Data Mining Industry: Emerging Trends and New Opportunities

By:

Walter Alberto Aldana

B.S. Electrical Engineering and Computer Science
MIT, 2000

Submitted to the Department of Electrical Engineering and Computer Science at the Massachusetts Institute of Technology.

Master of Engineering in Electrical Engineering and Computer Science at the Massachusetts Institute of Technology

May 2000

[June 2000]

Signature of Author:_____
Department of EE & CS
May 19, 2000

Accepted by:_____
Dr. Amar Gupta
Co-Director, "PROFIT" Initiative, Sloan School of Management
Thesis Supervisor

Accepted by:_____
Arthur C. Smith
Chairman, Department Committee on Graduate Theses

Table of Contents:

Acknowledgements:

I would like to thank all the people who kindly gave me the support to continue progressing through many long nights of research and steadfast dedication in providing an in depth study of data mining and its future applications. I would like to thank my family especially my brother Carlos Aldana, who is currently obtaining his PHD in EE at Stanford University, for giving me inspiration to complete this thesis. As my undeclared mentor, he really provided some thoughtful and very constructive feedback in improving my thesis. In addition, I would like to thank my peers for helping me in the revision process of my thesis. In addition, I would like to add a special thank you to my advisor, Dr. Amar Gupta, who through his leadership was able to guide me in the right direction when I needed assistance.

Data Mining Industry: Emerging Trends and New Opportunities

By
Walter Alberto Aldana

Submitted to the Department of Electrical Engineering and Computer Science on May 20, 2000 in Partial Fulfillment of the Requirements for the Degree of Master of Engineering in Electrical Engineering and Computer Science

Abstract:

Data mining is part of the knowledge discovery process that offers a new way to look at data. Data mining consists of the nontrivial extraction of implicit, previously unknown, and potentially useful information from data. It uses machine learning, statistical and visualization techniques to discover and present knowledge in a form that is easily comprehensible to humans [4]. Data mining is then the process of discovering meaningful new correlations, patterns and trends by sifting through vast amounts of data using statistical and mathematical techniques. As Fortune 500 organizations continue to amass substantial quantities of information into their respective databases, data mining can offer the opportunity to learn from this data. Furthermore, current trends indicate that more companies implementing Enterprise Resource Planning systems or contracting with ASP vendors, could further benefit in using data mining techniques. Integrating a data-mining technique alongside these two added value services can proof to be an optimum solution in understanding a company's data. An additional key challenge for the 21st

century data mining companies is the creation of real-time algorithms for web mining analysis in this age of the Internet.

Thesis Supervisor: Dr. Amar Gupta

Title: Co-Founder and Co-Director of the MIT PROFIT Initiative

Introduction:

An important new trend in information technologies is to identify meaningful data collected in information systems. As this knowledge is captured, this can be key to gaining a competitive advantage over competitors in an industry. This then creates an important need for tools to analyze and model these data. The value of data mining is to proactively seek out the trends within an industry and to provide this understanding to organizations that maintain substantial amounts of information. The goal then of data mining is to improve the quality of the interaction between the organization and their customers.

In earlier times, due to lack of existing information systems able to store the data and to analyze them, companies suffered. However, in this day and age, a new pattern of looking into data and extrapolating patterns has come about offered at many levels to organizations. Data mining usually denotes applications, under human control, of low-level data mining methods [82]. Large scale automated search and interpretation of discovered regularities belong to Knowledge Discovery in databases (KDD), but are typically not considered part of data mining. KDD concerns itself with knowledge discovery processes applied to databases. KDD deals with ready data, available in all domains of science and in applied domains of marketing, planning, control, etc. Typically, KDD has to deal with inconclusive data, noisy data, and sparse data [8]. Thus, KDD refers to the overall process of discovering useful knowledge from data while data mining refers to the application of algorithms for extracting patterns from data.

Data mining, if done right, can offer an organization a way to optimize its processing of its business data. In this day and age, new data mining companies are

springing up to the challenge of providing this service. Though data mining is improving the interaction between a business organization using data mining and its customers, there are many data mining companies that are trying to vertically integrate to offer the best services to broad markets. This is done by focusing on a particular industry and trying to understand the types of information collected by companies in that sector. Data mining is then the process of extracting out valid and yet previously unknown information from large databases and using it to make critical business decisions [3]. Data mining or exploratory data analysis with large and complex datasets brings together the wealth of knowledge and research in statistics and machine learning for the task of discovering new snippets of knowledge in very large databases.

Over the last three decades, increasingly large amounts of critical business data have been stored electronically and this volume will continue to increase in the future. Despite the growing volume of data collected by businesses, few have been able to fully capitalize on its value. This is due to the difficult task of fully analyzing these data and discerning the underlying patterns that can emerge. An example of a difficult problem that data mining will try to solve is the following. Suppose a retail company, such as Wal-Mart, which collects large quantities of information from every buyer that comes through the store, wants to investigate the problem of inventory management. Predicting inventory optimization for a large client who sells millions of products is not an easy problem. There are many sub-problems complicated enough to take sufficient time to figure out. One such sub-problem involves understanding and predicting Wal-Mart's customer and consumer preferences. A data-mining tool can be used in this example to discern the subtle patterns in customer behavior to help Wal-Mart stock the proper

amounts of inventory. Since an organization may hold data that can consume many gigabytes or terabytes of storage, data mining can probe through this mass of data and sort out all important pieces of information and present it to a CEO of a client to better understand his client's business structure.

An overview of the evolution of data mining is shown in Table 1. Data mining techniques are the result of a long process of research and product development [3]. This evolution began when business data was first stored on computers, continued with improvements in data access such as more recently generating technologies that allow users to navigate through their data in real time. Data mining is ready for application in the business community because it is supported by three technologies that are mature [3]:

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

Commercial databases are growing at unprecedented rates. A recent META Group survey of data warehouse projects found that 19% of respondents are beyond the 50 gigabyte level, while 59% expect to be there by second quarter of 1996 [3]. In some industries, such as retail, these numbers can be much larger. The accompanying need for improved computational engines can now be met in a cost-effective manner with parallel multiprocessor computer technology. Data mining algorithms embody techniques that have existed for at least 10 years, but have only recently been implemented as mature, reliable, understandable tools that consistently outperform older statistical methods [2]. In the evolution from business data to business information, each new step has built upon the previous one. For example, dynamic data access is critical for drill-through in data

navigation applications, and the ability to store large databases is critical to data mining. From the user's point of view, the four steps listed in Table 1 were revolutionary because they allowed new business questions to be answered accurately and quickly. The core components of data mining technology have been under development for decades, in research areas such as statistics, artificial intelligence, and machine learning. Today, the maturity of these techniques, coupled with high-performance relational database engines and broad data integration efforts, make these technologies practical for current data warehouse environments.

Table 1: Evolution of Data Mining

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

Source: An Introduction to Data Mining. Pilot Software Whitepaper. Pilot Software. 1998.

A continuing trend in the data mining industry is the presence of Enterprise Resource Planning (ERP) vendors and Application Service Providers (ASP). Many large companies (Global 2000) have benefited from implementing ERP systems [31]. ERP systems attempt to integrate all departments and functions across a company onto a single computer system that can serve all those different departments' particular needs. Application Service Provider companies on the other hand, seek to offer similar services as ERP vendors but to smaller organizations helping these companies leverage their data management. These two types of companies, ERP & ASP companies, could gain even greater footing in the marketplace with the addition of providing data mining services. Providing a software tool that can integrate a companies' existing data across many departments with the addition of providing a data-mining tool that is maximized to work most efficiently with that software package could yield significant benefit to client organizations.

Historical Perspective:

From Tools to Solutions:

Data mining has been the subject of many recent articles in business and software magazines. However, just a few short years ago, few people had not even heard of the term data mining. Though data mining is the evolution of a field with a long history, the term itself was only introduced relatively recently, in the 1990s.

The roots of data mining can be traced back along three family lines. The longest of these three is classical statistics. Without statistics, there would be no data mining, as these statistics are the foundation of most technologies on which data mining is founded upon. Classical statistics embrace concepts such as regression analysis, standard distribution, standard deviation, standard variance, cluster analysis, and confidence intervals, all of which are used primarily to study data and data relationships. These are the very building blocks on which more advanced statistical analyses are built upon. Even in today's data mining tools and knowledge discovery techniques, classical statistical analysis still plays a significant role.

The second longest family line for data mining is artificial intelligence, AI. This discipline, which is built upon heuristics as opposed to statistics, attempts to apply human thought-like processing to statistical problems. Because this approach requires vast amounts of computer processing power, it was not practical until the early 1980s, when computers began to offer useful power at reasonable prices [13]. AI found relatively few applications at the very high-end scientific and government markets, and the required supercomputers of the era priced AI out of the reach of virtually everyone else [13]. The

notable exceptions were certain AI concepts that were adopted by some high-end commercial products, such as query optimization modules for Relational Database Management Systems. Over time, this changed, as AI was used to create new ways in addressing and solving very complex and math-driven problems. At the Artificial Intelligence Laboratory at MIT, founded in the 1960s, there is extensive research in many aspects of intelligence. Their aim is two-fold: to understand human intelligence at all levels, including reasoning, perception, language, development, learning, and social levels; and to build useful artifacts based on intelligence.

The third family line of data mining is machine learning, which is more accurately described as the union of statistics and AI. While AI was not a commercial success, and is therefore primarily used as a research tool, its techniques were largely co-opted by machine learning. Machine learning, able to take advantage of the ever-improving price/performance ratios offered by computers of the 1980s and 1990s, found more applications because the entry price was lower than AI. Machine learning could be considered an evolution from AI because it blends AI heuristics with advanced statistical analysis. Machine learning attempts to let computer programs learn about the data they study, such that programs make different decisions based on the characteristics of the studied data, using statistics for fundamental concepts, and adding more advanced AI heuristics and algorithms to achieve its goals.

As such, data mining, in many ways, is fundamentally the adaptation of machine learning techniques to business applications. Data mining is best described as the union of historical and recent developments in statistics, AI, and machine learning. These techniques are used together to study data and find previously hidden trends or patterns

within. Data mining is finding increasing acceptance in science and business areas that need to analyze large amounts of data to discover trends that they could not otherwise find.

1st Generation:

The first generation of what is now called data-mining systems appeared in the 1980s and consisted primarily by research-driven tools focused on single tasks [11]. During this time, there was not much of a need to fully understand multi-dimensional layers of data as one-dimensional analysis tools could solve these tasks. This pre-Internet era consisted of analyzing single problems in a largely maintained database. These tasks included building a classifier using a decision-tree or a neural network tool, finding clusters in data, and data visualization (described later) [11]. These tools addressed a generic data-analysis problem and their intended user needed to be technically sophisticated to understand and interpret the results [11]. Furthermore, using more than one of these tools was very complicated and involved significant data and metadata transformations, which was not an easy task to achieve even for expert users.

2nd Generation:

Data-mining vendors developed the second-generation data-mining systems called suites around 1995 [11]. These second generation tools are driven largely by the realization that the knowledge discovery process requires multiple types of data analysis with most of the effort spent in data cleaning and preprocessing [11]. This discovery process generally consists of discovering patterns in data [82]. The suites, such as SPSS's Clementine, Silicon Graphics ' Mineset, IBM's Intelligent Miner, let the user

perform several discovery tasks (usually classification, clustering, and visualization) and support data transformation and visualization.

3rd Generation:

Despite the goals 2nd generation systems tried to solve, the problem with them lies in that business users cannot use these systems directly. Instead, they require significant statistical theory to be able to support multiple discovery tasks. This implies that in order to extract hidden patterns of information, substantial amounts of time need to be spent to understand what type of algorithm should be used and how it should be applied to generate useful results. As such, the third generation of systems came about. These third generation came about as a result of business users' needs resulting in vertical data-mining based applications and solutions in the 1990s [11]. These tools were primarily driven to solve specific business problems such as predicting future customer purchases or inventory optimization for a specific organization. This knowledge discovery process was done by sifting through piles of information stored in large databases to discover hidden patterns. The end results were pushed to front-end applications such as a decision support system to allow the business user to determine the strategy based to the specific problem the data mining tool was supposed to address and abstract away the data mining tool specifics.

Data mining vertical applications were developed to provide a high payoff for the correct decisions. These applications often solve the most popular and at the same time, the most critical business problems for managers. This implies that despite the results generated from applications, it is up to the manager of an organization to understand the data presented to him or her and base important and critical business decisions on this

data. Furthermore, with the Internet changing the ways in which well-established corporations battle with the competition, data mining offers new potential to client organizations. Now these established corporations equipped with vast amounts of information can value the rewards provided by a data-mining tool. These corporations largely having accumulated several orders of magnitude of data can now apply these data to mining tools helping corporations determine optimal decisions.

The Internet has infused many new types of data to be at the forefront of database storage. With increasing numbers, digital libraries are being used to store data such as voice, text, video, and images [88]. Web data mining or web mining presents a new set of challenges for data mining companies. Analyzing the click stream data found in customer logs to determine in real time the right advertisement or offer for a particular customer is a new problem faced by data mining companies [11]. To determine in real time what the right customer pop-up menu should be several technologies are becoming the forefront of providing this service to corporations. Researchers are developing new ways to make predictions for web sites. Collaborative filtering was originally developed at MIT and implemented in systems such as Firefly Network (later acquired by Microsoft in 1998) and NetPerceptions [11] are currently used on web sites in trying to predict what future customer purchasing patterns. Collaborative filtering is based on the premise that people looking for information should be able to make use of what others have already found and evaluated. As such, these systems make use of information on items purchased or selected to predict what future customers may want to purchase. Current collaborative filtering systems provide tools for readers to filter documents based on

aggregated ratings over a changing group of readers. Further analysis and discussion will follow in later sections.

Knowledge Discovery Process:

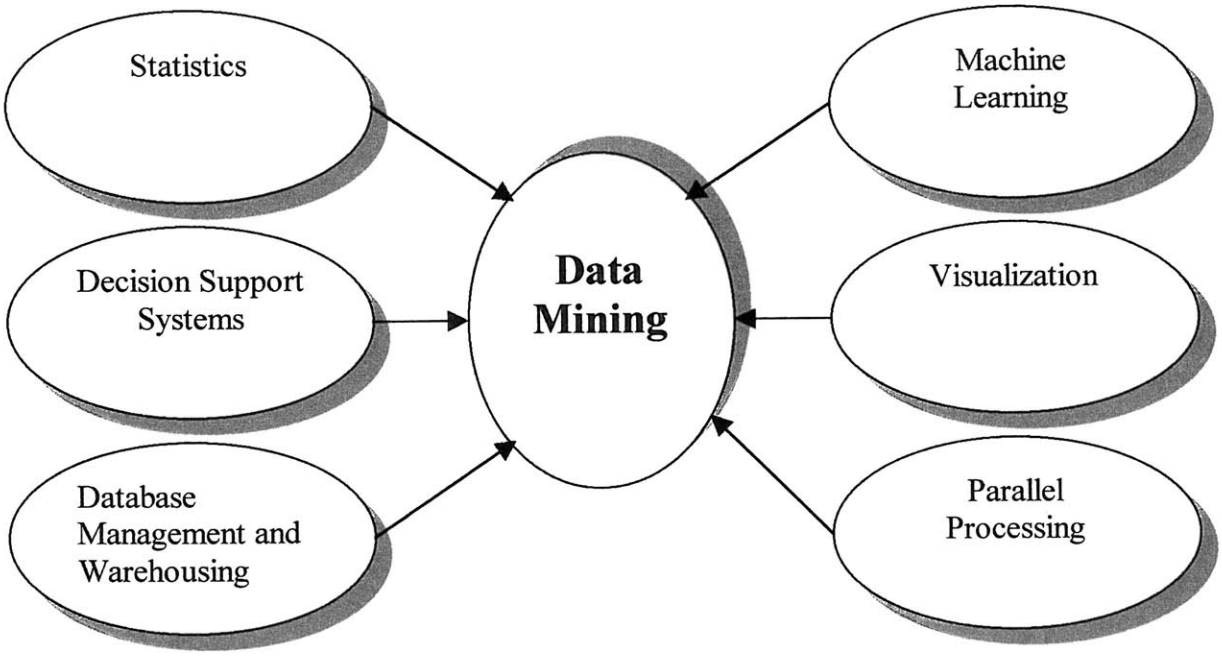
Data mining is part of a larger iterative process called knowledge discovery. The following summarizes the steps of the knowledge discovery process.

- Define the Problem. This initial step involves understanding the problem and figuring out what the goals and expectations are of the project.
- Collect, clean, and prepare the data. This requires figuring out what data are needed, which data are most important and integrating the information. This step requires considerable effort, as much as 70% of the total data mining effort [14].
- Data mining. This model-building step involves selecting data mining tools, transforming the data if the tool requires it, generating samples for training and testing the model, and finally using the tools to build and select a model.
- Validate the models. Test the model to ensure that it is producing accurate and adequate results.
- Monitor the model. Monitoring a model is necessary as with passing time, it will be necessary to revalidate the model to ensure that it is still meeting requirements. A model that works today may not work tomorrow and it is therefore necessary to monitor the behavior of the model to ensure it is meeting performance standards.

Data Mining Supporting Technologies Overview:

Data mining is an integration of multiple technologies as shown in Figure 1 [82]. Each of these technologies: Statistics, Decision Support Systems, Database Management and Warehousing, Machine Learning, Visualization, and Parallel Processing are all tools that interact and support a data mining tool [83]. Statistics and Machine Learning continue to be developed for more sophisticated statistical techniques [83]. Decision Support Systems, which are discussed in a separate section in greater detail, are a collection of tools and processes to help managers make decisions and guide them in management [84]. For example, tools used for scheduling meetings, organizing events, spreadsheets graph tools, and performance evaluation tools are example of support systems. Visualization techniques, also described in greater detail in the data mining techniques section, are used to aid the data mining process. Researchers from the computer visualization field are approaching data mining in a different way: present the data miner an interactive data-mining tool. Database management and data warehouses aid the data mining process in integrating various data sources and organizing the data in such a way so that this data can be effectively mined. Finally, achieving scalability of data mining algorithms is of primary concern. Techniques that use parallel processing is another key technology-supporting data mining progress.

Figure 1: Data Mining Technologies



Data Mining: Verification vs. Discovery

Data mining is not simply a query extraction verifying analysis tool. Decision support systems (DSS), executive information systems, and query writing tools are used mainly to produce reports about data. In this matter, these query tools serve mainly to access the records already existing in large databases. After the data are extracted, they are examined to detect the existence of patterns or other useful information that can be used in answering questions of interest. This kind of data extraction methodology that finds trends in existing data is called the verification method. Under this scheme the data miner must hypothesize the existence of information of interest, convert the hypothesis to a query, pose it to the warehouse, and interpret the returned results with respect to the decision being made [1]. The organization using this methodology will continually reissue hypotheses using query tools that support or negate it. This is a very systematic

approach in finding a potential solution to the problem posed. Little information is created using this retrieval approach [1].

The approach used by data mining is different. Instead, data mining uses a discovery methodology whereby based on the method used, this technology will try to detect trends and produce results about the data with very little guidance from the user. This technology is designed to find the most significant sources of data and draw conclusions from the data it selectively sifts through. In data mining, large amounts of data are inspected; facts are discovered and brought to attention of the person doing the mining. As such, data mining is a discovery tool where a more efficient mode of seeking relevant data and bringing forth this information to the data miner is accomplished.

Decision Support Systems:

A Decision Support System (DSS) is often interfaced with a data-mining tool to help executives make more informed decisions (See Figure 1). Though there are a variety of Decision Support Systems in the market today, their applications consist mostly of synthesizing the data to executives so that they can make more objective decisions based on the data analyzed. DSS technologies have sprung up since the mid 1980's and there has been a continuing trend to improve the analyzing tools of DSS. In this day and age of the Internet, on-line analytical processing or OLAP is slowly replacing aging support systems. Essentially if a computerized system is not an on-line transaction processing system (OLTP), then the system is regarded as a DSS [5]. Increasingly, OLAP and multi-dimensional analysis is used for decision support systems to find information from databases [85]. Executive Information Services (EIS), geographic information systems (GIS), OLAP, and knowledge discovery systems can all be classified into the category of systems referred to as DSS [5]. Two major categories of DSS include enterprise-wide and desktop. In enterprise-wide, the DSS is linked to a large data warehouse and usually serves many managers within an organization. This large infrastructure allows managers to access information quickly. In enterprise-wide DSS, the most sophisticated enterprise-wide analysis systems provide access to a series of decision-oriented databases or data marts, predefined models and charts, and instantaneous access to variables in the corporate data warehouse [5]. Furthermore, analytical tools such as data mining can manipulate the data further to help managers make more informed decisions. In contrast to enterprise-wide, desktop refers to a DSS

that is mostly used by a single manager. In most organizations today, there is a constant flow of communication between enterprise-wide, data warehouses, and desktop DSS [84]. There are a variety of DSS architectures that an organization can implement. For example, a single enterprise-wide DSS can exist to handle all the data and operations flow of an organization or can have multiple layers with other decision support systems such as a desktop DSS that resides on the desktop of a single user. A client-server architecture can further transmit information flow between client desktop and associated DSS tools that reside on a server.

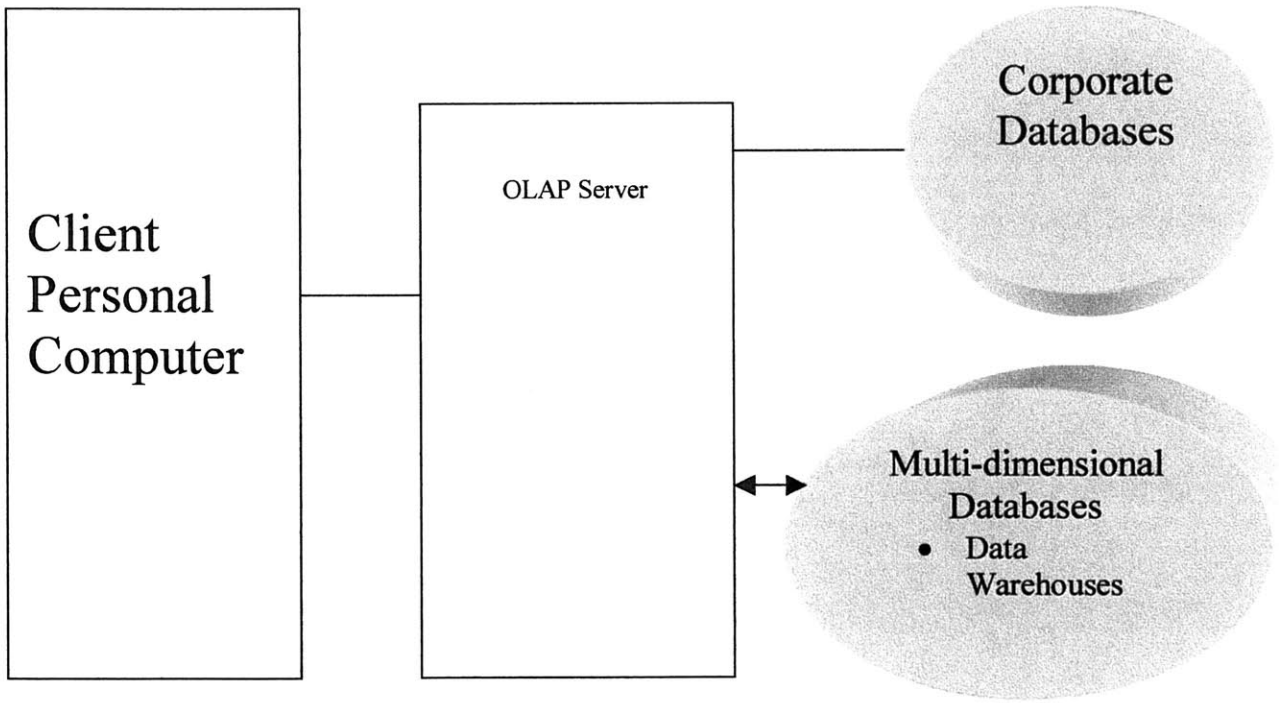
OLAP:

OLAP is a technology that enables the building of a multidimensional data cube from information stored in relational and tabular 2D databases [29]. Users can use the data to answer real-world complex business questions. On-line Analytical Processing (OLAP) is a software technology that enables fast analysis of shared multidimensional information [85]. Multi-dimensional data are data organized to allow for viewing and comparing across multiple dimensions simultaneously—as distinct from the two dimensions (horizontal and vertical) of a spreadsheet (See Appendix) [30]. While a spreadsheet allows a user to compare data in just two dimensions, multidimensional structures enable an almost infinite number of views and associations. Multidimensional analysis is necessary to obtain fast answers to such complex organizational questions as "What is the best-selling product line in North America during Q4, 1995, with a profit margin of at least 20 percent, and selling through the indirect sales channel?"

OLAP enables better decision making by giving business users quick, unlimited views of multiple relationships in large quantities of summarized data. This can result in

high performance access to large amounts of summarized data for multidimensional analysis. With OLAP, managers and analysts can rapidly and easily examine key performance data and perform powerful comparison and trend analyses, even on very large data volumes. Data comparisons can be used in a wide variety of business areas, including sales and marketing analysis, financial reporting, quality tracking, profitability analysis, manpower and pricing applications, and many others. OLAP uses data warehousing techniques to create a repository of information that is built from data from an organization's systems of enterprise-wide computing (see Figure 2). Regardless of where the data reside, they are accessible to requesting applications on any supported platform anywhere on the network, including Web-based applications.

Figure 2: Overview of OLAP



Desktop DSS:

Desktop single-user DSS are not as popular as enterprise-wide systems because they do not allow for multiple users to link to large data warehouses for information. However, desktop DSS are indeed useful. An individual user can analyze information using Microsoft Excel, Lotus 1 2 3, or some other specific DSS applications for individual managers. Expert Choice is an example of a specialized software Windows package that serves as a desktop DSS. Expert Choice is a software tool mainly used to support decision situations through an analytical hierarchical model consisting of a goal, possible scenarios, criteria, and alternatives. E.g., an analyst at Goldman Sachs using Excel to model a financial problem can present his/her findings to managers at large as programmed components in enterprise-wide DSS. Analysts can then proceed to conduct the analysis and once done can disclose their findings on the company's intranet.

In general, DSS has a well-defined purpose whether it is strategic or operational decision to be solved. The tools that are available for retrieving and analyzing data, the database, the variables included, and the time series of the data available determine the questions that can be posed and the decision-relevant information that can be generated [5]. In the long run, however, DSS can help managers retrieve, summarize, and analyze decision relevant data to make wiser and more informed decisions.

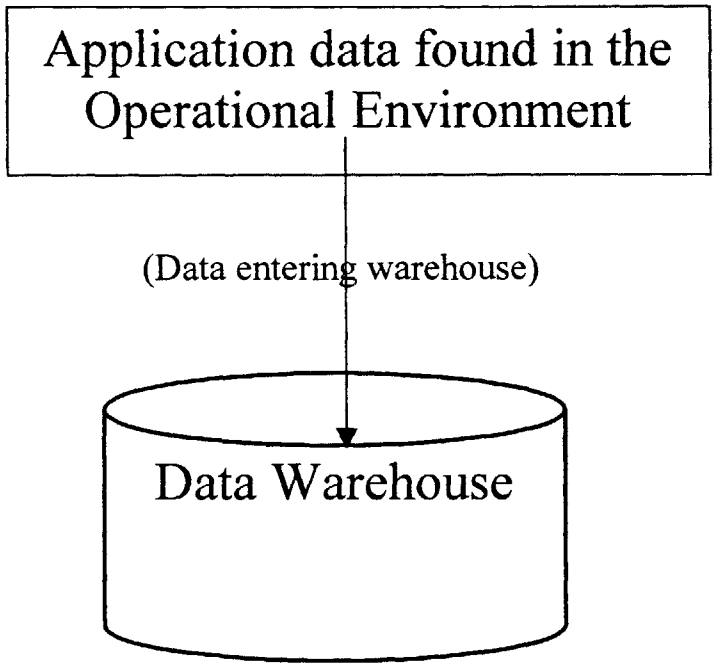
Data Warehouse:

Data mining and DSS are two kinds of applications that can make use of the data warehouse. Through a DSS alone a manager can make much more informed decisions. However, by interacting with a data warehouse a DSS can make better use of quality information and interpreting the information in a more efficient manner. Essentially, a data warehouse is a central repository for all or significant parts of the data that an enterprise or organization's business systems collect. Typically, a data warehouse is housed on an enterprise mainframe server. Data from various on-line transaction processing (OLTP) applications and other sources are selectively extracted and organized on the data warehouse database for use by analytical applications such as data mining tools. Thus, it is used for supporting the informational processing for organizations needing data management by providing a platform of integrated, historical data from which to do analysis [6].

Four Main Characteristics of a Warehouse:

A data warehouse consists of the following four main characteristics:

- Subject-oriented
- Integrated
- Time-Variant
- Nonvolatile



Data are entered into the data warehouse from the operational environment. That is the data warehouse is always logically a separate storage of data transformed from the application data found in the operational environment [6]. One of the primary features of the data warehouse is that it is subject-oriented. As an example, the data warehouse could be structured around an organization's customers, its products, or its R & D development. In contrast, the operational system can be mostly organized around applications and functions of a subject area. The main difference in the structuring of these two systems is that a data warehouse will exclude all information not used by a DSS tool [6]. However, an operational system will include all data regardless of whether or not a DSS tool will use the data.

Another distinguishing feature of a data warehouse is that the data contained within the bounds of a warehouse are consistent. For example, instead of allowing many date formats, one particular dating format will be used. This consistency will abound for all data entering into a warehouse; all variables, naming schemes, and encoding structures will all follow a particular pre-determined convention. For the DSS analyst studying the data, his or her concern will be on using the data in the warehouse as opposed to wondering about the consistency and credibility of the data [6].

The third characteristic of a data warehouse is that of time-variance. In the operational environment, data are expected to be accurate at the moment of access. In a data warehouse, however, data are accurate at some moments in time. This difference lies in the underlying fact that the time horizons for the operational vs. the warehouse environment vary significantly. In a data warehouse, data are kept over long periods of time, usually for a span of several years. In contrast, for the operational environment, data are maintained for a span of months [6]. Thus, applications in the operational environment must carry a high degree of flexibility as data are continuously updated. In a warehouse environment, vast numbers of records are maintained making it very difficult to update the data warehouse as the need arises.

Lastly, the fourth interesting characteristic of a warehouse is that it is nonvolatile. Whereas changes are updated regularly in an operational environment, in a data warehouse once data are loaded no further updates to the data can occur. Needing to support record-by-record house keeping, as is the case with the operational environment through backup and recovery, transaction and data integrity, and the detection of deadlock is no longer needed [86].

As such, these four underlying characteristics allow for an environment that is significantly different from an operational environment. Note however, that all the data that enter a data warehouse comes from the operational environment. The data warehouse functions as a tool that transforms these data into more a more useful way of analyzing and synthesizing the data. The maintenance of a warehouse and the simplicity of design allow for the data warehouse to be a significant factor in the relaying of higher quality information to data mining tools. Being subject-oriented, integrated, time-variant, and nonvolatile are all significant contributors to maintaining consistent data for other systems to interact with.

The Data Mining Process:

The goal of identifying and utilizing information hidden in data has three requirements [1]:

- The captured data must be integrated into organization-wide views instead of specific views.
- The information contained in the integrated data must be extracted.
- The obtained information must be organized in ways that enable decision-making.

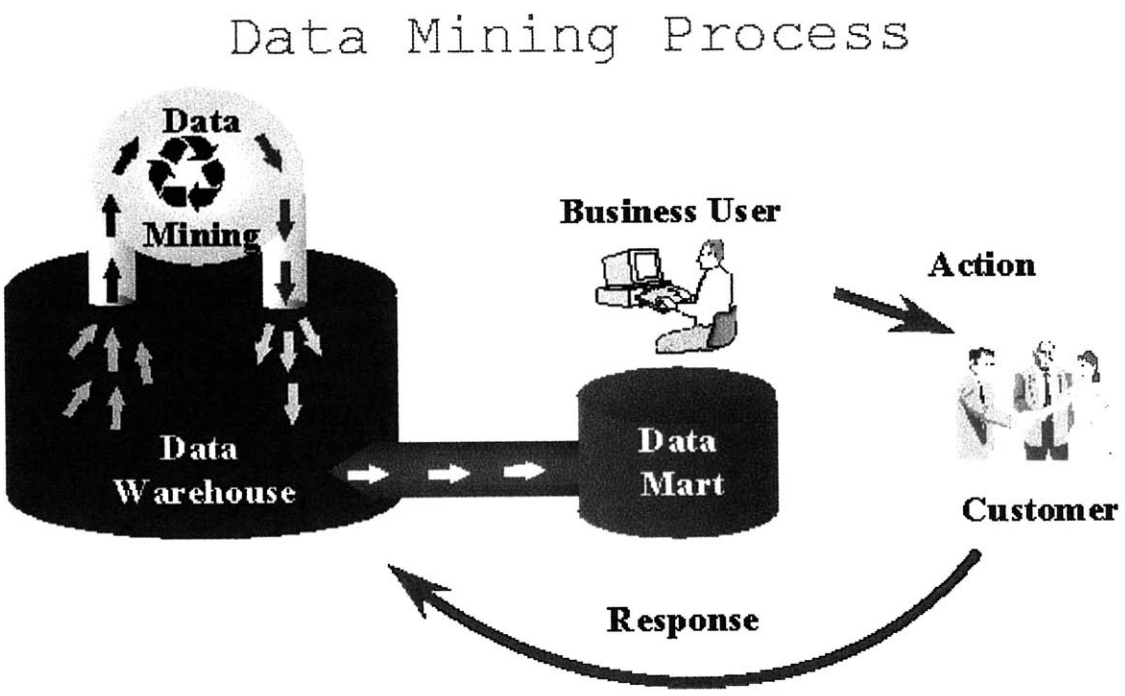
The data mining process can be classified as going through a series of four steps. This consists of transforming the already summarized data found in a data warehouse into information than can produce useful results. These four steps can be summarized into [1]:

- Data selection
- Data transformation
- Mining the Data
- Interpretation of Results

Data selection consists of gathering the data for analysis. Data transformation will then convert appropriate data to a particular format. Data mining will then extract the desired type of information yielding in results to be interpreted. In Figure 3, the data-mining tool will extract the relevant information from the data warehouse environment. In order for the data-mining tool to work, the sub-processes of data selection and transformation must take place prior to data mining. The results are then passed to a decision-oriented databases or data mart, where the user can make a recommendation

based on the results and put the recommendations into action. Of course this assumes that all of the four steps will be successfully completed, which is not always the case.

Figure 3: **Data Mining Process**



Data selection can be the most important step in the process. This is due to the complexity in finding and constructing pre-selection criteria before the extraction of data actually transpires. The variables selected and the range of each of them should be determined in this step. For example, a marketing executive wishing to improve sales figures will pre-select those customers that have been most active in making purchases and observe their behavior. An executive can mine all the data, but this can turn out to be a very costly operation because the data-mining tool will have to search through all this data and moreover if results are generated, they have more risk in predicting an optimal recommendation. Carefully choosing the data is therefore a very important step.

Once the data to be mined has been chosen the next step in the data mining process usually consists of transforming the data into the particular formats needed by the data-mining tool. Data are further synthesized by computing certain rations and applying algorithms to convert the data to a particular type suitable for future applied tools [1].

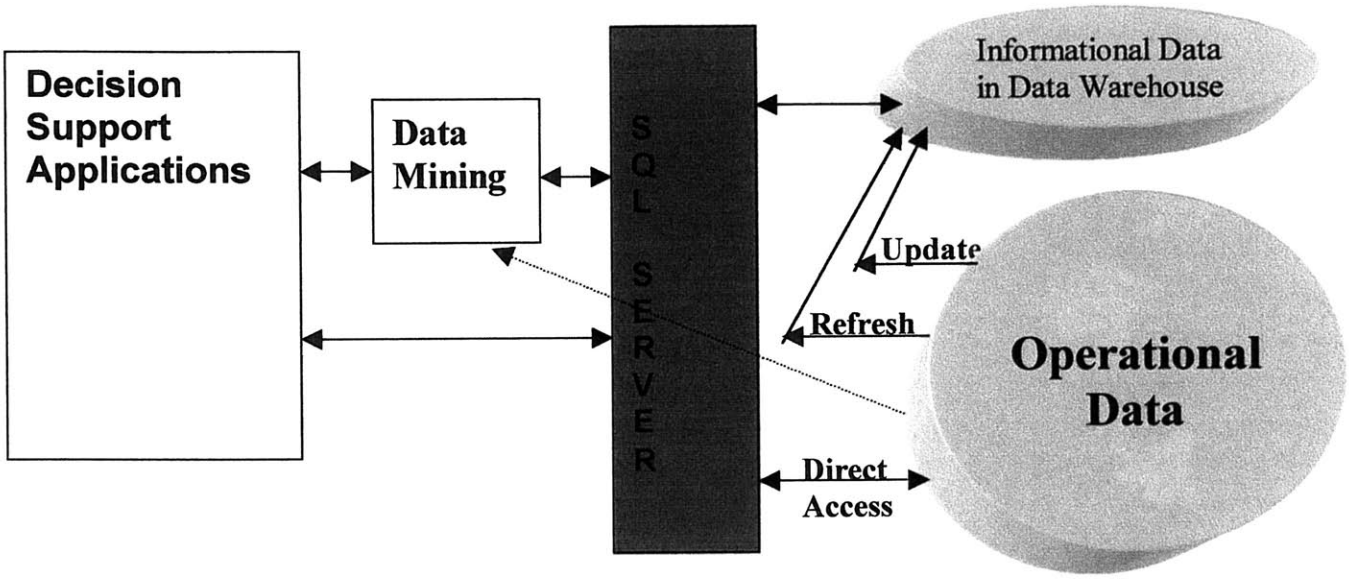
Once the data have been selected and required transformations done, a data-mining tool can now be applied. Specific predictions about futuristic events based on previous collected data can yield in significant hidden findings through the use of well designed algorithms, a topic of discussion in later sections. Using a data warehouse alongside with a mining tool is usually recommended as this allows for a more efficient organization of the collected data in ways that can facilitate and optimize analysis. Furthermore, the mining tool can also interface with a DSS for further interpretation of the data. However, a data mining system need not always interact with a data warehouse, and in fact, data mining can still extract pertinent information if given raw data from a database. The main advantage of using a data warehouse is that most of the data are already in already integrated in a suitable format of choice making it easier for a data-mining tool to extract the higher quality information.

The final step in the data mining process consists of interpreting the results. Once the extracted information is analyzed and interpreted, the most relevant information can be passed onto the decision-maker through a DSS. Result interpretation can consist not only of interpreting the output but also of further filtering the data and passing that information to the decision support system. In the case that the interpreted results are not satisfactory, it may be necessary to repeat any of the previous steps until the information generated contains the maximum added value to the data miner.

As such, data mining is a very complex process. Many steps need to be performed correctly before feeding of data to the data mining tool. Furthermore it is not guaranteed that the data-mining tool will yield significant results in any steps of the mining process. Certainly, performing many trials are recommended as this can reveal error corrections in any of the four steps. Any of the previously mentioned steps can be modified to continue investigating the data and searching for hidden patterns. This is the challenge of the data mining organization and though it can be a painstaking process, the more data that is mined, the more likely the data miner will learn from the process.

The use of tools such as DSS and a warehouse environment complement the data mining tools used to find useful facts buried in layers of data. To maximize the efficiency of data mining, both of these other tools must provide high quality delivery information to the data-mining tool. The use of good complementary tools to sift through data along with a powerful data-mining tool should be part of a well designed environment [1].

Figure 4: Integration of Tools



Source: Data Mining: Extending the Information Warehouse Framework, IBM. 2000.

The way in which a DSS along with a data warehouse all integrate with a data mining tool to extract hidden patterns in data can be summarized by Figure 4 seen above. Shown are all three major components necessary for a data mining framework extraction tool used today. A DSS, as discussed in the previous section, is needed to allow a manager to be able to examine the data and help him or her make a decision. A DSS will obtain the results passed from the data-mining tool [87]. The data fed to this data-mining tool will have been synthesized and integrated by the data warehouse system. Thus, the data-mining tool will interact with the DSS, presenting a final solution to the organization implementing a data mining strategy. Before this step can occur however, the data selected must be integrated and pre-analyzed before feeding these data to the data-mining tool. Integrating the data involves merging data that resides mainly in an operational

environment having multiple files or databases. Facilities for doing these data transformations are usually provided with mining tools. This is as a result of the mining developers also building tools that will integrate and merge well with the data mining functions. An important note is that for data mining to occur, it is not necessary to require that a data warehouse be built. In can be the case that data can be downloaded directly from the operational files to flat files that contain the data ready for the data mining analysis [1]. However, in most situations, the data fed into a data-mining tool will need to be synthesized and integrated. This is the task performed by a data warehouse. A data-mining tool will often access the data warehouse synthesized data through an SQL interface. Communication between the data warehouse to the data-mining tool to the decision support system will continue to operate with one another until an optimal solution deemed by the data miner is found.

Data Mining Specific Tool Trends:

Certain types of tools seem to be growing over other data mining tools. In a recent study conducted by KDNuggets.Com, a leading data mining news provider placed the following commercial and public-domain tools in the different categories shown in Table 2. This table shows that the growth of suites and other data mining tools continue to increase. Over 100% growth was experienced in suites over a 2 ½ year period. Furthermore, there is also strong evidence supporting continued research tool growth such as association and sequence algorithms, clustering, and visualization. Text and web mining is a growing field that can take advantage of the worldwide usage of the Internet and the many companies investing to develop online strategies.

Data on the Internet has to be effectively managed. This data could to be mined to gain insights into customer purchasing patterns and trends. The explosion of users on the Internet and the increasing number of World Wide Web servers are rapidly increasing digital libraries. Digital libraries are digitized information distributed across several Internet sites containing information such as multimedia data such as voice, text, video and images [88]. Thus, text and web mining are techniques employed to serve in finding trends amongst these piles of data. With information overload on the web, it is highly desirable to mine the data and extract patterns and relevant information to the user. This will inevitably make the task of browsing on the Internet easier for the user. Therefore, there has been considerable interest in studying algorithms that mine the web.

Table 2: The growth of commercial and public-domain data mining tools

Tool Type	Feb 1997	Feb 1998	Nov 1998	Aug 1999
Suites	18	30	37	37
Classification:				
Multi-approach	4	8	8	10
Decision Tree	14	18	18	19
Rule Discovery	10	11	12	14
Neural Network	50	60	76	100
Bayesian	0	0	10	20
Other	6	7	12	13
Total	84	104	126	176
Associations and sequences	0	0	6	13
Clustering	6	5	10	12
Visualization	5	12	26	31
Text and Web mining	0	5	7	15

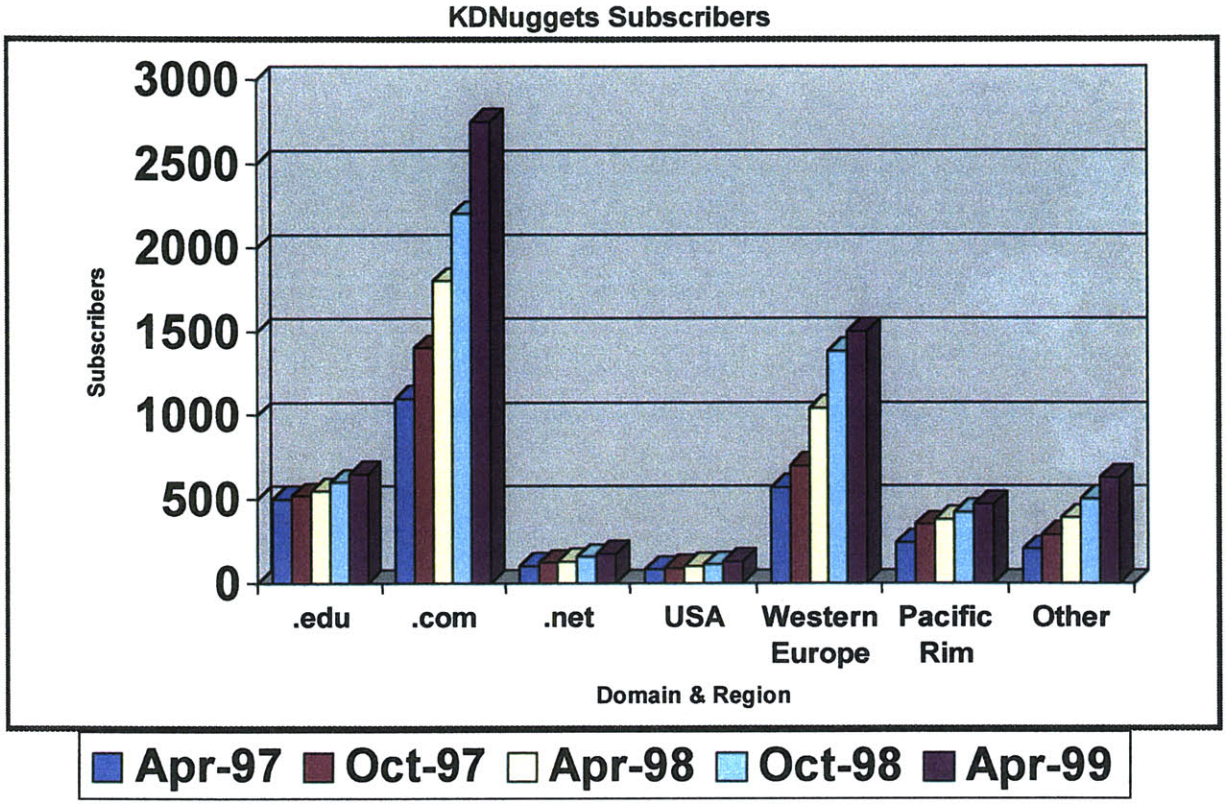
Source: Shapiro, Greg. The Data Mining Industry Coming of Age. IEEE Intelligent Systems. KDNuggets.Com. 1999.

Table 2 shows the growing trend of specific data mining tools. Suites seem to have saturated the market, as there was no increase in over an 8-month period. This stabilization in the number of suites shows that the market for commercial tools is growing [11]. Association and sequence algorithms, Bayesian networks, and data visualization are still growing strong and show no signs of a mature and saturated market.

Furthermore, the growth trend of data mining tools is also reflected in the number of users subscribing to the KDNuggets newsletter, which is today regarded as of the leading newsletter provider in data mining and knowledge discovery in Figure 5. As data mining continues to catch the public's eye, already there exists reason to prove why data mining will be a technology used and understood by more people. Figure 5 illustrates the growing acceptance of the data mining technologies. Researchers made up the majority of KDNuggets subscribers, but now most are from commercial domains. The region analysis indicates the growing acceptance of data mining worldwide. Already KDNuggets has over 8,000 subscribers for its newsletter all over the globe compared with a mere 50 people in 1993. Another interesting observation is that data mining is not a technology offering solely in the United States. Data mining is a global technology being developed to aid companies worldwide. However, with the increasing demand for web mining, most of the revenue generated in this particular sector will be predominant in the United States and Western Europe.

As the popularity of data mining continues to emerge, one of the main challenges for this technology in the 21st century will be generating algorithms that are able to cope with real-time predictions. Third generation data mining systems will have to be designed to meet the growing demand of analyzing real-time on-line information. Whereas today companies can implement data mining solutions that take many months to implement, offering solutions online and in real-time can be a very complex procedure as the dynamics of Internet web sites can be a hurdle to overcome.

Figure 5: KDNuggets Subscribers by Region and Domain



Source: KDNuggets

Software products are becoming a growing presence in the data-mining arena. To illustrate the point that more companies are trying to offer data mining products to different companies, one just needs to look at the distinct methodologies of applying the data-mining concept to solve classes of problems. Table 3 shows that suites are still an important part of data mining software products provided. Some of these leading commercial vendors include leading corporations such as IBM, SPSS, Angoss, and Oracle. Table 4 shows some of the leading companies in providing software products for web mining including some free services offered by other corporations.

Table 3: Leading Suites Supporting Full Knowledge Discovery Process

Commercial Products:	Main Features	Price	Released Date	MainClients/ Partnerships. Strategy
Alice d'Isoft, a streamlined version of ISoft's decision-tree-based AC2 data-mining product, is designed for mainstream business users.	- An on-line Interactive Data Mining tool. - Discovers hidden trends and relationships in the data and makes reliable predictions using decision trees	Estimated starting price at $1000.	Version 6.1 launched in 1999.	ISoft has a strong partnership strategy, illustrated by broad marketing and technical alliances with information technology companies.
Clementine from SPSS, leading visual rapid modeling environment for data mining. Now includes Clementine Server.	-Users create and interact with a stream: a visual map of the entire data mining process. -Range of applications including web-mining capabilities to client's sites.	Estimated starting price at $2499.	Updated version launched in 1998. Clementine was the first enterprise-strength data mining offering aimed at business users.	Strategic partnerships with partnering in the areas of application development, data management, analytical consultation, and solution implementation. More than 250 Clients.
Darwin (now part of Oracle), high-performance data mining software, optimized for parallel servers.	- Darwin enables improved marketing by segmenting customers and predicting their behavior. -Multiple algorithms used	$1,495	Updated version released in 1999.	-Oracle has joined forces with high-value warehouse companies that are recognized in their fields of expertise, including System Integrators, Independent Software Vendors
IBM Intelligent Miner for Data, enables users to mine structured data stored in conventional databases or flat files.	-Offers a suite of tools offering data processing, statistical analysis, and results visualization to complement a variety of mining methods. - Mining operations can be performed directly on DB2 .	-Price can widely vary depending on project needs. Estimated starting price at $2500 for software only.	Updated version released in 1998.	-Seeks to acquire application and service partners.
Data Detective has a modular design consisting of an associative analysis engine, a graphical user interface and interfaces to virtually all-common database formats. Offered by Sentient Machine Research.	- 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.	-Price dependent on per project objective.	-Founded in 1990 with the objective to turn adaptive techniques into solutions for business problems.	-Revenues come primarily from consultancy, carrying out data analysis projects and development of standard tools and tailor made information systems. -Clients include banks and financial institutions, insurance companies, and governments.

Table 3 Continued:

Data Engine, software tool for data analysis in which processing. Offered by Management Intelligenter Technologien.	-By using neural networks, logic and statistical methods, DataEngine provides techniques for data analysis. - Decision rules, clustering, neural networks and neural systems are offered in combination with mathematics, statistics and signal	-DateEngine: 5000 Euro-$4,500 -DataEngine ADL for Win95/98/NT: 800 Euro-$720 -DataEngine V.i: 1500 Euro-$1,350 -Maintenance Contract: The DataEngine maintenance contract amounts to 20% of the product price per year.	Developed in 1991. Updates to software tool made since. Latest version is DataEngine.	-Involved in a series of international projects to facilitate the effective exchange of information. -Examples include the ERUDIT Network of Excellence and various Eurupean projects.
DB Miner 2.0 (Enterprise) powerful and affordable tool to mine large databases; uses Microsoft SQL Server 7.0 Plato. Offered by DBMiner Technology Inc.	-Technology offers OLAP, association, classification and clustering techniques. -Provides powerful data mining query language: DMQL.	$999.00.	Releasted August 15, 1999.	-Currently looking to form partnerships with worldwide information technology providers and consumers in various areas.
Genio Miner integrates data acquisition and cleansing predictive models (Bayes, NN, Neural Nets, Decision Tree) and clustering. Offered by Hummingbird.	- Genio Miner provides an integrated environment to extract data from not only flat files, but also from multiple heterogeneous sources including direct database connections. -Uses predictive modeling, clustering, and decision trees.	Pricing to be determined at release date.	Available after May 15, 2000.	-Hummingbird offers complete global enterprise solutions from advanced host connectivity, through sophisticated data exchange, query & reporting and analytic applications, to information management at the desktop or on the Web.

Table 3 Continued:

KnowledgeMiner, has GMDH neural nets, fuzzy predictions and fuzzy rule induction (Mac only). Offered by Script Software.	-Main offerings include GMDH (global method of data handling), analog complexing and rule induction in one application. -KnowledgeMiner has been used sucessfully for analysis and prediction in the fields of finance, economics, imaging, ecology, health, biotechnology, chemistry, math and others.	-$780 for professionals -Discounted prices for students -Can download for free	Latest update in 1999.	-Used by: NASA, Boeing, MIT, Columbia, Notre Dame, Mobil Oil, Pfizer, Dean & Company, and many other corporations, universities, research institutes and individuals around the world.
Knowledge Studio from ANGOSS, featuring multiple data mining models in a visual interface.	-Knowledge Studio is a datamining tool that includes the power of decision trees, cluster analysis, and several predictive models. -Data mining algorithms supported include CHAID, XAID, K-Means, and Entropy decision tree algorithms, as well as multi-layered perceptron, radial basis formation and probabalistic neural nets. Under development are Kohonen, C4.5 and regression algorithms.	$7,500 per seat.	First released in 1998. Newest release 3.0 is due mid June in 2000.	- ANGOSS has a large and diversified customer base, with licensing transactions involving major financial services, telecom, retail, technology and manufacturing enterprises in North America and the European Community. -Data mining components licensed from ANGOSS are included in the decision support solutions of other vendors, such as COGNOS and Hummingbird product offerings.
NeoVista Suite, an enterprise-level suite of integrated data mining engines (Net, Tree, Cubist, Bayes, Cluster, Kmeans, AR), that includes support for predictions, descriptions, and transformations. Offered by Accrue Software Inc.	-Retail specific product - Provides powerful capabilities for generating prediction models including explaining associations in data, and identifying optimal business patterns.	Estimated at $2000.	Released in 1998.	-Include partnerships with ISP's including Organic Online and Ogilvy, and Solution Providers including Cambridge Technology Partners and JDA Software. - Mining product used in the "off-line" world by retailers such as JC Penny and Wal-Mart. -Over 550 clients.

Table 3 Continued:

Pilot Decision Support Suite provides high-performance analytical solution, combining OLAP and knowledge discovery capabilities. Offered by Pilot Software.	- Offers both high performance and the ability to analyze large data sets to manage all of an organization's divergent analytical needs. - Comprehensive solution with predictive data mining, high-performance OLAP, and flexible visualization	Estimated at $1000	Version 6.3 launched April 10, 2000.	-Strategic Alliances with Microsoft, SAP, Informatica, Compaq, ECS, and Sun Microsystems.
SAS Enterprise Miner, an integrated suite that provides a user-friendly GUI front-end to the SEMMA (Sample, Explore, Modify, Model, Assess) process.	-An environment that offers a broad set of data mining techniques such as decision trees, neural networks, associations, and data mining regression. -A GUI designed for business users.	Estimated starting price at $2000	Latest version launched in 1999.	SAS is the world's largest privately held software company. Founded in 1976, the Institute markets packaged business solutions for vertical industry and departmental applications. Thus, many SAS has worked with many clients and is involved in many partnerships.

Table 3 Continued:

Statistica, a comprehensive, integrated statistical data analysis, graphics, data base management, and custom application development system. Offered by Statsoft Corporation.	- Interactive operation including flexible output exploration options. -Advanced graphics display technology - Offers a flexible "workbench" on which input data, results, and graphs can be created and automatically arranged in a variety of ways to facilitate data exploration or hypothesis testing.	-Statistica for Windows: $1095. -Also offer training courses and workshops.	-1993-1998: The initial release of STATISTICA for Windows (4.0) was in 1993 followed by releases 4.5, 5.0, 5.1, '97 Edition, and '98 Edition. -The '99 Edition offers one the more comprehensive implementations of the GLM (General Linear Models), GSR (General Stepwise Regression), GLZ (Generalized Linear Models), and PLS (Partial Least Squares) methods available on the market.	- StatSoft Research and Development department is working with its corporate clients to provide "integrative" qualities of software. - The exact number of users of StatSoft products is estimated at over 600,000 users worldwide (Universities and major research institutions account for around 30% of StatSoft sales; corporations and manufacturing facilities: 60%; government agencies: 10%).
Xpertrule Miner 4.0, addressing the full data mining process for the professional and for the first time user. Offered by Attar Software.	See Table 5. - ActiveX data mining components for embedding within vertical applications. Scalable high performance client-server data mining.			

Table 4: Software Tools for Web Mining

Commercial Products:	Main Features	Price	Released Date	MainClients/ Partnerships, Strategy
Customer Interaction System's Micro Marketing module collects click streams at the application server level so it can generate very rich data, transform it automatically to the data warehouse, and provide mining operations. Offered by Blue Martini Software.	-Results are presented through Intelligence Reports designed for the business decision-maker that combine statistics, business rules, aggregated customer profiles, and data visualizations. -Customer segmentations	-Pricing is subscription-based calculated on the length of the engagement and on the volume of transaction and click stream data analyzed.	February 2, 2000.	-Build brand equity and raise revenues across all sales channels by interacting with customers through call centers and web sites for e-merchandising, e-marketing, and e-service.
Clementine offers sequence association and clustering used for Web data analysis. Offered by SPSS.		See Table 3.		
CustomerConversion from Quadstone, customer-centric analysis and graphical reporting of web and other data.	-Customer focused automated knowledge discovery and data mining -Integrated analysis of web and traditional customer data -A visual and statistical mining suite driven by the business user. -Tracks page performance: customer	Retails between £40,000-£250,000=$60,000-$373,000	-Company founded in 1995. -Product released in January 2000.	-Quadstone focus is on delivering customer behavior intelligence solutions to drive B2C business profitability.

Table 4 Continued:

EasyMiner, features Cross-session analysis; Click-stream analysis; Cross-sales by MINEit Software.	-Offers click stream analysis and visitor segmentation. - Claims to optimize web site for maximum commercial impact by understanding the dynamic behaviour of web site visitors.	Estimated starting price at $1500.	-Released in 1999.	-Current seeking strategic partnerships with resellers, systems integrators, consultancies, and Independent Software Vendors (ISVs).
HitList, powerful and flexible server logs analysis with over 300 report elements by Accrue Software. Captures information across multiple Web and application servers at the network or at the Web server level.	-Collects and updates database with Web traffic data in near real time. -Combines Web data with multiple enterprise data sources, such as transaction or demographic information onto a single report. -Builds custom reports utilizing more than 375 data elements.	Software licensing for Accrue Hit List begins at $10,000.	Company founded in 1996. Version 1.0 released in 1996. Version 4.51 released in December 1999.	-Alliances with Doubleclick, Vignette, Art Technology Group, Microstrategy, Informix, Oracle, IBM, Sun Microsystems, Microsoft, and Netscape.

Table 4 Continued:

KnowledgeWebMiner, combines ANGOSS KnowledgeSTUDIO with proprietary algorithms for click stream analysis, Acxiom Data Network, and interfaces to web log reporting tools and to design and deploy proprietary intelligent agents to drive content selection, visitor interaction, and transacting options and outcomes.	-Measuring the effectiveness of on-line and off-line promotion strategies in driving traffic, click-throughs and transaction volumes at client's Web sites. -Exploring and understanding differences in visitor usage patterns and behaviours. -Segmenting visitors, guests and members more precisely to enable more relevant, effective, and personalized interactions.	Introductory pricing for KnowledgeWebMiner starts at $85,000 for Windows NT/2000 and SUN Solaris platforms and includes a client-server based solution and comprehensive training for up to 5 e-business analysts.	-Company founded in 1984. -Knowledge WebMiner released in May 1, 2000.	-Partnerships with Whitecross Data Exploration, Tantau Software, Compaq, Microsoft, IBM, Oracle, Siebel, Sybase, NCR, Sequent, Informix, Polk Inc, Experian Database Marketing Solutions, Acxiom, Hummingbird Communications, Cognos, Sasi, and Customer Analytics.
Net.Analysis, e-business intelligence solution, providing the scalability required by large e-business enterprises. Offered by NetGenesis. This device-independent data collection architecture provides continuous streaming of site activity data in conjunction with rich sets of historical information and trend analysis	-High-end cross-platform enterprise database support -Provides high-end performance on NT 4.0/SQL Server 7.0; Oracle 8.i on Solaris 2.6; and Sun/Solaris and Sun/Sybase. -Provides more than 150 predefined reports with sophisticated drill-down capabilities to meet differing information needs of all members of an enterprise	-Price range varies according to consultation. Usually in ranges from $1000-$3000 for software product only.	-Founded in 1994 by Matt Cutler and Eric Richard, two MIT undergraduates, -Recognized worldwide as a leader in e-customer intelligence. - More than 400 Internet-enabled and Fortune 500 customers now use NetAnalysis worldwide. -Latest version of product released in 1999.	-Strategic partners with IBM, Oracle, Sun Microsystems, Microstrategy, Doubleclick, Microsoft, Art Technology Group -Involved with seeking alliances with ASP companies. -Over 50 clients. -Fiscal first quarter of 2000 quarterly revenues of $3.8 M.

Table 4 Continued:

NetTracker, powerful and easy-to-use Internet usage tracking programs by Sane Solutions.	- Each version of NetTracker contains a variety of standardized summaries, which provide detailed information about client's Internet traffic: -All versions of NetTracker analyze web server log files and create thirty web server summaries that provide detailed information about web site traffic.	-NetTracker 4.5 Professional: $495 -NetTracker 4.5 Professional Support Agreement, entitle users to telephone and e-mail support for a period of one year: $195 -NetTracker 4.5 Enterprise (analyze multiple web sites): $995 -NetTracker 4.5 eBusiness Edition: $5995	- Founded in March 1996 -Latest updated version released early 2000.	-Over 150 Clients -Predominantly Internet Partners including Sun, Cobalt Networks, SGI, and Compaq. -Strategy is to integrate NetTracker with other eBusiness solutions as well as to optimize our products on a wide variety of platforms.
Enterprise Suite, a suite for Data Mining of web traffic information. Offered by Web Trends.	-Provides web server traffic analysis -Exports results from WebTrends database to high-end Oracle, Microsoft SQL, Sybase, Informix and other ODBC-compliant databases for further data analysis.	-Enterprise Suite: $1999 -Suite Subscription: $599 -Suite & Subscription: $2398	-Company founded in 1995. -Enterprise Suite 3.0 released in December 1998.	Partnerships with AbleCommerce, Allaire, Covision, Hewlett Packard, Intershop, Lotus, Microsoft, Novell, Netscape, Oracle, and Vignette Technology Partner.

Table 4 Continued:

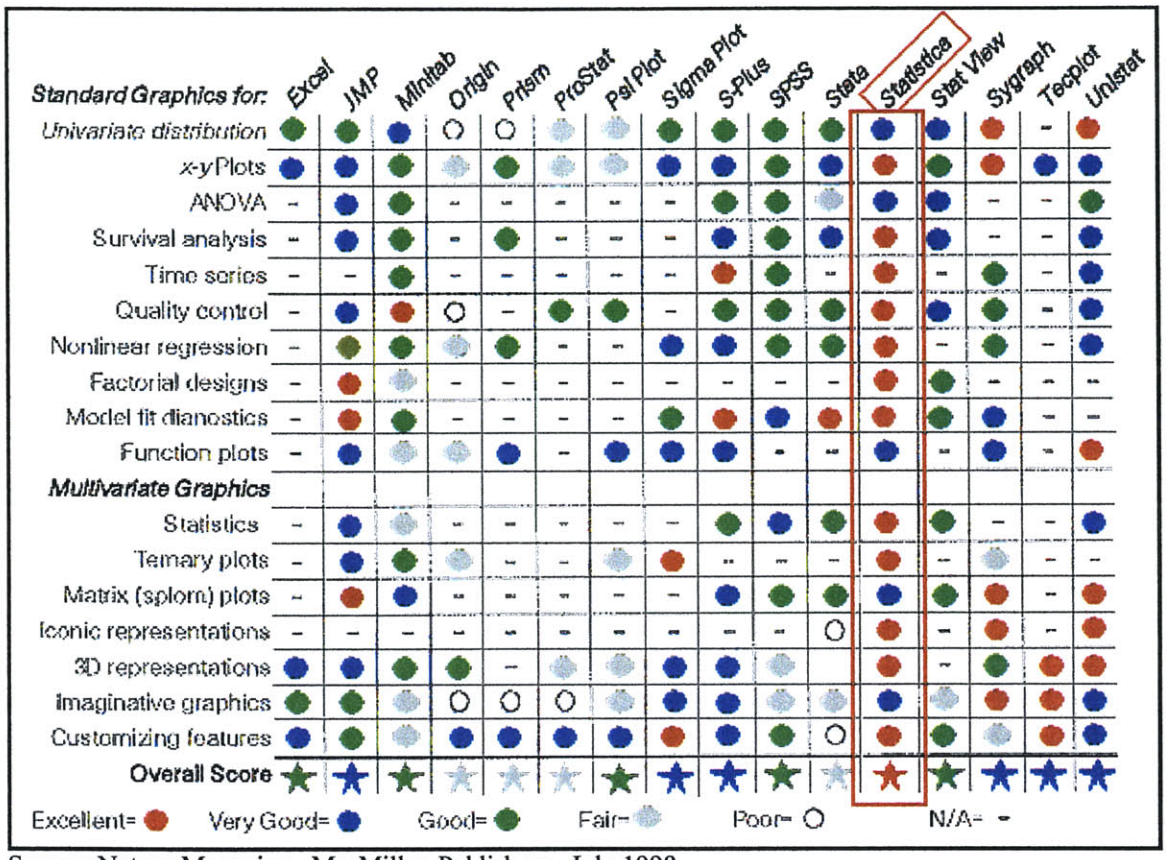
Free Products:	Main Features	Released Date	MainClients/ Partnerships, Strategy
Analog (from Dr. Stephen Turner), a free and fast program to analyze the web server log files (Win, Unix, more)	-Program designed to analyze the log files from your web server. Indicates which pages are most popular, which countries people are visiting from, which sites they tried to follow broken links from, etc.	Current version is 4.1 released in 1999.	Continuing research to develop new features.
Discovery, attaches to a browser to search the Web and to the host computer. Offered by Altavista.	-Utility that brings knowledge and information together from the Internet and the host computer. -Locates pages on the Net that are similar in content to the page a particular reading.	Launched in 1999.	Planning to launch new version by mid-June 2000.
WUM 5.0, an integrated environment for log preparation, querying and visualization. Offered by Humboldt University Berlin.	-WUM is a sequence miner. -Primary purpose is to analyze the navigational behavior of users in a web site, but it is appropriate for sequential pattern discovery in any type of log.	Last update March 2000.	Seeking potential research partnerships with educational institutions.

Software Statistical Tool Comparison:

For each section there has been an analysis of the main features of what some of these software products offer to customers. In this section, there is a further discussion of comparing some of these tools relative to one another. Statistics and analytics of a tool are as important to the user as the other capabilities a software package can offer. Statistics play an integral portion of any data mining software package offered to customers.

One of the major user criteria when evaluating a tool is the ease and convenience of a well-designed graphical user interface. Shown in Table 5 is a comparison of leading software tools and the results from evaluating the tools. According to Nature Magazine, results show that Statistica is the leading provider for graphical solutions in leading software packages based on these seventeen characteristics [89]. Statistica is a Windows package that offers a general-purpose statistical and graphics package featuring a wide selection of basic and advanced analytic procedures for science, engineering, business, and data mining applications. In comparison, StatView for Windows and Macintosh, offered by SAS, offers packs data management, statistical analyses, and presentation tools into a single software package. Price comparisons indicate that Statistica is $400 more expensive than StatView ($1095-695), potentially justifying the higher markup in price due to the array of graphical features and added functionality to the user.

Table 5: Comparison of Graphics of Leading Commercial Software Products



Source: Nature Magazine. MacMillan Publishers. July 1998.

In terms of performance, Statistica seems to meet user demands in comparison to other competing products. In a comprehensive comparative technical review in 1998 of the top five leading statistics software packages (*STATISTICA*, *SPSS*, *Minitab*, *Statgraphics*, and *S-PLUS*), including comparative benchmarks of computational accuracy and extensive quantitative ratings of various aspects of the functionality of the programs, STATISTICA was selected as the leading statistics package (See Table 6) [90]. Some of these quantitative ratings included features such as analytic features, ease of learning, ease of use, data visualization, data compatibility, installation, and other features including speed and access time [90]. From Tables 5 and 6, STATISTICA seem to offer the best statistical solution in the market today. In a separate study conducted by

Buyer's Guide, the highest rated statistics software published by *Scientific Computing and Automation* (September, 1998) also revealed that Statistica was the leading tool in statistic packages. Other leading comptetitors included SAS, SPSS, and Minitab. The results, shown in Table 7, rank the tools in a comparative listing of features of statistical packages.

Table 6: Performance Comparison of Five Leading Statistic Products

Top 5 Statistic Tools:	Recommended Quantitative Ratings of Software Tools
StatSoft, STATISTICA (release 5.1)	1
SPSS, Windows Student Version (release 7.5)	2
Minitab Inc, Statistical Software (release 12)	3
Statgraphics, Statgraphics Plus (release 3)	4
MathSoft, S-PLUS (release 4)	5

Source: Technology Special Statistics Report. CMP Media Incorporated. September 1998.

Table 7: Performance Comparison of Leading Statistic Products

StatSoft/STATISTICA	1
SAS/SAS Software	2
SPSS	3
Manugistics/Statgraphics	4
Minitab	5

Source: Comparative Study of Statistical Packages. Scientific Computing and Automation. September 1998.

Data Mining Techniques:

Introduction:

What kinds of business problems can data mining technology solve and what must users of these tools understand to apply these tools effectively? Questions such as: "Do sales of product X increase in the month of July?" or questions such as "Do sales of product X decrease when there is a promotion on Product Y?" are easily solved without the aid of data mining. Moreover existing tools such as OLAP and statistical techniques can be used in this situation to analyze those types of cases. In contrast, with data mining, a potential customer can ask questions such as: " What are the optimal factors in determining the sales of Product X?" However, not all data mining tools are necessarily optimal in solving certain kinds of problems. Some tools are better than others in specific types of problems.

With traditional tools, an analyst trying to come up with an answer to questions such as those above will arduously try to generate a model through trial and error. He or she will first pose a set of assumptions with a hypothesis, then test it, and finally propose additional hypothesis and go through the repeated process of testing it and in this iterative way, build a model. With data mining, though certainly a set of assumptions along with a hypothesis needs to be postulated, along with testing it, and revising the procedures, the advantage in using a data mining tool is that a tool shifts much of the work of finding an appropriate model from the analyst to the computer. Thus, generating a model requires less manual effort alongside the added advantage that using a computer allow for a larger

number of models to be evaluated increasing the odds of finding a proven and working model.

Classes of Tools [82]:

- Associations
- Sequential Patterns
- Classifiers/Regression Analysis
    - Decision Trees
    - Neural Networks
- Visualization
- Clustering
- Collaborative Filtering
- Data Transformation and Cleaning
- Deviation and Fraud Detection
- Estimation and Forecasting
- Bayesian and Dependency Networks
- OLAP and Dimensional Analysis
- Statistical Analysis
- Text Analysis
- Web Mining

The above list of methods of data mining applications includes most of the extensive technology uses of data mining. The more important applications of data mining to date are mostly association and sequential pattern tools, classifiers, visualization, and

clustering [14, 15]. Collaborative filtering is a fairly new application used mainly in web mining. Further analysis will follow on collaborative and web mining. Besides web mining, the remaining applications will be addressed and discussed in the context of other sections. In each of these applications, data mining differs in the approach taken to solve problems. Each application is usually geared in solving a particular type of problem. Usually a specific algorithm will be favored over others depending in what the problem posed by the data miner is. Each of these applications will be explained in detail in the following sections along with a discussion of the occasions where one tool should be favored over the use of another.

Associations:

In association tasks, a problem is solved by issuing a query against a database and finding out the affinities between existing variables. For example, these affinities can be expressed by stating "85% of customers who purchased items A, B, C also ended up purchasing D & E." As such, the goal of using associations is to find common relationship amongst the items, or variables, existing in a collection of records. Selling as many products to as many customers increases the potential revenue of a business and is an important component in understanding consumer preferences [14]. One of the ways in which this goal may be realized is by understanding what products or services customers tend to be purchased at the same time, or later as follow-up purchases. As such, determining consumer purchasing behavioral trends is a very common application of data mining, and association and sequencing techniques can perform this kind of analysis.

Sequential Patterns:

Sequential patterns are another form of a commonly used application in data mining. Sequential pattern mining functions are quite powerful and can be used to detect the set of customers associated with some frequent buying pattern. A sequential pattern function tries to detect frequently occurring patterns in the data. As an example, if customers increase their purchase of a particular product A by 20% and also increase their consumption of another product B by 25%, then product C is found to also increase by 10%. Whether product's C increase in purchases was a direct result of the increase in sales of products A & B collectively is uncertain; what is certain is that this relationship holds and it is up to the data miner to examine this trend further by issuing further queries to determine whether or not a particular relationship is a casual one. In this manner, sequential patterns are used in detecting frequency of events occurring and their relationship to other events.

Thus, association and sequencing tools analyze data to discover rules that identify patterns of behavior. Using an association or sequencing algorithm is frequently called market basket analysis. Business managers or analysts can use such market basket analysis to plan discounting products, product placement, and timing of promotional sales. The timing of such events is crucial for organizations. If customers tend to buy goods A and B, discounting A by an x% could yield in a higher probability of more customers purchasing both goods. Furthermore placing products in a store that are most likely to be purchased by the same customer is another area in which these algorithms can be highly effective. Lastly, deciding when to generate a promotion that will increase sales dramatically is also another way these tools can help organizations better time their strategic moves.

Association and sequencing mining tools in specific industries:

The industries in which association and sequencing data mining tools are mostly used include the retail, health care, financial, and insurance industries. Some of the examples above tied into how effective these tools can be for retailers. However, these tools can add value to healthcare companies as well. There are many applications in care management, procedure interactions and pharmaceutical interactions where these tools can find patterns in transaction data [14]. For example, when a patient consults a doctor, he might prescribe the patient 3 different types of medicine. However, it turns out that a doctor tends to recommend the same 3 doses to the same person suffering from a similar sickness. As such, patients taking drugs A and B are more likely to also be taking drug C.

In the financial service sector, finding a pattern that stocks in industry X go up after a certain percentage in industry Y go down can prove significant to a broker optimizing the value of stocks for clients. In the telecommunications and insurance sectors, detecting fraud is a serious concern for companies in these industries. Trying to find patterns in which fraud detectors take advantage of the telephone lines is valuable. For example, if a certain percentage of fraud crimes were known to occur between the hours of 1-3 AM in certain parts of the country, adding further security to those regions in which hidden patterns seem to emerge can save money and time to telephone companies. Another type of fraud common today is that of credit card fraud. Detecting when certain crimes in the credit card industry occur is important. It if often the case that accident claims involving certain parties, the Jones for example, turn out to be twice as likely to be fraud than other accident claims. An example of how associations and sequencing use data to compute

results is illustrated in Table 4. Shown is a database of information collected about purchases on detergent and bleach maintained by a pseudo-company X.

1,000,000 total transactions in year 1999
50,000 transactions involving detergent products in 1999 (5% of total purchases)
20,000 transactions involving bleach products in 1999 (2% of total purchases)
10,000 transactions involving both detergent and bleach products in 1999 (1% of total purchases)

Association and sequential algorithms often use the expected confidence, the ratio of the number of transactions of a particular item over the total sum of transactions, as a basic key statistic [14]. For detergent, the expected confidence will be 5% (see Table 4). Support (or prevalence) measures how often items occur together [14]. In this example, the support for detergent and bleach products would be 1% of the time. Support is not dependent on the direction of the rule; it is only dependent on the set of items in the rule. Confidence (or predictability) measures what the dependence is on one item to another [14]. Since 50,000 transactions were on detergent products and 10,000 transactions contained both detergent and bleach products, this implies that the confidence of people buying detergent products will also buy bleach products is 20%. Contrast this figure to the confidence rule that out of 20,000 people who purchase bleach products, out of those 10,000 also buy detergent products, giving a confidence rule of 50% of people purchasing bleach products, they also purchase detergent products. Usually a ratio, labeled as a lift, is commonly used to determine the strength between the confidence that consumers purchasing a particular product will buy other products divided by the

expected confidence. The stronger the ratio, the more significant the effect of purchasing a product will have on other products. These numbers are summarized in Table 8 illustrating the findings of using an association or sequential algorithm to find interesting consumer patterns.

Table 8: Association Table

Product	Future Product Consumption Patterns	Expected Confidence	Confidence	Support
Detergent	Purchase Bleach	5%	20%	1%
Bleach	Purchase Detergent	2%	50%	1%

Association and sequential algorithms work by taking ratios amongst different variables and signaling these variables as the significant variables to study. Using item hierarchies as above can facilitate the analysis of variables. Using a hierarchy of products can be used to group similar items together [14]. Hierarchies are very useful especially for retail companies that sell all kinds of products. By breaking up the products into similar categories, the data-mining tool can benefit from this higher-level abstraction to find patterns on the generic variables it analyzes. In the example above, many kinds of detergents exist, and if it is found that the lift ratio turns out to be greater than a threshold, finding all the brand name detergents that make up this group class will be helpful in further finding useful hidden patterns in the data. Similarly for the bleach general class of variables, conducting a further analysis on all relevant bleach products could result in even more significant results by signaling the products that are in fact causing a certain relationship to stand out. Using this type of analysis by breaking the

problem up makes careful use of the vast amounts of information usually found in databases.

The data-mining sequencing tool can then start to be finely attuned to be aware of specific types of data and potential patterns it may have discovered from previous clients when working with its current client. In Table 9 there is a listing of current leading companies devoted to developing association and sequential pattern knowledge discovery software tools. Table 10 is a listing of all major competitors involved with software development in the area of sequential patterns.

Table 9: Software Tools for Associations Knowledge Discovery

Commercial Products:	Main Features	Price	Released Date	MainClients/ Partnerships Strategy
Azmy SuperQuery, an association rule finder.	-Summaries using more than 20 statistical functions. -Reports facts as a Table to enable user to re-analyze the Fact database itself.	Office Edition: $149.95 Discovery Edition: $449.95	Both editions have been modified from their earlier versions. New versions released in 1999.	
Clementine, Suite from SPSS, includes market basket analysis	See Table 3.			
IBM Intelligent Miner for Data, through knowledge discovery tries to leverage hidden information in its data, uncovering associations, patterns, and trends that can lead to real competitive advantage.	See Table 3.			
Magnum Opus, a tool for finding association rules from data for Windows 95/98/NT platforms. Offered by RuleQuest.	-Designed to analyse substantial databases containing up to millions of records. - Allows the user to specify the maximum number of association rules to be found. -Software installed on a single PC	$740	1998.	

Table 9 Continued:

Nuggets suite, includes association rules. Offered by Data Mining Technologies.	-Rule induction data mining tool with validation, prediction, and rule query modules. Limited to 20 variables analysis.	Services offerings: -Consulting: $2,500/day -1 day class: $2,500 -Maintenance: 15% of license fee - Data Mining 30-Day Jump-Start Program $15,000 plus license fees	1998.	Industries served include: Internet Marketing Pharmaceutical Banking Insurance Government Education
Megaputer Polyanalyst Suite 4, includes market basket analysis engine. Offered by MegaPuter.	- Uses an algorithm that examines a long list of transactions in order to determine which items are most frequently purchased together.	-PolyAnalyst Pro 4.1: $9,960 -PolyAnalyst COM, a self-customization package using a distributed application: $16,780	Developed in 1998, Launched in 1999.	-Over 50 Clients. Fortune 100 companies and smaller companies: McKinsey & Co, 3M, Boeing, and Pioneer Securities
SGI MineSet, 3.0 Enterprise Edition includes market basket analysis. Offered by Silicon Graphics Incorporated.	-MineSet is an advanced scalable client/server tool set for extracting information from data warehouses, mining the data with sophisticated analytical algorithms, and revealing newly discovered patterns and connections through interactive visual displays.	Price not disclosed.	Launched in 1997.	-Main customers include: Ford, Price Waterhouse Coopers, Andersen Consulting, General Mills, and many other large corporations.
SRA KDD Explorer Suite includes algorithms for discovering associations, classifications, sequences, and clusters. Offered by Knowledge Discovery Solutions.	- Multi-strategy data mining algorithms for discovering Associations, Classifications, Sequences, and Clusters. -Claim 6 month turnaround time	$2,000	Launched in 1997.	Partners with Oracle, Microstrategy, and Sequent

Table 9 Continued:

WizRule and WizWhy find rules in data. Offered by WizSoft Incorporated.	-Report rules and cases deviating from normal. -Display content of deviated cases. -User-friendly interface based on Windows 95 standard.	WizRule: $1395 WizWhy: $3995	WizRule Version 3.5 was released in September 1999. WizWhy was released in 1999.	Irvine University
Xpertrule Miner 4.0, Attar Software's next generation product evolved from the established Profiler scalable client-server data mining software.	-Miner includes extensive data transformation, visualisation and reporting features - Solutions can now be built as stand-alone mining systems or embedded in other vertical applications under MS-Windows. -3 day course at customer site. -The number of rows/records is limited only by the server (2 billion max)	As of January 2000, $4995	Launched in 2000.	The Johnson & Johnson Consumer Products Company
Free Products:	Main Features	Price	Released Date	MainClients/ Partnerships Strategy
CBA, mines association rules and builds accurate classifiers using a subset of association rules.	-Developing algorithms at the National University at Singapore.	N/A	Currently still in progress.	Project funded by National Science and Technology Board – NSTB.
MDEP, a program for the discovery of multi-valued dependencies from relations	-The program MDEP includes implementations of the top-down and bottom-up algorithms.	N/A	Deployed 1999.	

Table 10: Software Tools for Sequential Patterns Knowledge Discovery

Commercial Products:	Main Features	Price	Released Date	MainClients/ Partnerships Strategy
CaprI (part of Clementine package) discovers different types of sequences across records (and time).	- CaprI is a data-mining algorithm that discovers sequences across time.	See Table 3.		
EasyMiner includes support for different types of sequences, including clickstreams, and concept hierarchies. Offered by MINEit Corporation.	-Optimises web sites for maximum commercial impact by understanding the dynamic behavior of web site visitors. -Tries to accurately determine the success of on-line marketing campaigns and effectively gauge the ROI of banner advertising.	-Starting at $1500	1999.	Partnerships with corporations such as Value Added Resellers (VARs), Systems Integrators and Consultancies, OEMs and Independent Software Vendors (ISVs).
IBM Intelligent Miner Suite, includes tools for discovery of associations and sequential patterns	See Table 3.			
Decision Series Suite (now part of Accrue), includes algorithms for sequence discovery. Offered by NeoVista Software.	- Model building using these sophisticated modeling tools is accessible to an even broader range of corporate decision makers–from the casual user to the expert. - Release 3.0 introduces three new data mining engines, DecisionKmeans, DecisionBayes and DecisionCubist.	The cost for Decision Series 3.0, depending on options, ranges from $25,000 to $200,000.	Release 3.0 launched in the 4th quarter of 1998.	Privately held, NeoVista is backed by leading venture capital investors including GE Capital Corporation, Index Securities S.A., Kleiner Perkins Caufield & Byers, New York Life Ventures, Perot Systems Corporation, and U.S. Trust Company of New York.

Table 10 Continued:

HMMpro 2.2, biological sequence analysis software. Offered by Net-ID Incorporated.	-Net-ID developed HMMpro to mine the wealth of biological information generated by genome and other sequencing projects. -HMMpro is a general purpose hidden Markov model (HMM) simulator for biological sequence analysis. It uses machine-learning techniques to automatically build statistical models of protein and DNA sequences.	Estimated at $1000	Launched in 1998.	Services include: -Consulting -Software development -Collaborative research and development -Construction of protein or DNA HMM libraries based on public and/or proprietary data.
SAS Enterprise Miner, an integrated software product that provides an end-to-end business solution for data mining.	See Table 3.			
SRA KDD Explorer Suite, includes algorithms for discovering associations, classifications, sequences, and clusters.	See Table 9.			

Table 10 Continued:

Free Products:	Main Features	Price	Released Date	MainClients/ Partnerships Strategy
WUM, finds clickstream sequences. Offered by Humboldt University at Berlin.	-WUM is a sequence miner. Its primary purpose is to analyze the navigational behaviour of users in a web site, but it is appropriate for sequential pattern discovery in any type of log. It discovers patterns comprised of not necessarily adjacent events and satisfying user-specific criteria.	-WUM is a powerful sequence miner: it can discover complex sequence patterns and has a friendly interface for guiding the mining process. - Web usage analysis is important for several applications, including educational hypermedia and electronic commerce.	Released in 1999.	-No Major Partnerships. Research-based study.

Classifiers and Regression:

Classification and regression are two common types of problems to which data mining is applied today [15]. Classification and regression analysis is used to predict customer behavior, to signal potentially fraudulent transactions, to predict store profitability, and to identify candidates for medical procedures to name a few of the applications in which this kind of algorithm can be used. Classification and regression trees (CART) methodology concerns the use of tree-structured algorithms to classify data into discrete classes. Breiman invented the terminology in the early 1980's [15]. The technique has found uses in medical, market research statistics, marketing and customer relations. For example, one tree-structured classifier uses blood pressure, age and previous medical conditions to classify heart patients as either risky or not. Another tool might use age related variables and other demographics to decide who should appear on a mailing list. Predicting direct mail response and identifying ways to control customer attrition in the telecommunication industry are other industry-specific applications. The variety of applications that use classification tools is plentiful.

The difference between classification and regression is the type of output that is predicted. Classification predicts class membership. For example, a model predicts that John Doe, a potential customer, will respond favorably to an offer. The predicted output (class) is therefore categorical; where categorical is a variable that has few possible values, such as "Yes" or "No", or "Low," Middle," or "High [15]." In contrast, regression predicts a specific value. Using the above example, this model predicts that John Doe will purchase $900 in products the upcoming year. Regression is used over classification where the predicted output can take on many possible values. Regression

and classification problems are really the same problem looked from two angles. It is a matter of notation and keeping track of what the variables' values represent. For example, to convert from a regression problem to a classification problem, to convert John Doe's response of potential monetary purchases of $900 into a classification table, a table could be set up such that all customers purchasing under a $1000 in products correspond to a value of "Low", while those from $1000-$2000 correspond to "Middle", while those over $2,000 correspond to a value of "High." In general, a regression problem can be turned into a classification problem by establishing categories in which certain set of values classify to every category. Looking at Figure 6 can show a further example of the difference in output modeling between a classification and regression algorithm.

Figure 6: Classification vs. Regression Modeling

Modelling

Given some examples $(y_1, x_1),\ldots,(y_i, x_i)$

Classification
$x \in R^n,\ y \in \{-1, +1\}$

Alcohol	Cigarettes	Retired
100	0	No
30	70	No
2	0	Yes
...	...	...

Regression
$x \in R^n,\ y \in R$

Alcohol	Cigarettes	Age
100	0	40
30	70	53
2	0	74
...	...	...

Classifiers are used to tag the data to be analyzed. Tagging data can include assigning attributes to particular characteristics of the data to produce classes of records of these data. Then the function of classifiers is to produce results based on these tagged classes of records. Thus far, models have used this schema to produce results including linear regression models, decision tree models that are based on rules and as neural network models. An example of a classifier might include a company, which has records about its customers. By creating a classification model, the company might tag its customers as those being loyal, indifferent, and disloyal customers. A classifier can then sort through this list of customers and product results about the trends of these customers based on these classes.

Furthermore, classifiers can be either explicit or implicit in form. Explicit classification refers to a set of rules to describe each of the classes of records found in a database. Implicit classification refers to mathematical functions, such as neural networks, that takes these sets of records and uses them as inputs to the functions to generate the classes of records. As such, various techniques are available for classification and regression problems. Each technique uses its own distinct model, however each technique generates a predictive model based on historical data [15]. The model then predicts a set of new outcomes of cases.

Predictive Modeling Techniques for Classification and Regression Tools:

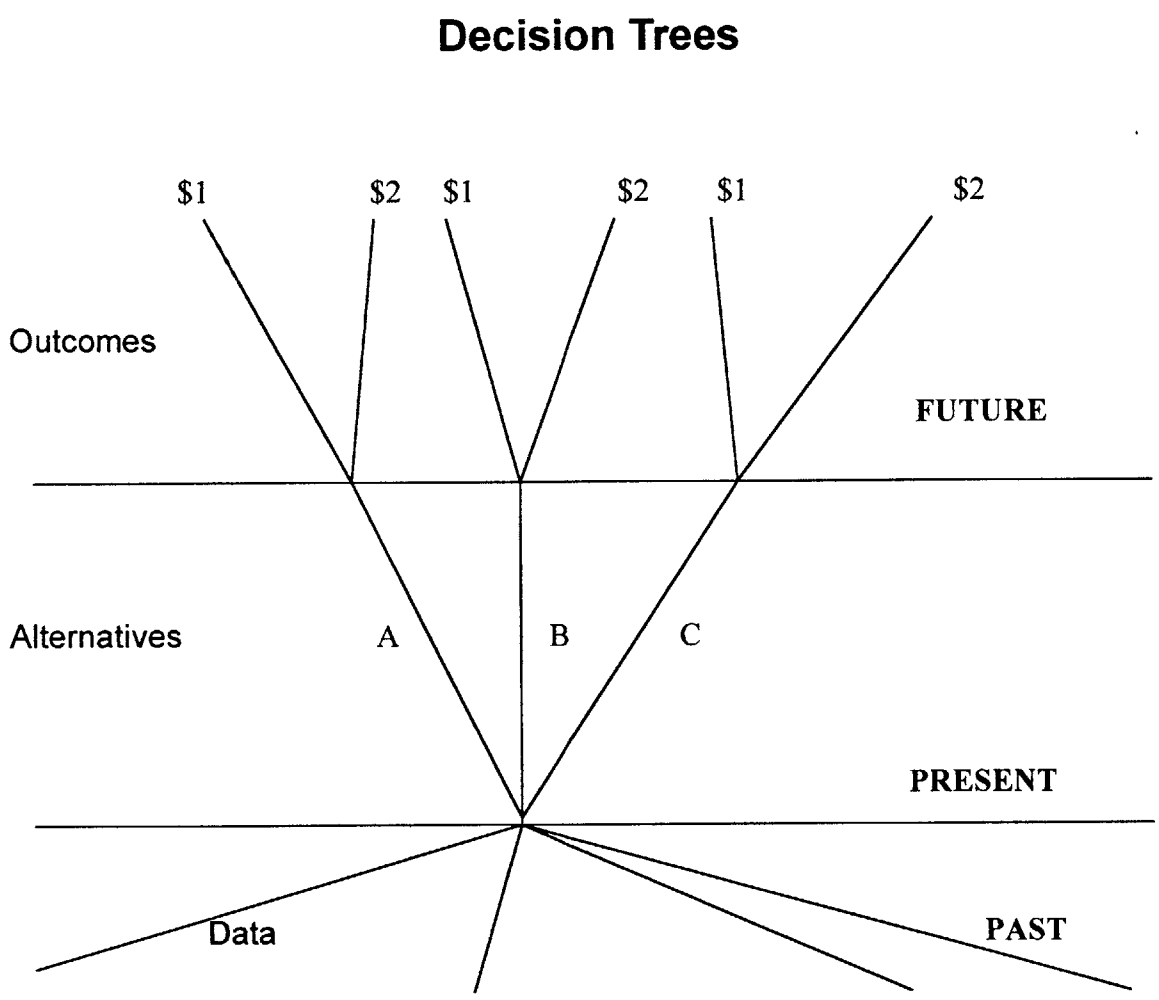
- Decision tree
- Neural Networks

A decision tree is a technique that produces a graphical analysis of the model it produces. The graphic output consists of a tree with nodes denoting decision points. The

decision tree method encompasses a number of specific algorithms including Chi-squared Automatic Interaction Detection (CHAID), Classification and Regression Trees (CART), C4.5 and C5.0. A decision tree is a model that is both predictive and descriptive. Decision trees are commonly used for classification but can also be used for regression analysis. Decision trees are advantageous tools for making corporate or financial decisions where a lot of complex information has to be considered. Decision trees provide a functional framework in which alternate decisions and the ramifications of making those decisions can be laid down and evaluated. Decision trees also help in forming a balanced, accurate picture of the risks and rewards that can result from a particular decision.

Decision trees separate data into sets of rules that are likely to have a different effect on a target variable. For instance, finding the characteristics of a person likely to respond to a direct mail piece can be an important consideration for an organization planning on a new promotion through mail. These characteristics can be translated into a set of rules. Typically this is done by looking for combinations of demographic variables that best distinguish those households who responded to the previous promotion from those that did not.

Figure 7: Decision Trees



Source: The illustration above is taken from: JL Riggs, Production Systems, Wiley, 1987. P208.

The illustration in Figure 7 shows a generic decision tree - rooted in the past, events and data come to a node in the present and then branch into a multitude of possibilities in the future. The decision tree can be used as a model simply to explain the complexity inherent in planning, prediction and strategic thought. It is also used to map future possibilities and alternatives. Thus, the primary output of a decision tree is the tree itself. The training process that goes about creating the actual tree and its corresponding nodes that tie past and present nodes together is called induction. In the above example,

each node represents a possibility or an outcome. A decision tree grows from a root node splitting the data at each level to form new nodes [80]. Branches in turn connect these nodes. Usually, the term given to nodes at the end of a branch is called leaf nodes. As such, the tree is a good way to break up the realm of possibilities into discrete points in time where a single path trace will result in a particular outcome. The more complex a tree gets, the more nodes, leaf nodes, and branches it will have.

Decision tree algorithms usually go through two phases: a tree-growing phase (splitting) followed by a pruning phase. The tree-growing phase consists of an iterative process that involves splitting the data into progressively smaller subsets [15]. Each iteration considers the data in only one node with the first iteration considering the root node that contains all data. Subsequent iterations will work on derivative nodes that will contain subsets of the data. The algorithm will work based on the defined sets of variables. Usually the data will consist of independent and dependent variables. The algorithm first begins by analyzing the data to find the independent variables used as a measure for the splitting rule that will result in nodes that are most different from each other with respect to the dependent variable [81]. The second phase, pruning, consists of making a tree more general by removing splits and the sub-trees created by them. This is done by comparing the performance at each node (measured by the accuracy between the independent and dependent variables) and determining based on the accuracy which nodes need to be pruned to attain the highest level of performance.

Decision trees are very popular classification tools [15]. Many users of decision trees find the tools easy to understand and use. As a result, more users trust decision tree models than they trust "black box" models such as those produced by neural networks

[15]. The future trend of decision tree vendors will be to continue to improve decision tree algorithms. In addition to improving algorithms, vendors are expanding their user interfaces to include tree graphical views with expanding and collapsing nodes that facilitate user exploration. Currently, there are a significant number of players in the classification software data mining industry. A listing is provided in Table 11.

Table 11: Software Tools for Decision Tree Classification Analysis

Commercial:
AC2, provides graphical tools for data preparation and builing decision trees.
Alice d'Isoft 6.0, a streamlined version of ISoft's decision-tree-based AC2 data-mining product, is designed for mainstream business users.
Business Miner, data mining product positioned for the mainstream business user.
C4.5, the "classic" decision-tree tool, developed by J. R. Quinlan
C5.0/See5, constructs classifiers in the form of decision trees and rule sets. Includes latest innovations such as boosting.
CART, decision-tree software, combines an easy-to-use GUI with advanced features for data mining, data pre-processing and predictive modeling.
Cognos Scenario, provides quick identifying and ranking factors that have a significant impact on key business variagles.
Decisionhouse, provides data extraction, management, pre-processing and visualization, plus customer profiling, segmentation and geographical display.
Kernel Miner, decision-tree-based classifier with fast DB access (2nd Place in KDD'99 CUP classifier contest)
KnowledgeSEEKER, high performance interactive decision tree analytical tool.
SPSS AnswerTree, easy to use package with four decision tree algorithms – including CHAID, and CART.
Free:
EC4.5, an efficient version of c4.5, which uses the best among three strategies at each node construction.
IND, provides CART and C4.5 style decision trees and more. Publicly available from NASA but with export restrictions.
LMDT, builds Linear Machine Decision Trees (based on Brodley and Utgoff papers).
ODBCMINE, shareware data-mining tool that analyzes ODBC databases using the C4.5, and outputs simple IF..ELSE decision rules in *ascii*.
OC1, decision tree system continuous feature values; builds decision trees with linear combinations of attributes at each internal node; these trees then partition the space of examples with both oblique and axis-parallel hyperplanes.
PLUS, Polytomous Logistic regression trees with Unbiased Splits, (Fortran 90).
SE-Learn, Set Enumeration (SE) trees generalize decision trees. Rather than splitting by a single attribute, one recursively branches on all (or most) relevant attributes. (LISP)

Neural Networks:

Artificial neural networks are a system loosely modeled on the human brain [17]. Neural networks are an attempt to simulate within specialized hardware or sophisticated software, the multiple layers of simple processing elements called neurons [17]. Each neuron is linked to certain of its neighbors with varying coefficients of connectivity that represent the strengths of these connections. Learning is accomplished by adjusting these strengths to cause the overall network to output appropriate results. The basic unit of neural networks, the artificial neurons, simulates the four basic functions of natural neurons. Artificial neurons are much simpler than the biological neuron; Figure 8 shows the basics of an artificial neuron.

Often, neural networks are considered to be "black boxes" because of their non-linear behavior and are usually more complicated than other techniques. Neural networks can be as effective as decision trees for classification and regression [15]. Training a neural network is a further challenge requiring setting numerous parameters. Furthermore, the output of a neural network is not as easily understood by the user as the output seen by a decision tree tool. Neural networks are used for feature extraction, association, optimization, function fitting, and modeling. Neural networks can be divided into two major categories:

- Supervised Neural Networks
- Unsupervised Neural Networks

Supervised Neural Networks consist of input patterns associated with known output patterns, while unsupervised Neural Networks have structures found in the input patterns.

Origins of Artificial Neural Networks

Artificial neural networks are the result of academic investigations that involve using mathematical formulations to model nervous system operations. The resulting techniques are being successfully applied in a variety of everyday business applications. Neural networks represent a meaningfully different approach to using computers in the workplace. A neural network is used to learn patterns and relationships in data. The data may be the result of a number of studies in which hidden patterns embedded in the data may have gone unnoticed. Regardless of the specifics involved, applying a neural network is a substantial departure from traditional approaches.

Neural networks offer overwhelming advantages in that raw data can simply be fed to a network and users can view the network as a "black box" that computes an output. This output based on the training of the neural network and the input data fed into it, will hopefully result in knowledge discovery for the data miner. Traditionally, a programmer would specifically have to code every facet of the problem in order for the computer to understand the situation. Neural networks do not require the explicit coding of the problem. For example, to generate a model that performs a sales forecast, a neural network only needs to be given raw data related to the problem. The raw data might consist of: history of past sales, prices, competitors' prices, and other economic variables. The neural network sorts through this information and produces an understanding of the factors impacting sales. The model can then be called upon to provide a prediction of future sales given a forecast of the key factors.

These advancements are due to the creation of neural network learning rules, which are the algorithms used to learn the relationships in the data. The learning rules enable the network to gain knowledge from available data and apply that knowledge to

assist a manager for example in making strategic decisions based on the results of the neural network analysis [18].

Neural Networks Main Uses:

Neural networks constitute a powerful tool for data mining. Applications of neural networks are numerous. Organizations have more and more data from which they need to extract key trends in order to run their businesses more efficiently and improve decision-making. Many organizations receive their first introduction by reading about the applications of the techniques in financial market predictions. Other successful applications of the techniques include: analysis of market research data and customer satisfaction, industrial process control, forecasting applications, and credit card fraud identification [1]. Currently, banks are installing neural network technology to detect credit card fraud. The realized savings are expected to pay for the new system sometimes in as few as six months [92]. These systems are able to recognize fraudulent use based on past charge patterns with greater accuracy than other available methods. Another example of using neural networks is to improve decisions in medical diagnosis. A neural network can be shown a series of case histories of patients, with a number of patient characteristics, symptoms, and test results. This process is often referred to as training the neural network. Once this process is done, the network can also given the diagnosis for a particular case from the attending physician. The network can then be shown information regarding new patients and the network will provide a diagnosis for the new cases. This essentially creates a system containing the expertise of numerous physicians, which can be called upon to give an immediate initial diagnosis of a case to medical personnel.

Given a steady increase in successful applications, neural networks offer substantial benefits. Successful applications share certain common characteristics that may be easily understood. First, there will exist interrelationships between the explanatory factors that are used to estimate potential outcomes. Having interrelationships in the data means that two or more factors work together to predict model outcome. For example, a chemical process in a production facility may be dependent on temperature and humidity. These two factors combine to affect the outcome of the process. The second condition in which neural networks excel is when there is a non-linear relationship between the explanatory factors and the outcome. This implies that the nature of the relationship between the factors and the outcome changes as the factors take on different values.

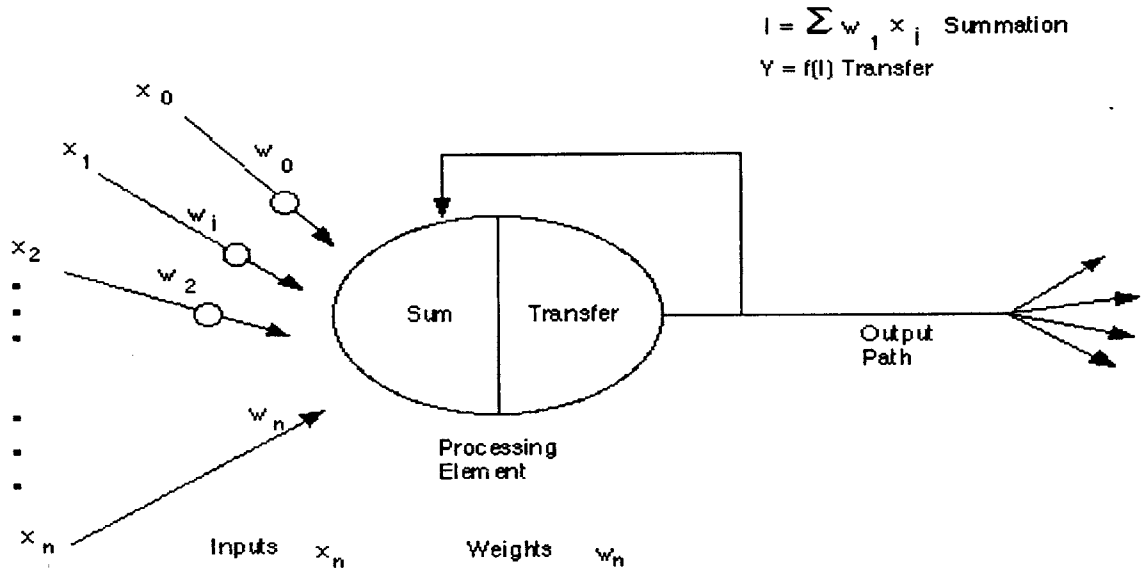
The need to improve processes by doing things better and cheaper is more important than ever in today's competitive business climate [18]. Likewise, the desire to develop computer systems that can learn by themselves and improve decision-making is an ongoing goal in AI. The neural network techniques used today may be vastly different in years to come. However, the goal of developing computers that learn from past experience and lead to better business decisions will remain a high priority. Neural networks now represent one of the best practices in achieving this goal. Furthermore, continued achievements toward this goal are likely to be inspired or generated from these technologies.

Neural networks for the most part are practical and cost effective, although finding documented proof of this can be a challenge. It is true that the techniques are relatively new and that experience with these techniques is not as extensive as with

traditional techniques [91]. Despite this, neural networks are proving their worth everyday in a wide variety of business applications, and saving their users time and money in the process.

Neural networks should be applied in situations where traditional techniques have failed to give satisfactory results, or where a small improvement in modeling performance can make a significant difference in operational efficiency or in bottom-line profits. Direct marketing is an excellent example of where a small improvement can lead to significant results. The response rate on direct marketing campaigns is usually quite low. A five percent response rate is often considered very good [18]. By reviewing the demographic data on those that respond it may be possible to identify characteristics that would produce a 6% response rate. If a neural network is used to analyze the demographic characteristics and a 7% response rate is produced, then the cost of the direct mail campaign can be reduced while maintaining the same desired level of positive response from prospects. The bottom line is that any manager interested in getting more useful information from available data should consider neural network technology as an option. They can be used by aggressive organizations to focus available resources more effectively, thus gaining a valuable competitive edge.

Figure 8: Representation of the mathematics of a neural network



Source: The illustration above is taken from: Klerfors, Daniel, Artificial Neural Networks, Saint Louis University, November 1998.

The above figure of a neural network shows that the various inputs to the network are represented by the mathematical symbol, x(n). Each of these inputs in turn is multiplied by a connection weight, these weights are represented by w(n). This weight attaches a credibility value to each input indicating how relevant each input can be in helping determine the optimal solution. In the simplest case, these products are simply summed, fed through a transfer function to generate a result, and then this result is fed as an output [17]. Even though all artificial neural networks are constructed from this basic building block the fundamentals may vary in these building blocks and there are differences.

Table 12 shows a listing of all major software products currently available for classification neural network tools. Artificial neural networks are one of the promises for the future in data mining [19]. They offer an ability to perform tasks outside the scope of

traditional algorithms. They can recognize patterns within vast data sets and then generalize those patterns into recommended courses of action. The main advantage of a network over traditional practices is that neural networks learn and are not programmed. Yet, even though they are not traditionally programmed, the designing of neural networks does require a skill. It involves understanding the selection of learning rules, transfer functions, summation functions, and how to connect the neurons within the network [18]. Because neural networks can be trained to recognize patterns in data up to several orders of magnitude in dimensions more than a human being can, neural networks are promising tools for the future. As technology speeds up and hardware accessories become cheaper and faster, neural networks can be able to become the intelligence behind technology that never tires nor becomes distracted.

Table 12: Classification Software Tools for Neural Networks

Commercial Products:	Description	Algorithm	Platform	Vendor	Price
Adaptive Logic Network (ALN)	-ALN is a commercial package that does least squares fitting to arbitrary data using linear pieces, and allows the user to input a priori knowledge prior to training. -A new *relational paradigm* is used which eliminates the need for coding real values into Booleans.	Adaptive Logic Networks	MS-Windows	Dendronic Decisions Limited	$100
Attrasoft Boltzmann Machine (ABM)	-Version 2.3 simulates neural networks. In this version, an interface is introduced which translates data to and from neural data.	Boltzmann Machine	Windows 95	Attrasoft	$99
Attrasoft Predictor	- Attempts to predict the Stock Market, Lottery Numbers, Markov Chains, and Dynamic Systems	N/A	Window 95, NT, and 3.1x	Attrasoft	$99
Braincel	- Braincel is an add-in to Excel for MS Windows 3.1 and up.	N/A	MS-Windows	Promised Land Technologies	$249
BrainMaker	- BrainMaker allows for business and financial forecasting, pattern recognition, medical diagnosis, sports handicapping.	Backpropagation	Windows (3.x, 95, NT), DOS, Macintosh	California Scientific Software Corporation	$195

Table 12 Continued:

BrainSheet	-Claims to predict future results based on past data, and analyses hidden relationships in data.	Backpropagation	Windows 95	BitStar International	$189
Clementine	-The Clementine technology supports a data mining approach that allows users to match data analysis and modeling technologies to their specific business questions.	MLP, RBF, SOM, K-means	Windows NT, Unix	SPSS Inc.	See Table 3.
DataEngine	-Tool for data analysis with fuzzy logic, neural networks, and statistical methods. Applications of DataEngine have been performed in the areas of e.g. quality control, process analysis, forecasting, and diagnosis.	N/A	i486 and Pentium PCs running Windows 3.1, 95, NT	Management of Intelligent Technologies	Euro 5000-$4575

Table 12 Continued:

DataMining Workstation (DWM) and DWM/Marksman	-DWM is a broadly focused neural network application that includes hardware and software. This product allows users to configure their own neural network to make full-scale predictive model. Typical applications include credit scoring, customer profiling, consumer payment behavior, etc.. DWM/Marksman is a later iteration of DMW and provides data analysis for direct marketing applications.	N/A	Windows NT	HNC Software Inc.	$100
DynaMind Developer Pro	-Emulates the ETANN chip and iDynamind versions.	Backpropagation, Madeline III	PC, Macintosh	InfoTech Software Engineering	$150
ECANSE - Environment for Computer Aided Neural Software Engineering	-Designed to develop HYBRID SYSTEMS. -Combines innovative approaches like Neural Nets, Fuzzy Logic, and Nonlinear Analysis using Chaos Theory, Genetic Algorithms to form a single, easily extendible tool with a common user interface.	N/A	UNIX (Sun, HP, SGI) and PC (Win NT)	Siemens Austria	$99

Table 12 Continued:

FlexTools	-Development of soft computing modules that integrate techniques such as fuzzy systems, evolutionary algorithms, neural networks, and chaos theory.	Neural Networks, Fuzzy Logic, Evolutionary Computation	Windows and Matlab	Flexible Intelligence Group	$499
KnowMan Basic Suite	- An expert system builder and data mining application based on n-tuple network techniques. Expert systems on the World wide Web	Enhanced n-tuple network. Self-optimizing system.	Windows 95, Windows NT	Intellix A/S	$200
Matlab: Neural Network Toolbox	-Provides a complete neural network-engineering environment within MATLAB. -Applications include signal processing, nonlinear control, and financial modeling.	-Perceptron, linear, backpropagation, Levenberg-Marquard, radial basis, Elman, Hopfield, learning vector quantization, Hebb, Kohonen, competitive, feature maps, and self-organizing maps	Matlab running under Windows, MacOS, UNIX	MathWorks	$900
Neural Connection	- -Efficient for market research, database marketing, financial research, operational analysis and scientific research. -Its tools for intelligent data analysis solve problems in prediction, classification, and time series analysis.	Multilayer perceptron with conjugate gradient, radial basis function, Bayesian neural network, Kohonen network (SOM).	Windows 3.1 or later	SPSS Inc.	$1949

Table 12 Continued:

NeuralWorks	-Developing environment for deploying real-time applications in forecasting, modeling and classification automatically.	Backpropagation, ART-1, Kohonen, Modular NN, General Regression, Fuzzy ART-map, Probabilistic Nets, Self-Organizing Map, LVQ, Boltzmann, BSB, SPR, etc.	PC, Macintosh, Sun, IBM RS/6000, SGI, Dec, NEC EWS-4800, HP 9000/700	NeuralWare Inc.	Profesional package: $1995
NeuroSolutions	-Combines a modular, icon-based network design interface with an implementation of advanced learning procedures, such as recurrent backpropagation and backpropagation through time.	Educator Product: MLP, Generalized Feedforward -Consultants Product: Time-Delay (TDNN), Time-Lag Recurrent (TLRN) -Developers Product: Dynamic Link Libraries	Windows 95, Windows NT	NeuroDimension Inc.	Educator: $195 Consultants: $995 Developers: $1495
STATISTICA: Neural Networks	-Comprehensive application capable of designing a wide range of neural network architectures, employing both widely used and highly specialized training algorithms.	-Back Propagation, Levenberg-Marquardt, Conjugate Gradient Descent, Sub-sampling, K-Means, K-Nearest Neighbour, Isotropic Deviation assignment, PNN training, GRNN training, Genetic Input Selection	Windows 95, NT, 3x	StatSoft Inc.	$1095

Table 12 Continued:

Trajan	-Includes support for a wide range of Neural Network types, training algorithms, and graphical and statistical feedback on Neural Network performance. -Uses a 32-bit architecture.	Levenburg-Marquardt, Conjugate Gradient Descent, Back Propagation, Quick Propagation, K-Means, K-Nearest, Principal Components Analysis, Automatic Network Design, Neuro-Genetic Input Selection, Weigend Weight Regularisation.	PC and Unix	Trajan Software LTD	$795

Visualization:

A significant implicit model frequently used is that of visualization. Visualization provides analysts with visual summaries of data from a database [1]. This can be extremely useful to a data mining tool that is trying to understand features that are difficult to understand when working with numbers but obvious when using graphical tools. This technology facilitates the ability to quickly and easily change the type of information displayed and the particular visualization chosen. Furthermore, visualization can help provide insight to a smaller subset of the data that may have gone unnoticed by the rest of the data. New sets of powerful data visualization tools have appeared in the marketplace and in the research community. This, combined with readily available computer memory, speed, and graphics capabilities, makes it possible to explore larger and larger data sets [19]. Visualization is easier for an analyst as he or she can notice any interesting phenomena by the visuals provided by this method instead of trying to pose questions and try to find answers to those postulated questions. This method allows the data miner to notice any interesting or unexpected phenomena through the ease of detailed displayed visuals. Visualization tools take advantage of human perception as a method for analysis. What numbers cannot show, corresponding pictures often can. For example, a linear trend in data might not be evident from a table of data. However, a scatterplot, which shows a series of points lined up on a straight line, provides immediate insight into data relations. Visualizing tools can therefore be multidimensional, displaying for the data miner a graphical representation of results. Such a tool can yield in significant findings as the software tool visually displays all-important changes to

significant parameters in solving the particular problem in real time. Figure 9 shows a sample output of a visualization tool, Advizor, from Visual Insights. Furthermore, the use of one of these data mining techniques discussed thus far can be used in conjunction with other tools to yield in maximum knowledge discovery. For example, the IBM Research tool SONAR has come up with a decision tree visualization tool, which shows how a decision tree can be embedded within a visualization tool to discover hidden patterns using the decision tree algorithm and displaying these results using a visualization tool that is easier for the user to interpret. Using a regression tree, which was trained using the New York markets data, the user is provided with a visualization tool in addition to the regression tree analysis tool using SONAR. The data contains numeric attributes such as "BPS" (British pound sterling), "GDM" (Deutschmark), "YEN" (Japanese yen), "TB3M" (3-month Treasury bill yields), "TB30Y" (30-years Treasury bill yields), "GOLD" (gold price per ounce) and "SP500" (Standard and Poors index). The tree is constructed so that variance of the SP500 is minimized. Visually portrayed through SONAR is the overall distribution of the entire database starting at the root node to the inner and outer nodes containing graphical distribution information at those particular nodes.

Table 13 shows the current leading providers of software visualization tools used in the knowledge discovery process. In particular, notice that Alta Analytics Netmap is a visualization tool specific to fraud detection applications. Furthermore, Siftware SphinxVision includes tools such as classification and clustering in addition to providing visualization. As such, this software product combines several knowledge discovery tasks.

Figure 9: Sample 3D visualization of Advizor provided by Visual Insights

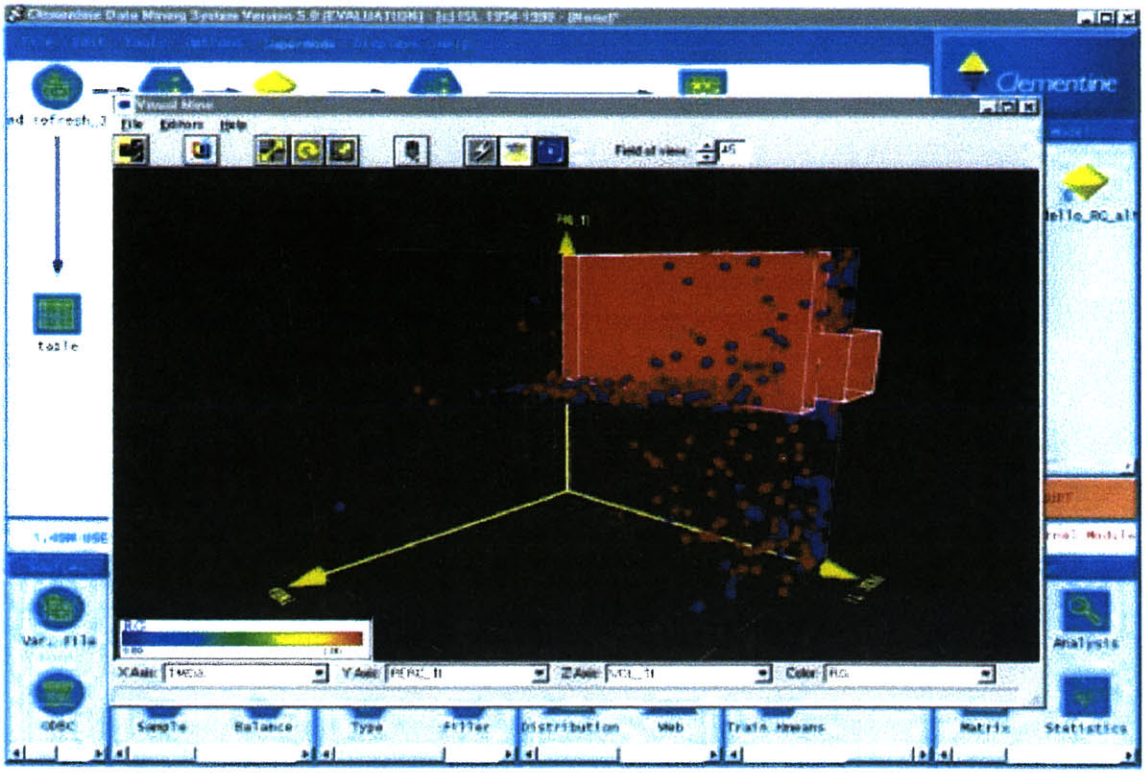


Table 13: Software Tools for Visualization Knowledge Discovery

Commercial:
CViz Cluster Visualization, a visualization tool designed for analyzing high-dimensional data in large, complex data sets.
Daisy, a graphical analysis and interactive investigation program (available as OCX).
DataScope 3.0, visualization-oriented tool. Offers 3D Views, Hierarchical Organization and Automatic Detection of Data Relationships.
High Tower TowerView presents huge quantities of data (up to 10,000 different parameters) in a three-dimensional graphical environment.
JWAVE, web-based software that allows you to quickly access and understand what your data means.
NETMAP, innovative combination of link analysis and data visualization, with applications to fraud detection and claims analysis.
SphinxVision, data mining tool using meta-net technique and highly performing interactive graphics. Discovery tasks include: Classification, Clustering, Visualization
Spotfire.net, the core offering of visualization provided by Spotfire Inc.
Thinx, a 32bit Windows application that is becoming the standard by which users connect graphics with data provided by Thinx Software.
VDI Discovery for Developers, 3-dimensional visualization toolset provided by Visible Decisions
Viscovery SOMine, a data mining system with powerful visualization features based on Self-Organizing Maps (Windows).
Visual Insights Advizor, a complete visual query and analysis based decision-support application. It combines patented interactive data visualization components with application templates.
VisualMine, the advanced 3D visualization gives new possibilities of analyzing large amounts of information, supporting data analysts and decision makers.
VRCharts, an interactive 3D Data Visualization Software for widespread business use.
WinViz, a data analysis tool utilizing visualization and supporting classification applications.
Free:
Graf-FX, a data mining toolkit that uses Microsoft Access for graphs/queries of tables/queries.
IRIS, the system provides an intelligent assistance in visual data exploration by the means of automatical generation of data presentation on maps. Used primarily for geographic discovery.

Table 13 Continued:

VisDB, a visual data mining and database exploration system.
Xmdv, software visualization package tool.

Clustering:

Clustering is an important technique in exploratory data analysis, with applications in image processing, object classification, target recognition, data mining etc. The aim is to partition data according to natural classes present in it, assigning data points that are "more similar" to the same "cluster". In clustering, no data are tagged before being fed to a function. Instead, the input to a clustering function is a collection of untagged records. The goal then of clustering, through a pre-established criterion, will sift through the data to produce a segmentation of the input records. Different clustering functions will hence yield different sets of sorted data.

Clustering techniques also work for managing risk -- against either unprofitable research ventures or even losses in the stock market. Marketing managers like the technique for identifying customer populations they may want to target, such as with a special promotion. Clustering tools and neural networks are by far the most popular learning algorithms because of their adaptive, predictive capabilities. Table 14 lists some of the leading commercial tools in today's market.

Table 14: Software Tools for Clustering Techniques

Commercial:
ClustanGraphics3, hierarchical cluster analysis from the top, with powerful graphics
CViz Cluster Visualization, for analyzing large high-dimensional datasets; provides full-motion cluster visualization.
Darwin 3.6 Suite , new features include clustering tools
IBM Intelligent Miner for Data, includes clustering algorithms
SGI MineSet, 3.0 offers associations, decision trees, statistics, data transformation, clustering, and visualization tools.
SOMine, a user-friendly Self-Organizing Map tool
Free:
Autoclass C, an unsupervised Bayesian classification system from NASA, available for Unix and Windows
ECOBWEB, a concept formation program for the creation of hierarchical classification trees (in Common Lisp).
MCLUST/EMCLUST, model-based cluster and discriminant analysis, including hierarchical clustering. Developed in Fortran with interface to S-PLUS.
Snob, a clustering tool, using an unsupervised concept learning and mixture modeling that can deal with missing data.

Evaluation of Data Mining Tools:

In evaluating which data mining tools are most efficient, no major studies have been undertaken to show a particular data mining technique being more efficient in others. There is little understanding in the performance in so many of these data mining techniques along with the features that they have to offer [79]. A recent study was conducted in which several leading software tools under specific data mining applications were compared to other applications using different software tools. The experiment consisted of obtaining a range of tools from leading vendors and comparing them to one another. The four technologies under experimentation were decision trees, rule induction, neural networks, and polynomial networks. The decision tree products analyzed in Table 14 consisted of CART, Scenario, See5, and S-Plus [79]. The rule induction tools WizWhy, Datamind, DMSK. Neural networks consisted of three programs with different algorithms coming from NeuroShell2, PcOLPARS, and PRW [79]. The polynomial network tools used were ModelQuest Expert, Gnosis, NeuroShell2, and KnowledgeMiner.

The basis for comparison amongst these tools were derived from the following variables:

- **Capability**
- **Learnability/Usability**
- **Interoperability**
- **Flexibility**
- **Accuracy**

Capability measures what a tool can do and how well it does it. Learnability and Usability refer to how easy it is to learn that particular tool along with using it. Interoperability refers to how a particular tool is able to interact with other computer applications. Flexibility denotes how easy it is to change parameters or change the working environment of an application. Finally, accuracy reflects to how a particular tool was able to perform for the data fed into the tools. The kind of data that was fed into each one of these applications was classification data sets. Three different sets of classification data were created and transmitted to each of these applications. Performance was then measured based on all these variables and how they performed as an aggregate performance with the three sets of data. Data were randomized in the sense that one classification set was about Breast Cancer, another containing Diabetes Patients' Information, and information about glasses.

With the result shown in Table 15, this demonstrates the assessment of each of the tools grouped according to tool technology (Y-Axis) and evaluation category (X-Axis). The results were such that the best tool under each data mining application were the following: S-Plus for decision trees, Datamind for Rule Induction, PRW for neural networks, and ModelQuest Expert with NeuroShell 2 for polynomial networks [79]. Though the results weigh a particular tool over another, the results for this experiment yielded in a specific answer as to which tool is better suited to solve the collection of data sets that were fed into each application. However, before a conclusion can be reached as to which application may be more efficient, a more relevant question is to pose what the goal is of each individual application. Undoubtedly some applications will be better at doing certain tasks over other applications. This is the challenge of data mining. There is

not a unique answer to a problem. Much time needs to be spent to ensure that a proper tool was selected. Though performance can be quantified to some extent, performance will be tied into what the problem at hand to be solved is all about and what kinds of data are being fed into the application. As there are so many variables to quantify, there is no explicit answer as to which tool is better suited to potential clients at large. The problem needs to be defined, and based on what the answer the problem seeks, then and only then can tools start being quantified. This example showed that applications can be quantified but only to the extent that the data is defined along with the range of tasks a data mining application is supposed to undergo [79].

Table 15: Data Mining Tool Evaluation Summary

Technology	Tool	Capability	Learnabilit/ Usability	Interoperability	Flexibility	Accuracy	Overall	Price ($)
Decision Trees	CART	+	√	-	√ +	+	√ +	995
	Scenario	√ -	+	+	-	--	√	695
	See5	√	√ -	√	√ -	+	√	440
	S-Plus	+	√ -	++	+	+	+	1,795
	Tree Average	√	√	√ +	√	√ +	√ +	Median=845
Rule Induction	WizWhy	√	√ +	√	√ -	-	√	4,000
	Datamind	√ +	++	+	√ -	√	√ +	25,000
	DMSK	-	--	√ -	-	+	-	75
	Rule Average	√	√	√ +	-	√	√	Median=4000
Neural Networks	Neuroshell 2	-	√	-	-	++	√ -	395
	PcOLPARS	√ -	-	-	√ -	√	√ -	495
	PRW	√ +	+	++	√	++	+	10000
	Neural Average	√ -	√	√	√ -	+	√ -	Median=495
Polynomial Network Tools	MQ Expert	+	√	√	√ +	+	√ +	5,950
	Neuroshell 2	√ -	√	√	√	+	√	495
	Gnosis	√ -	√	--	√ -	++	√ -	4,900
	Knowledge Miner	-	-	-	√ -	+	-	100
	Poly Net Average	√ -	√ -	-	√	+	√ -	Median=2,698
	Overall Average	√	√	√	√ -	√ +	√	Median=845

Source: King, Michel and Elder, John. Evaluation of fourteen desktop data mining tools. Systems, Man, and Cybernetics. IEEE International Conference. Volume: 3. 1998.

++=Excellent; +=Good; √ = Average; - = Needs Improvement; -- = Poor; None = Does Not Exist; NE = Not Evaluated. Source provided by UVA Research System Engineering.

Industry Wide Applications of Data Mining:

Thus far, the main industries where data mining is expected to add value to customers are in five industry applications. These include retail, banking, insurance, health care, transportation, and medicine. In each of these, there are specific factors in which data mining can prove a valuable asset. For example, in the retail industry, customers may want to use data mining to find out useful consumer purchasing trends. In banking, detecting patterns of fraudulent credit usage and identifying "loyal" customers are other useful characteristics that data mining could improve. In insurance and health care, predicting which customers will buy new policies is important in understanding. Determining loading patterns is useful in the transportation industry. Finally, in medicine, characterizing patient behavior to predict office visits is a valuable variable factor to analyze [1].

As seen above, one can apply data mining technology to many areas from making predictions about consumer preferences to finding out the number of office visits. The retail sector has many distinguishing characteristics that differentiate from other industries. For one, there are many different types of retailers and these differences will be key to understanding how a technology like data mining can prove to be more valuable in some areas of the retailer industry than others. Are there areas in which data mining is more of an efficient technology within the retailer market? At best and worst case scenarios, what is the likelihood of data mining's success? What are the payoffs for companies in the different retailer sectors? These are the kinds of interesting questions that remain to be answered in the following sections.

Specific Uses of Data Mining Applications (Credit Fraud):

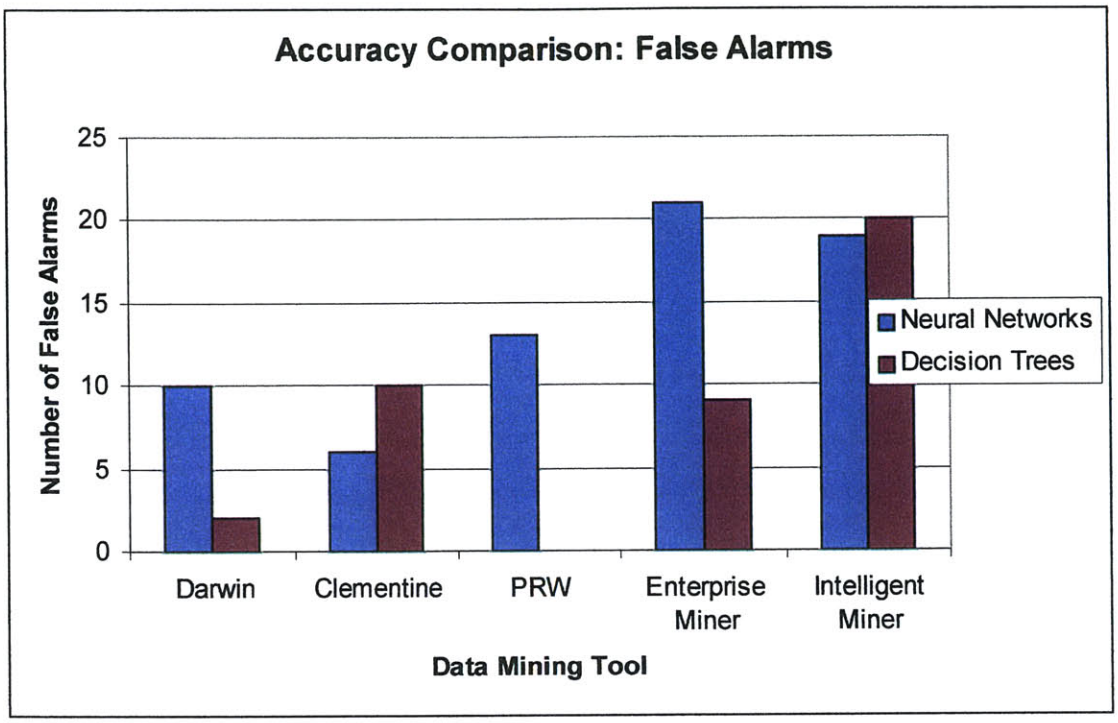
As more information is stored into databases and as the accumulation of data grows, the task of providing value to customers using data mining tools becomes more difficult. Though data mining has served as the tool to add value to organizations using data mining by finding new and important snippets (or nuggets) of knowledge, the problem of focusing on the most significant pieces of information becomes increasingly important [20]. It is not only necessary to find the most relevant information, but it is also necessary to be able to find the most important snippets from the collection of nuggets found. As such, data mining first extracts all relevant pieces of information, but then selects the more important pieces of information found therein [20]. Using this technique and methodology is commonly found in the use of detecting credit fraud where it is important to detect the perpetrators of a system.

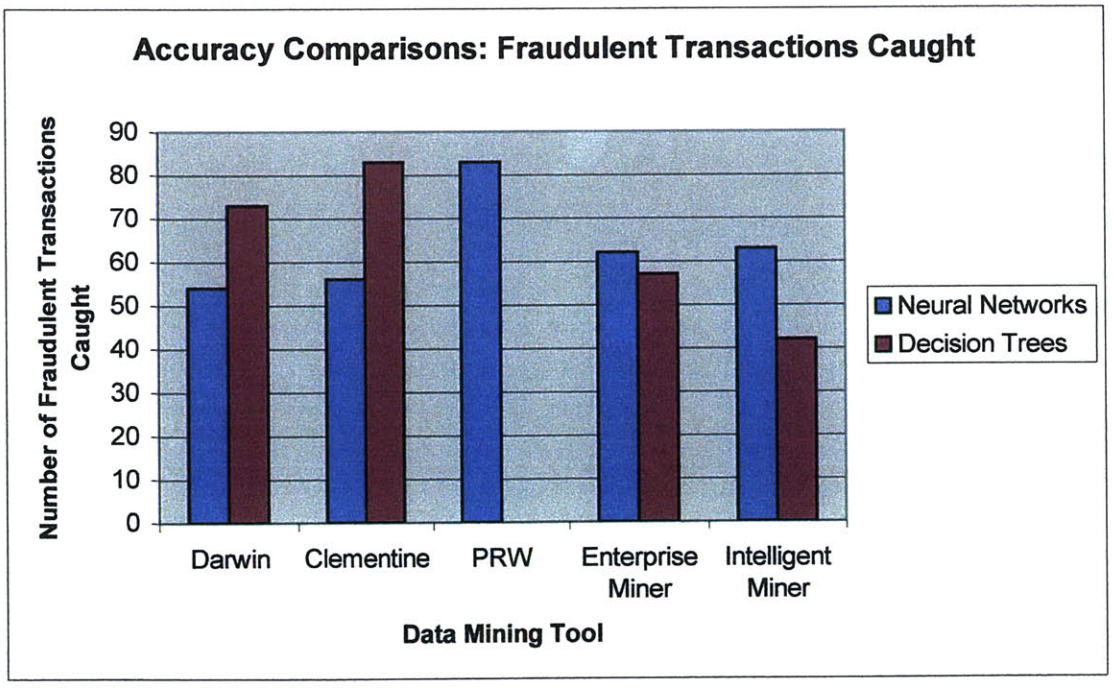
Fraud is becoming an increasing area of concern for corporations. In very large collections of data (such as that held by credit card companies, telephone companies, insurance companies, and the tax office), it is often expected that there is a small percentage of customers who are practicing fraudulent behavior. To illustrate this point further, consider that the Health Insurance Commission maintains a large database containing information relating to payments made to doctors and patients from the government's Medicare program. The database is measured in terms of terabytes. Using a data mining extraction tool against such a large database would result in significant findings, such as pointing out the handful of customers who are engaging in fraudulent behavior [28]. As such, data mining can be used to yield insights into who might be conducting this illegal behavior. An approach used for fraud detection is called hot spots

[20]. This methodology consists of finding the nuggets created by the data-mining tool. Of this set of nuggets, a further analysis will be conducted to narrow the set of nuggets into a handful of nuggets that truly capture the important data delivered by the mining tool. A hot spots evaluation function is used to mine information to find those nuggets that have significantly higher rates of claim than the overall average of nuggets [20]. This leads to finding the data points that can potentially solve the problem of fraud by finding those individuals whose claims are falsified.

In a recent study five of the most highly acclaimed data mining application were compared on a fraud detection application [100]. Forty tools were evaluated of which five emerged as the leading tools [100]. Curiously enough, four of five leading companies in 1997 produced these five tools. According to DataQuest in 1997, IBM was the data mining software leader with a 15% share of license revenue, Information Discovery was second with 10%, Unica was third with 9%, and SGI was fourth with 6% [101]. The comparison was between the Oracle's Darwin, SPSS's Clementine, Unica's PRW, SAS's EnterpriseMiner, and IBM's Intelligent Miner products. These products will be discussed in greater detail in a general-wide comparison of many other powerful commercial tools. The comparisons in Figure 10, shown in the next page, show that decision trees are generally better than neural networks at reducing false alarms. This was a result of non-fraudulent transactions given a higher priority in the data therefore reducing their number missed [100]. These commercial tools are adept in particular area; in general, they are not good for solving all kinds of problems. Thus, some of these packages are better suited for certain types of applications than others.

Figure 10: Accuracy Comparisons of Leading Tools





Data Mining in Medical Databases:

Medical institutions have collected vast amounts of data of all the patients that they have served. With each added entry, the database grows. Currently there exist few tools that can take advantage of a person's records by evaluating and analyzing a patient's data that is stored in a large database [77]. Data mining can be used in this situation to take advantage of this large collection of information not only to manage and store the data more efficiently, but in addition, to make predictions as to when particular customers will run out of medication and will require new prescriptions to be issued. Furthermore evaluation of stored clinical data may lead to the discovery of trends and patterns hidden within the data that could significantly enhance the understanding of disease progression and management. Specifically, this can done by using an association rule such as the following described below by Agrawal for large databases [78].

Let $I=\{i_1, i_2, \ldots, i_n\}$ be a set of items found in a database such as individual patient records in a medical database. Let DB be the database of transactions where each transaction denoted by T consist of a set of items such that $T \subseteq I$. Given an item set $X \subseteq I$, a transaction T contains X if and only if $X \subseteq T$. An association rule consists of the following type of expression $X \Rightarrow Y$, where $X \subseteq I$, $Y \subseteq I$, and $X \cap Y = \varnothing$. The association rule $X \Rightarrow Y$ holds in a database DB with confidence c if the probability of a transaction in the DB that contains X also contains Y is c. By setting a threshold and using association rules, the confidence level can be compared to the threshold indicating to the data mining application the probability of a certain sequence of items yielding in fruitful results. Using decision trees can help develop a method for finding association rules from a database and attaching certain confidence intervals to a set of items.

Comparison of Leading Data Mining Tools

Thus far there has been a discussion of leading data mining tools from classification tools to decision trees to clustering techniques to neural networks to name a few. In this section, an illustrative comparison of the leading software products in each of these specific fields is discussed. In Table 16 is a listing of leading products along with a criterion for evaluating these products. As shown, these products have certain strengths and also have certain weaknesses. The corresponding key strengths and weaknesses are shown in Table 17. Thus, as Tables 16 & 17 show, there is no clear winner of a better tool. However, certain tools are favored over others in specific situations. It is up to the client organization and the data mining company to figure out what is the optimal tool to use whether or not they provide it.

Table 16: Comparison of Leading Commercial Products

Product/ Company/ Version:	Researcher's Experience	Platform Independence	Algorithms	Learning Rate	Learning Rate Decay	Decision Tree Use	Regression/Stats	Data Loading/Manipulation	Model Building	Model Understanding	Technical Support	Overall Usability	Visualization
Clementine SPSS 4	√	√+	√+	√	√	√-	√	√+	√+	√+	√+	√+	√
Darwin Oracle 3.0.1	√	√	√	√		√		√	√	√+	√	√	√-
DataCruncher DataMind 2.1.1	√+	√	√					√+	√+	√	√	√	√
Enterprise Miner SAS Beta	√	√+	√	√		√	√+	√	√	√	√	√	√
GainSmarts Urban Science 4.0.3	√-	√	√+			√	√+	√+	√	√	√	√	√-
Intelligent Miner IBM 2	√-	√	√			√	√-	√	√	√	√	√	√
MineSet SGI 2.5	√-	√	√			√	√	√	√+	√+	√	√+	√
Model 1 Group 1 3.1	√	√	√+	√		√-	√+	√+	√+	√+	√+	√+	√

√-=some capability, √=good capability, √+=excellent capability. Source: Elder, John F. and Abbott, Dean W. A Comparison of Leading Data Mining Tools. Fourth Annual Conference on Knowledge Discovery & Data Mining. August 1998.

Table 16 Continued:

ModelQuest AbTech Corporation 1	√	√	√-			√-	√	√	√+	√+	√+	√+	√
PRW Unica Technologies 2.5	√+	√	√+	√		√	√+	√+	√+	√+	√+	√+	√
CART Salford Systems 3.5	√	√+	√			√+		√-	√	√	√	√	√
Scenario Cognos 2	√	√	√			√		√	√+	√+	√	√+	√
NeuroShell Ward Systems Group 3	√	√	√	√	√			√	√	√	√	√	√
OLPARS PAR Government Systems 8.1	√+	√	√	√				√	√-	√	√	√	√
See5 RuleQuest Search 1.07	√	√+	√			√+	√	√	√	√	√	√	√
Splus MathSoft 4	√+	√-	√+			√	√+	√	√	√+	√	√	√
WizWhy WizSoft 1.1	√	√	√					√	√	√+	√	√	√

√-=some capability, √=good capability, √+=excellent capability. Source: Elder, John F. and Abbott, Dean W. A Comparison of Leading Data Mining Tools. Fourth Annual Conference on Knowledge Discovery & Data Mining. August 1998.

Table 17: Strengths & Weaknesses of Leading Tools

Product	Key Strength	Key Weaknesses
Clementine	Visual Interface, algorithm breadth	Scalability
Darwin	Efficient client-server, intuitive interface options	Limited visualization
DataCruncher	Ease of use	Single algorithm
Enterprise Miner	Depth of algorithms, visual interface	Hard to use, new product added complexity issues
GainSmarts	Data transformations, algorithm option breadth	Limited visualization
Intelligent Miner	Algorithm breadth, graphical tree/cluster output	Few algorithm options
MineSet	Data visualization	Few algorithms, no model export
Model 1	Ease of use, automated model discovery	A vertical tool only
ModelQuest	Breadth of algorithms	Non-intuitive interface options for user
PRW	Extensive algorithms, automated model selection	Limited visualization
CART	Depth of tree options	Difficult file I/O, limited visualization
Scenario	Ease of use	Narrow analysis path
NeuroShell	Multiple neural network architectures	Unorthodox interface, only neural networks
OLPARS	Multiple statistical algorithms, class-based visualization	Dated-interface, difficult file I/O
See-5	Depth of tree options	Limited visualization, few data options
S-Plus	Depth of algorithms, visualization, programmable/extendable	Limited inductive methods, steep learning curve
WizWhy	Ease of use, ease of model understanding	Limited visualization

Source: Elder, John F. and Abbott, Dean W. A Comparison of Leading Data Mining Tools. Fourth Annual Conference on Knowledge Discovery & Data Mining. August 1998.

Current Data Mining Tools (Web mining):

In this day and age of the Internet, the data mining applications offered to companies are slowly shifting to web mining tools. Despite that most of the companies using data mining are companies with vast amount of collections which included Fortune 500 companies web mining can now be used by Internet companies accumulating substantial data and not limited in focus to well established organizations. Web-based organizations, which now include almost any organization in any market, often generate and collect large volumes of data in their daily operations. Analyzing such data can help these organizations to determine the life time value of clients, design cross marketing strategies across products and services, evaluate the effectiveness of promotional campaigns, and find the most effective logical structure for their Web space. This type of analysis involves the discovery of meaningful relationships from a large collection of primarily unstructured data, often stored in Web server access logs. Web mining applied with the applications of data mining and knowledge discovery techniques can now be applied to data collected in World Wide Web transactions.

The information revolution is generating exceeding amounts of data every year. This includes data from sources as diverse as credit card transactions, telephone calls, Web clickstreams, to research studies. As the pace of data generation speeds up, so too is the rate at which technology advances over time (dtechnology/dt >>0). Processor power is doubling every year while memory density doubles every 3 years [12]. Disk capacity doubles every three years while the seek rate doubles every 10 years [12]. While these dissimilar doubling rates occur for these technological components, so too does the drop in price. In January 1981, the $/Gbyte exceeded $100,000 [51]. This rate has slowly

been decreasing and was $40/Gbyte in January 2000 [51]. As such, faster and cheaper storage technology allows for greater amounts of data to be stored.

The Web is expanding the focus of data mining beyond the traditional analysis of structured data. There are now huge amounts of information in free text, images, sounds, and video that are hard to analyze. Current systems, such as the Altavista Photo and Media Finder and MP3 Search Engines can find an image or a sound based on the surrounding text, but they fail at more complex request such as finding similar images or sounds [93]. However, with the advent of XML, this will greatly facilitate the analysis of previous unstructured data. XML is a metadata language for describing markup languages. Metadata describes how and when and by whom a particular set of data was collected, and how the Data are formatted. MetaData are essential for understanding information stored in data warehouses. XML can provide more and better facilities for browser presentation and performance. Information will be more accessible and reusable, because the more flexible XML can be used by any XML software and is not restricted to specific manufacturers as is the case with HTML. As such, there will be software facilities created with the advancement of new technology to cope with data management.

Data Mining/Web Mining Trends:

According to Data Mining News, as of 1999, the data mining industry was valued at $1 billion [7]. Furthermore, according to Herb Edelstein, by 1998, the data mining tool market was at $45 million, a 100% increase from 1997 [8]. As such, the data mining industry is growing at a fast pace. The information revolution is generating vast amounts of data and the only way to keep up with providing steadfast growth is to understand the

data. Data mining can still offer potential solutions in many different areas of industries. With the age of the Internet, however, more data are being accumulated through on-line transactions than before. Increasing in numbers, on-line transactions are setting new records on a yearly basis. It has been estimated by Internet News that the number of adults using credit cards to purchase goods and services online more than doubled between 1998 and 1999 [10]. Furthermore, the report included that by the third quarter of 1999, 19.2 million adults used their credit cards to make online transactions, compared to 9.3 million in 1998 and only 4.9 million in 1997. As such, online transactions are increasing in record numbers.

In the Healthcare E-commerce industry, a recent study showed that by 2004, this industry would surpass the $370 Billion mark in online transactions. Compared with travel, finance, and even the steel industry, healthcare e-commerce is a late bloomer, says the report from Forrester Research [2]. The Web will become the foundation of a new healthcare industry where current large-scale operations between suppliers, distributors, and customers will have a new vehicle to carry their transactions in. Furthermore, Forrester Research indicated that already with 32% of online customers already shopping for health products on the Web, online sales show no signs of slowing down. In fact, Forrester expects that 8% of all retail health sales, $22 billion, will move to the Internet by 2004. Prescriptions drugs will dominate the market with $15 billion in online sales while pharmaceuticals will generate $3.3 billion in online sales as consumers take charge of their spending preferences. Over the counter drugs and health aids will account for $1.9 billion of the market and about $900 million, or 47% to the online retail industry added the report. As this report indicates, there is much awareness of the impact of the

Internet in this particular industry as online transactions are expected to continue its sharp increase.

A further interest in studying the Internet market in how data mining can be applied can be seen with the immense impact of the dot.com phenomena today. Web mining will be a growing technology as more companies are investing their futures in the Internet. Just in advertising alone, companies are spending most of their revenues trying to publicize their sites and try to attract as many potential customers as they can. With increasing online transactions and the potential of attaining a large set of customers through the Internet, it is no surprise that most Internet companies are devoting most of their sources of revenue to advertising costs.

In the recent years, Internet companies have accumulated over $1 billion in advertising [10]. Advertising by dot.com companies saw growth nearly triple in the third quarter of 1999 compared to the same time period from last year according to a study by Competitive Media Reporting. Including the holiday season, CMR predicted that dot.com's have more than doubled their spending on advertising than the $649.2 million spent offline for all of 1998. According to the CKS Group, they estimated online advertising spending to be $301 million in 1996. The holiday season, especially in the fourth quarter, has created sustained pressures on dot.com's to increase their spending to retain and attract a larger customer base.

In Table 18 shows the ever-growing presence of the Internet can be seen. Companies are flushing more money than ever before. As these numbers continue to reach new highs, there will be an increasing need for web mining. Understanding their respective customers for these dot.com's is a necessity. As well-established corporations

try to maintain their existing customer base away from these dot.com's, it will be important that whichever player remains in an industry especially where the Internet is concerned mine their data.

Online broker E*Trade lead the advertising Internet campaign, in spending $89 million in the first nine months (see Table 19). This table shows the increasing importance of the Internet to companies and the associated risk in spending these vast quantities of money to outperform existing competitors. Increases in spending have span across other mediums as well (see Table 20). Network television leads the dot.com spending in this sector with over $278 million in spending in the first nine months of 1999. As more data are found in the Internet, the leader in the web mining sector will be that company that can implement a viable solution to deal with not only vast amounts of data but also with different interfaces of the data.

Table 18: Dot Com Offline Advertising Spending

Category	Jan-Sept 1999	Jan-Sept 1998	% Change
Online & Internet Services	$288,860,000	$114,398,500	152.5%
E-Commerce	$1,077,201,400	$234,743,200	358.9%
Total Dot Com Spending	$1,366,061,400	$349,141,600	291.3%

Source: Competitive Media Reporting. Offline Spending by Internet Brands Passes $1 Billion. Internet Newsletter. 1999.

Table 19: Top Internet Brands Spending Offline

Rank	Brand	Jan-Sept 1999	Jan-Sept 1998	% Change
1	E*Trade	$88,985,000	$16,967,100	424.5%
2	Value America Stores	$46,538,200	$18,743,600	148.3%
3	Charles Schwab	$40,861,900	$442,100	9,142.7%
4	Snap.Com	$38,054,800	$3,179,700	1,096.8%
5	Ameritrade	$36,388,900	---	---
6	AT&T Business Network	$32,845,500	$21,382,900	18.6%
7	America Online	$30,051,100	$25,338,000	18.6%
8	Monster.Com	$20,613,700	$235,300	8,659.5%
9	Go.Com	$20,306,200	---	---
10	Priceline.Com	$20,116,600	$8,034,000	150.4%
11	Amazon.Com	$17,730,400	$14,028,500	26.4%
12	Discover Brokerage	$17,091,400	---	---
13	AT&T Worldnet	$16,073,900	$20,587,300	-21.9%
14	Microsoft Online	$16,067,200	$993,300	1,517.6%
15	Yahoo	$16,028,900	$4,274,200	275.0%

Source: Competitive Media Reporting. Offline Spending by Internet Brands Passes $1 Billion. Internet Newsletter. 1999.

Table 20: Dot Com Spending by Media

Rank	Measured Media	Jan-Sept 1999	Jan-Sept 1998	% Change
1	Network Television	$278,275,800	$60,184,500	362.4%
2	Magazines	$265,085,100	$91,401,600	190.0%
3	Cable TV	$202,627,000	$43,471,900	366.1%
4	Spot Television	$166,928,300	$44,120,000	278.4%
5	National Spot Radio	$154,621,400	$27,400,800	464.3%
6	National Newspapers	$148,659,900	$41,404,000	259.0%
7	Newspapers	$69,392,900	$17,522,500	296.0%
8	Network Radio	$43,137,500	$17,172,200	151.2%
9	Outdoor	$24,640,300	$3,972,400	520.3%
10	Sunday Magazine	$6,978,300	$581,000	1,101.1%
11	Syndication	$5,715,200	$1,910,700	199.1%

Source: Competitive Media Reporting. Offline Spending by Internet Brands Passes $1 Billion. Internet Newsletter. 1999.

Most Current Mining Techniques

Filtering:

Every year, more and more data are accumulated into databases. Currently, the amount of information that is being generated each year is exceeding the ability of properly processing the data. Though databases can be constructed to hold plentiful amounts of information, processing through these data are hampered by the volume of data. Algorithms are optimal to a certain level of processing. After a point, several layers of processing need to be done to the data. This requires the use of filtering, where the most types of information are extracted our to the user. In the case of web mining, web pages are filtered to a user to maximize the time a user spends in navigating through the site. The more filtering is done, the more beneficial it will be to the user in locating the information that is of interest to him or her. As such, three general techniques have been developed over the years to address the problem of information overflow. These techniques are information retrieval, information filtering, and collaborative filtering. Each of these technologies focuses on particular sets of tasks or problems. Information retrieval revolves around in fulfilling tasks such as fulfilling user interest queries. This essentially involves a query to a database for the extraction of information. Information filtering involves classifying streams of new content into categories such as finding any newly released soundtracks by Madonna or finding any newly released movies with the actor Sylvester Stallone. Lastly, collaborative filtering deals with focusing which items in a set should a user view based on the recommendations given by other users within the same community or group.

Information Retrieval:

Information Retrieval consists of issuing a query against a database requesting specific information to the user. This may involve indexing a list of documents using a query to capture the sub list of matching documents pertaining to the query. In general, information retrieval techniques are less valuable in the actual recommendation process since they capture no new information about user preferences and add no existing value to the existing information [23]. As such, for knowledge discovery processes, information retrieval is not considered part of data mining techniques as this technique merely involves issuing a query and does not involve digging through the layers of information trying to find hidden patterns in the data. One could indeed use a data-mining tool to issue multiple queries against a database to find interesting results, but information retrieval does not yield in any new insight.

Information Filtering:

Another use of a filtering protocol is that of information filtering. This type of filtering concerns itself with item content and the development of a personal user interest profile [23]. This differs from the focus of collaborative filtering where the goal is to identify users of similar tastes and the use of their opinions to predict the value of items for each user in the community. Computer users are connected via worldwide networks to an increasing number of data sources and other users. This interconnectivity provides users with previously unknown riches of available knowledge [24]. With today's increasing production of information coming from organizations, individuals, and society at large, even finding personal user interest profiles can be problematic. Imagine tailoring an Amazon.com site specific to every user. The amounts of information that

would need to be maintained for each customer from the overwhelming amounts of data stored in a database would not be an optimal solution. An alternative approach can be using both collaborative and information filtering to address the needs of user and groups collectively. In that manner, content needs to be filtered with respects to specific user groups and not to individuals themselves. As computers, communication, and the Internet make it easier for anyone and everyone to speak to a large audience, well-developed filtering techniques will need to be developed to meet reasonable performance standards [23].

Collaborative Filtering:

Collaborative filtering or recommender systems help users make choices based on the opinions of other users [25]. Systems that use collaborative filters help people find articles they will like in the huge stream of available articles. Collaborative filtering is a technique mostly used in the context of web mining. Though web mining itself is a loosely defined term, the definition given to the kinds of applications web mining is used for is mostly in the context that has been discussed so far. Wherever an application can make use of web log files, ad files, previous customer purchases and other significant purchasing information to discover knowledge in the data, then the application is considered a web mining application. Another application such as querying a web server and obtaining information is a primitive protocol that does not reveal any hidden information or extracts out certain patterns in the data. Instead, it provides a fixed answer to a problem and does not try to do anything further. These applications can be called web mining, in the context of this paper, these types of applications are not classified under the discussion of knowledge discovery web applications.

One of the best ways to find useful information is to find someone who has similar interests as another use and ask them for recommendations. Collaborative filtering is a way of mechanizing this form of information search [26]. Here are a few examples of collaborative filtering systems:

- Firefly (once known as Ringo) offers a "personal music recommendation agent". Thousands of people tell Firefly what music CDs they like; Firefly finds people with similar tastes and recommends music that other people in their similarity group like.
- Webhunter (once known as WebHound) does something similar for Web pages, finding pages for a current user from past users.
- GroupLens helps to filter Usenet News. After reading an item, each reader rates it according to how interesting it was. Subsequent readers see an "interest score" that is computed as a weighted average of previous readers. But the score is personalized: the weight a person's evaluation receives in this average depends on how often you have agreed with that person in the past [23].

One of the chief problems facing providers of information services is how to filter information; collaborative filtering offers a solution to this problem and appears to be fertile area for research and development. Collaborative filtering is a unique approach to information filtering that does not rely on the content or shape of objects, as it is the case with content-based filtering. For example, content-based filtering would allow for recommendation based on a movie genre (fiction, horror, comedy, romance, etc) and cast/credits (Mel Gibson, Arnold Schwarzenegger, Bruce Lee, etc) [23]. Collaborative filtering on the other hand would tend to recommend a movie selection based on what

group-minded individuals would tend to view. Collaborative filtering relies on meta-data (data containing information about other data) information pertaining to information about objects, such as CD's, movies, book, or web pages [21]. Data can either be collected automatically, by inferring from the user's interaction with the filtering system, or voluntarily collected where users supply the information.

A specific type of collaborative filtering application is called active collaborative filtering [22]. This approach is based on encouraging people to share information with one another rather than collecting ratings and modeling user interests in order to compile recommendations as in traditional collaborative filtering techniques. Active collaborative filtering builds on the following premises:

- Every person says what items they like and dislike
- New items are recommended to a user based on the opinions of people with similar taste
- Filtering can be applicable to music, movies, websites, news, TV programs, etc.

On the other hand, passive collaborative filtering in the Usenet domain is based on providing users with data about the news readings of other Usenet users [21]. This approach is based on the observation that experienced users use not only the subject of discussions with a group but moreover use the occurrence of contributions by other users as indicators for potentially interesting discussions. Essentially then, collaborative filtering uses a database about user preferences to predict additional topics or products a new user might like.

Hybrid Recommender Systems:

Several systems have tried to combine the information and collaborative filtering techniques in an effort to overcome the limitations in each technique [23]. As noted above, GroupLens is a hybrid system that combines provided by users, data inferred from user behavior, e.g., the time spent reading articles as indicator for interest, and content-based data extracted from the objects under investigation, such as the proportion of spelling errors and included text in documents [21].

The growth of Internet commerce has stimulated the use of collaborative filtering algorithms [27]. Such systems leverage knowledge about the known preferences of multiple users to recommend items of interest to other users. Microsoft Research Group has evaluated a new method called personality diagnosis [27]. Given a user's preferences for some items, they compute the probability that he or she is of the same "personality type" as other users, and, in turn, the probability that he or she will like new items. This new way of applying traditionally similarity weighting collaborative filtering approaches can be used in that all data are brought to bear on each prediction and new data can be added easily and incrementally.

Another research application of collaborative filtering is currently being at developed at the MIT Collaborative Ontology Department. Ontologies are a means of categorizing objects, such as features of a product, a person's interests, or Web pages. This department has researched that in general ontologies developed by a single organization are necessarily sparse or coarse and can be slow to add important new categorizations. As such, this group is investigating innovative ways in which ontology can be developed by many distributed individuals and organizations. Their first

prototype, tentatively called Mishmash, will evaluate one approach to how an easily extensible ontology can evolve collaboratively without a priori structure or centralized direction.

ERP & Data Mining Systems:

ERP & ASP Introduction:

Enterprise Resource Planning Systems (ERP) comprises a commercial software package that promises the seamless integration of all the information flowing through the company - financial, accounting, human resources, supply chain, and customer information [94]. ERP systems are a collection of software programs that tie together all of an enterprise's various functions--HR, finance, manufacturing, sales, etc. This software also provides for the analysis of this data to plan production, forecast sales, and analyze quality. Application Service Provider companies seek to offer competing services to smaller organizations helping these companies leverage their data management. These two types of companies, ERP & ASP companies, could gain even greater footing in the marketplace with the addition of providing data mining services. Providing a software tool that can integrate a companies' existing data across many departments with the addition of providing a data-mining tool that is maximized to work most efficiently with that software package could yield significant benefit to client organizations.

Examples of ERP packages are HRMS, Financials, Manufacturing, Distribution, and Sales. Each ERP Package may offer different functionality for different industries. Current targeted industries for ERP installations are Communications, Federal Government, Financial Services, Healthcare, Higher Education, Manufacturing, Public Sector, Retail, Service Industries, Transportation, and Utilities [35]. In these industries, only large companies have been targeted due to the length and cost of an implementation.

Recently, ERP software manufactures are offering reduced software versions, with fewer features, to medium-sized companies [35]. ERP companies are now facing the pressures of providing smaller scaled services to these smaller companies. Application Service Provider companies (ASP) are competing for this market space in offering systems that can integrate and consolidate companies' information systems.

Few mid-sized businesses have the capital to invest in, manage, and upgrade advanced e-commerce and e-marketing technologies in-house. In order to remain competitive, many companies are now turning to application service providers (detailed explanation and analysis in next section) organizations that serve the function of internal IT departments by developing and managing the technologies necessary to deliver key services over the Internet. The integration of the Internet is creating new dimensions of opportunities that will quickly and drastically affect the way that society functions [33]. One of the most innovative developments is occurring in the ASP market as these service organizations are allowing smaller companies to seek the advantages of ERP systems without actually having to implement anything in-house. ASP companies are now able to offer data mining services to smaller players in any industry. This is the advantage realized by corporations using an ASP model as they can provide leading services to companies not having enough capital to budget expensive ERP systems. Because ERP and ASP vendors offer significant advantages to corporations incorporating such solutions, integrating a data-mining tool alongside these services can provide an optimum solution in understanding a company's data.

Enterprise Resource Planning [ERP] Systems:

ERP systems try to build a single software program that serves the needs of people in finance as well as it does the people in human resources up to the managers making decisions. Each of those departments typically has its own computer system, each optimized for the particular ways that the department does its work. But ERP combines them all together into a single, integrated software program that runs off a single database so that the various departments can more easily share information and communicate with each other [32].

Enterprise Resource Planning software systems provide comprehensive management of financial, manufacturing, sales, distribution and human resources across the enterprise. The ability of ERP systems to support data drill down and to eliminate the need to reconcile across functions is designed to enable organizations to compete on the performances of the entire supply chain [37]. To utilize these capabilities managers have to learn how to manage processes in the ERP environment.

That integrated approach can have a tremendous payback if companies install the software correctly. For example, in a common scenario experienced by frustrated customers consists of an order bouncing around different departments before an order has been finalized. When a customer places an order, that order begins a mostly paper-based journey from in-basket to in-basket around the company, often being keyed and re-keyed into different departments' computer systems along the way. Meanwhile, no one in the company may not truly know what the status of the order is at any given point because there is no way for the finance department, for example, to get into the warehouse's computer system to see whether the item has been shipped. This is usually attributed to

company may not truly know what the status of the order is at any given point because there is no way for the finance department, for example, to get into the warehouse's computer system to see whether the item has been shipped. This is usually attributed to firewalls existing in the system. If a customer is interested in finding out the status of an order, he or she may have to go through several departments asking each one what the status of the order. Without an ERP system in place, this is a very inefficient system.

There are three major reasons why companies undertake ERP. These include integrating financial data, standardizing manufacturing and HR information [37]. Integrating financial data becomes a major reason to undertake the implementation for a system such as ERP as it simplifies the decision making for a top executive in a corporation. For example, as the CEO tries to understand the company's overall performance, he or she may find many different versions of the truth. The finance department has its own set of revenue numbers, the department of sales has another version, and other business units may each have their own versions of how much they contributed to revenues. ERP creates a single version of the truth that cannot be questioned because everyone is using the same system.

The second major reason to undertake ERP involved standardizing manufacturing processes. Manufacturing companies often find that multiple business units across the company produce the same product using different methods and computer systems. Standardizing those processes and using a single, integrated computer system can save time, increase productivity and reduce headcount.

Finally standardizing HR information is a final reason to implement an ERP system. In companies with multiple business units, HR may not have a unified, simple

method for tracking employee time and communicating with them about benefits and services. ERP can serve to provide a more integrated approach to facilitate easier communication.

ERP systems have emerged because the past decade the business environment has changed dramatically. Today, organizations are confronting new markets, new competition and increasing customer expectations. This has put a tremendous demand on manufacturers to deal with current day problems such as:

- Lower total costs in the complete supply chain
- Shorten throughput times
- Reduce stock to a minimum
- Enlarge product assortment
- Improve Product quality
- Provide more reliable delivery dates and higher service to the customer
- Efficiently coordinate global demand, supply and production.

Trying to grasp all these challenging goals becomes increasingly difficult if the necessary facilities are not available to attain those goals. Today's organizations have to constantly re-engineer their business practices and procedures to be more responsive to customers and competition [39].

Table 21 shown below illustrates the important historical dates in the evolutions of ERP Systems. As shown, in the 1960's the focus of manufacturing systems was on inventory control. Most of the software packages then (usually customized) were designed to handle inventory based on traditional inventory concepts [39]. In the 1970's the focus shifted to MRP (Material Requirement Planning) systems. MRP systems are a

time phased priority-planning technique that calculates material requirements and schedules supply to meet changing demand across all products and parts in one or more plants. Essentially, MRP consists of a computer-based system for managing inventory and production schedules. In the 1980's the concept of MRP-II (Manufacturing Resources Planning) evolved which was an extension of MRP to management activities [39]. MRP-II is a computer-based system in which the entire production environment is evaluated to allow schedules to be adjusted and created based on feedback from current production and purchase conditions. In the early 1990's, MRP-II was further extended to cover areas like Engineering, Finance, Human Resources, and Projects Management to complete a gamut of activities within any business enterprise. Hence, the term ERP (Enterprise Resource Planning) was coined [39].

Table 21: Timeline of ERP Technologies

1960s	Enterprise Resource Planning (ERP) is born in the early 1960s from a joint effort

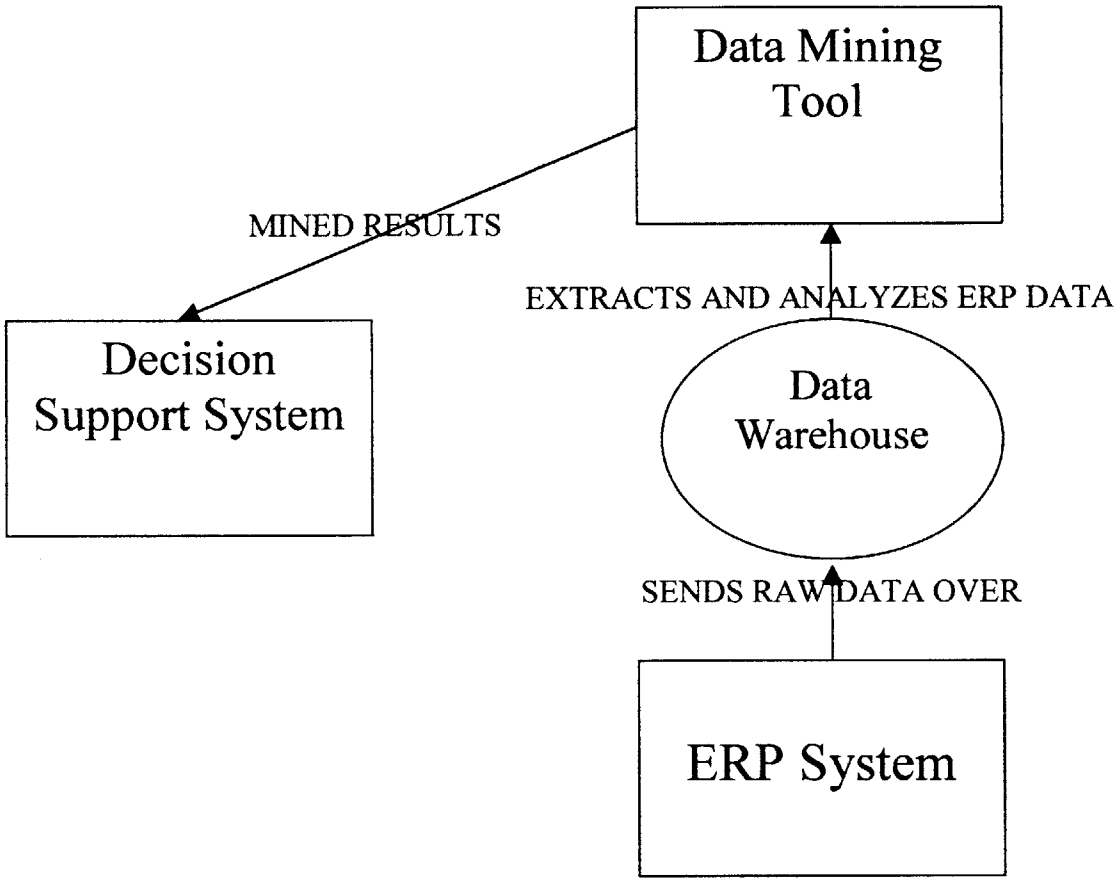
	between J.I. Case, the manufacturer of tractors and other construction machinery, and partner IBM. Material Requirements Planning or MRP is the initial effort. This application software serves as the method for planning and scheduling materials for complex manufactured products.
1970s	Initial MRP solutions are big, clumsy and expensive. They require a large technical staff to support the mainframe computers on which they run.
1972	Five engineers in Mannheim, Germany begin the company, SAP. The purpose in creating SAP is to produce and market standard software for integrated business solutions.
1975	Richard Lawson, Bill Lawson, and business partner, John Cerullo begin Lawson Software. The founders see the need for pre-packaged enterprise technology solutions as an alternative to customized business software applications.
1976	In the manufacturing industry, MRP (Material Requirements Planning) becomes the fundamental concept used in production management and control.
2000 and Beyond	Most ERP systems are enhancing their products to become "Internet Enabled" so that customers worldwide can have direct access to the supplier's ERP system.

ERP System Main Advantages:

ERP systems leverage existing data information and manage data more efficiently allowing an organization to maximize the use of personnel, customer, and organizational time [45]. ERP automates the tasks involved in performing a business process-such as order fulfillment, which involves taking an order from a customer, shipping it and billing for it. With an ERP system, when a customer service representative takes an order from a customer, he or she has all the information necessary to complete the order [35]. For example, this includes having the customer's credit rating and order history, the company's inventory levels, and the shipping dock's trucking schedule. Everyone else in the company sees the same computer screen and has access to the single database that holds the customer's new order. Thus to find out where the order is at any point, one need only log into the ERP system and track it down.

Probably the most important advantage in using an ERP system is the integration of the data. The Internet and data warehousing are two key technologies transforming ERP systems [96]. Shown in Figure 11 is ERP system that uses a data warehouse. The advantage in doing so is that a data warehouse can best solve the integration of distributed information systems [96]. Many ERP vendors are starting to incorporate data warehouse applications to complement their ERP systems to solve this particular problem only [96]. Furthermore, because of the advantages inherent in maintaining a separate highly integrated storage system (see Data Warehouse discussion), ERP systems incorporating a warehouse can gain by having a packaged application solution further consolidating the data. A further realizable advantage in employing such a scheme is that now a data-mining tool can be incorporated. If the trend is to continue using data warehouses for further efficiency of storing and managing a corporation's data, this additional step in providing a data-mining tool that detects hidden trends would be a consolidated effort in trying to provide an organization optimal solutions.

Figure 11: Data Warehouse integrated with an ERP System



Time horizons for ERP projects:

Companies that install ERP do not have an easy time in doing so. The reasons are mostly due to the overwhelming complexity in installing a system and ensuring that the system is meeting performance standards [36]. Typically, a project will take a span of months to complete. A project lasting under six months is considered a short time frame as these type of implementations have a catch of one kind or another: the company was either small, the implementation was limited to a small area of the company, or the company only used the financial pieces of the ERP system (in which case the ERP system is nothing more than a very expensive accounting system) [37]. To do ERP right,

the ways in which the business organization carries itself, along with the tasks given to employees, may all change [38]. Only if the policies undertaken by an organization are effective and optimal, (orders all shipped on time, productivity higher than all your competitors, customers completely satisfied), there is no reason to even consider ERP [36]. Requiring substantial changes in the way an organization carries out its practices and its data management has significant implications. As such, an organization not only needs to adapt to the changes and accept them, but it must also use these changes to help it optimize its strategic choices in the marketplace. The important point for organization implementing ERP is not to focus on how long it will take—real transformational ERP efforts usually run between one to three years, on average—but rather to understand why an organization needs it and how it can be used to improve an organization's business [37].

Reasons for Undertaking ERP Projects

It is extremely critical for companies to figure out if their ways of doing business will fit within a standard ERP package optimally before the checks are signed and the implementation begins. ERP Systems can be very expensive systems not only monetarily but also to the very people carrying out the actual implementation processes [40]. The most common reason that companies walk away from multimillion-dollar ERP projects is that they discover that the software does not support one of their important business processes. At that point there are two things companies can do: change the business process to accommodate the software, which will mean deep changes in long-established ways of doing business (that often provide competitive advantage) and shake up

important peoples' roles and responsibilities. Or they can modify the software to fit the process, which will slow down the project, potentially introduce bugs into the system and make upgrading the software to the ERP vendor's next release more difficult [41]. Thus the move to ERP can be a project of breathtaking venture. In addition to budgeting for software costs, financial executives should plan for a long laundry list of other expenses before the benefits of ERP start to manifest themselves [40]. Furthermore, failure to consider data warehouse integration requirements and the cost of extra software to duplicate the old report formats can be another major hurdle for corporations if ERP is not implemented in the correct manner [42]. Thus, a few oversights in the budgeting and planning stage can send ERP costs spiraling out of control faster than oversights in planning almost any other information system undertaking.

Even more risk in the implementation of ERP systems is occurring in this day and age as companies are purchasing off the shelf ERP solutions for IT planning [43]. Whether or not this is a good idea is uncertain at this point. Certainly, the larger and the better-established companies undertake in using off-the-shelf products, the more likely they are to suffer from lack of proper optimal implementation. Where software can be uniquely designed for a corporation potentially resulting in an optimal ERP system, the tradeoff in using an off-the-shelf product is that it is not maximized to suit an individual corporation's data management needs. Thus, an effective IT infrastructure can support a business vision and strategy whereas a poor decentralized one can break a company [43].

Costs of ERP Systems

Meta Group recently did a study looking at the Total Cost of Ownership (TCO) of ERP, including hardware, software, professional services, and internal staff costs [96]. The TCO numbers include installing the software and two years after implementation, which is when the real costs of maintaining, upgrading and optimizing the system for an organizations business are felt. Among the 63 companies surveyed—including small, medium and large companies in a range of industries—the average TCO was $15 million (the highest was $300 million and lowest was $400,000) [97]. While it is hard to estimate exactly what the expense is for an ERP project, the Meta Group came up with one statistic that proves that ERP can be expensive no matter what kind of company is using it. The TCO for an average user over that period exceeded $53,320 [97]. In the same study, out of the 63 companies, it took on average eight months after the new system was in (31 months total accounting for implementation and due-diligence time) to see any benefits. But the median annual savings from the new ERP system was $1.6 million per year. Although different companies will find different land mines in the budgeting process, those who have implemented ERP packages agree that certain costs are more commonly overlooked or underestimated than others [44].

E-Commerce and ERP Systems:

Today, the hottest areas for outward-looking Internet post-ERP work are electronic commerce, planning and managing an organization's supply chain, and tracking and serving customers [46]. Most ERP vendors have been slow to develop offerings for these areas, and they face stiff competition from niche vendors. ERP

vendors have the advantage of an existing installed base of customers and a virtual stranglehold on common office function such as order fulfillment. Recently ERP vendors have begun to shrink their ambitions and focus on being the back-office engine that powers electronic commerce, rather than trying to own all the software niches that are necessary for a good electronic commerce website [46]. As the niche vendors make their software easier to hook into electronic commerce web sites, and as middleware vendors make it easier for IS departments to hook together applications from different vendors, there is wonder to how much longer ERP vendors can claim to be the primary software platform for the Fortune 500.

Developing ERP-based e-Commerce applications must include a good balance of the following three major components: a well-defined and integrated Internet strategy approach, a proven technology base, and a well-engineered design [52]. A defined clear strategy goes a long way in understanding what the proper steps and the proper vision in fulfilling that strategy. A well-thought out strategy will have taken into account multiple pathways and chosen a particular path based on the strategic tradeoffs that are most suitable to that organization. Having a solidified technology base is important in that it allows an organization to gain proper footing in today's sector where technology changes in a very rapid manner. As technology changes, the easier it is for a company to adapt to those changes if it has been involved with the particular technology in the past. Lastly, the final component to a successful ERP e-commerce solution is having an implementation design that works. It is not enough to engineer a design that meets some goals while not fulfilling others. Thus, meeting the needs of a design requires a successful implementation of all the requirements of the system. Based on these general

characteristics, a company can lead in the proper steps in achieving a well-thought out and well-archived strategic business model.

Future of ERP Systems:

The future of ERP continues to lie in automating the tasks of the major functional areas of a company finance, HR, sales and distribution and stores all the data from those different areas in a single database, accessible by all. This is the main advantage realized by ERP systems. A centralized database containing all key data components of an organization provides reliability, time-efficiency, and better storage of information for large corporations. The problem lies in that there can be many downsides to implementing an ERP. Consolidating different department's database into one can be a major problem to an organization as it adds an extra layer of complexity that is not the case with having separate data storage warehouses for each large department. As such, consolidating all databases into one, providing a clean user interface, and providing the necessary access control list while maintaining confidentiality, authentication, and reliability is a major challenge to ERP vendors to say the least. The advantages that ERP systems provide are worth the risk and time for corporations that do not have a centralized database. Finally, the Internet represents the next major technology enabler that allows rapid supply chain management between multiple operations and trading partners. Most ERP systems are enhancing their products to become Internet enabled so that customers worldwide can have direct to the supplier's ERP system. Recognizing the need to go beyond the MRP-II and traditional ERP systems, vendors are busy adding to their product portfolio.

ASP Solutions:

For smaller, second-tier companies, ERP systems are expensive, subject to delay and missed deadlines, and require significant unavailable management oversight. The applications hosting market is a new approach that promises mid-tier companies the benefit of Enterprise-wide Resource Planning (ERP) systems without the implementation overheads. The size of the market (frequently termed Application Service Provider or ASP) has been widely estimated. In a study conducted by International Data Corporation (IDC) projected a strong future for ASP. IDC is the world's leading provider of information technology data, industry analysis and strategic and tactical guidance to builders, providers and users of information technology. Worldwide spending on ASP will approach $8 billion by 2004, up from a meager $296 million last year, which amounts to a 92% compound annual growth rate are recent estimated by IDC analysts [47]. ASP companies are typically distinguished as providing the use of packaged software applications, implementation services, required computing hardware, secure network connectivity, on-going system operation, and scaling and upgrading on a "fee for usage" basis. This allows businesses to process higher volumes of increasingly complex applications, achieve economies of scale, and avoid the inherent risks of purchasing and implementing an expensive in-house system. Low customer awareness and the relative immaturity of ASP offerings inhibited spending in 1999 [48]. However, customers are realizing they can gain access to applications without initial investments in application licenses, servers, people and other resources. Also boosting market growth is the emergence of collaborative and personal segments within the overall ASP market [48].

As such, ASPs can offer smaller companies a value realization in what larger companies gain through the implementation of ERP systems.

Application Outsourcing

Application Outsourcing (AO) providers manage and maintain software applications. The provider assumes the responsibilities associated with the application. Application Service Provider (ASP) and Application Maintenance Outsourcing are sub-sectors of the AO market [46]. One important distinction between the two relationships is the actual ownership of the application. The ASP is the newest concept emerging from the foundation established in the outsourcing market. The ASP remotely hosts and delivers a packaged application to the client from an off-site, centralized location. The client does not claim ownership of the application but instead "rents" the application, typically on a per user basis. Application Maintenance Outsourcing providers manage a proprietary or packaged application from either the client or provider's site.

Characteristics of Application Outsourcing

Information Utilities and Business Process Outsourcing (BPO) providers focus on economic and efficient outsourcing solutions for complex but repetitive daily business processes. These could be as sophisticated as finance and accounting business functions or more repetitive processes, such as disbursements and payroll. The provider assumes all responsibilities associated with the entire business process or function. Platform IT Outsourcing providers offer a range of data center services, including hardware facilities management, onsite and offsite sup-port services, server-vaults and data security and disaster recovery capabilities. These relationships typically involve the transfer of IT

facilities, staff or hardware. However, it should be noted that many companies participate in multiple settings and are not necessarily defined to reside exclusively in a particular assigned category.

Estimates of Growth/Trends of ASP

At $296 million in 1999, worldwide spending on application service provider (ASP) services was minute in proportion to the amount of media attention ASP received [47]. IDC predicts that ASP spending will grow to $7.8 billion by 2004 [47]. This surge in spending translates to a 92% compound annual growth rate from 1999 to 2004. Estimates from the Gartner Group that the ASP market alone is approximately $2.7 billion and it is expected to grow to $22.7 billion by 2003 [65]. Furthermore, among factors fueling ASP growth has been the strong endorsement of technology mainstays AT&T, IBM, Microsoft, Sun and others who have developed applications for ASP and, in some case, have launched ASP initiatives of their own, IDC says. Their commitment is resulting in the applications, infrastructure, and services needed to drive growth in the market. In addition, the industry is already seeing a larger number of acquisitions and mergers in this industry sector [49]. The increase in customer acquisitions is being brought about by customers' realization that they can gain access to new applications without initial investments in application licenses, servers, people, and other resources.

According to IDC, another factor contributing to the ASP market's growth is the emergence of the collaborative and personal segments. Three critical trends are currently driving ASP companies: rapid technological changes, the worldwide shortage of IT personnel, and the desire for companies to offload what is not core to their business [65]. In 1999, spending on enterprise applications that include industry-specific and analytical

applications dominated the mix of ASP services accounting for approximately 66% of ASP spending [47]. Collaborative along with personal applications combined to account for one-third of ASP spending in 1999. Collaborative applications include groupware, document management, and email, while personal applications consist of office suites and consumer-oriented applications. By 2004, their combined share will jump to 50% [47].

Realizable Advantages of ASP

The telephone is connected to the wall and provides access to the entire world. No high initial investment is involved, but only a monthly charge, per user. The Internet can be responsible for a new business model – a model that allows companies to use applications by way of the public Internet by taking a subscription with what is known as an 'Application Service Provider'. Essentially, the goal is to use the Internet and its capabilities to allow companies to use applications that run on their particular Intranets to gain advantage of the use of the services these applications provide. The ultimate result is then the elimination of the cost of developing and managing IT aids along with the added benefits of high-speed implementation with reasonable performance and security [51]. In addition, the Internet can bring worldwide access and high reliability of the operational environment [50]. The only thing required is an Internet entry and a web browser an organization's staff. An Application Service Provider manages applications on central servers. They make use of the architecture of the Internet and development environments based on web technology [53]. This means that instead of installing numerous applications on large numbers of PCs, an organization has its staff use applications and data stored on the public Internet. With the increasing availability of

higher bandwidths at lower and lower cost, Application Service Providing has become a serious alternative for companies that are having more and more difficulty handling the increasing investments in their own information technology [56]. In addition, the protection of this central system against emergencies and unauthorized access to data are extremely effective.

With ASP companies no longer purchase a system at all: they take a subscription on an application. This means that the subscription includes a service level agreement (that lays down agreements on availability, performance and guaranteed security) and a help desk and primary and secondary support. Apart from this, implementation services are provided for the migration of data from old systems and the provision of a connection to the Internet. That is the advantage realized in ASP over implementing larger systems such as ERP systems.

It is important to note that ASP companies are not like traditional outsourcing companies. The ASP does not take over an organization's business process; it instead seeks to give an organization customized front-end applications that it runs on its hardware [98]. ASP differs from traditional network hosting in that it offers more than management of a network of servers; it also offers the applications that run on the network and the support for implementing those applications [98].

The Added Value of ASP

If organizations are to deal with today's ever-increasing competition, they need to operate cleverly and effectively. By making use of an Application Service Provider, a company guarantees itself access to information that is critical to the business (even in the case of rapid growth). Use can moreover be made of state-of-the-art technology and

IT support, which put the firm in a position to compete quickly and effectively in its own market and to move into new segments [61]. The Application Service Provider ensures that the applications provided stay up-to-date and continue to offer the latest possibilities. Thus, in addition to competitive power, this system provides advantages of cost and efficiency. By making use of an Application Service Provider, organizations can work with shorter implementation cycles, which are also less complex than that associated with their own system implementation [63]. The costs entailed in the use of the applications become more predictable (subscriptions) and the total costs of owning applications and information services for the entire organization can be lowered [63]. An Application Service Provider furthermore offers greater flexibility, and no risk-bearing implementation of in-house solutions is necessary. For example, some applications hosted by leading ASP company Siennax offers the following to its existing customer base:

- Intranet and Extranet
- Account management
- Microsoft Exchange
- Microsoft Office 2000
- Lotus Learningspace

Thus, ASP companies try to offer the latest product offerings to its customers. These companies can take advantage of not having the necessary resources or the desire to implement an ERP system, but can use such a system, customize it, and apply data mining tools to this system. ASP companies then offer a cost effective way for smaller

companies to store and manage their data while still being able to apply data mining tools to the data collected.

Partner Relationship Management & Customer Relationship Management

In comparison to the potential ASP software can have in the future, partner relationship management (PRM) software and customer relationship management (CRM) software are already reaching new highs [58]. Partner Relationship Management is an e-business solution for sales channel management enhancing partner relations, cutting costs as a whole and increasing revenues. PRM solutions are aimed specifically at managing vendor-partner relationships and bring new functionality to this neglected area. Customer relationship management (CRM) solutions capture customer data from systems such as enterprise resource planning system, analyze the consolidated data, and then distribute the results to all the contact points around the enterprise, as well as to suppliers and customers [58]. Customer relationship management provides in-depth research and analysis of the services required to provide sales, marketing, and customer care solutions [59]. This can produce cost-saving and revenue-enhancing results [58]. Rapid competitive change in financial services is underpinned by new technology, allowing radically improved customer relationship management at much lower channel cost [66].

PRM-CRM Growth & Trends

The ASP/PRM market is continuously reaching new highs. The PRM market is expected to reach $1 billion within the next five years. The PRM market is being driven by the growing dependence on channel partners to leverage sales and the fact that

knowledge is now a major corporate asset. The result is a convergence of ASP and PRM. Combining these in a single offering will lead to a dominant market position for the company that is able to do so [65].

Worldwide revenue for PRM packaged software reached $36.9 million in 1999, more than doubling the 1998 figure of $15.6 million. IDC predicts this new software category, a segment of the multibillion-customer relationship management (CRM) software market, will double in size annually over our forecast period, reaching $497.3 million in 2003. The PRM market is in its infancy and currently served by a range of companies most included in the four sectors:

- New, dedicated PRM players
- Niche companies whose products can be used for elements of vendor-partner relationships
- Broad-based Customer Relationship Management suppliers
- Traditional Enterprise Relationship Management Companies (ERM)

Although dedicated PRM suppliers have received much attention of late for focusing on the vendor-partner relationships, it is too premature in the industry to state whether or not they will dominate the still-developing market. Most of the dedicated players have only a small number of customers and are still in the venture capital stage [65]. They are highly reliant on strategic alliances, often with larger CRM or ERM companies that may decide to target this market. Many of these more-established suppliers are bringing functionality to the market and have the resources to overwhelm some of the niche suppliers should they so choose. On the other hand, the market is growing rapidly enough that there is room for a variety of solutions. Some customers

will seek point solutions instead of highly integrated ones due to corporate needs and other requirements.

The predominant leaders currently found in each of these four sectors are in Table 12 listing all the major players in each sector. Recently, dedicated PRM players that have been launched within the last few years include companies such as those listed in Table 22.

Table 22: Companies in Four Major Sectors of PRM Market

Dedicated PRM Players	Niche companies	Leading CRM Companies	Leading ERM Companies
Allegis	Asera	Siebel	Oracle
ChannelWave	Backweb	Pivotal	PeopleSoft
Partnerware	Intelic	Onyx	SAP
Radnet	MarketSoft	Saratoga	Management Alchemy
Ten North	Callidus	SalesLogix	Maxit Corporation
Webridge	Trilogy	YouCentric	Softworks

Other niche software companies that bring products to the partner relationship market include Asera, Backweb, Intelic, MarketSoft, Callidus, and Trilogy. CRM players, such as Siebel, Pivotal, Onyx, Saratoga, SalesLogix, and YouCentric (formerly Sales Vision), also offer functionality aimed at managing partner relationships. In addition, established ERM players, such as Oracle, PeopleSoft (which recently acquired Vantive Corp.) and SAP, are beginning to broaden their focus beyond direct sales to include indirect channels [62]. The new realities of the Internet economy and the resulting customer-first emphasis on customer-facing support and productivity, together with expanding ERM functionality, means the Enterprise Relationship Management consulting market continues its rapid growth. The blending with ERP for order fulfillment adds further impetus.

In this supply-side research, IDC analyzes the 1998 and 1999 PRM software market, providing market size, discussion of key market drivers, profiles of selected vendors, and a forecast through 2003. The market is growing rapidly, and new business models are being put into place to meet customer needs. In particular, the Internet has caused a dislocation in both supply and demand, and customers are eagerly rushing to take advantage of the power and versatility that this new technology offers. The delivery of applications through the application service provider (ASP) model also promises to stimulate demand as customers can offload many IT responsibilities and often achieve a positive return on investment (ROI) sooner than with traditional licensing models [67]. Because the Internet had demanded that products and services be delivered in a fast and timely manner, technology companies must deliver the solutions that help their customers meet their corresponding demands [60].

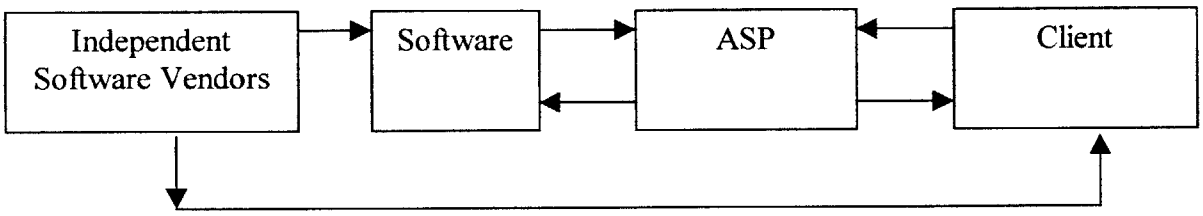
Emergence of ASPs

The convergence of software and IT infrastructure toward an Internet integrated environment has enabled the ASP concept to emerge [68]. Software has evolved from proprietary applications to pre-packaged or off-the-shelf applications and now to the development of Internet centric applications. Net-centric software allows for Web-enabled commerce, communication and the management of information content. Likewise, IT infrastructure has evolved from a closed, mainframe environment to a distributed computing approach resulting in a net-centric infrastructure linking all stakeholders [68]. There will need to be continual advances, particularly in software and broadband technologies to further propel growth in the ASP market. These are two requirements for the future of the ASP market. Without either of these requirements being fulfilled, the ASP future looks dim.

Figure 12 illustrates the many relationships ASPs get involved with. All the software services provided by an ASP can be leased from an Independent Software Vendor. Services are for the most part hosted through centrally located servers of the ASP to the client. It can be the case that an independent software vendor interacts directly with the client to 'rent' out the services the client deems of necessity. Thus, this higher-level diagram depicts the usual representation model of how an ASP company dealing interacting with software vendors will obtain services that clients will be interested in leasing out for a license fee. Thus, the potential exists that everyone from the client to the ASP company to the software vendor all benefit from such transactions. Typically the case is that an ASP aligns with a particular software vendor performing the

initial application implementation and integration along with controlling the data center management and providing continuous uninterrupted connectivity and support [55]. In this fashion, the ASP manages the client relationship acting as a complete end-to-end solution provider. In the end, no matter how the ASP is structured, the ultimate objective is a smooth and no "hiccups" service, in which the client interacts only with the ASP [68].

Figure 12: ASP Relationships



Source: Cherry Tree & Company Reports.

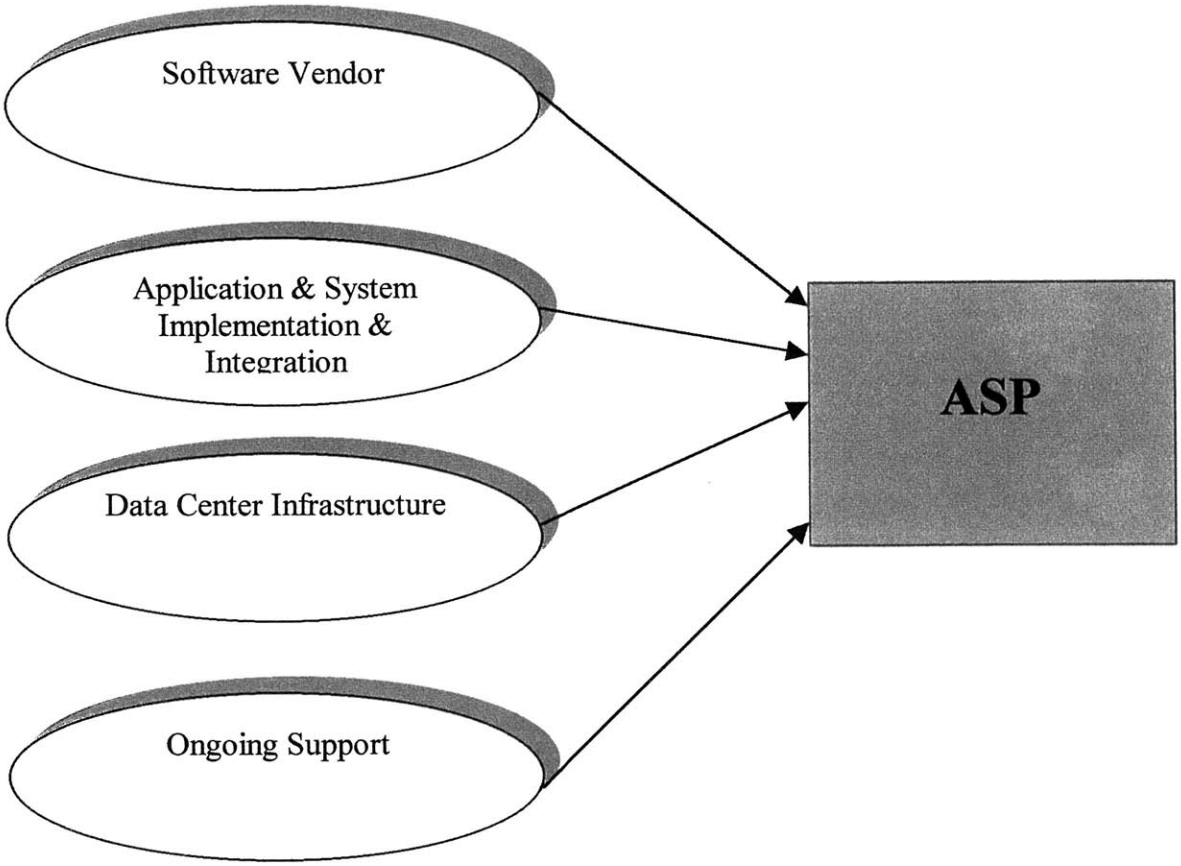
The ASP Business Model

In this section, the overall business model of an ASP will be discussed. Channel relationships are the key to a successful ASP. Without proven software vendors, the future of an ASP looks gloomy. The subsection of ASP specific characteristics discussed the overall strategy of an ASP. Finally, in this section, pricing models will be addressed.

Channel Relationships:

The critical aspects of the ASP channel include software vendors, systems implementation and integration, data center infrastructure and ongoing support (see Figure 13) [68]. As noted previously, software vendors are necessary to provide the software that ASPs will lease out to corresponding clients. To ensure that each application is implemented and integrated smoothly with the rest of system, ASP must support a system to deal with the full spectrum of implementing and integrating all applications. A data center infrastructure is needed to store and maintain a client's processing data. Finally, supporting infrastructure for an ASP is necessary, as more clients will use the resources and services provider by an ASP. Thus, each of the channel components encompasses unique responsibilities that, in aggregate, are necessary to effectively administer a comprehensive ASP solution [68]. The complexity of these fragmented and diverse responsibilities has impacted strategic development among ASP participants. The scope and magnitude of these services has also created new opportunities for IT service providers.

Figure 13: Channel Relationships in ASPs



ASP Characteristics

ASP companies have typically negotiated short-term, non-exclusive licensing agreements with independent software vendors [68]. As illustrated earlier, the advantages of an ASP over ERP systems is that an ASP can deliver any type of software application, from basic e-mail and messaging applications to a complete ERP system that manages, controls and reports on the multiple aspects of an enterprise [68]. ASP provides the pre-packaged application; infrastructure capabilities, the initial and ongoing support services and some degree of customization for each client if they deem further

customization necessary. Today, the level of customization done by ASP is minimal measured by current day standards [68]. In fact, several of the leading ASP have publicly acknowledged the lack of high-level customization. Early ASP market leaders have limited their implementations to core application functionality and, on several occasions, have publicly expressed a disinterest in building highly customized solutions.

ASP Pricing Model

The ASP receives a multi-year contract, normally ranging from 18 to 36 months [68]. Typically a client relationship includes a fixed monthly payment structure ordinarily based on the number of users. However, new technologies are permitting payment schemes based on variable terms such as the number of transactions, the number of screen clicks and amount of usage time. Pricing of the ASP service is a composite of each of the channel responsibilities and their relative costs. At the present time, there seems to be a high degree of uncertainty regarding pricing structures and precisely where the market will assign a price ceiling. Table 17 provides general estimates for direct cost relationships and profitability levels for the ASP. However, cost and profitability will vary based on the complexity of the hosted applications. During the development stage, ASPs will require significant investments to put in place the various resources necessary to manage the ASP relationship. Consequently, pricing and direct cost relationships will vary substantially during this period. Pricing and profitability should improve as economies of scale can be achieved by spreading sunk costs such as data center expenditures across each new client.

A leading ASP company, Interpath, used AcceleratedSAP (ASAP), a rapid-implementation methodology for SAP's flagship R/3 Enterprise Resource Planning (ERP)

product. ASAP is a structured environment that guides, supports, and integrates the activities of an SAP implementation. In the mid-1990s, a typical R/3 implementation spanned 12 to 36 months [76]. This was too long of a process for organizations as these needed to first engage in an internal reengineering effort, then implement R/3, and, customize the solution at the code level.

Table 23 illustrates that current estimates place gross margins in the 30% to 45% range once economies of scale can be recognized [68]. The Phillips Group says the applications hosting market, which last year was pegged at just under $1 billion, is expected to grow to more than $20 billion by 2003 [69]. The Phillips Group further says that average revenue growth of the ASP market is expected to reach nearly 250 percent in 2000 [69]. Meanwhile, Cahners In-Stat Group based Scottsdale, Ariz., has come up with its own ASP industry forecast that also projects rapid growth. Specifically, Cahners In-Stat predicts that some 3 million small businesses will spend more than $7 billion on applications services by 2004 [69]. In 1999, they will spend less than $10 million Cahners In-Stat says. According to researchers at Cahners In-Stat, much of the growth will be driven by small businesses increasing understanding of what the ASP industry has to offer and as ASP form new relationships with established technology firms, especially well-known telecom service providers. Thus, to reach these small companies, smaller players in the industry will have to merge to compete against the more established players in the industry.

Another interesting phenomena occurring in this era of ASPs is the fact that ASPs are obtaining large amounts of venture funding [70]. For instance Intacct Corporation, which bills itself as the leading Internet accounting ASP, late in March, announced a $10

million investment by Hummer Winblad Venture Partners. IBT Technologies Inc., a firm concentrating on online corporate training software and services landed $6.4 million in third round funding led by Counsel Corp. of Toronto. More important than brand name in the ASP industry is actually getting the word out to interested customers. It is more important to provide the functionality than it is to provide a proven brand name [70].

Table 23: Estimates of Costs as % of Revenues of ASP

Costs	% Cost of Revenues Accrued
Systems Implementation & Integration Costs	Incurred in initial stage and are generally charged upfront
Direct Ongoing Costs	
Software License	15%-20%
Data Center & Network Infrastructure	25%-30%
Ongoing Support	15%-20%
Total Cost of Revenue	55%-70%
Targeted Gross Margin	30%-45%

Source: Cherry Tree & Company Reports. 2000.

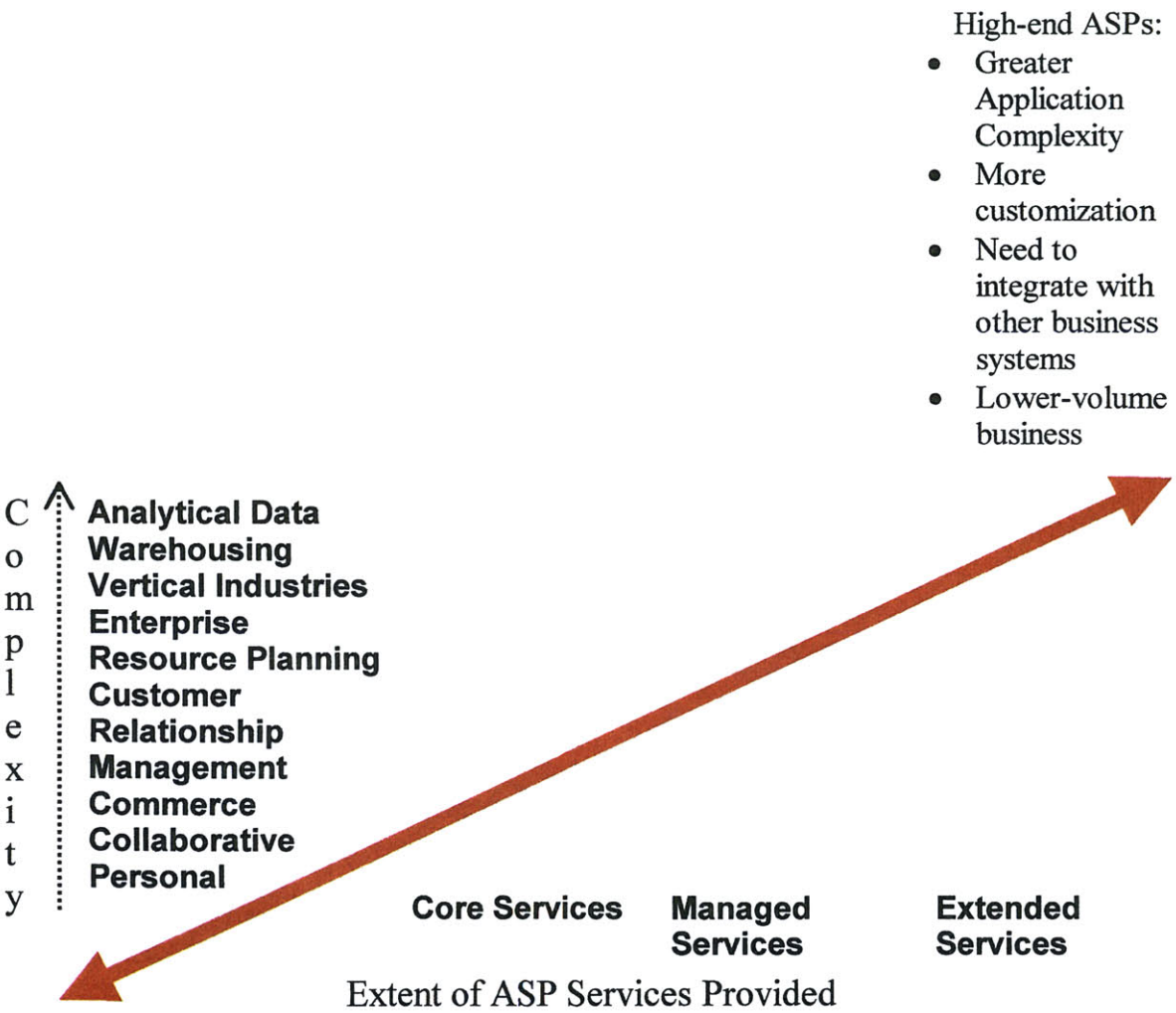
ASP & ERP & Data Mining Tradeoffs:

Figures 14 & 15 summarize some of the key offerings from ASPs along with the demand of applications customers seek. If data-mining companies can find ways to integrate their services with ASPs, then Figure 11 can be redrawn replacing the ERP systems with ASP applications. As Figure 14 shows, since ASPs have control of the number of applications they provide, it may be easier to integrate data-mining tools with ASPs than it is with ERP systems. This is because ERP systems are complex systems that aim to integrate many services. In contrast, ASPs do not have to deal with the concern of integrating existing services. Instead they aim to provide certain types of applications required by clients. Thus, the sole goal of ASPs is to let organizations implement new applications without major staff effort or expense. This relieves overextended IT departments and also gives smaller companies access to sophisticated business applications they could not otherwise afford [98]. Figure 15 shows that IT managers surveyed by InformationWeek Research listed ERP as a top leasing candidate. This survey suggests that companies want ERP but seek to avoid the prohibitive costs, massive IT effort, and high failure rates associated with ERP implementation [98]. Therefore, in general, the ASP delivery model creates a lower-cost with a relatively more rapidly implemented solution.

Though integrating data mining tools with ASPs would seem a clear winner, there are indeed some limitations and problems. First, ASPs may not have all the information necessary to develop a data-mining tool as a back-end solution. Instead, in order for a data-mining tool to work alongside an ASP all critical data and information would need to be passed onto the data-mining tool for analysis. A solution could lie in the approach

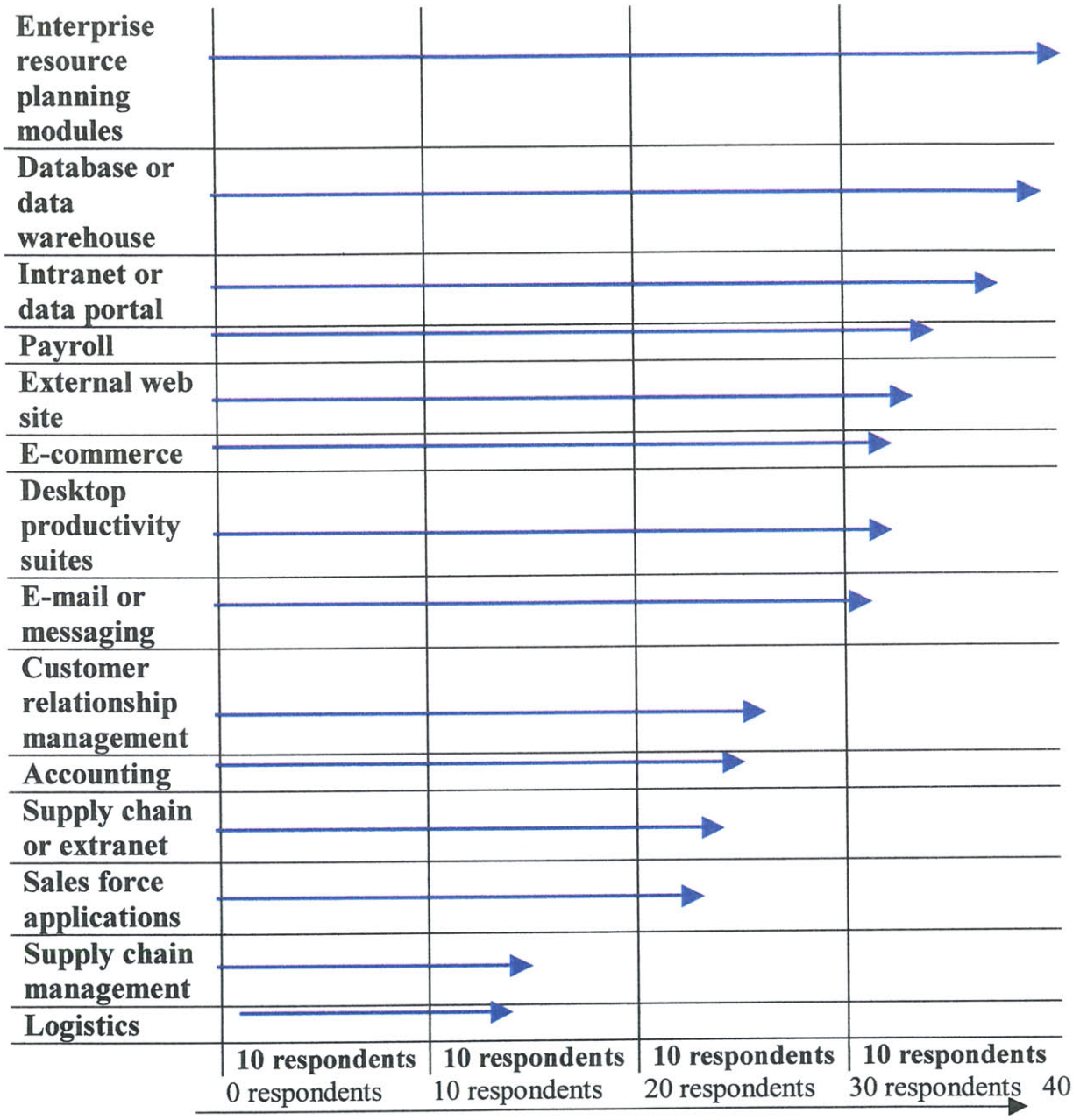
currently being undertaken by ERP providers. Using a data warehouse, ASPs could integrate the critical data and pass that information to the data-mining tool. However, assurance that all critical data was processed would seem to continue to be a problem. It could be the case that because only certain applications are hosted by the ASP, it does not have access to all the information. In this case, given that most if not all critical data can be presented to a data-mining tool from ERP systems, ERP systems using a data-mining tool would seem more reliable than using an ASP with data-mining solutions. Another challenge for ASP integrating with data-mining companies is whether or not this would cause fewer companies to form a contract with ASPs. The reason ASP companies have been successful is because they are leveraging ERP technology to smaller companies. Increasing the price in providing this value-added service like data mining could stir away some companies, as data mining is not a first time guaranteed success. It takes many trials and runs before an optimal solution is found. The exact balance point taking into considerations these factors is unknown at this point. Perhaps the first company to figure out an optimal strategy will gain first movers advantage and offer unparallel services by data mining, ASP, and ERP competing companies.

Figure 14: ASPs Vary in the Types of Applications and Services They Offer



Source: International Data Corporation. Shepard, Susan J. It Shops Take Stock of Application Service Providers. IEEE IT Professional. Volume: 2 Issue: 2. March-April 2000.

Figure 15: Applications Most Likely to be Leased in Year 2000



Respondents expecting to lease via ASPs in the next 12 months (percentage)
Base: 102 ASP adopters. Multiple responses allowed.

Source: InformationWeek Research study of 250 IT managers. Shepard, Susan J. It Shops Take Stock of Application Service Providers. IEEE IT Professional. Volume: 2 Issue: 2. March-April 2000.

Barriers to the ASP Concept:

There are many important issues when it comes to potential barriers that ASP Companies must face. These are:

- Security Concerns
- Quality of Service
- Scalability
- Flexibility
- Adaptability of Software

Each of these issues exemplifies the tough challenge for a successful ASP player. It must first tackle the important concern that there can be security breaches in the central servers of an ASP [60]. This can be catastrophic to a system as intruders can not only obtain proprietary information, but also can modify and delete key information of clients. Quality of service that meets the expectations of client's will be sought after. Guaranteeing a certain percentage of down time and guaranteeing fault isolation of different modules of a client can be a hard to obtain goal. As experience has shown in many system engineering designs, ensuring and meeting a high performance level can be challenging to say the least. Furthermore, ensuring that availability of servers be guaranteed adds to the cost as to ensure availability, replication of servers can be a solution in guaranteeing availability [71]. Scalability is another important consideration in designing an ASP as this allows further functionality to be added to the system [72]. If there is a bottleneck in the network, employing further functionality can only be achieved once the network bandwidth is either increased or system calls are reduced. Flexibility is another challenge, as application demand will change over time resulting in continuous

changes to the system. If the system is configured to be flexible to allow changes promptly, the quicker the response to the customer, and the more satisfied an organizations' customers would be. If changes take an unnecessary amount of time to process, organizations using as ASP will inevitably back out of their agreement with the ASP or even worse, they will stay with the ASP at the expense of losing customers. Lastly, adaptability of the software is another important criteria preventing an ASP from achieving successful implementation. Today, most software is not truly web-enabled [68]. Thus, software applications need to be developed for the Internet to employ an ASP that is truly able to handle with such a challenge as providing latitude of services over the Internet.

ASP Applications

The applications made available by ASP include a wide variety of services. According to IDC, most applications fall into one of the following segments [75]:

- **Analytic applications** are applications built to analyze business problem such as financial analysis, customer churn analysis, and Web site analysis
- **Vertical applications** include industry-specific applications such as MRP in the manufacturing industry
- **ERM applications** include accounting, human resouce management, and materials management
- **CRM applications** include segments such as sales force automation, customer service, and marketing applications
- **Collaborative applications** include groupware, email, and conferencing applications

• **Personal applications** include Office suites such as Microsoft Office and consumer applications

Leading Companies in ASP Market Niche:

Currently it is difficult to predict which companies are going to set the standard by which future ASP companies will follow. It appears that the market is still undergoing many changes before any one particular standard can settle in. Furthermore there are so many services that taken as a whole ASP companies are trying to offer. It is further unclear which of these services will be prominent in the future. To complicate matters worse, companies that may be involved with ERP systems may engage in providing ASP services to smaller companies. ASPs are gaining credibility even with large enterprises now eager to outsource applications, not just within a department but also across the entire enterprise [76]. As such, one of the more established ERP players can take hold of the industry by consolidating smaller players and creating a solution that meets most of the needs of clients. Through 2001, ASP winners and losers will be determined by their ability to manage their cost structures, their respective choice of partners, their name recognition, and the repeatable solutions along with their ability to execute successfully on required service levels [74]. Thus, it is too early to properly state what the proper solution or what company has the advantage in its development efforts. Companies that are currently focusing on providing only ASP services include:

- **USinternetworking (NASDAQ: USIX)**, perhaps the most recognized ASP, became publicly traded in late May of 1999. The investment community immediately assigned a billion-dollar market valuation to the company with a 12-

month trailing revenue base of $15 million [68]. The company currently offers one of the most comprehensive sets of enterprise hosted solutions (see Table 24). This means that they provide enterprise wide applications to smaller organizations wanting the services offered by these enterprise systems. As such, the wide range of services provided by this leading ASP company is impressive. Specific details of each of these services can be found at their corresponding web site: http://www.usinternetworking.com. Contract terms range from 18 to 36 months, with average fixed monthly recurring fees of $33,000 for an enterprise solution [68]. Their client list currently consists of 32 companies. Those can also be found in the web site.

Table 24: Usinternetworking ASP Services

Application Solution	Independent Software Vendor Partner
Financial Management and Human Resources	PeopleSoft Lawson Software
Enterprise Customer Relationship Management/Resouce Planning	Siebel
E-Commerce	Broadvision Microsoft
General Business Services	Sagent Technology
Enterprise Messaging	Microsoft
Professional Services Automation	Niku
Application Hosting Program	Actuate

Souce: Cherry Tree Company Reports. 2000.

- **Research Online International** (ROI): Founded in 1994, ROI Direct.com offers e-commerce site development and hosting, e-marketing program design and execution, and online customer care solutions tailored to clients' needs. Leading provider of application services, 24x7 systems management, account support, feature upgrades, and shared use of $1Million+ systems infrastructure for less

than our clients typical costs of a single IT employee. By customizing a claimed reusable set of advanced applications, they are able to deliver these solutions in a fraction of the time of in-house or competing third-party solutions.

- **ChannelWave** is an application service provider (ASP) focused on providing partner eCertification solutions. The company provides Web-hosted applications that allow companies to grow revenue by increasing the effectiveness and mindshare of their channel partners.
- **Corio** founded in 1998, is a privately held company that also has received substantial notoriety as a pioneering ASP. Corio is currently one of the only pure-play ASP offering an integrated solution featuring Siebel and PeopleSoft [73]. Corio is considered to be an enterprise ASP [75]. Its focus is to provide a fully integrated suite of enterprise applications spanning customer relationship management, electronic commerce, accounting, human resources, supply-chain and vertical specific applications such as manufacturing applications. Corio determines the most appropriate software vendor for its customer base by evaluating the leading software providers in each category and then selecting a partner that it determines to be the most appropriate for Corio's target customers [68]. Management believes establishing third-party channel partnerships and focusing on a few, select applications provide a distinct competitive advantage. Corio has focused exclusively on PeopleSoft and Siebel enterprise applications and has partnered with Concentric Networks and Exodus Communications for the management of its data center infrastructure [73]. Corio intends to deliberately

build strategic relationships with third-party ISVs by selectively partnering with "best of breed" enterprise application solutions.

- **FutureLink**, founded in 1997, is positioned as an end-to-end ASP solution internally managing its ASP channel. The company claims to tailor its hosted applications to the specific needs of the customer rather than offer a limited hosted application portfolio [68]. FutureLink has established software vendor relationships with Great Plains Software, Applix, Galleon Distributed Technologies, Microsoft and Onyx.
- **Interpath** is a network-based Application Service Provider (ASP) serving midsize and large enterprise customers in the mid-Atlantic and southeast regions of the United States. Interpath offers a range of enterprise applications, data services, and e-Commerce solutions. It also provides Internet business consulting. To deliver end-to-end capabilities backed by meaningful service level agreements (SLAs), Interpath has developed strong - even groundbreaking - partnerships with SAP, Sun Microsystems, IBM DB2, and IBM Global Services [76]. Interestingly enough, Interpath was the first ASP to enter into a ground-breaking business relationship with SAP by licensing software directly from SAP for rental to Interpath customers. Negotiated in the spring of 1999, it was indeed a milestone in ASP history. Since that time, SAP has introduced a monthly lease option, giving customers the flexibility of monthly license payments.
- **Telecomputing ASA**, founded in Norway in 1997, TeleComputing is one of the world's leading ASP, now supporting more than 100 customers in 200 locations

throughout the US and Europe. The company has a complete hosted application solution and also has several applications specific to vertical markets [68].

- **AristaSoft** is a leading industry-focused application service provider (ASP), offering integrated IT solutions designed to meet the specific needs of emerging high tech equipment companies.
- **ServiceNet** is a joint venture between Andersen Consulting and GTE Internetworking. The company is in its early stages and is currently seeking to establish software partner vendors.
- **Global Recruiting Solutions** is an HR applicant tracking and hiring process management vendor, whose product is completely Web-enabled.

For a complete listing of all ASP companies with a brief description, please refer to Table 25.

Table 25: Complete Listing of ASPs

• **Agiliti, Inc.** – allows businesses to procure, administer, and control business applications.
• **AirFlash.com** – application service provider for mobile relevant services.
• **Air2Web** – offers a mobile application platform that allows business to deliver their services to users regardless of device or network. Supports audio and voice recognition.
• **Applicast, Inc.** – providing implementation, outsourcing, and management services for enterprise applications, including SAP R/3 and Siebel.
• **Applications2GO** – service that gives Independent Software Vendors (ISVs) immediate access to the skills and technologies to an ASP business solution.
• **AppServ, Inc.** – provides and hosts full featured applications, such as mailing lists, message boards, and chat rooms.
• **Breakaway Solutions** – full service provider of e-business solutions to

growing enterprises, including strategy solutions, software application solutions, customer relationship management solutions and application hosting.
• **Citrix Systems, Inc.** – supplier of application server products and technologies that enable the effective and efficient enterprise-wide deployment and management of applications designed for Microsoft Windows operating systems. The Company's MetaFrame and WinFrame product lines, developed under license and strategic alliance agreements with Microsoft Corporation, permit organizations to deploy Windows applications without regard to location, network connection, or type of client hardware platforms.
• **CmeRun** – offers consumer applications online.
• **Concur Technologies, Inc.** – a leading provider of workplace eCommerce solutions that automate costly and inefficient purchasing and business processes. Concur's suite of solutions include corporate procurement, human resources self-service, and travel and expense management, integrated through the Concur eWorkplace enterprise business portal. Concur's solutions are provided in licensed and ASP models. Concur's global trading network, Concur Commerce Network, links buyers with suppliers to conduct business-to-business eCommerce for corporate goods and services.
• **Corio** – application service provider (ASP) of integrated business applications and services.
• **eAlity** – offers a library of rentable online applications that automate business processes.
• **eHNC Inc.** – applications service provider (ASP) of customer interaction management products. Formerly Aptex Software Inc.
• **Eloquent, Inc.** – offers products and services for transferring knowledge to large, geographically dispersed audiences, including Web-based on-demand video, audio, text, and graphics.
• **eOnline** – focused on delivering the enterprise application solutions.
• **Esoft** – ESOFT Global utilizes its specialist server-based computing, Internet, systems management and security skills to ASP enable, deliver and manage leading line-of-business applications over the Internet or leased lines to Small to Medium Enterprises (SME) and Corporate business users.
• **HostLogic, Inc.** – Managed Application Provider (MAP) that specializes in hosting industry-specific enterprise software solutions.

• **InsynQ** – services include hosted applications, documents and file sharing, security, file back-up and data storage.
• **Interliant** – provider of a wide range of hosting and enhanced Internet services that enable its customers to deploy and manage their Web sites and network-based applications more effectively than internally developed solutions. Interliant's hosting services store its customers' Web sites, software applications, and data on servers typically housed in its data centers so that others on the Internet can access and interact with its customers' Web sites and network-based applications.
• **Interpath** – Application Service Provider (ASP) providing network services and virtual private applications (VPAs).
• **Jamcracker** – aggregates and integrates web-based IT and business services, applications, and tools.
• **MyCIO.com** – delivers managed, hosted network security and availability services. From Network Associates.
• **NaviSite** – provider of outsourced Web hosting and application services for companies conducting mission-critical business on the Internet. The Company helps customers focus on their core competencies by outsourcing the management and hosting of mission critical Web sites and applications. NaviSite is a majority-owned operating company of CMGI, Inc with minority investment from Microsoft Corporation. NaviSite's SiteHarbor solutions provide secure, reliable co-location and high-performance hosting services, including high-performance Internet access and high-availability server management solutions through load balancing, clustering, mirroring and storage services.
• **Oracle Business OnLine** – specializes in web-enabled applications.
• **OutBlaze** – provides free private label portal services.
• **Spirian Technologies, Inc.** – technology deployment specialists providing IT reliability services through the ASP model for large-scale users.
• **TeleComputing, Inc.** – provides application hosting and computing services, and also manages the entire IT infrastructure for the end user.
• **Ten North, Inc.** – full-service Application Service Provider (ASP) for making third-party sales channels.
• **Tendo Corporation** – offers a suite of Internet based workflow applications.

• **Thinter.net** – application hosting for software developers and resellers.
• **Usinternetworking Inc.** – internet managed applications provider.
• **WebOnTap** – provides online applications, for both web authoring and back-office tasks, delivered and maintained through a Web browser.
• **Xuma Technologies** – software development company dedicated to engineering mission critical applications for businesses on the Web.

Souce: Yahoo Corporation. Data Mining Companies. 2000.

Sample ASP Infrastructure

A leading ASP provider, Interpath specializes in delivering end-to-end support of business applications. Interpath is an integrating ASP, assembling best-of-breed components to provide its customers with brand-name solutions. Interpath possesses substantial communications infrastructure, including a 2,000 fiber-mile network, a high-speed ATM core, and a national backbone with Internet connectivity at over 100 peering points for swift transit of data over the Internet. Interpath also maintains exceptional data center facilities.

ASP Commercial Products of Choice

Aberdeen research confirms that Sun servers power the core of the Internet [76]. Sun's platform including its servers, storage, and the Solaris operating system offer high reliability and scalability [76]. Solaris is the standard for mission-critical database as SAP's largest customers run their applications on the Sun platform. As a result, that reliability has drawn many ASPs to the Sun platform. Similar to the preference of Solaris platforms, DB2 is preferred as a database to other competitors. IBM's DB2 Universal Database has features optimized across a range of applications - Internet, intranet,

extranet, data warehousing, data mining OLTP and OLAP [76]. SAP uses the DB2 Universal Database as the primary database on IBM, Sun, and Linux platforms for internal development and production systems. Sun Enterprise series servers provide the processing power for both the SAP application and DB2 database servers [76].

Future Thoughts of ASPs Integrating with Data Mining:

As has been stated thus far, the future possibilities of ASP are endless. Integrated with a data mining solution, a company can emerge as the leader of providing not only the services needed to consolidate a client's proprietary information, but in addition can provide a tool that can extract relevant patterns in the data. Data mining offers many ways to optimize findings in data and as such the techniques discussed in earlier sections can be applied to yield in knowledge discovery for even small to middle sized companies using the ASP model. If larger companies can afford the implementation costs of ERP systems, then these companies can also benefit from the added value of using data mining techniques applied to the collected data. As was illustrated earlier, the more data that is collected, the better it is for clients using data mining applications. In the case of using an ASP, a client can run into trouble in having to rely on outside sources mine their data, as clients can potentially not directly be overseeing the collected data generated by the applications being used. Thus, clients using an ASP along with wanting to seek data mining applications will need to trust their corresponding ASP to deliver the correct and optimal solutions for them.

ASP Conclusion:

The promising ASP market represents a dramatic change in application delivery. While today the market is merely an extension of applications outsourcing, in the long term, it will become a dramatically different way to deliver application functionality and related services. As products and tools become available to enable Web-based application delivery, the ASP market will find customers everywhere availing themselves of these services. Yet, many vendors will not be able to provide satisfactory levels of service as their customer base grows. Recent Aberdeen research reveals that ASPs are gaining credibility among even the largest enterprises that have chosen to outsource complex applications across the entire company - not simply within a single division. For these companies, the benefits of the ASP proposition are simply too numerous to ignore.

ASPs are helping to level the playing field by making applications that support customer relationship management available to all organizations and allow them to act quickly to offer their products at Internet speed. This empowers businesses to provide customers and prospects with faster, more responsive, and personalized services that increase loyalty and result in more profitable, long-term relationships.

Conclusions:

Data mining is a technology that has emerged to provide organizations whether large or small the opportunity to discovery-hidden trends and patterns in their data. This realization has come about as a result of the increasing loads of data being stored in organization's databases. To take advantage of this storage data mining can use a data warehouse to manage the data before applying a data mining application. The reasons data mining has caught the attention of so many companies is that data mining has proven itself as a satisfactory tool. With the advent of ERP and ASP companies making progress in providing leading products and services, consolidation of data mining services alongside these services is a challenging path that can lead to very promising results. The future is still very uncertain. Because of the value that ERP and ASP companies can provide to organizations through their respective tools, an even greater benefit to companies is providing a data-mining tool that further analyzes the data.

In summary, data mining works by simply learning from data it can. There is no dependence on the skills or intuition of a programmer to synthesize a model. Models are created automatically from the data and represent an unbiased distillation of the business experience. Decision models are generated automatically. This is a fast, cost-effective procedure. Models are easily updated by re-learning. Many of the techniques can handle many input factors (for example, Clementine has used 7000/record) while the same techniques can ignore input factors that do not contribute to a particular decision [2]. Some of the techniques, notably neural networks, can discover complex non-linear models. Lastly, good data mining tools allow users to mix and match many techniques to

solve problems. Users' confidence in the results is increased if multiple techniques are deployed, and they all provide similar predictions.

In conclusion, through the combinations of ERP and ASP systems, organizations can better manage and store their data. The advantage of using an ERP system is that it allows a convenient consolidation of a client's data. However, the drawback to ERP systems is that they can take a long time to implement along with being very costly to a customer. Thus, an emerging technology are ASP companies that seek to provide services to mid-tier companies interested in obtaining the same benefits of an ERP system. Consolidating these services provided by either of these two tools with a data mining solution is a valuable service to any client interested in obtaining both a more efficient data management system and having the ability of finding hidden patterns within the data. The wide range of applications of data mining will continue to progress into potentially web-mining tools where applications will be developed towards a net-centric approach where data storage and transfers can be done through the Internet. At this point in time the possibilities are endless as there is no clear market leader and there are many leading packages such as ERP, PRM, and CRM that could benefit from consolidating their products with leading data mining tools. As the market continues to grow, ASPs could make the leap in being able to consolidate these services sooner than ERP companies because of the smaller size of their respective clients. Only if ERP companies can find a way to easily integrate their services and offer a bundled package can they take over the market.

Bibliography/References:

1. Data Mining: Extending the Information Warehouse Framework: Data Management Solutions. IBM Whitepaper. IBM. 2000.

2. Data Mining: An Introduction. SPSS Whitepaper. SPSS. 2000.

3. An Introduction to Data Mining. Pilot Software Whitepaper. Pilot Software. 1998.

4. Jain, A.K. Duin, P.W. and Jianchang, Mao. Statistical Pattern Recognition: A Review. Pattern Analysis and Machine Intelligence, IEEE Transactions. January 2000.

5. Alter, S.L. Decision Support Systems: Current Practice and Continuing Challenge. Addison-Wesley. 1980.

6. W.H. Inmon. Tech Topic: What is a Data Warehouse? Prism Solutions. Volume 1. 1995.

7. Data Mining News. Intelligent Data Analysis Group. 2000.

8. Edelstein, Herb. Data Mining News "Two Crows Releases 1999 Technology Report". Volume 2, number 18. 10 May 1999.

9. Cisco, General Growth Properties Link Mall Retailers Online. InternetNews.Com. May 12, 2000.

10. Offline Spending by Internet Brands Passes $1 Billion. Internet Newsletter. 1999.

11. Piatetsky-Shapiro, Gregory. The Data-Mining Industry Coming of Age. IEEE Intelligent Systems. 2000.

12. Kaashoek, Frans. 6.033 Lessons. MIT. 2000.

13. Piatetsky-Shapiro, Gregory. Knowledge Discovery in Databases: An Overview. AI Magazine, Fall 1992. Pages 213-228.

14. Brands, Estelle and Gerritsen, Rob. Assocation and Sequencing. DBMS, Data Mining Solutions Supplement. Miller Freeman, Inc. 1998.

15. Brand, Estelle. Classification and Regression. DBMS, Data Mining Solutions Supplement. Miller Freeman, Inc. 1998.

16. Riggs, JL. Production Systems. Wiley. 1987.

17. Klerfors, Daniel. Artificial Neural Networks. Saint Louis University. Nov 1998.

18. Data & Analysis Center for Software. Artificial Neural Networks Technology. August 1992.

19. Thinx Technology White Paper. Visualization Technology. Thinx Software. 1998.

20. Williams, Graham and Huang, Zhexue. Mining the Knowledge Mine: The Hot Spots Methodology for Mining Large Real World Databases. Artificial Intelligence, Volume 1342, Advanced Topics in Artificial Intelligence. 1997.

21. Lueg, Christopher. Issues in Understanding Collaborative Filtering. AI-Lab, Department of Computer Science. University of Zurich. 1999.

22. Maltz, David and Ehrlich, Kate. Pointing the Way: Active Collaborative Filtering. Proceedings of the Annual ACM Sigchi Conference on Human Factors in Computing Systems. ACM Press. May 1995.

23. Good, Nathaniel Schafer, Ben J. Konstan, Joseph A. Borchers, Al Sarwar, Badrul Herlocker, Jon, and John Riedl. Combining Collaborative Filtering with Personal Agents for Better Recommendations. Grouplens Research Project. American Association for Artificial Intelligence. 1999.

24. Messinger, Eli Shoens, Kurt Thomas, John and Luniewski, Allen. Rufus: The Information Sponge. IBM Research. 1991.

25. Resnick, Paul Iacovou, Neophytos, Suchak, Mitesh Bergstrom, Peter and Riedl, John. GroupLens: An Open Architecture for Collaborative Filtering of Netnews. Association for Computing Machinery. 1999.

26. Breese, Jack Heckerman, David Kadie, Carl. Empirical Analysis of Predictive Algorithms for Collaborative Filtering. Microsoft Research. 1999.

27. Pennock, David M. Horvitz, Eric. Collaborative Filtering by Personality Diagnosis: A Hybrid Memory- and Model-Based Approach. Microsoft Research. 1999.

28. Berry, Michael J.A. The Privacy Backlash. Information Access Company. Oct. 1999.

29. SPSS Employees. Report OLAP: Technology Options for Decision Makers. SPSS Inc. 2000.

30. Pedersen, Torben B. Jensen, Christian S. Multidimensional Data Modeling for Complex Data. Proceedings of the 15th International Conference on Data Engineering. Institute of Electrical and Electronics Engineers, Inc. 1998.

31. Barker, Paul. The Brains and Drawn of Business: Managing Strategic Business Directions Through ERP. ERP World 1999 Proceedings. 1999.

32. Maude, Dan. Balancing Time, Scope, and Cost in ERP Implementations. ERP World 1999 Proceedings. 1999.

33. Pallesen, Soren. How to Use the Internet to Reduce Sales Costs and Sell More. ERP World Proceedings. 1999.

34. Bransby, Michele. ERP and Automated Data Collection. ERP World Proceedings. 1999.

35. Ripma, Mark. Vendor Care and Feeding. ERP News. 1999.

36. Welti, Norbert. Successful SAP R3 Implementation: Practical Management of ERP Projects. Addison-Wesley Pub Co. 1999.

37. Shtub, Avraham. Enterprise Resource Planning: The Dynamics of Operation Management. Kluwer Academic Publishing. 1999.

38. Whiting, John T. Putting "e" technology into perspective: The integration of ERP and IT e-business components. TechRepublic. 2000.

39. Shankarnarayanan S. ERP Systems -- Using IT to gain a competitive advantage. Baan Infosystems India Pvt.Ltd. 1998.

40. Holland, Christopher and Light, Ben. Global Enterprise Resource Planning Implementation. Systems Sciences. Proceedings of the 32nd Annual Hawaii International Conference. 1999.

41. Fitzgerald, A. Enterprise Resource Planning-Breakthrough or Buzzword? Factory 2000. Competitive Performance Through Advanced Technology, Third International Conference. 1992.

42. Krasner, Herb. Ensuring e-business success by learning from ERP failures. IT Professional. 2000.

43. Holland, Christopher and Light, Ben. A Critical Success Factors Model for ERP Implementation. IEEE. 1999.

44. Ross, Jeanne. Surprising Facts About Implementing ERP. IT Professional, IEEE. July 1999.

45. Tjoa, Bhieng, Raman, Ramesh, Itou, Toshiaki, and Natori Yukikazu. Impact on Enterprise Wide Supply-Chain Management Techniques on Process Control. Proceedings of the 1999 IEEE International Conference. 1999.

46. Fleisch, Edgar and Powell, Stephen. The Value of Information in a Business Network. Proceedings of the 32nd Hawaii International Conference on System Sciences. 1999.

47. Mizoras, Amy. Worldwide ASP Market Size and Forecast, 1999-2004. IDC Publishers. 1999.

48. Big Prospects for the ASP Market. EMarketer Newsletter. 2000.

49. EGlobal Report. Emarketer Reports. 2000.

50. Saltzer, Jerome H. The Network of a System and as a System Component. MIT. 2000.

51. Kaashoek, Frans. Protection of Information in Computer Systems. MIT. 2000.

52. Raghunath, Ramanujam, Copp, Craig, and Prock, Darren. Key Components of a Successful ERP-Based e-Commerce Implementation. ERP World Organization. 2000.

53. Mizoras, Amy. A Tale of Two ASP-Delivered Applications: Case Study of USinternetworking's Customer, PSDI. International Data Corporation. March 2000.

54. Mason, Paul. The Management Service Provider Appears: Some Early Offerings. International Data Corporation. March 2000.

55. Presti, Ken. The ASP Model: How Resellers Fit In. International Data Corporation. March 2000.

56. Wilson, Gisela. Internet-Based Materials Management: eCommerce Creates New Opportunities. International Data Corporation. April 2000.

57. Whalen, Meredith. eCommerce ASPs: How Will eCommerce Projects Become the Next Battleground for ASPs and Internet Service Firms?. International Data Corporation. April 2000.

58. Hodges, Judy and Shiang, David. Partner Relationship Management Software: 2000 Worldwide Markets and Trends. International Data Corporation. March 2000.

59. Ebbesen, Anders. CRM and ASP in the Middle Market: A Window of Opportunity. International Data Corporation. March 2000.

60. Garone, Steve and Cusack, Sally. Application-Enabling Technologies. International Data Corporation. February 2000.

61. Murray, Steve. Web Hosting Trends: ISPs Jump on the ASP Bandwagon. International Data Corporation. June 1999.

62. McCarty, Meredith and Gillan, Clare. Hewlett-Packard, SAP and Qwest Partner for ASP-delivered Solutions. International Data Corporation. May 1999.

63. McHale, Steve, Gillan, Clare, and Symonds, Steven. The Emerging ASP: A Look Inside the Business Model. International Data Corporation. September 1998.

64. Murphy, Cynthia, McCarty, Meredith, Tan, Susan, and Gere, Traci. Application Outsourcing: An Emerging Service Opportunity. International Data Corporation. December 1998.

65. Partwise Inc. ASP/PRM Market Analysis. Saratoga Group. 1999.

66. Skipper, John. Electronic Banking and Payments. IEEE. 1998.

67. Arjunan, Mallik. Maximizing ROI by Minimizing Process Variability. IEEE. Reliability and Maintainability Symposium. 1991.

68. Application Service Providers Spotlight Report. Cheerytree & Company. 1999.

69. Fairchild, David. Studies forecast massive opportunity for ASP industry as demand for enterprise and consumer applications grows. BizSpace Inc. 2000.

70. ASP Industry Review. BizSpace Inc. 2000.

71. Saltzer, Jerome H. The Protection of Information in Computer Systems. IEEE. September 1975.

72. Birrell, Andrew D, Levin, Roy, Needham, Roger M., and Schroeder, Michael D. Grapevine: An Exercise in Distributed Computing. Communications of the ACM. April 1982.

73. Nairne, Christine and Lim, Victor. APPS for all, E*Offerings Take on the ASP Sector. E*Offering Corporation. October 1999.

74. ASP Trends: The ASP Model Moves Closer to 'Prime Time'. Gartner Group. 2000.

75. Gillan, Clare and McCarty, Meredith. ASPs Are for Real, But What's Right for You? International Data Corporation. 1999.

76. Aberdeen White Paper:ASP Model. Aberdeen Group, Inc. 2000.

77. Saraee, Mohammad, Koundourakis, George, and Theodoulidis, Babis. EasyMiner: Data Mining in Medical Databases. The Institution of Electrical Engineers. 1998.

78. Agrawal, Rakesh. Mining Quantitative Association Rules in Large Relational Tables. Proceedings of the ACM International Conference on Very Large Databases. September 1995.

79. King, Michel and Elder, John. Evaluation of fourteen desktop data mining tools. Systems, Man, and Cybernetics. IEEE International Conference. Volume: 3. 1998.

80. Chen, Ming. On the Evaluation of Attribute Information for Mining Classification Rules. Tools with Artificial Intelligence. Tenth IEEE International Conference. Nov. 1998.

81. Hambaba, Mohamed. Intelligent Hybrid System for Data Mining. IEEE. 1998.

82. Advances in Knowledge Discovery and Data Mining. American Association for Artificial Intelligence. Press/ The MIT Press. 1996.

83. Clarkson, Philip and Moreno, Pedro. On the use of support vector machines for phonetic classification. Acoustics, Speech, and Signal Processing. 1999 IEEE International Conference. Volume 2. 1999.

84. Lu, J., Quaddus, M.A. and Williams, R. Developing a Knowledge-Based Multi-Objective Decision Support System. System Sciences. Proceedings of the 33rd Annual Hawaii International Conference. 2000.

85. Goil, Sanjay and Choudhary, Alok. Design and Implementation of a Scalable Parallel System for Multidimensional Analysis and OLAP. 13th International and 10th Symposium on Parallel and Distributed Processing. 1999.

86. Qiming Chen and Meichun Hsu. A Data-Warehouse / OLAP Framework for Scalable Telecommunication Tandem Traffic Analysis. Data Engineering Proceedings. 16th International Conference. 2000.

87. Proceedings of the Thirty-First Hawaii International Conference on System Sciences. Proceedings of the Thirty-First Hawaii International Conference. Volume: 2. 1998.

88. Manjunath, B.S. Huang, T. Teklap, A.M. and Zhang, H.J. Guest Editorial: Introduction to the Special Issue on Image and Video processing for Digital Libraries. Image Processing, IEEE Transactions. Volume: 9 Issue: 1. Jan. 2000

89. Nature, International Weekly Journal of Science. Macmillan Publishers. July 1998.

90. Technology Special Statistics Report. BYTE, CMP Media Incorporated. September 1998. Spanish Edition.

91. Comparative Study of Statistical Packages. Scientific Computing and Automation. September 1998.

92. Brause, R.; Langsdorf, T.; Hepp, M. Neural data mining for credit card fraud detection. Tools with Artificial Intelligence. 11th IEEE International Conference. 1999.

93. Reinartz, Thomas. Focusing Solutions for Data Mining: Analytical Studies and Experimental Results in Real-World Domains. Springer-Verlag Berlin Heidelberg. 1999.

94. Davenport, T.J. Enterprise Resource Planning Systems. Harvard Business Review. July-Aug 1998.

95. Gibson, Nicola Holland, Christopher and Light, Ben. Enterprise Resource Planning: A Business Approach to Business Development. Proceeding of the 32nd Hawaii International Conference on System Sciences. 1999.

96. Yurong, Yao and Houcon, He. Data Warehousing and the Internet's Impact on ERP. IEEE IT Professional. March-April 2000.

97. Burris, Peter and Bradley, Anthony. Database Pricing Trends. META Group. August 1999.

98. Shepard, Susan J. It Shops Take Stock of Application Service Providers. IEEE IT Professional. Volume: 2 Issue: 2. March-April 2000

99. Elder, John F. and Abbott, Dean W. A Comparison of Leading Data Mining Tools. Fourth Annual Conference on Knowledge Discovery & Data Mining. August 1998.

100. Abbott, Dean Matkovsky, Philip Elder, John. Systems, Man, and Cybernetics. An Evaluation of High-end Data Mining Tools for Fraud Detection. IEEE International Conference. Volume: 3. 1998.

101. Data Mining Market Share. Data Mining News. Volume 1, Number 18. May 1998.

Appendix:

Glossary of Data Mining Terms [3]:

analytical model	A structure and process for analyzing a dataset. For example, a *decision tree* is a model for the *classification* of a dataset.
anomalous data	Data that result from errors (for example, data entry keying errors) or that represent unusual events. Anomalous data should be examined carefully because it may carry important information.
artificial neural networks	*Non-linear predictive models* that learn through training and resemble biological neural networks in structure.
CART	Classification and Regression Trees. A decision tree technique used for *classification* of a dataset. Provides a set of rules that you can apply to a new (unclassified) dataset to predict which records will have a given outcome. Segments a dataset by creating 2-way splits. Requires less data preparation than CHAID.
CHAID	Chi Square Automatic Interaction Detection. A decision tree technique used for *classification* of a dataset. Provides a set of rules that you can apply to a new (unclassified) dataset to predict which records will have a given outcome. Segments a dataset by using chi square tests to create multi-way splits. Preceded, and requires more data preparation than, CART.
classification	The process of dividing a dataset into mutually exclusive groups such that the members of each group are as "close" as possible to one another, and different groups are as "far" as possible from one

	another, where distance is measured with respect to specific variable(s) you are trying to predict. For example, a typical classification problem is to divide a database of companies into groups that are as homogeneous as possible with respect to a creditworthiness variable with values "Good" and "Bad."
clustering	The process of dividing a dataset into mutually exclusive groups such that the members of each group are as "close" as possible to one another, and different groups are as "far" as possible from one another, where distance is measured with respect to all available variables.
data cleansing	The process of ensuring that all values in a dataset are consistent and correctly recorded.
data mining	The extraction of hidden predictive information from large databases.
data navigation	The process of viewing different dimensions, slices, and levels of detail of a *multidimensional database*. See OLAP.
data visualization	The visual interpretation of complex relationships in multidimensional data.
data warehouse	A system for storing and delivering massive quantities of data.
decision tree	A tree-shaped structure that represents a set of decisions. These decisions generate rules for the *classification* of a dataset. See CART and CHAID.
dimension	In a flat or relational database, each field in a record represents a dimension. In a *multidimensional database*, a dimension is a set of similar entities; for example, a multidimensional sales database might include the dimensions Product, Time, and City.
exploratory data analysis	The use of graphical and descriptive statistical techniques to learn about the structure of a dataset.

genetic algorithms	Optimization techniques that use processes such as genetic combination, mutation, and natural selection in a design based on the concepts of natural evolution.
linear model	An *analytical model* that assumes linear relationships in the coefficients of the variables being studied.
linear regression	A statistical technique used to find the best-fitting linear relationship between a target (dependent) variable and its predictors (independent variables).
logistic regression	A linear regression that predicts the proportions of a categorical target variable, such as type of customer, in a population.
multidimensional database	A database designed for *on-line analytical processing*. Structured as a multidimensional hypercube with one axis per *dimension*.
multiprocessor computer	A computer that includes multiple processors connected by a network. See *parallel processing*.
nearest neighbor	A technique that classifies each record in a dataset based on a combination of the classes of the k record(s) most similar to it in a historical dataset (where k is greater than or equal to 1). Sometimes called a k-nearest neighbor technique.
non-linear model	An *analytical model* that does not assume linear relationships in the coefficients of the variables being studied.
OLAP	On-line analytical processing. Refers to array-oriented database applications that allow users to view, navigate through, manipulate, and analyze *multidimensional databases*.
outlier	A data item whose value falls outside the bounds enclosing most of the other corresponding values in the sample.

	May indicate *anomalous data*. Should be examined carefully; may carry important information.
parallel processing	The coordinated use of multiple processors to perform computational tasks. Parallel processing can occur on a *multiprocessor computer* or on a network of workstations or PCs.
predictive model	A structure and process for predicting the values of specified variables in a dataset.
prospective data analysis	Data analysis that predicts future trends, behaviors, or events based on historical data.
RAID	Redundant Array of Inexpensive Disks. A technology for the efficient parallel storage of data for high-performance computer systems.
retrospective data analysis	Data analysis that provides insights into trends, behaviors, or events that have already occurred.
rule induction	The extraction of useful if-then rules from data based on statistical significance.
SMP	Symmetric multiprocessor. A type of *multiprocessor computer* in which memory is shared among the processors.
terabyte	One trillion bytes.
time series analysis	The analysis of a sequence of measurements made at specified time intervals. Time is usually the dominating dimension of the data.