# Mining Software Engineering Data: A Survey

## A DACS State-of-the-Art Report

Contract Number SPO700-98-D-4000
(Data & Analysis Center for Software)

**Prepared for:**
**Air Force Research Laboratory -**
**Information Directorate (AFRL/IF)**
**525 Brooks Road**
**Rome, NY 13441-4505**

**Prepared by:**
**Manoel Mendonca**
**University of Maryland**
**Department of Computer Science**
**A.V. Williams Building #3225**
**College Park, MD 20742**

**and**

**Nancy L. Sunderhaft**
**DoD Data & Analysis Center for Software (DACS)**
**ITT Industries**
**Griffiss Business & Technology Park**
**775 Daedalian Drive**
**Rome, NY 13441-4909**

**DoD Data & Analysis Center for Software (DACS)**
**P.O. Box 1400**
**Rome, NY 13442-1400**
**(315) 334-4905, (315) 334-4964 - Fax**
**cust-laisn@dacs.dtic.mil**
**http://www.dacs.dtic.mil**

The Data & Analysis Center for Software (DACS) is a Department of Defense (DoD) Information Analysis Center (IAC), administratively managed by the Defense Technical Information Center (DTIC) under the DoD IAC Program. The DACS is technically managed by Air Force Research Laboratory Information Directorate (AFRL/IF) Rome Research Site. ITT Industries - Systems Division manages and operates the DACS, serving as a source for current, readily available data and information concerning software engineering and software technology.

# REPORT DOCUMENTATION PAGE

*Form Approved*
*OMB No. 0704-0188*

Public reporting burden for this collection is estimated to average 1 hour per response including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Washington Headquarters Services, Directorate for Information Operations and Reports. 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project, (0704-0188). Washington, DC 20503.

1. AGENCY USE ONLY *(Leave Blank)*

2. REPORT DATE
14 November 1999

3. REPORT TYPE AND DATES COVERED
N/A

4. TITLE AND SUBTITLE
A State-of-the-Art-Report
Mining Software Engineering Data: A Survey

5. FUNDING NUMBERS
SPO700-98-D-4000

6. AUTHORS
Manoel Mendonca - University of Maryland
Nancy L. Sunderhaft- DACS

7. PERFORMING ORGANIZATIONS NAME(S) AND ADDRESS(ES)
University of Maryland, Department of Computer Science
A.V. Williams Building #3225, **College Park, MD 20742**
ITT Industries, Systems Division, 775 Daedalian Drive
Rome, NY 13441-4909, (301) 405-1226,

8. PERFORMING ORGANIZATION REPORT NUMBER
DACS-SOAR-99-3

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Defense Technical Information Center (DTIC)/ AI
8725 John J. Kingman Rd., STE 0944, Ft. Belvoir, VA 22060
and Air Force Research Lab/IFTD
525 Brooks Rd., Rome, NY 13440

10. SPONSORING/MONITORING AGENCY REPORT NUMBER
N/A

11. SUPPLEMENTARY NOTES
Available from: DoD Data & Analysis Center for Software (DACS)
775 Daedalian Drive, Rome, NY 13441-4909

12a. DISTRIBUTION/ AVAILABILITY STATEMENT
Approved for public release, distribution unlimited

12b. DISTRIBUTION CODE
UL

13. ABSTRACT *(Maximum 200 words)*

This report discusses the state-of-the-art, as well as recent advances in the use of data mining techniques as applied to software process and product information. This report includes:

- A discussion on data mining techniques and on how they can be used to analyze software engineering data.
- A bibliography on data mining with special emphasis on data mining of software engineering information.
- A survey of the data mining tools that are available to software engineering practitioners.
- A listing of web resources for data mining information.

14. SUBJECT TERMS
Datamining, Datasets, Database Management, Data Warehouse, DACS, State-of-the-Art-Report, Datamining Tools, Bibliography

15. NUMBER OF PAGES
193

16. PRICE CODE
N/A

17. SECURITY CLASSIFICATION OF REPORT
Unclassified

18. SECURITY CLASSIFICATION OF THIS PAGE
Unclassified

19. SECURITY CLASSIFICATION OF ABSTRACT
Unclassified

20. LIMITATION OF ABSTRACT
UL

# Abstract

This report discusses the state-of-the-art, as well as recent advances in the use of data mining techniques as applied to software process and product information. This report includes:

- A discussion on data mining techniques and on how they can be used to analyze software engineering data.
- A bibliography on data mining with special emphasis on data mining of software engineering information.
- A survey of the data mining tools that are available to software engineering practitioners.
- A listing of web resources for data mining information.

# Table of Contents


**1. Introduction ..... 6**

**2. Data Mining and Software Engineering Measurement ..... 7**
- 2.1 Software Engineering Measurement and Knowledge Discovery ..... 7
- 2.2 Data Mining ..... 8

**3. Data Mining Tasks and Levels of the Data Mining Process ..... 10**
- 3.1 Data Mining Tasks ..... 10
- 3.2 Levels of the Data Mining Process ..... 11

**4. Data Selection and Pre-processing ..... 12**
- 4.1 Data Selection and Extraction ..... 12
- 4.2 Data Pre-Processing ..... 13
  - 4.2.1 Data Transformation ..... 13

**5. Mining Techniques ..... 15**
- 5.1 Classification Trees ..... 15
  - 5.1.1 Numeric Fields ..... 17
  - 5.1.2 The Order of the Node Splitting and Fan-Out Effect ..... 17
- 5.2 Association Discovery Techniques ..... 17
  - 5.2.1 Market Basket Analysis ..... 17
  - 5.2.2 Deriving an Interestingness Function ..... 19
  - 5.2.3 Attribute Focusing ..... 21
- 5.3 Clustering Techniques ..... 22
  - 5.3.1 Calculating Distances ..... 23
  - 5.3.2 Clustering Algorithms ..... 25
  - 5.3.3 Using Clustering for Classification and Prediction ..... 28
- 5.4 Artificial Neural Networks ..... 29
  - 5.4.1 Backpropagation ..... 30
  - 5.4.2 Local Maximums and Overfitting ..... 30
- 5.5 Optimized Set Reduction ..... 31
- 5.6 Bayesian Belief Networks ..... 32
- 5.7 Visualization and Visual Data Mining ..... 33
  - 5.7.1 Visualization of Multivariate Data ..... 33
  - 5.7.2 Visual Data Mining ..... 34

**6. Interpretation and Assimilation (Knowledge Extraction) ..... 37**
- 6.1 Patterns, Models, and Knowledge ..... 37
- 6.2 Interpreting Patterns through Visualization ..... 37
- 6.3 Evaluating Models ..... 39
- 6.4 Interpretable Models ..... 40

**7. Bibliography on Mining Software Engineering Data ..... 41**
- 7.1 Classification Trees ..... 42
- 7.2 Artificial Neural Networks ..... 43
- 7.3 Association Discovery ..... 43
- 7.4 Clustering ..... 44
- 7.5 Optimized Set Reduction ..... 44
- 7.6 Bayesian Belief Networks ..... 45
- 7.7 Visualization ..... 45
- 7.8 Others ..... 46

**8. Concluding Remarks ..... 47**

**9. References ..... 48**

# Appendixes

A. Bibliography ..... 52
B. Tool Descriptions ..... 60
C. Data Mining Resources ..... 168

# List of Figures

Figure 1. Machine Learning Process ..... 8
Figure 2. Data Mining Process ..... 8
Figure 3. Levels of Data Mining ..... 11
Figure 4. A Classification Tree for “Error Likelihood” of Software Modules ..... 16
Figure 5. AF Diagram: Adjusted Error Type Distribution for SS Modules ..... 21
Figure 6. Single Link Clustering of Software Modules Based on # of Modifications ..... 23
Figure 7. Euclidean and Manhattan Distances in Two Dimensions ..... 24
Figure 8. K-means Method ..... 26
Figure 9. Distances Measures Used by the Agglomerative Methods ..... 27
Figure 10. A Neuron and a Sigmoid Function ..... 28
Figure 11. A Neural Network for Software Development Effort Estimation ..... 29
Figure 12. An OSR Hierarchy ..... 31
Figure 13. A BNN for Software Reliability Prediction ..... 32
Figure 14. A Multivariate Display Built Using DataMiner ..... 34
Figure 15. An Interactive Visual Data Mining Display Built with Spotfire ..... 35
Figure 16. An Animated Visual Data Mining Display Built using SGI’s MineSet ..... 36
Figure 17. A Data Pattern Presented in Graphical Format ..... 38
Figure 18. An Interpretable Classification Tree Built Using ALICE d’ISoft ..... 41

# List of Tables

Table 1. Cross-correlation Matrices Focusing on Interface Errors ..... 18
Table 2. Number of Modifications and Cyclomatic Number Metrics ..... 23
Table 3. Data Records Describing Software Modules on Several Attributes ..... 33
Table 4. A Data Pattern Presented in a Tabular Format ..... 38
Table B–1 Level of Data Mining Process ..... 61
Table B–2 Mining Techniques ..... 63
Table B–3 Algorithms ..... 65
Table B–4 Supplemental Tool Information ..... 69

# Overview

Software organizations have often collected volumes of data in hope of better understanding their processes and products. Useful information has been extracted from those large volumes of data, but it is commonly believed that large amounts of useful information remains hidden in software engineering databases.

Data mining has appeared as one of the tools of choice to better explore software engineering data. Data mining can be defined as the process of extracting new, non-trivial, and useful information from databases. This broad definition covers a wide spectrum of methods, techniques, and tools. This State of the Art Report (SOAR) discusses how data mining can be, and how it has been, used to analyze software engineering data.

# Introduction

The goal of this report is to survey the use of data mining in software engineering. Data mining represents a shift from verification-driven data analysis approaches to discovery-driven data analysis approaches. In the former approach, a decision maker must hypothesize the existence of information of interest, collect this information, and test the posed hypothesis against the information collected. Due to the size and complexity of data repositories nowadays, this approach is not sufficient to efficiently explore the data available in an organization. Discovery-driven approaches sift through large amounts of data and automatically (or semi-automatically) discover important information hidden in the data.

Discovery-driven approaches are not new to data analysts. Traditional query and decision support systems can be considered data mining in its infant age. The main reasons for the recent boom of data mining are:

1. software and hardware infrastructures have matured to handle the intensive computation required by discovery-driven data analysis;
2. advances in and ability to use techniques such as pattern recognition, neural networks, association measures, and decision trees have advanced dramatically in recent years;
3. cheap data storage and repositories integration have made large amounts of data readily available in modern organizations.

These three issues have not remained unnoticed by software development organizations. Software engineering research has been dealing with discovery-driven data analysis for some time now. Although, only recently have the buzzwords “data mining” been mentioned in software engineering publications.

This report discusses data mining and its applications to software engineering data analysis. Section 2 introduces the definition of data mining and its relation to software engineering measurement. Section 3 discusses the levels associated with data mining and the steps necessary to plan and prepare for data mining. Section 4 discusses data selection and pre-processing. Section 5 discusses types of data mining techniques with examples on software engineering data analysis. Section Error! Reference source not found. discusses the interpretation of mined information and knowledge assimilation. Section Error! Reference source not found. surveys the available literature on the use of data mining in software engineering. Appendix A contains a bibliography of data mining materials, sorted by publication date. Appendix B contains descriptions for a collection of data mining and knowledge discovery tools. Appendix C contains descriptions of data mining resources available on the World Wide Web (WWW).

# 2. Data Mining and Software Engineering Measurement

This section presents the basic concepts and terms that will be used throughout this report. Section 2.1 presents a basic terminology that combines concepts from data mining and software engineering measurement. This terminology is an adaptation of the data mining terminology proposed by Klösgen and Zytkow Error! Reference source not found. and the software engineering measurement terminology proposed by Fenton Error! Reference source not found.Error! Reference source not found.. We feel that this combination will help readers with expertise in software engineering, but with little knowledge of data mining techniques. Section 2.2 formally introduces data mining and the whole process behind it.

## 2.1 Software Engineering Measurement and Knowledge Discovery

We define **application domain** as the real or abstract system a software organization wants to analyze using software engineering data. An **entity** (object, event, or unit) is a distinct member of an application domain. Similar entities can be grouped into classes such as persons, transactions, locations, events, products, and processes. Entities are characterized by *attributes* and *relations* to other *entities*. An **attribute** (field, variable, feature, property, and magnitude) is a single characteristic of all entities in a particular entity class, for instance "usability" of software products or "size" of source code. In the case of a measurement framework, an attribute defines "what" one wants to measure. A **relation** is a set of entity tuples which has a specific meaning, for instance "a is married to b" (for person entities "a" and "b"). We measure entity attributes to empirically define relations between entities, for instance we can determine the relation "software module a is more complex than software module b" by measuring the complexity of entities "software module a" and "software module b."

**Measurement** is the process of assigning a value to an attribute. A **metric** is the mapping model used to assign values to a specific *attribute* of an entity class. A metric states "how" we measure something. It usually includes a measurement instrument, a value domain, and a scale. **Data** is a set of measured (collected, polled, surveyed, sensed, observed) attribute values produced by specific *metrics*.

**Domain knowledge** is non-trivial and useful empirical information specific to the *application domain* believed to be true by the data users. **Background knowledge** is the domain knowledge that data users had before analyzing the data. And, **new** or **discovered knowledge** is the new domain knowledge that data users gain by analyzing the data. Domain experts are data users that have a sizeable amount of expertise in a specific *application domain*.

Knowledge can be gained by *induction* or *deduction*. **Deduction** infers information that is a logical consequence of the information in a data set. This information is always true provided that the data set content is true. **Induction** infers information by generalization of the information contained in a data set. This information is believed to be true and is supported by data patterns in the data set. Consider a data set describing ten similar software systems developed and maintained by the same group of people. Suppose that they have measures of the development and maintenance costs of these systems, and each project used one of the following two languages: FORTRAN or Ada. In this case, one can deduce the average cost of the projects done in FORTRAN and in Ada. However, one can only induce information on which language is more costly to use in general. The process of discovering *new domain knowledge* through *induction* is referred to as **inductive learning** [28]. The automation of *inductive learning* processes has been originally researched in an artificial intelligence area called **machine learning** [50 ].

## 2.2 Data Mining

A typical machine learning system does not interact directly with its environment. It uses "coded observations" of this environment to learn about it. Figure 1 depicts the machine learning process. It samples facts from the environment that we want to model. It codes these facts as "coded observations" of the environment. These coded observations are fed into a machine learning mechanism to produce a model of the environment. The model can then be used to derive unknown and interesting information (i.e. *new knowledge*) about the environment under study [1].

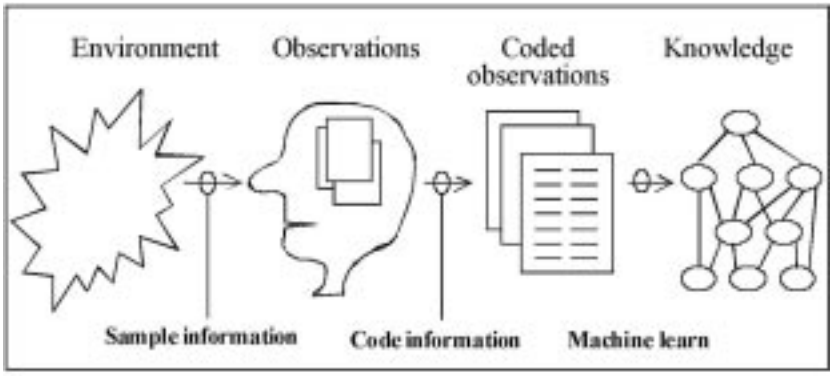



**Figure 1. Machine Learning Process**

The study of automated and semi-automated learning systems that draw coded observations directly from a database is called data mining. Formally, data mining has been defined as the process of inducing previously unknown, and potentially useful, information from databases [21][25]. In this report, we will modify the scope of this definition. For us, **data mining** is the process of inducing previously unknown, and potentially useful, information from software engineering data sets. By that we mean that the software engineering data sets used to induce knowledge are not necessarily located in databases.

Figure 2 shows a data mining process (adapted from [20]). Although the framework for data mining and machine learning — as shown in Figures 1 and 2 may seem similar, there is a key distinction between them. The raw data in Figure 1 was derived for purposes other than data mining. The entities and attributes represented in the data have been collected to meet the needs of the applications that use it rather than the needs of the data mining process [29]. In our case, the data was collected to meet the needs of a software organization having their own product and process improvement goals and possibly using diverse software metrics to collect this data. This means that the data is not organized in a way that will facilitate automated or semi-automated knowledge induction. Also, there may be irrelevant, missing, noisy, and uncertain data in this raw data set.

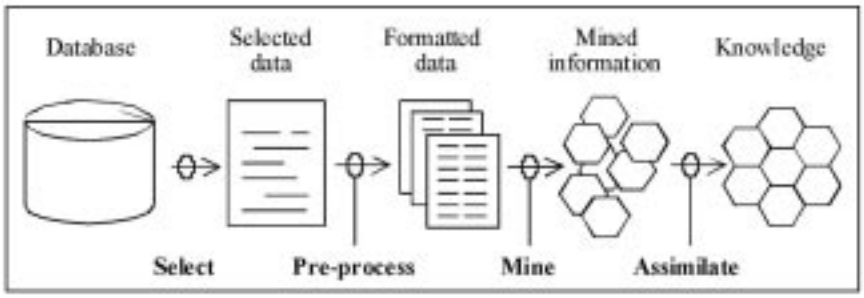



**Figure 2. Data Mining Process**

As shown in Figure 2, a typical data mining process has four steps. The first step is to select the types of data that will be used by the mining algorithm. Raw data sets usually contain a variety of diverse data, not all of which is necessary to achieve a data mining goal. In this step, the data analyst will have to identify where the desirable data is, gain access to it, and transport the data from its original location to the data repository with which he/she will work. These operations are usually simple in concept but complex in execution. Frequently, it is not easy to

gain access to data sources that are not your own. Also, frequently software engineering data is not organized in well structured databases or centralized in one location. A fair amount of effort is usually necessary to collect sizeable amounts of data and transport it to one centralized location where it can be mined. Section 4.1 of this report will discuss data selection and extraction in more detail.

The second step is to pre-process the data for analysis. Usually the data has to be formatted, sampled, adapted, and sometimes transformed for the data mining algorithm. Formatting usually involves treating noisy or missing data. Adaptation is frequently necessary to make the data work with the data mining technique that will be used. For example, some algorithms only work with nominal data. In those situations, numeric data has to be mapped into categories based on chosen value ranges. More sophisticated data pre-processing may involve data transformation. This operation is very common in software engineering data and happens when one uses the data set to derive new attributes for data mining. For example, suppose a software organization has an error reporting database. Among this data are the dates of each fault introduction, report, and removal in a set of software projects. A typical data transformation would be to use those dates to calculate the attribute "seriousness" of the fault as a function of the fault introduction, time to discover, and time to fix. Section 4.2 will discuss data pre-processing in more detail.

After pre-processing, the data is finally ready to be mined by a data mining algorithm. The data mining step aims to extract interesting – potentially useful, unknown, non-trivial – information from this data. This step may involve very different data mining techniques. For example, one may use the pre-processed data to:

1. develop an accurate classification model for fault prediction on software modules
2. automatically produce charts containing interesting associations between software project characteristics and software faults profiles;
3. or, gain a high level view of the projective characteristics and fault profiles by using interactive data visualization tools.

In the first case, the mined information is contained in the derived model itself. In the second case, the mined information will be extracted when a domain expert examines the patterns contained in interesting charts. In the third case, the mined information will be extracted when a domain expert interactively explores visual displays of the data. These examples represent the major areas of data mining: model building, automatic pattern extraction, and visual data exploration. Section 5 will discuss the main techniques used in these data mining areas.

The last step of the data mining process is to assimilate the mined information. This is done by interpreting and assimilating the information the data mining technique considered "interesting." In the case of model building, this step consists of evaluating the robustness and effectiveness of the produced models by using techniques such as resubstitution and cross-validation [17 ]. If approved, the model can then be incorporated into the organization's business processes.

In the case of pattern extraction and visual data exploration, this step consists of actually trying to make sense of the information extracted by the data mining algorithm. This is usually done by a domain expert who will look at the mined information and use his *background knowledge* to check if this information is indeed something new, useful, and non-trivial (i.e., *new knowledge*). Section 6 will discuss interpretation and assimilation of mined information in more detail.

The data mining process usually is highly interactive. The data analyst has to re-think data selection whenever the mined information is not interesting. He has to further refine data pre-processing if the formatted data is not completely adequate to the mining algorithm. The algorithm frequently needs to be re-tuned when too few interesting facts are identified in the data assimilation step. By its nature, the data mining process has feedback loops. The quality of the information it discovers is dependent on the quality of the raw data, the mining algorithm, and other techniques it uses throughout the data mining process. But, it is also very much dependent on the way the data mining process is put together as a whole, and how well the data analysts interact between its steps. The next three sections discuss the three major phases of data mining: planning and preparing for data mining; mining; and, knowledge assimilation.

# 3. Data Mining Tasks and Levels of the Data Mining Process

The choice of which data mining technique one should use at a given point in time depends on the domain expert's goals and the tasks one wants to perform to achieve those goals. Section 3.1 discusses the basic data mining tasks one can execute on software engineering data. Section 3.2 discusses how these tasks can be classified according to the level of information they produce.

## 3.1 Data Mining Tasks

The first step in any data mining endeavor is to map the domain expert's business - in our case, software management and engineering - goals into one or more data mining tasks. Data mining tasks can be roughly classified into six categories:

1. Estimation and Prediction. Estimation consists of examining attributes of a set of entities (products, processes, and resources) and, based on these attribute values, assigning values to an unknown attribute that one wants to quantify. The term prediction is sometimes used when an estimation is done to predict the future outcome of an attribute value. A typical example of an estimation task is to use attributes that characterize a project to estimate (predict) its costs. Figure 11 in Section 5.4 shows an artificial neural network model being used to predict software development effort.
2. Classification. Classification consists of examining the attributes of a given entity and, based on these attribute values, assigning it to a predefined category or class. Figure 4 in Section 5.1 shows a classification tree used to assign software modules into two categories "likely" and "unlikely to have errors associated with them."
3. Association Discovery. Association discovery consists of identifying which attributes are associated with each other in a given environment. A typical example would be to try to find out which attribute of the software development team (e.g., experience attributes, training attributes, domain knowledge attributes, etc.) are associated with the final software product attributes (e.g., usability attributes, maintainability attributes, reliability attributes, etc.). Section 5.2 discusses association discovery techniques.
4. Clustering. Clustering is the task of segmenting a heterogeneous population into a set of more homogeneous subgroups. Clustering differs from classification, as it does not rely on predefined classes. Clustering splits a population into classes on the basis of self similarity between the class members. It produces a high level description of the population based only on distance measures between its elements, no classification model is explicitly built. Section 5.3 discusses clustering techniques.

5. Visualization of Data. Data visualization is the task of describing complex information through visual data displays. Visualization is based on the premise that a good description of an entity (be it a software resource, product, or process) will often improve a domain expert's understanding of this entity and its behavior. Section 5.7 discusses the use of visualization to describe and interpret software engineering data.
6. Visual Data Exploration. An extension to the visual data description task, visual data exploration is the task of inspecting large volumes of data through interactive control of visual displays. The goal of this task is to allow domain experts to quickly examine "what if" scenarios in multivariate visual displays. This task employs "visual data mining" tools with advanced user interfaces and interactive data query resources to fly through the data being explored. Section [ ] discusses visual data exploration.

## 3.2 Levels of the Data Mining Process

The six data mining tasks listed in the previous section work at very different levels with regard to knowledge discovery. The first two tasks, classification and prediction, aim to build explicit models that are ready to be employed in the software organization. The next two tasks, association discovery and clustering, aim to automatically identify useful patterns in data. Those patterns have to be interpreted by a domain expert in order to produce business insights for an organization. The last two techniques, visualization and visual data exploration, aim to help domain experts to easily find, by themselves, useful patterns in the explored data. These patterns can be used to produce business insights or to simply help the experts to better understand what is going on in the data. Those tasks define three levels of data exploration and knowledge discovery:

1. Model Building
2. Automatic Pattern Extraction
3. Interactive Visual Data Exploration

As shown in Figure 3, the three levels of data exploration are quite related to each other. They can be seen as a pyramid where basic and visual data exploration at the bottom creates the basis for good pattern extraction and model building at the top. It should be noted, however, that data exploration and pattern extraction do more than just support model building. Application domains, technologies, and organizations are continually changing over time. Basic data exploration and pattern extraction are tools for domain experts to keep up to date on what is happening (and changing) in the data monitored by his organization.

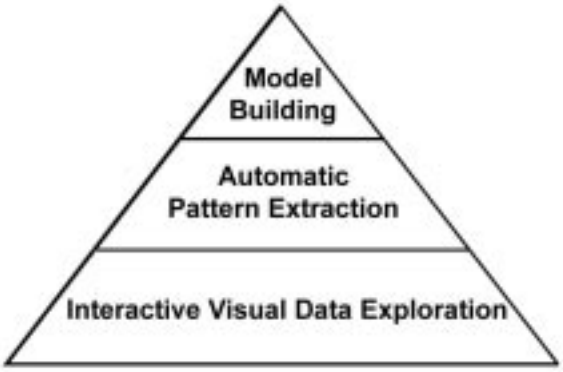



**Figure3. Levels of Data Mining**

It is our belief that all three levels of data exploration are very important for an organization. *Interactive visual data exploration* is needed to:

1. characterize entities (resources, products, and processes),
2. gain a high level view of the data being explored,
3. better understand how entities behave in a particular application domain,
4. identify entity classes, and
5. gain a basic understanding of how those entities relate to each other.

*Automatic pattern extraction* is needed to:

1. identify complex entity classes, and
2. better characterize and understand complex relationships between entities.

*Model building* is needed to code those relationships in explicit models ready to be used for classification, estimation, or prediction in the organization.

# 4. Data Selection and Pre-processing

This section discusses the first two steps of the data mining process shown in Figure 2: data selection and pre-processing. Fundamental for successful data mining endeavors, these two steps are sometimes regarded as minor tasks in the data mining process. There can be no reliable information extracted from bad data. Data has to be understood and cleaned before it can be successfully mined.

## 4.1 Data Selection and Extraction

Data mining is based on the premise that non-trivial, unknown, and valuable information lies in an existing data repository. The goal of data mining is to sift through the repository data and extract interesting information from it. The first requirement for successful data mining is to identify and gain access to an information abundant data repository.

Frequently, software engineering data is not stored in a unique data repository or an integrated set of data repositories. In this case, the data analyst must seek out possible data sources in various groups or departments of the software organization, gain access to this data, and use it to assemble their own data repository.

These tasks are not simple and many times will consume a fair amount of the data mining effort. Frequently, the data analyst has to spend a significant amount of effort to gain access to a departmental data repository. Once this access is granted, the data analyst usually finds a repository that is designed for the department's own particular purpose. Extracting data suitable to the data analyst's knowledge discovery goals from this repository may require a fair amount of effort.

First, the data analyst must understand exactly what type of data is stored in the repository. Effort data, for example, may be logged in very diverse forms. A department may log effort by hours by task by person, another may log it by team by day by project. The data analyst must understand the semantics of the data before he can start any effort to extract it from the data repository. The next step is to understand how the data is organized in the repository. For example, if the repository is in a database system, the data analyst may have to study its schema, nomenclature, metadata, and query mechanisms. If the repository is in a file system, the data analyst must understand exactly where the files are stored and how the data is stored in each of those files.

The last step of this process is the data extraction enactment. In this step, the data is transported from the departmental data repository to the data analyst's own repository. Depending on the complexity of the data, the data source, the data destination, and the data semantics, this procedure may range from a simple query to a data base management system to a complex data migration procedure involving intricate data type conversions.

## 4.2 Data Pre-Processing

The data extracted from departmental data sources is usually not ready for data mining. During and after the data extraction process, the data has to be formatted to a data mining ready format. Some of the operations that may be needed during this phase are:

*Capitalization*. It is common to find data mining algorithms that are case sensitive when working with nominal or ordinal data. In such cases, the data analyst may transform characters of a data stream to either uppercase or lowercase.

*Concatenation*. It is common for multiple data fields in a data source to be transformed into just one data field in the data mining repository. Suppose, for example, that the source repository stores the following fields `<system_version=2; system_release=15>`. The data analyst may want to concatenate those two fields in a new field `<version_release=2.15>` in the data mining repository.

*Representation Format*. It is common for certain types of information to be represented in different formats by different data sources. Date is probably the most common case. The data analyst should make sure that this type of information is represented in a consistent and known format in the data mining repository. Consider, for example, a data field extracted from a European data source in the format DD-MM-YYYY, the data analyst may want to transform it to a data field in the format MM-DD-YYYY to facilitate assimilation of the data mining results.

*Character Clean-up*. Sometimes extraneous characters must be removed from the data so it can be used by certain data mining algorithms. A common case is the dollar ($) character. Data fields representing cost frequently have to be treated as numeric fields. The dollar sign has to be removed from the data fields representing cost in order to make it usable as a numeric field.

*Data Clean-up*. Frequently the data source has fields with missing values, extraneous, or plainly wrong values. These fields have to be fixed. The data analyst can do that by interpolating values, by entering special codes in those fields (e.g., N/A), or by eliminating the records that contains these fields.

*Data Set Reduction*. Although this is not common in software engineering, some data sets may be too large for certain data mining algorithms. In this case, the data analyst has two options. The data set can be split into smaller, more specific, data sets or the data set can be sampled before it is put into the data mining repository.

### 4.2.1 Data Transformation

A more complex type of data pre-processing is data transformation. This type of operation is frequently needed to adapt data to particular data mining techniques. Some techniques, such as market basket analysis, work essentially with categorical data. In this case, attributes measured in numerical (interval, ratio, or absolute) scales must be mapped to categories defined by numeric value ranges. Other techniques, like neural networks, work essentially with numeric scales. In this case, attributes measured in categorical data must be mapped to numbers.

*Scale Reduction*. Some data mining algorithms deal only with nominal or ordinal data. In such cases, the data analyst must transform numeric fields into character fields. Consider for example that the "number of terminal installations" is used to measure the attribute "size of the installation" of Internet products. Suppose that a software engineer wants to take this attribute into account while analyzing a database on software fault reports; however, he wants to map the numeric value "number of terminal installations" to a categorical scale <*small, medium, large*>. The mapping between the two scales can be done in several different ways. The most common is to consider expert opinion or the distribution of the data values in the repository. In the expert opinion method, an expert would be asked to subjectively determine what range of values should be considered small, medium, or large. This approach is very useful when an organization already uses a similar mapping in their day to day operation. In our example, the expert may know that senior management refers to an installation as small if it has less than 100 terminals, as medium if it has more than 100 but less than 300 terminals, and as large if it has 300 or more terminals.

Scale augmentation. Other data mining algorithms deal only with numerical (rational, interval, or absolute) scales. In such cases, the data analyst must transform character fields into numeric fields. A common approach for this is to transform an attribute into a set of different attributes. Each new attribute will correspond to a different value of the original attribute. The new attributes are then assigned values 0 or 1 depending on the original attribute value. Consider for example the attribute "system type" with the following possible values: <*CPU intensive*>, <*real time*>, and <*transaction intensive*>. This attribute could be substituted by the following three attributes, "CPU intensive," "real time," and "transaction intensive." The attribute "real time" is assigned value 1 if the original record has value <*real time*> for "system type," and 0 otherwise.

*Unit Conversion*. It is common for different data sources to represent the same attribute in different scales. Consider for example the attribute <effort>, which is frequently measured in people-years, people-months, or people-hours. The data analyst must ensure that attributes such as these are consistently recorded in the data mining repository. Heterogeneous scales must be transformed to a common unit of measurement, usually the unit employed locally. This apparently simple task can be quite complex when identical attributes are being measured by different metrics. Consider for example that two data sets have different measures for source code size. The first data set count source lines of code including comments to measure code size. The second counts source lines of code without including comments to measure code size. In order to use those two data sets together, one has to convert the units of the first data set into the one used in the second or vice versa. This conversion may require a lot of work if the data sets are large. In the example, one might have to empirically determine the average percentage of comment lines per line of code based on some sampling of the measured code, in order to do the unit conversion.

*Value Normalization*. Some mining techniques require attribute values to be normalized on a certain range, usually from 0 to 1. The minimum and maximum values of an attribute are mapped to 0 and 1 respectively. All other values are then mapped to a value in this normalized range. This is done mostly with numeric attributes, but in some cases, categorical attributes are also normalized. This is usually done through a matching criterion such as "if an attribute is equal to value X map it to 1, otherwise map it to 0." Some value normalization operations may also be quite complex as sometimes one wants to smooth down outliers and large value variations through non-linear transformations.

*Data Set Adaptation*. Some data outliers or unbalanced data sets can be modified to improve the effectiveness of certain data mining techniques. Those operations can be quite complex and are usually dependent on the mining technique one wants to use. Common operations include:

1. outliers elimination and
2. merging sets of records on "average equivalent records" that represents several data records at once.

# 5. Mining Techniques

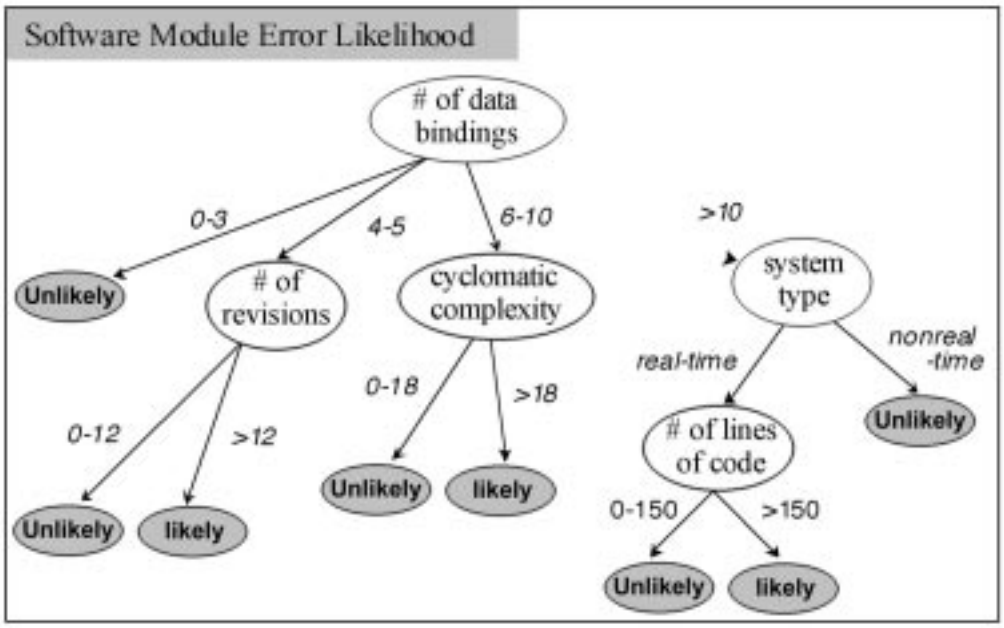
This section presents the data mining algorithms and techniques most commonly used to produce patterns and extract interesting information from software engineering data. The techniques are organized in seven sections: classification trees, association discovery, clustering, artificial neural networks, optimized set reduction, bayesian belief networks, and visual data mining.

## 5.1 Classification Trees

Classification or decision trees are induction techniques used to discover classification rules for a chosen attribute of a data set by systematically subdividing the information contained in this data set. They have been one of the tools of choice for building classification models in the software engineering field [31] [46] [47][51][52][54][55]. Figure 4 shows an example of a classification tree extracted from [38]. In this fictitious example, the goal is to identify risky software modules based on attributes of the module and its system. Consider as an example the right most path from root to leaf in Figure 4's tree, this path is saying that: IF a module has more than 10 data bindings AND it is part of a non real-time system THEN this module is unlikely to have errors.

The algorithms used to build classification trees seek to find those attributes and values that provide maximum segregation of data records in the data set at each level of the tree. In Figure 4, "# of data bindings" was selected first because this is the attribute that most equally divides records for "error likelihood" in the data set. In terms of information theory, this is the attribute that provides most information by reducing the most uncertainty about the "error likelihood" value. The reasoning is that the more information a tree has at each node the smaller this tree will be. Below, we have ID3, a classification tree induction algorithm proposed by Quinlan in the eighties [49 ].

1. Select an attribute as the root of the tree, make branches for all values this attribute can have;
2. Use generated tree to classify the training set. If all examples at a particular leaf node have the same value for the attribute being classified (e.g., error likely module); this leaf node is labeled with this value. If all leaves are labeled with a value, the algorithm terminates.
3. Otherwise, label the node with an attribute that does not occur on the path to the root, branch for all possible values, and return to step 2.




**A Classification Tree for "Error Likelihood" of Software Modules**

The key to making the above algorithm work is to select a suitable attribute at each node of the tree. ID3 uses an information-based heuristic to execute this selection. As hinted before, it tries to select an attribute that minimizes the information needed in the remaining subtrees to classify the data.

The attribute selection works as follows. The data set is divided into two subsets *P* and *N*. Subset *P* contains all positive elements (e.g., all modules that are likely to have faults). Subset *N* contains all negative elements (e.g., all modules that were unlikely to have faults). For each attribute *A* (e.g., # data bindings) and value set $V_i$ (e.g., #data binding >10), determine the subset $S_i$ of elements for which A is in $V_i$ (e.g., all modules for which #data binding >10). Count the number $n_i$ of elements that $S_i$ has in *N* and the number pi of elements that $S_i$ has in P. Calculate the amount of information needed to decide if an arbitrary element in $S_i$ belongs to P or N using the following formula derived from information theory:

$$I(p_i, n_i) = -\frac{p_i}{p_i + n_i} \log_2\left(\frac{p_i}{p_i + n_i}\right) - \frac{n_i}{p_i + n_i} \log_2\left(\frac{n_i}{p_i + n_i}\right)$$

Assuming that attribute A parts a data set *S* in the subsets $\{S_1, S_2, ..., S_v\}$ for its value sets $\{V_1, V_2, ..., V_v\}$. And, knowing that the amount of information needed to decide if an element of $S_i$ belongs to *P* or to *N* is $I(p_i, n_i)$. The information needed to classify an element of data set S using attribute A is:

$$E(A) = \sum_{i=1}^{v} \frac{p_i + n_i}{p + n} I(p_i, n_i)$$

That is the weighted average of the information needed to classify elements in the subtrees $S_i$ defined by A values $\{V_1, V_2, ..., V_v\}$. Attribute A is then selected for a node of the classification tree if it has the minimal information needed, *E(A)*, among all attributes that can be selected at this node.

### 5.1.1 Numeric Fields

A condition not addressed by ID3 was how to determine $\{V_1, V_2, ..., V_v\}$ for numerical attributes. An extension to ID3, called C4.5 [ ], was later proposed to allow tests of inequality of numerical attributes, such as A<N and A_N. The information gain in this case is computed as follows. The elements are sorted on the values of the attribute being considered. This number is finite. Let us say that there are t different values for the attribute A in the data set $\{V_1, V_2, ..., V_t\}$. In this case, there are t-1 possible binary splits on this attribute, all of which are examined for minimal information needed. The best split is then selected to calculate the attribute E(A)

### 5.1.2 The Order of the Node Splitting and Fan-Out Effect

In 0, one may notice that the "# of data bindings" split is not binary but quaternary (a 4-way split). Higher order splits can be done up to the number of different values an attribute has in the data set. On the previous example, we can have up to a t-way split on attribute A. The number of possible splits of order *n* (*n-way split*) on an attribute with *t* different values in the data set is:

$$Splits\ (n,t) = \frac{(t\ \text{-}1)!}{(n\ \text{-}1)!\ (t\ \text{-n})!}$$

The larger the order of a split the smaller is E(A), the information needed in the node's sub-trees. Many times, however, it is not a good policy to arbitrarily raise the order of splits of a classification tree. If an attribute has many different values represented in the data set, the order of splits can be very high. This creates a huge fan-out effect at that node, generating trees with very broad horizontal levels. These trees are usually difficult to interpret. Even worse, these trees have many leaf nodes, possibly with very few data elements represented in them. In these cases, the rules produced by the path that goes from the tree's root to its leaves would have little significance because they are supported by very small subsets of data. Good decision tree algorithms must create a balance between the uncertainty reduction at each of its nodes and the final confidence level at each of its leaves.

## 5.2 Association Discovery Techniques

Association discovery extracts information from coincidences in the data set. Knowledge discovery takes place when these coincidences are previously unknown, non-trivial, and interpretable by a domain expert.

### 5.2.1 Market Basket Analysis

Market basket analysis techniques allow one to discover correlations or co-occurrences of transactional events. Market basket analysis uses cross-correlation matrices in which the probability of an event occurring in conjunction with every other event is computed. Consider the hypothetical example of a company that logs errors discovered at its software modules in a transaction system. Each error is classified in one or more of the following categories: (1) interface error - Ie; (2) missing or wrong functionality - Fe; (3) algorithm or data structure error - Ae; (4) initialization or assignment error - INIe; (5) cosmetic or documentation error - De. The company also logs possible "error inducing events" based on the software module features: (1) large size - LS; (2) small size - SS; (3) large complexity - LC; (4) large number of revisions - LNR; (5) small number of revisions - SNR; (6) large fan-out - LFO; and (7) large fan-in - LFI.

Table 1 shows three cross-correlation matrices focusing on interface errors (Ie). The first row of the table shows the matrix for one dimension associations. It contains the percentage of errors in which: (1) the error was classified as an interface error - Ie ; (2) the module that originated the error had large size - LS; (2) had small size - SS; (3) large complexity; (4) large number of revisions; (5) small number of revisions; (6) large fan-out; and (7) large fan-in, respectively. The second row of the table shows the matrix for two-dimensional associations related to interface errors. It shows the percentages of errors that were interface errors AND had the error inducing events - LS, SS, LNR, SNR, LFO, and LFI, respectively. The rest of the table, rows 3 through 8, shows the three dimensional associations related to interface errors.

The first thing to notice in these matrices is that some relations make sense and can be easily interpreted. For example, modules with large fan-out (LFO) and a large fan-in (LFI) should indeed be highly associated with interface errors (22.7%). Unfortunately, this fact is expected and consequently not interesting for most software engineers. Other relations point to facts that are unknown but difficult to interpret. For example, modules with large complexity (LC) and large size (LS) are associated with interface errors (13.9%). This type of fact cannot be easily interpreted and effectively transformed into knowledge by a domain expert.

Still other relations show facts that are interesting and can be transformed in real knowledge by a domain expert. For example, Table 1 shows that modules with short size (SS) and large number of revisions (LNR) are associated with interface errors (10.1%). An expert in the software domain may interpret this as a sign that modifications in small modules tend to affect its interface and introduce new errors in the software. In this case, maintainers of small modules might want to pay closer attention to its interface to other modules of the system.

**Table 1. Cross-Correlation Matrices Focusing on Interface Errors**

| | Ie | LS | SS | LC | LNR | SNR | LFO | LFI |
|---|---|---|---|---|---|---|---|---|
| | 38.0% | 53.0% | 15.0% | 55.0% | 67.0% | 18.0% | 45.0% | 48.0% |
| Ie^ | | 22.0% | 12.0% | 23.0% | 25.0% | 8.0% | 31.0% | 33.0% |
| Ie^LS^ | | | 0.0% | 13.9% | 7.3% | 1.5% | 11.5% | 13.9% |
| Ie^SS^ | | | | 0.1% | 10.1% | 0.9% | 7.6% | 7.5% |
| Ie^LC^ | | | | | 7.5% | 2.1% | 11.9% | 14.4% |
| Ie^LNR^ | | | | | | 0.0% | 10.1% | 13.4% |
| Ie^SNR^ | | | | | | | 2.6% | 2.8% |
| Ie^LFO^ | | | | | | | | 22.7% |

Table 1 only focuses on interface errors and shows up to three-dimensional matrices. Matrices of higher order can be derived just as easily. A four-dimensional matrix might show an “interesting” association between IE, LFO, LFI, and LNR. A five-dimensional matrix might show an even more “interesting” association between IE, LFO, LFI, LNR, and LC. So, sometimes it is desirable to produce higher order matrices when using association discovery techniques. There are, however, two problems with producing cross-correlation matrices for higher order associations. The first is the significance of obtained results. The number of data elements supporting associations gets smaller as the association dimension grows. This creates a phenomenon similar to the classification tree fan-out effect. The significance of the associations decreases with the growth of the cross-correlation matrix dimension.

The second issue is the difficulty to review all associations. The number of cells in a cross-correlation matrix grows very fast with the dimension of the association represented in it. An n-dimensional matrix with $t$ attributes has $\binom{t}{n}$ cells, in other words:

$$Number_of_Cells = \frac{t!}{n!\,(t\text{ -n})!}$$

In such cases, reviewing the percentages shown in the matrices is quite difficult, if not impossible. Real world association discovery analyses should have mechanisms to automatically select candidates associations to be examined by the domain experts. Good mechanisms for association selection and presentation can significantly increase the number and importance of discoveries.

### 5.2.2 Deriving an Interestingness Function

In association discovery data mining, "interestingness functions" are usually used to *quantify* how interesting an association would be to a domain expert (a software engineer expert in our case.) Just like the algorithm for choosing good splits in a classification tree, these functions are key to successful association discovery.

An association is "interesting" if it is unlikely, significant, novel, valuable, and interpretable. It is unlikely if it is neither common knowledge, nor expected by the domain expert. It is significant if it is supported by a sizeable subset of the data. It is novel if it covers regions of the association space that are not covered by other associations. It is valuable if it has real business impact. It is interpretable if it can be transformed into domain knowledge, building on the domain expert background knowledge and intuitions.

Interestingness functions usually employ cross-correlation percentages, together with other criteria defined by the data analysts as mechanisms to quantify unlikely, significant, novel, valuable, and interpretable associations. Once a good interestingness function is defined for a data set, it can be used to sort the associations extracted from the data sets. Domain experts can then look at the associations deemed the most interesting by the interestingness function and try to derive domain knowledge from them.

Let us first select a function that gives us an unlikely event. The simplest way to do this is to identify attributes that present associations that are unlikely when compared with the random chance of their association in the data. Consider two attributes A and B, a possible measure of the interestingness of their association could be:

$$Interestingness_1\,(A,B) = \left|P(A|B) - P(A)\right| = \left|\frac{P(A|B)}{P(B)} - P(A)\right| = \left|\frac{P(A|B) - P(A)P(B)}{P(B)}\right|,$$

where $P(A)$ is the percentage of occurrence of $A$ in the data set $P(A \mid B)$ is the percentage of occurrence of $A \mid B$ in the data set.

This function is close to zero when the attributes A and B are independent and larger otherwise. However, this function is only a mathematical measure of association and disassociation between attributes. A good interestingness measure should also include some of the domain expert's knowledge about possible associations. Consider for example the disassociation between small size and large complexity in Table 1. This disassociation is expected because almost all small modules have small complexity. In order to avoid mining expected results such as the disassociation between small size and large complexity, a good interestingness function should include a filter with which the data analysts can sift out expected (dis)associations. A simple filter is presented below. It uses a list of expected

associations entered by a domain expert prior to the analysis to identify associations that are expected and, consequently, have little value.

$$Interestingness_2\ (A, B) = Interestingness_1\ (A, B) * Association\ Filter\ (A, B)$$

$$Association\ Filter\ (A, B) = \begin{cases} 0, & \text{if association } (A, B) \text{ is the list of expected associations} \\ 1, & \text{otherwise} \end{cases}$$

The previous function does not consider how significant an association is. Suppose, for example, that event A and event B occur only once in the data set. If they occur together, they will have a 100% association. Unfortunately, this association has no significance as it is based on only one data point. An interestingness function must factor the significance of the association into its formula. A way to do this is by creating a function that takes into account the support the association has in the data set:

$$Interestingness_3\ (A, B) = Interestingness_2\ (A, B) * Support\ (A, B)$$

$$Support\ (A, B) = FunctionOf\ (\text{Number of occurrences of } A \mid B \text{ in the data set})$$

The novel factor should also be considered by an interestingness function. This can be done by considering the number of appearances of the association's attributes in other mined associations. Consider, for example, the top ranked associations produced by $Interestingness_3\ (A, B)$. One can interactively adjust the interestingness score produced by $Interestingness_3\ (A, B)$ based on the number of appearances of the attributes A and B on other associations ranked high by this function. For example, the interestingness rank of Table 1 will include several associations involving LFO and LFI, and just a few associations involving SS and LNR. This fact indicates novelty for associations involving SS and LNR when compared with associations involving LFO and LFI. The following recursive approach factors this criterion into our interestingness function:

$$Interestingness_4\ (A, B) = Interestingness_3\ (A, B) * Novel\ (A, B)$$

$$Novel\ (A, B) = FunctionOf\ \frac{1}{\text{appearances of A and B on } Interestingness_3\ (A, B) \text{ top rank}}$$

It is not easy to factor the association "business value" and "ease of interpretation" in an interestingness function. These factors are usually subjective and quite dependent on domain knowledge and attribute features. Domain expertise must be the main tool used to evaluate these factors. An interestingness function can, however, include some criteria to, at least partially, consider the associations' "business value" and "ease of interpretation." For quantifying an association's value, an interestingness function could use a rank of attribute importance, as defined by a domain expert. Consider that in our example the domain expert knows that interface errors are harder to fix than cosmetic or documentation errors. Thus, associations involving those types of errors would be considered more interesting by an interestingness function:

$$Interestingness_5\ (A, B) = Interestingness_4\ (A, B) * Value\ (A, B)$$

$$Value\ (A, B) = FunctionOf\ (\text{A and B in importance rank according to the domain expert})$$

For quantifying ease of interpretation, the interestingness function may consider that associations with a smaller number of attributes should be easy to interpret. This does not specifically apply to our functions as they are only dealing with two-dimensional associations. However, real world interestingness functions must deal with N-dimensional associations. In these, scenarios, simpler associations should be given priority over higher dimensional associations with similar interestingness.

### 5.2.3 Attribute Focusing

Attribute Focusing (AF) [4] is a data mining technique that has spun off from IBM's work on Orthogonal Defect Classification (ODC) [1][14]. It has been used in several different applications - including software process measurement [4][5][6].

AF is a typical association discovery technique and involves one or more experts in the knowledge discovery process. The AF tool searches an attribute-value (measurement) database for interesting facts. An interesting fact is characterized by an unexpected correlation between values of a set of attributes. The facts are presented in easily interpretable bar chart diagrams. The diagrams are sorted by interestingness level and presented to the experts. Knowledge discovery takes place when the experts address the questions raised by the diagrams.

The diagram presented in Figure 5 involves two attributes, Error Type and Module Size. The X-axis of the diagram displays the distribution of error by error type. The overall distribution of errors is shown on the lighter bars and the distribution of errors for short-sized modules is shown on the darker bars. In this case, the diagram was produced because it shows a strong association between interface errors and short-sized modules. The percentage of interface errors was 38% overall but it jumps to 80% in short-sized modules. In a typical AF analysis many diagrams of this type will be produced and ordered by the interestingness degree of the strongest association between their attribute values.

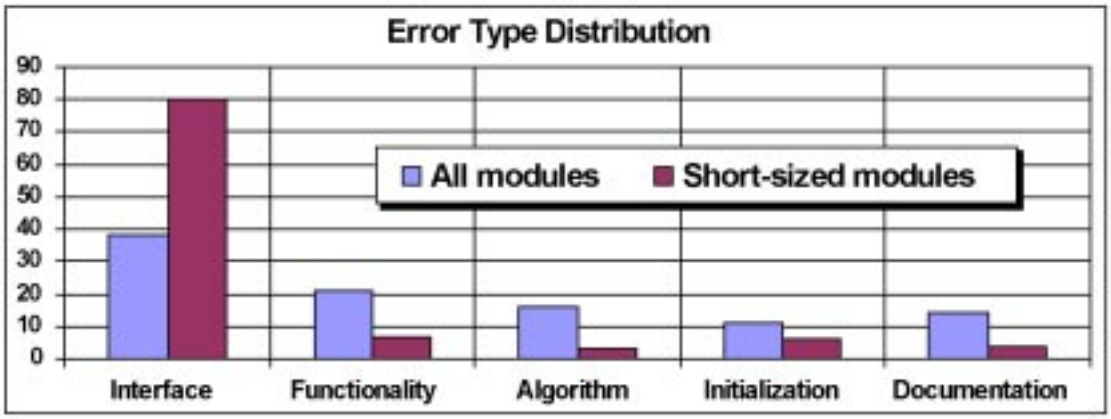



**Figure 5. AF Diagram: Adjusted Error Type Distribution for SS Modules**

Please notice that contrary to the normal market basket analysis, which focuses on events and specific attribute values, AF displays distributions of attribute values. Still, the function used to calculate the interestingness level in AF focuses on associations between discrete attribute values. Its interestingness function is quite similar to the interestingness functions discussed in Section 5.2.2. For 2-dimensional associations, AF uses the following function:

$$Interestingness(A, B) = \forall_u \forall_v (\max | P(A = u) * P(B = v) - P(A = u \mid B = v) ),$$

where $P(A = u)$ is the percentage of occurrence of value $u$ for attribute $A$ in the data set and $P(A=u \mid B\text{-}v)$ is the percentage of occurrence of values $u \mid v$, together, in the data set.

A *Support* (significance) function is hardwired into this interestingness function as the formula above doesn't divide (*P(A)P(B)-P(A_B)*) by *P(B)*. This gives more weight to associations supported by a high number of data points.

AF also filters known and/or undesired associations. For that, attributes are classified into classes. Generic relationship questions (GRQs) are then used to define which classes attributes should be compared to each other. A typical GRQ has the form: "How *attribute class A* relates *attribute class B*?" In our example, this would be expressed as: "How module *features relates* to *error types*?" The attribute classes are then used to define which types of attribute associations should be considered unexpected during the AF analyses. In our example, the associations must include one or more attributes from class A (module features) and one, and only one, attribute from class B (error type.) The attribute classes together with GRQs effectively implement association filters by avoiding the computation of associations between the attributes that are put in the same classes.

A minimum interestingness cutoff value is also set for each data analysis to establish an interestingness threshold. The data analyst also controls the maximum number of interesting diagrams that can be produced by the tool. This is very useful when the expert's data analysis time is limited or costly. In AF, the previous interestingness function for two-way associations is extended to three-way associations as follows:

$$Interestingness(A, B, C) = \forall_u \forall_v \forall_z \ \max \left( \left| \begin{array}{l} P(A = u \wedge B = v) * P(C = z) - \\ P(A = u \wedge B = v \wedge C = z) \end{array} \right| \right)$$

and,

$$Interestingness(A, B, C) > Interestingness(A, C) \wedge$$
$$Interestingness(A, B, C) > Interestingness(B, C)$$

In other words, three-way associations are considered interesting (with *C* as the focus attribute) if the absolute value of the association is greater than any of the 2-way associations established between (*A,C*) and (*B,C*). Implicit to this algorithm is the selection of the strongest associations between attributes of an existing set of attributes and the selection of the optimum length of the associations. That is, the three-way evaluation compares the results of the various two-way evaluations, if the three-way evaluation is greater than (more interesting than) the two-way evaluations, this pattern is output, otherwise not. The three-way association formula above can just as easily be extended to N-way associations.

## 5.3 Clustering Techniques

Clustering techniques are among the oldest data mining techniques. Unfortunately, we are aware of just a few works in which they are used to analyze software engineering data [36][45]. The concept of clustering is very simple; consider the following example. Suppose that one is moving and wants to pack all his belongings. One wants to group material with similar characteristics together so he knows how to handle them during transportation. Fragile objects should be packaged together because they require careful handling. Cooking utensils should be packaged together because they will go to the kitchen. In this example, objects were clustered together because they have attributes in common about the way they behave. The same is true for data or information clustering. One wants to group data records with

similar attributes together so information can be abstracted. Data clustering can be used to: (1) produce a high-level view of what is going on in the data; (2) automatically identify data outliers; or (3) classify or predict the value of new records using a technique called nearest neighbor classification.

### 5.3.1 Calculating Distances

The most important concept used in clustering algorithms is the concept of distance between data records. Suppose that a database has records on software modules of a software system and that one wants to cluster those records based on the "# of modifications" made on those modules. In order to do that, one has to calculate the distance between the modules based on the "# of modifications" metric. This distance can then be used to determine the similarity between the modules based on the "# of modifications" criterion. Table 2 shows the "# of modifications" for a set of modules labeled from A to O.

**Table 2. Number of Modifications and Cyclomatic Number Metrics**

| Modules | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Number of Modifications | 29 | 25 | 5 | 21 | 19 | 2 | 8 | 3 | 12 | 14 | 35 | 30 | 9 | 15 | 27 |
| Cyclomatic Number | 122 | 132 | 21 | 85 | 87 | 23 | 19 | 24 | 34 | 84 | 134 | 110 | 124 | 89 | 129 |

The distance between the records can be calculated as the numerical difference on the attribute "# of modifications." Figure 6 illustrates a possible clustering of the software modules using such distance measure. An agglomerate hierarchical clustering algorithm is used to group the software modules, labeled from "A" to "O," based on their number of modifications.

The algorithm uses the distance measure to agglomerate single records in clusters and interactively group these clusters into higher level clusters until all the records are clustered together at the highest level cluster. The distance measure used to group clusters together is the minimum distance between the records of the lower level clusters. Consider cluster <FHC> at level 2 as an example. It is clustered with cluster <GM> at level 3, because the distance between module C and module G is 3, i.e., 8 minus 5.

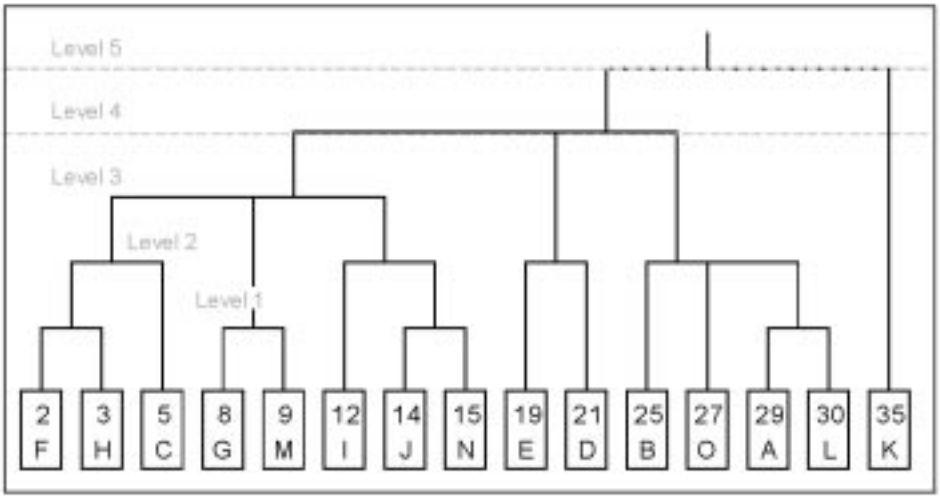



**Figure 6. Single Link Clustering of Software Modules Based on Number of Modifications**

Other distance measures could be used to cluster the modules on the "# of modifications" attribute. One could, for example, normalize the "# of modification" values to range between zero and one. The distance between clusters could be calculated as the maximum or average distances among its member records. These clustering approaches will be discussed later in this section.

Let us now consider the problem of clustering the same modules using not one but two attributes. Suppose that one wants to use the cyclomatic number as well as the # of modifications to cluster the modules. One now has to calculate the distance between two records in a two dimensional space. The problem is equivalent to calculating the distance between two points in a scatter plot chart. There are a number of metrics that can be used to do this. The two most common are the Euclidean and the Manhattan distances. In both cases, one calculates the distance between the records in each of the attributes. The Euclidean distance is calculated as the square root of the sum of the squared distances. The Manhattan distance is calculated simply as the sum of the distances. Consider the attributes values for modules B and D in Table 2. The Euclidean and Manhattan distances of those attributes are calculated as follows:

$$\text{Euclidean}(B, D) = \sqrt{(21\text{-}25)^2 + (85\text{-}132)^2} = 47.17$$

$$\text{Manhattan}(B, D) = 4 + 47 = 51$$

Both measures can just as easily be extended to an N-dimensional space. In this case, each record is described by N attributes. The distances between the records are calculated as:

$$\text{Euclidean}(B, D) = \sqrt{(a_1(B) - a_1(D))^2 + \ldots + (a_n(B) - a_n(D))^2}$$

$$\text{Manhattan}(B, D) = |a_1(B) - a_1(D)| + \ldots + |a_n(B) - a_n(D)|$$

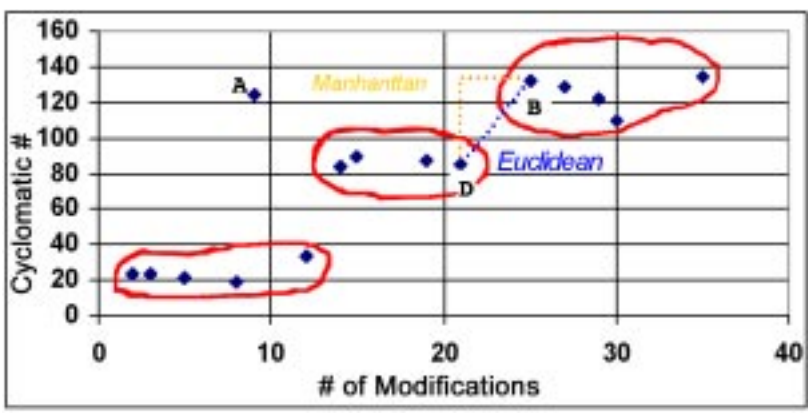



**Figure 7. Euclidean and Manhattan Distances in Two Dimensions**

Figure 7 plots the software modules according to the "# of modifications" and "cyclomatic number" attributes. The Euclidean and Manhattan distances between records "B" and "D" are pictured in it. A good observer will note that "cyclomatic number" has most of the weight on the distance calculation. This happens because this attribute has a larger range of values than the "# of modifications" attribute. Normalizing the attribute values so they vary between a defined range of values can solve this problem.

Rescaling of values can also be used to give different weights to different attributes. This is useful when the data analysts consider some attributes more important than others with respect to the data analysis objectives.

Another problem appears when attributes are measured in ordinal or nominal scale. In this case, attributes are not measured in numeric values. We have to map them to numeric scales in order to use them in distance measures like the one we discussed. The best solution for nominal attributes is to use a perfect match police. When two records have the same nominal value in an Attribute "X," they are assigned distance 0 in the "X" dimension. When they have different values, they are assigned distance 1 (or another predefined value) in this dimension. Consider, for example, that the software modules of our example are classified in hard real time or not. In this case, the attribute "real time module" will have the values "yes" and "no." If the value for attribute "real time" is "yes" for module "B" and "no" for module "D," the distance between them will be 1 with regard to this attribute. After this mapping, the calculation of Manhattan or Euclidean distances is straightforward. One can use the same approach to map ordinal scales to numeric values, but a more useful approach is to map the ordinal values to an ordinal numeric scale. An attribute using an ordinal scale of values like <low, medium, high> can be mapped to the values <1,2,3>. The distance between records can be directly calculated by the difference between those numbers. The advantage of this approach over the previous one is that it captures the fact that the distance between "high" and "low" is higher than the distance between "high" and "medium" values. Adjusting the numeric scales to normalize attribute values may further refine both the nominal and ordinal scale approaches. Adjustments may also be done to compensate for different frequency of appearance of values in nominal and ordinal scales.

### 5.3.2 Clustering Algorithms

There are two main types of clustering algorithms: hierarchical and non-hierarchical. Figure 6 shows an example of hierarchical clustering. This type of technique creates a hierarchy of clusters from small to big. Consider Figure 6 as an example. At level 3, there are 4 clusters in the cluster hierarchy: <F,H,C,G,M,I,J,N>; <E,D>; <B,O,A,L>; and <K>. At level 4, there are only two clusters in the cluster hierarchy: <F,H,C,G,M,I,J,N,E,D,B,O,A,L> and <K>. The advantage of this approach is that an expert can subjectively choose what is the best clustering for a given problem based on his domain knowledge.

Nonhierarchical techniques require some a priori decision on the number of desired clusters or the minimum required "nearness" for records in a cluster. This limits somewhat the expert participation in the mining process. The expert cannot choose the best clustering based on his subjective opinion like he can on hierarchical clustering. However, nonhierarchical clustering usually requires less computational power to create clusters, and this may be a critical issue in large data sets.

#### *5.3.2.1 Nonhierarchical Clustering*

The most common nonhierarchical techniques are based on reallocation methods. The basic algorithm for a reallocation method, also known as K-means method, was proposed by MacQueen [40]. It works as follows:

1. Select a number K of desired clusters.
2. Pick a record to be the centroid (center) of each of the chosen clusters.
3. Go through the data set and assign it to the nearest cluster.
4. Recalculate the centroid (or means) of the new clusters.
5. Repeat steps 3 and 4 until there is minimum relocation of records between the clusters.

Figure 8 shows the K-means method graphically. K is equal to 3 in this case. Three clusters, named A, B, and C, are shown. Part (i) of the figure shows steps 2 and 3 of the algorithm. In step 2, three data records are chosen and their positions assigned as centroids for the Clusters A, B, and C. In step 3, the other records are assigned to the cluster with the nearest centroid. In step 4, shown in Part (ii), the new centroids for clusters A, B, and C are calculated based on the means of the record associated with them in Part (i).

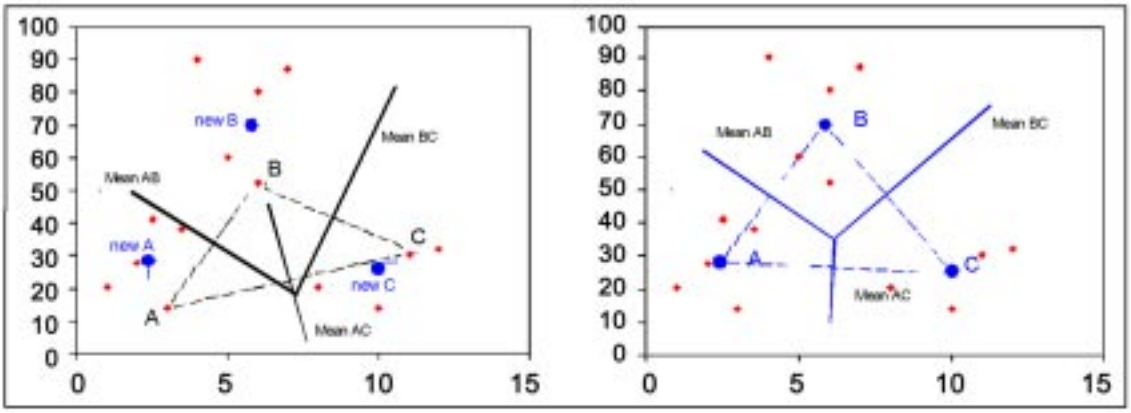



**Figure 8 K-means Method**

It is important to remark that the number of clusters that will be produced by the K-means method is established a priori in step 1 of the algorithm. This is a drawback when the data sets split gracefully in a number of clusters different than the number that was selected a priori. Figure 8 shows a data set that splits gracefully in three clusters. If a data analyst had selected a two-cluster split for this data set, the algorithm would ungracefully try to fit the data set in two clusters instead of using the more elegant solution with three clusters. However, this type of situation is not common with sizeable data sets. In those cases, there are many ways to cluster a data set. One seldom would end up with a wrong answer.

### *5.3.2.2 Hierarchical Clustering*

Hierarchical clustering creates hierarchy of clusters on the data set. This hierarchical tree shows levels of clustering with each level having a larger number of smaller clusters. Figure 6 shows an example of such a tree. Hierarchical clustering algorithms work by agglomerative or divisive clustering. Agglomerative algorithms start from the bottom of the hierarchical tree, having as many clusters as there are records in the data set. At each level of the tree, the algorithm merges smaller clusters into larger ones. The process is repeated until all the records in the data set are members of one cluster at the top level of the tree. Divisive algorithms work in the opposite direction. It starts at the top of the tree, with all records in one cluster. It then divides larger clusters into smaller ones as is goes down the hierarchical tree.

Agglomerative algorithms are more commonly used in practice because they are usually faster than the divisive ones. A general algorithm for agglomerative clustering work as follows:

1. Create a cluster for each record in the data set.
2. Merge the nearest clusters into large ones.
3. Repeat the process until only one cluster remains.

The key step of the agglomerative algorithm is to choose which clusters will be merged in Step 2. This decision depends on the way the distance between clusters is calculated. There are basically four major ways of doing this:

1. *Single link method*: The distance between two clusters is equal to the distance between the two closest records in them.
2. *Complete link method*: The distance between two clusters is equal to the distance between the two most distant records in them.
3. *Centroid method*: The distance between two clusters is equal to the distance between their centroids.
4. *Ward's method*: The total distance between the cluster's records and its centroid is computed for each possible merge. The merge with the smallest total resulting distance is selected as the next merge.

Figure 9 shows the distances used in the single link, complete link, and centroid methods. The example shows three clusters that would be merged in three different ways by those three methods. The single link method would merge clusters B and C next because they have the minimum single link distance. On a similar fashion, the complete link method would merge clusters A and C. And, the centroid link method would merge clusters A and B.

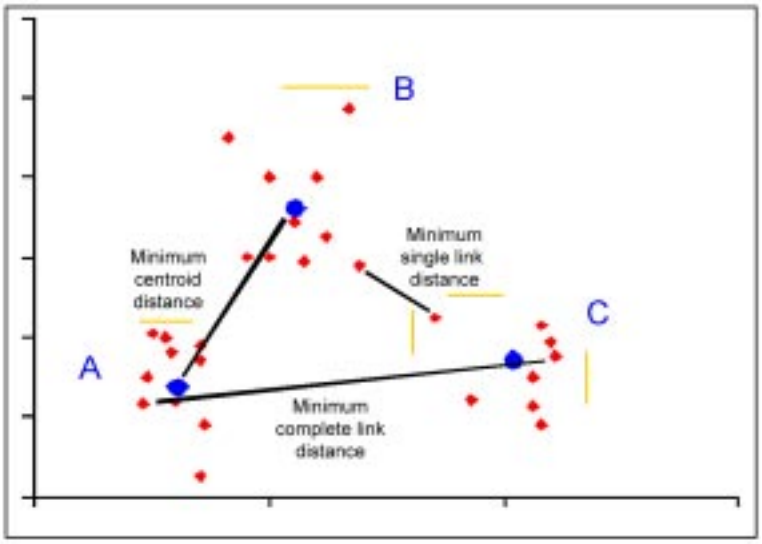



**Figure 9. Distance Measures used by the Agglomerative Methods**

Studies have shown that there is no single method that is superior for all types of data [42]. The single link method can create elongated clusters because it merges based on just a single pair of near records. The complete link method favors the creation of small compact clusters. The centroid method produces clusters that are balanced in between the elongated single-link clusters and the tight single-link clusters. The Ward's method tends to produce symmetric hierarchies and is considered very effective in producing balanced clusters. It has, however, difficulty in dealing with outliers and elongated clusters.

Contrary to nonhierarchical clustering, hierarchical clustering does not require the data analyst to enter the number of clusters that will be produced during data analysis. Hierarchical clustering is solely defined by the data records and clustering algorithm. Also, hierarchical clustering allows the data analyst to choose the best cluster distribution by allowing him to look at the hierarchical clustering tree after data analysis. At this point in time, data analysts can employ their domain expertise to make the best cluster selection for the problem at hand.

### 5.3.3 Using Clustering for Classification and Prediction

The example in Figure 7 shows three clusters of software modules. By interpreting this chart, a domain expert would conclude that these modules can be classified into three categories: (1) low complexity low volatility modules, lower left corner of the chart; (2) medium complexity medium volatility modules, middle of the chart; and (3) high complexity high volatility modules, right upper corner of the chart. The chart also shows that module "A" is an outlier, a module with low complexity and high volatility. This type of interpretation highlights two of the most important applications of clustering techniques: (1) producing a high-level view of what is going on in the data; and (2) automatically identifying data outliers. The third application is to use clustering for classification and prediction.

Suppose that a data record, ER, has an attribute ER(Y) that one wants to classify (or predict) and a set of defined attributes $ER(X_1,X_2,\ldots,X_N)$ that will be used as independent variables to classify it. Suppose that the historical data set has data records with values for all attributes $\{X_1,X_2,\ldots,X_N,Y\}$. One way to estimate ER(Y) is by finding the K-nearest neighbors of $ER(X_1,X_2,\ldots,X_N)$ according to a predefined distance measure. The value of ER(Y) is then estimated to be the average Y value for the K-nearest neighbors of ER. This method tends, however, to overfit the data as the whole historical data set composes the estimation model. In order to produce a simpler and more accurate model, one has to simplify the data set by doing the following operations:

1. Merge nearby historical records together on an averaged record if this operation does not decrease the accuracy of the predictions.
2. Remove historical records that do not affect the accuracy of the predictions.

What these two operations do is to produce small clusters of records that are homogeneous with respect to the attribute one wants to predict (Y). This creates a model that is a generalization of the overfitted model created previously by the whole data set.

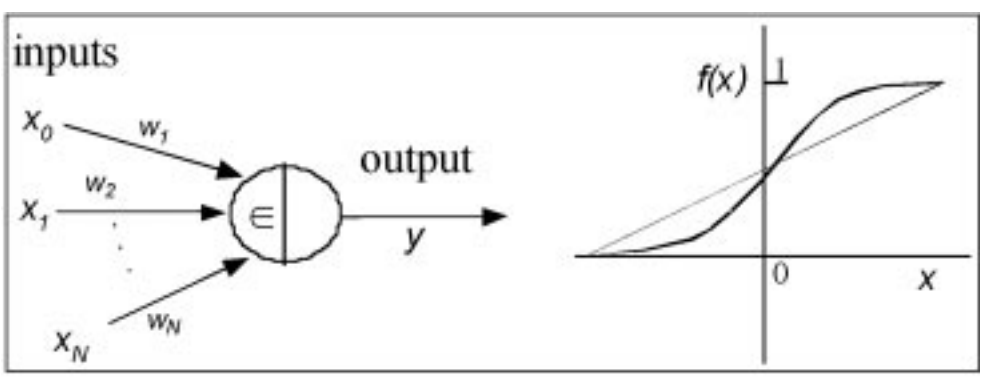



**Figure 10. A Neuron and a Sigmoid Function**

## 5.4 Artificial Neural Networks

Neural networks have been one of the tools of choice for building predictive software engineering models [30][32][33][37][52]. They are heavily interconnected networks of simple computational elements [7][53]. An example of such an element, often called a neuron, is shown in Figure 10. The neuron has N inputs $x_1, x_2, \ldots, x_N$ and one output y, all having continuous values in a particular domain, usually [0,1]. Each neuron input also has a weight ($w_1, w_2, \ldots, w_N$) that determines how much each input contributes to the neuron output y.

The neuron computes its output by calculating the weighted sum of its inputs and passing it through a non-linear filtering function f(x). Figure 10 shows a sigmoid, a function commonly used for this purpose. The output is calculated as:

$$y = f\left(\sum_{i=0}^{N} w_i x_i\right), \text{ where the sigmoid function is } f(x) = \frac{1}{1+e^{-x}}$$

$$\therefore \qquad y = \frac{1}{1+e^{-\sum_{i=0}^{N} w_i x_i}}$$

Neural networks are built by connecting the output of a neuron to the input of one or more neurons. Input connections are then assigned to a layer of nodes, called input nodes, and outputs are assigned to another layer of nodes, called output nodes. Figure 11 shows a neural network adapted from [52]. In this example, the network architecture aims to build a software effort estimation model. It uses inputs derived from COCOMO's cost drivers and other important software attributes. The COCOMO cost drivers are discussed in depth in [8] and [9]. The attributes shown as inputs here are: adjusted delivered source instructions (AKDSI); total delivered source instructions (TKDSI); execution time constraints (TIME-const); storage time constraints (STOR-const); and, computer language (L-Cobol, L-Fortran, and L-PL1). The output is an effort estimate based on the input values and the weights of the network connections.

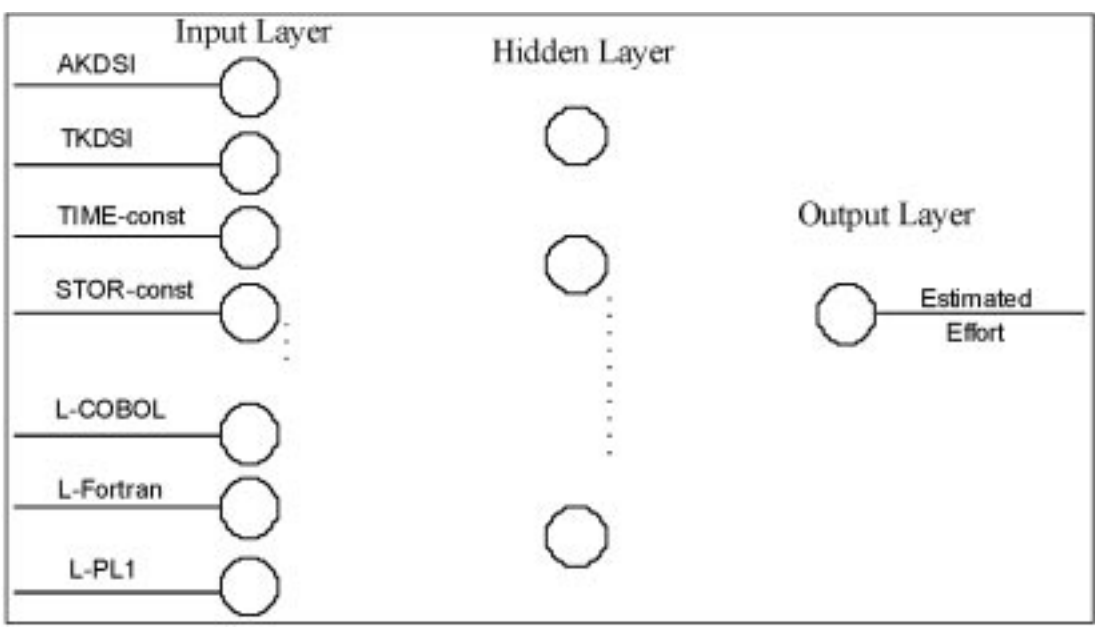



**Figure 11. A Neural Network for Software Development Effort Estimation**

The main steps in building a neural network for classification or prediction, such as the one in Figure 11, are: (1) identify the network inputs and outputs; (2) process the input and output values so that they fall into a numeric range, usually between 0 and 1; (3) choose an appropriate topology for the network by defining the number of hidden layers; (4) train the network on a representative set of examples; (5) test the network on a test set independent of the training set and retrain the network if necessary; (6) apply the generated model to predict outcomes in real situations.

The main challenge in building a neural network model is to train the network (step 4). This is achieved by setting the network connection weights so that the network produces the appropriate output patterns (effort estimation, in our case) for corresponding input patterns (software attributes and cost drivers values, in our case). The idea is to use a set of examples, called a training set, to adjust the network weights to the right predictive values.

### 5.4.1 Backpropagation

The most common approach to train a neural network is backpropagation. This approach is discussed in detail in [27] and [38]. The idea of backpropagation is to start with a random set of weights, use an example in the training set to estimate the output, and compare the estimate with the actual value. The backpropagation algorithm uses this comparison to calculate the estimation or classification error. The error is then fedback through the network and the weights in the network are adjusted to minimize the error. The bigger the error the more the weights are modified. Each node sees their input node as advisors. The more an input node contributes with the wrong advice, the more its weight is downgraded. The actual weight calculation also includes the slope of the filtering function and a learning rate value.

After, the adjusted weights are backpropagated from the output to the input nodes. A new example is showed to the network. After being shown enough training examples, the weights on the network no longer change significantly. This is when the training stops and the network is said to have learned the concept. Critical to this process is the learning rate, i.e., how much the weights are adjusted to compensate the error at each training cycle. The learning rate controls how quickly the network reacts to a particular error. If the learning rate is too aggressive, the network will act quickly to correct one type of error but may corrupt the weight structure previously assembled to prevent other types of error. This might bar the network from converging to a specific set of weights. On the other hand, if the learning rate is too conservative, the network will take a long time to converge to a specific set of weights. This is especially troublesome when one has a limited training set, as is frequently the case in software engineering.

Usually, the best approach for the learning rate is to start aggressive and decrease it slowly as the network is being trained. Initially, the weights are random, so large oscillations are useful to get in the vicinity of their best values. However, as the network gets closer to a solution, the learning rate should be decreased so the weights can be fine-tuned to an optimum solution.

### 5.4.2 Local Maximums and Overfitting

Two dangers of neural networks training is building models that arrive at a local maximum or overfitting the training set. A local maximum or local optimum solution happens when the network produces good results and adjusting the weights no longer improves the performance of the model. However, there is some other combination of weights, usually quite different from those previously derived, that yields a much better model. Training the network several times using different initial weight settings mitigates this problem, as one can select the best model among the training trials.

Overfitting happens when a model works very well on its training set, but has poor performance on new data. Although overfitting occurs in all predictive modeling, it is particularly problematic in neural networks. Larger neural networks can easily overfit small training data sets. This is quite problematic because one cannot interpret how the model works (like classification trees, for example). Solutions for this problem include: (1) using training sets that are large when compared with the number of input nodes; (2) using a limited number of hidden nodes (some suggest 2/3 of the number of input nodes) and a few hidden layers; and (3) always using test sets to validate the obtained neural network model.

## 5.5 Optimized Set Reduction

Optimized Set Reduction (OSR) is a technique that was specifically developed in the realms of software engineering data analysis [11][12]. Its approach is to determine what subsets of data records provides the best characterization for the entities being assessed. It works by successive decompositions of the training set into subsets. At each step of decomposition an attribute is selected and records having the same values on the selected attribute are extracted from the training set to form a new subset. This is done recursively on the subsets until a termination criteria is met. Prediction and classification can then be done based on the average value of the dependent variable on the terminal subsets.

A simplified example of an OSR process is seen in Figure 12. The example, adapted from [30], shows part of a model for maintenance effort prediction. Subset1 is a subset of the training set for which the maintainers confidence on the task to be performed is HIGH. Similarly, Subset2 is extracted from Subset1 by limiting the type of maintenance task to corrective (CORR) activities. In the figure, Subset2 meets the termination criterion and the effort prediction is done based on the record contained in this subset.

Like classification trees, OSR produces models that can be interpreted by a domain expert. However, unlike classification trees, OSR does not select a unique attribute at each decomposition level. In the above example, the technique does not have to use the attribute confidence to derive others subsets from the training set. This helps the technique to work well in small data sets.

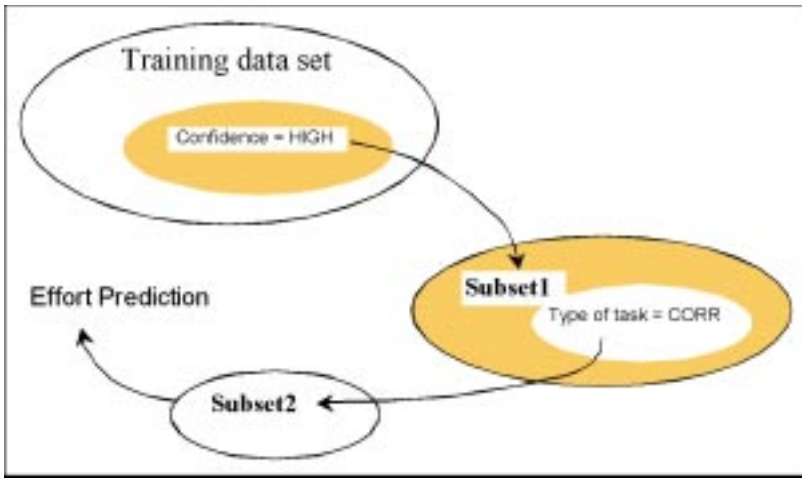



**Figure 12. An OSR Hierarchy**

## 5.1 Bayesian Belief Networks

A Bayesian Belief Networks (BBN) are graphical networks that represent probabilistic relationship between variables [13][16]. Recently, BBNs have attracted attention from the software engineering community [43][44]. BBNs can be used by domain experts to articulate their beliefs about the dependencies between different process and product attributes. With them one can propagate consistently the impact of known attribute values on to probabilities of uncertain outcomes.

Figure 13 shows a BBN from an example kindly provided by Norman Fenton [22]. In the example a BBN is built to predict software reliability. The nodes represent continuous and discrete software attributes. The arcs represent influential relationships between the attributes. For example, reliability is defined by the number of latent faults and frequency of operational usage. Similarly, the coder's performance is defined by their experience, the problem complexity, and the use of IEC 1508 (a safety integrity standard). Each node is associated with a probability table (PT) that maps the inputs value distribution into an output value distribution. The probabilities table may be derived based on subjective expert opinion or based on objective measured data. This flexibility is one of the main strengths of this technique. Once the PTs are defined the BBN model is ready to be used. The BBN will produce a probability distribution for its output attributes once its input attributes are measured.

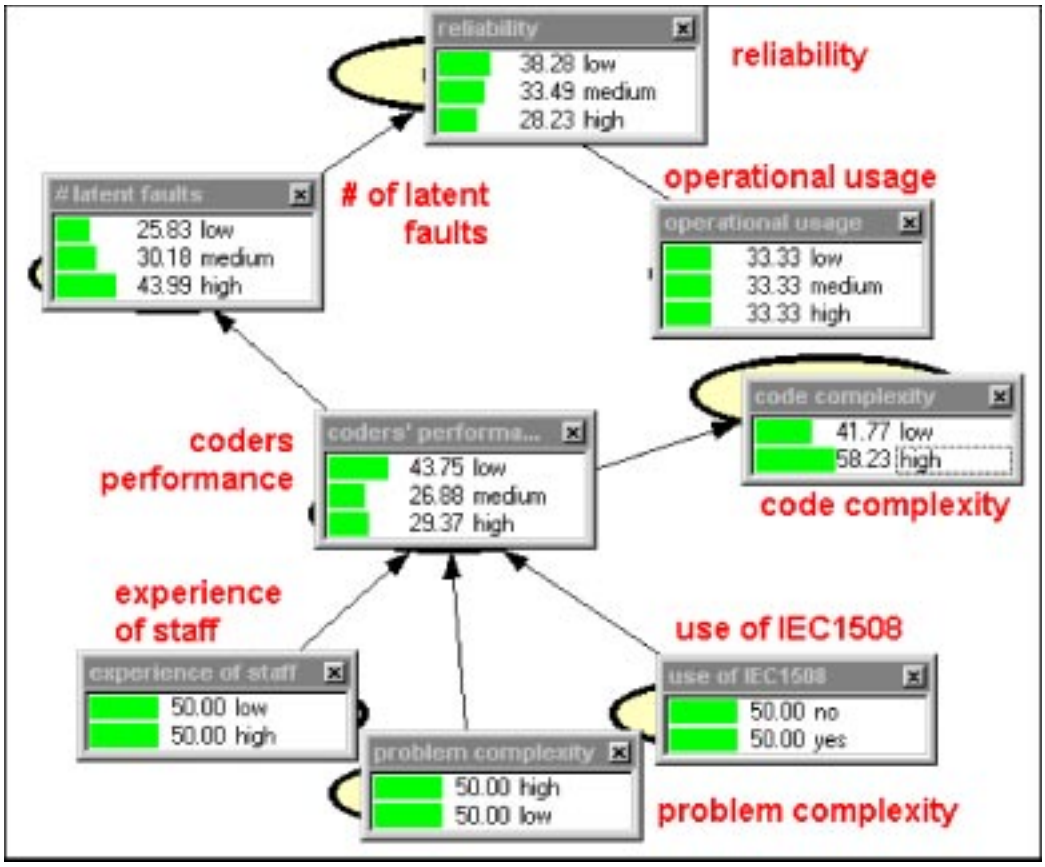



**Figure 13. A BNN for Software Reliability Prediction**

Current BBN tools support the construction of large networks and complex node probability tables for both discrete and continuos attributes. BBNs produce interpretable models that allow experts to understand complex chain of events through visual graphical displays. They also model uncertainty explicitly in their estimates. For these reasons, BBN is a promising technique for supporting decision making and forecasting in the software engineering field.

## 5.7 Visualization and Visual Data Mining

Data visualization can be thought of as the science of mapping volumes of multidimensional data into two dimensional computer screens. Visualization is an important technique for data mining because humans excel at processing visual information. Humans can extract important features of complex visual scenes in a matter of milliseconds. Good visualization techniques play with this human strength by displaying complex information in a form that can be quickly processed by the human brain. Bell Labs work on source code visualization is an excellent example of how visualization can be used in software engineering [2][18][19].

### 5.7.1 Visualization of Multivariate Data

There are several ways that data can be visually displayed. The challenge is to display multidimensional information in a two dimensional screen. This is achieved by associating data records or set of data records with a series of "visual attributes." Each visual attribute is then associated with a dimension in the real data. Consider the example in Table 3. In this table, the data records, representing software modules, are described in five dimensions: fan-out, fan-in, coupling, number of modifications, and cyclomatic number.

**Table 3. Data Records Describing Software Modules on Several Attributes**

| Modules | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Fan-out | 7 | 8 | 4 | 6 | 5 | 1 | 1 | 2 | 2 | 5 | 5 | 6 | 6 | 4 | 6 |
| Fan-in | 4 | 5 | 2 | 3 | 3 | 1 | 1 | 4 | 3 | 3 | 2 | 7 | 6 | 3 | 7 |
| Coupling | 14 | 22 | 7 | 8 | 4 | 4 | 3 | 5 | 5 | 6 | 12 | 11 | 10 | 7 | 13 |
| Number of Modifications | 29 | 25 | 5 | 21 | 19 | 2 | 8 | 3 | 12 | 14 | 35 | 30 | 9 | 15 | 27 |
| Cyclomatic Number | 122 | 132 | 21 | 85 | 87 | 23 | 19 | 24 | 34 | 84 | 134 | 110 | 124 | 89 | 129 |

In order to display those five dimensions at the same time in a visual display, the visualization application has to map each software module attribute to a visual attribute. Figure 14 shows a screen shot of a display built using a data mining tool called DataMiner. A description of DataMiner can be found in Appendix B. The picture displays the data records of Table 3 mapping the software module attributes to the following visual attributes: fan-out is shown as size, fan-in is shown as color, coupling is shown as X-position, number of modification is shown as Y-position, and cyclomatic number is shown Z-position.

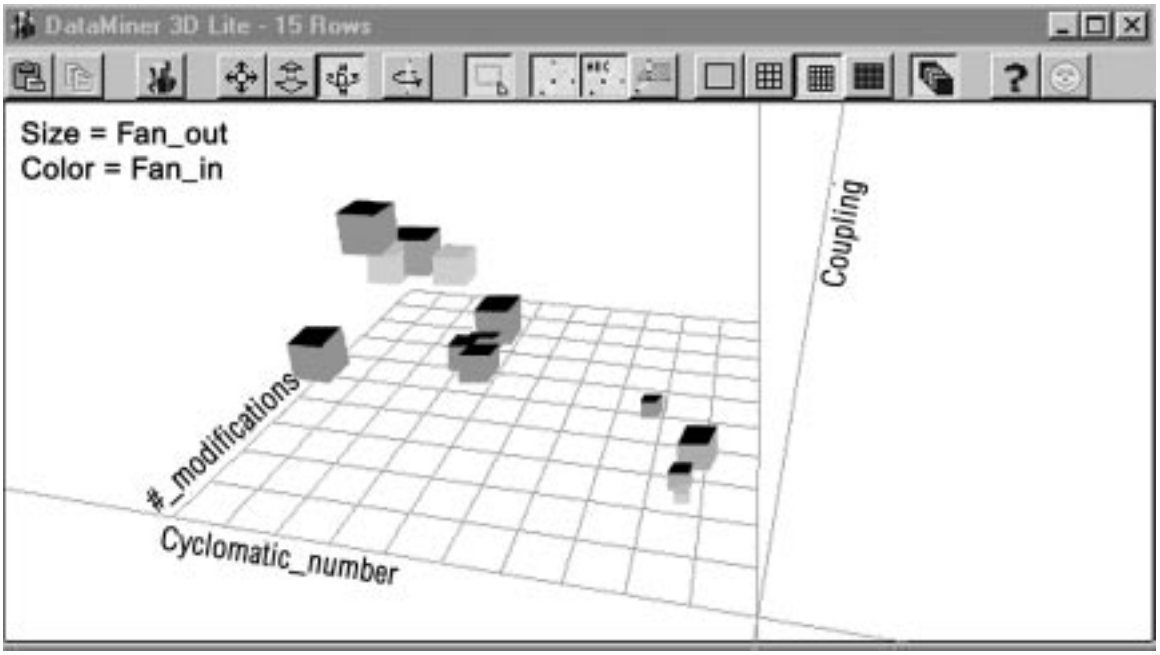



**Figure 14. A Multivariate Display Built Using DataMiner**

### 5.7.2 Visual Data Mining

Many modern data visualization tools combine powerful visual displays with easy to operate data selection and display controls. These functionalities allow domain experts to interactively explore data so efficiently that they are able to find interesting data patterns without using automated data mining algorithms. This type of data mining is sometimes called visual data mining. A good visual data mining tool has the following functionalities:

- Ability to interactively navigate on the visual canvas allowing zooms, rotations, and scans over the displayed data.
- Ability to interactively control display formats and the visual attributes of the displayed data.
- Ability to interactively control the granularity in which the data is visualized, allowing the domain expert to look at it from a high level perspective or to drill down to particular data sets. This enables domain experts to analyze the big picture or to focus on details and singularities of the displayed information.

Figure 15 shows a screen shot from a visual data mining tool called Spotfire. A description of Spotfire can be found in Appendix B. The picture displays the data records on software errors extracted from a real world software engineering database. The right side of the screen shows the error attributes. Some of the shown attributes are: the date the error was detected (DATE_DETECTED); date the error was finally fixed (COMPLETION); number of components affected (COMPONENT); source of the error (ERR_SOURCE); and error class (ERR_CLASS). Associated with those attributes, there is a query device, a widget that can be interactively manipulated to select or deselect the data records being examined. In the particular example, the error source (ERR_SOUR) is deselected for errors originated in "(4)CODE," and the date of detection (DATE_DETE) was selected for errors detected between July 94 (94-07-11) and December 97 (97-12-19).

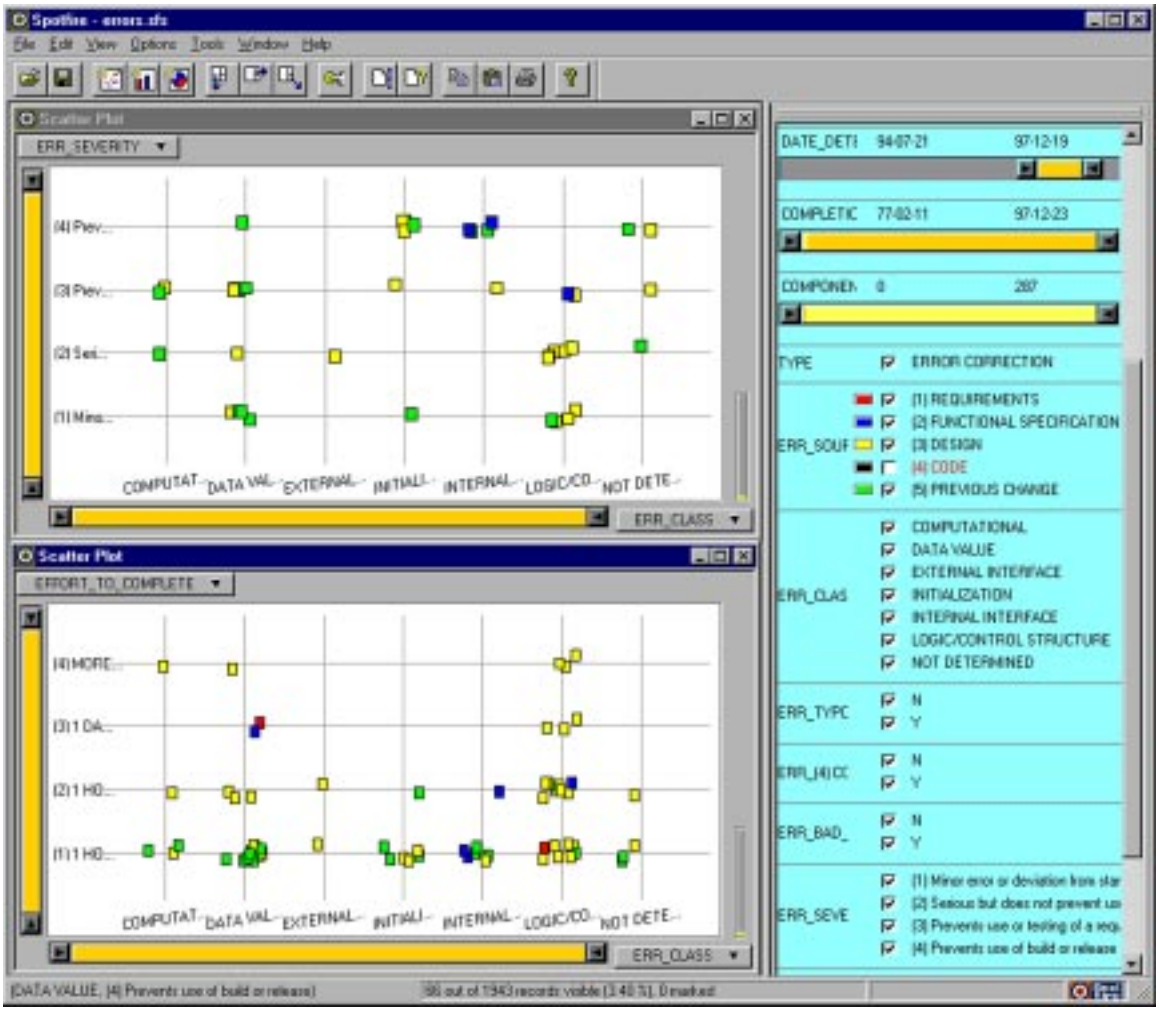



**Figure 15. An Interactive Visual Data Mining Display Built with Spotfire**

The left side of the screen shows two charts. The first plots effort to fix the error versus error class (EFFORT_TO_COMPLETE x ERR_CLASS) and the second plots error severity versus error class (ERR_SEVERITY x ERR_CLASS). Both charts are colored by error source (ERR_SOURCE).

In this particular tool, the chart displays on the left are linked to the query devices on the right, so the domain expert can interactively explore the data. In the example, the charts show no black squares because the domain expert deselected ERR_SOURCE=(4)CODE on the right side of the screen. Likewise, there are no data records for errors detected before July 94 or after December 97 shown ( 94-07-21≤ DATE_DETECTED ≤97-12-19) on the charts. The strength of visual data mining approaches like the one we just described lies in the fact that domain experts can explore the multivariate data in real time. This enables them to understand trends, and raise and test hypotheses with a few mouse clicks.

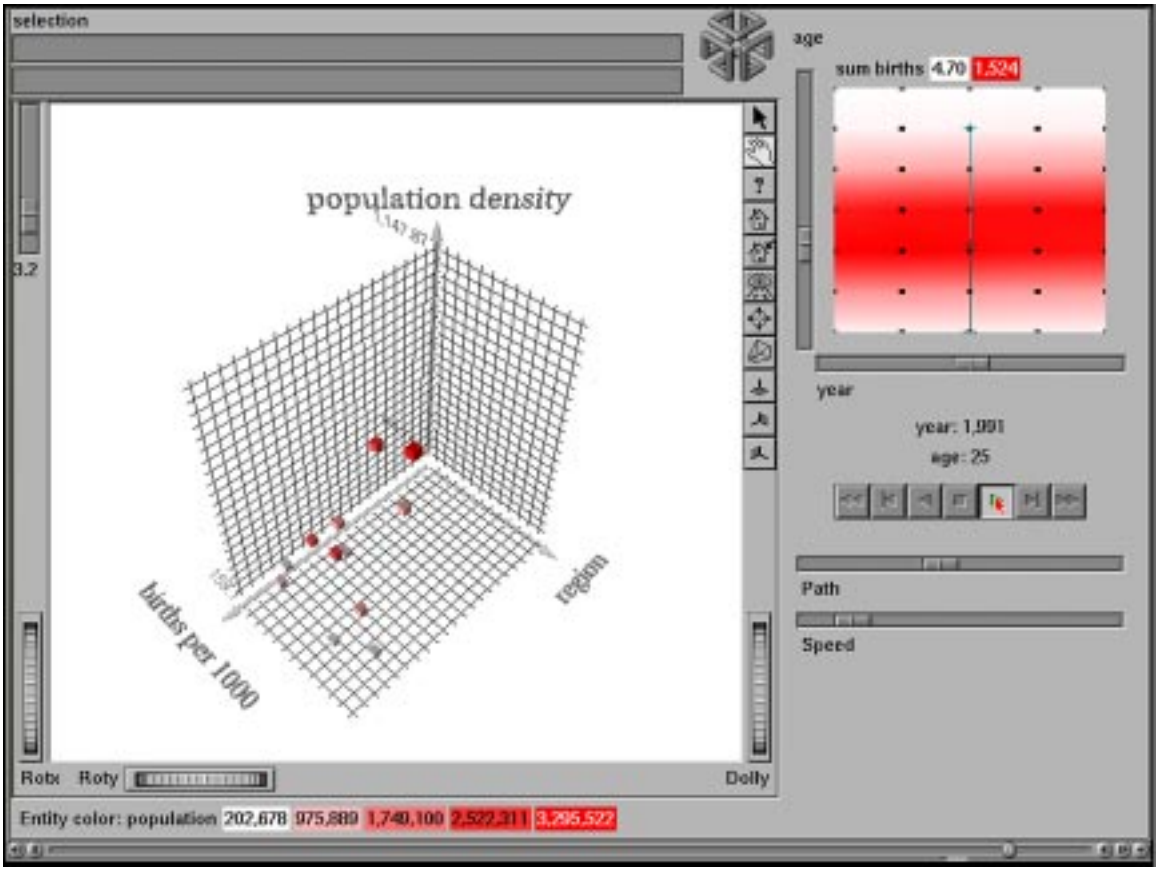



**Figure 16. An Animated Visual Data Mining Display Built Using SGI's MineSet**

### *5.7.2.1 Animation*

The most recent data visualization tools also include data animation features. Common visualization techniques play with the human ability for processing complex visual information. Animation techniques play with the human ability for detecting motion of visual information. For this reason, animation can be a powerful tool for analyzing data, especially temporal data.

Data animation usually works by choosing criterion variables and letting the animation tool display the behavior of the data records for different values of these variables. Figure 16 shows a screen shot of an animation display built by a tool called MineSet. A description of MineSet can be found in Appendix B. This particular example is not in software engineering but in social sciences. The left side of the screen shows a scatter plot containing birth rate, population density, region, and population (color shading). The right side of the picture shows the widgets used to control the animation. The control for animation speed is shown at the bottom right side. The controls for playing, stopping, and pausing the animation are shown in the middle. The controls for selecting values for the criterion variables are shown at the top. This particular example uses two criterion variables, women's age and year of birth. The little chart on the top right side is used to define the values to be animated on the criterion variables. These values are determined when the data analyst chooses an animation path in this control chart. In the example, the path, a vertical line, determines that the records will be animated on the variable "age." The picture shows a snap shot of the animation at the point where the variable "age" is equal to "25." The real animation movie for this example can be downloaded from:
http://www.sgi.com/software/mineset/movies/scatterviz.mov

# 6. Interpretation and Assimilation (Knowledge Extraction)

The techniques presented in Section 5 are used to mine interesting information to domain experts, software engineering experts, in our case. This information has to be assimilated by those experts in order to be transformed into useful knowledge.

## 6.1 Patterns, Models, and Knowledge

The information mined by the techniques described in Section 5 is expressed as patterns or models. Consider that the following sequence of values is mined by a mining technique: *ABABAB*… This sequence of values describes a *pattern* in the data. If this pattern is generic enough, it can be abstracted into a *model*, such as: if A then B will follow. A model is thus an abstraction of the original data set that is ready to be applied for decision making, classification, or prediction in the organization.

There are basically two types of data mining techniques: (1) those that extract patterns, and (2) those that produce models from the mined information. Techniques like neural networks and classification trees produce classification or prediction models directly from the data. These models code knowledge in a form that is readily usable by an organization. Once they have been mined, they can be used for classification of new information, for example.

Techniques like clustering (Section 5.3.2) and attributes focusing (Section 5.2.3) only extract patterns from the data. These patterns have to be examined by a domain expert in order to be transformed into useful knowledge. In this case, models are created in the domain expert's head when he interprets the mined pattern.

Both classes of techniques have advantages and disadvantages. Neural networks and classification trees can produce models that represent complex knowledge that cannot be abstracted by an expert from a data pattern display. On the other hand, experts can use their background knowledge about the application domain when they are interpreting the patterns extracted from a database. This can improve knowledge discovery by bringing domain specific information to the knowledge discovery process.

Consider the following pattern as an example: A, AB, ABC, ABCD, … The best model that could be automatically extracted from this pattern would say something like: if string <y> then string <y>X will follow. This model can only predict that a string <ABCDx> will follow <ABCD>. It doesn't have the knowledge that E follows D because this information is not present in the database. However, any person that has "background knowledge" of how letters are ordered in the alphabet can predict that ABCDE will follow ABCD. This happens because this "background knowledge" is automatically incorporated into the model one creates by looking at the above pattern.

## 6.2 Interpreting Patterns through Visualization

It is key that the techniques that produce patterns present these patterns in a format that can be easily interpreted by a domain expert. This is usually achieved by presenting patterns visually as graphs or charts. Patterns expressed in text and table formats are usually difficult to interpret. Consider the following data pattern as an example.

**Table 4. A Data Pattern Presented in a Tabular Format**

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Y | 0 | 0.259 | 0.5 | 0.707 | 0.866 | 0.966 | 1 | 0.966 | 0.8866 | 0.707 |
| X | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 |
| Y | 0.5 | 0.259 | 0 | -0.259 | -0.5 | -0.707 | -0.866 | -0.966 | -1 | -0.966 |
| X | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 | 31 |
| Y | -0.866 | -0.707 | -0.5 | -0.259 | 0 | 0.259 | 0.5 | 0.707 | 0.866 | 0.966 |
| X | 32 | 33 | 34 | 35 | 36 | 37 | 38 | 39 | 40 | 41 |

When compared with the pattern <ABABAB> presented at the beginning of this section, the pattern above appears to be quite difficult to interpret. Figure 17 shows the above pattern in a graphical format. This pattern can now be mapped to a sine like model and easily interpreted by a domain expert. The graphical display shows that the above pattern is simple and even comparable, by its periodical nature, to pattern <ABABAB>. Nevertheless, simple patterns like the one in Figure 17 are not the rule. Mined patterns can be quite complex and vary in several dimensions. Section 5.7 showed how data mining techniques try to reduce this complexity by using advanced data visualization techniques. Unfortunately, the data mining tools cannot eliminate all this complexity for domain experts. Besides relying on the tools to control the way information is presented to them, domain experts should be aware of their own limitations and be well prepared to try to interpret mined patterns.

Humans have limitations on the amount of information that they can hold in their short-term memory at one time. That is why the pattern in the tabular data presented in Table 4 is difficult to interpret. Usually, after a few pairs of numbers are read, any new incoming information displaces the previous information. In fact, it is well-known that this number is around seven plus or minus two items [41]. This small short-term memory capacity makes the performance of domain experts suffer as soon as they are exposed to patterns displayed in textual or tabular format. As seen in Section 5.7, visualization techniques can significantly mitigate this problem. However, good visualization support has its own limitations in dealing with the short-term memory problem. The problem re-appears when humans try to keep track of too many graphical patterns at once.

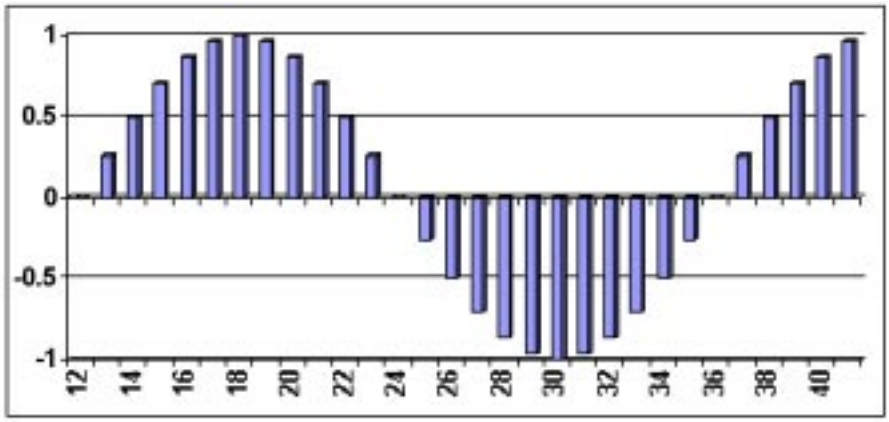



**Figure 17. A Data Pattern Presented in Graphical Format**

Suppose that a group of displays present a pattern only when compared to each other. Although each display is interpretable individually, the expert will miss the pattern expressed by the whole group if he does not display them in one screen. The problem is essentially the same as before, experts quickly lose track of individual items as they progress through the extracted information. Data mining techniques can mitigate this problem by:

1. sorting the results on some interestingness criteria, see discussion in Section 5.2.2.
2. displaying mined patterns in networks or hierarchies of related results, see Section 5.3.2.2.
3. incorporating interactiveness and/or animation to visual displays, see Section 5.7.2.

Besides using these techniques, domain experts must work methodically while trying to interpret data patterns. Whenever possible, they should work top-down, trying to look at the big picture before progressing to more fine-grained information. They should exam patterns involving global and important variables before going into more particular and less important variables. Experts should also focus on one reasoning thread at a time while "drilling down" the results. Lastly, experts should log interesting patterns and discoveries as soon as they are observed.

## 6.3 Evaluating Models

It is important to make sure that the knowledge gained through data mining is trustworthy. The section above discussed how patterns can be transformed into useful knowledge through the use of visualization. In many cases, knowledge gained through visualization is grounded on such strong patterns that no further investigation is needed. In other cases, the reliability of the mined information has to be statistically tested. A common approach is to express the discovered knowledge as hypotheses. The domain expert can then use traditional hypothesis-testing to test if the discovered information is indeed trustworthy. This report does not discuss hypothesis-testing techniques. This subject is covered by any good book on inferential statistics [39].

Some other data mining techniques, such as neural networks and classification trees, produce explicit classification or prediction models ready for use by a software organization. In this case, it is also important to validate the produced models. Cross-validation is a technique commonly used for this purpose. It works by dividing the data set into two chunks. The first chunk, called the training set, is used to produce the model. The second chunk, called the test set, is used to test the model's accuracy. In this approach, the model produced with the training set is used to estimate the value of each record in the testing set. The average accuracy is then considered as an estimate of the model's future accuracy. This validation approach has a major drawback. It reserves half of the data records for validation. These records are wasted from the model building point of view as they are never used in the model training procedure. A more complex validation model, called v-fold cross-validation [17], can be used to avoid this problem. It works as follows:

1. Divide the data set in v subsets of similar size, v_10.
2. Set one of the subsets as the test set.
3. Build a model with the remainder sets.
4. Use the model to estimate the values in the test set and calculate its accuracy.
5. Set the next subset as the test data set and return to Step 2.
6. After all data sets have been used, calculate the average accuracy of the models. This value is an estimate of the accuracy of the model built with all the data sets.

The advantage of v-fold cross-validation over common cross validation is that the former does not reserve any records for the validation process. All data records are used to build the model. This represents a clear advantage in domains like software engineering, where data sets are frequently small.

## 6.4 Interpretable Models

From the domain expert's point of view, model building techniques can be classified into two groups:

1. Techniques that produce models that can be interpreted by a domain expert.
2. Techniques that produce models that are not easily interpretable by a domain expert.

In the first category, we have techniques like classification trees and bayesian belief networks. In the second, we have techniques like artificial neural networks and genetic algorithms [26]. As hinted before in this report, the advantage of the first group over the second is that domain experts can use their background knowledge to interpret the model produced. This has two benefits. First, the domain expert can sanity check the produced model. Overfitting and significance problems can sometimes be avoided this way. Second, the expert may use the information explicitly described in the models to gain new knowledge about the domain being analyzed. In other words, the way a data mining technique builds a model can shed some new light on the problem at hand.

Some of the more modern model building tools are well aware of these possibilities. They append a lot of visual information to their models in order to facilitate their interpretation. Figure 18 shows a classification tree built using a tool called ALICE d'ISoft. This tree shows the values of the classified attribute "Success" graphically in colored bars and in percentages. It also shows the number of records supporting each node of the tree in order to avoid significance problems.

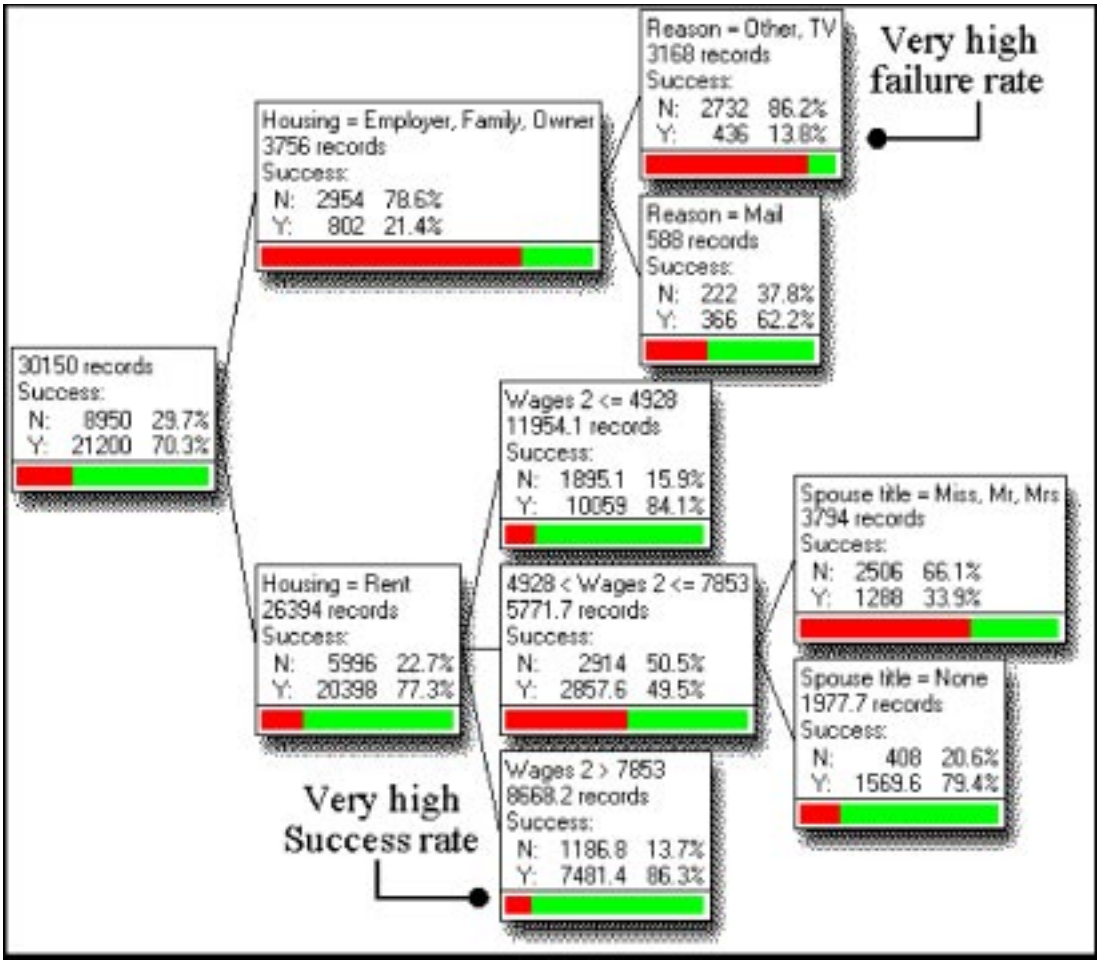



**Figure 18. An Interpretable Classification Tree Built Using ALICE d'Isoft**

# 7 Bibliography on Mining Software Engineering Data

This section surveys the software engineering literature on data mining. It focuses on the use of data mining related techniques to analyze software engineering data. Papers on the use of statistical-based techniques to build software engineering models are not included here.

This section focuses on key papers in the several sub-areas of data mining that we discussed in this report. In particular, it focuses on papers published in recognized software engineering journals. The list of papers discussed here gives a broad overview of what was published in the area. However, this list is not definitive or comprehensive by any means. Interested readers should be able to find many other interesting papers elsewhere.

## 7.1 Classification Trees

Richard W. Selby and Adam A. Porter. Learning from Examples: Generation and Evaluation of Decision Trees for Software Resource Analysis. *IEEE Trans. on Soft. Eng.*, 14(12), pp. 1743-1757, December 1988.

> This paper uses Quinlan's ID3 system to build classification trees to identify software modules with high development effort or high number of faults. The data - collected from a NASA production environment - included 4,700 software modules with 74 measured attributes. This is a seminal article on the application of machine learning-based techniques to build software engineering classification models. It has an excellent background section and good references.

Adam A. Porter and Richard W. Selby. Empirically Guided Software Development Using Metric-Based Classification Trees. *IEEE Software*, 7(2), pp. 46-54, March 1990.

> This paper is a follow up of the previous one. It presents a description of the authors' work on classification trees that is more oriented to software engineering practitioners. It focuses on the data analysis methodology and lesson learned by the authors.

Adam Porter and Richard Selby. Evaluating Techniques for Generating Metric-based Classification Trees. *J. Systems Software*, pp. 209-218, December 1990.

> This paper expands on the authors' previous work [51] by presenting a comparison of five variations on their original classification tree algorithm. It considers the accuracy and complexity of the generated trees resulting from different techniques for partitioning the attribute data values during the tree generation process.

Jeff Tian. Integrating Time Domain and Input Domain Analyses of Software Reliability Using Tree-Based Models. *IEEE Trans. on Soft. Eng.*, 21(12), pp. 945-958, December 1995.

> This paper presents the use of classification trees to build software reliability models. The tree-based model combine time domain - the software's ability to perform correctly over time - and input domain - the software's ability to perform correctly for different inputs - data to provide reliability assessment and improvement guidance. The approach was successfully tested in several commercial products developed at IBM Software Solutions Toronto Laboratory.

Jeff Tian and Joe Palma. Analyzing and Improving Reliability: A Tree-based Approach. *IEEE Software*, pp. 97-104, 15(2), March-April 1998.

> This paper is a follow up of the previous one. It presents the authors' work on tree-base reliability models in a form that is more oriented to software engineering practitioners. It focuses on the data analysis methodology and lesson learned by the authors.

Krishnamoorthy Srinivasan and Douglas Fisher. Machine Learning Approaches to Estimating Software Development Effort. *IEEE Trans. on Soft. Eng.*, 21(2), pp. 126-137, February 1995.

> This paper presents a study on the application of neural networks (using backpropagation) and classification trees (using the CART algorithm[10]) to estimate software development effort. Both techniques are discussed in detail and compared to other machine learning techniques previously used for software development cost estimation.

Taghi M. Khoshgoftaar and Edward B. Allen. Modeling Software Quality with Classification Trees. In *Recent Advances in Reliability and Quality Engineering*, Hoang Pham Editor. World Scientific, Singapore, 1999.

> This paper presents a study on the use of CART (a classification tree algorithm, [10]) to identify fault prone software modules based on product and process metrics. The data is drawn from large telecommunication software systems at Nortel.

## 7.2 Artificial Neural Networks

Taghi M. Khoshgoftaar and D. L. Lanning. A Neural Network Approach for Early Detection of Program Modules Having High Risk in the Maintenance Phase. *J. Systems Software*, 29(1), pp. 85-91, 1995.

> This paper describes the use of neural networks to classify software modules into high or low risk. Software product attributes based on complexity metrics are used to train the network. The authors argue that prediction techniques such as regression and statistical analysis are too sensitive to random anomalies in the data or are too dependent on assumptions that are not always met.

Taghi M. Khoshgoftaar, Edward B. Allen, John P. Hudepohl, and Stephen J. Aud. Neural Networks for Software Quality Modeling of a Very Large Telecommunications System. *IEEE Trans. On Neural Networks*, (8)4, pp. 902-909, July, 1997.

> This paper is a natural progression from the authors' previous work. It describes the use of neural networks to predict the number of faults in software modules at Nortel. Software product attributes based on complexity and size metrics are used to train the networks. The results are compared to models built using discriminant analysis.

Krishnamoorthy Srinivasan and Douglas Fisher. Machine Learning Approaches to Estimating Software Development Effort. *IEEE Trans. on Soft. Eng.*, 21(2), pp. 126-137, February 1995.

> This paper presents a study on the application of neural networks (using back propagation) and classification trees (using the CART algorithm [10]) to estimate software development effort. Both techniques are discussed in detail and compared to other machine learning techniques previously used for software development cost estimation.

## 7.3 Association Discovery

Inderpal S. Bhandari, M. J. Halliday, E. Tarver, D. Brown, J. Chaar, and R. Chillarence. A Case Study of Software Process Improvement During Development. *IEEE Trans. on Soft. Eng.*, 19(12), pp. 1157-1170, December 1993.

> This paper effectively introduced the attribute focusing technique to the software engineering community. It discusses the use of association discovery for exploring software defect data and process improvement. The authors discuss the results of their technique in a case study executed at IBM.

Inderpal S. Bhandari, M. J. Halliday, J. Chaar, R. Chillarence, K. Jones, J. S. Atkinson, C. Lepori-Costello, P. Y. Jasper, E. D. Tarver, C. C. Lewis, and M. Yonezawa. In-Process Improvement through Defect Data Interpretation. *IBM Syst. Journal*, 33(1), pp. 182-214, 1994.

> This paper is an extension of the previous one. It presents a description of the authors' work on attribute focusing that is more oriented to software engineering practitioners. It focuses on the data analysis methodology and lesson learned by the authors.

Manoel G. Mendonça, Victor R. Basili, Inderpal S. Bhandari, and Jack Dawson. An Approach to Improving Existing Measurement Frameworks. *IBM Syst. Journal*, 37(4), pp. 484-501, 1998.

> This paper discusses two approaches for improving existing measurement and data analysis in software organizations. The first approach works top-down based on goal-oriented measurement planning. The second approach works bottom-up by extracting new information from the legacy data already available in the organization. For the later approach, the authors use association discovery to gain new insights into the data that already is present in the organization.

## 7.4 Clustering

Uwe Krohn and Cornelia Boldyreff. Application of Cluster Algorithms for Batching of Proposed Software Changes. *J. Softw. Maint: Res. Pract.* 11, 151-165. May-June 1999.

> This paper presents an interesting application of cluster analysis to software maintenance planning. The authors apply hierarchical clustering to identify software changes that may be batched together. The technique uses a binary distance measure based on impact analysis of which modules will be affected by a proposed software change. Similar changes are batched based on the cluster of modules that they will affect.

Andy Podgurski, Wassim Masri, Yolanda McCleese, and Francis G. Wolff. Estimation of Software Reliability by Stratified Sampling. *ACM Trans. on Soft. Eng. and Methodology*, (8)3, pp. 263-283, July 1999.

> This paper presents a methodology to estimate operational software reliability by stratified sample of beta testers' code execution profiles. Cluster analysis is used to group code executions into dissimilar profiles. The authors show that more accurate estimates of failure frequencies can be drawn by stratified samples of those clustered execution profiles.

## 7.5 Optimized Set Reduction

Lionel C. Briand, Victor R. Basili, and William Thomas. A Pattern Recognition Approach for Software Engineering Data Analysis. *IEEE Trans. on Soft. Eng.*, 18(11), pp. 931-942, November 1992.

> This paper introduces optimized set reduction. It presents the application of OSR for software cost estimation. The authors uses a combination of data sets available in the literature to compare the accuracy of their model against models derived from COCOMO [9]**.** and built by stepwise regression. They also go over a detailed qualitative discussion on the advantages of their modeling technique.

Lionel C. Briand, Victor R. Basili, and Christopher J. Hetmanski. Developing Interpretable Models with Optimized Set Reduction for Identifying High-Risk Software Components. *IEEE Trans. on Soft. Eng.*, 19(11), pp. 1028-1044, November 1993.

> This paper presents the use of OSR to build models for classifying high risk software models. This paper has a better description of OSR than the previous one. This time the authors use data from 146 components of a large software system written in Ada. The model produced by OSR is compared to models produced by classification tree and logistic regression using a v-fold cross validation procedure.

## 7.6 Bayesian Belief Networks

Norman Fenton and Martin Neil. A Critique of Software Defect Prediction Models. To appear in the *IEEE Trans. on Soft. Eng.*, 1999. Available for download at: http://www.agena.co.uk/resources.html

> This paper presents the use of Bayesian Belief Networks (BBN) to build defect prediction models. These are the preliminary results of an interesting work. The paper has an wonderful discussion on the limitation of traditional defect prediction models. The authors argue that BBN models are interpretable and can include contextual software process information in them. This allows domain experts to analyze how defect introduction and detection variables affect the defect density counts in the model.

## 7.7 Visualization

Christof Ebert. Visualization Techniques for Analyzing and Evaluating Software Measures. IEEE *Trans. on Soft. Eng.*, 18(11), pp. 1029-1034. November 1992.

> This paper discusses the visualization of software engineering data but it was unfortunately published as a concise paper. The paper describes six approaches for displaying data and compares them on criteria such as clarity and comprehensibility.

Stephen G. Eick, Joseph L. Steffen, and Eric E. Summer, Jr. SeeSoft – A Tool for Visualizing Line-Oriented Software Statistics. *IEEE Trans. on Soft. Eng.*, (18)11, pp. 957-968. November 1992.

> This is an excellent paper on a seminal work on visualization of software source code. Developed at Bell Labs, SeeSoft is a visualization system that allows one to look at tens of thousands of lines of code simultaneously. Those lines of code are displayed as thin lines on the computer screen. Colors are used to visualize statistics on several attributes of interest in the code. Interactive controls allow users to drill up and down the source code display.

Thomas A. Ball and Stephen G. Eick. Software Visualization in the Large. *IEEE Computer*, (29)4, pp. 33-43, April 1996.

> This paper is a follow up of the previous one. It presents the authors' work on software visualization in a form that is more oriented to software engineering practitioners. Besides discussing visualization through line representation, the authors introduce pixel, file summary, and hierarchical representations for software visualization. The paper also describes five examples of real world applications of their visualization tool.

Stephen G. Eick, Audris Mockus, Tood L. Graves, Alan F. Karr. A Web Laboratory for Software Data Analysis. *World Wide Web*, 12, pp. 55-60, 1998.

> This paper describes how the authors' ideas on software visualization are being ported to a distributed system based on the World Wide Web. The system accesses data from central repositories enabling the users to visualize the most up to date data. The authors also argue that the system encourages collaborative research as observations and displays can be easily replicated and studied in detail by teams working geographically apart.

## 7.8 Others

Barbara Kitchenham. A Procedure for Analyzing Unbalanced Datasets. *IEEE Trans. on Soft. Eng.*, 24(4), pp. 278-301, April 1998.

> This paper uses statistical-based techniques, namely residual analysis and analysis of variance, to analyze unbalanced data sets. The proposed method is used to build a predictive model for software productivity and compared with the results obtained by CART (a classification tree algorithm) on the same data set. The method was especially developed for small data sets with many different attributes, a common scenario in software engineering. The paper is interesting because it proposes a data transformation technique that can itself be used to analyze software engineering data.

Taghi M. Khoshgoftaar, Matthew P. Evett, Edward B. Allen, and Pei-Der Chien. An Application of Genetic Programming to Software Quality Prediction. In *Computational Intelligence in Software Engineering*, Witold Pedrycz and James F. Peters editors. World Scientific Series *on Advances in Fuzzy Systems – Applications and Theory*, 16, pp. 176-195, 1998.

> This paper describes the use of genetic programming to build predictive models for software quality estimation. Genetic programming [26]. is a data mining technique not covered in this paper. This is one of the few papers we have found in the literature on this subject.

F. Lanubile and G. Visaggio, Evaluating Predictive Quality Models Derived from Software Measures: Lessons Learned," *The Journal of Systems and Software*, 38:225-234, 1997.

> This paper compares predictive quality models built using several different techniques. Among the techniques discussed are principal component analysis, discriminant analysis, logistic regression, classification trees, and artificial neural networks. Besides discussing the techniques and its validation, the paper has a good set of references on software engineering model building.

Magne Jørgensen. Experience with the Accuracy of Software Maintenance Task Effort Prediction Models. IEEE TSE, 21(8), pp. 674-681, August 1995.

> This paper reports on the development and use of several software maintenance effort prediction models. The models were developed applying regression analysis, neural networks, and OSR. The prediction accuracy of models was measured and compared to each other.

Lastly, The *Software Engineering and Knowledge Engineering Journal* (IJSEKE) is preparing a special issue on mining software engineering data. This issue, entitled "Knowledge Discovery from Empirical

Software Engineering Data" should be out by the end of 1999 or the beginning of 2000. It should make a good reading to complement the issues discussed in this report.

# 8. Concluding Remarks

Data mining has appeared as one of the tools of choice to better explore software engineering data. The constant increase on software and hardware infrastructures will only increase the availability of data in software organizations. The recent boom on data mining research will more than likely increase the number and quality of tools available for data analysis. One does not need to be a visionary to predict that during the next few years the use of data mining techniques will increase sharply among software engineering practitioners and data analysts.

The availability of new data analysis methods and tools should be met with caution by software engineering professionals. First of all, these techniques can only be as good as the data one collects. Having good data is the first requirement for good data exploration. There can be no knowledge discovery on bad data. This report does not discuss data collection improvement, but this issue should be on the top of the list for any organization planning to start a data mining program.

Good data is, however, just the first step. The second requirement for successful data mining is to understand what is being analyzed. One has to understand what is being sifted through a data mining technique in order to recognize (and use) it as "new knowledge." Data analysts and domain experts must understand the semantics of the data they propose to mine. If the data is being extracted from an external source, one must make sure that he knows what he is getting. One should also recognize the data limitations, treat missing values, and identify noisy information. Section 4 of this report discussed data selection and pre-processing. Those two activities are fundamental for successfully mining data. Unfortunately, many times they take the back seat for the more glamorous "mining" activities (see Figure 2.) Data analysts should avoid this trap. Data extraction and pre-processing may require a lot of effort but they are key to the success of any data mining endeavor.

Now, assuming that one has good data and has successfully extracted and preprocessed it. What next? Well, this will depend on the problem at hand and the task one wants to perform. Most of the time, the most attractive task to perform is to build a model that solves the problem at hand. Model building is indeed the most common application of data mining nowadays. In software engineering, reports on model building are almost as old as the discipline itself. It just happens that building good models is very hard. This is especially true in software engineering, a discipline that is very complex, human intensive, and many times abstract. It is our belief that model building should be the last step of the ladder. Model building tasks should be grounded on data exploration, see Figure 3. Software engineers should try to understand their object of analysis – be it a resource, a product, or a process – before trying to model it.

Data analysts should also consider that every data mining technique has strengths and weaknesses. The effectiveness of a technique is very dependent on the type of data at hand. Some techniques – such as neural networks – are well suited to analyzing numeric data, others – such as classification trees – are well suited to analyzing categorical data. A technique's effectiveness also depends on the number of data points available for analysis. OSR, for example, appears to work well with small data sets. This can be a clear advantage on some software engineering applications – such as software development effort prediction – on which only a few historical data points are available. Whenever possible, the data analyst should compare or even combine available techniques in order to obtain the best possible results.

Domain experts should strive to acquire new domain knowledge whenever possible. Domain knowledge is the basis for software process improvement. Techniques that produce interpretable models – like classification trees and bayesian belief networks – can be a valuable source of new domain knowledge. These techniques allow domain experts to use their background domain knowledge to interpret the models built by the automated algorithms. This activity is both a tool for sanity checking the produced models and a valuable source of new insights on the objects being modeled.

Lastly, one should know that statistics is alive and well. Statistical-based techniques are very useful and should not be discarded by any data analyst. Descriptive statistics can be used to characterize data and should stay in the front line of data exploration. Hypothesis testing should be used whenever there is doubt about the significance of a "discover." Techniques based on regression analysis are still heavily used in software engineering model building. In fact, there probably are far more papers published on statistical-based model building than on machine learning-based model building. Data analysts should always consider statistical-based techniques as tools that can significantly enhance the data mining process.

## 9. References


[1] D. Andenberg. *Cluster Analysis for Applications*. Academic Press, NY, 1973.

[2] T. A. Ball and S. G. Eick. Software Visualization in the Large. *IEEE Computer*, (29)4, pp. 33-43, April 1996.

[3] K. A. Bassin, T. Kratschmer, P. Santhanam. Evaluating Software Development Objectively. *IEEE Software*, 15(6), pages 66-74, Nov/December 1998.

[4] I. S. Bhandari, M. J. Halliday, E. Tarver, D. Brown, J. Chaar, and R. Chillarenge. A Case Study of Software Process Improvement During Development. *IEEE Trans. On Soft. Eng.*, 19(12), pages 1157-1170, December 1993.

[5] I. S. Bhandari, M. J. Halliday, J. Chaar, R. Chillarenge, K. Jones, J. S. Atkinson, C. Lepori-Costello, P. Y. Jasper, E. D. Tarver, C. C. Lewis, and M. Yonezawa. In process improvement through defect data interpretation. *IBM System Journal*, 33(1), pages 182-214, January 1994.

[6] I. S. Bhandari, B. Ray, M. Y. Wong, D. Choi, A. Watanabe, R. Chillarege, M. Halliday, A. Dooley, and J. Chaar. An Inference Structure for Process Feedback: Technique and Implementation. *Software Quality Journal*, 3(3), pages 167-189, September 1994.

[7] J. P. Bingus. *Data Mining With Neural Networks: Solving Business Problems – From Application Development to Decision Support*. McGraw-Hill, New York, 1996.

[8] B. W. Boehm. *Software Engineering Economics*. Prentice-Hall, Englewood Cliffs, NJ, 1981.

[9] B. W. Boehm. Software Engineering Economics. *IEEE Trans. On Soft. Eng.*, 10(1), pp. 4-21, January 1994.

[10] L. Breiman, J. Friedman, R. Olshen, and C. Stone. Classification and Regression Trees. Wadsworth Inc., Belmont, California, 1984.

[11] L. C. Briand, V. R. Basili, and C. J. Hetmanski. Developing Interpretable Models with Optimized Set Reduction for Identifying High-Risk Software Components. *IEEE Trans. on Soft. Eng.*, 19(11), pp. 1028-1044, November 1993.

[12] L. C. Briand, V. R. Basili, and W. Thomas. A Pattern Recognition Approach for Software Engineering Data Analysis. *IEEE Trans. on Soft. Eng.*, 18(11), pp. 931-942, November 1992.

[13] W. L. Buntine. Operations for Learning with Graphical Models. *Journal of Artificial Intelligence Research*, 2, pp. 159-225, 1994.

[14] R. Chillarenge, I. S. Bhandari, J. Chaar, M. J. Halliday, D. S. Moebus, B. K. Ray, and M. Wong. Orthogonal Defect Classification – A Concept for In-Process Measurements. *IEEE Trans. On Sofware Engineering*, 18(11), pages 943-956, November 1992.

[15] J. G. Carbonell and R. S. Michalski and T. M Mitchell. An overview of machine learning. In J. G. Carbonell and R. S. Michalski and T. M Mitchell, editors, *Machine Learning, an Artificial Intelligence Approach*, volume 1, pages 3-24. Morgan Kaufmann, San Mateo CA, 1983.

[16] G. F. Cooper and E. Herskovitz. A Bayesian Method for the Induction of Probabilistic Networks from Data. *Machine Learning*, 9, pp. 309-347, 1992.

[17] B. Efron. Estimating the Error Rate of a Prediction Rule: Improvement on Cross-Validation. Journal of the American Statistical Association, 78(382), pages 316-331, June 1983.

[18] S. G. Eick, A. Mockus, T. L. Graves, A. F. Karr. A Web Laboratory for Software Data Analysis. *World Wide Web*, 12, pp. 55-60, 1998.

[19] S. G. Eick, J. L. Steffen, and E. E. Summer, Jr. SeeSoft – A Tool for Visualizing Line-Oriented Software Statistics. *IEEE Trans. on Soft. Eng.*, (18)11, pp. 957-968. November 1992.

[20] U. Fayyad and G. Piatetsky-Shapiro and P. Smyth. The KDD process for extracting useful knowledge from volumes of data, *Communications of the ACM*, 39(11), pages 27-34, November 1996.

[21] U. Fayyad and R Uthurusamy. Data mining and knowledge discovery in databases, *Communications of the ACM*, 39(11), pages 24-26, November 1996.

[22] N. E. Fenton. Bayesian Belief Networks – An Overview Web Article. In WWW: http://www.agena.co.uk/bbn_article/bbns.html. Agena Ltd, 1999.

[23] N. E. Fenton. Software measurement: A necessary scientific basis. *IEEE Transactions on Software Engineering*, 20(3), March 1994.

[24] N. E. Fenton. *Software Metrics: A Rigorous Approach*. Chapman Hall, 1991.

[25] W. J. Frawley, G. Piatetsky-Shapiro, and C. J. Mattheus. Knowledge discovery in databases: An overview. *AI Magazine*, pages 57-70, Fall 1992.

[26] D. E. Goldberg. *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison-Wesley Publishing Company, 1989.

[27] G. E. Hinton. Connectionist Learning Procedures. In Kodratoff and Michalski [35], pages 555-610.

[28] J. H. Holland, K. J. Holyoak, R. E. Nisbett, and P. R. Thagard. *Induction: Processes of Inference, Learning, and Discovery*. MIT Press, Cambridge, MA, 1986.

[29] M. Holsheimer and A. P. J. Siebes. Data mining: the search of knowledge in databases. Technical Report CS-R9406, CWI – Department of Algorithms and Architecture, Amsterdam, The Netherlands, 1994.

[30] M. Jørgensen. Experience With the Accuracy of Software Maintenance Task Effort Prediction Models. IEEE TSE, 21(8), pp. 674-681, August 1995.

[31] T. M. Khoshgoftaar and E. B. Allen. Modeling Software Quality with Classification Trees. In *Recent Advances in Reliability and Quality Engineering*, Hoang Pham Editor. World Scientific, Singapore, 1999.

[32] T. M. Khoshgoftaar, E. B. Allen, J. P. Hudepohl, and S. J. Aud. Neural Networks for Software Quality Modeling of a Very Large Telecommunications System. *IEEE Trans. On Neural Networks*, (8)4, pp. 902-909, July, 1997.

[33] T. M. Khoshgoftaar and D. L. Lanning. A Neural Network Approach for Early Detection of Program Modules Having High Risk in the Maintenance Phase. *J. Systems Software*, 29(1), pp. 85-91, 1995.

[34] W. Klösgen and J. M. Zytkow. Knowledge discovery in databases terminology. In U. M. Fayyad, G. Piatetsky-Shapiro, P. Smith, and R. Uthurusamy, editors, *Advances in Knowledge Discovery and Data Mining*. AAAI Press/ The MIT Press, Cambridge, MA, 1996.

[35] Y. Kodratoff and R. S. Michalski, editors. *Machine Learning, an Artificial Intelligence Approach, Volume 3*. Morgan Kaufmann, San Mateo, California, 1990.

[36] U. Krohn and C. Boldyreff. Application of Cluster Algorithms for Batching of Proposed Software Changes. *J. Softw. Maint: Res. Pract.* 11, 151-165. May-June 1999.

[37] F. Lanubile and G. Visaggio, Evaluating predictive quality models derived from software measures lessons learned", *The Journal of Systems and Software*, 38:225-234, 1997.

[38] R. P. Lippmann. An Introduction to Computing with Neural Nets. *IEEE Acoustical, Speech, and Signal Processing Magazine*, 4, pp. 4-22, 1987. Reprinted in *Neural Networks: Theoretical Foundations and Analysis*, Edited by Clifford Lau, IEEE Press, 1992. Also reprinted in *Optical Neural Networks*, Edited by S. Jutamulia, SPIE Optical Engineering Press, 1994.

[39] H. J. Loether and D. G. McTavish. *Descriptive and Inferential Statistics: An Introduction. Part VI - Inferential Statistics*. Allyn and Bacon, Inc. Needham Heights, MA, 1988.

[40] J. B. MacQueen. Some Methods For Classification and Analysis of Multivariate Observations. In L. M. LeCam and J. Neyman, editors, *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistic and Probability*, pages 281-297, University of California Press, Berkley, CA, 1967.

[41] G. Miller. The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information. Psychological Review, 101(2), pp. 343-352, April 1994.

[42] G. W. Milligan. An Examination of the Effect of Six Types of Error Perturbation of Fifteen Clustering Algorithms. Psychometrika, 45(3), pp. 325-342, September 1980.

[43] M. Neil and N. E. Fenton. Predicting software quality using Bayesian belief networks. *Proc 21st Annual Software Eng Workshop*, NASA Goddard Space Flight Centre, pp. 217-230, Dec, 1996.

[44] M. Neil, B. Littlewood, and N. E. Fenton. Applying Bayesian belief networks to systems dependability assessment, *in Proceedings of 4th Safety Critical Systems Symposium*, Springer Verlag, pp. 71-93, 1996.

[45] A. Podgurski, W. Masri, Y. McCleese, and F. G. Wolff. Estimation of Software Reliability by Stratified Sampling. *ACM Trans. on Soft. Eng. and Methodology*, (8)3, pp. 263-283, July 1999.

[46] A. A. Porter and R. W. Selby. Empirically Guided Software Development Using Metric-Based Classification Trees. *IEEE Software*, 7(2), pp. 46-54, March 1990.

[47] A. A. Porter and R. W. Selby. Evaluating Techniques for Generating Metric-based Classification Trees. *J. Systems Software*, pp. 209-218, December 1990.

[48] J. R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufman, 1992.

[49] J. R. Quinlan. Induction of Decision Trees. *Machine Learning*, 1(1), pages 81-106, 1986.

[50] J. C. Schilimmer and P. Langley. Machine Learning. In S. C. Shapiro, editor, *Encyclopedia of Artificial Intelligence*, volume 1, pages 785-805. John Wiley & Sons, 1992.

[51] R. W. Selby and A. A. Porter. Learning from Examples: Generation and Evaluation of Decision Trees for Software Resource Analysis. *IEEE Trans. on Soft. Eng.*, 14(12), pp. 1743-1757, December 1988.

[52] K. Srinivasan and D. Fisher. Machine Learning Approaches to Estimating Software Development Effort. *IEEE Trans. On Soft. Eng.*, 21(2), pp. 126-137, February 1995.

[53] K. Swingler. *Applying Neural Networks: A Practical Guide*. Academic Press, London, 1996.

[54] J. Tian. Integrating Time Domain and Input Domain Analyses of Software Reliability Using Tree-Based Models. *IEEE Trans. on Soft. Eng.*, 21(12), pp. 945-958, December 1995.

[55] J. Tian and J. Palma. Analyzing and Improving Reliability: A Tree-based Approach. *IEEE Software*, pp. 97-104, 15(2), March-April 1998.

# A. Bibliography

## 1999


Bayardo Jr., R. J. and R. Agrawal, and D. Gunopulos. “Constraint-Based Rule Mining in Large, Dense Databases.” In Proceedings of the 15th International Conference on Data Engineering, 188-197, 1999.

Bayardo Jr., R. J. and R. Agrawal, “Mining the Most Interesting Rules.” In Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1999.

Berry, M. Mastering Data Mining. John Wiley & Sons, Incorporated; 1999. ISBN 0471331236.

Berson, Alex and Stephen J. Smith. Building Data Mining Applications. McGraw Hill Text; 1999. ISBN: 0071344446.

Bhandari, Inderpal S. and Edward Colet. Data Mining. Prentice-Hall Engineering/Science/ Mathematics,-1999. ISBN 0-13-083004-6.

Chapman, Hall, Ed. Data Mining and Reverse Engineering. Chapman & Hall; 1999. ISBN 0412822504.

Crestana, Viviane and Nandit Soparkar. “Mining Decentralized Data Repositories.” University of Michigan Department of Electrical Engineering and Computer Science , February 1, 1999. CSE-TR-385-99.

Data Mining. Prentice Hall; 1999. ISBN 0130830046.

Grath, R. Data Mining. Prentice Hall; 1999. ISBN 0130862711.

Han, J. “Characteristic Rules,” to appear in W. Kloesgen and J. Zytkow (Eds.), Handbook of Data Mining and Knowledge Discovery, Oxford University Press, 1999.

Han, J. “Data Mining.” in J. Urban and P. Dasgupta (Eds.), Encyclopedia of Distributed Computing, Kluwer Academic Publishers, 1999.

Han, Jiawei and Micheline Kamber. Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers, 1999. ISBN 1558604898.

Kambayashi, Y., Ed. Advances in Database Technologies: Workshops on Data Warehousing and Data Mining, Mobile Data Access, and Collaborative Work Support, International Conference on Conceptual Modeling. Springer Verlag; 1999. ISBN 3540656901.

Koperski, Krzysztof. “Progressive Refinement Approach to Spatial Data Mining,” Ph.D. Thesis, Computing Science, Simon Fraser University, April 1999.

Lamb, Ceferino. "Business Mining for Decision Support Insights." Business Objects, SA, 1999. http://www.businessobjects.com/decision/library/index.htm

McHugh, Linda. "Data Mining." Teradata Review; Summer 1999.
http://www.teradatareview.com/summer99/mchugh.html

PMML 1.0 — Predictive Model Markup Language. Data Mining Group. 1999.
http://www.dmg.org/public/techreports/pmml-1_0.html

Pyle, Dorian. Data Preparation for Data Mining. Morgan Kaufmann Publishers, 1999.
ISBN 1558605290.
http://www.amazon.com/exec/obidos/ISBN%3D1558605290/thedataminersA/002-8122587-3880614

Reingruber, Michael C. The Data Mining Handbook: A Best Practice Approach to Building Quality Data Models, 2nd Edition. John Wiley & Sons, Incorporated; 1999. ISBN 047135452X.
http://shop.barnesandnoble.com/booksearch/isbnInquiry.asp?userid=3LH0TVJMJB&mscssid=X9HK6MJRAUSH2NL0001PQUN3XR3QE0BB&sourceid=00008330780237151532&bfdate=08%2D17%2D1999+14%3A23%3A00&pcount=0&srefer=&isbn=047135452X

Tung, A. K. H., H. Lu, J. Han, and L. Feng. "Breaking the Barrier of Transactions: Mining Inter-Transaction Association Rules." In Proceedings of the 1999 International Conference on Knowledge Discovery and Data Mining (KDD'99), San Diego, California, August 1999.
ftp://ftp.fas.sfu.ca/pub/cs/han/kdd/kdd99.ps

Williams, Graham J. "The Terminology Rescue Kit Knowledge Discovery in Databases and Data Mining." 1999.
http://www.cmis.csiro.au/Graham.Williams/DataMiner/Dictionary.html

Witten, J. Tools for Data Mining. Morgan Kaufmann Publishers; 1999. ISBN 1558605525.
http://shop.barnesandnoble.com/booksearch/isbnInquiry.asp?userid=3LH0TVJMJB&mscssid=X9HK6MJRAUSH2NL0001PQUN3XR3QE0BB&sourceid=00008330780237151532&bfdate=08%2D17%2D1999+14%3A23%3A00&pcount=0&srefer=&isbn=1558605525

Proceedings Pacific-Asia Conference on Knowledge Discovery and Data Mining 1999.
Springer-Verlag New York, Incorporated; 1999. ISBN 3540658661.
http://shop.barnesandnoble.com/booksearch/isbnInquiry.asp?userid=3LH0TVJMJB&mscssid=X9HK6MJRAUSH2NL0001PQUN3XR3QE0BB&sourceid=00008330780237151532&bfdate=08%2D17%2D1999+14%3A23%3A00&pcount=0&srefer=&isbn=3540658661

Zhou, X., D. Truffet, and J. Han. "Efficient Polygon Amalgamation Methods for Spatial OLAP and Spatial Data Mining." In Proceedings of the 6th International Symposium on Large Spatial Databases (SSD'99), Hong Kong, July 1999.

# 1998


Agrawal, Rakesh and Paul Stolorz, Eds. Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD-98). AAAI Press, 1998. ISBN 1-57735-070-7.

Agrawal, Rakesh , Johannes Gehrke, Dimitrios Gunopulos, and Prabhakar Raghavan. "Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications." In Proceedings of the ACM SIGMOD International Conference on Management of Data, Seattle, Washington, June 1998.

Anand, Sarabot S. and Alex G. Buchner. Decision Support Using Data Mining. Pitman Publishing; 1998. ISBN 0273632698.

Bayardo Jr., R. J. "Efficiently Mining Long Patterns from Databases." In Proceedings of the ACM SIGMOD Conference on Management of Data, Seattle, Washington, June 1998.
http://www.almaden.ibm.com/cs/quest/papers/sigmod98_max.pdf

Cios, Krzysztof J., Witold Pedrycz, and Roman Swiniarski. Data Mining Methods for Knowledge Discovery. Kluwer Academic Publishers; 1998. ISBN 0792382528.
http://shop.barnesandnoble.com/booksearch/isbnInquiry.asp?userid=3LH0TVJMJB&mscssid=X9HK6MJRAUSH2NL0001PQUN3XR3QE0BB&sourceid=00008330780237151532&bfdate=08%2D17%2D1999+14%3A23%3A00&pcount=0&srefer=&isbn=0792382528

Ebecken, N.F. , Ed. Data Mining. Computational Mechanics, Incorporated; 1998. ISBN 1853126772.
http://shop.barnesandnoble.com/booksearch/isbnInquiry.asp?userid=3LH0TVJMJB&mscssid=X9HK6MJRAUSH2NL0001PQUN3XR3QE0BB&sourceid=00008330780237151532&bfdate=08%2D17%2D1999+14%3A23%3A00&pcount=0&srefer=&isbn=3540643834

Edelstein, Herbert A. Introduction to Data Mining and Knowledge Discovery. Two Crows Corporation, 1998. ISBN 1892095009.
http://www.amazon.com/exec/obidos/ISBN%3D1892095009/thedataminersA/002-8122587-3880614

Elder, John F. and Dean W. Abbott. "A Comparison of Leading Data Mining Tools," In Fourth Annual Conference on Knowledge Discovery & Data Mining, New York, NY, August 28, 1998.
http://www.datamininglab.com/pubs/kdd98_elder_abbott_nopics.pdf

Information Discovery, Inc. "A Characterization of Data Mining Technologies and Processes." The Journal of Data Warehousing, January 1998.
http://www.datamining.com/dm-tech.htm

Liu, Huan and Hiroshi Motoda. Feature Selection for Knowledge Discovery and Data Mining. Kluwer International Series in Engineering and Computer Science, 1998. ISBN 79238198X.
http://www.amazon.com/exec/obidos/ISBN%3D079238198X/002-8122587-3880614

Megiddo, N. and R. Srikant, “Discovering Predictive Association Rules.” In Proceedings of the 4th International Conference on Knowledge Discovery in Databases and Data Mining, New York, August 1998.

Michalski , Ryszard, Ivan Bratko, and Avan Bratko, Eds. Machine Learning and Data Mining: Methods and Applications. John Wiley & Sons, Incorporated; 1998. ISBN 0471971995.

Research and Development in Knowledge Discovery and Data Mining: Second Pacific-Asia Conference, Melbourne, Australia, April 15-17, 1998. Springer Verlag, 1998. ISBN 3540643834.

Sarawagi, S., S. Thomas, and R. Agrawal. “Integrating Association Rule Mining with Databases: Alternatives and Implications.” In Proceedings of the ACM SIGMOD International Conference on Management of Data, Seattle, Washington, June 1998.
http://www.almaden.ibm.com/cs/quest/papers/sigmod98_dbi.pdf

Thuraisingham, Bhavani M. Data Mining: Technologies, Techniques, Tools, and Trends. CRC Press, 1998. ISBN 0849318157.
http://www.amazon.com/exec/obidos/ISBN%3D0849318157/thedataminersA/002-8122587-3880614

Westphal, Chris and Teresa Blaxton. Data Mining Solutions: Methods and Tools for Solving Real World Problems. John Wiley & Sons, 1998. ISBN 0471253847.
http://www.amazon.com/exec/obidos/ISBN%3D0471253847/thedataminersA/002-8122587-3880614

Williams, Graham J. “High Performance Data Management Issues in Data Mining.” In Workshop on Parallel and Distributed Data Mining, Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD-98), April 1998.
http://www.cmis.csiro.au/Graham.Williams/papers/pddm98.html

Williams, Graham J. “The Data Miner’s Arcade: Pluggable Data Mining.” Technical report, CSIRO Mathematical and Information Sciences, 1998.
http://www.cmis.csiro.au/Graham.Williams/dataminer/Arcade.html

Zaki, M. J. , C. T. Ho, and R. Agrawal. “Parallel Classification for Data Mining on Shared-Memory Multiprocessors.” IBM Research Report, 1998
http://www.almaden.ibm.com/cs/quest/papers/sprint_smp.pdf

Zaki, M. J. “Scalable Data Mining for Rules.” University of Rochester; Computer Science Technical Report. July 1998.
http://cstr.cs.cornell.edu:80/Dienst/UI/1.0/Display/ncstrl.rochester_cs/TR702

# 1997


Ali, K., S. Manganaris, and R. Srikant. "Partial Classification using Association Rules." In Proceedings of the 3rd International Conference on Knowledge Discovery in Databases and Data Mining, Newport Beach, California, August 1997.
http://www.almaden.ibm.com/cs/quest/papers/kdd97_class.pdf

Berry, Michael J. and Gordon Linoff. Data Mining Techniques for Marketing, Sales, and Customer Support. John Wiley. 1997. ISBN 0-471-17980-9.

Berson, Alex and Stephen J. Smith. Data Warehousing, Data Mining, & OLAP. McGraw-Hill, 1997. ISBN 0-07-006262-2.

Brooks, Peter. "Visualizing Data — Sophisticated Graphic Visualization and Development Tools Tailored for Business Applications." DBMS Magazine, August 1997.
http://www.dbmsmag.com/9708d13.html

Cabena, Peter, Pablo Hadjinian, Rolf Stadler, Jaap Verhees, and Alessandro Zanasi. Discovering Data Mining from Concept to Implementation. Prentice Hall, 1997. ISBN 0137439806.
http://www.amazon.com/exec/obidos/ASIN/0137439806/qid=928955544/sr=1-25/002-8122587-3880614

Cabena, Peter, Pablo Hadjinian, Rolf Stadler, Jaap Verhees, and Alessandro Zanasi. Discovering Data Mining, IBM Red Book SG24-4839-00. ISBN 0738404187.
http://www.redbooks.ibm.com/

Chen, M. S., J. Han, and P.S. Yu. "Data Mining: an Overview from Database Perspective." *IEEE Transactions on Knowledge and Data Engineering*, 1997.
ftp://ftp.fas.sfu.ca/pub/cs/han/kdd/survey97.ps

Darling, Charles B. "Datamining for the Masses." Datamation Plugin, February 1997.
http://www.datamation.com/PlugIn/issues/1997/feb/02mine.html

Dhar, Vasant and Roger Stein. Seven Methods for Transforming Corporate Data into Business Intelligence. Prentice Hall Computer Books; 1997. ISBN 0132820064.
http://www.amazon.com/exec/obidos/ISBN%3D0132820064/thedataminersA/002-8122587-3880614

Edelstein, Herb. "Mining for Gold — A Raft of Data Mining Tools Offers a Wide Range of Features for Digging up Business Opportunities." Information Week; April 21, 1997.
http://www.techweb.com/se/directlink.cgi?IWK19970421S0046

Ferguson, Mike. "Evaluating and Selecting Data Mining Tools." InfoDB; November 1997.
http://www.dbaint.com/pdf/v11n21.pdf

Fong, J., Ed. Data Mining, Data Warehousing and Client-Server Databases: Proceedings of the 8th International Hong Kong Computer Society Database Workshop. Springer-Verlag New York, Incorporated; 1997. ISBN 9813083549.
http://shop.barnesandnoble.com/booksearch/isbnInquiry.asp?userid=649A0VH565&mscssid=CX6465JL3WSH2JS500JP424CRB4X48K2&sourceid=00008330780130724843&bfdate=06%2D09%2D1999+16%3A28%3A15&pcount=0&srefer=&isbn=9813083549

Freitas, Alex A. and Simon H. Lavington. Mining Very Large Databases with Parallel Processing. Kluwer Academic Publishers; 1997. ISBN 0792380487.
http://shop.barnesandnoble.com/booksearch/isbnInquiry.asp?userid=3LH0TVJMJB&mscssid=X9HK6MJRAUSH2NL0001PQUN3XR3QE0BB&sourceid=00008330780237151532&bfdate=08%2D17%2D1999+14%3A23%3A00&pcount=0&srefer=&isbn=0792380487

Groth, Robert. Data Mining: A Hands-On Approach for Business Professionals. Prentice Hall PTR (ECS Professional), 1997. ISBN 0-13-756412-0.
http://www.prenhall.com/books/ptr_0137564120.html

Heckerman, David, Heikki Mannila, Daryl Pregibon, and Ramasamy Uthurusamy, Eds. Proceedings of the Third International Conference on Knowledge Discovery and Data Mining (KDD-97). AAAI Press; 1997. ISBN 1-57735-027-8.
http://www.aaai.org/Press/Proceedings/KDD/1997/

Keim, Daniel. Visual Database Exploration Techniques. University of Munich.
http://www.dbs.informatik.uni-muenchen.de/~daniel/KDD97.pdf

Kennedy, Ruby L., Ed., Yuchun Lee, Benjamin Van Roy, Christopher Reed, and the Staff of Unica Technology, Inc. Solving Data Mining Problems through Pattern Recognition. Data Warehousing Institute Series from Prentice Hall, 1997. ISBN 0130950831.
http://www.amazon.com/exec/obidos/ASIN/0130950831/qid=928955544/sr=1-38/002-8122587-3880614

Komorowski, Jan and Jan Zytkow, Eds. Principles of Data Mining and Knowledge Discovery: First European Symposium, Pkdd '97 Trondheim, Norway, June 24-27, 1997: Proceedings (Lecture Notes). Springer Verlag, 1997. ISBN 3540632239.
http://www.amazon.com/exec/obidos/ISBN%3D3540632239/thedataminersA/002-8122587-3880614

Lin, T. Y., Ed. and N. Cercone. Rough Sets and Data Mining: Analysis for Imprecise Data. Kluwer International, 1997. ISBN 0792398076.
http://www.amazon.com/exec/obidos/ISBN%3D0792398076/thedataminersA/002-8122587-3880614

Lu, H-J J., H. J. Lu, H. Motoda, Ed. Knowledge Discovery and Data Mining: Techniques and Applications. World Scientific Publishing Company, Incorporated; 1997. ISBN 9810230729.
http://shop.barnesandnoble.com/booksearch/isbnInquiry.asp?userid=3LH0TVJMJB&mscssid=X9HK6MJRAUSH2NL0001PQUN3XR3QE0BB&sourceid=00008330780237151532&bfdate=08%2D17%2D1999+14%3A23%3A00&pcount=0&srefer=&isbn=9810230729

Moxon, Bruce. "Data Mining: The Golden Promise." OReview, June 1997.
http://www.oreview.com/9706moxn.htm

Roth, Pam, Ed. Data Mining: Data Warehousing Tools for the New Generation. Spiral Books; 1997. ISBN 1571090169.
http://shop.barnesandnoble.com/booksearch/isbnInquiry.asp?userid=3LH0TVJMJB&mscssid=X9HK6MJRAUSH2NL0001PQUN3XR3QE0BB&sourceid=00008330780237151532&bfdate=08%2D17%2D1999+14%3A23%3A00&pcount=0&srefer=&isbn=1571090169

Siu, Brian, Paul K.M. Kwan, Benedict Lam, and Peter de Vries, Eds. Data Mining, Data Warehousing & Client/Server Databases. Springer-Verlag New York, Incorporated; 1997. ISBN 9813083530.
http://shop.barnesandnoble.com/booksearch/isbnInquiry.asp?userid=3LH0TVJMJB&mscssid=X9HK6MJRAUSH2NL0001PQUN3XR3QE0BB&sourceid=00008330780237151532&bfdate=08%2D17%2D1999+14%3A23%3A00&pcount=0&srefer=&isbn=9813083530

Small, Robert D. "Debunking Data Mining Myths." *Information Week*; January 20, 1997.
http://www.techweb.com/se/directlink.cgi?IWK19970120S0042

Solving Data Mining Problems through Pattern Recognition. Unica Technology, Inc.; Prentice Hall PTR (ECS Professional), 1997. ISBN 0-13-095083-1.
http://www.prenhall.com/allbooks/ptr_0130950831.html

Srikant, R., Q. Vu, and R. Agrawal, "Mining Association Rules with Item Constraints." In Proceedings of the 3rd International Conference on Knowledge Discovery in Databases and Data Mining, Newport Beach, California, August 1997.
http://www.almaden.ibm.com/cs/quest/papers/kdd97_const.pdf

Stolorz, Paul and Ron Musick, Eds. Scalable High Performance Computing for Knowledge Discovery and Data Mining. Kluwer Academic, 1997. ISBN 0792380975
http://www.amazon.com/exec/obidos/ASIN/0792380975/qid=928957629/sr=1-88/002-8122587-3880614

Weiss, Sholom M. and Nitin Indurkhya. Predictive Data Mining: A Practical Guide. Morgan Kaufman Publishers, 1997. ISBN 1558604030.
http://www.amazon.com/exec/obidos/ISBN%3D1558604030/thedataminersA/002-8122587-3880614

Williams, Graham J. and Zhexue Huang. "Mining the Knowledge Mine: The Hot Spots Methodology for Mining Large, Real World Databases." In Abdul Sattar, Ed., Advanced Topics in Artificial Intelligence, volume 1342 of Lecture Notes in Computer Science, pages 340-348. Springer-Verlag, December 1997. http://www.cmis.csiro.au/Graham.Williams/papers/ai97.html

## 1996

Adriaans, Pieter and Dolf Zantinge. Data Mining. Addison-Wesley. 1996. ISBN 0201403803.
http://www.amazon.com/exec/obidos/ISBN%3D0201403803/thedataminersA/002-8122587-3880614

Agrawal, R. and J.C. Shafer. "Parallel Mining of Association Rules." *IEEE Transactions on Knowledge and Data Engineering*, Vol. 8, No. 6, December 1996.

Association for Computing Machinery; Special Issue of the Communications of the ACM; "Data Mining and Knowledge Discovery;" November, 1996, Vol. 39, number 11. Guest Editors: Usama Fayyad and Ramasamy Uthurusamy.

Bigus, Joseph P. Data Mining with Neural Networks: Solving Business Problems from Application Development to Decision Support. McGraw Hill Text, 1996. ISBN 0070057796

Edelstein, Herb. "Technology How-To — Mining Data Warehouses." *Information Week*; January 8, 1996.

English, Larry P. "Help for Data Quality Problems — A Number of Automated Tools can Ease Data Cleansing and Help Improve Data Quality." *Information Week*; October 7,1996.

Fayyad, Usama M., Gregory Piatetsky-Shapiro, Padhraic Smyth, and Ramasamy Uthurusamy (Eds.). Advances in Knowledge Discovery and Data Mining. MIT Press, 1996. ISBN 0262560976.

Ganesh, M., Eui-Hong (Sam) Han, Vipin Kumar, Sashi Shekhar, and Jaideep Srivastava. "Visual Data Mining." University of Minnesota Computer Science Technical Report, TR 96-021. 1996.

Simoudis, Evangelos, Jiawei Han, and Usama Fayyad, Eds. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96). AAAI Press, 1996. ISBN 1-57735-004-9.

Srikant, R. and R. Agrawal. "Mining Quantitative Association Rules in Large Relational Tables." In Proceedings of the ACM SIGMOD Conference on Management of Data, Montreal, Canada, June 1996.
http://www.almaden.ibm.com/cs/quest/papers/sigmod96.pdf

## 1995

Dilly, Ruth. "Data Mining — An Introduction." The Queen's University of Belfast, Parallel Computer Centre, 1995.
http://wwwpcc.qub.ac.uk/tec/courses/datamining/stu_notes/dm_book_1.html

Fayyad, M. Usama and Ramasamy Uthurusamy, Ed. Proceedings of the Third International Conference on Knowledge Discovery and Data Mining. AAAI Press, 1995. ISBN 0929280822
Fayyad, Usama. "Tutorial MA1 — Knowledge Discovery and Data Mining." Tutorial Notes KDD (Knowledge Discovery in Databases) Data Mining. 1995.
http://shop.barnesandnoble.com/booksearch/isbnInquiry.asp?userid=3LH0TVJMJB&mscssid=X9HK6MJRAUSH2NL0001PQUN3XR3QE0BB&sourceid=0000833078023715153 2&bfdate=08%2D17%2D1999+14%3A23%3A00&pcount=0&srefer=&isbn=0929280822

Srikant, R. and R. Agrawal. "Mining Generalized Association Rules." In Proceedings of the 21st International Conference on Very Large Databases, Zurich, Switzerland, September 1995.

# B. Data Mining Tool Descriptions

This appendix contains descriptions of a collection of data mining and knowledge discovery tools. It summarizes information to allow a reader to browse for an appropriate tool description based upon tool name, tool vendor, the level of data mining process, mining technique, algorithms, and supplemental tool information (the availability of case studies, tutorials, white papers, newsletters, user groups, demonstration copies of the tools, and evaluation copies of the tool), and keywords.

Inclusion in this report does not imply endorsement of the tool by the DoD

Data & Analysis Center for Software (DACS). The effectiveness of each tool is not evaluated in this report.

- Table B–1 summarizes the level of support for the data mining process for each tool.
- Table B–2 summarizes the mining techniques employed by each tool.
- Table B–3 summarizes the data mining algorithms used in given tools.
- Table B–4 summarizes the supplemental information available for each tool.

**Table B–1 summarizes the level of support for the data mining process for each tool.**

# Level of Data mining Process

| | Basic Data Exploration | Discovery of Data Patterns | Support Model building | Perform Data Cleanup | Perform Data Transformation | Perform Model Evaluations |
|---|---|---|---|---|---|---|
| **AC2** | | | | | | |
| **Aira Data Mining Tool** | X | X | | | | |
| **ALICE d'ISoft** | X | | X | X | X | X |
| **Alterian Nucleus** | | | | | | |
| **AnswerTree** | | | | | | |
| **AT Sigma Data Chopper** | | | | | | |
| **AVS/Express Visualization Edition** | | | | | | |
| **Blue Data Miner** | X | | | | | |
| **Broadbase EPM (Enter Perf Mgmt)** | | | | | | |
| **BusinessMiner** | | | | | | |
| **Caprl** | | | | | | |
| **CART** | | X | X | X | X | X |
| **Clementine** | X | X | X | X | X | X |
| **CrossGraphs** | | | | | | |
| **Cubist** | | | X | | | |
| **Darwin** | X | X | X | X | X | X |
| **Data Detective** | X | X | X | | | |
| **Data mining Suite** | X | X | X | X | X | X |
| **Data SURVEYOR** | | | | | | |
| **Data Logic/RDS** | | X | X | X | X | X |
| **DataMiner 3D** | X | X | | X | X | X |
| **DataMite** | X | X | | X | X | X |
| **DataScope** | X | | | | | |
| **dbProbe** | | | | | | |
| **DecisionWORKS** | | | | | | |
| **DeltaMiner** | X | X | | X | X | X |
| **IBM Intelligent Miner for Data** | X | X | X | X | X | X |
| **IDL (Interactive Data Language)** | X | X | | X | X | X |
| **Informix Red Brick Formation** | | | | | | |
| **KATE-DataMining** | X | X | X | X | X | X |

| | Basic Data Exploration | Discovery of Data Patterns | Support Model building | Perform Data Cleanup | Perform Data Transformation | Perform Model Evaluations |
|---|---|---|---|---|---|---|
| **KnowledgeSEEKER** | | | | | | |
| **KnowledgeSTUDIO** | X | X | X | X | X | X |
| **MARS** | | X | X | X | X | X |
| **MineSet** | | | | | | |
| **MODEL 1** | | | | | | |
| **Nested Vision3D** | | | | | | |
| **Nuggets** | | X | X | X | X | X |
| **OMNIDEX** | | | | | | |
| **Open Visualization Data Explorer** | | | | | | |
| **PolyAnalyst** | X | X | X | X | X | X |
| **PrudSys Discoverer** | X | | X | X | X | X |
| **S-PLUS** | X | X | X | X | X | X |
| **SAS Enterprise Miner** | | | | | | |
| **Scenario** | | | | | | |
| **See 5/C5.0** | | | X | | | |
| **sphinxVision** | X | X | X | X | X | X |
| **Spotfire Pro** | | | | | | |
| **SRA KDD Explorer** | X | X | X | X | X | X |
| **Syllogic Data Mining** | | | | | | |
| **SyNERA** | | | | | | |
| **The Easy Reasoner (TER)** | X | | X | | | |
| **Viscovery SOMine** | X | X | X | X | X | X |
| **Visual Insights ADVIZOR** | | | | | | |
| **VisualMine** | X | X | X | | | |
| **XpertRule Miner** | X | X | X | X | X | X |

**Table B–2 Data Mining Techniques**

# Mining Techniques

| | Supervised Induction | Association Discovery | Sequence Discovery | Clustering | Visualization | Other |
|---|---|---|---|---|---|---|
| **AC2** | | | | | | |
| **Aira Data Mining Too** | | | | | X | |
| **ALICE d'ISoft** | | | | X | | |
| **Alterian Nucleus** | | | | | | |
| **AnswerTree** | | | | | | |
| **AT Sigma Data Chopper** | | | | | | |
| **AVS/Express Visualization Edition** | | | | | | |
| **Blue Data Miner & Blue Open** | X | | | | | |
| **Broadbase EPM (Ent Perf Mgmt)** | | | | | | |
| **BusinessMiner** | | | | | | |
| **Caprl** | | | | | | |
| **CART (Class & Regress Trees)** | X | | | | | |
| **Clementine** | X | | | | | |
| **CrossGraphs** | | | | | | |
| **Cubist** | X | | | | | |
| **Darwin** | X | | | | | |
| **Data Detective** | X | | | | | |
| **Data Mining Suite** | X | | | | | |
| **Data SURVEYOR** | | X | | | | |
| **Data Logic/RDS** | | | | | | |
| **DataMiner 3D** | X | | | | | |
| **DataMite** | X | | | | | |
| **DataScope** | | | | X | | |
| **dbProbe** | | | | | X | |
| **DecisionWORKS** | | | | X | | |
| **DeltaMiner** | | X | | | | |
| **IBM Intelligent Miner for Data** | X | | | | | |
| **IDL (Interactive Data Language)** | | | | | X | |
| **Informix Red Brick Formation** | | | | | | |
| **KATE-DataMining** | X | | | | | |

| | Supervised Induction | Association Discovery | Sequence Discovery | Clustering | Visualization | Other |
|---|---|---|---|---|---|---|
| **KnowledgeSEEKER** | | | | | X | |
| **KnowledgeSTUDIO** | X | | | | | |
| **MARS** | X | | | | | |
| **MineSet** | | X | | | | |
| **MODEL 1** | | | | | | |
| **Nested Vision3D** | | | | | X | |
| **Nuggets** | X | | | | | |
| **OMNIDEX** | | | | | | |
| **Open Visualization Data Explorer** | | | | | | |
| **PolyAnalyst** | X | | | | | |
| **PrudSys Discoverer** | X | | | | | |
| **S-PLUS** | | X | | | | |
| **SAS Enterprise Miner** | | | | X | | |
| **Scenario** | | X | | | | |
| **See 5/C5.0** | X | | | | | |
| **sphinxVision** | | | | | X | |
| **Spotfire Pro** | | | | | | |
| **SRA KDD Explorer** | X | | | | | |
| **Syllogic Data Mining** | | | | X | | |
| **Synera** | | | | | | |
| **The Easy Reasoner** | X | | | | | |
| **Viscovery SOMine** | | | | | | X |
| **Visual Insights ADVIZOR** | | | | | | |
| **VisualMine** | | | | | X | |
| **XpertRule Miner** | X | | | | | |

**Table B–3 Data Mining Algorithms**

# Algorithms

| | Decision Trees | Genetic | K-Nearest Neighbor | Learning Clustering | Neural Networks | Rule-based |
|---|---|---|---|---|---|---|
| **AC2** | X | | | | | |
| **Aira Data Mining Too** | | | | | | X |
| **ALICE d'ISoft** | X | | | X | | |
| **Alterian Nucleus** | | | | | | |
| **AnswerTree** | X | | | | | |
| **AT Sigma Data Chopper** | | | | | | |
| **AVS/Express Visualization Edition** | | | | | | |
| **Blue Data Miner & Blue Open** | | | | | | |
| **Broadbase EPM (Ent Perf Mgmt)** | | | | | | |
| **BusinessMiner** | X | | | | | |
| **Caprl** | | | | | | |
| **CART (Class & Regress Trees)** | X | | | | X | |
| **Clementine** | X | | | X | X | X |
| **CrossGraphs** | | | | | | |
| **Cubist** | | | | | | X |
| **Darwin** | X | X | X | X | X | X |
| **Data Detective** | | | X | X | | |
| **Data Mining Suite** | | | | | | X |
| **Data SURVEYOR** | X | | | | | X |
| **Data Logic/RDS** | | | | | | X |
| **DataMiner 3D** | | | X | | | |
| **DataMite** | | | X | X | | X |
| **DataScope** | | | | | | |
| **dbProbe** | | | | | | |
| **DecisionWORKS** | | X | | X | X | |
| **DeltaMiner** | | | | | | |
| **IBM Intelligent Miner for Data** | X | | | | X | X |
| **IDL (Interactive Data Language)** | | | | | X | |
| **Informix Red Brick Formation** | | | | | | |
| **KATE-DataMining** | X | | | X | | |

| | Decision Trees | Genetic | K-Nearest Neighbor | Learning Clustering | Neural Networks | Rule-based |
|---|---|---|---|---|---|---|
| **KnowledgeSEEKER** | X | | | | | X |
| **KnowledgeSTUDIO** | X | | | X | X | |
| **MARS** | | | | | | |
| **MineSet** | X | | | X | | |
| **MODEL 1** | | | | | | |
| **Nested Vision3D** | | | | | | |
| **Nuggets** | | | | X | | X |
| **OMNIDEX** | | | | | | |
| **Open Visualization Data Explorer** | | | | | | |
| **PolyAnalyst** | | X | X | X | X | X |
| **PrudSys Discoverer** | X | | | | | |
| **S-PLUS** | X | | | X | | |
| **SAS Enterprise Miner** | X | | | X | X | |
| **Scenario** | X | | | | | |
| **See 5/C5.0** | X | | | X | | X |
| **sphinxVision** | | | | | | |
| **Spotfire Pro** | | | | | | |
| **SRA KDD Explorer** | | | | | | |
| **Syllogic Data Mining** | X | | | | | X |
| **Synera** | | | | | | |
| **The Easy Reasoner** | | | X | | | |
| **Viscovery SOMine** | | | | | | |
| **Visual Insights ADVIZOR** | | | | | | |
| **VisualMine** | | | | X | | X |
| **XpertRule Miner** | X | | | X | | X |

Table B–4 Supplemental Tool Information

# Supplemental Tool Information

| | Case Studies | Tutorial | White Papers | Newsletter | User Group | Demo | Evaluation Copy |
|---|---|---|---|---|---|---|---|
| AC2 | | | | | | X | |
| Aira Data Mining Too | | | | | | X | |
| ALICE d'ISoft | | | | | | | X |
| Alterian Nucleus | X | | | | | | |
| AnswerTree | X | | X | | | | |
| AT Sigma Data Chopper | | | | | | X | |
| AVS/Express Visualization Edition | X | | X | X | | | |
| Blue Data Miner & Blue Open | | | X | | | X | X |
| Broadbase EPM (Ent Perf Mgmt) | X | | | | | X | |
| BusinessMiner | | X | X | | | | |
| Caprl | | | | | | | |
| CART (Class & Regress Trees) | X | X | X | | | X | X |
| Clementine | X | | | | X | | X |
| CrossGraphs | | X | | X | | X | |
| Cubist | X | X | | | | X | X |
| Darwin | X | | | | | X | |
| Data Detective | X | X | X | | X | X | X |
| Data Mining Suite | | | X | | | | |
| Data SURVEYOR | X | | | | X | | |
| Data Logic/RDS | | | X | | | | |
| DataMiner 3D | | | | | | | X |
| DataMite | | | | | | | |
| DataScope | | | | | | X | |
| dbProbe | | | | | | X | X |
| DecisionWORKS | | | | | | X | |
| DeltaMiner | | | | | | | |
| IBM Intelligent Miner for Data | X | | X | | | X | |
| IDL (Interactive Data Language) | X | | X | | | X | |
| Informix Red Brick Formation | | | | | | | |
| KATE-DataMining | | | | | | X | |

| | Case Studies | Tutorial | White Papers | Newsletter | User Group | Demo | Evaluation Copy |
|---|---|---|---|---|---|---|---|
| KnowledgeSEEKER | X | | X | | | X | X |
| KnowledgeSTUDIO | | | X | | | X | X |
| MARS | | X | X | | | X | X |
| MineSet | | X | X | | | X | X |
| MODEL 1 | | | | | | | |
| Nested Vision3D | | | | | | | |
| Nuggets | | | X | | | | |
| OMNIDEX | | | X | | | | |
| Open Visualization Data Explorer | | | | | | | |
| PolyAnalyst | X | X | X | | X | X | X |
| PrudSys Discoverer | | | X | | | | |
| S-PLUS | X | X | X | | X | X | |
| SAS Enterprise Miner | | | X | | | | |
| Scenario | X | | | | | | |
| See 5/C5.0 | X | X | | | | X | X |
| sphinxVision | | | X | | | X | X |
| Spotfire Pro | | | X | | | X | |
| SRA KDD Explorer | | | | | | | |
| Syllogic Data Mining | | | | | | | |
| Synera | | | X | | | | X |
| The Easy Reasoner | | | | | | | X |
| Viscovery SOMine | | X | X | | | X | X |
| Visual Insights ADVIZOR | | | | | | X | |
| VisualMine | X | | | | | | |
| XpertRule Miner | X | | X | | | | X |

## AC2

http://www.alice-soft.com/products/ac2.html

ISoft
Chemin de Moulon
F-91190 Gif-sur-Yvette France

Raphaelle Thomas
+33 1 6935 3737
+33 1 6935 3739 (FAX)
info@isoft.fr
http://www.alice-soft.com/products/index.html

Keywords
Programmable; Toolkit; Algorithm — Decision Tree; C++; GUI (Graphical User Interface); Knowledge Discovery; C; Decision Support; Demo

Platforms
Windows 95; Windows NT; Unix

Description
AC2 is a comprehensive data mining toolkit, aimed at expert users or developers who want to use Data Mining functionalities with their own interfaces. AC2 is primarily a set of libraries that developers can use to build data mining solutions on the server side.

Levels of Exploration

• Basic Data Exploration through Visualization — Not Known

• Provides Discovery of Data Patterns — Not Known

• Provides Support for Model Building — Not Known

Levels of Data Mining Process
Not Known

Data Mining Techniques
Not Known

• Functionality to Clean up Data — Not Known

• Tool Executes Data Transformations — Not Known

• Tool Handles Null/Missing Data Values — Not Known

• Tool Performs Model Evaluation — Not Known

## AIRA Data Mining Tool

http://www.hycones.com.br/

Hycones IT
Av. Victor Barreto, 2288 - 3o andar
CEP 92010-000
Canoas-RS Brazil

Rodrigo Leal
+55 51 4728108 r: 222; +55 51 9825124
+55 51 4728108 r: 222 (FAX)
hycones@conex.com.br
http://www.hycones.com.br/

Keywords
Algorithm — Rule-Based; Demo; GUI (Graphical User Interface); Knowledge Discovery; Decision Support; Data Warehouse

Platforms
Windows 95; Windows NT

Description
AIRA Data Mining Tool is a tool able to: discover new, useful and interesting knowledge from databases; generate rules; summarize information; detect irregular behavior in the database; and represent the discovered knowledge in different ways, making it easier to understand.

Levels of Exploration

• Basic Data Exploration through Visualization — Yes
The 'circle view,' which is a circular histogram implemented as a complementary tool in AIRA Data Mining, may be included in this category.

• Provides Discovery of Data Patterns — Yes
The core of AIRA was implemented around a hybrid architecture for associative rule discovery and classification. This 'core' is covered by 'layers' of knowledge visualization tools. These tools are complementary and may be more or less usable accordingly to the task at hand, as it has been verified in our applications. In terms of interestingness, AIRA discovers all rules that match user criteria, and user may choose to filter these rules by the presence of certain attribute(s).

In the rule browser, the user interaction, choosing classes, and then attributes, and values, is also a way for the user to express his 'interest.' This view as been useful to explore databases with large number of attributes. AIRA Data Mining provides a set of complementary tools that could fit in different 'classes'. AIRA was built around a rule discovery algorithm, but this was complemented by several auxiliary tools, which will be briefly described.

AIRA is goal-oriented, as the user is responsible for defining which attribute is the 'target' (which indicates a class, for example). All rules that match user criteria should be discovered. These rules can be viewed, browsed, and explored through these tools:

Rule list: a simple textual list of discovered rules, which can be set to be presented in natural language format, may include exceptions, and can be filtered according to the presence certain attribute(s), by each target-attribute value, as well as by minimum support and confidence levels.

Rule browser: A hierarchical structure that allows user to browse into a certain class, viewing only the attributes that actually appear in rules for that class, choose any of these attributes to see the relevant values, and going on until rules are formed. This view allows user to query the data that support each rule, the exceptions, the class distribution for each rule, etc.

HTML: Rules according to user settings can be exported to HTML format and published into the intranet. AIRA includes different templates that can be used, one based on frames, and a frame-less version.

Knowledge Graph: A graph structure can represent discovered rules, with classes on the left (or top) of the screen, and findings (cause-attribute(s) values) on the right (or the bottom). The links in the middle represent rules. This view allows the user to check if certain findings lead to different classes according to the combination with other findings, which cannot be easily viewed in a rule list. This view also allows to zoom into each rule, viewing additional statistics on each finding (how specific a finding is to a class, and how general it is). Besides all these alternatives to represent rules, AIRA also induces frame-like structures, and includes a circular histogram (circle view).

Frame View: For each class, this view provides the most specific findings (the triggers that would lead to some conclusion), and the most sensible (most supported) findings, which would be like 'essential findings' for a certain diagnostic.

Circle View: values of attributes chosen by the user are presented as slices of a circle. The association of different findings is represented by lines linking these slices, with different color, according to the strength of the association.

• Provides Support for Model Building — Not Known

<u>Levels of Data Mining Process</u>
Extracting Relationships and Data Associations

<u>Data Mining Techniques</u>
Visualization

• Functionality to Clean up Data — No
• Tool Executes Data Transformations — No
• Tool Handles Null/Missing Data Values — Yes
• Tool Performs Model Evaluation — No

Future Enhancements
AIRA for Excel 2000; Aira Client-Server

## ALICE d'ISoft

http://www.alice-soft.com/products/alicev50.html

ISoft
Chemin du Moulon
F-91190 Gif-sur-Yvette France

Herve Perdrix
+33 1 6935 3737
+33 1 6935 3739 (FAX)
info@isoft.fr
http://www.alice-soft.com/products/index.html

Keywords
OLAP (On-Line Analytical Processing); Algorithm — Decision Tree; Visual Data Mining; Algorithm — Learning/Clustering; GUI (Graphical User Interface); Pattern Discovery; SQL; Evaluation Copy; Predictive Modeling

Platforms
Windows 95; Windows 98; Windows NT

Description
ALICE d'ISoft is designed for the nontechnical business user. It explores databases through interactive decision trees and creates queries, reports, charts, and rules for predictive models. ALICE d'ISoft gives business users access to the knowledge hidden in their databases, discovering the trends and relationships in their data and making predictions using that information. ALICE d'ISoft covers the whole analysis, from raw data to the model deployment, with its large range including: ALICE d'ISoft V6.0 with its data preparation module and OLAP functionality, ALICE/Stat (descriptive statistics plugin), ALICE/Crosstab (dynamic cross tables), ALICE/Correlation (dynamic correlation tables), ALICE/Report (reporting plugin), ALICE/Scoring (scoring plugin) and ALICE/Segment (segmentation plugin). ALICE d'ISoft V6.0 integrates an OLAP engine and advanced dynamic data aggregation functions. ALICE d'ISoft is also available on a server version: ALICE/SERVER.

Levels of Exploration

• Basic Data Exploration through Visualization — Yes
ALICE d'ISoft provides graphics online; data visualization (histogram, plot, density curve, pie chart, etc.), and 3D visualization.

ALICE d'ISoft achieves exploration for each subgroup of data, during any phase of the analysis. It then provides an high degree of interactivity.

• Provides Discovery of Data Patterns — No

• Provides Support for Model Building — Yes
ALICE d'ISoft provides interactive decision tree and regression tree and includes five different algorithms.

Levels of Data Mining Process
Basic Data Exploration; Extracting Relationships and Data Associations; Model Building

Data Mining Techniques
Clustering; Visualization; Supervised Induction

• Functionality to Clean up Data — Yes
• Tool Executes Data Transformations — Yes
• Tool Handles Null/Missing Data Values — Yes
• Tool Performs Model Evaluation — Yes

White Papers and Resources
"Data Mining for Business Users"; http://www.byte.com/art/9611/sec18/art13.htm

## Alterian Nucleus

http://www.alterian.com/nucleus.htm

Alterian Limited
Century Place
Newfoundland Street
Bristol, BS2 9AG England

+44 (0) 117 970 3200
+44 (0) 117 970 3201 (FAX)
info@alterian.com
http://www.alterian.com

Keywords
ActiveX; Case Studies; Intranet; Scalable; SQL; WWW Interface; Data Visualization

Platforms
Windows 95; Windows 98; Windows NT

Description
Alterian Nucleus is an analysis database that is optimized for analysis, it processes tens of millions of records per second on a standard PC platform. In common with a relational database management system, Nucleus stores the actual value of every field for every record, but it achieves speeds of analysis usually associated with an On Line Analytical Processing Engine (OLAP).

Levels of Exploration

• Basic Data Exploration through Visualization — Not Known

• Provides Discovery of Data Patterns — Not Known

• Provides Support for Model Building — Not Known

Levels of Data Mining Process
Not Known

Data Mining Techniques
Not Known

• Functionality to Clean up Data — Not Known
• Tool Executes Data Transformations — Not Known
• Tool Handles Null/Missing Data Values — Not Known
• Tool Performs Model Evaluation — Not Known

## AT Sigma Data Chopper

http://www.atsigma.com/datamining/index.htm

Advanced Technologies
2601 Oswell Street
Suite 206
Bakersfield  CA  93306

(661) 872-4807
(661) 872-3316 (FAX)
info@atsigma.com

Keywords
Demo; Knowledge Discovery; ODBC–Compliant

Platforms
Not Known

Description
AT Sigma Data Chopper will scan through mountains of data to find significant relationships between variables in a database and display the results in tables and graphs.

Levels of Exploration

• Basic Data Exploration through Visualization — Not Known

• Provides Discovery of Data Patterns — Not Known

• Provides Support for Model Building — Not Known

Levels of Data Mining Process
Extracting Relationships and Data Associations

Data Mining Techniques

Not Known

• Functionality to Clean up Data — Not Known
• Tool Executes Data Transformations — Not Known
• Tool Handles Null/Missing Data Values — Not Known
• Tool Performs Model Evaluation — Not Known

## AVS/Express Visualization Edition

http://www.avs.com/products/ExpVis/ExpVis.htm

Advanced Visual Systems, Inc.
300 Fifth Avenue
Waltham MA 02451

(800) 728-1600
(781) 890-4300
(781) 890-8287 (FAX)
http://www.avs.com/

Keywords
C; C++; Case Studies; Data Visualization; GUI (Graphical User Interface); OLAP (On-Line Analytical Processing); Visual Programming; White Papers; Newsletter

Platforms
IBM (AIX); Hewlett-Packard (HP-UX); Silicon Graphics (IRIX); Unix (Digital Unix); Unix (Solaris); Windows 95; Windows NT; Unix (SunOS)

Description
The AVS/Express Visualization Edition offers scientists, researchers, and other technical professionals a suite of data visualization and analysis capabilities. It provides end users with state-of-the-art technology for advanced graphics, imaging, data visualization, and presentation. AVS/Express Visualization Edition's visual programming environment makes it easier for users to quickly and interactively visualize their data.

Levels of Exploration

• Basic Data Exploration through Visualization — Not Known

• Provides Discovery of Data Patterns — Not Known

• Provides Support for Model Building — Not Known

Levels of Data Mining Process
Not Known

Data Mining Techniques
Not Known

• Functionality to Clean up Data — Not Known
• Tool Executes Data Transformations — Not Known
• Tool Handles Null/Missing Data Values — Not Known
• Tool Performs Model Evaluation — Not Known

## Blue Data Miner and Blue Open

http://www.bluedatainc.com/bdm.htm

Blue Data, Inc.
5889 Oberlin Drive
Suite 105
San Diego CA 92121

G. Graham Murray
(619) 558-3338
(619) 558-3341 (FAX)
contact@bluedatainc.com; GGMURRAY2@aol.com
http://www.bluedatainc.com/

Keywords
Data Warehouse; Decision Support; Demo; RDBMS (Relational Database Management Systems); SQL; White Papers; WWW Interface; C++; Data Mart; EIS (Executive Information System); Evaluation Copy; Flat Files; OLAP (On-Line Analytical Processing); Visual Data Mining

Platforms
Windows NT

Description
Blue Open converts any table from a database into a data mart and permits data mining it. Blue Data Miner has the additional feature that it can extract database tables from reports. These are PC tools. Enterprise Reporting and Analysis (ERA) is the Enterprise version running on a Windows NT Server. Each product allows executives, managers, and other key personnel to filter and summarize those tables. With ERA this is done over the Internet or intranet. All systems use only a standard web browser. Blue Open and Blue Data Miner function as stand-alone systems. ERA is a 3-tier system using a Windows NT Server.

Levels of Exploration

• Basic Data Exploration through Visualization — Yes
The Blue Data products offer various color graphical outputs, which are easily modified to show other views.

Blue Data Miner has as a goal being the simplest possible data mining tool, and so the user can within minutes rearrange data and view it in a variety of color graphic displays.

• Provides Discovery of Data Patterns — No

• Provides Support for Model Building — No

Levels of Data Mining Process
Basic Data Exploration; Extracting Relationships and Data Associations

Data Mining Techniques
Supervised Induction; Visualization

Blue Data Miner can extract a table from a report (using an Extractor software routine). This table can then be data mined.

- Functionality to Clean up Data — No
- Tool Executes Data Transformations — No
- Tool Handles Null/Missing Data Values — No
- Tool Performs Model Evaluation — No

White Papers and Resources
http://www.bluedatainc.com/khome.htm

## Broadbase EPM (Enterprise Performance Management)

http://www.broadbase.com/products/broadbaseepm.asp

Broadbase
http://www.broadbase.com/contact/
http://www.broadbase.com/

Keywords
EIS (Executive Information System); WWW Interface; Decision Support; Demo; Case Studies; Object Oriented User Interface

Platforms
Not Known

Description
Broadbase EPM allows a user to: measure what is happening; track performance against key indicators; analyze performance; identify process bottlenecks, critical success factors and key trends to recognize present — and future — opportunities and exposures; and continuously improve results by identifying adjustments, prioritizing resources, and accelerating decision-making.

Levels of Exploration

• Basic Data Exploration through Visualization — Not Known

• Provides Discovery of Data Patterns — Not Known

• Provides Support for Model Building — Not Known

Levels of Data Mining Process
Not Known

Data Mining Techniques
Not Known

• Functionality to Clean up Data — Not Known
• Tool Executes Data Transformations — Not Known
• Tool Handles Null/Missing Data Values — Not Known
• Tool Performs Model Evaluation — Not Known

Future Enhancements
Soon to be released Broadbase EPM application modules include: E-Commerce, Field Service, Finance, Logistics, Call Center, Human Resources, and Balanced Scorecard.

## BusinessMiner

http://www.businessobjects.com/products/enduser_tools/businessminer/index.htm

Business Objects Americas
Jacqueline Biggio
(703) 708-9661
jbiggio@dlt.com; webleads@businessobjects.com
http://www.businessobjects.com/

Keywords
Data Warehouse; Algorithm — Decision Tree; Algorithm — Rule-Based; Tutorial; White Papers; Decision Support; OLAP (On-Line Analytical Processing); Pattern Discovery; RDBMS (Relational Database Management Systems); Glossary

Platforms
Not Known

Description
BusinessMiner provides a data mining solution designed for mainstream business users. Based on intuitive decision tree technology, BusinessMiner is a desktop data mining tool that lets nontechnical business users find useful information hidden in their business data. BusinessMiner is a data mining tool whose automatic analysis lets a user find useful information hidden in business data, whether in relational databases, data warehouses, OLAP servers, or personal files. BusinessMiner discovers trends hidden in data, then displays them for analysis in the form of a decision tree.

Levels of Exploration

• Basic Data Exploration through Visualization — Not Known

• Provides Discovery of Data Patterns — Not Known

• Provides Support for Model Building — Not Known

Levels of Data Mining Process
Not Known

Data Mining Techniques
Not Known

• Functionality to Clean up Data — Not Known
• Tool Executes Data Transformations — Not Known
• Tool Handles Null/Missing Data Values — Not Known
• Tool Performs Model Evaluation — Not Known

White Papers and Resources
http://www.businessobjects.com/decision/library/index.htm

## Caprl

http://www.mineit.com/

MINEIT Software Limited
info@mineit.com
http://www.mineit.com/

Keywords
Knowledge Discovery

Platforms
Not Known

Description
Caprl is a data mining algorithm that discovers different types of sequences in databases from within the Clementine environment. Caprl allows the specification of domain knowledge, such as start and end pages, as well as various time-related constraints. Caprl can handle numeric and symbolic data input values, and is able to produce three different types of sequences.

Levels of Exploration

• Basic Data Exploration through Visualization — Not Known

• Provides Discovery of Data Patterns — Not Known

• Provides Support for Model Building — Not Known

Levels of Data Mining Process
Not Known

Data Mining Techniques
Not Known

• Functionality to Clean up Data — Not Known
• Tool Executes Data Transformations — Not Known
• Tool Handles Null/Missing Data Values — Not Known
• Tool Performs Model Evaluation — Not Known

## CART

http://www.salford-systems.com/products.html

Salford Systems
8880 Rio San Diego Drive
Suite 1045
San Diego  CA  92108

Kerry Martin
(619) 543-8880
(619) 543-8888 (FAX)
mkerry@salford-systems.com
http://www.salford-systems.com/

Keywords
Algorithm — Decision Tree; Algorithm — Neural Networks; Demo; GUI (Graphical User Interface); Pattern Discovery; Predictive Modeling; White Papers; Algorithm — Regression; Scalable; Case Studies; Data Visualization; Decision Support; Evaluation Copy; Knowledge Discovery; ODBC–Compliant; Tutorial

Platforms
Unix; Windows 95; Windows NT; Windows 3.x; MacOS; IBM VMS & CMS

Description
CART (Classification and Regression Trees) is a widely used decision tree classification and regression system from the statistical community with many applications to data mining, predictive modeling, and data preprocessing.  CART is a robust, easy-to-use decision tree tool that automatically sifts large, complex databases, searching for and isolating significant patterns and relationships.  Based on over a decade of research, CART ensures stable performance and reliable results. CART's easy to use GUI, intelligent default settings, and interactive Tree Navigator empower nontechnical users to develop a highly intuitive understanding of their data — to tell a story about what's driving the results and why.  For power users, CART's unique advanced features coupled with its batch production mode deliver versatility, speed, and accuracy.

Levels of Exploration

• Basic Data Exploration through Visualization — No

• Provides Discovery of Data Patterns — Yes
CART decision-tree software automatically discovers cause-and-effect relationships, reveals significant patterns, and forecasts trends.  Discovered relationships and patterns in the data — even in massively complex data sets with hundreds of variables — are presented as a tree-shaped diagram, much like a flow chart.  The visual display enables users to see the hierarchical interaction of the variables; and it often confirms previous knowledge about

important data relationships, which adds confidence in the reliability and utility of the CART model. Further, because simple if-then rules can be read right off the tree, models are easy to grasp and easy to apply to new data.

CART is a state-of-the-art classification tool that, as a standalone package, can investigate any classification task and provide a robust, accurate predictive model. The software tackles the core data-mining challenges by accommodating classification — for categorical variables, such as responder and non-responder — and regression for continuous variables, such as sales revenue.

CART searches for questions that split nodes into relatively homogenous child nodes, such as a group consisting largely of responders, or high credit risks, or people who bought sport-utility vehicles. As the tree evolves, the nodes become increasingly more homogenous, identifying important segments. The set of predictor variables used by CART to split the nodes into segments — read directly off the tree and summarized in the variable importance tables — are thus the key drivers of the target variable.

CART is also an excellent preprocessing complement to data-mining packages. In the first stage of a data-mining project, CART can extract the most important variables from a very large list of potential predictors. Focusing on the top variables from the CART model can significantly speed up neural networks and other data-mining techniques. For neural nets in particular, CART bypasses 'noise' and irrelevant variables, quickly and effectively selecting the best variables for input. The result is significant reductions in neural-net training speeds and more accurate and robust neural networks. In addition, the CART outputs, or predicted values, can be used as inputs to the neural net.

The analytical engine driving CART is the only true and complete implementation of the original code by CART's developers, Breiman, Friedman, Olshen and Stone. Their landmark work created the modern field of sophisticated, mathematically- and theoretically-founded decision trees. The CART methodology solves a number of performance, accuracy, and operational problems that still plague many current decision-tree methods. CART's innovations include:

• solving the 'how big to grow the tree' problem;
• using strictly two-way (binary) splitting;
• incorporating automatic testing and tree validation, and
• providing a completely new method for handling missing values.

• Provides Support for Model Building — Yes

CART is a classification and regression tool based on a decade of research assuring stable performance and reliable results. CART's proven methodology is characterized by: Reliable pruning strategy — CART's developers determined definitively that no stopping rule could be relied on to discover the optimal tree, so they introduced the notion of overgrowing trees and then pruning back; this idea, fundamental to CART, ensures that important structure is not overlooked by stopping too soon.

Other decision-tree techniques use problematic stopping rules. Powerful binary-split search approach — CART's binary decision trees are more sparing with data and detect more structure before too little data is left for learning. Other decision-tree approaches use multi-way splits that fragment the data rapidly, making it difficult to detect rules that require broad ranges of data to discover. Automatic self-validation procedures — in the search for patterns in databases it is essential to avoid the trap of 'over fitting,' or finding patterns that apply only to the training data. CART's embedded test disciplines ensure that the patterns found will hold up when applied to new data. Further, the testing and selection of the optimal tree are an integral part of the CART algorithm. Testing in other decision-tree techniques is conducted after-the-fact and tree selection is left up to the user. In addition, CART accommodates many different types of real-world modeling problems by providing a unique combination of automated solutions:

• surrogate splitters intelligently handle missing values;
• adjustable misclassification penalties help avoid the most costly errors;
• multiple-tree, committee-of-expert methods increase the precision of results, and
• alternative splitting criteria make progress when other criteria fail.

Levels of Data Mining Process

Basic Data Exploration; Extracting Relationships and Data Associations; Model Building

Data Mining Techniques

Supervised Induction

Tree-based clustering and association discovery are under development.

• Functionality to Clean up Data — Yes
• Tool Executes Data Transformations — Yes
• Tool Handles Null/Missing Data Values — Yes
• Tool Performs Model Evaluation — Yes

Future Enhancements

Among the new features to be included in the next version of CART are: handling of character variables; sampling weights; Entropy/Chi-Squared Splitting Rules; and enhanced model reporting functionality.

White Papers and Resources

http://www.salford-systems.com/whitepaper.html

## Clementine

http://www.spss.com/software/clementine

SPSS Inc.
233 South Wacker Drive
11th Floor
Chicago IL 60606–6307

Peter Caron
(800) 543-2185, (312) 651-3000
(312) 651-3444 (FAX)
pcaron@spss.com
http://www.spss.com

Keywords
Case Studies; Oracle; Ingres; Sybase; Informix; Data Visualization; EIS (Executive Information System); Visual Programming; Evaluation Copy; User Group; CRISP-DM (CRoss-Industry Standard Process for Data Mining); Algorithm — Neural Networks; Algorithm — Rule-Based; Visual Data Mining; Algorithm — Decision Tree;
Algorithm — Learning/Clustering; Object Oriented User Interface; ODBC–Compliant; Pattern Discovery; Predictive Modeling

Platforms
Windows NT; Hewlett-Packard Workstations; SunSPARC; IBM RS6000; Unix (NCR); Silicon Graphics Workstations; Unix (Digital Unix); Data General AViiON

Description
Clementine is a rapid modeling environment that combines enterprise-strength data mining and business knowledge to discover solutions that a user otherwise would not. The visual interface makes data mining an interactive process that invites a user's business expertise at every step in the data mining process, from data access and preparation to models and results.

Levels of Exploration

• Basic Data Exploration through Visualization — Yes
Clementine is a rapid modeling workbench that represents steps in the data mining process as nodes for a visual representation through graphs.

Clementine includes graphs such as Web diagrams, plots, distributions, histograms and other graphical data representations. Graphs in Clementine are interactive. For example, spokes of a Web diagram, showing associations between variables, can be hidden to highlight more interesting relationships. In addition, individual plots or bars on a graph can be selected to derive a subset of data that can then be used for further visualization or analysis.

• Provides Discovery of Data Patterns — Yes
Clementine provides algorithms for discovering clusters, associations, and sequential patterns. Outliers can be detected and handled by specifying limits on values, using visualization to manually tag or eliminate outliers, using balance nodes to boost rare or attenuate swamping events, or let learning algorithms screen them out.
For example, Clementine's Web diagram can automatically detect the strength of associations between variables and display the strength with weighted lines.
Clustering can be performed with Kohonen Network or K-Means Clustering. Association discovery can be performed with Apriori and GRI algorithms. Web Visualization also shows associations. Sequential patterns discovery can be performed with Apriori and GRI algorithms.

• Provides Support for Model Building — Yes
Classification can be performed by Backprop Multi-Layer Perception (MLP) and Radial Basis Function (RBF) neural networks, and by C5.0 trees and rule induction. Regression can be performed by MLP and RBF neural networks, Build Rule (extended ID3), and Linear Regression. Time Series Forecasting can be performed by the neural networks and rule induction.

Levels of Data Mining Process
Basic Data Exploration; Model Building

Data Mining Techniques
Supervised Induction; Association Discovery; Sequence Discovery; Clustering; Visualization; Regression and Time Series

• Functionality to Clean up Data — Yes
• Tool Executes Data Transformations — Yes
• Tool Handles Null/Missing Data Values — Yes
• Tool Performs Model Evaluation — Yes

Future Enhancements
Clementine Solution Publisher, distributed architecture version, deployment vehicles, and updated versions.

## CrossGraphs

http://www.belmont.com/cg.html

PPD Informatics Belmont Research
84 Sherman Street
Cambridge MA 02140

Jeff Millstein
(617) 868-6878
(617) 868-2654 (FAX)
cginfo@belmont.com; jeff.millstein@belmont.ppdi.com
http://www.belmont.com/

Keywords
Data Visualization; Demo; ODBC–Compliant; OLE; Oracle; SAS Analytical Tool; Programmable

Platforms
Hewlett-Packard (HP-UX); Unix (Solaris); Unix (SunOS); Windows 95; Windows 98; Windows NT; Power MacOS

Description
CrossGraphs combines statistical graphics with cross-tabulation, without programming. Use CrossGraphs to organize data into arrays of graphs to highlight trends and show relationships among many variables. Produce reports with one graph or thousands with the same simple interface. Drill-down on graphs to see and export underlying data. Run CrossGraphs in batch mode for production reporting. Import data from ASCII, ODBC, SAS, dBase, and Oracle. Relate multiple files for enhanced data exploration.

Levels of Exploration

• Basic Data Exploration through Visualization — Not Known

• Provides Discovery of Data Patterns — Not Known

• Provides Support for Model Building — Not Known

Levels of Data Mining Process
Not Known

Data Mining Techniques
Not Known

• Functionality to Clean up Data — Not Known
• Tool Executes Data Transformations — Not Known
• Tool Handles Null/Missing Data Values — Not Known
• Tool Performs Model Evaluation — Not Known

## Cubist

http://www.rulequest.com/cubist-info.html

RuleQuest Research Pty Ltd.
30 Athena Avenue
St Ives NSW 2075 Australia

Ross Quinlan
+61 2 9449 6020
+61 2 9440 9272 (FAX)
quinlan@rulequest.com
http://www.rulequest.com/

Keywords
Tutorial; C; Demo; Evaluation Copy; Algorithm — Rule-Based; Decision Support; Pattern Discovery; Programmable; Case Studies; Flat Files; GUI (Graphical User Interface); Knowledge Discovery; Pattern Discovery; Predictive Modeling

Platforms
Windows 95; Windows 98; Windows NT; Unix (Solaris); Unix (Irix); Unix (Digital Unix); Unix (Linux)

Description
Cubist produces rule-based models for numerical prediction. Each rule specifies the conditions under which an associated multivariate linear sub-model should be used. The result — powerful piecewise linear models. Cubist builds rule-based predictive models that output values. Cubist can effectively process datasets containing tens of thousands of records and hundreds of fields (attributes). Public C source code is provided so that models developed by Cubist can be displayed in other applications.

Levels of Exploration

• Basic Data Exploration through Visualization — No

• Provides Discovery of Data Patterns — No

• Provides Support for Model Building — Yes
Cubist models a user-selected dependent variable as a function of the other variables. A Cubist model consists of a set of rules of the form if (conditions) then (linear model) where the conditions are a conjunction of Boolean expressions on the independent variables. The rules are not necessarily mutually exclusive — some data may be covered by multiple rules, in which case the predicted values provided by the rules are averaged.

Levels of Data Mining Process
Model Building

Data Mining Techniques
Supervised Induction

- Functionality to Clean up Data — No
- Tool Executes Data Transformations — Yes
- Tool Handles Null/Missing Data Values — Yes
- Tool Performs Model Evaluation — Yes

Future Enhancements
Import of data via ODBC; network licensing

## Darwin

http://www.think.com/html/products/dartechdatasht.htm#one

Thinking Machines Corporation
16 New England Executive Park
Burlington MA 01803

Nikki Houck
(877) 677-1400
(781) 238-3400
(781) 238-3440 (FAX)
nhouck@appliedcom.com
http://www.think.com

Keywords
ActiveX; Algorithm — Decision Tree; Algorithm — Neural Networks; C; C++; Demo; Algorithm — Genetic; GUI (Graphical User Interface); Algorithm — Regression; Algorithm — Rule-Based; Java; ODBC–Compliant; Oracle; Programmable; RDBMS (Relational Database Management Systems); SAS Analytical Tool; Case Studies; Algorithm — Learning/Clustering; Algorithm — K-Nearest Neighbor; Evaluation Copy

Platforms
Unix (Solaris); Windows NT; Hewlett-Packard (HP-UX); Windows 95; Windows 98

Description
Darwin is scalable, enterprise data mining software that helps organizations to rapidly transform large amounts of data into actionable business intelligence. Darwin helps to find meaningful patterns and correlations in corporate data for understanding and prediction of customer behavior.

Levels of Exploration

• Basic Data Exploration through Visualization — Yes
In addition to viewing summary results in MicroSoft Excel, users may view interactive model results including support for mouse 'fly over' that displays the value where the mouse is pointing.

• Provides Discovery of Data Patterns — Yes
Darwin release 3.6 provides support for clustering.

• Provides Support for Model Building — Yes
Darwin supports neural networks, CART, and memory-based reasoning using the K-Nearest Neighbor technique.

Levels of Data Mining Process
Basic Data Exploration; Extracting Relationships and Data Associations; Model Building

Data Mining Techniques
Supervised Induction; Sequence Discovery; Clustering; Visualization; Other
K-Nearest Neighbor, CART, Naieve Bayes. A Model Seeker wizard automatically runs multiple models and recommends the best model. A Key Fields wizard sifts through the data and identifies the most relevant fields.99

- Functionality to Clean up Data — Yes
- Tool Executes Data Transformations — Yes
- Tool Handles Null/Missing Data Values — Yes
- Tool Performs Model Evaluation — Yes

Future Enhancements
Future versions of Darwin will include k-means clustering and ODBC write-back to the database, as well as integration with Pilot software. They will include the addition of the Windows NT Server platform, native database access, and the ability to easily score a database, faster algorithms using 'threads' for better parallelism, a Naive Bayes algorithm, and enhanced clustering using self-organizing maps (SOM). A new family of packaged applications powered by Darwin, initially for Response Modeling, Churn, Cross-Selling, Customer Profiling, and Profitability, will be offered.

White Papers and Resources
"Why Mine Data? An Executive Guide"
http://www.think.com/html/products/execguide.htm

"Scalable Data Mining"
http://www.think.com/html/products/darwin/scal_dat.htm

"Big Data — Better Returns: Leveraging Your Hidden Data Assets to Improve ROI"
http://www.think.com/html/products/darwin/r_roi.htm

## Data Detective

http://www.smr.nl/

Sentient Machine Research B.V.
Baarsjesweg 224
1058 AA Amsterdam The Netherlands

Peter van der Putten
+31 20 6186927
+31 20 6124504 (FAX)
info@smr.nl
http://www.smr.nl/

Keywords
Fuzzy Logic; GUI (Graphical User Interface); Algorithm — Learning/Clustering; Algorithm — K-Nearest Neighbor; Case Studies; Demo; Evaluation Copy; Tutorial; Tutorial; White Papers; Wizards; Data Visualization; CRISP-DM (CRoss-Industry Standard Process for Data Mining); ODBC–Compliant; Predictive Modeling; Programmable; RDBMS (Relational Database Management Systems); WWW Interface; ActiveX; C

Platforms
Windows 95; Windows 98; Windows NT

Description
DataDetective data mining tool has a modular design consisting of an associative analysis engine, a graphical user interface and interfaces to common database formats. Several analysis tasks are supported such as normal queries, fuzzy queries (selections defined by soft criteria instead of hard criteria), profile analyses and extensive graphical report facilities. The user always has direct access to the relevant data on which analyses are based.

Levels of Exploration

• Basic Data Exploration through Visualization — Yes
Functionality: descriptives, series, frequencies, histograms, scatter plots, surface plots, 3D scatter plots. Formats: In 2D — pie, polar, bubble, scatter, line, bar, area, hi-lo, candlestick, box-whisker, and time series. In 3D — pie, bar, tape, area, scatter, and surface.

Functionality: descriptives, series, frequencies, histograms, scatter plots, surface plots, 3D scatter plots. Formats: In 2D — pie, polar, bubble, scatter, line, bar, area, hi-lo, candlestick, box-whisker, and time series. In 3D — pie, bar, tape, area, scatter, and surface. Fully Interactive: a user can change any part of a diagram, remove series and points simply by clicking on it, etc.

• Provides Discovery of Data Patterns — Yes
Correlations Profiling engine to discover interesting characteristics of a certain selection compared to some reference group. Highly interactive. Offers absolute and relative selectivity

indices. Includes several statistical tests to check significance of findings. Segmentation — discover subgroups in data by projection of high-dimensional data (up to thousands of dimensions) onto two dimensions. Use profiling engine to discover interesting characteristics of the clusters. Highly non-linear and local projection algorithm can model very complex data.

The profiling engine discovers interesting characteristics of a certain selection compared to some reference group. Highly interactive. Offers absolute and relative selectivity indices.

Algorithms under the hood — statistical tests and proprietary K-Nearest Neighbor projection algorithms. Data Display Formats — descriptives, series, frequencies, histograms, scatter plots, surface plots, and 3D scatter plots. In 2D — pie, polar, bubble, scatter, line, bar, area, hi-lo, candlestick, box-whisker, time series. In 3D — pie, bar, tape, area, scatter, and surface. Fully Interactive: a user can change any part of a diagram, remove series and points simply by clicking on it, etc.

• Provides Support for Model Building — Yes
Wizard to construct a prediction model. All parameter choices including train and test set selection can be made automatically. The underlying algorithm is the K-Nearest Neighbor, but the tool supports plugging in any other algorithm.

Levels of Data Mining Process
Basic Data Exploration; Extracting Relationships and Data Associations; Model Building

Data Mining Techniques
Supervised Induction; Association Discovery; Clustering; Visualization; Other
Fuzzy Matching, e.g., for matching applicants and vacancies on a website, finding suspects in a database of known offenders, etc.

• Functionality to Clean up Data — No
• Tool Executes Data Transformations — Yes
• Tool Handles Null/Missing Data Values — Yes
• Tool Performs Model Evaluation — Yes

Future Enhancements
Further (WWW) Client/Server development and further development of vertical data mining solutions.

White Papers and Resources
van der Putten, Peter in Data Mining in Direct Marketing Databases; Complexity and Management: A Collection of Essays, Walter Baets, Editor. World Scientific Publishing. ISBN 981-02-3714-6.

"Judgmental Computers, Sentient Machine Research: Applying AI to Real Business Problems." Tornado-insider.com. June 1999. http://www.Tornado-insider.com

## Data Mining Suite

http://www.datamining.com/dmsuite.htm

Information Discovery, Inc.
Marketing Communications
703-B Pier Avenue, Suite 169
Hermosa Beach CA 90254

Diana Lin
(310) 937-3600
(310) 937-0967 (FAX)
datamine@ix.netcom.com
http://www.datamining.com/

Keywords
SQL; Algorithm — Rule-Based; GUI (Graphical User Interface); Java; OLAP (On-Line Analytical Processing); Pattern Discovery; White Papers; C; Decision Support; Intranet; RDBMS (Relational Database Management Systems)

Platforms
Unix; Windows NT; Windows; IBM (AIX)

Description
The Data Mining Suite provides a solution for enterprise-wide, large scale decision support. It provides the ability to directly mine large multi-table SQL databases. The Data Mining Suite currently consists of these modules: Rule-based Influence Discovery; Dimensional Affinity Discovery; Trend Discovery Module; incremental Pattern Discovery; Comparative Discovery; and the Predictive Modeler.

Levels of Exploration

• Basic Data Exploration through Visualization — Yes
The tool provides for the browsing of data, schema, table layout, value selection, etc. to allow the users to have the first look at the data characteristics.
Query, indexing, and SQL function calls are used to achieve basic data exploration through visualization.

• Provides Discovery of Data Patterns — Yes
The tool finds patterns among segments, values, and identifies the patterns with variation and scores. Rule induction, logical reasoning, prediction, etc.

• Provides Support for Model Building — Yes
The tool allows users to build models of data infrastructure.

Levels of Data Mining Process
Basic Data Exploration; Extracting Relationships and Data Associations; Model Building

Data Mining Techniques
Supervised Induction; Association Discovery; Sequence Discovery; Clustering; Visualization

- Functionality to Clean up Data — Yes
- Tool Executes Data Transformations — Yes
- Tool Handles Null/Missing Data Values — Yes
- Tool Performs Model Evaluation — Yes

White Papers and Resources
http://www.datamining.com/datamine/loginb.htm

## Data SURVEYOR

http://www.ddi.nl/products/main.html

Data Distilleries
Kruislaan 419
1098 VA Amsterdam The Netherlands

Ienske Meindertsma
+31 20 562 0020
+31 20 562 0030 (FAX)
info@ddi.nl
http://www.ddi.nl/main.html

Keywords
ODMG (Object Database Management Group); C; C++; Intranet; Java; JavaBeans; ODBC–Compliant; WWW Interface; CORBA (Common Object Request Broker Architecture); Algorithm — Decision Tree; Algorithm — Rule-Based; Glossary; GUI (Graphical User Interface); Programmable; SQL; Toolkit; Visual Data Mining; Case Studies

Platforms
Unix (Solaris); Windows NT; Windows 95; Windows 98; Silicon Graphics (IRIX)

Description
Data SURVEYOR is an interactive data mining product line for business users and analysts enabling: data mining solutions for business users; an expert toolkit (Expert Suite) for analysts and domain experts providing extensive data mining functionality; and a solution factory for third parties allowing new solutions to be built and maintained.

Levels of Exploration

• Basic Data Exploration through Visualization — Not Known

• Provides Discovery of Data Patterns — Not Known

• Provides Support for Model Building — Not Known

Levels of Data Mining Process
Basic Data Exploration; Extracting Relationships and Data Associations; Model Building

Data Mining Techniques
Association Discovery; Visualization

• Functionality to Clean up Data — Not Known
• Tool Executes Data Transformations — Not Known
• Tool Handles Null/Missing Data Values — Not Known
• Tool Performs Model Evaluation — Not Known

## DataLogic/RDS

http://www.reduct.com/About_the_Company/products.htm

REDUCT & Lobbe Technologies
P.O. Box 3570
Regina, SK  S4P 3L7  Canada

(306) 586-9400
(306) 586-9442 (FAX)
info@reduct.com
http://www.reduct.com/

Keywords
Algorithm — Rule-Based; White Papers; Decision Support; Knowledge Discovery; Pattern Discovery

Platforms
Windows 95; Windows 98; Windows NT

Description
DataLogic/RDS is a tool for knowledge acquisition, classification, predictive modeling, expert system development and database mining. DataLogic extracts rules from databases by finding the best representation for the knowledge. DataLogic extracts knowledge from disorganized, incomplete and ambiguous data and presents this knowledge in the form of simple rule statements.

Levels of Exploration

• Basic Data Exploration through Visualization — No

• Provides Discovery of Data Patterns — Yes

• Provides Support for Model Building — Yes

Levels of Data Mining Process
Extracting Relationships and Data Associations; Model Building

Data Mining Techniques
Not Known

• Functionality to Clean up Data — Yes
• Tool Executes Data Transformations — No
• Tool Handles Null/Missing Data Values — Yes
• Tool Performs Model Evaluation — Yes

Future Enhancements
Real-time systems.

White Papers and Resources
http://www.reduct.com/About_the_Company/papers.htm

## DataMiner 3D

http://www.dimension5.sk/products/products.htm

DIMENSION 5, Ltd.
Hurbanova 36
SK-920 01 Hlohovec Slovakia

Dusan Toman
+421 905 409604
+421 804 7421223 (FAX)
dusan@dimension5.sk; info@dimension5.sk
http://www.dimension5.sk/

Keywords
Visual Data Mining; Data Visualization; Decision Support; Evaluation Copy; GUI (Graphical User Interface); C; Data Extraction; Flat Files; Knowledge Discovery; Microsoft SQL; ODBC–Compliant; Oracle; Pattern Discovery; SQL; Sybase; Scalable; Algorithm — K-Nearest Neighbor

Platforms
Windows 95; Windows 98; Windows NT

Description
The DataMiner 3D family of products provides a flexible data visualization for rapid visual analysis of large multidimensional data sets.

Levels of Exploration

• Basic Data Exploration through Visualization — Yes
The system displays data by mapping of real data to graphic attributes of a 3D data model. Rich sets of graphic attributes are available to allow the analyst to develop understandable multidimensional models (up to eight different dimensions simultaneously).

An interactive and intuitive user interface helps users to develop, adjust, validate, and maintain data models. Users can at anytime connect numerical (integer, real, scientific, or currency), logical, or textual data series to most selective graphic attributes (sizes, positions, rotations, colors, opacity, …), scale or change the data mapping method.

• Provides Discovery of Data Patterns — Yes
The system displays multiple data series at once, each mapped to another graphic attribute. The user selects data series included in the model, maps them to graphics, and finally visually explores and finds data correlations and data patterns.
Visual validation by the analyst is necessary.

Visual validation by the analyst is necessary.

The user can visually select a data subset at a criteria. The selection remains active also in another data model, where the same data are visualized under different rules. This provides the opportunity to see data subsets (clusters) from another scope and visually identify data patterns.

• Provides Support for Model Building — No

Levels of Data Mining Process
Basic Data Exploration; Extracting Relationships and Data Associations

Data Mining Techniques
Supervised Induction; Association Discovery; Sequence Discovery; Clustering; Visualization

• Functionality to Clean up Data — Yes
• Tool Executes Data Transformations — No
• Tool Handles Null/Missing Data Values — Yes
• Tool Performs Model Evaluation — Yes

## DataMite

http://www.lpa.co.uk/dtm.html

Logic Programming Associates (LPA) Ltd.
Studio 4
Trinity Road
London SW 18 3 SX England

Clive Spenser
(800) 949-7567
+44 181 871 2016
+44 181 874 0449 (FAX)
sales@lpa.co.uk
http://www.lpa.co.uk/

Keywords
ODBC–Compliant; Algorithm — Learning/Clustering; Algorithm — Rule-Based; Data Visualization; RDBMS (Relational Database Management Systems); Decision Support; Algorithm — K-Nearest Neighbor; GUI (Graphical User Interface); Microsoft SQL; Programmable; Toolkit; Wizards

Platforms
Windows 95; Windows 98; Windows NT

Description
DataMite enables rules and knowledge to be discovered in ODBC-compliant relational databases. DataMite requires neither programming skills nor specialized expertise, merely an understanding of the data to be explored. A user can point it at a file, tell it which outcomes he is interested in and it will discover patterns in the data that lead to relationships between columns. These relationships can vary from being exact to tenuous. The patterns are discovered through the synthesis of clusters utilizing the power of the database engine to do the counting. The resulting rules can be used to gain better insight into existing databases, predict future trends, build accurate models or fine tune business operations.

Levels of Exploration

• Basic Data Exploration through Visualization — Yes
Charting tools for visualizing relationship between variables and target.
Select a column; divide the column into intervals; chart frequencies of records in intervals that coincide with target. The charting package offers a high degree of interaction to control the resulting display.

• Provides Discovery of Data Patterns — Yes
Set threshold parameters; DataMite will ‘discover’ values/value ranges that are ‘interesting.’ A user can then combine these items to find intersections and hence rules, IF A & B & C & D ->E.

Entropy based measure of ‘interestingness.’
Induction-based counting algorithms; ODBC/SQL-based.

• Provides Support for Model Building — No

Levels of Data Mining Process
Basic Data Exploration; Extracting Relationships and Data Associations

Data Mining Techniques
Supervised Induction; Association Discovery; Clustering; Visualization

• Functionality to Clean up Data — Yes
• Tool Executes Data Transformations — No
• Tool Handles Null/Missing Data Values — Yes
• Tool Performs Model Evaluation — Not Known

## DataScope

http://www.cygron.hu

Cygron Research & Development, Ltd.
Pf.: 727
Szeged, H-6701 Hungary

Norbert Somlai
+36 62 435 505
+36 62 435 505 (FAX)
cygron@cygron.hu
http://www.cygron.hu

Keywords
Data Visualization; Decision Support; Demo; ODBC–Compliant

Platforms
Windows 95; Windows 98; Windows NT

Description
DataScope is a data mining tool that enables a user to visually analyze the contents of an arbitrary database and extract the knowledge hidden behind the numbers. Using special visualization techniques that support human thinking and intuition in analyzing the data, a user can recognize trends and patterns, or exceptions from these, simultaneously examine the data from different points of view and query the data without using special commands or formulas.

Levels of Exploration

• Basic Data Exploration through Visualization — Yes
DataScope displays up to 16 windows, each displaying a database field or a relation of up to seven fields. These windows operate synchronically, allowing the user to analyze the data from several points of view simultaneously. For numeric fields, DataScope employs a special distribution function that transforms data values to a percentage value that helps the users to transform the raw data into subjective judgments. DataScope also supports visual data query: query conditions can be specified without using commands or formulas and the query results are visible in all windows.

Determining the database fields to visualize: fully interactive. Selecting records: fully interactive and synchronic. Specifying query conditions: fully interactive. Looking for trends and exceptions in data: fully interactive. Manipulating the query results: fully interactive. Creating calculated fields: fully interactive.

• Provides Discovery of Data Patterns — No

• Provides Support for Model Building — No

Levels of Data Mining Process
Extracting Relationships and Data Associations

Data Mining Techniques
Clustering; Visualization

- Functionality to Clean up Data — No
- Tool Executes Data Transformations — Yes
- Tool Handles Null/Missing Data Values — Yes
- Tool Performs Model Evaluation — No

## dbProbe

http://www.InterNetivity.com/products.html

InterNetivity Inc.
1545 Carling Avenue
Suite 404
Ottawa  Ontario K1Z 8P9  Canada

Mickey Gill
(613) 729-4480
(613) 729-6711 (FAX)
info@InterNetivity.com
http://www.InterNetivity.com/

Keywords
OLAP (On-Line Analytical Processing); Demo; GUI (Graphical User Interface); ODBC–Compliant; WWW Interface; Visual Data Mining; Evaluation Copy; Flat Files; Java; Scalable; Decision Support; Informix; SQL

Platforms
Windows 95; Windows 98; Windows NT; Unix (Linux); Unix (Solaris); Hewlett-Packard (HP-UX); Digital Dec Alpha

Description
dbProbe is a business intelligence (OLAP and reporting) tool that combines powerful data analysis and scalability to thousands of users, with simple deployment for administrators. Users can drill down, slice-and-dice, graph, filter, create, and share reports and more. Data sources include MS OLE DB for OLAP, Informix Metacube, and others.

Levels of Exploration

• Basic Data Exploration through Visualization — Not Known

• Provides Discovery of Data Patterns — Not Known

• Provides Support for Model Building — Not Known

Levels of Data Mining Process
Basic Data Exploration; Model Building

Data Mining Techniques
Visualization

• Functionality to Clean up Data — Not Known
• Tool Executes Data Transformations — Not Known
• Tool Handles Null/Missing Data Values — Not Known
• Tool Performs Model Evaluation — Not Known

Future Enhancements

Integration with various report writer tools (Actuate, Sqribe, Seagate Crystal); integration with other backend server architectures; and additional reporting and analysis capabilities.

White Papers and Resources

"Get OLAP ASAP — Using dbProbe to add Web-based Reporting and Decision Support to Your Application" http://www.internetivity.com/oem.pdf

## DecisionWORKS

http://www.asacorp.com/product/index.html

Advanced Software Applications
333 Baldwin Road
Pittsburgh PA 15205

(800) 295-1938
(412) 429-1003
(412) 429-0709 (FAX)
http://www.asacorp.com/

Keywords
Demo; Algorithm — Learning/Clustering; Decision Support; Predictive Modeling; Pattern Discovery; Algorithm — Neural Networks; Algorithm — Genetic; Data Visualization; Workbench

Platforms
Not Known

Description
DecisionWORKS is composed of several tools including dbPROFILE (creates clusters and segmentations of data), ModelMAX (a predictive modeling application), ScorXPRESS (powerful pattern recognition capabilities), and DecisionPOSTM (decision support).

Levels of Exploration

• Basic Data Exploration through Visualization — Not Known

• Provides Discovery of Data Patterns — Not Known

• Provides Support for Model Building — Not Known

Levels of Data Mining Process
Model Building

Data Mining Techniques
Clustering; Visualization

• Functionality to Clean up Data — Not Known
• Tool Executes Data Transformations — Not Known
• Tool Handles Null/Missing Data Values — Not Known
• Tool Performs Model Evaluation — Not Known

## DeltaMiner

http://194.152.41.50/english/Delta/Products/Miner/delta_miner.htm

MIS AG
325 Columbia Turnpike
Florhan Park  NJ  07932

Christopher Peron
(800) 647-3177
(973) 765-0405
(973) 765-0305 (FAX)
cperon@mis-ag.com
http://194.152.41.50/english/english.htm

Keywords
ODBC–Compliant; OLE; GUI (Graphical User Interface); OLAP (On-Line Analytical Processing)

Platforms
Windows 95; Windows 98; Windows NT

Description
DeltaMiner accelerates typical tasks by imitating the way human experts investigate data.  It integrates over 15 predefined analyses such as Navigation, Time Series, Power Search, Cross Table, ABC Analysis, and Meta Search.  The analysis techniques are based on a combination of OLAP technology, data mining, and statistical heuristics and methods.  DeltaMiner explains deviations, variances, and exceptions.  It also detects compensations and helps end-users to navigate through their financial, sales, or web database.

Levels of Exploration

• Basic Data Exploration through Visualization — Yes
Portfolio Analysis is used to position Object in the four-quarter matrix with regard to two variables.  Data clusters and exceptions (Objects, that have a different variable-share with regard to the average share in the total group of analysis objects) can be identified, the classification set by the user can be integrated as a new structure in the database.

ABC-Analysis is used to classify objects in A, B, and C classes and to analyze the concentration of the objects with regard to the selected variable.  All calculation steps are executed automatically, a diagram (concentration curve) showing the level of concentration is automatically drawn and the classification (set by the user) can be integrated as a new structure in the database.

Crosstables are used to create standard reports.  To identify the most important exception very rapidly, each crosstable can be switched to a diagram.  The complete distribution of values becomes obvious — extreme exceptions, data-clusters, and remarkable values distributions

can be identified. Movement analysis is used to analyze life cycles or the dynamic structure changes in a certain time periods. Timelines are used to analyze the history behind a data-phenomenon. Moreover, different timelines (timelines for different analysis objects or different variables) can be compared and analyzed simultaneously.

Object Analysis shows the most important variables of the analysis objects in several charts. Buttons are used to switch between different charts. Each variable that is included in one of the charts can be used for analysis. Pure data mining with the Comparator automatically identifies the most striking distribution differences between variables for any analysis object. Analysis of all available objects is completed in a single step.

The Profiler automatically identifies the most significant profiles (characteristics) for analysis objects with regard to the selected variable. Comparing the average variable share of the object with all possible combinations of object-characteristics shows significant profiles. Power-Search provides ranking functionality. Creating ascending (winning) or descending (losing) hit lists of objects with regard to a variable. Moreover, Power-Search can easily be used to identify compensations within a dataset.

All methods described earlier are available for each variable that is in one of the object-analysis charts and all of the data objects. The user only has to select the appropriate method by pressing a button, selecting the key figure by dragging and dropping it in the analysis window and selecting the objects to be analyzed. For executing one of the analysis functions, there are only a few clicks necessary. Moreover, for power data analysis it is necessary to build and follow-up analysis chains — to link different analysis methods to each other (e.g., identifying the cause of a variance and analyzing the history behind the phenomenon using a timeline analysis).

• Provides Discovery of Data Patterns — Yes
Data Mining in CrossTables — Analysis assistants automatically identify data clusters, remarkable distributions of values, outliers, index of expectancy, etc. Automatic Identification of associations. Automatic Navigation to identify and explain variances, outliers, or correlations (Descriptor, Comparator, and Profiler, described earlier). Automatic Identification of distribution differences (Descriptor, Comparator, and Profiler, described earlier). Automatic Identification of profiles (Profiler, described earlier).

DeltaMiner is capable of identifying the most important data patterns automatically. This is because the main idea behind DeltaMiner is to show the user automatically and very rapidly, the most important information. Using the different built-in methods of MIS DeltaMiner a user can let the data speak for itself. This conception is called ‘active information management.’ There are many ‘interestingness’ measures calculated by DeltaMiner in the background — most of them are based on statistical measures such as scattering, degree of heterogeneity, variance, etc. Calculating these measures during analysis, DeltaMiner imitates the behavior of a human expert in statistics and filters the important information out of the database. The only thing that the user has to set up, is the selection of variables and the objects to be used for analysis. This approach requires no in-depth knowledge of building statistical models or data mining models on the part of the user.

The algorithms are based on statistical measures and imitate human behavior. This is why DeltaMiner's methods do not fit in the typical data mining classifications. The data display format (diagrams, color-coding, line diagrams) are method-specific. In addition, all patterns that are found are described in textual format.

• Provides Support for Model Building — No

Levels of Data Mining Process
Basic Data Exploration

Data Mining Techniques
Association Discovery; Sequence Discovery; Clustering; Visualization

Discovery of Distribution Differences using Comparator and Descriptor — Identifying significant distribution differences between data segmentations or different variables (e.g., used to analyze the result of a mailing or marketing campaign by comparing the number of mails and the characteristics of the responses). Automatic Deviation Analysis — A typical analysis task is to explain a variance between different variables (e.g., actual vs. budget). Identifying the source of such a variance, the automatic navigation of DeltaMiner suggests a data-driven drill-down path through the various data dimensions and the different levels of the detail unit the main cause of the variance is clearly identified. Instead of searching for explanations, DeltaMiner leads the user directly to the most significant exceptions.

• Functionality to Clean up Data — Yes
• Tool Executes Data Transformations — Yes
• Tool Handles Null/Missing Data Values — Yes
• Tool Performs Model Evaluation — No

## IBM Intelligent Miner for Data

http://www.software.ibm.com/data/iminer/fordata/index.html

IBM
7100 Highlands Parkway
Smyrna GA

Linda Davis
(800) 426-2255
(770) 863-1825
(800) 242-6329
lkdavis@us.ibm.com; ibm_direct@vnet.ibm.com
http://www.software.ibm.com/data/intelli-mine/

Keywords
Algorithm — Neural Networks; Algorithm — Rule-Based; Case Studies; Demo; Flat Files; GUI (Graphical User Interface); Java; Knowledge Discovery; Pattern Discovery; RDBMS (Relational Database Management Systems); Visual Data Mining; White Papers; Scalable; Algorithm — Decision Tree; Algorithm — Regression

Platforms
Windows 95; Windows NT; IBM (AIX); IBM MVS; OS/2

Description
The IBM Intelligent Miner family helps to identify and extract high-value business intelligence from data assets. Through a process of 'knowledge discovery,' an organization can leverage hidden information in its data, uncovering associations, patterns, and trends that can lead to competitive advantage.

Levels of Exploration

• Basic Data Exploration through Visualization — Yes
IM's function 'Bivariate Statistics' computes statistical information about the data. A comparison is done between distributions of all data records and those with a specific field value or within a specific range (bivariate statistics). Statistics include discrete distributions and distributions by range (histograms), modal values, mean values and standard deviation, as well as Chi-square and F-Test results. In addition, IM provides the possibility to register external visualization or data exploration tools that are then launched from within IM. Principal Component Analysis and Factor Analysis find linear dependencies between numeric variables and generate a new equivalent (or similar) set of linearly independent variables.

Results are presented graphically, with histograms and pie charts to show the distributions for all selected variables. The (bivariate) comparison is done by overlaying two distributions in the same histogram or as two rings of a pie. A 'Detail' menu option shows the additional statistical values. Principal Component Analysis and Factor Analysis show factor and variable

relationships, factor loadings, rotated factor loadings, correlation coefficients, principal attributes, eigenvectors etc.

• Provides Discovery of Data Patterns — Yes
Discover Clusters, Discover Associations, Discover Sequential Patterns, Discover Similar Sequences

Discover Clusters: Notion of record similarity (based on value distributions), notion of importance of a field for the definition of a cluster. Discover Associations: Support, Confidence, and Lift of rules and item sets, Rules at a meaningful level of generality, by using a taxonomy (hierarchy) of items. Discover Sequential Patterns: Support of sequential patterns. Discover Similar Sequences: Match fraction for pairs of sequences.
Discover Clusters: IM provides two algorithms, Neural Net (Kohonen feature maps) and Demographic (relational data analysis, Condorcet criterion). Displayed as bivariate statistics with respect to clusters. All Clusters, Single Cluster, and Single Field views. Details views of statistical information (field importance, distributions etc.) Discover Associations: IM provides an Associations Discovery algorithm (generate large item sets). Graphical display of rules and frequent item sets, Rules as plain English text, Statistical information. Discover Sequential Patterns: IM provides a Sequential Patterns Discovery algorithm (generate frequent sequences). Display of sequential patterns. Statistical information. Discover Similar Sequences: Given a database of (time) sequences, pairs of similar sequences or subsequences are found. Display of pairs of sequences with matching parts colored.

• Provides Support for Model Building — Yes
IM provides Classification and Numeric Prediction as well as Time Series forecasting. The following techniques are employed: Back propagation (neural network); Radial Basis Functions (RBF); Decision Tree; Modified CART; regression tree (tree induction); Linear Regression; Polynomial Regression; Logistic regression (with neural net); and Univariate Curve Fitting.

Levels of Data Mining Process
Not Known

Data Mining Techniques
Supervised Induction; Association Discovery; Clustering; Visualization; Other
Discover Similar Sequences; Principal Component Analysis; and Factor Analysis

• Functionality to Clean up Data — Yes
• Tool Executes Data Transformations — Yes
• Tool Handles Null/Missing Data Values — Yes
• Tool Performs Model Evaluation — Yes

White Papers and Resources
http://www.software.ibm.com/data/iminer/fordata/library.html

## IDL (Interactive Data Language) and IDL DataMiner

http://www.rsinc.com/idl/idl_dataminer.cfm

Research Systems, Inc.
4990 Pearl East Circle
Boulder CO 80301

Mark Goosman
(303) 786-9900
(303) 786-9909 (FAX)
mgoosman@rsinc.com; info@rsinc.com
http://www.rsinc.com/

Keywords
Case Studies; Data Visualization; Demo; GUI (Graphical User Interface); ODBC–Compliant; White Papers; Algorithm — K-Nearest Neighbor

Platforms
Windows 95; Windows 98; Windows NT; Unix (Solaris); MacOS; Hewlett-Packard (HP-UX); IBM (AIX); Silicon Graphics (IRIX); Compaq (OpenVMS)

Description
IDL, the Interactive Data Language, enables in-depth data analysis through visualization. It can be used for cross-platform application development. The optional IDL DataMiner allows a user to connect directly to a database for easy access, query and edit actions from one or multiple ODBC databases.

Levels of Exploration

• Basic Data Exploration through Visualization — Yes
IDL provides a cross platform, 4GL which provides an environment for easy data access, analysis, and visualization as well as a complete development environment for the development of related applications. Specific functionality includes: Contour plots, automatic boundary close; Filled contours; Line plots, scatter plots, histograms, bar graphs, polar plots, error bars; Linestyle, color and marker attributes; Log, semi-log and linear scaling; Overplot multiple data sets; Vector flow diagrams; Surface Plotting and 3D Graphics; 3D transformations; 4D data display of gridded elevations with overlaid image or user-specified shading; Mesh surface plots with hidden line removal; Regular and irregular gridding; Shaded surface representations of solids and gridded elevations; Surface interpolation of irregularly gridded data points; Volume contouring; Voxel rendering; Graphics Architecture; Direct and object graphics; Efficient rendering algorithms; Elements and graphic device; Fine-grained access and control; Independent interface table; OpenGL accelerated 3D graphics; Z-buffered graphics; Graphic Effects; 3D plot symbols and text; Flat shading; Gouraud shading; Linestyles, patterns, plot symbols; Multiple, colored lights, point, directional, ambient, spot light sources; Color Systems; RGB, HLS, HSV, indexed available on all graphic devices; CMYK, HSV, HLS to RGB color value control; and Opacity (RGB model only).

IDL provides a hardware accelerated OpenGL graphics system for the quick display of 2D, 3D, up to n dimensional data display. In addition, IDL's flexible graphics system allow users to define just about any type of custom visualization required.

• Provides Discovery of Data Patterns — Yes
IDL provides a rich set of data access and analysis functions including the following: Curve and Surface Fitting; Gradient-expansion nonlinear least-squares; Levenberg-Marquardt nonlinear least-squares; Multiple linear regression; Polynomial spatial warping; Polynomial surface; Singular-value-decomposition nonlinear least-squares; Weighted and unweighted least-squares polynomial; Image and Signal Processing; 1-, 2- and 3D convolution; Adaptive Fast Fourier transform; Bi-level, pseudo- and true-color thresholding; Block convolution; Convert true-color to pseudo-color; Color systems: RGB, HLS and HSV, Indexed; Convolution and frequency-domain block convolution; Fourier Transform: 1 to 7 dimensions, any number of points; Frequency domain filtering & analysis; Generalized image arithmetic; Geometric transformations: magnification, reduction, rotation, polynomial warping with regular or irregular grids; High- and low- pass filtering; Histogram equalization and processing; Image annotation; Interactive contrast enhancement; Lomb periodogram; Median filtering; Morphological operators: erode and dilate; Region of Interest (ROI) selection; Roberts edge enhancement; Signal editing; Sobel edge enhancement; Spectral analysis; Time-series analysis; Waveform generation; Wavelet transform using Daubechies' coefficients; Zoom and pan; Integration; Iterated Gaussian quadrature; Modified Romberg integration over an open interval; Newton-Cotes integration of tabulated data; Romberg integration over a closed interval; Simpson integration over a closed interval; Eigensystems; QR, QL, and TQ algorithms; Tridiagonal forms; Subspace iteration; Linear Systems; Cholesky, Gauss-Seidel, LU, SVD, and tridiagonal methods; Complex LU decomposition and backsubstitution; Condition number; Cramer's algorithm; Determinant; Generalized inverse; Gauss-Seidel iteration; Infinity and Euclidean norms; Perturbed solution iteration; Transpose; Tridiagonal forms; Sparse Linear Systems; Dense-to-sparse and sparse-to-dense conversions with thresholds; Iterative biconjugate-gradient algorithm for solving linear equations; Multi-dimensional optimization; Row-indexed sparse storage format; Sparse format file I/O; Sparse matrix-matrix and matrix-vector multiply; Nonlinear Systems and Root Finding; Broyden's and Newton's globally-convergent algorithms for systems of nonlinear equations; Differential equations; Laguerre's algorithm for polynomial root-finding; Muller's algorithm for real and complex root-finding; Multi-Dimensional Optimization; Davidon-Fletcher-Powell minimization; Gradient-free Powell minimization; Special and Transcendental Functions; Beta and incomplete beta functions; Error and exponential integral functions; Exponentials and logarithms; Forward and inverse Chebyshev polynomial expansion; Gamma, incomplete gamma, and logarithmic gamma functions; I-Bessel, J-Bessel, and Y-Bessel functions; Trigonometric, inverse trigonometric, and hyperbolic functions; Correlation Analysis and Forecasting; Auto and cross covariances/correlation; Autoregressive modeling/forecasting; Cluster analysis; Differencing/box-car smoothing; Discrete auto/cross correlation; Exponential, geometric, Gompertz, hyperbolic, logistic, and logsquare growth models; Gradient-expansion nonlinear least-squares; Kendall and Spearman rank correlations; Lagged auto and cross correlations; Least-absolute-deviation fitting; Levenberg-Marquardt curve fitting algorithm; Linear, multiple and partial correlations; Moving averages/smoothing; Multiple linear regression; Multiple correlation; Nonlinear least-squares fitting; Partial correlation; Polynomial surface fitting; Principal components; Statistical fitting of

data; Weighted and unweighted polynomial fitting; Hypothesis Testing; Chi-square, F, Gaussian (normal) and Student's T tests; Chi-squared model validation; Contingency test for independence; Cumulative binomial (Bernouli); F-variances test; Kruskal-Wallis H-test; Lomb frequency test; Mann-Whitney U-test; Median delta test; Normality test; Normally- and uniformly-distributed pseudo-random numbers; Runs test for randomness; Sign test; T-means test; Wilcoxon rank-sum test; Multi-Dimensional Gridding and Interpolation; 1-, 2- and 3D nearest-neighbor and linear; 1-, 2- and 3D cubic convolution; 2D parametric cubic splines; 2D Delaunay triangulation; 2D linear and quintic gridding; 2D quintic extrapolation; 2D Voronoi polygon; 3D interpolation using Kriging; 3D minimum curvature surfaces; 3D polar (r, theta, z) to rectangle; 4D smooth fit; Cubic splines; Spherical gridding; Supports non-uniformly gridded data; Mapping; High-resolution map database; 19 geographic mapping transformations; and Warp image data onto arbitrary projections.
Not by default, but could be designed.

• Provides Support for Model Building — No

Levels of Data Mining Process
Basic Data Exploration

Data Mining Techniques
Visualization

• Functionality to Clean up Data — Yes
• Tool Executes Data Transformations — Yes
• Tool Handles Null/Missing Data Values — Yes
• Tool Performs Model Evaluation — No

White Papers and Resources
http://www.rsinc.com/inprint/index.cfm

## Informix Red Brick Formation

http://www.redbrick.com/products/formation/vformbtm.htm

Informix Software, Inc.
4100 Bohannon Drive
Menlo Park  CA  94025

(800) 331-1763
(650) 926-6300
http://www.informix.com/cgi-bin/contact.pl

Keywords
Data Mart; Data Warehouse; Data Extraction; GUI (Graphical User Interface); Informix

Platforms
Not Known

Description
Red Brick Formation is a powerful and flexible data extraction and transformation tool that substantially reduces the time and complexity of building and maintaining very large data warehouses and data marts.

Levels of Exploration

• Basic Data Exploration through Visualization — Not Known

• Provides Discovery of Data Patterns — Not Known

• Provides Support for Model Building — Not Known

Levels of Data Mining Process
Extracting Relationships and Data Associations

Data Mining Techniques
Not Known

• Functionality to Clean up Data — Not Known
• Tool Executes Data Transformations — Not Known
• Tool Handles Null/Missing Data Values — Not Known
• Tool Performs Model Evaluation — Not Known

## KATE-DataMining

http://www.acknosoft.com/fTools.html

AcknoSoft
1814 Riverbend Crossing
Sugar Land  TX  77478

Nathalie Frezouls
(281) 265-5360
(281) 265-5360 (FAX)
nfrezouls@compuserve.com
http://www.acknosoft.com/

Keywords
CBR (Case Based Reasoning); Algorithm — Decision Tree; Decision Support; GUI (Graphical User Interface); C++; Intranet; WWW Interface; Algorithm — Learning/Clustering; EIS (Executive Information System); JavaBeans; Knowledge Discovery; Oracle; Visual Data Mining; Demo; Workbench

Platforms
Windows 3.1; Windows 95; Windows NT; Unix (Solaris)

Description
As a Case-Based Reasoning (CBR) tool, KATE recalls past experience that is similar to the current problem and adapts the solution that worked in the past in order to solve the current problem.  Using induction technology, KATE-DataMining extracts knowledge that is hidden in the data. The tool automatically generates decision trees where the essential information for efficient decision making is presented. The developer may also choose to optimize the time required as well as the cost of decision making and introduce his expertise on the relative importance of the descriptors. He can use several interactive graphical tools to detect hidden dependencies, parameter shifts, and anomalies in the data or predict trends. He can also choose to modify the decision trees by hand. The auto-consultation module is used to test the tree automatically.  With KATE-Editor, KATE-CBR, KATE-Data Mining, and KATE-Runtime, the KATE suite of software tools eases the development of powerful decision support systems.  As an option, clients may add KATE-Call Tracking, KATE for R/3 Service Management, or KATE-WebServer.  KATE-Editor contains three modules: The model editor allows the definition of the relevant terms that will be used to describe cases and the creation of the case structure using objects. The questionnaire generator supports the development of electronic questionnaires to acquire cases. Dynamic, the questionnaire only asks questions that are relevant in a given context. Questionnaires can include pictures, drawings, audio, or video clips. The database import facility is used to automatically create a KATE model and case base from an existing database or a spreadsheet.  KATE-CBR works well even in domains that are poorly understood because CBR does not need to know why a solution worked in the past.  Exploiting the power of case based reasoning, KATE-CBR contains two modules: The nearest neighbor module is used to compare the current problem with ones that have already been solved, to

retrieve the most similar cases and to adapt their known solutions. The similarity measure can be customized and the relative weights of descriptors adjusted to suit the specific application. During a consultation, the dynamic induction module enables discovery of the most discriminating questions and retrieves relevant cases efficiently. One can start working with few cases. New cases are added over time to enhance the content of the knowledge base. KATE-DataMining, using induction technology, extracts knowledge that is hidden in data. The tool automatically generates decision trees wherever the essential information for efficient decision making is presented. The user may also choose to optimize downtime, or the financial cost of decision making, by favoring quick and cheap tests. He may also input his own expertise on the relative importance of the descriptor. Visual data mining can also be performed by using a wide range of interactive graphical tools to detect hidden dependencies, parameter shifts, anomalies in the data or to predict trends. The user can also choose to modify the decision trees by hand. The auto-consultation module tests the tree automatically. KATE-Runtime is used to distribute applications. KATE-CallTracking provides simple call tracking facilities for call centers. KATE for R/3 Service Management is an optional tool that is integrated with R/3 Service Management, the call tracking module of SAP. KATE-WebServer allow access to the decision support system over an Intranet or over Internet.

Levels of Exploration

• Basic Data Exploration through Visualization — Yes
KATE's basic aim is to group data. Additionally, KATE enables users to visualize how these groups of data achieve certain properties (enabling the user to evaluate the quality of the groups). With respect to a specific value of a specific attribute, this feature enables a user to see at a glance whether a group of data achieves more or less a specific property. With respect to all values of a specific attribute (spectrum), the visualization will provide for each group of data in the group.

The basic technique used for the visualization procedure is frequency counted. The procedure is completely customizable in the front-end interface by the user. The procedure applies to any group of data proposed by KATE. The visualization is displayed in a browser window, which is printable, exportable, etc.

• Provides Discovery of Data Patterns — Yes
KATE creates automatically data clusters and enables users to identify anomalies in the data with respect to the identified patterns. The technique used for identifying with clusters is called induction. The data clusters are identified both in a global graphical view and in a detailed statistical manner (i.e., each cluster can be analyzed automatically for extracting the common features, etc.).

The induction process is based on a measure called the information gain that extracts, at each stage of the building process, the set of most interesting, or discriminating, attributes amongst the initial attributes.

The algorithm uses a hill-climbing strategy. The input is a set of data, described by attribute (any type of attribute one may find in a standard relational database may be used). The information gain theory estimates the best candidates for building the clusters and proposes

the best one (the user can choose another one or influence on to the algorithm in various ways). The process is iterated until some stopping criteria are found (e.g., minimal number of data in a cluster, etc.). The basic algorithm is described in the literature (see for instance Quinlin, R. Induction of Decision Trees and its Application to Chess and Games, in Machine Learning 1, an Artificial Intelligence Approach, Morgan Kauffmann, 1986). This basic algorithm has been improved so that it uses an object-oriented language to describe the cases; uses background domain knowledge; minimizes the overall cost of decisions (financial cost, reliability of tests, time required to perform a test, etc.); and handles continuous tests by computing thresholds or intervals.

• Provides Support for Model Building — Yes
The clusters provided by the algorithms are organized in a classification tree. When the user has a new problem, he can use the tree for decision-making purposes. The algorithm follows the branches from the root to the leaves of the tree, by using the attributes that have been recognized as the most interesting during the building phase. Once one or several leaves have been found, KATE reuses the patterns of the leaves to predict the pattern of the new problem. The whole tree can be turned into a set of decision rules.

Levels of Data Mining Process
Extracting Relationships and Data Associations; Model Building

Data Mining Techniques
Supervised Induction; Clustering; Visualization; Visualization; Other
Dynamic Induction — This is a combination between the traditional induction, which builds static decision trees, and approaches from the Case-Based Reasoning (CBR) field. It uses together the information gain measure and similarity measures to evaluate the analogy between data and groups of data. The dynamic induction may be driven by the system (the system proposes at each step the most relevant attribute and then looks for similar cases depending on the attribute value) or user-driven (the system proposes at each step the list of the most relevant attributes and the user chooses the best one from his point of view).

• Functionality to Clean up Data — Yes
• Tool Executes Data Transformations — Yes
• Tool Handles Null/Missing Data Values — Yes
• Tool Performs Model Evaluation — Yes

Future Enhancements
The Kate suite will soon become available for Linux.

White Papers and Resources
Bergmann R., S. Breen, M. Goker, M. Mango, and S. Wess Editor. Developing Industrial Case Based Reasoning Applications, Springer-Verlag

## KnowledgeSEEKER

http://www.angoss.com/ksprod/kspage.htm

ANGOSS Software Corporation
34 St. Patrick Street
Suite 200
Toronto ON Canada M5T-1V1

Barry de Ville
(416) 593-1122
(416) 593-5077 (FAX)
bdeville@angoss.com
http://www.angoss.com/

Keywords
White Papers; Algorithm — Decision Tree; Programmable; Intranet; Case Studies; Evaluation Copy; Knowledge Discovery; Visual Data Mining; Algorithm — Rule-Based; Demo; GUI (Graphical User Interface); OLAP (On-Line Analytical Processing); Pattern Discovery; SAS Analytical Tool; SQL

Platforms
Windows 95; Windows 98; Unix; Unix (Solaris); Unix (Irix); Unix (Digital Unix); Unix (Linux); Unix (Sinux); Windows 3.x; IBM (AIX); Hewlett-Packard (HP-UX); SCO

Description
At the heart of KnowledgeSEEKER is a Decision Tree Induction process, which in simplified terms acts as an automated query generator, so users do not have to manually construct the queries. This Decision Tree Induction process has the mathematical power and crunch power to construct and run the queries required. This process shows the combined dependencies between multiple predictors. The analysis results are presented in a highly intuitive colored classification tree. Decision trees allow for effective data visualization and are extraordinarily easy to understand and manipulate.

KnowledgeSEEKER findings can also be translated into a knowledge base of rules or a set of executable programming statements.

Levels of Exploration

• Basic Data Exploration through Visualization — Not Known

• Provides Discovery of Data Patterns — Not Known

• Provides Support for Model Building — Not Known

Levels of Data Mining Process
Extracting Relationships and Data Associations

Data Mining Techniques
Visualization; Association Discovery

- Functionality to Clean up Data — Not Known
- Tool Executes Data Transformations — Not Known
- Tool Handles Null/Missing Data Values — Not Known
- Tool Performs Model Evaluation — Not Known

White Papers and Resources
http://www.angoss.com/kstudio/whitepaper/whitepaper.htm

## KnowledgeSTUDIO

http://www.angoss.com/kstudio/studio.htm

ANGOSS Software Corporation
34 St. Patrick Street
Suite 200
Toronto ON Canada M5T-1V1

Barry de Ville
(416) 593-1122
(416) 593-5077 (FAX)
bdeville@angoss.com; info@angoss.com, support@angoss.com
http://www.angoss.com/

Keywords
Object Oriented User Interface; Predictive Modeling; White Papers; Algorithm — Decision Tree; Algorithm — Learning/Clustering; Algorithm — Neural Networks; SAS Analytical Tool; ActiveX; Intranet; Programmable; Java; Visual Data Mining; Demo; Evaluation Copy; Workbench; ActiveX (MicroSoft);; Decision Support

Platforms
Unix; Windows 95; Windows 98; Windows NT; Windows 3.x;

Description
The KnowledgeSTUDIO product line is a new generation of data mining software from ANGOSS Software Corporation. These new technologies focus on integrating advanced data mining techniques into corporate environments so that business can achieve the maximum benefit from their investment in data.

Levels of Exploration

• Basic Data Exploration through Visualization — Yes
KnowledgeSTUDIO supports basic data exploration through data value profiling and a chart wizard that provides functionality to the level of chart options in Microsoft Excel. The charts can be easily edited and cut and pasted into other applications. It is easy to move from field to field in this form of interaction.

All data fields may be displayed as a graph object and can be edited through the graph object editing procedure or through the application of a graph object wizard. This functionality is similar to the functionality provided in the graph functions of Microsoft Excel.

• Provides Discovery of Data Patterns — Yes
KnowledgeSTUDIO supports data pattern discovery through decision tree analysis (CHAID and CART-like XAID analysis), three forms of neural networks and two forms of cluster analysis.

Generally the automatic discovery of patterns is done through the application of an adjusted test of significance or through some variance reduction and testing technique such as least squares.

KnowledgeSTUDIO supports decision tree analysis using a CHAID-based k-way decision tree approach that supports both categorical (CHAID) and continuous (XAID) outcomes. Merging and splitting algorithms, based on a simple test statistic, are used to identify the branches of a decision tree for each field in the analysis while a generalized, Bonferonni-adjusted statistical test is used for each branch on the decision tree to determine significance. KnowledgeSTUDIO supports the development of a multilayered neural network. This is a two-weight layer neural network with specified hidden and output layer activation functions. Radial Basic Functions are also identified. These are approximated with Gaussian basis functions. Probabilistic Neural Networks are identified. Cluster analysis by means of the K-means algorithm is undertaken. Cluster analysis by means of the expectation-maximization algorithm is undertaken.

• Provides Support for Model Building — Yes
Classification and regression tree models are held in internal representation to be applied against a novel data set or, alternatively, the predictive or classification attributes can be output as text rules, as SQL CASE or SELECT statements, as SAS or SPSS code or as Java code. All other modeling representations, including the neural networks and clustering algorithms are held in an internal representation for subsequent scoring of an external data set through the application of the KnowledgeSTUDIO scoring environment.

Levels of Data Mining Process
Basic Data Exploration

Data Mining Techniques
Supervised Induction; Association Discovery; Sequence Discovery; Clustering; Visualization

• Functionality to Clean up Data — Yes
• Tool Executes Data Transformations — No
• Tool Handles Null/Missing Data Values — Yes
• Tool Performs Model Evaluation — Yes

Future Enhancements
Version 3.0 (Q4 1999); Non Microsoft-dependent Java solutions, for model deployment and data exploration; support for OLE DB, Algorithm Developers Kit (ADK); enhanced data visualization; and connectivity options for data import from external sources via the Internet.

White Papers and Resources
http://www.angoss.com/kstudio/whitepaper/whitepaper.htm

## MARS

http://www.salford-systems.com/products.html

Salford Systems
8880 Rio San Diego Drive
Suite 1045
San Diego, CA  92108

Kerry Martin
(619) 543-8880
(619) 543-8888 (FAX)
kerry@salford-systems.com
http://www.salford-systems.com/

Keywords
GUI (Graphical User Interface); Algorithm — Regression; Decision Support; Demo; White Papers; Evaluation Copy; Knowledge Discovery; ODBC-Compliant; Pattern Discovery; Predictive Modeling; Regression Splines; Tutorial

Platforms
Windows 95; Windows NT; Unix; IBM VMS & CVS

Description
MARS is a multivariate non-parametric regression procedure introduced in 1991 by Stanford statistician and physicist, Jerome Friedman.  Salford Systems' MARS, based on the original code, has been substantially enhanced with new features and capabilities in exclusive collaboration with Dr. Friedman.  MARS can automatically find optimal variable transformations and interactions, the complex data structure that often hides in high-dimensional data.

Levels of Exploration

• Basic Data Exploration through Visualization — No

• Provides Discovery of Data Patterns — Yes
MARS excels at automatically finding optimal variable transformations and interactions, the complex data relationships that often go undiscovered in high-dimensional data.  This flexible approach to regression modeling can be used for a wide variety of analyses, such as when you need to accurately predict how much? How many?  How long?  Key drivers of business outcomes are graphically depicted, enabling a more intuitive understanding of the significant relationships automatically revealed by MARS.

MARS automatically separates relevant from irrelevant predictor variables, transforms predictor variables exhibiting a nonlinear relationship with the target variable, and determines interactions between variables needed for predictive accuracy.

MARS or “Multivariate Adaptive Regression Splines” was developed in the early 1990s by Jerome Friedman, a world-renowned Stanford statistician and one of the co-developers of CART. Inherently a regression tool, this new methodology automates the process of building accurate predictive regression models for continuous and binary outcome response variables.

MARS enables analysts to rapidly search through all possible models and quickly identify the optimal solution. Because the software can be exploited via intelligent default settings, for the first time, analysts at all levels can easily access MARS’ innovations.

• Provides Support for Model Building — Yes
The MARS methodology automates the process of building accurate predictive regression models for continuous and binary outcome response variables. In doing so, this new approach to data mining effectively uncovers complex data patterns and relationships that are difficult, if not impossible, for other methods to reveal.

Given a target variable and a set of candidate predictor variables, MARS automates all aspects of model development, including
• separating relevant from irrelevant predictor variables,
• transforming predictor variables exhibiting a nonlinear relationship with the target variable,
• determining interactions between variables needed for predictive accuracy,
• handling missing values with new nested variable techniques, and
• conducting extensive self-tests to protect against over fitting.

To oversimplify, MARS builds flexible models by fitting piecewise linear regressions; that is, the non-linearity of a model is approximated through the use of separate regression slopes in distinct intervals of the predictor variables. The variables to use and the end points of the (possibly many) intervals for each variable are found via a fast intensive search procedure. In addition to searching variables one by one, MARS also searches for interactions between variables, allowing any degree of interaction to be considered.

The ‘optimal’ MARS model is selected in a two-stage process. In the first stage, MARS constructs an overly large model by adding ‘basis functions’ — the formal mechanism by which variable intervals are defined. Basis functions represent either single variable transformations or multivariable interaction terms. As basis functions are added the model becomes more flexible and more complex, and the process continues until a user-specified maximum number of basis functions is reached.

In the second stage, basis functions are deleted in order of least contribution to the model until an optimal model is found. By allowing for any arbitrary shape for the function as well as for interactions, and by using this two-stage model selection method, MARS is capable of reliably tracking very complex data structures that often hide in high-dimensional data.

Levels of Data Mining Process
Basic Data Exploration; Extracting Relationships and Data; Model Building

Data Mining Techniques
Supervised Induction

- Functionality to Clean up Data — Yes
- Tool Executes Data Transformations — Yes
- Tool Handles Null/Missing Data Values — Yes
- Tool Performs Model Evaluation — Yes

Future Enhancements
Sophisticated handling of time series data; handling of character variables; optimizing MARS models for logit (binary response variables) and extending MARS to multinomial logit (multi-class dependent variables).

White Papers and Resources
http://www.salford-systems.com/whitepaper.html

Friedman, J.H. (1991) “Multivariate Adaptive Regression Splines,” Annals of Statistics, 19, 1-141 (March).

## MODEL 1

http://www.unica-usa.com

Unica Technologies, Inc.
55 Old Bedford Road
Lincoln MA 01773

Diane Robinson
(781) 259-5900
(781) 259-5901 (FAX)
dianerob@unica-usa.com; unica@unica-usa.com
http://www.unica-usa.com

Keywords
GUI (Graphical User Interface); Programmable; Scalable; Workbench

Platforms
Windows 95; Windows NT

Description
Model 1 is a suite of targeted data mining solutions. It can be used effectively by both statisticians, and modelers. Model 1 produces solid, actionable marketing knowledge that represents the ROI from data collection/storage activities such as data warehousing.

Levels of Exploration

• Basic Data Exploration through Visualization — Not Known

• Provides Discovery of Data Patterns — Not Known

• Provides Support for Model Building — Not Known

Levels of Data Mining Process
Not Known

Data Mining Techniques
Not Known

• Functionality to Clean up Data — Not Known
• Tool Executes Data Transformations — Not Known
• Tool Handles Null/Missing Data Values — Not Known
• Tool Performs Model Evaluation — Not Known

## Nested Vision3D

http://www.nvss.nb.ca/products/index.html

NVision Software Systems, Inc.
10 Knowledge Park Drive
Suite 110
Fredericton  New Brunswick  Canada E3C 2M7

Hans Galldin
+1 (506) 460-6220
+1 (506) 460-6221 (FAX)
Hans.Galldin@ibm.net;  sales@nvss.nb.ca
http://www.nvss.nb.ca/

Keywords
Data Visualization; ActiveX; Visual Data Mining; Object Oriented User Interface

Platforms
Windows 98; Windows 95; Windows NT

Description
Nested Vision3D is a toolset that enables a user to create 3D interactive visualization models of large complex information structures.  The toolkit leverages Microsoft's COM and ActiveX technologies to allow embedding in any COM-aware application.

Levels of Exploration

• Basic Data Exploration through Visualization — Not Known

• Provides Discovery of Data Patterns — Not Known

• Provides Support for Model Building — Not Known

Levels of Data Mining Process
Extracting Relationships and Data Associations

Data Mining Techniques
Visualization

• Functionality to Clean up Data — Not Known
• Tool Executes Data Transformations — Not Known
• Tool Handles Null/Missing Data Values — Not Known
• Tool Performs Model Evaluation — Not Known

## Nuggets

http://WWW.DATA-MINE.COM/Products.htm

Data Mining Technologies, Inc.
2 Split Rock
Melville NY 11747

Michael Gilman
(516) 692-4400
info@data-mine.com; mgilman@data-mine.com
http://WWW.DATA-MINE.COM

Keywords
Algorithm — Learning/Clustering; Algorithm — Rule-Based; Toolkit; White Papers

Platforms
Windows 95; Windows 98; Windows NT

Description
Nuggets utilizes SiftAgent™ technology to autonomously probe every facet of an organization's data. Nuggets then characterizes the 'behavior' of the data. Nuggets was designed to be used by business analysts, scientists, researchers, and nontechnical people.

Levels of Exploration

• Basic Data Exploration through Visualization — No

• Provides Discovery of Data Patterns — Yes
Extracts patterns in the form of IF-THEN rules with associated probabilities.
Nuggets uses a unique proprietary set of discovery algorithms. These are searching methods that employ meta knowledge gleaned from the search process itself. No statistical assumptions are required and the methods deal naturally with nominal variables (i.e. non-numerical). This insures that all possible patterns can be found unlike tree building methods.

• Provides Support for Model Building — Yes
Models are built in the form of a set of IF-THEN rules in English. These rules can be used by Nuggets to forecast for new data, generalize the patterns for better understanding of the model, and validate the data by using Nuggets supplied functionality on holdout data. An attribute significance report is also supplied.

Levels of Data Mining Process
Basic Data Exploration

Data Mining Techniques

Supervised Induction; Association Discovery; Sift Agents — proprietary non-statistical, non-tree building method.

- Functionality to Clean up Data — Yes
- Tool Executes Data Transformations — No
- Tool Handles Null/Missing Data Values — Yes
- Tool Performs Model Evaluation — Yes

## OMNIDEX

http://sun2.disc.com/products.html

Dynamic Information Systems Corporation
5733 Central Avenue
Boulder CO 80301

Michael Lin
(303) 444-4000
(303) 444-7460 (FAX)
mlin@disc.com; info@disc.com
http://www.disc.com/home

Keywords
Microsoft SQL; Oracle; Sybase; Informix; RMS; C–ISAM; Image/SQL; Flat Files; White Papers; Data Warehouse; Decision Support; Intranet; WWW Interface; SQL; RDBMS (Relational Database Management Systems); ODBC–Compliant; Data Mart; Java; Toolkit; Query Accelerator

Platforms
Windows 95; Windows 98; Windows NT; Hewlett-Packard (HP-UX); Unix (Solaris); IBM (AIX); Unix (Digital Unix); VMS

Description
OMNIDEX is a database search engine that uses advanced indexing to deliver fast answers to complex queries without database tuning. OMNIDEX is designed for large/very large databases where unpredictable queries are common. OMNIDEX performs instantaneous keyword searches; unlimited multidimensional analysis; and high-speed, dynamic aggregations or data summaries.

Levels of Exploration

• Basic Data Exploration through Visualization — Not Known

• Provides Discovery of Data Patterns — Not Known

• Provides Support for Model Building — Not Known

Levels of Data Mining Process
Not Known

Data Mining Techniques
Not Known

• Functionality to Clean up Data — Not Known
• Tool Executes Data Transformations — Not Known
• Tool Handles Null/Missing Data Values — Not Known
• Tool Performs Model Evaluation — Not Known

White Papers and Resources
http://sun2.disc.com/whtpaper.html

## Open Visualization Data Explorer

http://www.research.ibm.com/dx/dxDescription.html

IBM
ibmdx@watson.ibm.com
http://eagle.almaden.ibm.com/dx/

Keywords
Data Visualization; Visual Programming; Open Source; Object Oriented User Interface

Platforms
Not Known

Description
Open Visualization Data Explorer is a full visualization environment that gives users the ability to apply advanced visualization and analysis techniques to their data. These techniques can be applied to help users gain new insights into data from applications in a wide variety of fields including science, engineering, medicine and business. Data Explorer provides a full set of tools for manipulating, transforming, processing, realizing, rendering and animating data and allow for visualization and analysis methods based on points, lines, areas, volumes, images or geometric primitives in any combination. Data Explorer is discipline-independent and adapts to new applications and data. The integrated object-oriented graphical user interface is intuitive to learn and easy to use.

Levels of Exploration

• Basic Data Exploration through Visualization — Not Known

• Provides Discovery of Data Patterns — Not Known

• Provides Support for Model Building — Not Known

Levels of Data Mining Process
Not Known

Data Mining Techniques
Not Known

• Functionality to Clean up Data — Not Known
• Tool Executes Data Transformations — Not Known
• Tool Handles Null/Missing Data Values — Not Known
• Tool Performs Model Evaluation — Not Known

## PolyAnalyst

http://www.megaputer.com/pasystem.html

Megaputer Intelligence, Inc.
1518 E Fairwood Drive
Bloomington IN 47408

Sergei Ananyan
(812) 325-3026
(812)-339-1646 (FAX)
s.ananyan@megaputer.com
http://www.megaputer.com

Keywords
GUI (Graphical User Interface); Data Visualization; Workbench; Algorithm — Rule-Based; Algorithm — K-Nearest Neighbor; Algorithm — Neural Networks; Algorithm — Regression; Algorithm — Genetic; ActiveX; Algorithm — Learning/Clustering; Case Studies; Demo; Evaluation Copy; Tutorial; User Group; White Papers; Wizards; Programmable; C++; Object Oriented User Interface

Platforms
Windows 95; Windows 98; Windows NT

Description
PolyAnalyst is a multi-strategy data mining suite. A broad selection of exploration engines allows the user to predict values of continuous variables, explicitly model complex phenomena, determine the most influential independent variables, solve classification and clustering tasks, and deliver explanations of the found relationships. PolyAnalyst features an object-oriented design, point-and-click GUI, versatile data manipulation, visualization, and reporting capabilities, a minimum of explicit statistics, and a simple interface to various data storage architectures.

Levels of Exploration

• Basic Data Exploration through Visualization — Yes
Visualization including histograms, 2D and 3D charts, snake diagrams, frequency charts, and rule graphs. Snake diagrams provide convenient visualization and comparison of descriptive statistics for compared data sets. Frequency charts provide robust visualization of relative frequencies of categorical and Yes/No variables present in the investigated data set. Rule graphs allow the user to visualize the behavior of multidimensional models. By visually changing values of independent variables involved in the model the user observes the behavior of the model. Interactive visual selection of data points for further exploration from a graph is available.

• Provides Discovery of Data Patterns — Yes
Cluster: Finds the best set of variables for selecting groups of records similar between themselves and very different from the rest of the data, and collects these similar records in a separate data set for further exploration. Find Dependencies ('Liberal' modification): Identifies most important variables and 'outliers' with respect to the selected target variable. Market Basket Analysis: Finds association rules expressing frequently occurring relationships in transactional data. Visual Selection of data from a graph: Provides point-and-click visual selection of data points from graphs for further investigation or presentation. The records corresponding to selected data points are placed in a separate new data set.

Two of PolyAnalyst's data exploration algorithms, Cluster and Market Basket Analysis, are designed for unsupervised learning. Cluster: Localization of anomalies algorithm. Find Dependencies: Comparison of data distributions in equally populated hypercubes in the space of all independent variables. Market Basket Analysis: Association Rules generator for identifying groups of items often co-occurring in the considered transactions, and presenting relations between these items in the form of rules. Statistical analysis utilizing support, confidence, and improvement measures of results.

• Provides Support for Model Building — Yes
Algorithms: Symbolic Knowledge Acquisition: Unique proprietary algorithm for presenting found relations explicitly in a simple but expressive analytical form, with a possible inclusion of IF-THEN statements. This form of presentation is especially appealing to human analysts. Neural Network: GMDH (group method of data handling) hybrid: Automatically builds a self-learning binary hierarchical structure of connected nodes with up to third power polynomial built out of input variables in each node. Fuzzy Logic Classification: Assigns different cases to one of two classes. Models a continuous belonging function with the help of Neural Network, Symbolic Knowledge Acquisition, or Linear Regression, and then selects a threshold for classification that maximizes the number of correct classifications. Memory Based Reasoning: K-Nearest Neighbor algorithm that incorporates selecting the best metric in the space of all independent variables with the help of Genetic Algorithms. Can be used for classification of cases into multiple categories, as well as for predicting numeric variables. Find Dependencies ('Strict' modification): Table based numeric variable prediction tool utilizing comparison of distributions in equally populated hypercubes in the space of all independent variables. Stepwise Linear Regression: Only most influential independent variables are included in the model. Categorical and Yes/No variables are included correctly with the help of IF-THEN statements. Market Basket Analysis: Association Rules generator for identifying groups of items often co-occurring in the considered transactions, and presenting relations between these items in the form of rules. Statistical analysis involving support, confidence, and improvement measures of results.

<u>Levels of Data Mining Process</u>
Extracting Relationships and Data Associations; Model Building

<u>Data Mining Techniques</u>
Supervised Induction; Association Discovery; Sequence Discovery; Clustering; Visualization Symbolic Knowledge Acquisition, Fuzzy Logic Classification, and Memory Based Reasoning.

Algorithms:
Symbolic Knowledge Acquisition Technology
Neural Network – GMDH hybrid
Memory Based Reasoning (K-Nearest Neighbor + Genetic Algorithms)
Clustering (Localization of Anomalies)
Fuzzy Logic Classification — Discrimination
Association Rules
Stepwise Linear Regression
Dependency Detection (n-Dimensional Distribution Analysis)
Descriptive Statistics
Visualization: Histograms, 2D and 3D charts, Snake Diagrams, Frequency charts, Rule graphs

- Functionality to Clean up Data — Yes
- Tool Executes Data Transformations — Yes
- Tool Handles Null/Missing Data Values — Yes
- Tool Performs Model Evaluation — Yes

Future Enhancements
Dedicated Data Import Wizard for easy user controlled import of data held in various storage architectures.

DCOM modules and web-based front end to PolyAnalyst; and full support for visual data mining.

White Papers and Resources
http://www.dmreview.com

## PrudSys Discoverer

http://www.prudsys.com/discoverer

Prudential Systems Software GmbH
c/o Technologiezentrum Chemnitz
Annaberger Str. 240
D-09125 Chemnitz Germany

Elke Bruckner
+49 (0) 3 71 / 53 47 - 1 23
+49 (0) 3 71 / 53 47 - 1 26 (FAX)
info@prudsys.com
http://www.prudsys.com

Keywords
GUI (Graphical User Interface); Knowledge Discovery; Scalable; Algorithm — Decision Tree; Data Extraction; Data Visualization; Decision Support; Flat Files; ODBC–Compliant; Predictive Modeling; White Papers

Platforms
Windows 95; Windows 98; Windows NT

Description
PrudSys Discoverer is the prototype of an intelligent data mining system whose heart is a new method for classification and segmentation of the underlying data (Nonlinear Decision Trees, NDT). PrudSys Discover provides the user the opportunity to apply a wide variety of different models to data.

Levels of Exploration

• Basic Data Exploration through Visualization — Yes
The PrudSys Discoverer offers different methods to data visualization such as box plots, histograms, scatter plots, pie charts and line plots. Additionally, the PrudSys Discoverer offers an extensive collection of statistical measures.

The statistic module can be used for data preprocessing. Cross tables or correlation tables can be displayed for all features of a database. Missing values or outliers can be identified. With the statistical charts one can get a better overview of its data. Simple structures can be immediately made visible.

• Provides Discovery of Data Patterns — No

• Provides Support for Model Building — Yes
PrudSys Discoverer utilizes the universal Non-linear Decision Tree (NDT) algorithm which was developed by Prudential Systems and which covers a width range of advanced classification algorithms. In every node of the tree different discrimination functions can be applied in order to adapt the model optimally to the domain. Especially, this gives rise to an extremely fast handling of large databases.

Levels of Data Mining Process
Model Building

Data Mining Techniques
Supervised Induction; Clustering; Visualization

• Functionality to Clean up Data — Yes
• Tool Executes Data Transformations — Yes
• Tool Handles Null/Missing Data Values — Yes
• Tool Performs Model Evaluation — Yes

Future Enhancements
Import from text-based flat files; statistic module with explorative statistics; visualization module (bar, pie, line, scatter, decision trees); report generator; local remote (ODBC) or native access to databases.

White Papers and Resources
http://www.prudsys.com/de/news/papers/

## S-PLUS Professional

http://www.mathsoft.com/splus/

MathSoft, Inc.
Data Analysis Products Division
1700 Westlake Ave North, Suite 500
Seattle WA 98109-9891

Cheryl Mauer
(800) 569-0123
(206) 283-8802 EXT 264
(206) 283-8691 (FAX)
cmauer@mathsoft.com
http://www.mathsoft.com/

Keywords
Algorithm — Decision Tree; Algorithm — Learning/Clustering; Case Studies; C; C++; Decision Support; Demo; Flat Files; GUI (Graphical User Interface); Knowledge Discovery, Object Oriented User Interface; ODBC–Compliant; Pattern Discovery; Predictive Modeling; Programmable; Toolkit; Tutorial; User Group; Visual Data Mining; Visual Programming; White Papers; Wizards

Platforms
Unix; Linux; Windows 95; Windows NT

Description
S-PLUS Professional is a solution for advanced data analysis, data mining, and statistical modeling. S-PLUS combines an intuitive graphical user interface with an extensible data analysis environment to offer ease of use and flexibility. Import data from virtually any source including ASCII, Excel, SAS, and SPSS. Access powerful statistical functions through convenient menus, toolbars, and dialogs. Statistics include linear and nonlinear regression, generalized linear models, generalized additive models, tree models, smoothing splines, survival analysis, time series, multiple comparisons and more. Customize any function or create your own methods to suit your analysis needs using the S programming language. Choose from over 80 2D and 3D graph types. Drag-and-drop data to create intelligent graphs interactively. Point-and-click to control every detail of your graphs and produce publication-quality output. Change line weights, axes, colors, labels, fonts, symbol types and more. Reveal hidden meanings in complex, multidimensional data with Trellis graphics, exclusively available in S-PLUS. Export graphs to Word and PowerPoint for papers and presentations.

At the core of the S-PLUS System is S, the only language designed specifically for data visualization and exploration, statistical modeling and programming with data. S provides a rich, object-oriented environment designed for interactive data discovery. With a huge library of functions for all aspects of computing with data, S offers unmatched power, flexibility, and extensibility. MathSoft is the exclusive licensee of the core S System technology that is the platform for MathSoft's S-PLUS product line.

Levels of Exploration

• Basic Data Exploration through Visualization — Yes
S-PLUS offers advanced data visualization capabilities including over 80 2D and 3D plot types, brush and spin techniques, and conditioning plots. The S-PLUS plot types include basic scatter and line plots, histograms, box plots, linear and nonlinear regression plots, classification and regression trees and 3D contours, surfaces and point clouds. S-PLUS also offers a unique visualization technique called Trellis graphics. Trellis graphics allow users to see how two variables change with variations in one or more "conditioning" variables. For example, users can analyze sales of a particular product conditioned on geographic regions. S-PLUS graphics are object-oriented and highly interactive. Users can point-and-click on any object on a graph to edit the object. Users can display multiple data variables on the same graph and experiment with different plot types.

• Provides Discovery of Data Patterns — Yes
This tool provides extensive 2D and 3D visualization capabilities; Trellis graphics (conditioning plots); classification and regression trees; cluster analysis; factor analysis; and linear (standard and robust), nonlinear (parametric and non-parametric) regression plots.

• Provides Support for Model Building — Yes
The tool provides over 3,500 data analysis functions including robust and modern methods such as:

a. Predictive Models
   i. Classification — Classification trees, logistic regression
   ii. Regression — (i.e., forecasting continuous values)

      Linear regression, ANOVA, constrained regression, nonlinear regression, robust regression, generalized linear models, generalized additive models,

      linear and nonlinear mixed effect models, parametric and nonparametric survival, regression trees, smoothing, ACE, AVAS, projection pursuit
   iii. Time Series Forecasting — AR, MA, and ARIMA models, classical and

      robust smoothers and filters.
b. Descriptive Models
   i. Clustering — Agglomerative nesting, divisive analysis, fuzzy analysis, monothetic analysis, partitioning arould medoids, k-means, model-based clustering.
   ii. Association discovery, factor analysis, principal components.
   iii. Sequential patterns discovery, autocorrelations and autocovariances, periodograms, ARIMA modeling, quality control charts.
c. Neural net extensions for S-PLUS are freely available on the web.

Levels of Data Mining Process
Extracting Relationships and Data Associations, and Model Building

Data Mining Techniques
Supervised Induction, Association Discovery, Sequence Discovery, Clustering, and Visualization

• Functionality to Clean up Data — Yes
S-PLUS has a full suite of data manipulation functions and dialogs for tasks such as merges, subsetting, and transformation. Computational and graphical summaries are available to detect outliers.

• Tool Executes Data Transformations — Yes

• Tool Handles Null/Missing Data Values — Yes
Missing values are supported as a special data type, with missingness propagated through computations.

• Tool Performs Model Evaluation — Yes

White Papers and Resources
http://www.mathsoft.com/splus/whitepapers/index.htm

## SAS Enterprise Miner

http://www.sas.com/software/components/miner.html

SAS Institute Inc.
SAS Campus Drive
Cary NC 27513-2414

(919) 677-8000
(919) 677-4444 (FAX)
software@sas.com

Keywords
GUI (Graphical User Interface); SEMMA (Sample, Explore, Modify, Model, Assess); Algorithm – – Decision Tree; Algorithm — Learning/Clustering; Algorithm — Neural Networks; Visual Data Mining; White Papers; Data Visualization

Platforms
Not Known

Description
SAS Enterprise Miner is an integrated software product that provides an end-to-end business solution for data mining. A graphical user interface (GUI) provides a user-friendly front-end to the SEMMA (Sample, Explore, Modify, Model, Assess) process. Statistical tools include clustering, decision trees, linear and logistic regression, and neural networks. Data preparation tools include outlier detection, variable transformations, random sampling, and the partitioning of data sets (into train, test, and validate data sets). Advanced visualization tools enable a user to quickly and easily examine large amounts of data in multidimensional histograms, and to graphically compare modeling results.

Levels of Exploration

• Basic Data Exploration through Visualization — Not Known

• Provides Discovery of Data Patterns — Not Known

• Provides Support for Model Building — Not Known

Levels of Data Mining Process
Basic Data Exploration; Extracting Relationships and Data Associations; Model Building

Data Mining Techniques
Clustering; Visualization

• Functionality to Clean up Data — Not Known
• Tool Executes Data Transformations — Not Known
• Tool Handles Null/Missing Data Values — Not Known
• Tool Performs Model Evaluation — Not Known

White Papers and Resources
http://www.sas.com/ads/wp_form.hsql?code=sdm

## Scenario

http://www.cognos.com/scenario/

Cognos, Inc.

Fracois Ajestat
(800) 637-7447
(613) 738-1338 EXT 3264
(613) 738-7402 (FAX)
Marketing.Service.Desk@cognos.com; support@cognos.com
http://www.cognos.com/

Keywords
Case Studies; Decision Support; Tutorial; GUI (Graphical User Interface); Informix; ODBC–Compliant; Oracle; Pattern Discovery; RDBMS (Relational Database Management Systems); SQL; Sybase; Algorithm — Decision Tree; Knowledge Discovery; Visual Data Mining; Predictive Modeling

Platforms
Windows 95; Windows NT; Windows 98

Description
Scenario allows the user to segment and classify their data, quickly identifying the factors and profiles most impacting the business. Users can choose from different views — segment graph, classification tree, data spreadsheet, and literal explain — and multiple strategies to clearly 'see' their results. Scenario 2.0 allows for both continuous and categorical targets, letting users process survey results just as easily as key performance indicators. Users can tailor the interface to include/exclude detailed statistical information, identify data exceptions, and benchmark against alternate data. Most significantly, profiles established in Scenario can be used to dynamically filter impromptu reports, or dynamically create multidimensional cubes for PowerPlay analysis.

Levels of Exploration

• Basic Data Exploration through Visualization — Not Known

• Provides Discovery of Data Patterns — Not Known

• Provides Support for Model Building — Not Known

Levels of Data Mining Process
Basic Data Exploration; Extracting Relationships and Data Associations; Model Building

Data Mining Techniques
Association Discovery

- Functionality to Clean up Data — Not Known
- Tool Executes Data Transformations — Not Known
- Tool Handles Null/Missing Data Values — Not Known
- Tool Performs Model Evaluation — Not Known

## See 5/C5.0

http://www.rulequest.com/see5-info.html

RuleQuest Research Pty Ltd.
30 Athena Avenue
St Ives NSW 2075 Australia

Ross Quinlan
+61 2 9449 6020
+61 2 9440 9272 (FAX)
quinlan@rulequest.com
http://www.rulequest.com/

Keywords
Algorithm — Decision Tree; Tutorial; C; Demo; Evaluation Copy; Algorithm — Rule-Based; Algorithm — Learning/Clustering; Decision Support; Flat Files; GUI (Graphical User Interface); Knowledge Discovery; Case Studies; Predictive Modeling; CRISP-DM (CRoss-Industry Standard Process for Data Mining)

Platforms
Windows 95; Windows 98; Windows NT; Unix (Solaris); Unix (Irix); Unix (Digital Unix); Unix (Linux)

Description
This system constructs classifiers in the form of decision trees and sets of 'if-then' rules. See5/C5.0 has been designed to operate on large databases and incorporates innovations such as boosting. See5 (Windows 95/98/NT) and its Unix counterpart C5.0 are sophisticated data mining tools for discovering patterns that delineate categories, assembling them into classifiers, and using them to make predictions. Public C code is provided to allow classifiers constructed by See5/C5.0 to be used by other applications.

Levels of Exploration

• Basic Data Exploration through Visualization — No

• Provides Discovery of Data Patterns — No

• Provides Support for Model Building — Yes
See5/C5.0 constructs classification models expressed either as classification trees or sets of 'if-then' rules. See5/C5.0 supports boosting — the construction of multiple models using adaptive reweighting of the training cases.

Levels of Data Mining Process
Model Building

Data Mining Techniques
Supervised Induction

- Functionality to Clean up Data — No
- Tool Executes Data Transformations — No
- Tool Handles Null/Missing Data Values — Yes
- Tool Performs Model Evaluation — Yes

Future Enhancements
Import of data via ODBC; attribute relevance filtering; network licensing

## sphinxVision

http://www.asoc.de/main2.html

Associative Computing (ASOC) AG
In der Spöck 10
77656 Offenburg Germany

+49 781 969296 0
+49 781 969298 0 (FAX)
support@asoc.com
http://www.asoc.com

Keywords
Visual Data Mining; White Papers; Fuzzy Logic; GUI (Graphical User Interface); Visual Programming; WWW Interface; Knowledge Discovery; Semantic Model; Semantic Model Application; Demo; Evaluation Copy; Workbench; Data Visualization

Platforms
Silicon Graphics (IRIX)

Description
The sphinxVision product family, KDT-SEM-SEMA, includes the Knowledge Discovery Tool (KDT), Semantic Models (SEM), and Semantic Model Applications (SEMA). KDT employs 'best-of-breed' visualization techniques to provide solutions directly to end users. KDT empowers the user to interact with a three-dimensional representation in order to recognize multidimensional relationships hidden in data. These relationships — in fact, semantic variables — can be named, marked and further refined to deliver increasingly useful knowledge. They can also be exported and embedded in other applications.

Levels of Exploration

• Basic Data Exploration through Visualization — Yes
sphinxVision supplies a plot of correlations contained in single input variables. The variables can be interactively selected ( SGI Version only).

sphinxvision supplies a 2D correlation plot of fields and a 3D bar chart, with which a user can interactively scroll through all datasets of the data to be analyzed. Field selections and ordering are interactively selectable. The 3D display responds to mouse actions over a displayed item showing further details.

• Provides Discovery of Data Patterns — Yes
sphinxVision uses Kohonen based algorithms to cluster data according to their similarity. The results are then visualized in a highly interactive 3D display using a special visualization paradigm typical for sphinxVision. Anomalous patterns and outliers become directly visible.

sphinxVision highlights the most important patterns, if you are able to specify the 'interestingness' measure by an analytical formula.

The pattern discovery algorithms are hierarchies of Kohonen based networks. The topology of this hierarchy is interactively built during the visualization process. Results are displayed using a highly interactive 3D visualization interface containing objects representing data patterns. There is an interactive drill down (down to single record level) using the 3D bar chart feature. Results can be outputted in VRML2.0 JPEG ( images, screens) as an HTML report, or data tables (ASCII, database table).

• Provides Support for Model Building — Yes
The user can interactively select patterns (subspaces of the high-dimensional data space) using the visualizer and save a model of this subspace (also multiple subspaces supplied). Using a special runtime module (SEMBA = semantic business application) a user can load this model and run as classifier: applying a new dataset ( same fields as the data set the model is building from) to this SEMBA. It will output a value that indicates, with which probability this new dataset belongs to the selected subspace, and outputting a confidence value. Building different topologies of these subspace classifiers a user can solve a wide range of problem classes.

Levels of Data Mining Process
Model Building; Extracting Relationships and Data Associations

Data Mining Techniques
Visualization; Association Discovery; Clustering; Other
Other' is the generating of 'knowledge models' which are rather based upon the knowledge of the user than performing a supervised Induction. In addition, 'Other' is the feedback of the result sets of generated models into sphinxVision to perform a mining of the Model performance (Model Mining).

• Functionality to Clean up Data — Yes
• Tool Executes Data Transformations — Yes
• Tool Handles Null/Missing Data Values — Yes
• Tool Performs Model Evaluation — Yes

Future Enhancements
A completely redesigned version with substantially enhanced functionalities and open interfaces is under development.

White Papers and Resources
http://www.asoc.com/d_whitepapers.html

## Spotfire Pro

http://www.spotfire.com

Spotfire, Inc.
One Broadway, Eighth Floor
Cambridge MA 02142

Andy Paquin
(617) 621-0340 EXT 114
(617) 621-0381 (FAX)
andy@spotfire.com; support@spotfire.com
http://www.spotfire.com

Keywords
C++; Decision Support; Demo; Informix; Java; Oracle; Pattern Discovery; SQL; Sybase; Visual Data Mining; White Papers; Scalable

Platforms
Unix; Windows

Description
Spotfire Pro reads large amounts of multivariable data originating from disparate data sources and automatically generates intelligent, interactive query devices for rapid identification of trends, anomalies, outliers and patterns. Spotfire Pro is database independent and extracts data (up to 150,000 records) from commercial data sources such as Oracle, SQL-Server, Informix, and Sybase. Users can interactively query data and immediately receive response presentations as scatterplots, bar charts, and pie charts.

Levels of Exploration

• Basic Data Exploration through Visualization — Not Known

• Provides Discovery of Data Patterns — Not Known

• Provides Support for Model Building — Not Known

Levels of Data Mining Process
Not Known

Data Mining Techniques
Not Known

• Functionality to Clean up Data — Not Known
• Tool Executes Data Transformations — Not Known
• Tool Handles Null/Missing Data Values — Not Known
• Tool Performs Model Evaluation — Not Known

Future Enhancements
Interactive 3D query, full-text search, and web-based solutions.

## SRA KDD Explorer

http://www.knowledgediscovery.com/home/product/toolset.html

SRA International, Inc.
4300 Fair Lakes Court
Fairfax VA 22033

Jim Hayden
(703) 803-1689
(703) 803-1793 (FAX)
jim_hayden@sra.com; david_vennergrund@sra.com
http://www.knowledgediscovery.com/

Keywords
Data Visualization; Java; Scalable; KDD (Knowledge Discovery Database); White Papers

Platforms
Sequent Dynix; Sun; Windows 98; Windows NT

Description
SRA's KDD Explorer includes: high-speed rule and sequence-based algorithms for known pattern detection; multi-strategy data mining algorithms for discovering Associations, Classifications, Sequences, and Clusters; a comprehensive set of Java user interfaces to visualize data mining results for analysis and interpretation; and algorithms that are scalable to take advantage of multiprocessor platforms for rapid analysis of extremely large data sets.

Levels of Exploration

• Basic Data Exploration through Visualization — Yes
Link analysis wheel diagrams which use notes and lines to represent relationships between entities (e.g., loans are linked to properties; buyers are linked to properties; etc.).

Users can select any two attributes available on any entity to plot in the scatter plot tool. Users can move a slider bar to change the thresholds on a link, thus you might only display doctors with more than 30 patients OR doctors with more than 100 claims pending.

• Provides Discovery of Data Patterns — Yes
Association Finder Algorithm finds unusual correlations between attribute value pairs (e.g., 23% of the time an older patient has disease A they also have disease B). It is an unsupervised learning technique. Peer Comparison Algorithm takes a target entity (good client, fraudster, poor performer) and identifies what attributes distinguish it from other peers. It can then be used to find all peers that are similar to this target. We report associations by confidence and support, two measures that reflect how often a relationship holds and how often it is present in a data set. Display formats include trees with pie charts, rules, hierarchical sequences, and plots.

• Provides Support for Model Building — Yes
Decision Tree Algorithm is a classifier that will read a labeled data set and build a predictive model that relates independent attributes to the dependent attribute (a.k.a. labeled outcome).

Levels of Data Mining Process
Basic Data Exploration; Extracting Relationships and Data Associations; Model Building

Data Mining Techniques
Supervised Induction; Association Discovery; Clustering; Visualization; Other
Tool can detect sequences given a pattern to match against, referred to as Sequence Detection.

• Functionality to Clean up Data — Yes
• Tool Executes Data Transformations — Yes
• Tool Handles Null/Missing Data Values — Yes
• Tool Performs Model Evaluation — Yes

Future Enhancements
Peer identification to identify nearest neighbors in n-dimensional space. Useful for finding actors that behave in a manner similar to targeted actors (good clients, fraudsters, etc.)

White Papers and Resources
http://www.knowledgediscovery.com

## Syllogic Data Mining

http://www.syllogic.nl/aboutsyllogic/productdataminingtool.html

Syllogic B.V.
Hoefseweg 1
3821 AE Amersfoort
The Netherlands

+31 33 453 4545
+31 33 453 4550 (FAX)
info@syllogic.com
http://www.syllogic.nl

Keywords
Algorithm — Decision Tree; Algorithm — Rule-Based; Data Visualization; Knowledge Discovery; Pattern Discovery

Platforms
Not Known

Description
Syllogic combines different techniques in a toolbox approach from different fields of database analysis. This toolbox approach enables the user to approach the database from different perspectives. A range of visualization techniques improves insight into the patterns that exist in the database. The latest techniques include Clustering, Visualization, Decision Trees, Association Rules and K-nearest neighbor.

Levels of Exploration

• Basic Data Exploration through Visualization — Not Known

• Provides Discovery of Data Patterns — Not Known

• Provides Support for Model Building — Not Known

Levels of Data Mining Process
Not Known

Data Mining Techniques
Clustering

• Functionality to Clean up Data — Not Known
• Tool Executes Data Transformations — Not Known
• Tool Handles Null/Missing Data Values — Not Known
• Tool Performs Model Evaluation — Not Known

## Synera

http://www.syneracorp.com/

Synera Corporation
4600 South Ulster Street
Suite 700
Denver CO 80237-2882

Jeff Stutz
(303) 846 3054
(303) 486 0848 (FAX)
jstutz@syneracorp.com

Keywords
ActiveX; Evaluation Copy; Java; Knowledge Discovery; ODBC–Compliant; White Papers; Flat Files; Decision Support

Platforms
Not Known

Description
Synera is a knowledge base that enables unlimited access to information without the need for navigation. Synera can be linked to most existing analytical tools in a variety of ways. Synera can perform the initial selects of data from a very large database and extract the result to any third-party application through ODBC or exported flat files.

Levels of Exploration

• Basic Data Exploration through Visualization — Not Known

• Provides Discovery of Data Patterns — Not Known

• Provides Support for Model Building — Not Known

Levels of Data Mining Process
Not Known

Data Mining Techniques
Not Known

• Functionality to Clean up Data — Not Known
• Tool Executes Data Transformations — Not Known
• Tool Handles Null/Missing Data Values — Not Known
• Tool Performs Model Evaluation — Not Known

White Papers and Resources
http://www.syneracorp.com/pages/downloads2.html

## The Easy Reasoner (TER)

http://www.haley.com/framed/TheEasyReasoner.html

The Haley Enterprise, Inc.
1108 Ohio River Boulevard
Sewickley PA 15143

Doug Lauffer
(800) 233-2622
(412) 741-6420
(412) 741-6457 (FAX)
sales@haley.com; info@haley.com
http://www.haley.com

Keywords
Algorithm — Decision Tree; Algorithm — Rule-Based; CBR (Case Based Reasoning); Evaluation Copy; ODBC–Compliant; SQL; Algorithm — K-Nearest Neighbor

Platforms
Unix; Windows

Description
The Easy Reasoner (TER) is a Case-Based Retrieval (CBR) capability for the Eclipse inference engine product, also developed by The Haley Enterprise. TER uses a variety of machine learning techniques including inductive learning of decision trees (or rules) and nearest neighbor classification. CBR can classify new information (a case) and it can retrieve old information based on the conceptual distance between the new information and the existing information (the case base). After the relevant information has been identified and retrieved from the existing case-base, Eclipse rules can leverage a variety of knowledge sources while adapting the retrieved cases to the new situation. Thus, the Case-Based Retrieval of the Easy Reasoner together with the rule-based reasoning capabilities of Eclipse provide a complete Case-Based Reasoning solution.

Levels of Exploration

• Basic Data Exploration through Visualization — Yes
The Easy Reasoner uses proven inductive techniques to construct decision trees that can classify new information based on rules that are statistically and information theoretically discovered across records in a database of classified information. Given a new record, its classification can be determined algorithmically by traversing the decision tree. Similar records that may have a variety of classifications can also be retrieved by traversing a decision tree index constructed for a database. If the query is incomplete, The Easy Reasoner can still complete this form of hierarchical retrieval. Retrieved records can then be considered by rule-based or other types of programs, or in simple cases, directly by end-users of an application to complete the classification of new information. Once classified, a new experience may be easily handled by applying methods used successfully for similar experiences that have

occurred in the past. Not only does The Easy Reasoner automatically learn to classify, it also can retrieve without classification using nearest neighbor techniques. In nearest neighbor retrieval, a weighted distance function measures the distance between a new case (record) and existing cases stored in a database. This 'Query By Similarity' (QBS) is in sharp contrast to the Boolean nature of SQL queries or Query By Example (QBE). In effect, QBS lets you retrieve and rank records based on their distance from a specified point in an N-dimensional 'similarity space'. SQL and QBE only retrieve a set of records that satisfy Boolean constraints; they do not support a notion of distance or ranking. The Easy Reasoner (TER) supports dBase databases. The customer can explore his database with any third party tool that can view a dBase database. The current version of the Easy Reasoner does not have a GUI to support Basic Data Exploration through Visualization, but the API would support the development of such a GUI. The next release of The Easy Reasoner will provide such a GUI.

• Provides Discovery of Data Patterns — No

• Provides Support for Model Building — Yes
Decision Trees and Nearest Neighbor Techniques: The Easy Reasoner is able to build a decision tree for a set of databases, but once again it does not provide a GUI for doing this. However the API would support the development of such a GUI. The next release of The Easy Reasoner will provide such a GUI.

Levels of Data Mining Process
Extracting Relationships and Data Associations; Model Building

Data Mining Techniques
Supervised Induction

• Functionality to Clean up Data — Not Known
• Tool Executes Data Transformations — Not Known
• Tool Handles Null/Missing Data Values — Yes
• Tool Performs Model Evaluation — Not Known

## Viscovery SOMine

http://www.eudaptics.com/somine.htm

Eudaptics Software GmbH
Helferstorferstrasse 5/8
A-1010 Vienna Austria

+43 1 532 0570
+43 1 532 0570 21 (FAX)
office@eudaptics.com
http://www.eudaptics.com

Keywords
Algorithm — Neural Networks; Data Visualization; Demo; GUI (Graphical User Interface); Pattern Discovery; Evaluation Copy; Flat Files; Predictive Modeling; Tutorial; Visual Data Mining; White Papers; Wizards; Algorithm — Self-Organizing Maps

Platforms
Windows 95; Windows 98; Windows NT

Description
Viscovery SOMine is a tool for exploratory data analysis and data mining. Employing Self-Organizing Maps (SOMs), a very robust form of unsupervised neural networks first introduced by Tuevo Kohnen, it puts complex data into order based on its similarity and shows a map from which the features of the data can be identified and evaluated. The result is presented in a track breaking graphical way that allows the user to intuitively discover, analyze, and interpret nonlinear relationships within the data without requiring profound statistical knowledge.

Levels of Exploration

• Basic Data Exploration through Visualization — Yes
Data records are ordered on a map according to their similarity. Each component (variable) is represented by a separate map window, showing the distribution of the respective component over the map. The user may choose between a Self-Organizing Maps (SOM) and a SOM-Ward Clustering and determine significance or the number of clusters for a separate clusters window. The user may draw paths, select nodes and the range of the components displayed.

The visualization technique used is based on Self-Organizing Maps. Individual preprocessing and weight settings for each components may be set. For the training of the map, the number of nodes, the map ratio, and the training schedule may be chosen.

• Provides Discovery of Data Patterns — Yes
The tool provides two different clustering methods: A SOM-clustering and a SOM-Ward clustering.

• Provides Support for Model Building — Yes
Predictive models may be built by association and recall of target components and by a new method combining the ordering of the map with local regressions.

Levels of Data Mining Process
Not Known

Data Mining Techniques
Self-Organizing Maps (SOM)

• Functionality to Clean up Data — Yes
• Tool Executes Data Transformations — Yes
• Tool Handles Null/Missing Data Values — Yes
• Tool Performs Model Evaluation — Yes

Future Enhancements
New clustering method: SOM-Ward; new powerful prediction feature through local regressions; common statistical analysis (principal component analysis, regression, correlation); splitting of nominal variables; automatic priority setting; Active-X interface; input file formats enhanced by SPSS files (*.sav), Business Object Files (*.rep) and others; and many improvements to the GUI some of which include project environment with an improved workflow, and changeable scale in map windows, node selection by selecting a range on the color scale.

White Papers and Resources
http://www.eudaptics.com/whitepaper/viscovery-whitepaper.html

## Visual Insights ADVIZOR

http://www.visualinsights.com/products/

Lucent Technologies
(630) 713-0800
sales@visualinsights.com

Keywords
Data Visualization; Decision Support; Demo

Platforms
Windows 95; Windows 98; Windows NT

Description
Visual Insights ADVIZOR is an interactive data visualization-based decision support application. Visual Insights ADVIZOR enables companies to create and deploy interactive visual analysis applications and application templates. Replacing complex query languages demanded by other decision support applications, interactive data visualization users can easily identify and select specific areas of interest and drill down to transaction level detail.

Levels of Exploration

• Basic Data Exploration through Visualization — Not Known

• Provides Discovery of Data Patterns — Not Known

• Provides Support for Model Building — Not Known

Levels of Data Mining Process
Not Known

Data Mining Techniques
Not Known

• Functionality to Clean up Data — Not Known
• Tool Executes Data Transformations — Not Known
• Tool Handles Null/Missing Data Values — Not Known
• Tool Performs Model Evaluation — Not Known

## VisualMine

http://www.visualmine.com/Datasheet/datasheet.htm

Artificial Intelligence Software (AIS) SpA
Via Calabria 56
00187 Rome Italy

Stefano Tornaghi
+39 06-42874610;
+39 06-42874611 (FAX)
visualmine@ais.it
http://www.visualmine.com/

Keywords
Oracle; Sybase; ODBC–Compliant; Case Studies; RDBMS (Relational Database Management Systems); Visual Data Mining; Algorithm — Rule-Based; Flat Files; Data Visualization; Algorithm — Learning/Clustering; C; C++; Data Extraction; Scalable; Pattern Discovery

Platforms
Windows 95; Windows 98; Windows NT; Unix

Description
Based on the employment of advanced 3D visualization technology, VisualMine delivers visual data mining to the analyst's desktop. 3D visualization enables data analysts to quickly analyze large quantities of information (millions of entries), providing the ability to quickly understand data distributions and detect patterns early.

Levels of Exploration

• Basic Data Exploration through Visualization — Yes
VisualMine enables users to quickly analyze large quantities of information, with the ability to quickly understand data distributions and detect patterns early. This activity usually takes place before the detailed specific analysis, which may employ statistical and other data mining tools. In addition, visualization techniques are a powerful tool to understand the results of other data mining algorithms, such as regression, clustering, and multivariate analysis.

VisualMine combines advanced graphical capabilities with a totally flexible and easy to use management environment. Data can be loaded from files or directly from RDBMSs, they can be interactively selected and mapped to a choice of tens of different visualization metaphors; all this can be done on the fly.

From a general point of view, VisualMine supports Data Mining activity in an fully interactive environment where the user selects one of the available visualization techniques and specifies the variables he wants to visualize. The way users exploit the VisualMine capabilities is highly interactive: a.) users can move, scale and rotate the views; b.) users can access detailed data information by picking the visualized objects; c.) users can change the visualization on the fly:

they can modify the current visualization by selecting different variables or they can select a different visualization technique from those available; d.) users can control the layout of the view by adding labels, annotation, etc.; and e.) users can apply processing modules or filters performing specific actions on data used to build the view.

From a technical point of view, the available visualization techniques belong to three different types: basic 2D graphs; advanced 3D visualizations; and geographical visualizations. Basic 2D graphs implement the well known Business Charts where variables are plotted in a 2D space. VisualMine supports different types of charts where the variables are rendered by using different techniques (scattered points, lines, areas, staircase, stair area).

In general, these kind of views are used to analyze variable correlations but are also useful in discovering outliers or to drill down to a detailed level after having analyzed large datasets by using more advanced 3D representations. Basic 2D graphs allow the users to map one variable on the X axis and one or more variables on the Y axis. Advanced 3D visualizations supports multivariate analysis, cluster analysis, pattern identification and dynamic analysis of a large quantity of data.

VisualMine presents four different types of 3D advanced visualizations: scatter 3D viewer; 3D histogram viewers; pies and torus 3D viewers; and cluster visualizer. The Scatter 3D Viewer is one of the richest and most complete visualization techniques VisualMine offers. It actually allows the visualization of a large number of variables (up to twenty) and it is made of three complementary components: major or primary graph; additional graphs based on virtual cutting planes; and interpolations. The primary graph is a 3D scatter points representation where a set of points is placed; the points are colored and dimensioned in a 3D Cartesian space. The additional graphs are additional copies of 3D scatter points representation placed in the major 3D. Usually they are used to enhance the number of visualized variables. In the additional graphs, the scattered points in the 3D space are actually hidden (in this way they are virtual): but it is possible to project their intersections with a cursor panel moving in the 3D Cartesian space. The Interpolations create geometrical models derived from the spatial distribution of the scattered points. These models can be further visualized using a huge number of special Visualizers, such as isolines, isosurface, and clustering. The Scatter 3D Viewer presents three types of graphical entities: geometric, pictorial and pattern. The variables mapped to the geometrical entities (X Axis, Y Axis and Z Axis) are used to reference points both in the primary and the additional graphs, while the variables linked to the pictorial entities (Point Color and Object Dimension) are used to compute their color and their dimension. 3D histogram viewers (Bars 3D, Planes and Surface) are sophisticated visualizations very useful to display data distribution and to discover the relationships among variables. The basic concept of the 3D histograms visualizations is to populate a 2D grid with a set of objects. Their color and shape are proportioned to the values of two distinct variables, which are associated to the Point Color entity and to the Height (Z Axis) entity. The shape of the bidimensional grid is computed according to the values of two other variables, which are associated to the X Axis and to the Depth (Y Axis) entities in the data mapping. In other words, the values used to build the bidimensional grid (X Axis and Depth entities) define a matrix. Each item of the matrix is characterized with a value of the x variable and with a value of the depth variable. This classification is used to compute the aggregated values of the color and of the height variables

for each item of the matrix. Finally, the matrix will be rendered showing different shapes that depend on the specific visualization techniques. Once the 2D grid is determined, the rendering displays its values using different techniques. In the Bars 3D visualization, the renderer creates a block on each node of the 2D grid. The color and height of the blocks are determined independently by the aggregated values. When the user does not associate any variable to the Point or to the Height entities, the system automatically computes the number of records for every item of the grid. In this case, the color (or the height) of each item of the grid will represent this value. In the Planes visualization. the renderer visualizes ribbons by treating the height values as heights above the surface of the 2D grid, and the color values as color to be applied. The surfaces so created are orthogonal to one axis (selectable as X or Y) and parallel along the other axis. So they look like a set of colored ribbons draping over a surface. Finally, in the Surface visualization, the renderer creates a continuous surface over the 2D grid. The elevation and the color of the surface are computed as described above for the Planes visualization. The Pies and Torus 3D viewers build 3D graphs that visualize a discrete amount of correlated variables, through the replication of 3D pies (or torus). The replication of pies and the use of colors enforce the correlation. Actually, the color entity (which rules the color of each slice of the pies) is shared among the pies. This means that the slices showing the same color are related to the same record of the current dataset. The other two entities (Slice Dimension and Slice Height) are used to calculate the size of each slice and its height. Geographical visualization supports the visualization of data in respect to their geographical context. VisualMine provides two different types of geographical visualization. The first one, the Absolute Geographical viewer, supports data visualization on thematic maps (at different level of geographical granularity). According to the Data Elevation Model visualization technique, VisualMine builds views where every territory in the map is colored and elevated: the color and the height of each territory is computed on the variables specified by the user. The second type of geographical visualization is again a Data Elevation Model, but, in this case, it is aimed to the visual representation of flows. In the views, the flows are depicted as colored arcs where the color and the height of the flows are computed on the variables specified by the user. The basic usage of the Cluster Visualizer should be in the phase following the application of a clustering model, when a black box clustering model exists and identification of each cluster's characteristics is required. Basically, after application of a clustering method, each record data is added with a column specifying which cluster the record belongs to (cluster variable); in this situation, VisualMine will represent data belonging to a specific cluster by using a 3D torus. The system will create a view composed up to four 3D torus. When the number of clusters is greater than four, the VisualMine the user is allowed to specify which cluster/s must be visualized. Optionally, examination variables can be used for cluster analysis. To better understand the composition of each cluster the following information can be retrieved: a.) number of records in the cluster; b.) percentage of records in the cluster on total number of records; and c.) mean values of each variable. Details for each clusters can be obtained by a drill down mechanism, by the Cluster visualizer the user should be able to see cluster characteristics (most important variables of the cluster) and their average values.

• Provides Discovery of Data Patterns — Yes
The interpolation capability builds a model of the visualized data and then shows the model by using sophisticated visualization techniques.

The process is piloted by the user (who decides which are the variables to be interpolated and visualized), so VisualMine is not capable of finding important patterns automatically. Moreover, there is no measure of 'interestingness.'

The algorithm used in data interpolation is the Nearest Neighbor. Then the interpolated fields can be investigated by visualizing isolines, isosurfaces, orthoslices, and segment visualizers.

• Provides Support for Model Building — Yes
The Selector 3D facility supports the interactive definition of regions in the 3D space of a Scatter 3D view. Regions define rules on data and it is possible to save such rules and apply them on a different set of data. Since the process is aimed to perform data clustering, when the model identified by the rules is applied to the new set of data, VisualMine adds a new column tagging the cluster to which each record belongs,

Levels of Data Mining Process
Basic Data Exploration; Extracting Relationships and Data Associations; Model Building

Data Mining Techniques
Visualization; Clustering

• Functionality to Clean up Data — No
• Tool Executes Data Transformations — Yes
• Tool Handles Null/Missing Data Values — Yes
• Tool Performs Model Evaluation — No

Future Enhancements
Wizard interface to support less expert users, automatic data clustering, vertical tools for Internet data analysis, visual environment to perform data manipulation.

## XpertRule Miner

http://www.attar.com/

Attar Software Ltd.
Two Deerfoot Trail
On Partridge Hill
Harvard MA 01451

Robert Keller
(800) 456-3966
(978) 456-3946
(978) 456-8383 (FAX)
rkeller@attar.com; info@attar.com
http://www.attar.com/

Keywords
Case Studies; Evaluation Copy; White Papers; GUI (Graphical User Interface); ActiveX; Intranet; WWW Interface; Visual Data Mining; Programmable; Algorithm — Decision Tree; Algorithm — Learning/Clustering; CRISP-DM (CRoss-Industry Standard Process for Data Mining); Knowledge Discovery; ODBC–Compliant; Pattern Discovery; SQL; Algorithm — Rule-Based; Scalable

Platforms
Windows 95; Windows 98; Windows NT

Description
Using ActiveX technology, the XpertRule Miner client can be deployed in a variety of ways. Solutions can now be built as stand-alone mining systems or embedded in other vertical applications under MS-Windows. Deployment can also be over Intranets or the Internet. The ActiveX Miner client works with Attar's high performance data mining servers to provide multi-tier client-server data mining against very large data bases. Mining can be performed either directly against the data in situ, or by high performance mining against tokenized cache data tables. XpertRule Miner includes extensive data transformation, visualization and reporting features. Data can be manipulated using a drag and drop interface. Users can graphically design their customized data manipulation, mining and reporting processes. Software developers can also directly control the application using the exposed methods and properties of the Miner's objects. This enables Miner to be seamlessly integrated as part of vertical applications, which could have been built in any environment. All this is achieved without compromising scalability or performance.

Levels of Exploration

• Basic Data Exploration through Visualization — Yes
User defined reporting and automated graphs/plots on variables, including: Field vs. Field; Time Series Graphs; Coherence Reports; and Lift Gain Charts.
All of the reporting tools (Field vs. Field, etc.) are fully customizable by the user.

• Provides Discovery of Data Patterns — Yes
a.) Rule induction of graphical decision trees; b.) Discovery of associations between fields; c.) Discovery of clusters of related fields.

The tool uses entropy (information value or content level) of each variable, and field distribution counts and relevancy. a.) Hybrid rule induction with parts from ID3, statistics — handling discrete, continuous data and discrete or continuous outcomes; b.) Association rules, Dynamic Item set Counting (DIC) algorithm; c.) Clustering using Hybrid rule induction and DIC.

• Provides Support for Model Building — Yes
a.) Graphical decision trees; b.) Association rules via English-like rule statements; c.) Clusters via English-like rule statements.

Levels of Data Mining Process
Basic Data Exploration; Extracting Relationships and Data Associations; Model Building

Data Mining Techniques
Supervised Induction; Association Discovery; Clustering; Visualization

• Functionality to Clean up Data — Yes
• Tool Executes Data Transformations — Yes
• Tool Handles Null/Missing Data Values — Yes
• Tool Performs Model Evaluation — Yes

White Papers and Resources
http://www.attar.com/pages/info_xm.htm

# C. Data Mining Resources

This appendix contains descriptions of data mining resources available on the World Wide Web (WWW). Inclusion in this report does not imply endorsement of the resources by the DACS.

Data mining and knowledge discovery in databases are often used synonymously. A standard definition for data mining is the non-trivial extraction of implicit, previously unknown, and potentially useful knowledge from data. Another definition is that data mining is a variety of techniques used to identify nuggets of information or decision-making knowledge in bodies of data, and extracting these in such a way that they can be put to use in areas such as decision support, prediction, forecasting, and estimation. The data is often voluminous but, as it stands, of low value as no direct use can be made of it; it is the hidden information in the data that is useful.

## DACS Data Services

**DACS Data Services Brochure**
http://www.dacs.dtic.mil/about/services/pdf/Data-Brochure.pdf
The DACS gathers software engineering experience data, as well as documented scientific information. This information is stored in computerized databases for easy retrieval. This is a link to a brochure in PDF format, describing these services.

**DACS Product & Services**
http://www.dacs.dtic.mil/about/services/services.shtml
The DoD DACS technical area of focus is software technology and software engineering, in its broadest sense. DACS supports the development, testing, validation, and transitioning of software engineering technology. This page describes in general the services available from the DACS.

## Data Mining & Data Warehousing Service Providers

**Agena**
http://www.agena.co.uk/resources.html
This site has an excellent overview on the application of Bayesian Belief Networks (BBN) in software data analysis and decision making.

**ANGOSS Software Corporation**
http://www.angoss.com/
ANGOSS Software Corporation is an international publisher of software with offices in North America and subsidiaries in Europe. ANGOSS specializes in knowledge engineering, that is the process of gathering business value from data using knowledge discovery tools. Knowledge discovery and data mining technology that turn data resources into actionable information best demonstrate this process. ANGOSS offers a full service approach for businesses requiring assistance in setting up their knowledge discovery and data mining practices, and who may require on-going training and consultation as ANGOSS' technology is rolled out across their organizations. Special tutorials are also offered by ANGOSS.

**AbTech Corporation**
http://www.abtech.com/
AbTech's mission is to enable its direct marketing and data mining customers to realize major benefits and a substantial return on investment by providing them with predictive modeling tools and services. The result of over 50 person-years of research and development, their advanced software delivers predictive data modeling capabilities.

**Catalog of Data Mining Tools and Service Providers**
http://www.dit.csiro.au/~gjw/dataminer/Catalogue.html
This catalog provides pointers to data mining tool vendors and service providers and is maintained by Graham J. Williams of the Commonwealth Scientific and Industrial Research Organisation (CSIRO) in Australia.

**Center for Data Insight (CDI)**
http://insight.cse.nau.edu/
The CDI is a university based applied research center in data mining and knowledge discovery. It is a partnership between Northern Arizona University and KPMG LLP. The CDI lab has an extensive suite of world class data mining tools; as well as an experienced academic and business staff trained in the latest data mining and knowledge discovery technologies.

**Corporate Information Factory & Vendor List**
http://www.datawarehouse.com/factory/
This site presents an interactive view of a conceptual architecture for business intelligence. Take a tour complete with section definitions and links to vendors and solution providers. You may also order printed posters of this concept.

**Data Mart/Data Warehousing**
http://direct.boulder.ibm.com/bi/tech/datamart.htm
An IBM page.

**Information Builders**
http://www.ibi.com/
Information Builders offers software and services that are used in the design of enterprise reporting and decision support systems, data warehouses, cross-platform application development, and integrated application solutions. These software solutions all share a common middleware architecture for enterprise data access.

**KDNuggets™ Directory: Data Mining and Knowledge Discovery Resources**
http://www.kdnuggets.com/
This information rich site offers software tools (siftware), companies, jobs, courses, research projects, reference materials, meetings, and dataset resources.

**Knowledge Discovery Associates**
http://www.knowledge-discovery.com/
The mission of Knowledge Discovery Associates is to add provable value to its clients' operations and processes, by analyzing and discovering relevant business knowledge in client data, and by implementing intelligent data analysis enhancements for existing databases. Knowledge Discovery Associates is a consulting practice based in the Boston area, serving clients nationwide. It is an affiliation of data analysts and expert system implementers, who are experienced in applying knowledge discovery, data mining, and intelligent data analysis to business data in a variety of settings.

**NeoVista Software, Inc.**
http://www.neovista.com/
NeoVista is a provider of software and services that empower business executives and managers to discover relationships and trends in corporate data, then leverage this knowledge to implement strategies that improve profitability and efficiency. NeoVista's Decision Series is an integrated suite of scalable data mining tools that can be assembled into powerful, automated predictive business analysis solutions for decision support. The Decision Series consists of an extensible set of selectable tools that are configured to best suit each computing environment. The Decision Series is highly scalable, providing the flexibility to address the dynamic changes in the size and nature of data or processing needs.

**Partek, Incorporated**
http://www.partek.com/
Partek is a company dedicated to providing their customers with software and services for data analysis and data modeling. They provide a combination of statistical analysis and modeling techniques and modern tools such as neural networks, fuzzy logic, genetic algorithms, and data visualization.

**Pilot Software Data Mining Site**
http://www.pilotsw.com/datamining/index2.htm
Pilot Software offers data mining tools and training in their use. This site also provides a video overviewing data mining.

**SPSS & Data Mining**
http://www.spss.com/datamine/
SPSS Inc. is a multinational company that delivers reporting, analysis and modeling software products. The company's primary markets are marketing research, business analysis/data mining, scientific research and quality improvement analysis. The SPSS mission is to drive the widespread use of statistics. This page provides information about SPSS products and services for data mining.

**Syllogic**
http://www.syllogic.nl/
An internationally operating IT company with knowledge and expertise of system and network management, data warehousing, data mining and groupware.

**The Data Mining Corporation**
http://www.dataminer.demon.co.uk/
This British company provides data mining services. They survey and map data. Based on the form in which the data is held ( SQL database or report text), they select the appropriate reporting tool. In discussion with senior financial management, they design management reports. Finally, they present the necessary data, either on paper, or to the PC desktop of the people who need it.

**The Data Warehouse Institute (TDWI)**
http://www.dw-institute.com/
The Data Warehousing Institute (TDWI) is dedicated to helping organizations increase their understanding and use of business intelligence by educating decision makers and I/S professionals on the proper deployment of data warehousing strategies and technologies. In addition, TDWI helps its membership advance their professional development as data warehousing managers and practitioners. Membership is international, representing more than 40 countries.

**Thinking Machines Corporation**
http://www.think.com/
Thinking Machines Corporation is a provider of knowledge discovery software and services. Darwin, TMC's high-end data mining software suite, enables users to extract meaningful information from large databases — information that reveals hidden patterns, trends, and correlations — and allows them to make predictions that solve business problems. Darwin's power of prediction enables businesses to increase return on investment, expand market share, improve the effectiveness and efficiency of marketing programs, and maximize the quality of their customer service.

**Trajecta**
http://www.trajecta.com/trajecta_home.htm
Trajecta provides products and services for data mining solutions for optimizing sales and marketing decisions.

**Crows Corporation**
http://www.twocrows.com/
Two Crows Corporation is a consulting firm specializing in knowledge discovery: the process of extracting information from data. The newest, hottest technology for knowledge discovery is data mining, which uses sophisticated statistical analysis and modeling techniques to uncover patterns and relationships hidden in organizational databases — patterns that ordinary methods might miss. But data mining is not a stand-alone task. Successful knowledge discovery also includes identifying the problem to be solved, collecting and preparing the right data, interpreting and deploying models, and monitoring the results.

**Yahoo! — Data Mining**
http://www.yahoo.com/Business_and_Economy/Companies/Computers/Software/Databases/Data_Mining/
A directory of companies providing tools and services for data mining.

## Data Mining Education, Training, Courses, and Conferences

**Data General — Introduction to Data Mining (Course No. 17124)**
http://dg.com/education/html/17124.html
This five hour, CD-ROM training provides an introduction to the emerging concept of data mining. It covers the data mining process, its methods, and model.

**KDNuggets™ Directory: Data Mining and Knowledge Discovery Resources**
http://www.kdnuggets.com/
This information rich site offers software tools (siftware), companies, jobs, courses, research projects, reference materials, meetings, and dataset resources.

**Pilot Software Data Mining Site**
http://www.pilotsw.com/datamining/index2.htm
Pilot Software offers data mining tools and training in their use. This site also provides a video overviewing data mining.

**Vision Deep Technology**
http://domino8.visionnet.com/webdev.nsf/pages/datamine.html
The QuickVision Data Mining Workshop is a three-day workshop for a company's business managers, IT professionals, and executives to identify and select potential data mining opportunities for the organization. Data mining experts from the Vision Associates assist company personnel in evaluating potential data mining opportunities and identifying: business value to the organization, data requirements to support the data mining effort, availability and quality of required data, and the estimated cost of solution.

## Data Mining Literature

**Algorithms for Collecting and Analyzing Data for Decision Support (Acolade)**
http://www.cise.ufl.edu/~hyoon/acolade/acolade.html

**An Introduction to Data Mining**
http://www.pilotsw.com/dmpaper/dmindex.htm
A white paper from Pilot Software.

**DM Review**
http://www.dmreview.com/
A monthly publication of Powell Publishing, Inc., covering data warehousing issues and solutions for executives and Information Technology management through columns by top industry experts, informative and timely articles, data warehouse success stories, executive interviews, and third-party product reviews.

**Data Mining Developments Gain Attention**
http://www.kdnuggets.com/press/wt97/
An online version of an article by John Makulowich from Washington Technology Online, a biweekly supplement to the Washington Post.

**Data Mining News**
http://www.idagroup.com/
Published by the Bethesda, Maryland-based Intelligent Data Analysis Group, Data Mining News provides concise, timely news and analysis for tracking the tools, products, and players in the data mining industry. This site provides subscription information and a sample copy of the publication.

**Data Mining People & Papers**
http://cs.bilkent.edu.tr/~fayan/dm_people_papers.html

**Data Mining and Knowledge Discovery**
http://www.wkap.nl/journalhome.htm/1384-5810
A Kluwer journal, intended to be the premier technical publication in the field, providing a resource collecting relevant common methods and techniques and a forum for unifying the diverse constituent research communities. The journal publishes original technical papers in both the research and practice of DMKD, surveys and tutorials of important areas and techniques, and detailed descriptions of significant applications. Short (2–4 pages) application summaries are published in a special section.

**Data Mining and Knowledge Discovery**
http://www.research.microsoft.com/datamine/

**Information Discovery, Inc. — Publications**
http://www.datamining.com/papers.htm
This page provides a collection of published articles on data mining, decision support and data warehousing.

**Journal of Intelligent Information Systems**
http://www.isse.gmu.edu/JIIS/
The mission of the Journal of Intelligent Information Systems (JIIS) is to foster and present research and development results focused on the integration of artificial intelligence and database technologies to create next generation information systems — Intelligent Information Systems. The categories of papers published by JIIS include: research papers, invited papers, meeting , workshop and conference announcements and reports, survey and tutorial articles, and book reviews. Short articles describing open problems or their solutions are also welcome.

**Knowledge Discovery Nuggets**
http://www.kdnuggets.com/nuggets/
A popular electronic newsletter for the data mining and knowledge discovery community, focusing on the latest research and applications. KDNuggets is free and comes out 2–3 times each month.

**Machine Learning Online**
http://mlis.www.wkap.nl/
Kluwer Academic Publishers, the publishers of the journal Machine Learning, make this service available. This site aims to provide readers with one-stop shopping for information related to machine learning. Most of the information is available free of charge, however, access to the full text articles of the Machine Learning Journal is restricted to users with an electronic subscription. A subscription order form, free sample articles, and links to other machine learning resources are provided.

**Machine Learning and Inference Laboratory, George Mason University — Publications**
http://www.mli.gmu.edu/pubs.html

**SIGKDD Explorations**
http://research.microsoft.com/datamine/sigkdd/
This newsletter is a publication of the ACM's Special Interest Group (SIG) on Knowledge Discovery and Data Mining.

**Sandwich Paradigm**
http://www.datamining.com/sandwich.htm
This article introduced a methodology for using data mining to assist data warehouse development. It shows how the warehousing effort can be sandwiched between two layers of data mining to avoid a "toxic data dump" with data structures that will not easily lend themselves to analysis.

**The Business Intelligence and Data Warehousing Glossary**
http://www.sdgcomputing.com/glossary.htm

**The Terminology Rescue Kit Knowledge Discovery in Databases and Data Mining**
http://www.cmis.csiro.au/Graham.Williams/DataMiner/Dictionary.html
With the explosion of interest in data mining and the fact that data mining is the fusion of many disciplines, terminology is being thrown around with gay abandon. Here the author collects terms that are used (some very loosely used) in the context of data mining and knowledge discovery from databases.

**White Papers on Data Warehousing**
http://pwp.starnetinc.com/larryg/whitepap.html
Provides links to many papers discussing data warehousing.

# Data Mining People, Programs and Organizations

## Data Mining Experts

### Biswas, Gautam

http://www.vuse.vanderbilt.edu/~biswas/homepage.html

An associate professor at Vanderbilt University, Dr. Biswas is also an associate director of the Center for Intelligent Systems at the Vanderbilt University School of Engineering. His work in knowledge discovery includes a conceptual clustering system and a knowledge-based equation discovery system that uses clustering techniques in deriving analytical equations for the response variable.

### Fayyad, Usama

http://www.research.microsoft.com/~fayyad/

Realizing that data mining has a big role to play beyond the analysis of science data sets, Usama joined Microsoft in early 1996. He hopes to push the research front of this new and growing field as well as help develop data mining capabilities to make computers easier to use and more effective tools for dealing with the data glut they helped create in the first place. In addition to data mining, Usama's research interests include knowledge discovery in large databases, machine learning, statistical pattern recognition and clustering. Usama received his Ph.D. in computer science and engineering from the University of Michigan, Ann Arbor in 199. In 1994 and 1995, Usama was program co-chair of the International Conference on Knowledge Discovery and Data Mining (KDD). He served as general chair of the KDD-96 conference, is an editor-in-chief of the new technical journal Data Mining and Knowledge Discovery, and co-edited the MIT Press book: Advances in Knowledge Discovery and Data Mining.

### John, George H.

http://robotics.stanford.edu/~gjohn/

George H. John is the Data Mining Guru at Epiphany Marketing Software, a startup company developing enterprise automation software for marketing, where he is developing third-generation data mining technology and applications for marketing. Prior to joining Epiphany, he was a Senior Data Mining Analyst in the Global Business Intelligence Solutions division of IBM. He earned a Ph.D. with Distinction in Teaching from the Computer Science Department of Stanford University, where his research was supported by a National Science Foundation fellowship. His professional work in data mining during the past three years has covered direct marketing, customer retention, loan delinquency, stock selection, and many other business problems; his research in data mining has resulted in many publications covering issues in classification trees, Bayesian models, neural networks, outlier detection, cross-validation, attribute selection, SIPping, and reinforcement learning.

**Langley, Pat**
http://robotics.stanford.edu/users/langley/bio.html

Pat Langley's research focuses on machine learning - the study of algorithms that improve their performance based on experience. Dr. Langley received his Ph.D. from Carnegie Mellon University in 1979. Since then, he has worked in academia (at Carnegie Mellon and the University of California, Irvine), in government (NASA Ames Research Center), and in industry (Siemens Corporate Research). He currently serves as Director of the Institute for the Study of Learning and Expertise (a nonprofit research center), as Head of the Intelligent Systems Laboratory at Daimler-Benz Research and Technology, and as Consulting Professor of Symbolic Systems at Stanford University, where he continues his learning research in the areas of planning, perception, and control.

**ACM Special Interest Group on Knowledge Discovery in Data and Data Mining (SIGKDD)**
http://www.acm.org/sigkdd
The primary focus of the SIGKDD is to provide the premier forum for advancement and adoption of the science of knowledge discovery and data mining.

**Aerospace Data Miner (ADAM)**
http://ai.iit.nrc.ca/IR_public/ADAM/ADAM2.html
The aim of the ADAM project is to develop an easy to use domain specific software system that integrates data mining and monitoring techniques to aid maintenance and operation of commercial aircraft.

**Austrian Research Institute for Artificial Intelligence, Vienna**
http://www.ai.univie.ac.at/oefai/ml/project_descriptions/fwf-data-mining.html

The central goal of this project is to provide solutions to data mining problems by developing an abstract framework that allows one to reason about the data mining process at an appropriate level of detail.

**Center for Data Insight (CDI)**
http://insight.cse.nau.edu/
The CDI is a university based applied research center in data mining and knowledge discovery. It is a partnership between Northern Arizona University and KPMG LLP. The CDI lab has an extensive suite of world class data mining tools; as well as an experienced academic and business staff trained in the latest data mining and knowledge discovery technologies.

**Data Mining Group at CWI**
http://dbs.cwi.nl:8080/cwwwi/owa/cwwwi.print_themes?ID=3
CWI is the National Research Institute for Mathematics and Computer Science in the Netherlands. This page provides information on the people in this group and some of their papers.

**Data Mining Group at the University of Helsinki**
http://www.cs.helsinki.fi/research/pmdm/datamining/
The Data Mining group works at the Department of Computer Science at the University of Helsinki. The Data Mining group is part of the larger From Data to Knowledge group. Their research topic, data mining (or knowledge discovery in databases), is a new research area developing methods and systems for extracting interesting and useful information from large sets of data. Data mining methods can be used in a variety of application areas, such as commercial databases, telecommunication alarm sequences, epidemiological data, etc. The area combines techniques from databases, statistics, and machine learning.

**Data Mining People & Papers**
http://cs.bilkent.edu.tr/~fayan/dm_people_papers.html

**Data Mining Resources (Purdue University)**
http://www.cs.purdue.edu/homes/ayg/CS590D/resources.html
This collection of data mining resources includes links to research groups, tools and systems, and publications.

**George Mason University — Knowledge Discovery in Databases: INLEN**
http://www.mli.gmu.edu/projects/inlen.html
This project is concerned with the development of a large-scale multi-type reasoning system, called INLEN, for extracting knowledge from databases. The system assists a user in discovering general patterns or trends, meaningful relationships, conceptual or numerical regularities or anomalies in large databases.

**Kensington Enterprise Data Mining**
http://ruby.doc.ic.ac.uk/kensington/index.html
The Parallel Computing Research Centre at Imperial College, University of London, is developing the Kensington enterprise data mining system in intensive collaboration with partners from various industrial sectors. Kensington applies advanced distributed object technology to provide a fully scalable system that enables distributed collaborative data mining using high performance servers, including massively parallel computers.

**Kurt Thearling, Ph.D.**
http://www3.shore.net/~kht/index.shtml
This site offers references to data mining white papers, books, and a tutorial.

**QUEST Data Mining Project**
http://www.almaden.ibm.com/cs/quest/
The Quest data mining project at the IBM Almaden Research Center has developed innovative technologies to discover useful patterns in large databases. IBM's technologies include mining for association rules, sequential patterns, classification, and time-series clustering. IBM is making these technologies available through its data mining product, IBM Intelligent Miner.

**Rakesh Agrawal**
http://www.almaden.ibm.com/cs/people/ragrawal
Rakesh Agrawal is a researcher at the IBM Almaden Research Center in San Jose, California. His current research interests include data mining, text and web mining, OLAP, and electronic commerce.

**The CSIRO and ANU Data Mining Group**
http://www.cmis.csiro.au/Graham.Williams/DataMiner/
The Commonwealth Scientific and Industrial Research Organization (CSIRO) is the Australian Government's independent research organization. The Cooperative Research Centre for Advanced Computational Systems (ACSys), a joint venture between CSIRO, the Australian National University (ANU), Digital, Fujitsu, and Sun, provides further focus for the Data Mining Program. Located in the Canberra and Sydney Laboratories the Data Mining Program is performing research, development, and consulting in the area of knowledge discovery in databases, data mining, and analysis of large and complex data sets. The Data Mining Program brings together expertise in machine learning, artificial intelligence, computational statistics, neural networks, evolutionary computation, databases, visualization, classification, and high performance computing. Information is available from this page on research, papers, and staff in the group. Also, information is provided about other data mining sites.

**The Data Mining Group**
http://www.dmg.org
A consortium of industry and academics formed to facilitate the creation of useful standards for the data mining community.

**The Data Mining Institute**
http://www.datamining.org
The Data Mining Institute is a non-profit organization that aims to bring together data mining corporate users and the main players in the market (tool and service providers). Acting both as an observer and a consulting institute, the association has two main aims: to promote data mining best practices and assist users through the different stages of the data mining process.

**The KESO Project**
http://nathan.gmd.de/projects/ml/keso.html
The KESO project (Knowledge Extraction for Statistical Offices) is funded by the European Union's ESPRIT program. The central goal of the project is to construct a versatile, efficient, industrial strength knowledge extraction/data mining system prototype that satisfies the needs of providers of large-scale statistical data.

## Data Mining Related Sites

**Adaptive Automation Resources**
http://www.geocities.com/SiliconValley/Lakes/6007/
This web site categorizes adaptive automation links. This site addresses the following major areas: algorithms, statistics, operations research, graph analysis, expert systems, fuzzy logic, neural networks, and evolutionary computation. It organizes links to information that a broad audience should find understandable and useful within the problem solving technology

continuum of advanced heuristic methods. Here is a place to find FAQs, Newsgroups, Software, Books, Electronic Journals, and Hot Lists. Adaptive automation is an exciting technology filled with opportunity to solve seemingly intractable problems. It is also a powerful tool for developing models of processes and systems that do not yield to traditional analysis or constructive techniques.

**Andy Pryke's Data Mining Resources**
http://www.cs.bham.ac.uk/~anp/sites.html
This site provides a rich collection of data mining resources including general data mining/ knowledge discovery in databases information, data mining research groups/projects and commercial information.

**Data Mining & Knowledge Discovery Resource Page**
http://www.partek.com/html/rsc/dm.html
This site is maintained by Partek, Incorporated.

**Data Mining and Knowledge Discovery Information**
http://www.kdnuggets.com/
This site is a catalog of data mining resources.

**Data Warehouse Information Center**
http://pwp.starnetinc.com/larryg/
Larry Greenfield of LGI Systems Incorporated maintains this site.

**Data Warehouse Survival Kit**
http://www.datawarehouse.com/sigs/survival/

**Data Warehousing Resources**
http://www.credata.com/infopage.htm
This site is maintained by Creative Data, Incorporated.

**KDNuggets™ Directory: Data Mining and Knowledge Discovery Resources**
http://www.kdnuggets.com/
This information rich site offers software tools (siftware), companies, jobs, courses, research projects, reference materials, meetings, and dataset resources.

**Know Directory for Data Mining (Data Sets)**
http://www.iscs.nus.edu.sg/~ngkians1/KDD_Anno/KDD_Data.html

**The Data Mine**
http://www.cs.bham.aac.uk/~anp/TheDataMine.html
The Data Mine provides information about data mining and knowledge discovery in databases (KDD), also known as knowledge acquisition from databases and knowledge discovery.

**University of California, Irvine (UCI) — Knowledge Discovery in Databases Archive**
http://kdd.ics.uci.edu/
This is an online repository of large data sets which encompasses a wide variety of data types, analysis tasks, and application areas. The primary role of this repository is to enable researchers in knowledge discovery and data mining to scale existing and future data analysis algorithms to very large and complex data sets.

**comp.databases.olap**
ws:comp.databases.olap; news:comp.databases.olap
This is an Internet newsgroup devoted to On-line Analytical Processing (OLAP).

## Data Mining Software Tools

**AC2**
http://www.alice-soft.com/products/ac2.html
AC2 is a comprehensive data mining toolkit, aimed at expert users or developers who want to use data mining functionalities with their own interfaces. AC2 is primarily a set of libraries that developers can use to build data mining solutions on the server side.

**ALICE d'ISoft**
http://www.alice-soft.com/products/alicev50.html
ALICE d'ISoft is designed for the nontechnical business user. It explores databases through interactive decision trees and creates queries, reports, charts, and rules for predictive models. ALICE d'ISoft gives business users access to the knowledge hidden in their databases, discovering the trends and relationships in their data and making predictions using that information.

**ANGOSS Software Corporation**
http://www.angoss.com/
ANGOSS Software Corporation is an international publisher of software with offices in North America and subsidiaries in Europe. ANGOSS specializes in knowledge engineering, that is the process of gathering business value from data using knowledge discovery tools. Knowledge discovery and data mining technology that turn data resources into actionable information best demonstrate this process. ANGOSS offers a
full-service approach for businesses requiring assistance in setting up their knowledge discovery and data mining practices, and who may require on-going training and consultation as ANGOSS' technology is rolled out across their organizations. Special tutorials are also offered by ANGOSS.

**AT Sigma Data Chopper**
http://www.atsigma.com/datamining/index.htm
AT Sigma Data Chopper will scan through mountains of data to find significant relationships between variables in a database and display the results in easy to read tables and graphs.

**AVS/Express Visualization Edition**
http://www.avs.com/products/ExpVis/ExpVis.htm
The AVS/Express Visualization Edition offers scientists, researchers, and other technical professionals a comprehensive suite of data visualization and analysis capabilities. It provides end users with state-of-the-art technology for advanced graphics, imaging, data visualization, and presentation. AVS/Express Visualization Edition's visual programming environment makes it easy for users to quickly and interactively visualize their data.

**Aira Data Mining Tool**
http://www.hycones.com.br/
The AIRA Data Mining Tool is a tool able to: discover new, useful and interesting knowledge from databases; generate rules; summarize information; detect irregular behavior in the database; and represent the discovered knowledge in different ways, making it easier to understand.

**AnswerTree**
http://www.spss.com/software/spss/AnswerTree/
AnswerTree offers four powerful algorithms that enable a user to build the best model, for any type of data: CHAID, Exhaustive CHAID, Classification and Regression Trees (C&RT) and QUEST.

**Blue Data Miner**
http://www.bluedatainc.com/bdm.htm
Blue Data Miner is a decision support tool that extracts database tables from reports. Furthermore, it allows executives, managers, and other key personnel to filter and summarize those tables over the Internet via a standard web browser. Blue Data Miner functions as a stand-alone system, or it provides an easy, cost effective, and risk-free complement to the enterprise data warehouse.

**Broadbase EPM (Enterprise Performance Management)**
http://www.broadbase.com/products/broadbaseepm.asp
Broadbase EPM allows a user to: measure what is happening; track performance against key indicators, ensuring that managers can respond to important events, such as deals in danger or escalating service cases; analyze why it is happening; identify process bottlenecks, critical success factors and key trends to recognize present — and future — opportunities and exposures; optimize what will happen; and continuously improve results by identifying adjustments, prioritizing resources, and accelerating decision–making.

**BusinessMiner**
http://www.businessobjects.com/products/enduser_tools/businessminer
BusinessMiner is a data mining tool distributed by Business Objects. BusinessMiner permits the business person to take advantage of powerful data mining technology right on the desktop. BusinessMiner is a data mining product positioned for the mainstream business user. BusinessMiner goes beyond the analytical power of OLAP by automatically detecting patterns in data. Using BusinessMiner, for example, managers can discover the attributes that best determine customer profitability or pinpoint those segments of a population most likely to respond to a given promotion.

**CART (Classification and Regression Trees)**
http://www.salford-systems.com/products.html
CART is a decision tree classification and regression system from the statistical community with many applications to data mining, predictive modeling, and data preprocessing. CART is a robust, easy-to-use decision tree tool that automatically sifts large, complex databases, searching for and isolating significant patterns and relationships. Based on over a decade of research, CART ensures stable performance and reliable results. CART's easy to use GUI, intelligent default settings, and interactive Tree Navigator empower nontechnical users to develop a highly intuitive understanding of their data — to tell a story about what is driving the results and why. For power users, CART's unique advanced features coupled with its batch production mode deliver versatility, speed, and accuracy.

**Caprl**
http://www.mineit.com/
Caprl is a data mining algorithm that discovers different types of sequences in databases from within the Clementine environment. Caprl allows the specification of domain knowledge, such as start and end pages, as well as various time-related constraints. Caprl can handle numeric and symbolic data input values, and is able to produce three different types of sequences.

**Catalog of Data Mining Tools and Service Providers**
http://www.dit.csiro.au/~gjw/dataminer/Catalogue.html
This catalogue provides pointers to data mining tool vendors and service providers. It is maintained by Graham J. Williams of the Commonwealth Scientific and Industrial Research Organisation (CSIRO) in Australia.

**Clementine**
http://www.spss.com/software/clementine
Clementine is a rapid modeling environment that combines enterprise-strength data mining and business knowledge to discover solutions. The powerful visual interface makes data mining an interactive process that invites a user's business expertise at every step in the data mining process, from data access and preparation to models and results.

**Cognos**
http://www.cognos.com/
Cognos is a supplier of business intelligence software, allowing users to extract critical information from corporate data assets through analysis, reporting and forecasting. Cognos products fall into two basic categories: business intelligence tools and 4GL tools. Though the two categories are aimed at different user markets, they share a common purpose: to streamline business processes and increase productivity.

**CrossGraphs**
http://www.belmont.com/cg.html
This tool helps users to visualize large quantities of complex data to discover and report patterns, trends, and relationships spanning many variables and subgroups.

**Cubist**
http://www.rulequest.com/cubist-info.html
Cubist produces rule-based models for numerical prediction. Each rule specifies the conditions under which an associated multivariate linear sub-model should be used. The result — powerful piecewise linear models. Cubist builds rule-based predictive models that output values. Cubist can effectively process data sets containing tens of thousands of records and hundreds of fields (attributes).Public C source code is provided so that models developed by Cubist can be displayed in other applications.

**Darwin**
http://www.think.com/html/products/dartechdatasht.htm#one
Darwin is scalable, enterprise data mining software that helps organizations to rapidly transform large amounts of data into actionable business intelligence. Darwin helps to find meaningful patterns and correlations in corporate data for understanding and prediction of customer behavior.

**Data Detective**
http://www.smr.nl/
The DataDetective data mining tool has a modular design consisting of an associative analysis engine, a graphical user interface and interfaces to common database formats. Several analysis tasks are supported such as normal queries, fuzzy queries (selections defined by soft criteria instead of hard criteria), profile analyses and extensive graphical report facilities. The user always has direct access to the relevant data on which analyses are based.

**Data Miner Software Kit (DMSK)**
http://www.data-miner.com/
The software kit implements the data-mining techniques presented in Predictive Data Mining: A Practical Guide published by Morgan Kaufmann. The software kit has programs for data preparation, data reduction or sampling, and prediction. Of special note are the advanced sampling techniques that can yield near-optimal predictive results for data mining.

**Data Mining Resources (Purdue University)**
http://www.cs.purdue.edu/homes/ayg/CS590D/resources.html
This collection of data mining resources includes links to research groups, tools and systems, and publications.

**Data Mining Software Vendors**
http://www.data-miners.com/products/vendors.html
This page, maintained by Michael Berry and Gordon Linoff, provides links to vendors of many data mining products and reviews of some of these products.

**Data Mining Suite**
http://www.datamining.com/dmsuite.htm
The Data Mining Suite provides a solution for enterprise-wide, large scale decision support. It has the ability to directly mine large multi-table SQL databases. The Data Mining Suite currently consists of these modules: Rule-based Influence Discovery; Dimensional Affinity Discovery; Trend Discovery Module; Incremental Pattern Discovery; Comparative Discovery; and the Predictive Modeler.

**Data SURVEYOR**
http://www.ddi.nl/products/main.html
Data SURVEYOR is an interactive data mining product line for business users and analysts enabling: data mining solutions for business users; an expert toolkit (Expert Suite) for analysts and domain experts providing extensive data mining functionality; and a solution factory for third parties allowing new solutions to be built and maintained easily.

**Data Warehouse Quality (DWQ) Project**
http://www.dbnet.ece.ntua.gr/~dwq/
Data warehousing has become an important strategy to integrate heterogeneous information sources in organizations, and to enable on-line analytic processing. However, the wide variety of product and vendor strategies, combined with a weak understanding of the foundations of data warehousing, make the design and evolution of data warehouses haphazard and prone to failure. The DWQ Project develops techniques and tools to support the rigorous design and operation of data warehouses. The DWQ Project is an
Esprit project.

**DataCruncher**
http://www.rightpoint.com/products/datacruncher.html
Based on the use of data mining technology, DataCruncher automatically sifts through large volumes of customer data to determine the best target audience for each new campaign. With built-in statistics and lift curve analysis, DataCruncher provides pinpoint accuracy for finding market-of-one opportunities. DataCruncher is distributed by DataMind, a developer of enterprise applications for real-time marketing.

**DataMiner 3D**
http://www.dimension5.sk/products/products.htm
The DataMiner 3D family of products provides a flexible data visualization for rapid visual analysis of large multidimensional data sets.

**DataMite**
http://www.lpa.co.uk/dtm.html
DataMite enables rules and knowledge to be discovered in ODBC-compliant relational databases. DataMite is distributed by Logic Programming Associates, a British firm.

**DataScope**
http://www.tiszanet.hu/cygron/contact.htm
DataScope is a data mining tool that enables a user to visually analyze the contents of an arbitrary database and extract the knowledge hidden behind the numbers. Using special visualization techniques that support human thinking and intuition in analyzing the data, a user can easily recognize trends and patterns, or exceptions. From these, a user can simultaneously examine the data from different points of view and query the data with just a mouse click, no special commands or formulas being required.

**Datasage**
http://www.datasage.com/products/products.html
Datasage is a product for production data mining. It automatically and continually analyzes large, complex data sets at high speed to deliver accurate and useful information that would otherwise remain hidden in masses of data. Datasage has a complete architecture for successful enterprise-scale data mining. It is database-centric, scalable and offers a complete set of open APIs. The Database-Centric Architecture allows high-speed, atomic level data mining and directly supports third party data access and visualization tools.

**dbProbe**
http://www.itivity.com/products.html
dbProbe is a business intelligence (OLAP and reporting) tool that combines powerful data analysis and scalability to thousands of users, with simple deployment for administrators (no client software to install). Users can drill down, slice-and-dice, graph, filter, create, and share reports and more. Data sources include MS OLE DB for OLAP, Informix Metacube, and others.

**DecisionWORKS**
http://www.asacorp.com/product/index.html
DecisionWORKS is composed of several tools including dbPROFILE (creates clusters and segmentations of data), ModelMAX (a predictive modeling application), ScorXPRESS (powerful pattern recognition capabilities), and DecisionPOSTM (decision support).

**DeltaMiner**
http://194.152.41.50/english/Delta/Products/Miner/delta_miner.htm
DeltaMiner accelerates typical tasks by imitating the way human experts investigate data. It integrates over 15 predefined analyses such as Navigation, Time Series, Power Search, Cross Table, ABC Analysis, and Meta Search. The analysis techniques are based on a combination of OLAP technology, data mining, and statistical heuristics and methods. DeltaMiner explains deviations, variances, and exceptions. It also detects compensations and helps end-users to navigate through their financial, sales, or web database.

**Dynamic Information Systems Corporation (DISC)**
http://www.disc.com/home
DISC makes and distributes the OMNIDEX database search engine. The OMNIDEX query accelerator provides fast data access for Data Warehousing, Decision Support, OLAP and Web applications with instant, up-front qualifying counts; multidimensional analysis; high-speed data summaries; and fast keyword searches.

**IBM Intelligent Miner for Data**
http://www.software.ibm.com/data/iminer/fordata/index.html
The IBM Intelligent Miner family helps to identify and extract high-value business intelligence from data assets. Through a process of “knowledge discovery,” an organization can leverage hidden information in its data, uncovering associations, patterns, and trends that can lead to competitive advantage.

**IDL (Interactive Data Language) Data Miner**
http://www.rsinc.com/idl/idl_dataminer.cfm
IDL enables in-depth data analysis through visualization. It can be used for cross-platform application development. The optional IDL Data Miner allows a user to connect directly to a database for easy access, query and edit actions from one or multiple databases.

**Information Discovery, Inc.**
http://www.datamining.com/
Information Discovery, Incorporated is a leading provider of large scale data mining oriented decision support software and solutions. All your decision support needs are served with pattern discovery and data mining software, strategic consulting and warehouse architecture design.

**Informix Red Brick Formation**
http://www.redbrick.com/products/formation/vformbtm.htm
Red Brick Formation is a powerful and flexible data extraction and transformation tool that substantially reduces the time and complexity of building and maintaining very large data warehouses and data marts.

**KATE-DataMining**
http://www.acknosoft.com/fTools.html
As a case-based reasoning tool, KATE recalls past experience that is similar to the current problem and adapts the solution that worked in the past in order to solve the current problem. Using induction technology, KATE-DataMining extracts knowledge that is hidden in the data. The tool automatically generates decision trees where the essential information for efficient decision making is presented. The developer may also choose to optimize the time required as well as the cost of decision making and introduce his expertise on the relative importance of the descriptors. He can use several interactive graphical tools to detect hidden dependencies, parameter shifts, and anomalies in the data or predict trends. He can also choose to modify the decision trees by hand. The auto-consultation module is used to test the tree automatically.

**KDD Explorer**
http://www.knowledgediscovery.com/home/product/toolset.html
KDD Explorer is a multi-strategy data mining tool set that discovers relationships, trends, and behaviors hidden in terabyte sized databases. KDD Explorer includes: high-speed rule and sequence-based algorithms for known pattern detection; multi-strategy data mining algorithms for discovering associations, classifications, sequences, and clusters; a comprehensive set of Java user interfaces to visualize data mining results for analysis and interpretation; algorithms that are scalable and paralleled to take advantage of multiprocessor platforms for rapid analysis of extremely large data sets.

**KDNuggets™ Directory: Data Mining and Knowledge Discovery Resources**
http://www.kdnuggets.com/
This information rich site offers software tools (siftware), companies, jobs, courses, research projects, reference materials, meetings, and dataset resources.

**KnowledgeSEEKER**
http://www.angoss.com/ksprod/kspage.htm
At the heart of what makes KnowledgeSEEKER such a powerful and easy-to-use tool is its decision tree Induction process, which in simplified terms acts as an automated query generator, so managers do not have to manually construct the queries. This decision tree Induction process has the mathematical power and crunch power to construct and run the queries required. This process shows the combined dependencies between multiple predictors and the analysis results are presented in highly intuitive colored classification tree. Decision trees allow for effective data visualization and are extraordinarily easy to understand and manipulate. KnowledgeSEEKER findings can also be translated into a knowledge base of rules or a set of executable programming statements.

**KnowledgeSTUDIO**
http://www.angoss.com/kstudio/studio.htm
The KnowledgeSTUDIO product line is a new generation of data mining software from ANGOSS Software Corporation. These new technologies focus on integrating advanced data mining techniques into corporate environments so that business can achieve the maximum benefit from their investment in data.

**MARS**
http://www.salford-systems.com/products.html
MARS is a multivariate non-parametric regression procedure introduced in 1991 by Stanford statistician and physicist, Jerome Friedman. Salford Systems' MARS, based on the original code, has been substantially enhanced with new features and capabilities in exclusive collaboration with Dr. Friedman. MARS eliminates the time consuming, trial-and-error process of building accurate predictive models. MARS automatically finds optimal variable transformations and interactions, the complex data structure that often hides in high-dimensional data. This new-generation approach to regression modeling effectively uncovers business-critical data patterns and relationships that are difficult, if not impossible, for other approaches to uncover.

**MODEL 1**
http://www.unica-usa.com
Model 1 is a suite of targeted data mining solutions. It can be used effectively by both statisticians and modelers. Model 1 produces solid, actionable marketing knowledge that represents the ROI from data collection/storage activities such as data warehousing. Model 1 is composed of four modules which can be used individually to solve specific problems or together to form a complete marketing system.

**MineSet**
http://www.sgi.com/software/mineset/
MineSet is a suite of tools for data mining, visualization, and exploratory data analysis. MineSet is distributed by Silicon Graphics.

**NeoVista Software, Inc.**
http://www.neovista.com/
NeoVista is a provider of software and services that empower business executives and managers to discover relationships and trends in corporate data, then leverage this knowledge to implement strategies that improve profitability and efficiency. NeoVista's Decision Series is an integrated suite of scalable data mining tools that can be assembled into powerful, automated predictive business analysis solutions for decision support. The Decision Series consists of an extensible set of selectable tools that are configured to best suit a given computing environment. The Decision Series is highly scalable, giving an organization the flexibility to address the dynamic changes in the size and nature of its data or processing needs.

**Nested Vision3D**
http://www.nvss.nb.ca/products/index.html
Nested Vision3D was designed as a tool for visualizing networks of information, in particular, but not limited to, object oriented source code. The NestedVision3D system graphically shows information structure as a 3D graph of nodes (3D boxes), and interconnecting arcs. The nodes represent the objects (for code, these are software objects such as files, classes, and variables), and the arcs represent the relationships between the objects.

**Nucleus**
http://www.alterian.com/nucleus.htm
Nucleus is an analysis database that is optimized for analysis, it processes tens of millions of records per second on a standard PC platform. In common with a relational database management system, Nucleus stores the actual value of every field for every record, but it achieves speeds of analysis usually associated with an On Line Analytical Processing Engine (OLAP).

**Nuggets**
http://WWW.DATA-MINE.COM/Products.htm
Nuggets utilizes SiftAgent™ technology to autonomously probe every facet of an organization's data. Nuggets more easily, completely and accurately characterizes the behavior of the data. Nuggets was designed to be used by business analysts, scientists, researchers, and nontechnical personnel.

**OMNIDEX**
http://sun2.disc.com/products.html
OMNIDEX is a database search engine that uses advanced indexing to deliver fast answers to complex queries without database tuning. OMNIDEX is designed for large/very large databases where unpredictable queries are common. OMNIDEX performs instantaneous keyword searches; unlimited multidimensional analysis; and high-speed, dynamic aggregations or data summaries.

**Open Visualization Data Explorer**
http://www.research.ibm.com/dx/dxDescription.html
Open Visualization Data Explorer is a full visualization environment that gives users the ability to apply advanced visualization and analysis techniques to their data. These techniques can be applied to help users gain new insights into data from applications in a wide variety of fields including science, engineering, medicine and business. Data Explorer provides a full set of tools for manipulating, transforming, processing, realizing, rendering and animating data and allows for visualization and analysis methods based on points, lines, areas, volumes, images or geometric primitives in any combination. Data Explorer is discipline-independent and easily adapts to new applications and data. The integrated object-oriented graphical user interface is intuitive to learn and easy to use.

**QueryObject Systems Corporation**
http://www.queryobject.com/
QueryObject-based data marts employ advanced mathematics to create compact and accurate representations of huge data sets. In real-world applications, a QueryObject-based data mart can be tens, hundreds, even thousands of times smaller than the source data, thereby making terabyte-class databases small enough to transport on a conventional PC.

**Red Brick Systems, Incorporated**
http://www.redbrick.com/
Red Brick Systems, Inc. is a provider of data warehousing solutions to help businesses transform data into information for superior decision-making. The company develops and markets a comprehensive, integrated platform of data warehousing products and services. This page also provides accesses to some white papers on data warehousing and data mining.

**S-PLUS**
http://www.mathsoft.com/splus/
The S System, developed by Dr. John M. Chambers of Bell Labs, is a software program pioneered for data visualization and interactive statistical computing. MathSoft is the exclusive licensee of the core S System technology that is the platform for MathSoft's S-PLUS® and Axum® product lines. At the core of the S-PLUS System is S, the only language designed specifically for data visualization and exploration, statistical modeling and programming with data.

**SAS Enterprise Miner**
http://www.sas.com/software/components/miner.html
Enterprise Miner is an integrated software product that provides an end-to-end business solution for data mining. A graphical user interface (GUI) provides a user-friendly front-end to the SEMMA (Sample, Explore, Modify, Model, Assess) process. Statistical tools include clustering, decision trees, linear and logistic regression, and neural networks. Data preparation tools include outlier detection, variable transformations, random sampling, and the partitioning of data sets (into train, test, and validate data sets). Advanced visualization tools enable a user to quickly and easily examine large amounts of data in multidimensional histograms, and to graphically compare modeling results.

**SAS Institute Data Mining Page**
http://www.sas.com/software/data_mining/
SAS Institute develops, markets, and supports data warehousing and decision support software. SAS Institute's enhanced data mining solution offers an integrated environment for businesses that need to conduct comprehensive analyses of customer data. Data mining allows one to explore large quantities of data and discover relationships and patterns that lead to proactive decision making.

**SAS Institute Data Warehouse Page**
http://www.sas.com/software/data_warehouse/
SAS Institute develops, markets, and supports data warehousing and decision support software. The SAS Warehouse Solution enables IT to deliver reliable information for empowering business users to drive the company forward.

**SLP InfoWare**
http://www.slp-infoware.com/
SLP InfoWare provides telecommunications, banking and insurance companies with innovative, data mining-enabled customer retention and churn management solutions.

**sphinxVision**
http://www.asoc.de/main2.html
The sphinxVision product family, KDT-SEM-SEMA, includes the Knowledge Discovery Tool (KDT), Semantic Models (SEM), and Semantic Model Applications (SEMA). KDT employs best-of-breed visualization techniques to provide solutions directly to end users. KDT empowers the user to interact with a three-dimensional representation in order to recognize multidimensional relationships hidden in data. These relationships — in fact, semantic variables — can be named, marked and further refined to deliver increasingly useful knowledge. They can also be exported and embedded in other applications. SEM and SEMA allow ASOC's cutting-edge technologies to be deployed in a cost-effective manner by professional knowledge workers, systems developers and researchers in diverse industrial sectors.

**SPSS & Data Mining**
http://www.spss.com/datamine/
SPSS Inc. is a multinational company that delivers reporting, analysis and modeling software products. The company's primary markets are marketing research, business analysis/data mining, scientific research and quality improvement analysis. The SPSS mission is to drive the widespread use of statistics. This page provides information about SPSS products and services for data mining.

**SRA Knowledge Discovery Solutions**
http://www.knowledgediscovery.com/
This site provides marketing information on software tools and services provided by SRA International, Incorporated. SRA's KDD Explorer discovers novel patterns in terabyte sized databases and presents them for analysis in an intuitive Java interface.

**Scenario**
http://www.cognos.com/scenario/
Scenario allows the user to segment and classify their data, quickly identifying the factors and profiles most impacting the business. Users can choose from different views — segment graph, classification tree, data spreadsheet, and literal explain — and multiple strategies to clearly see their results. Scenario allows for both continuous and categorical targets, letting users process survey results just as easily as key performance indicators. Users can tailor the interface to include/exclude detailed statistical information, identify data exceptions, and benchmark against alternate data. Most significantly, profiles established in Scenario can be used to dynamically filter impromptu reports, or dynamically create multi-dimensional cubes for PowerPlay analysis.

**See 5 / C5.0**
http://www.rulequest.com/see5-info.html
This system constructs classifiers in the form of decision trees and sets of if-then rules. See5/C5.0 has been designed to operate on large databases and incorporates innovations such as boosting. See5 (Windows 95/98/NT) and its Unix counterpart C5.0 are sophisticated data mining tools for discovering patterns that delineate categories, assembling them into classifiers, and using them to make predictions. Public C code is provided to allow classifiers constructed by See5/C5.0 to be used by other applications.

**Siftware: Tools for Data Mining and Knowledge Discovery**
http://www.kdnuggets.com/siftware.html
Links to public-domain and shareware, research prototype, and commercial systems for data mining and knowledge discovery in databases.

**Spotfire Pro**
http://www.spotfire.com
Spotfire Pro reads large amounts of multivariable data originating from disparate data sources and automatically generates intelligent, interactive query devices for rapid identification of trends, anomalies, outliers, and patterns. Spotfire Pro is database independent and extracts data (up to 150,000 records) from commercial data sources such as Oracle, SQL-Server, Informix, and Sybase. Users can interactively query data and immediately receive response presentations as scatterplots, bar charts, and pie charts.

**Sybase's Data Warehouse**
http://www.sybase.com/products/dataware/index.html
This page provides information about data warehouse products available from Sybase, a leading database vendor.

**The Easy Reasoner**
http://www.haley.com/framed/TheEasyReasoner.html
The Easy Reasoner enables decision support systems to resolve problems by remembering solutions to previously encountered problems. It is appropriate for tasks where performance is affected by experience or where knowledge can be acquired from existing or accumulating examples. The Easy Reasoner facilitates diagnosis by recalling previous problems with similar symptoms and can discriminate between these cases so as to resolve the current problem. It uses statistical techniques to automatically discover the conceptual structure within records stored in databases. The resulting decision trees can be used to predict the value for a field given incomplete data.

**Thinking Machines Corporation (TMC)**
http://www.think.com/
Thinking Machines Corporation is a leading provider of knowledge discovery software and services. Darwin, TMC's high-end data mining software suite, enables users to extract meaningful information from large databases — information that reveals hidden patterns, trends, and correlations — and allows them to make predictions that solve business problems. Darwin's power of prediction enables businesses to increase return on investment, expand market share, improve the effectiveness and efficiency of marketing programs, and maximize the quality of their customer service.

**Torrent Systems, Incorporated**
http://www.torrent.com
Torrent provides technology for scalable enterprise computing. Their ORCHESTRATE product family hides the complexity of parallel programming, thereby supporting the creation of parallel, high-performance data-processing solutions. In addition, ORCHESTRATE provides fully parallel software components for building data warehousing and data-mining systems.

**Trajecta**
http://www.trajecta.com/trajecta_home.htm
Trajecta provides products and services for data mining solutions for optimizing sales and marketing decisions.

**Viscovery SOMine**
http://www.eudaptics.com/somine.htm
Viscovery SOMine is a tool for exploratory data analysis and data mining. Employing Self-Organizing Maps (SOMs), a very robust form of unsupervised neural networks first introduced by Tuevo Kohnen, it puts complex data into order based on its similarity and shows a map from which the features of the data can be identified and evaluated. The result is presented in a track breaking graphical way that allows the user to intuitively discover, analyze, and interpret non-linear relationships within the data without requiring profound statistical knowledge.

**Visual Insights ADVIZOR**
http://www.visualinsights.com/products/
Visual Insights ADVIZOR is an interactive data visualization-based decision support application. Visual Insights ADVIZOR enables companies to create and deploy interactive visual analysis applications and application templates. Replacing complex query languages demanded by other decision support applications, interactive data visualization users can easily identify and select specific areas of interest and drill down to transaction level detail.

**VisualMine**
http://www.visualmine.com/Datasheet/datasheet.htm
Based on the employment of advanced 3D-visualization technology, VisualMine delivers visual data mining to the analyst's desktop. 3D visualization enables data analysts to quickly analyze large quantities of information (millions of entries), providing the ability to quickly understand data distributions and detect patterns early.

**XpertRule Miner**
http://www.attar.com/
Using ActiveX technology, the XpertRule Miner client can be deployed in a variety of ways. Solutions can now be built as stand-alone mining systems or embedded in other vertical applications under MS-Windows. Deployment can also be over Intranets or the Internet. The ActiveX Miner client works with Attar's high performance data mining servers to provide multi-tier client-server data mining against very large databases. Mining can be performed either directly against the data in situ, or by high performance mining against tokenized cache data tables. Miner includes extensive data transformation, visualization and reporting features. Data can be manipulated using a drag and drop interface. Users can graphically design their customized data manipulation, mining and reporting processes. Software developers can also directly control the application using the exposed methods and properties of the Miner's objects. This enables Miner to be seamlessly integrated as part of vertical applications, which could have been built in any
environment. All this is achieved without compromising scalability or performance.

**Yahoo! — Data Mining**
http://www.yahoo.com/Business_and_Economy/Companies/Computers/Software/Databases/Data_Mining/
This is a directory of companies providing tools and services for data mining.