining; preprocessing, pattern discovery, and pattern analysis.

Preprocessing consists of converting the usage, content, and structure contained in the various available data sources into the data abstractions necessary for pattern discovery. It is typically the most difficult task in the web usage mining process due to the incompleteness of the available data. Some of the typical problems include:

- single IP address/multiple server sessions,
- multiple IP address/single server sessions,
- multiple IP addresses/single user and
- multiple agent/single user.

Other challenges associated with data collection include:

- Raw click-stream data needs to be collected from multiple servers
- Individual customer data is usually buried in a mass of other data regarding pages served, hosts, referring pages, and browser types
- A single page request can generate multiple entries in server logs
- Taking a sequence of log records and creating a session of page views involves lots of data cleansing to eliminate extraneous information
- Identifying the sessions contained in the data stream requires the use of cookies or embedding session identification numbers in URLs
- The use of proxy servers, where servers other than the home server fulfill customer requests, makes it difficult to identify the end of a session and the reason the session ended.

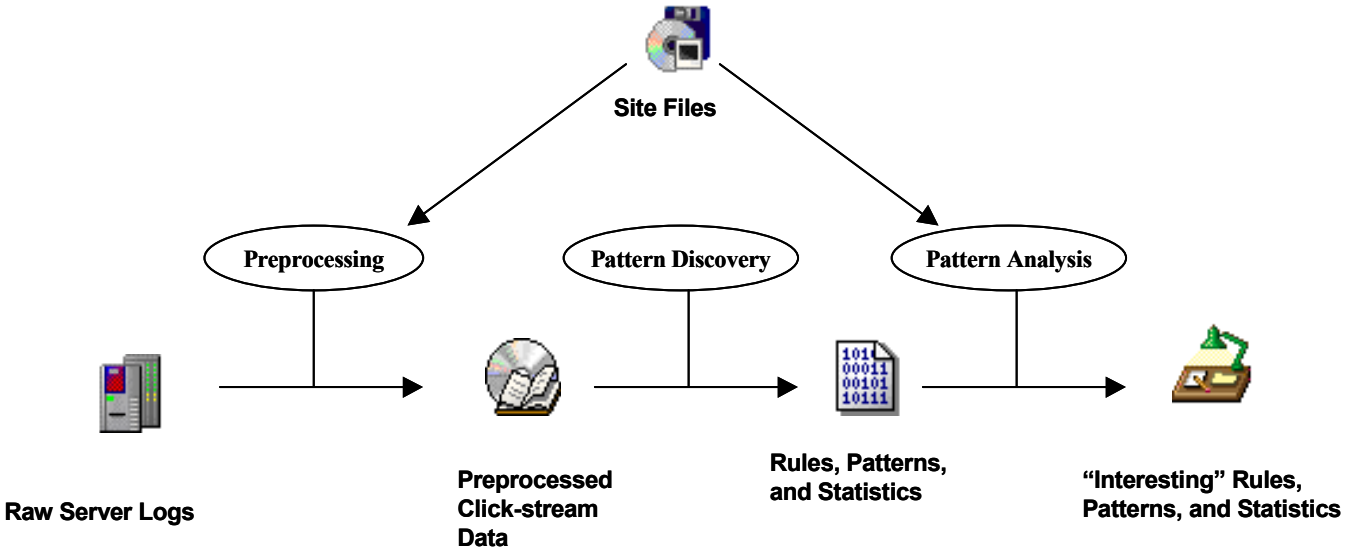


Figure 3. A High Level Web Usage Mining Process
(Adapted from Srivastava, et al 2000).

Pattern discovery draws upon methods and algorithms developed from several fields such as statistics, data mining, machine learning and pattern recognition (discussed in Section III). The methods and algorithms are similar to those developed for non-Web domains such as statistical analysis, clustering, and classification, but those methods must take into consideration the different kinds of data abstractions and prior knowledge available for Web Mining. For example, in association rule discovery, the notion of a transaction for market-basket analysis does not take into consideration the order in which items are selected. However, in Web Usage Mining, a server session is an ordered sequence of pages requested by a user.

Pattern analysis is the last step in the overall Web Usage mining process. The motivation behind pattern analysis is to filter out the uninteresting rules or patterns from the dataset found in the pattern discovery phase. The exact methodology used for analysis is usually governed by the application for which Web mining is to be done. The most common form of pattern analysis consists of a knowledge query mechanism such as SQL. Another method is to load usage data into a data cube to perform OLAP operations. Visualization techniques, such as graphing patterns or assigning colors to different values, can highlight patterns. The content and structure information can be used to filter out patterns which contain pages of a certain use type or content, or pages that match a certain hyperlink structure.

Despite being a rich source for data mining, the Web poses challenges for effective resource and knowledge discovery particularly in terms of data collection. The Web seems to be too huge for effective data warehousing and data mining. Also, Web pages are complex and lack a unifying structure. The highly dynamic nature of the Web as an information source poses challenges as well.

V. METHODOLOGICAL CONSIDERATIONS

Many data mining process methodologies are available. However, the various steps do not differ much from methodology to methodology. Two popular methodologies used by data mining tools are the SEMMA process for SAS Enterprise Miner and the 5 A's process for SPSS

Clementine. However, CRISP-DM evolved to become the *de facto* industry standard. CRISP-DM was conceived in mid-1997 and is non-proprietary, documented, and freely available. It was developed using input from more than 200 data mining users and data mining tool and service providers and is designed to provide a generic process model that can be specialized according to the needs of any particular company or industry.

SAS - THE SEMMA ANALYSIS CYCLE

SAS developed a data mining analysis cycle known by the acronym SEMMA. This acronym stands for the five steps of the analyses that are generally a part of a data mining project:

1. **S**ample,
2. **E**xplore
3. **M**odify
4. **M**odel
5. **A**ssess

as illustrated in Figure 4. The SEMMA analysis cycle guides the analyst through the process of exploring the data using visual and statistical techniques, transforming data to uncover the most significant predictive variables, modeling the variables to predict outcomes, and assessing the model by testing it with new data.

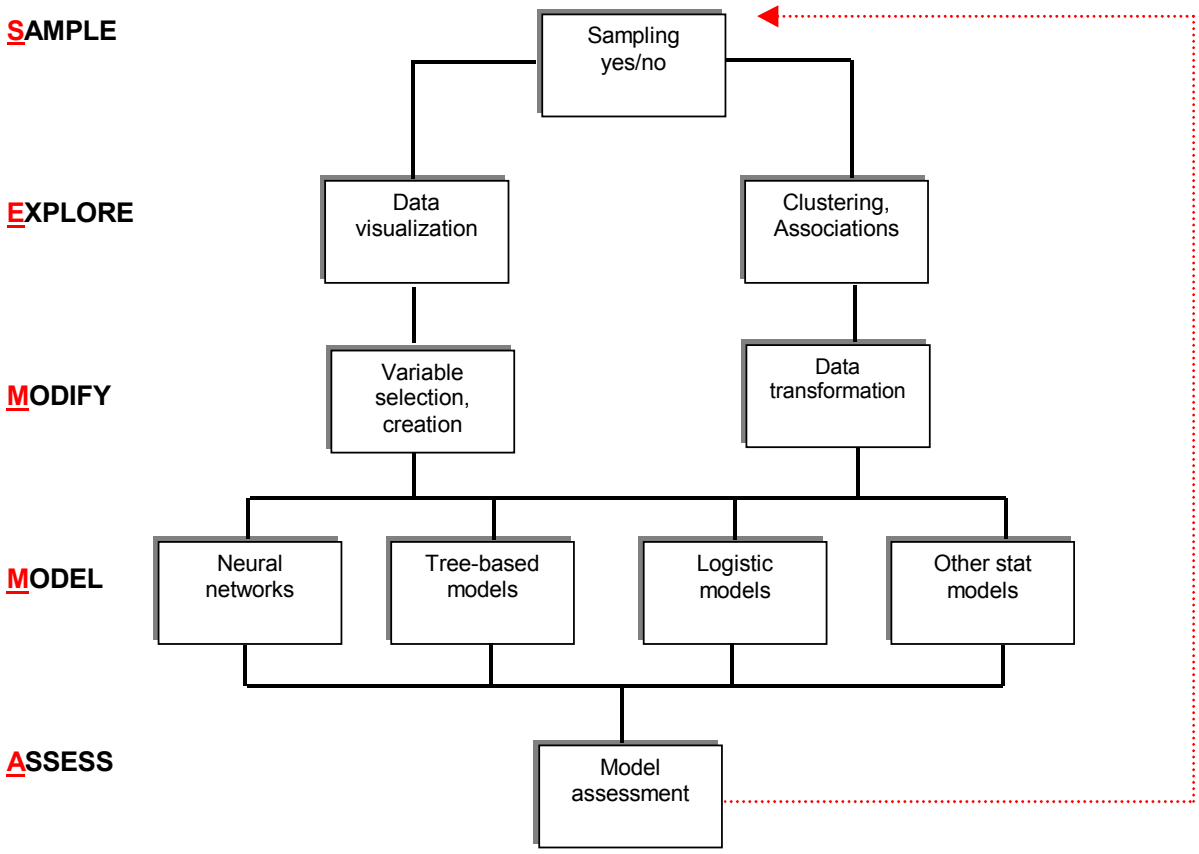


Figure 4. The SEMMA Analysis Cycle.

Sample: the first step in is to create one or more data tables by sampling data from the data warehouse. Mining a representative sample instead of the entire volume drastically reduces the processing time required to obtain business information.

Explore: after sampling the data, the next step is to explore the data visually or numerically for trends or groupings. Exploration helps to refine the discovery process. Techniques such as factor analysis, correlation analysis and clustering are often used in the discovery process.

Modify: modifying the data refers to creating, selecting, and transforming one or more variables to focus the model selection process in a particular direction, or to modify the data for clarity or consistence.

Model: creating a data model involves using the data mining software to search automatically for a combination of data that predicts the desired outcome reliably.

Assess: the last step is to assess the model to determine how well it performs. A common means of assessing a model is to set aside a portion of the data during the sampling stage. If the model is valid, it should work for both the reserved sample and for the sample that was used to develop the model.

SPSS - THE 5 A'S PROCESS

SPSS originally developed a data mining analysis cycle called the 5 A's Process[5]. The five steps in the process are

- **A**ssess
- **A**ccess
- **A**nalyze
- **A**ct
- **A**utomate

As illustrated in Figure 5. The 5 A's process methodology is similar to that of the SEMMA analysis cycle.

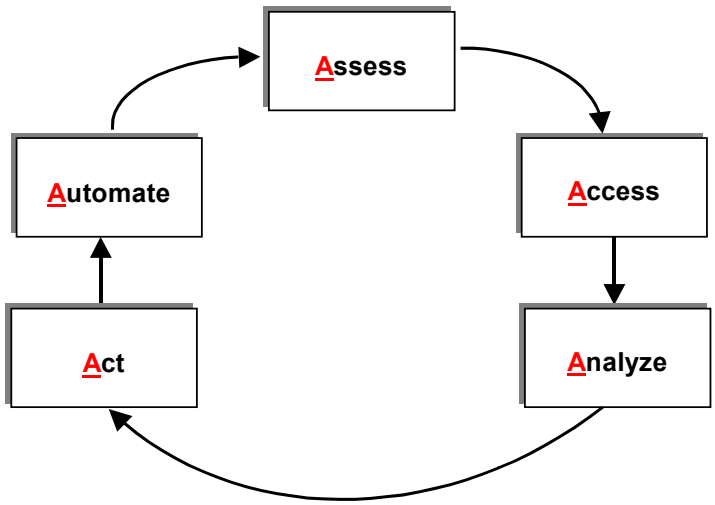


Figure 5. The 5A's Process

CRISP-DM – THE DE FACTO STANDARD FOR INDUSTRY

The CRISP-DM project began in mid-1997 and was funded in part by the European commission. The leading sponsors were:

[5] Note that during the second quarter of 2001, SPSS removed all references to the 5 A's process methodology from it's web site. SPSS now actively supports the CRISP-DM process model. SPSS uses the CRISP-DM model in it's consulting practice, offers a data mining training series built around CRISP-DM and provides support for CRISP-DM in Clementine, SPSS's data mining workbench. The Clementine support for CRISP-DM makes it convenient for Clementine users to structure data mining projects according to the CRISP-DM process methodology.

- NCR,
- DaimlerChrysler,
- Integral Solutions Limited (ISL) (now a part of SPSS), and
- OHRA, a Netherlands' independent insurance company

The goal of the project was to define and validate an industry- and tool-neutral data mining process model that which would make the development of large as well as small data mining projects faster, cheaper, more reliable and more manageable.

The project started in July 1997 and was planned to be completed within 18 months. However, the work of the CRISP-DM received substantial international interest, which caused the project to put emphasis on disseminating its work. As a result, the project end date was pushed back to and completed on April 30, 1999. The CRISP-DM model is illustrated in Figure 6.

BUSINESS UNDERSTANDING

This initial phase focuses on understanding the project objectives and requirements from a business perspective, and then converting this knowledge into a data mining problem definition, and a preliminary plan designed to achieve the objectives.

DATA UNDERSTANDING

The data understanding phase starts with an initial data collection and proceeds with activities to become familiar with the data, to identify data quality problems, to discover first insights into the data, or to detect interesting subsets to form hypotheses for hidden information.

DATA PREPARATION

The data preparation phase covers all activities to construct the final dataset (data that will be fed into the modeling tool(s)) from the initial raw data.

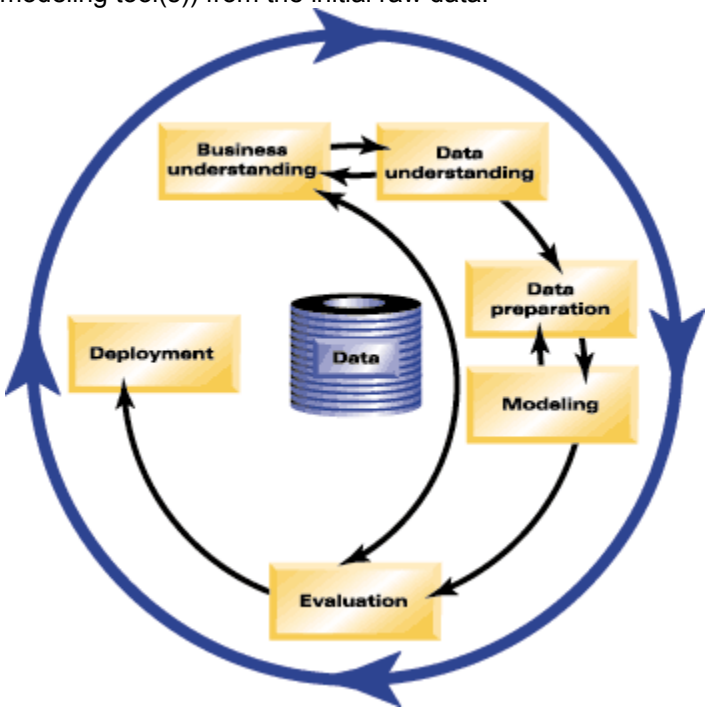


Figure 6. The CRISP-DM Model

Data preparation tasks are likely to be performed multiple times, and not in any prescribed order. Tasks include table, record, and attribute selection as well as transformation and cleaning of data for modeling tools.

MODELING

In this phase, various modeling techniques are selected and applied, and their parameters are calibrated to optimal values. Typically, there are several techniques for the same data mining problem type. Some techniques have specific requirements on the form of data. Therefore, stepping back to the data preparation phase is often needed.

EVALUATION

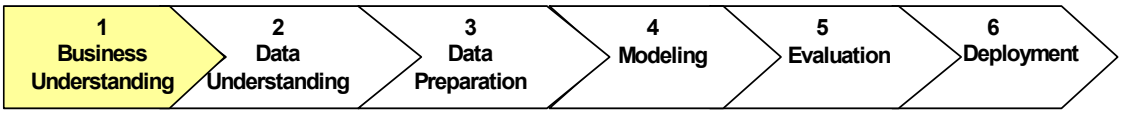
At this stage in the project the model (or models) built appears to have high quality from a data analysis perspective. Before proceeding to final deployment of the model, it is important to evaluate the model more thoroughly, and review the steps executed to construct the model, to be certain it properly achieves the business objectives. A key objective is to determine if there is some important business issue that has not been considered sufficiently. At the end of this phase, a decision on the use of the data mining results should be reached.

DEPLOYMENT

Creation of the model is generally not the end of the project. Even if the purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that the client can use. Depending on the requirements, the deployment phase can be as simple as generating a report or as complex as implementing a repeatable data mining process. In many cases it will be the client, not the data analyst, who will carry out the deployment steps. However, even if the analyst will not carry out the deployment effort it is important for the client to understand up front what actions will need to be carried out to make use of the models created.

V. ILLUSTRATION OF A DATA MINING PROCESS METHODOLOGY

The general form of the CRISP-DM data mining process methodology is further detailed and illustrated from a "how to" perspective in Figure 7 (shown on the next page). CRISP-DM does not provide details on two critical areas in the modeling process; building and assessing the model. While these two procedures are generally automated and most data mining tools provide support for them, it is important to gain an understanding of the purpose and focus of these steps. And to that end, (step 4) the Build Model and Assess Model portions of the Modeling phase of the CRISP-DM model, is supplemented with additional detail and illustrations.



1. Business Understanding

This initial phase focuses on understanding the project objectives and requirements from a business perspective, and then converting this knowledge into a data mining problem definition, and a preliminary plan designed to achieve the objectives.

1.1 Determine Business Objectives

The first objective of the data analyst is to understand thoroughly, from a business perspective, what the client really wants to accomplish. Often the client has many competing objectives and constraints that must be properly balanced. The analyst's goal is to uncover important factors, at the beginning, that can influence the outcome of the project. A possible consequence of neglecting this step is to expend a great deal of effort producing the right answers to the wrong questions.

1.2 Assess Situation

This task involves more detailed fact-finding about all of the resources, constraints, assumptions and other factors that should be considered in determining the data analysis goal and project plan. In the previous task, the objective is to get to the crux of the situation quickly. Here, the analyst wants to flesh out the details.

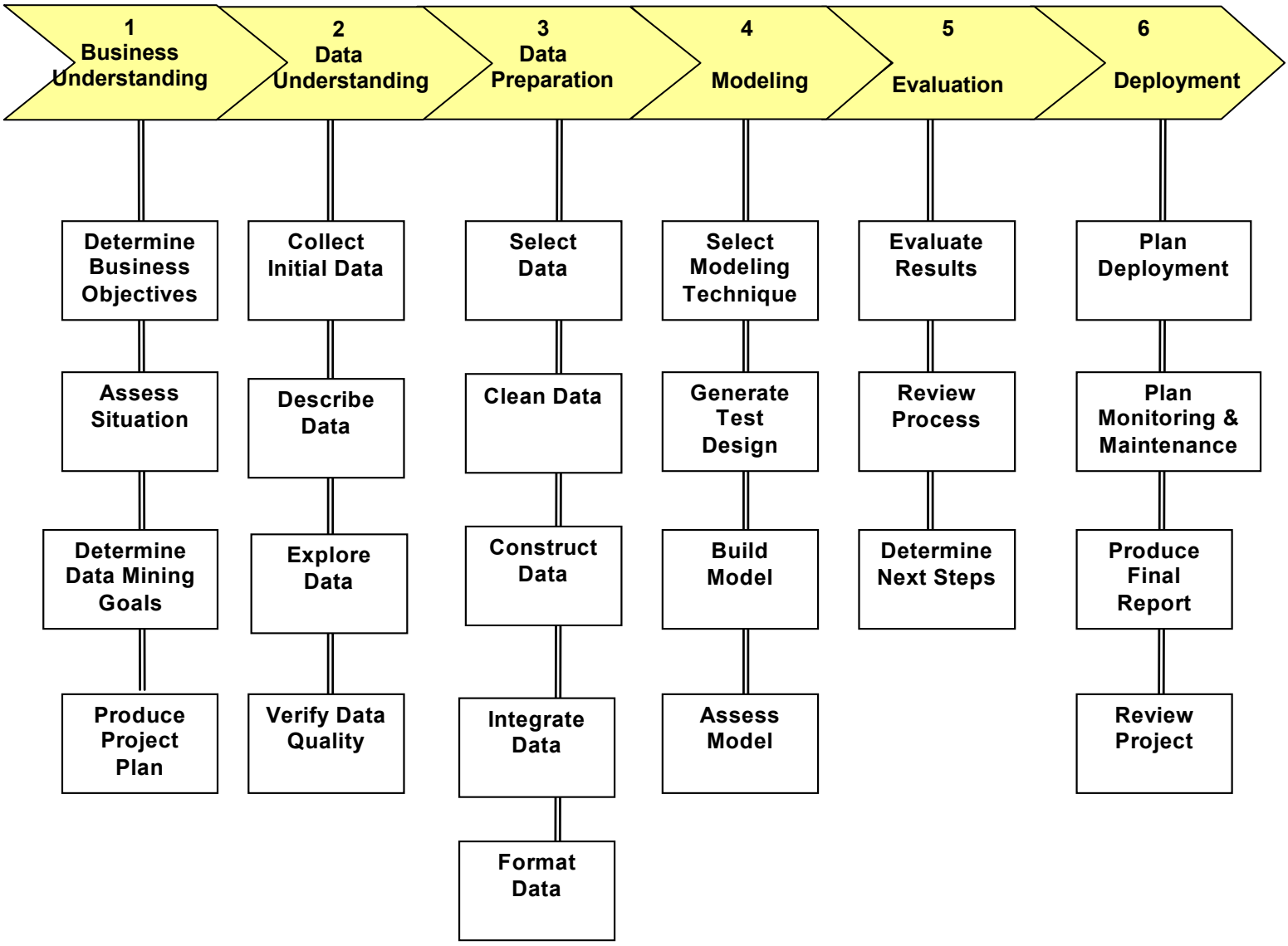


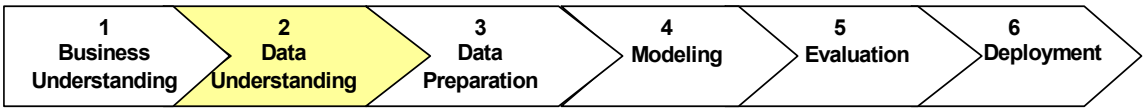
Figure 7. - CRISP-DM Data Mining Process Methodology

1.3 Determine Data Mining Goals

A *business goal* states objectives in business terms. A *data mining goal* states project objectives in technical terms. For example, the business goal might be "Increase catalog sales to existing customers." A data mining goal might be "Predict how many widgets a customer will buy, given their purchases over the past three years, demographic information (e.g., age, salary, city, Zip code), and the price of the item."

1.4 Produce Project Plan

Describe the intended plan for achieving the data mining goals and thereby achieving the business goals. The plan should specify the anticipated set of steps to be performed during the rest of the project including an initial selection of tools and techniques.



2. Data Understanding

The data understanding phase starts with an initial data collection. It proceeds with activities

- to get familiar with the data,
- to identify data quality problems,
- to discover first insights into the data, or to
- detect interesting subsets to form hypotheses for hidden information.

2.1 Collect Initial Data

Acquire within the project the data (or access to the data) listed in the project resources. This initial collection includes data loading if necessary for data understanding. For example, if applying a specific tool for data understanding, it makes perfect sense to load the data into this tool. This effort may lead to initial data preparation steps.

2.2 Describe Data

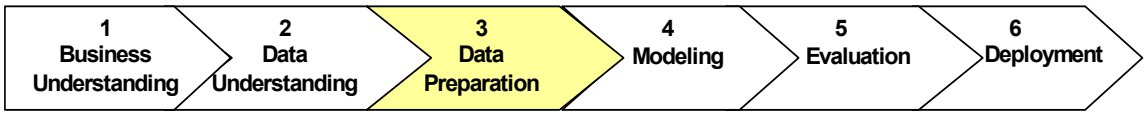
Examine the "gross" or "surface" properties of the acquired data and report on the results.

2.3 Explore Data

This task tackles the data mining questions that can be addressed using querying, visualization and reporting. These analyses may address the data mining goals directly. They may also contribute to or refine the data description and quality reports and feed into the transformation and other data preparation needed for further analysis.

2.4 Verify Data Quality

Examine the quality of the data, addressing questions such as: is the data complete? Is it correct? Are these missing values? If so how are they represented, where do they occur and how common are they?



3. Data Preparation

The data preparation phase covers all activities to construct the final dataset (data that will be fed into the modeling tool(s)) from the initial raw data. Data preparation tasks are likely to be performed multiple times, and not in any prescribed order. Tasks include table, record, and attribute selection as well as transformation and cleaning of data for modeling tools.

3.1 Select Data

Decide on the data to be used for analysis. Criteria include relevance to the data mining goals, quality and technical constraints such as limits on data volume or data types. Note that data selection covers selection of attributes (columns) as well as selection of records (rows) in a table.

3.2 Clean Data

Raise the data quality to the level required by the selected analysis techniques. Problems that can occur with "dirty data" include missing data, empty values, non-existent values, and incomplete data. Data cleaning may involve selection of clean subsets of the data, the insertion of suitable defaults or more ambitious techniques such as replacing the dirty data with derived values, or building separate models for those entities that possess dirty data. However, these approaches can introduce additional problems. Specifically, filtering the problematic data can introduce sample bias into the data and using data overlays could introduce missing values

3.3 Construct Data

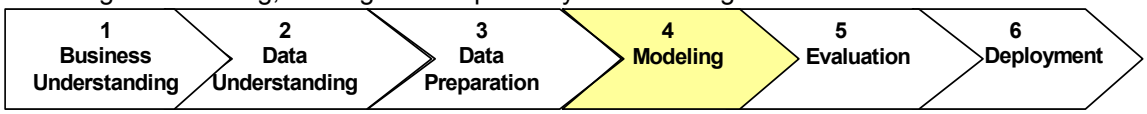
This task includes constructive data preparation operations such as the production of derived attributes, entire new records, or transformed values for existing attributes.

3.4 Integrate Data

Two methods used for integrating data are merging data and generating aggregate values. In these methods information is combined from multiple tables or other information sources to create new records or values. For example, merging tables refers to joining together two or more tables that have different information about the same objects; generating aggregate values refers to computing new values computed by summarizing information from multiple records, tables or other information sources.

3.5 Format Data

Formatting transformations refer to primarily syntactic modifications made to the data that do not change its meaning, but might be required by the modeling tool.



4. Modeling

In this phase, various modeling techniques are selected and applied, and their parameters are calibrated to optimal values. Typically, several techniques can be applied to the same data mining problem type. Some techniques require a specific form of data. Therefore, stepping back to the data preparation phase is often needed.

4.1 Select Modeling Technique

As the first step in modeling, select the actual modeling technique to be used. If a tool was selected in business understanding (Phase 1), this task refers to selecting the specific modeling technique, e.g., building decision trees or generating a neural network.

4.2 Generate Test Design

Prior to building a model, a procedure needs to be defined to test the model's quality and validity. For example, in supervised data mining tasks such as classification, it is common to use error rates as quality measures for data mining models. Therefore, if the test design specifies that the dataset should be separated into training and test sets, the model is built on the training set and its quality estimated on the test set.

4.3 Build Model

The purpose of building models is to use the predictions to make more informed business decisions. The most important goal when building a model is stability, which means that the model should make predictions that will hold true when it's applied to yet unseen data. Regardless of the data mining technique being used, the basic steps used for building predictive models are the same.

As shown in Figure 8, the model set first needs to be split into three components: (1)the training set, (2) the test set, and (3) the evaluation set.

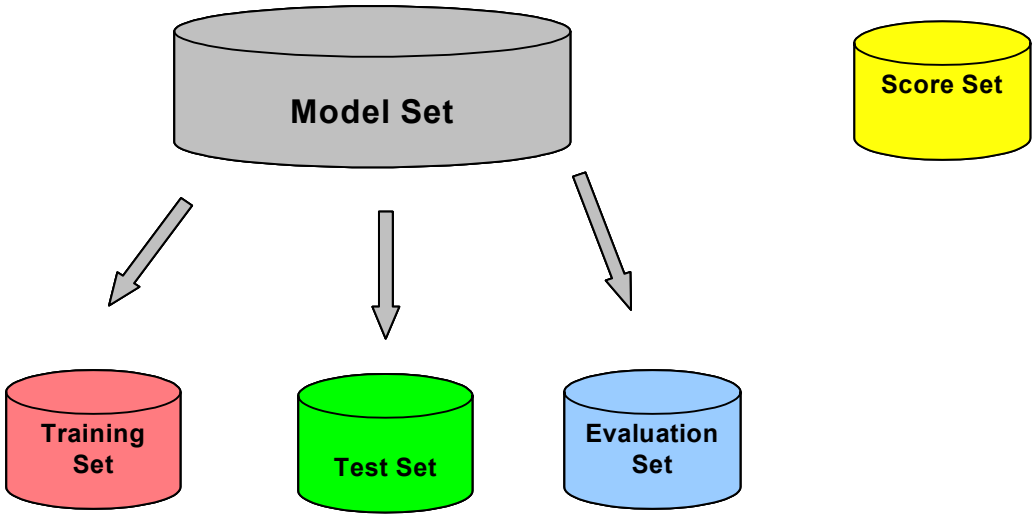


Figure 8. Data Sets

A fourth dataset, the score set, is not part of the model set.

Each of these components should be totally separate; that is, they should not have any records that are in common since each set performs a distinct purpose.

Models are created using data from the past in order for the model to make predictions about the future. This process is called training the model. In this step, the data mining algorithms find patterns that are of predictive value. Next, the model is refined using the test set. The model needs to be refined to prevent it from memorizing the training set. This step ensures that the model is more general (i.e. stable) and will perform well on unseen data. Next, the performance of the model is estimated using the evaluation set. The evaluation set is entirely separate and distinct from the training and test sets. The evaluation set (or hold out set) is used to assess the expected accuracy of the model when it is applied to data outside the model set. Finally, the model is applied to the score set. The score set is not pre-classified and is not part of the model set used to create the data model. The outcomes for the score set are not known in advance. The final model is applied to the score set to make predictions. The predictive sores will, presumably, be used to make more informed business decisions. The process is summarized in Figure 9.

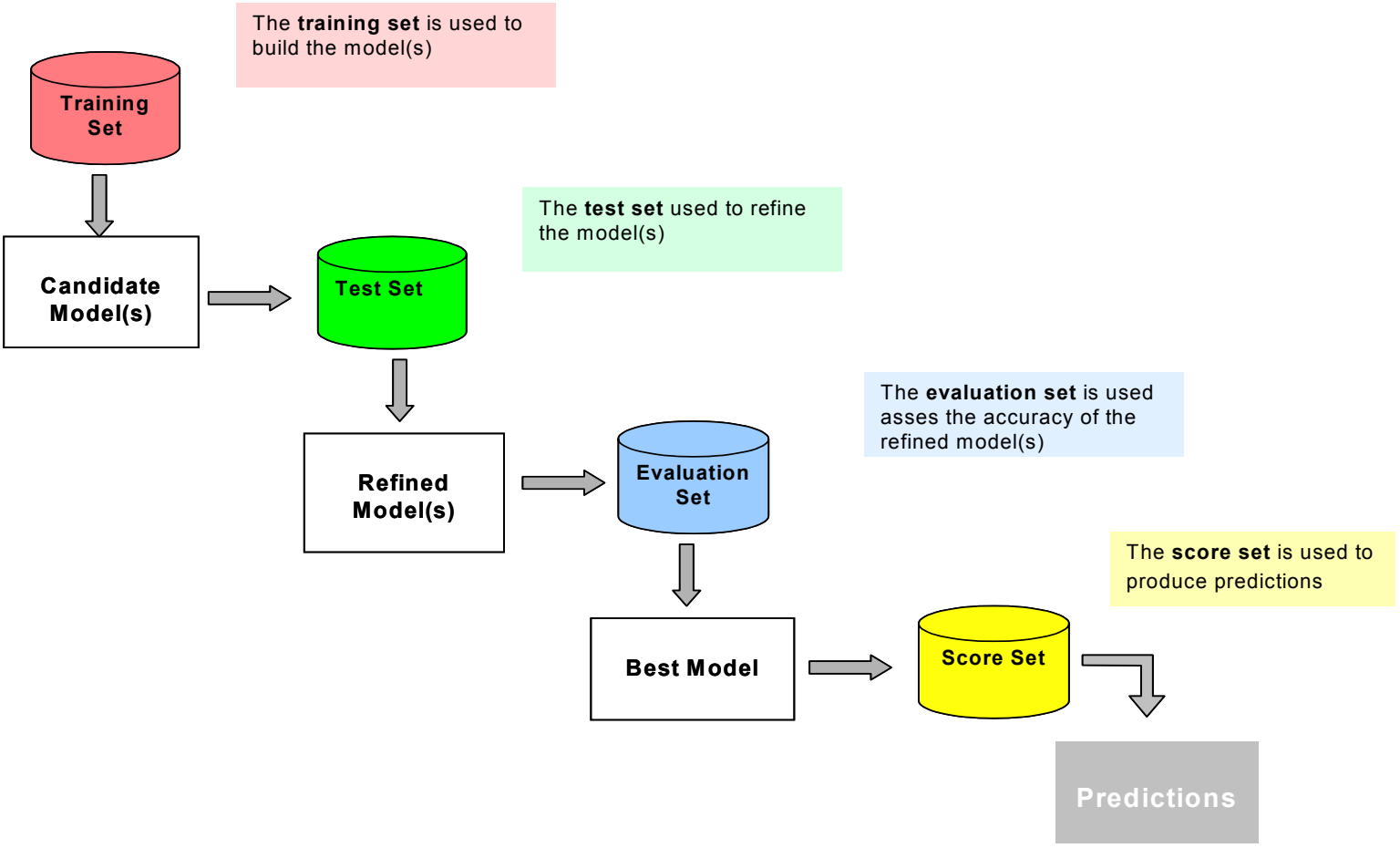


Figure 9. The Process of Building a Predictive Model

Overfitting. A problem that can occur is that the model created can overfit the data. As described in Section III, overfitting means that the specification of a model is in large part an artifact of the idiosyncrasies of the data set used to build it (i.e., the training set). Overfitting occurs when a model essentially memorizes the data on which is was built. The model should learn the patterns in order to recognize them in future unseen datasets, but the model should not memorize the patterns. The problem with the model memorizing the training set, is that when the model scores an unknown record, it will use the results from the model set if there is a match, and if not, it will produce a random guess. In that case the model is entirely unstable, i.e. it will do no better than random for the score set.

4.4 Assess Model

The model should now be assessed to ensure that it meets the data mining success criteria and passes the desired test criteria. This step is a purely technical assessment based on the outcome of the modeling tasks.

Two tools commonly used to assess the performance of different models are the lift chart and the confusion matrix.

A lift chart, sometimes called a cumulative gains chart, or a banana chart, is a measure of model performance. It shows how responses, (i.e., to a direct mail solicitation, or a surgical treatment for instance) are changed by applying the model. This change ratio, which is hopefully, the increase in response rate, is called the "lift". A lift chart indicates which subset of the dataset contains the greatest possible proportion of positive responses. The higher the lift curve is from the baseline, the better the performance of the model since the baseline represents the null model, which is no model at all.

To explain a lift chart, suppose we had a two-class prediction where the outcomes were yes (a positive response) or no (a negative response). To create a lift chart, instances in the dataset are sorted in descending probability order according to the predicted probability of a positive response. When the data is plotted, we can see a graphical depiction of the various probabilities. While the example shown in Figure 10 plots the results of different datasets for a single model, a lift chart can also be used to plot the results of a single dataset for different models.

Note that the best model is not the one with the highest lift when it is being built. It is the model that performs the best on unseen, future data.

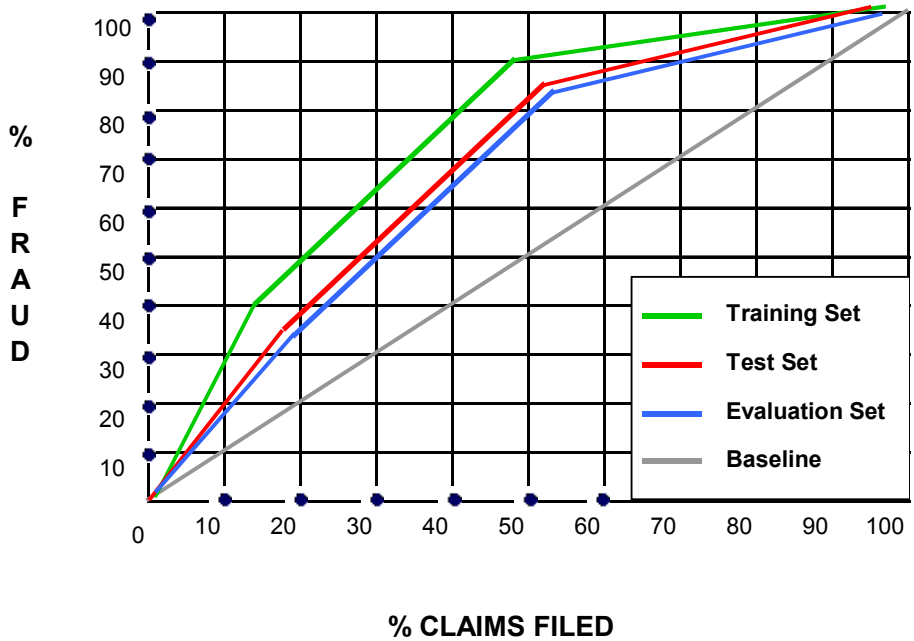


Figure 10. Example of a "Good" Lift Chart.

The results of the lift chart shown in Figure 10 are interpreted as follows: the model results for the **training** set, indicates that 20% of the claims filed account for 50% of known fraud cases.

The model results for the test and evaluation sets can be interpreted similarly.

The **baseline** indicates the expected result if no model were used at all, i.e. the naive probability of 50-50.

Figure 10 is an example of a good lift chart because not only does the chart exhibit good lift, but the results on the **test** and **evaluation** sets are similar. As should be the case, the performance on the training set is better than the performance on the test set, which in turn is better than performance on the evaluation set.

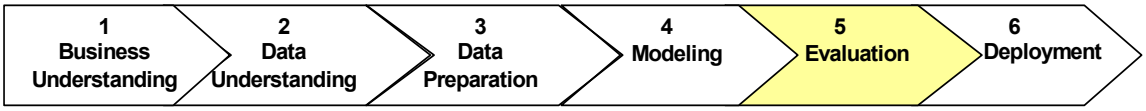
A confusion matrix, sometimes called a classification matrix, is used to assess the prediction accuracy of a model. It measures whether a model is confused or not; that is, whether the model is making mistakes in its predictions. Various classification rules are used in creating a confusion matrix. The classification rules that incorporate prior probabilities, posterior probabilities and misclassification costs are based on Bayesian statistical decision theory. Bayesian theory essentially revises prior probabilities based on additional available information [Sharma 1996]. The format of a confusion matrix for a two-class case with classes yes and no is shown in Figure 11.

		predicted class	
		yes	no
actual class	yes	true positive	false negative
	no	false positive	true negative

Figure 11. A Confusion Matrix for a Two-Class Case

The actual vales in a confusion matrix are often represented as percentages. Whether or not a confusion matrix is “good” depends on the costs of misclassification.

At the conclusion of the model building and assessment processes, the most appropriate model will be the model that meets the business objectives.



5. Evaluation

Previous evaluation steps dealt with factors such as the accuracy and generality of the model. This step assesses the degree to which the model meets the business objectives and seeks to determine if there is some business reason why the model is deficient. It compares results with the evaluation criteria defined at the start of the project.

A good way of defining the total outputs of a data mining project is to use the equation:

results = f(models, findings) (1)

In Equation 1, we define the total output of the data mining project as not just the models, but also the findings which can be defined as anything (apart from the model) that is important in meeting objectives of the business.

5.1 Evaluate Results

Previous evaluation steps dealt with factors such as the accuracy and generality of the model. This step assesses the degree to which the model meets the business objectives and seeks to determine if there is some business reason why this chosen model is deficient. Another option of evaluation is to test the model(s) on test applications in the real application if time and budget permits.

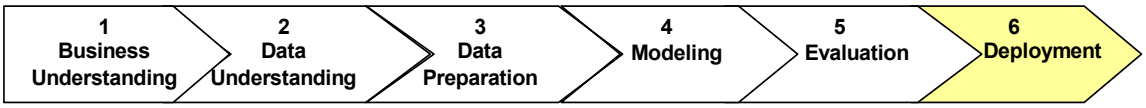
5.2 Review Process

At this point the resultant model appears to be satisfactory and appears to satisfy business needs. It is now appropriate to make a more thorough review of the data mining project in order to determine if there is any important factor or task that has somehow been overlooked. At this stage of Data Mining, the Review Process takes on the form of a Quality Assurance Review.

5.3 Determine Next Steps

According to the assessment results and the process review, the analyst decides how to proceed at this stage. The analyst needs to decide whether

- to finish the project and move on to deployment (Phase 6) or
- to initiate further iterations or
- to set up new data mining projects.



6. Deployment

Creation of the model is generally not the end of the project. Even if the purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that the client can use. Depending on the requirements, the deployment phase can be as simple as generating a report or as complex as implementing a repeatable data mining process. In many cases it will be the client, not the data analyst, who will carry out the deployment steps. However, even if the analyst will not carry out the deployment effort it is important for the client to understand up front what actions will need to be carried out to make use of the models created.

6.1 Plan Deployment

To deploy the data mining result(s) into the business, this task takes the evaluation results and develops a strategy for deployment. If a general procedure was identified to create the relevant model(s), this procedure is documented here for later deployment.

6.2 Plan Monitoring and Maintenance

Monitoring and maintenance are important issues if the data mining result becomes part of the day-to-day business and its environment. A careful preparation of a maintenance strategy helps to avoid unnecessarily long periods of incorrect usage of data mining results. To monitor the deployment of the data mining result(s), the project needs a detailed plan on the monitoring process. This plan takes into account the specific type of deployment.

6.3 Produce Final Report

At the end of the project, the project leader and the team write up a final report. Depending on the deployment plan, this report may be only a summary of the project and its experiences (if they have not already been documented as an ongoing activity) or it may be a final and comprehensive presentation of the data mining result(s).

6.4 Review Project

Assess what went right and what went wrong, what was done well and what needs to be improved.

VII. CONCLUSIONS AND DIRECTIONS FOR FURTHER RESEARCH

CONCLUSIONS

Today, most enterprises are actively collecting and storing data in large databases. Many of them have recognized the potential value of these data as an information source for making business decisions. The dramatically increasing demand for better decision support is answered by an extending availability of knowledge discovery, and data mining is one step at the core of the knowledge discovery process. This tutorial has illustrated how data mining centers about developing algorithms for extracting structure from data and how that structure can take the form of statistical patterns, models, and relationships. These structures provide a basis within which to predict and anticipate when certain events occur and when viewed at this level, one begins to understand the fundamental importance of data mining. Opportunities for further research abound particularly as the Internet provides businesses with an operational platform for interaction with their customers around the clock without geographic or physical boundaries. Therefore, from a strategic perspective, the need to navigate the rapidly growing universe of digital data will rely heavily on the ability to effectively manage and mine the raw data.

DIRECTIONS FOR FURTHER RESEARCH

The following is a (naturally incomplete) list of issues that warrant further investigation in the emerging field of data mining:

- *Privacy:* With such enthusiasm and opportunity for data mining the Internet, the serious issue of privacy needs to be handled effectively. Although privacy is not only an issue with data mining and the Internet, data mining researchers and practitioners need to be constantly aware of the implications of tracking and analysis technologies on privacy. Without properly addressing the issue of privacy on the Internet, the abundance of data may eventually flow much slower due to regulations, and other corrective or preventive restrictions.
- *Progress toward the development of a theory*: Progress toward the development of a theory regarding the correspondence between techniques and the specific problem domain to which they apply is needed. Questions regarding the relative performance of the various data mining algorithms remain largely unresolved. With a myriad of algorithms and problem sets to which they are applied, a systematic investigation of their performance is needed to guide the selection of a data mining algorithm for a specific case.
- *Extensibility:* Different techniques outperform one another for different problems. With the increasing number of proposed data analysis techniques as well as reported applications, it appears that any fixed set of algorithms will not be able to cover all potential problems and tasks. It is therefore important to provide an architecture that allows for easy syntheses of new methods, and for the adaptation of existing methods with as little effort as possible.
- *Integration with databases:* Most of the cost of data mining is not in the modeling algorithms; rather it is in data cleaning and preparation, and in data maintenance and management. The development of a standard application programming interface (API) and the subsequent integration with a database environment could reduce the costs associated with these tasks. The issues regarding data cleaning, preparation, maintenance and management are challenges that face databases, data warehouses, and decision support systems in general.
- *Managing changing data:* In many applications, particularly in the business domain, the data is not stationary, but rather changing and evolving. This changing data may make previously discovered patterns invalid and as a result, there is clearly a need for incremental methods that are able to update changing models, and for strategies to identify and manage patterns of temporal change in knowledge bases.

- *Non-standard data types:* Today's databases contain not only standard data such as numbers and strings, but also large amounts of non-standard and multi-media data, such as free-form text, audio, image and video data, temporal, spatial, and other data types. These data types contain special patterns, which cannot be handled well by the standard analysis methods, and therefore, require special, often domain-specific, methods and algorithms.
- *Support for both analysis experts and novice users:* With the current focus on technology and the automated techniques, rather than on the actual processes of exploration and analysis, many people perceive data mining as a product rather than as a discipline that must be mastered. Further most of the available tools are aimed at analysis experts and require an unaffordable amount of training before being useful to novice end users, who, while being less skilled in complex data analysis, have a thorough understanding of their knowledge domain.
- *Pattern Evaluation:* Several challenges remain regarding the development of techniques to assess the interestingness of discovered patterns as a data mining system can uncover thousands of patterns, but many of the patterns discovered may be uninteresting to the given user, i.e. representing common knowledge or lacking novelty. The use of interestingness measures, to guide and constrain the discovery process as well to reduce the search space is an active area for research.

FINAL REMARKS

While there are fundamental problems to be solved and challenges to be addressed, the benefits of data mining have been demonstrated in a broad range of application domains. The combination of urgent practical needs and the strong research interests seem to indicate that rather than it being a technology that receives early adopter enthusiasm, then eventually wanes; data mining seems certain to become mainstream and enjoy a wide adoption.

Editor's Note: This article is based on a tutorial presented by the author at AMCIS 2001 in Boston. The manuscript was received on July 1, 2001. It was with the author for approximately four and a half months for two revisions. It was published on March ,2002

REFERENCES

Berry, M. J., Linoff, G. S. (2000), "Mastering Data Mining: The Art and Science of Customer Relationship Management". Wiley Computer Publishing, New York.

Chung, H. M., Gray, P. (1999), "Special Section: Data Mining". *Journal of Management Information Systems*, (16:1),11-17.

Colin, S. (2000), "The CRISP-DM Model: The New Blueprint for Data Mining", *Journal of Data Warehousing*, (5:4), Fall, 13-22.

Fayyad, U., Piatetsky-Shapiro, G., and Smyth, R (1996). "The KDD Process for Extracting Useful Knowledge from Volumes of Data," *Communications of the ACM*, (39:11), pp. 27-34.

Fayyad, U., (2001), "The Digital Physics of Data Mining", *Communications of the ACM*, March, (44:3), 62-65.

Glymour, C., Madigan D., et al (1996), "Statistical Inference and Data Mining". *Communications of the ACM*, (39:11), 35-41.

Goebel, M., Gruenwald, L. (1999), "A Survey of Data Mining and Knowledge Discovery Software Tools", *ACM SIGKDD (Special Interest Group on Knowledge Discovery and Data Mining) Explorations*, June, (1:1), 20-28.

Gray P., Watson, H.J. (1998a), "Professional Briefings...Present and Future Directions in Data Warehousing", *Database for Advances in Information Systems*, Summer, (29:3), 83-90.

Gray, P., Watson, H.J. (1998b), Decision Support in the Data Warehouse, Upper Saddle River, N.J.

Gray, P. (1997) " Mining for Data Warehousing Gems," *Information Systems Management*, Winter, 82-86.

Han, J., Kamber, M. (2001), Data Mining: Concepts and Techniques, Morgan-Kaufmann Academic Press, San Francisco.

Hand, D. J. (1998), "Data Mining: Statistics and More?", *The American Statistician*, May (52:2), 112-118.

Johnson, R. & Wicheren,D.W. (1998). Applied Multivariate Statistical Analysis. Prentice Hall, New York.

Kennedy, R. L., Lee, Y. Roy, B. V. Reed, C. D. & Lippman, R. P. (1997). Solving Data Mining Problems Through Pattern Recognition. New Jersey: Prentice Hall Professional Technical Reference.

Kosala, R., Blockeel, H. (2000), "Web Mining Research: A Survey", *ACM SIGKDD (Special Interest Group on Knowledge Discovery and Data Mining) Explorations*, June, (2:1), 1-10.

Langley, P., Simon, H. (1995), "Applications of Machine Learning and Rule Induction", *Communications of the ACM*, November, 55-65.

Moeller, R. A. (2001), "Distributed Data Warehousing Using Web Technology", AMACOM, New York .

Peacock, P. R. (1998a) "Data Mining in Marketing: Part 1", *Marketing Management*, Winter, 9-18.

Peacock, P. R. (1998b) "Data Mining in Marketing: Part 2", *Marketing Management*, Spring, 15-25.

Rajagopalan, B., Krovi, R. (2002), "Benchmarking Data Mining Algorithms", *Journal of Database Management*, Jan-Mar, 13, 25-36

Ranjit, B., Sugumaran, V. (1999), "Application of Intelligent Agent Technology for Managerial Data Analysis and Mining", *Database for Advances in Information Systems*, (30:1), 77-94.

Sharma, S., "Applied Multivariate Techniques", John Wiley & Sons, Inc. (1996).

Srivastava, J., Cooley, R., Deshpande, M., Tan, P., "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data", *ACM SIGKDD (Special Interest Group on Knowledge Discovery and Data Mining) Explorations*, January, (1:2)

Tegarden, D.J. (1999) "Business Information Visualization" *Communications of AIS* (1)4

Wells, M. T. (1999), "Feature Extraction Construction and Selection: A Data Mining Perspective", *Journal of the American Statistical Association*, (94:448), 1390.

White, H.,"A Reality Check for Data Snooping" (2000), *Econometrica*, (68:5), September, 1097-1126.

Witten, I. H. (2000), Data mining : practical machine learning tools and techniques with Java implementations, Morgan Kaufman, San Francisco.

ADDITIONAL READING

Listed below is a sampling of resources for additional information related to data mining**.**

DATA MINING CONFERENCES

International Conference on Knowledge Discovery and Data Mining (ACM SIGKDD)

Pacific Asian conference on Knowledge Discovery and Data Mining (PAKDD)

The European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD)

The International Conference on Data Warehousing and Knowledge Discovery

GENERAL READING ON DATA MINING

ACM SIGKDD Explorations – Newsletter of the Special Interest Group (SIG) on Knowledge Discovery & Data Mining

Berry, M. J., Linoff, G. S. (2000), Mastering Data Mining: The Art and Science of Customer Relationship Management. Wiley Computer Publishing, New York.

Cabena, P., Hadjinian, P., Stadler, R., Verhees, J., Zanasi, A. (1998), Discovering Data Mining: From Concept to Implementation. Upper Saddle River, NJ: Prentice Hall.

Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R, Editors (1996), Advances in Knowledge Discovery and Data Mining, MIT Press, Cambridge, MA

Fayyad, U., Piatetsky-Shapiro, G., and Smyth, R (1996). "The KDD Process for Extracting Useful Knowledge from Volumes of Data," *Communications of the ACM*, (39:11), pp. 27-34.

Groth, R. (1998) Data Mining: A Hands-on Approach for Business Professional, Upper Saddle River, NJ

Piatetsky-Shapiro, G and Frawley W. J. (1991), Knowledge Discovery in Databases, MIT Press, Cambridge, MA

A SLIGHTLY MORE TECHNICAL APPROACH TO DATA MINING

Agrawal, R., Imielinski, T., Swami, A.(1993), "Database Mining: A Performance Perspective, *IEEE Transactions Knowledge and Data Engineering*, (5), 914-925.

Chen, M.-S., Jan, J., Yu, P.S. (1996) "Data Mining: An Overview from a Database Perspective", *IEEE Transactions on Knowledge and Data Engineering*", (8:6), 866.883.

Fayyad, U., (2001), "The Digital Physics of Data Mining", *Communications of the ACM*, March, (44:3), 62-65.

Han, J., Kamber, M. (2001), Data Mining: Concepts and Techniques, Morgan-Kaufmann Academic Press, San Francisco.

Lee, C. (2001), "The GeneMine System for Genome/Preteome Annotation and Collaborative Data Mining", *IBM Systems Journal*, (40:2), 592-604.

Moore, A., Lee, M. S. (1998), "Cached Sufficient Statistics for Efficient Machine Learning with Large Datasets", *Journal of Artificial Intelligence Research*, (8), 67.91.

Witten, I. H. (2000), Data Mining : Practical Machine Learning Tools and Techniques With Java Implementations, Morgan Kaufman, San Francisco.

EMPIRICAL DATA MINING STUDIES

Lee, K.C. Hall, I. & Kwon, Y. (1996). "Hybrid Neural Network Models For Bankruptcy Predictions", *Decision Support Systems*, 18(1), 63-73.

Kumar, N. K., R. & Rajagopalan, B. (1997), "Financial Decision Support With Hybrid Genetic And Neural Based Modeling Tools", *European Journal of Operational Research*, 103(2), 339-349.

Nazem, S. & Shin, B. (1999), "Data Mining: New Arsenal For Strategic Decision Making", *Journal of Database Management*. 10(1), 39-42.

Brachman, R.J. Khabaza, T. Kloesgen, W. Piatetsky-Shapiro, G. & Simoudis, E. (1996). "Mining Business Databases", *Communications of the ACM*, 39(11), 42-48.

Perkowitz, M. (1999), "Towards Adaptive Web Sites: Conceptual Framework and Case Study", *Computer Networks*, 1245-1261.

Ranjit, B., Sugumaran, V. (1999), "Application of Intelligent Agent Technology for Managerial Data Analysis and Mining", *Database for Advances in Information Systems*, (30:1), 77-94.

Spangler, W. E.; May, J. H., Vargas, L. G. (1999), "Choosing Data-Mining Methods For Multiple Classification: Representational And Performance Measurement Implications For Decision Support ",*Journal of Management Information Systems*, Summer, 37-62.

Tam, K.Y. & Kiang, M.Y. (1992). "Managerial Applications Of Neural Networks: The Case Of Bank Failure Predictions", *Decision Sciences*, 38(7), 926-948.

DATA MINING FROM A STATISTICAL PERSPECTIVE

Glymour, C., Madigan D., et al (1996), "Statistical Inference and Data Mining". *Communications of the ACM*, (39:11), pp. 35-41.

Hand, D. J. (1998), "Data Mining: Statistics and More?", *The American Statistician*, May (52:2), 112-118.

Moore, A., Lee, M. S. (1998), "Cached Sufficient Statistics for Efficient Machine Learning with Large Datasets", *Journal of Artificial Intelligence Research*, (8), 67.91.

Wells, M. T. (1999), "Feature Extraction Construction and Selection: A Data Mining Perspective", *Journal of the American Statistical Association*, (94:448), p. 1390.

Ye, Jianming (1998), "On Measuring and Correcting the Effects of Data Mining and Model Selection", *Journal of the American Statistical Association*, (93:441), 120-131

DATA MINING FROM A DATABASE/DATA WAREHOUSING PERSPECTIVE

Agrawal, R., Imielinski, T., Swami, A.(1993), "Database Mining: A Performance Perspective, IEEE Transactions Knowledge and Data Engineering, (5), 914-925.

Chen, M.-S., Jan, J., Yu, P.S. (1996) "*Data Mining: An Overview from a Database Perspective*", IEEE Transactions on Knowledge and Data Engineering", (8:6), 866.883.

Moeller, R. A. (2001), "Distributed Data Warehousing Using Web Technology", AMACOM, New York .

Gray P., Watson, H.J. (1998a), "Professional Briefings...Present and Future Directions in Data Warehousing", *Database for Advances in Information Systems*, Summer, (29:3), 83-90.

Gray, P., Watson, H.J. (1998b), Decision Support in the Data Warehouse, Upper Saddle River, N.J.

Gray, P. (1997)"BOOKISMS: Mining for Data Warehousing Gems," *Information Systems Management*, Winter, 82-86.

Moeller, R. A. (2001), Distributed Data Warehousing Using Web Technology, American Management Association (AMACOM) , New York.

DATA MINING FROM A MACHINE LEARNING PERSPECTIVE

Burges, C. J. (1998), "A Tutorial on Support Vector Machines for Pattern Recognition", Data Mining and Knowledge Discovery, (2:2),

Kennedy, R. L. Lee, Y. Roy, B. V. Reed, C. D. & Lippman, R. P. (1997). Solving Data Mining Problems Through Pattern Recognition. New Jersey: Prentice Hall Professional Technical Reference.

Langley, P., Simon, H. A. (1995), "Application of Machine Learning and Rule Induction", *Communications of the ACM*, (38:11), 55.64.

Moore, A., Lee, M. S. (1998), "Cached Sufficient Statistics for Efficient Machine Learning with Large Datasets", *Journal of Artificial Intelligence Research*, (8), 67.91.

Witten, I. H. (2000), Data Mining : Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufman, San Francisco.

DATA MINING FROM AN INTERNET/WEB-BASED PERSPECTIVE

Edelstein, H., A. (2001), "Pan for Gold in the Clickstream", *Informationweek.com*, March 12,2001, 77-91

Lee, C. (2001), "The GeneMine System for Genome/Preteome Annotation and Collaborative Data Mining", IBM Systems Journal, (40:2), 592-604.

Moeller, R. A. (2001), Distributed Data Warehousing Using Web Technology, American Management Association (AMACOM) , New York.

Perkowitz, M. (1999), "Towards Adaptive Web Sites: Conceptual Framework and Case Study", *Computer Networks*, 1245-1261.

Srivastava, J., Cooley, R., Deshpande, M., Tan, P., " Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data", ACM SIGKDD (Special Interest Group on Knowledge Discovery and Data Mining) Explorations, January, (1:2).

Tao, G. (1999), “KPS A Web Information Mining Algorithm”, Computer Networks, (31), 1495-1508.

ABOUT THE AUTHOR

Joyce Jackson is a doctoral student in MIS at the University of South Carolina. She received a BS in Computer Science from Point Park College and an MBA from Penn State University. Prior to returning to school to pursue her doctorate, Joyce spent ten years designing and developing business information systems for General Electric, Champion International, and Mellon Bank. In addition, she designed and taught courses in Applications of Distributed Systems, C/C++ and Micro-Focus COBOL. Her current research interests include data mining, knowledge management from a data mining perspective, and organizational strategy in the electronic exchange environment.

Communications of the Association for Information Systems

ISSN: 1529-3181