101

JOURNAL OF DATA WAREHOUSING

Volume 5
Number 4
Fall 2000

A 101communications Publication

CONTENTS

THE DATA WAREHOUSING INSTITUTE™

JOURNAL OF DATA WAREHOUSING

Volume 5 Number 4 Fall 2000

THE **DATA WAREHOUSING** INSTITUTE™

The Data Warehousing Institute's Mission

The Data Warehousing Institute™ (TDWI), a division of 101communications, is the premier provider of in-depth, high quality education and training in the data warehousing and business intelligence industry. TDWI is dedicated to educating business and information technology professionals about the strategies, techniques, and tools required to successfully design, build, and maintain data warehousing implementations, and also to the advancement of data warehousing research, knowledge transfer, and the professional development of its Members. TDWI sponsors and promotes a worldwide membership program, annual educational conferences, regional educational seminars, onsite courses, solution provider partnerships, awards programs for the best practices in data warehousing and innovative technologies, resourceful publications, an in-depth research program, and a comprehensive Web site.

The Data Warehousing Institute supports, develops, and distributes a wide range of publications to keep its Members up-to-date on the new techniques, events, and issues in data warehousing, as well as the trends within the vendor community. These publications are available to Members at a special discounted rate and include textbooks, journals, reports, white papers, and newsletters in both paper and electronic formats. These publications include: semiannual *Data Warehousing: What Works?*™ Corporate Case Study Compendium, annual *Information Strategies Resource Guide,* annual *Data Warehousing Perspectives,* quarterly *Journal of Data Warehousing,* quarterly *TDWI Member Newsletter*, semimonthly *FlashPoint* Electronic Bulletin, Annual *Data Warehousing Salaries, Roles, and Responsiblities* Report, Quarterly *Ten Mistakes to Avoid* Series, Data Warehousing Textbooks, and Coursebooks from TDWI's four annual Education and Training Conferences.

The *Journal of Data Warehousing* (article submission inquiries)
Theresa Johnston
The Data Warehousing Institute
5200 Southcenter Boulevard, Suite 250
Seattle, WA 98188-2356
206.246.5059, Ext. 109
Fax: 206.246.5952
Email: tjohnston@dw-institute.com
www.dw-institute.com/journal.htm

The Data Warehousing Institute (subscription inquiries)
The Data Warehousing Institute
Post Office Box 15896
North Hollywood, CA 91615-5896
Toll Free: 877.902.9760
Local: 818.487.4574
Fax: 818.487.4550
Email: publications@dw-institute.com
www.dw-institute.com

JOURNAL OF DATA WAREHOUSING
Volume 5 Number 4 Fall 2000

Statement of Purpose

Hugh J. Watson

Many organizations are committing considerable human, technical, and financial resources to build and use data warehouses. The primary purpose of these efforts is to provide easy access to specially prepared data that can be used with decision support applications, such as management reporting, queries, decision support systems, executive information systems, and data mining.

Data warehousing is broad in scope, including: extracting data from legacy systems and other data sources; cleansing, scrubbing and preparing data for decision support; maintaining data in appropriate data stores; accessing and analyzing data using a variety of end user tools; and mining data for significant relationships. It obviously involves technical, organizational, and financial considerations.

Research in this area is limited, and the findings are scattered over a variety of publications, many of them difficult to identify and locate. The purpose of the *Journal of Data Warehousing* is to provide a focal point to support and disseminate knowledge on data warehousing.

The *Journal of Data Warehousing* encourages submissions, including surveys of current practices, opinion pieces, conceptual frameworks, case studies that describe innovative practices or provide important insights, tutorials, technology discussions, technology forecasts, annotated bibliographies, and other data warehousing-related topics. We are inclusive in outreach: every submission is reviewed and considered. If it is important to the data warehousing community, the *Journal of Data Warehousing* is interested in publishing it.

THE DATA WAREHOUSING INSTITUTE™

JOURNAL OF DATA WAREHOUSING

Volume 5 Number 4 Fall 2000

The CRISP-DM Model: The New Blueprint for Data Mining

Colin Shearer

Abstract

This article describes CRISP-DM (CRoss-Industry Standard Process for Data Mining), a non-proprietary, documented, and freely available data mining model. Developed by industry leaders with input from more than 200 data mining users and data mining tool and service providers, CRISP-DM is an industry-, tool-, and application-neutral model. This model encourages best practices and offers organizations the structure needed to realize better, faster results from data mining.

CRISP-DM organizes the data mining process into six phases: business understanding, data understanding, data preparation, modeling, evaluation, and deployment. These phases help organizations understand the data mining process and provide a road map to follow while planning and carrying out a data mining project. This article explores all six phases, including the tasks involved with each phase. Sidebar material, which takes a look at specific data mining problem types and techniques for addressing them, is provided.

In 1996, while interest in data mining was mounting, no widely accepted approach to data mining existed. There was a clear need for a data mining process model that would standardize the industry and help organizations launch their own data mining projects. The development of a non-proprietary, documented, and freely available model would enable organizations to realize better results from data mining, encourage best practices in the industry, and help bring the market to maturity.

CRISP-DM (CRoss-Industry Standard Process for Data Mining) was conceived in late 1996 by four leaders of the nascent data mining market: Daimler-Benz (now DaimlerChrysler), Integral Solutions Ltd. (ISL), NCR, and OHRA. At the time, Daimler-Benz led most industrial and commercial organizations in applying data mining in its business operations. ISL (which SPSS Inc. purchased in 1998) first provided services based on data mining principles in 1990 and launched Clementine—the first commercial data mining workbench—in 1994. NCR, aiming to deliver added value to its Teradata data warehouse customers, met its clients' needs with teams of data mining consultants. OHRA, one of the largest Dutch insurance companies, provided a valuable testing ground for live, large-scale data mining projects.

A year later, a consortium formed with the goal of developing CRISP-DM. As CRISP-DM was intended to be industry-, tool-, and application-neutral, the consortium solicited input from a wide range of practitioners and others (such as data warehouse vendors and management consultants) with a vested interest in data mining. To gain this insight, the CRISP-DM Special Interest Group, or SIG, was created with the goal of developing a standard process model to service the data mining community.

During the next several years, the CRISP-DM SIG developed and refined the model. Several trials took place in live data mining projects at Daimler-Benz and OHRA, and commercial data mining tools began adopting CRISP-DM. The SIG proved invaluable, growing to more than 200 members and holding workshops in London, New York, and Brussels.

In 2000, the presentation of the next generation of CRISP-DM—version 1.0—reflects significant progress in the development of a standardized data processing model. While future extensions and improvements are certainly expected, industry players are quickly accepting the CRISP-DM methodology.

The CRISP-DM Model, *continued*

The CRISP-DM Reference Model

Put simply, CRISP-DM is a comprehensive data mining methodology and process model that provides anyone—from novices to data mining experts—with a complete blueprint for conducting a data mining project. CRISP-DM breaks down the life cycle of a data mining project into six phases: business understanding, data understanding, data preparation, modeling, evaluation, and deployment.

Figure 1. Phases of the CRISP-DM Reference Model

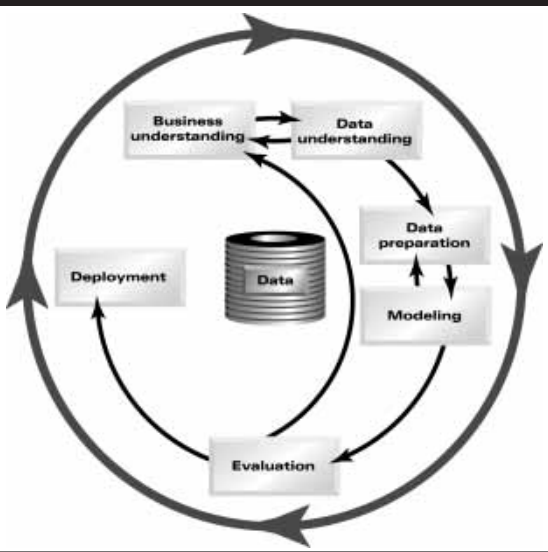

Figure 1 shows the phases of a data mining process. The arrows indicate the most important and frequent dependencies between the phases, while the outer circle symbolizes the cyclical nature of data mining itself and illustrates that the lessons learned during the data mining process and from the deployed solution can trigger new, often more focused business questions. Figure 2 outlines each phase of the data mining process.

Phase One: Business Understanding

Perhaps the most important phase of any data mining project, the initial business understanding phase focuses on understanding the project objectives from a business perspective, converting this knowledge into a data mining problem definition, and then developing a preliminary plan designed to achieve the objectives. In order to understand which data should later be analyzed, and how, it is vital for data mining practitioners to fully understand the business for which they are finding a solution.

The business understanding phase involves several key steps, including determining business objectives, assessing the situation, determining the data mining goals, and producing the project plan.

Figure 2. Tasks and Outputs of the CRISP-DM Reference Model

Business Understanding

Determine Business Objectives
- *Background*
- *Business Objectives*
- *Business Success Criteria*

Assess Situation
- *Inventory of Resources*
- *Requirements, Assumptions, and Constraints*
- *Risks and Contingencies*
- *Terminology*
- *Costs and Benefits*

Determine Data Mining Goals
- *Data Mining Goals*
- *Data Mining Success Criteria*

Produce Project Plan
- *Project Plan*
- *Initial Assessment of Tools and Techniques*

Data Understanding

Collect Initial Data
- *Initial Data Collection Report*

Describe Data
- *Data Description Report*

Explore Data
- *Data Exploration Report*

Verify Data Quality
- *Data Quality Report*

Data Preparation

Data Set
- *Data Set Description*

Select Data
- *Rationale for Inclusion / Exclusion*

Clean Data
- *Data Cleaning Report*

Construct Data
- *Derived Attributes*
- *Generated Records*

Integrate Data
- *Merged Data*

Format Data
- *Reformatted Data*

Modeling

Select Modeling Technique
- *Modeling Technique*
- *Modeling Assumptions*

Generate Test Design
- *Test Design*

Build Model
- *Parameter Settings*
- *Models*
- *Model Description*

Assess Model
- *Model Assessment*
- *Revised Parameter Settings*

Evaluation

Evaluate Results
- *Assessment of Data Mining Results w.r.t. Business Success Criteria*
- *Approved Models*

Review Process
- *Review of Process*

Determine Next Steps
- *List of Possible Actions*
- *Decision*

Deployment

Plan Deployment
- *Deployment Plan*

Plan Monitoring and Maintenance
- *Monitoring and Maintenance Plan*

Produce Final Report
- *Final Report*
- *Final Presentation*

Review Project
- *Experience Documentation*

Determine the Business Objectives
Understanding a client's true goal is critical to uncovering the important factors involved in the planned project—and to ensuring that the project does not result in producing the right answers to the wrong questions. To accomplish this, the data analyst must uncover the primary business objective as well as the related questions the business would like to address.

For example, the primary business goal could be to retain current customers by predicting when they are prone to move to a competitor. Examples of related business questions might be, ''How does the primary channel (e.g., ATM, branch visit, Internet) of a bank customer affect whether they stay or go?'' or ''Will lower ATM fees significantly reduce the number of high-value customers who leave?'' A secondary issue might be to determine whether lower fees affect only one particular customer segment.

Finally, a good data analyst always determines the measure of success. Success may be measured by reducing lost customers by 10 percent or simply by achieving a better understanding of the customer base. Data analysts should beware of setting unattainable goals and should make sure that each success criterion relates to at least one of the specified business objectives.

Assess the Situation
In this step, the data analyst outlines the resources, from personnel to software, that are available to accomplish the data mining project. Particularly important is discovering what data is available to meet the primary business goal. At this point, the data analyst also should list the assumptions made in the project—assumptions such as, ''To address the business question, a minimum number of customers over age 50 is necessary.'' The data analyst also should list the project risks, list potential solutions to those risks, create a glossary of business and data mining terms, and construct a cost-benefit analysis for the project.

Determine the Data Mining Goals
The data mining goal states project objectives in business terms such as, ''Predict how many widgets a customer will buy given their purchases in the past three years, demographic information (age, salary, city, etc.), and the item price.'' Success also should be defined in these terms—for instance, success could be defined as achieving a certain level of predictive accuracy. If the business goal cannot be effectively translated into a data mining goal, it may be wise to consider redefining the problem at this point.

Produce a Project Plan
The project plan describes the intended plan for achieving the data mining goals, including outlining specific steps and a proposed timeline, an assessment of potential risks, and an initial assessment of the tools and techniques needed to support the project. Generally accepted industry timeline standards are: 50 to 70 percent of the time and effort in a data mining project involves the Data Preparation Phase; 20 to 30 percent involves the Data Understanding Phase; only 10 to 20 percent is spent in each of the Modeling, Evaluation, and Business Understanding Phases; and 5 to 10 percent is spent in the Deployment Planning Phase.

Phase Two: Data Understanding
The data understanding phase starts with an initial data collection. The analyst then proceeds to increase familiarity with the data, to identify data quality problems, to discover initial insights into the data, or to detect interesting subsets to form hypotheses about hidden information. The data understanding phase involves four steps, including the collection of initial data, the description of data, the exploration of data, and the verification of data quality.

Collect the Initial Data
Here a data analyst acquires the necessary data, including loading and integrating this data if necessary. The analyst should make sure to report problems encountered and his or her solutions to aid with future replications of the project. For instance, data may have to be collected from several different sources, and some of these sources may have a long lag time. It is helpful to know this in advance to avoid potential delays.

Describe the Data
During this step, the data analyst examines the ''gross'' or ''surface'' properties of the acquired data and reports on the results, examining issues such as the format of the data, the quantity of the data, the number of records and fields in each table, the identities of the fields, and any other surface features of the data. The key question to ask is: Does the data acquired satisfy the relevant requirements? For instance, if age is an important field and the data does not reflect the entire age range, it may be wise to collect a different set of data. This step also provides a basic understanding of the data on which subsequent steps will build.

The CRISP-DM Model, *continued*

Explore the Data

This task tackles the data mining questions, which can be addressed using querying, visualization, and reporting. For instance, a data analyst may query the data to discover the types of products that purchasers in a particular income group usually buy. Or the analyst may run a visualization analysis to uncover potential fraud patterns. The data analyst should then create a data exploration report that outlines first findings, or an initial hypothesis, and the potential impact on the remainder of the project.

Verify Data Quality

At this point, the analyst examines the quality of the data, addressing questions such as: Is the data complete? Missing values often occur, particularly if the data was collected across long periods of time. Some common items to check include: missing attributes and blank fields; whether all possible values are represented; the plausibility of values; the spelling of values; and whether attributes with different values have similar meanings (e.g., low fat, diet). The data analyst also should review any attributes that may give answers that conflict with common sense (e.g., teenagers with high income).

Phase Three: Data Preparation

The data preparation phase covers all activities to construct the final data set or the data that will be fed into the modeling tool(s) from the initial raw data. Tasks include table, record, and attribute selection, as well as transformation and cleaning of data for modeling tools. The five steps in data preparation are the selection of data, the cleansing of data, the construction of data, the integration of data, and the formatting of data.

Select Data

Deciding on the data that will be used for the analysis is based on several criteria, including its relevance to the data mining goals, as well as quality and technical constraints such as limits on data volume or data types. For instance, while an individual's address may be used to determine which region that individual is from, the actual street address data can likely be eliminated to reduce the amount of data that must be evaluated. Part of the data selection process should involve explaining why certain data was included or excluded. It is also a good idea to decide if one or more attributes are more important than others are.

Clean Data

Without clean data, the results of a data mining analysis are in question. Thus at this stage, the data analyst must either select clean subsets of data or incorporate more ambitious techniques such as estimating missing data through modeling analyses. At this point, data analysts should make sure they outline how they addressed each quality problem reported in the earlier "Verify Data Quality" step.

Construct Data

After the data is cleaned, the data analyst should undertake data preparation operations such as developing entirely new records or producing derived attributes. An example of a new record would be the creation of an empty purchase record for customers who made no purchases during the past year. Derived attributes, in contrast, are new attributes that are constructed from existing attributes, such as Area = Length x Width. These derived attributes should only be added if they ease the model process or facilitate the modeling algorithm, not just to reduce the number of input attributes. For instance, perhaps "income per head" is a better/easier attribute to use than "income per household." Another type of derived attribute is single-attribute transformations, usually performed to fit the needs of the modeling tools. These transformations may be necessary to transform ranges to symbolic fields (e.g., ages to age bands), or symbolic fields ("definitely yes," "yes," "don't know," "no") to numeric values. Modeling tools or algorithms often require these transformations.

Integrate Data

Integrating data involves combining information from multiple tables or records to create new records or values. With table-based data, an analyst can join two or more tables that have different information about the same objects. For instance, a retail chain has one table with information about each store's general characteristics (e.g., floor space, type of mall), another table with summarized sales data (e.g., profit, percent change in sales from previous year), and another table with information about the demographics of the surrounding area. Each of these tables contains one record for each store. These tables can be merged together into a new table with one record for each store, combining fields from the source tables.

Data integration also covers aggregations. Aggregations refer to operations where new values are computed by summarizing information from multiple records and/or tables. For example, an aggregation could include converting a table of customer purchases, where there is one record for each purchase, into a new table where there is one record for each customer. The table's fields could include the number of purchases, the average purchase amount, the percent of orders charged to credit cards, the percent of items under promotion, etc.

Format Data
In some cases, the data analyst will change the format or design of the data. These changes might be simple—for example, removing illegal characters from strings or trimming them to a maximum length—or they may be more complex, such as those involving a reorganization of the information. Sometimes these changes are needed to make the data suitable for a specific modeling tool. In other instances, the changes are needed to pose the necessary data mining questions.

Phase Four: Modeling
In this phase, various modeling techniques are selected and applied and their parameters are calibrated to optimal values. Typically, several techniques exist 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 may be necessary. Modeling steps include the selection of the modeling technique, the generation of test design, the creation of models, and the assessment of models.

Select the Modeling Technique
This task refers to choosing one or more specific modeling techniques, such as decision tree building with C4.5 or neural network generation with back propagation. If assumptions are attached to the modeling technique, these should be recorded.

Generate Test Design
After building a model, the data analyst must test the model's quality and validity, running empirical testing to determine the strength of the model. In supervised data mining tasks such as classification, it is common to use error rates as quality measures for data mining models. Therefore, we typically separate the data set into train and test set, build the model on the train set, and estimate its quality on the separate test set. In other words, the data analyst develops the model based on one set of existing data and tests its validity using a separate set of data. This enables the data analyst to measure how well the model can predict history before using it to predict the future. It is usually appropriate to design the test procedure before building the model; this also has implications for data preparation.

Build the Model
After testing, the data analyst runs the modeling tool on the prepared data set to create one or more models.

Assess the Model
The data mining analyst interprets the models according to his or her domain knowledge, the data mining success criteria, and the desired test design. The data mining analyst judges the success of the application of modeling and discovery techniques technically, but he or she should also work with business analysts and domain experts in order to interpret the data mining results in the business context. The data mining analyst may even choose to have the business analyst involved when creating the models for assistance in discovering potential problems with the data.

For example, a data mining project may test the factors that affect bank account closure. If data is collected at different times of the month, it could cause a significant difference in the account balances of the two data sets collected. (Because individuals tend to get paid at the end of the month, the data collected at that time would reflect higher account balances.) A business analyst familiar with the bank's operations would note such a discrepancy immediately.
In this phase, the data mining analyst also tries to rank the models. He or she assesses the models according to the evaluation criteria and takes into account business objectives and business success criteria. In most data mining projects, the data mining analyst applies a single technique more than once or generates data mining results with different alternative techniques. In this task, he or she also compares all results according to the evaluation criteria.

Phase Five: Evaluation
Before proceeding to final deployment of the model built by the data analyst, it is important to more thoroughly evaluate the model and review the model's construction to be certain it properly achieves the business objectives. Here it is critical to determine if some important business issue has not been sufficiently

The CRISP-DM Model, *continued*

considered. At the end of this phase, the project leader then should decide exactly how to use the data mining results. The key steps here are the evaluation of results, the process review, and the determination of next steps.

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 determines if there is some business reason why this model is deficient. Another option here is to test the model(s) on real-world applications—if time and budget constraints permit. Moreover, evaluation also seeks to unveil additional challenges, information, or hints for future directions.

At this stage, the data analyst summarizes the assessment results in terms of business success criteria, including a final statement about whether the project already meets the initial business objectives.

Review Process
It is now appropriate to do a more thorough review of the data mining engagement to determine if there is any important factor or task that has somehow been overlooked. This review also covers quality assurance issues (e.g., did we correctly build the model? Did we only use allowable attributes that are available for future deployment?).

Determine Next Steps
At this stage, the project leader must decide whether to finish this project and move on to deployment or whether to initiate further iterations or set up new data mining projects.

Phase Six: Deployment
Model creation is generally not the end of the project. The knowledge gained must be organized and presented in a way that the customer can use it, which often involves applying "live" models within an organization's decision-making processes, such as the real-time personalization of Web pages or repeated scoring of marketing databases.

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 across the enterprise. Even though it is often the customer, not the data analyst, who carries out the deployment steps, it is important for the customer to understand up front what actions must be taken in order to actually make use of the created models. The key steps here are plan deployment, plan monitoring and maintenance, the production of the final report, and review of the project.

Plan Deployment
In order to deploy the data mining result(s) into the business, this task takes the evaluation results and develops a strategy for deployment.

Plan Monitoring and Maintenance
Monitoring and maintenance are important issues if the data mining result is to become part of the day-to-day business and its environment. A carefully prepared maintenance strategy avoids incorrect usage of data mining results.

Produce Final Report
At the end of the project, the project leader and his or her 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). This report includes all of the previous deliverables and summarizes and organizes the results. Also, there often will be a meeting at the conclusion of the project, where the results are verbally presented to the customer.

Review Project
The data analyst should assess failures and successes as well as potential areas of improvement for use in future projects. This step should include a summary of important experiences during the project and can include interviews with the significant project participants. This document could include pitfalls, misleading approaches, or hints for selecting the best-suited data mining techniques in similar situations. In ideal projects, experience documentation also covers any reports written by individual project members during the project phases and tasks.

Conclusion

CRISP-DM was designed to provide guidance to data mining beginners and to provide a generic process model that can be specialized according to the needs of any particular industry or company. The industry's initial use of the methodology confirms that it is a valuable aid to beginners and advanced data miners alike. At OHRA, a new employee used the CRISP-DM process model to plan and guide a highly successful data mining project. In addition, DaimlerChrysler has adapted CRISP-DM to develop its own specialized customer relationship management (CRM) tool to improve customer marketing.

SPSS's and NCR's Professional Services groups have adopted CRISP-DM and have used it successfully on numerous customer engagements covering many industries and business problems. Service suppliers from outside the consortium have adopted CRISP-DM; repeated references to the methodology by analysts have established it as the *de facto* standard for the industry; and customers have exhibited a growing awareness of its importance (CRISP-DM is now frequently referenced in invitations to tender and RFP documents).

CRISP-DM was not built in a theoretical, academic manner, working from technical principles; nor did elite committees of gurus create it behind closed doors. CRISP-DM succeeds because it is soundly based on practical, real-world data mining experience. In that respect, the data mining industry is overwhelmingly indebted to the many practitioners who contributed their efforts and their ideas throughout the CRISP-DM project.

The CRISP-DM process model is not meant to be a magical instruction book that will instantly make the most inexperienced novice succeed in data mining. However, combined with training in data mining methodology and techniques, as well as assistance from more experienced practitioners, it can be a valuable tool to help less experienced data mining analysts understand the value and the steps involved in the entire data mining process.

A Look at Data Mining Problem Types

Outlined below are several different types of data mining techniques that together can be used to solve a business problem.

Data Description and Summarization

Data Description and Summarization provides a concise description of the characteristics of data, typically in elementary and aggregated form, to give users an overview of the data's structure. Data description and summarization alone can be an objective of a data mining project. For instance, a retailer might be interested in the turnover of all outlets, broken down by categories, summarizing changes and differences as compared to a previous period. In almost all data mining projects, data description and summarization is a sub-goal in the process, typically in early stages where initial exploratory data analysis can help to understand the nature of the data and to find potential hypotheses for hidden information. Summarization also plays an important role in the presentation of final results.

Many reporting systems, statistical packages, OLAP, and EIS systems can cover data description and summarization but do not usually provide any methods to perform more advanced modeling. If data description and summarization is considered a stand-alone problem type and no further modeling is required, these tools also are appropriate to carry out data mining engagements.

Segmentation

The data mining problem type *segmentation* separates the data into interesting and meaningful subgroups or classes that share common characteristics. For instance, in shopping basket analysis, one could define segments of baskets, depending on the items they contain. An analyst can segment certain subgroups as relevant for the business question, based on prior knowledge or based on the outcome of data description and summarization. However, there also are automatic clustering techniques that can detect previously unsuspected and hidden structures in data that allow segmentation.

Segmentation can be a data mining problem type of its own when the detection of segments is the main purpose. For example, all addresses in ZIP code areas with higher than average age and income might be selected for mailing advertisements on home nursing insurance. However, segmentation often is a step

The CRISP-DM Model, *continued*

toward solving other problem types where the purpose is to keep the size of the data manageable or to find homogeneous data subsets that are easier to analyze.

Appropriate techniques
- Clustering techniques
- Neural nets
- Visualization

Example
A car company regularly collects information about its customers concerning their socioeconomic characteristics. Using cluster analysis, the company can divide its customers into more understandable subgroups, analyze the structure of each subgroup, and deploy specific marketing strategies for each group separately.

Concept Descriptions
Concept description aims at an *understandable* description of concepts or classes. The purpose is not to develop complete models with high prediction accuracy, but to gain insights. For instance, a company may be interested in learning more about their loyal and disloyal customers. From a description of these concepts (loyal and disloyal customers), the company might infer what could be done to keep customers loyal or to transform disloyal customers to loyal customers. Typically, segmentation is performed before concept description. Some techniques, such as conceptual clustering techniques, perform segmentation and concept description at the same time.

Concept descriptions also can be used for classification purposes. On the other hand, some classification techniques produce understandable classification models, which then can be considered concept descriptions. The important distinction is that classification aims to be complete in some sense. The classification model needs to apply to *all* cases in the selected population. On the other hand, concept descriptions need not be complete. It is sufficient if they describe important parts of the concepts or classes.

Appropriate techniques
- Rule induction methods
- Conceptual clustering

Example
Using data about the buyers of new cars and using a rule induction technique, a car company could generate rules that describe its loyal and disloyal customers. Below are simplified examples of the generated rules:

If SEX = male and AGE > 51 then CUSTOMER = loyal
If SEX = female and AGE > 21 then CUSTOMER = loyal

Classification
Classification assumes that there is a set of objects—characterized by some attribute or feature—which belong to different classes. The class label is a discrete (symbolic) value and is known for each object. The objective is to build classification models (sometimes called classifiers) that assign the correct class label to previously unseen and unlabeled objects. Classification models are mostly used for predictive modeling.

Many data mining problems can be transformed to classification problems. For example, credit scoring tries to assess the credit risk of a new customer. This can be transformed to a classification problem by creating two classes—good and bad customers. A classification model can be generated from existing customer data and their credit behavior. This classification model then can be used to assign a new potential customer to one of the two classes and hence accept or reject him or her. Classification has connections to almost all other problem types.

Appropriate techniques
- Discriminant analysis
- Rule induction methods
- Decision tree learning
- Neural nets
- K Nearest Neighbor
- Case-based reasoning
- Genetic algorithms

Example
Banks generally have information on the payment behavior of their credit applicants. By combining this financial information with other information about the customers, such as sex, age, income, etc., it is possible to develop a system to classify new customers as good or bad customers, (i.e., the credit risk in acceptance of a customer is either low or high, respectively).

Prediction

Another important problem type that occurs in a wide range of applications is *prediction*. Prediction is very similar to classification, but unlike classification, the target attribute (class) in prediction is not a qualitative discrete attribute but a continuous one. The aim of prediction is to find the numerical value of the target attribute for unseen objects. This problem type is sometimes called regression. If prediction deals with time series data, then it is often called forecasting.

Appropriate techniques

- Regression analysis
- Regression trees
- Neural nets
- K Nearest Neighbor
- Box-Jenkins methods
- Genetic algorithms

Example

The annual revenue of an international company is correlated with other attributes such as advertisement, exchange rate, inflation rate, etc. Having these values (or their reliable estimations for the next year), the company can predict its expected revenue for the next year.

Dependency Analysis

Dependency analysis finds a model that describes significant dependencies (or associations) between data items or events. Dependencies can be used to predict the value of a data item, given information on other data items. Although dependencies can be used for predictive modeling, they are mostly used for understanding. Dependencies can be strict or probabilistic.

Associations are a special case of dependencies that have recently become very popular. Associations describe affinities of data items (i.e., data items or events that frequently occur together). A typical application scenario for associations is the analysis of shopping baskets. There, a rule such as ''in 30 percent of all purchases, beer and peanuts have been bought together,'' is a typical example of an association. Algorithms for detecting associations are very fast and produce many associations. Selecting the most interesting ones is often a challenge.

Dependency analysis has close connections to prediction and classification, where dependencies are implicitly used for the formulation of predictive models. There also is a connection to concept descriptions, which often highlight dependencies. In applications, dependency analysis often co-occurs with segmentation. In large data sets, dependencies are seldom significant because many influences overlay each other. In such cases, it is advisable to perform a dependency analysis on more homogeneous segments of the data.

Sequential patterns are a special kind of dependencies where the order of events is considered. In the shopping basket domain, associations describe dependencies between items at a given time. Sequential patterns describe shopping patterns of one particular customer or a group of customers over time.

Appropriate Techniques

- Correlation analysis
- Regression analysis
- Association rules
- Bayesian networks
- Inductive Logic Programming
- Visualization techniques

Example

Using regression analysis, a business analyst might find a significant dependency between the total sales of a product and its price and the amount of the total expenditures for the advertisement. Once the analyst discovers this knowledge, he or she can reach the desired sales level by changing the price and/or the advertisement expenditure accordingly.

The CRISP-DM Model, *continued*

CRISP-DM Glossary

Activity – Part of a task in User Guide; describes actions to perform a task.

CRISP-DM methodology – The general term for all concepts developed and defined in CRISP-DM.

Data mining context – Set of constraints and assumptions such as problem type, techniques or tools, and application domain.

Data mining problem type – Class of typical data mining problems such as data description and summarization, segmentation, concept descriptions, classification, prediction, and dependency analysis.

Generic – A task that holds across all possible data mining projects as complete; i.e., applicable to both the whole data mining process and all possible data mining applications; i.e., valid for unforeseen developments such as new modeling techniques.

Model – Ability to apply to a data set to predict a target attribute; executable.

Output – Tangible result of performing a task.

Phase – High-level term for part of the process model; consists of related tasks.

Process instance – A specific project described in terms of the process model.

Process model – Defines the structure of data mining projects and provides guidance for their execution; consists of reference model and user guide.

Reference model – Decomposition of data mining projects into phases, tasks, and outputs.

Specialized – A task that makes specific assumptions in specific data mining contexts.

Task – Series of activities to produce one or more outputs; part of a phase.

User guide – Specific advice on how to perform data mining projects.

BIOGRAPHY

Colin Shearer *is Vice President, data mining business development with, SPSS Business Intelligence. Since 1984, he has been involved in applying advanced software solutions to solving business problems. Previously with SD-Scicon and Quintec Systems, he was one of the founders of Integral Solutions Ltd. (ISL) in 1989. A pioneer of data mining in the early 1990s, Shearer was the architect of ISL's award-winning Clementine system, the first data mining tool aimed at non-technologist end users, and led a team which tackled numerous successful data mining applications in areas including finance, broadcasting, market research, and defense. In December 1998, ISL was acquired by SPSS Inc., the world's leading supplier of analytical solutions for enterprises.*

SPSS Inc. Contact:
Matthew Martin
233 S. Wacker Drive
11th Floor
Chicago, IL 60606
312.651.3066
Email: mmartin@spss.com

The Data Warehousing Institute

The Data Warehousing Institute™ (TDWI), a division of 101communications, is the premier provider of in-depth, high quality education and training in the data warehousing and business intelligence industry. TDWI is dedicated to educating business and information technology professionals about the strategies, techniques, and tools required to successfully design, build, and maintain data warehousing implementations, and also to the advancement of data warehousing research, knowledge transfer, and the professional development of its Members. TDWI sponsors and promotes a worldwide membership program, annual educational conferences, regional educational seminars, onsite courses, solution provider partnerships, awards programs for the best practices in data warehousing and innovative technologies, resourceful publications, an in-depth research program, and a comprehensive Web site.

In the evolving and developing field of data warehousing, it is necessary for data warehousing and information technology professionals to connect and interact with one another. The Data Warehousing Institute provides these professionals with the opportunity to learn from each other, network, share ideas, and respond as a collective whole to the challenges and opportunities in the data warehousing industry.

Through Membership in TDWI, these professionals make positive contributions to the data warehousing industry and advance their professional development. TDWI Members benefit through increased knowledge of all the hottest trends in data warehousing, which makes TDWI Members some of the most valuable data warehousing professionals in the industry. TDWI Members are able to avoid common pitfalls, quickly learn data warehousing fundamentals, and network with peers and industry experts to give their data warehousing and companies a competitive edge in deploying a data warehousing solution.

The Data Warehousing Institute Membership includes more than 4,000 Members who are data warehousing and information technology professionals from Fortune 1000 corporations, consulting organizations, and governments in 45 countries.

THE **DATA WAREHOUSING** INSTITUTE™
5200 Southcenter Blvd., Suite 250, Seattle, Washington 98188, Phone: 206-246-5059, Fax: 206-246-5952
Email: info@dw-institute.com, Web: www.dw-institute.com

JOURNAL
OF DATA WAREHOUSING

Information About Subscription Rates

All subscription orders and changes of address should be sent to:
The Data Warehousing Institute
Post Office Box 15896
North Hollywood, CA 91615-5896
Toll Free: 877.902.9760
Local: 818.487.4574
Fax: 818.487.4550
Email: publications@dw-institute.com
www.dw-institute.com

THE **DATA WAREHOUSING** INSTITUTE™

849-J Quince Orchard Boulevard
Gaithersburg, MD 20878
301.947.3730
Fax: 301.947.3733
Email: info@dw-institute.com
www.dw-institute.com