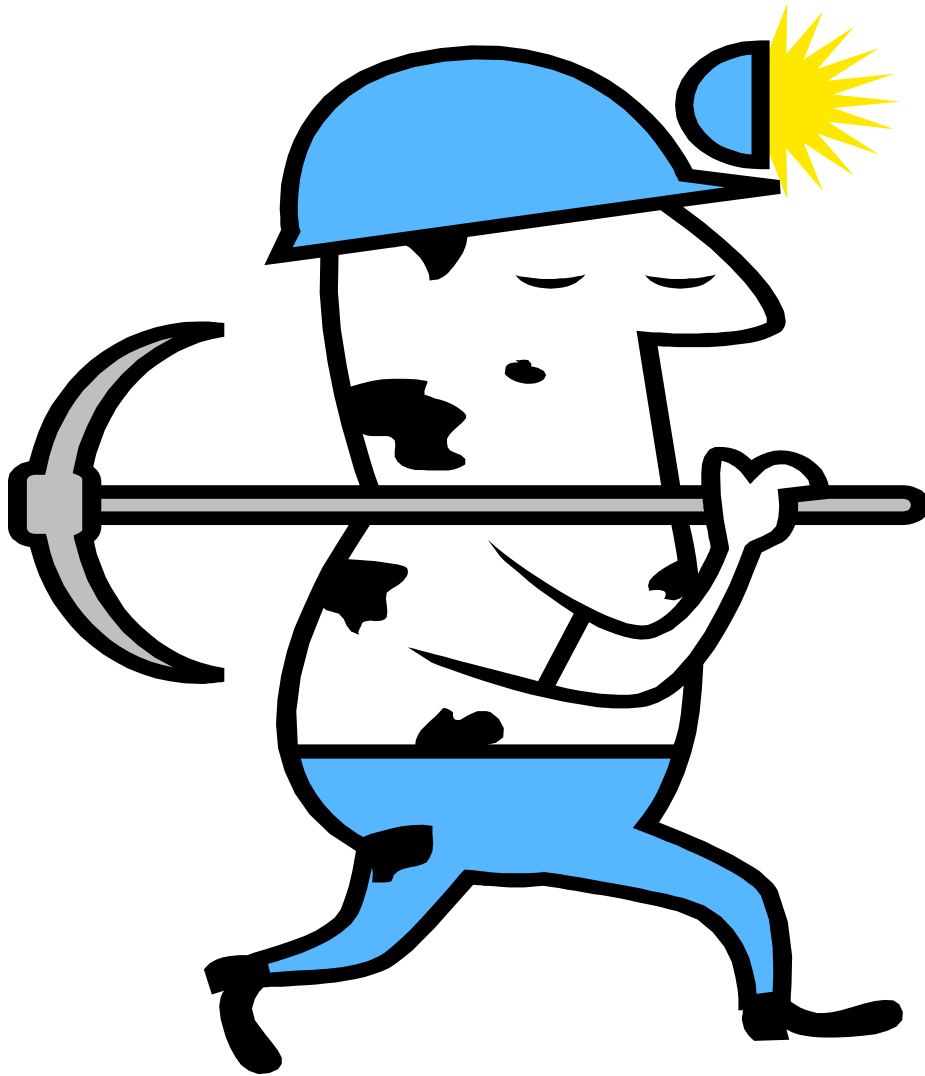


# A Fast Emerging Technology

Corey Toennies

April 19, 2004

CSI 490



## Table of Contents

1. Data Mining – What Is It?	1
2. History of Data Mining	1
3. Overview of Data Mining	2
4. Who Uses Data Mining	3
5. Using Data Mining to Capture Terrorists?	4
6. Succession with Data Mining	4
7. Relationships Sought	7
8. Data Mining vs. OLAP	8
9. Different Types of Techniques	9
9.1. Neural Networks	10
Figure 1. (Neural Network)	10
9.2. Decision Trees	12
Figure 2. (Decision Tree)	12
9.3. Genetic Algorithms	13
9.4. K-Nearest Neighbor	14
10. How Data Mining Works	15
10.1. Define Business Problem	16
10.2. Build a Data Mining Database	16
10.2.1. Data Collection and Description	17
10.2.2. Selection	18
10.2.3. Quality Assessment and Cleansing	18
10.2.4. Integration and Consolidation	19
10.2.5. Metadata Construction	19
10.2.6. Load the Data Mining Database	19
10.2.7. Maintain the Data Mining Database	20
10.3. Explore the Data	20
10.4. Prepare Data for Modeling	20
10.5. Data Mining Model Building	21
10.6. Evaluate the Results	22
10.7. Deploy the Model and Results	23
11. Mining the U.S. Voters	24

Figure 3. (Privacy Invasion?)	25
12. Disadvantages with Data Mining	25
12.1. Consumers	27
12.2. Organizations	28
12.3. Government	29
12.3.1. Congress Involvement in Data Mining Projects	30
12.4. Analysis of Ethical Issues	31
13. Data Mining Software	32
14. Employment in Data Mining	33
15. Conclusion	34
Works Cited	35

## **1. Data Mining – What Is It?**

With data in company data warehouses reaching into the terabytes, it is becoming significantly harder to manually analyze this data and get useful output. Data mining is the extraction of hidden predictive information from large databases. Data mining can uncover correlations that were previously invisible. Data mining has led to the creation of massive databases that can track what's selling and where, as well as who's buying it and what they will buy next. Companies are now utilizing data mining techniques to examine their databases looking for trends, relationships, and outcomes to enhance their overall operations and discover new patterns that will allow them to better serve their customers. This process helps companies focus on the most important information in their data warehouses.

Data mining tools can also predict future trends and behaviors allowing businesses to make proactive, knowledge-driven decisions. They scour databases for hidden patterns, finding predictive information that experts who manually search the databases may miss. Data mining provides numerous benefits to businesses, government, society, as well as individual persons.

## **2. History of Data Mining**

The start of data mining was in 1992 when data mining pioneer Thomas Blischok did a study of sales for an American drugstore. Blischok found a correlation between sales of beer and diapers between the hours of 5:00 and 7:00 p.m. He found that when men stopped to pick up diapers, they also tended to buy beer. On Thursdays and Saturdays, he found this trend tended to escalate regarding the beer and diaper sales. This finding is known to be the humble beginning of data mining. (Pigg)

### 3. Overview of Data Mining

Although data mining is relatively a new term, the technology is not. With technology advancing rapidly and computational power increasing, data mining is emerging in today's businesses. Human analysts with no special tools can no longer make sense of the huge volumes of data that require processing in order to make informed business decisions. Data mining is largely supported by three technologies:

- Massive data collection(into data warehouses)
- Powerful multiprocessor computers
- Data mining algorithms

Data mining uses elements of statistics, artificial intelligence, machine learning, and advance modeling techniques to predict future business trends and customer behavior patterns from large data warehouses and other forms of data resources. This is accomplished by running software applications to convert vast amounts of data into actionable, proactive, and knowledge-driven decisions.

There are 2 major capabilities of data mining that generate new business opportunities. The first one is automated prediction of trends and behaviors. Data mining automates the process of finding predictive information in large databases. Questions can now be answered directly from the data, which previously required extensive hands-on analysis. Don't be misled, it is still important to understand the data , even though it is an automated process. An example of this capability is using data on past promotional mailings to identify targets most likely to maximize return on investment in future mailings. Another example is forecasting bankruptcy.

The other capability of data mining is the automated discovery of previously unknown patterns. Data mining tools are able to sweep through databases and identify

previously hidden patterns. Examples of pattern discovery are the analysis of retail sales data to identify seemingly unrelated products that are often purchased together, and detecting fraudulent credit card transactions. (Manley)

Data mining can help answer a variety of question such as:

- What goods should be promoted to a customer?
- What is the probability that a certain customer will respond to a planned promotion?
- Will the customer default on a loan or pay back on schedule?
- What medical diagnosis should be assigned to a certain patient?
- Why a facility suddenly starts to produce defective goods?

#### **4. Who Uses Data Mining**

Data mining is used today by companies with a strong consumer focus, primarily companies in retail, financial, communication, and marketing. These companies determine relationships among internal factors such as price, product positioning, or staff skills, and external factors such as economic indicators, competition, and customer demographics. These factors help determine impact on sales, customer satisfaction, and corporate profits.

Data mining projects need to be managed by intelligent and qualified employees. These individuals need to understand the business, the data, and the nature of the analytical methods involved. Data mining involves scientists, statisticians, as well as those specialized with machine learning and pattern recognition. (Chhay) These individuals need to understand the algorithms being used, the hardware and software running the projects, and the results generated. Data mining isn't a simple task to run, organizations need committed employees who have knowledge and understand the various aspects on data mining.

## **5. Using Data Mining to Capture Terrorists?**

A few days after the September 11 attacks, FBI agents visited one of the largest providers of consumer data. They did this to see if any of the terrorists were in the database, and discovered five of them. One of the terrorists had been in the country for less than two years with 30 credit cards and a quarter of a million dollars in debt. Mohammed Atta, the ringleader of the operation, had also been here less than two years and had twelve addresses under different names. (Edelstein, Data Mining in Depth: Using Data Mining to Find Terrorists)

Using data mining to capture terrorists before they commit crimes isn't that easy though. Searching databases for people who have a number of credit cards, large debts, and multiple addresses would yield both criminals and perfectly innocent citizens. Finding terrorists with multiple addresses would be nearly impossible to do. As Atta did, all a terrorist would have to do to get around this search would be to use different aliases. Currently we don't have enough known terrorists or a consistent set of behaviors to use data mining to pin down a terrorist before he/she commits an evil act.

## **6. Succession with Data Mining**

Wal-Mart is pioneering the use of data mining to transform its supplier relationships. They capture point-of-sale transactions from over 2900 stores in 6 countries and continuously transmit this data to its massive 7.5 terabyte data warehouse. Wal-Mart allows more than 3500 suppliers to use this data to identify customer buying patterns at the store display level. Suppliers are able to see how each of the 70,000 products in the stores is selling. Wal-Mart uses this information to manage local store inventory and identify new merchandising opportunities. The information has the

capability to show customer buying habits such as product associations. Stores in colder climates may want to take notice what customers buy when a parka is purchased. With newly discovered product associations, stores can re-arrange displays so complementary products are located next to each other. Also, they are able to customize each store's offerings based on what products have a high turnover and which ones don't. (Data Mining: What is Data Mining)

Anheuser-Busch, the world's largest beer distributor, is finding enormous success with data mining. In 1997, chairman August Busch III vowed to make his company a leader in mining its customer's buying habits. It relies on data from customer purchases to better market its products. Sales representatives working for the company go from store to store in their region to track sales numbers. Each sales rep enters in detail about sales, shelf stocks, and shelf displays into a handheld PC. Not only do reps enter data about Anheuser products, but also data about competitors' product displays. By simply browsing through the aisles reps are able to see how competitors display their merchandise and what kind of promotions are being offered. Although Anheuser's competitors may disagree, tracking the competitors product displays isn't as big of an ethical issue as other issues that arise from data mining.

For each store, the handheld shows an inventory screen with a four week history and displays numbers on how much sales they did and where they had displays. Once the data is all entered into the handheld, sales reps plug the handheld into their cell phones and send off new orders to their warehouse along with the data they gathered.

Anheuser-Busch is mastering the science of finding out what beer lovers are buying, as well as where, when, and why. For each beer purchase Anheuser servers'

record price, date brewed, purchased warm or chilled, and whether you could have gotten a better deal somewhere else. Anheuser then mines this data to constantly change marketing strategies, designing promotions to suit the ethnic makeup of a market, and as early detection of where competitors may have an edge. Data from stores and wholesalers is becoming the backbone of the corporation.

Anheuser collects data from its distributor's servers every night about what brands are selling in which packages using which medley of displays, discounts, or promotions. Anheuser then mines this data to send information back to its distributors on how to increase sales. Anheuser's data mining savvy is far more advanced than any of the other breweries throughout the world. Anheuser's data mining techniques has helped their share of the \$74.4 billion U.S. beer market to reach record highs of 50.1%. They have a better understanding of what makes their product sell, thanks to data mining. By concentrating on customer buying habits, Anheuser can determine when their product is sold (weekend or weekday) and why (holiday or weather). Data mining uncovers these correlations, making it easier for management to determine what product(s) to deliver to each store and how much.

The results Anheuser receives from data mining are astonishing. Anheuser knows what images or ideas to push in its ads and what new products to unveil, such as the low-carb Michelob Ultra. They are also able to tailor marketing campaigns with precision to all of its different markets, being city, state, or country. In addition to mining the entire network of data they receive, Anheuser can segment the data into the different markets and then run data mining applications. This shows customized and unique reports for

each market, thus giving Anheuser the distinct advantage it has over its competitors.

Data mining revealed trends for Anheuser by:

- City - Bud Light is hot in Peoria, while Tequila fares better in San Antonio
- Holiday – 4<sup>th</sup> of July is a big seller in Atlanta, while St. Patrick's Day isn't
- Class – Cans for blue-collar stores, while bottles for white-collar

Since Anheuser began using data mining, it has posted double-digit profit gains for 20 straight quarters. This shows just how beneficial data mining is to Anheuser and that it's competitors may want to start following their lead. (Kelleher)

## **7. Relationships Sought**

Data mining looks to find four different types of relationships:

- Classes
- Clusters
- Associations
- Sequential patterns

Classes are the most common form of data mining and consist of shared characteristics. Stored data is used to locate data in predetermined groups. A data mining tool uses pattern recognition to create classes. Classes could be as simple as high medium or low. Once the desired classes are set and its unique parameters, data mining stores the results under the appropriate title based on the classes properties. For example, a restaurant chain could mine customer purchases to determine what customers order and when. They could then use this discovered information to have daily specials.

Clusters are a subset of classes that consist of patterns and relationships that have not been predefined or were not previously have known to existed. Data mining finds these relationships even though the user was not previously looking for them. Data is generally grouped into clusters based on logical relationships. Data mining can discover

several inputs that all have a similar attribute and then output a cluster to the user to show the previously hidden relationships

Associations deal with events. This means that the occurrence of one event leads to the occurrence of another event. Knowing associations can boost profits for corporations, because with the newly acquired information a sale of one item could also lead the sale of its complement. An example of an association is when somebody buys a light fixture, 75% of the time they buy light bulbs. Sequential patterns deal with events also, but are linked over time. In essence they are associations linked over a time period. For example, when credit card holders request a higher spending limit, they tend to buy a large item within the next two weeks. This information is valuable because if a store sells an abundance of printers it may want to stock up on replacement toners in the near future. (Data Mining: What is Data Mining?)

## **8. Data Mining vs. OLAP**

One of the most common questions from data processing professionals is about the difference between data mining and On-Line Analytical Processing (OLAP). They are different tools that can also complement each other. OLAP is used to discover why certain things are true within a database. A user will make a hypothesis about a relationship and then try to verify it by running queries against the data. OLAP works great if you only have a couple of variables to analyze against a database, such as individuals with low income and high debt are bad credit risks. When you use dozens or even hundreds of variables, it becomes much harder and more time consuming to find a good hypothesis and analyze the database with OLAP to verify or disprove your hypothesis. The more variables you use the harder it is to analyze each one and come up

with a universal hypothesis. It would be difficult to look at each variable and see how it relates to the dozens of other variables, whereas with a few variables relationships are more easily seen, therefore making it easier to depict a hypothesis.

Data mining is different from OLAP because rather than verify hypothetical patterns, it uses the data itself within the database to uncover patterns. With OLAP you have to know beforehand which variables may solve your business question, because you feed the database these variables, but data mining finds the variables for you. The number of variables involved doesn't affect data mining like it does OLAP. Data mining finds the variables for you, while OLAP involves testing databases with variables to find the answer to your problem. OLAP involves you guessing and testing which variables lead to bad credit risks, while data mining finds the variables for you.

OLAP is sometimes used with data mining because it is an easy process to explore your data. The better you understand your data the more effective the data mining process will be. It is important to remember that data mining finds the patterns for you, while OLAP involves speculating what the patterns are and then query a database to test your assumptions. (Introduction to Data Mining and Knowledge Discovery)

## **9. Different Types of Techniques**

There are four common techniques used in data mining, according to Two Crows' tutorial data mining publication.

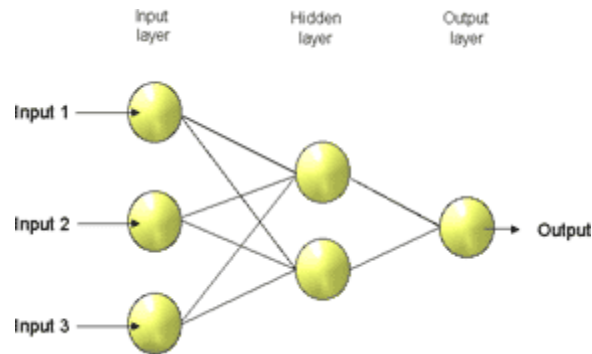
- Artificial neural networks
- Decision trees
- Genetic algorithms
- K-Nearest neighbor method

These techniques used in data mining are also referred to as modeling. Modeling is the act of building a model based off a situation that you know the answer and then applying it to a situation where you don't know the answer. Data mining software runs through the data and gathers its characteristics to put into the model. Once the model is built it can be ran in similar situations where the answer is unknown.

## 9.1. Neural Networks

Neural networks are non-linear predictive models that learn through training. This behavior model is built through learning, like a student learning Spanish. Here is a pictorial example of a very simple neural network.

Figure 1: Neural Network



Neural networks comprise of input, hidden, and output layers. Each node in the input layer represents a predictor variable. Each input node is connected to every node in the hidden layer. The majority of work in a neural network occurs in the hidden layers, where information is processed based on input variables received. After the information has been processed the output is sent either to another hidden layer for additional processing or sent to the output layer, where the information is then analyzed by a professional.

The McCulloch-Pitts processing element is an algorithm used in neural networks. The algorithm is:

$$f(\text{net}) = f\left(\sum_{i=1}^D w_i x_i + b\right)$$

In the algorithm “D” is the number of inputs in the neural network. Using Figure 1 as an example “D” would be three. The variable “ $x_i$ ” are the inputs to the hidden layers, whether it be green, 7, or \$15,000. The symbol “ $\sum$ ” stands for the summation. The variable “ $w_i$ ” is the weights of the input variables to the hidden layers. The weights are represented on the connecting lines between the input layer and the hidden layer. The weights are an important variable because they represent how strong or influential an input is. If an input is a big influence on the outcome of the neural network, the weight would be adjusted accordingly. The variable “b” represents any biased term that you may use in the neural network. It could also represent any outliers that you may be using. The entire algorithm, represented as  $f(\text{net})$  would be 1 for a net greater or equal to 0 and -1 for a net less than zero. (Euliano 102)

With neural networks, we are able to do a bank loan risk predictor with the obtainment of borrower’s personal yet descriptive data. The first job is to gather a few thousand past borrower files, including ones who paid back loans on time and those who did not. These files include the borrower’s description: which comprises of family makeup, income, nature of employment, etc... The files also contain amount of loan, duration of loan, and its outcome (good or bad).

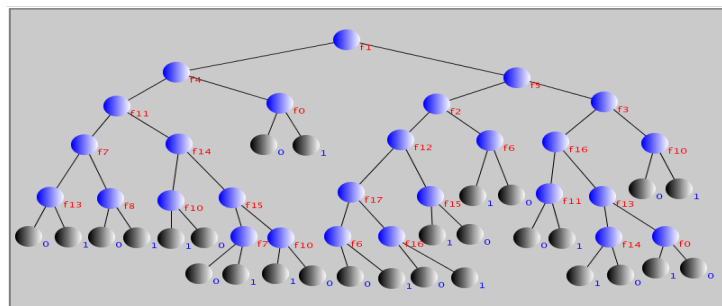
The data is then sent to the hidden layers of the neural network where the learning process begins. It is here, in the hidden layers, that the network will process the complex relationships between the borrower’s descriptive data and the loans outcome. The

learning process's result is a statistical model. This model now is able to estimate the outcome of a loan when given borrower's characteristics before the loan is ever granted. Lending institutions will now be able to calculate their risk of a loan being paid back as long as they have the potential borrower's personal information. (Neural Networks: Advanced tutorial)

## 9.2. Decision Trees

Decision trees are a way of representing a series of rules that require decisions made upon them that lead to a class or a value. A decision tree is comprised of the decision node, branches, and leaves. The first component of the decision is tree is called the top decision node, or root node, which specifies a test to be carried out. The results of this test cause the tree to split up into branches, each representing an answer from the decision node. Depending on the algorithm, each node may have two or more branches. Nodes that have exactly two branches are part of binary tree. Multiway trees have more than two branches coming of any given node. Each branch will either lead either to another decision node or to the bottom of the tree, called the leaf node. In Figure 2, decision nodes are coded in blue and leaf nodes are coded in gray.

Figure 2: Decision Tree



When a decision node is reached, the data from the case which you are mining chooses the appropriate branch to take. Examples of decision nodes would be: Is income

above \$50,000?, did customer respond to promotional mailing?, or is person in debt?

Decision trees are grown using iterative splitting of data into discrete groups, where the goal is to maximize the distance between each group. (Introduction to Data Mining and Knowledge Discovery)

One downfall of decision trees is that you can't look ahead when making decisions at a decision node. If you are at a decision node and realize that you could possibly branch off either way, you can't look at lower decision nodes to see which ones are more suitable to your case. This is a downfall because depending on which way you branch off, your solution will be vastly affected. An instance of this problem is if a decision node asks "Income above \$50,000?", and the income is \$49,995. Since the income difference is only five dollars, you could possibly split off either way, but each way leads to a totally different solution. Decision trees require that you make a decision at each node without really knowing where you will go next, which may lead to less accurate solutions.

### **9.3. Genetic Algorithms**

Genetic algorithms aren't used to find patterns as the other models do, they are used to guide the learning process of data mining algorithms. These genetic algorithms are primarily used with neural nets. Genetic algorithms are similar to biological evolution in which the members of one generation of models compete to pass on their characteristics to the next generation of models until the best model is found. The information passed between models is stored in chromosomes which also contain the parameters for building the model. These chromosomes in genetic algorithms show how to build a better model.

With neural nets, genetic algorithms can be used to help adjust the weights. In neural nets, we are always trying to adjust the weights to give a more accurate solution and genetic algorithms can be used to show how the weights need to be adjusted( such as higher or lower). Genetic algorithms also might be used to find the best architecture. Here, the chromosomes would contain information on the number of hidden layers and the number of nodes in each layer. It is important to remember that genetic algorithms aren't a model by themselves, rather they assist in finding the best model to use with data mining.

#### **9.4. K-Nearest Neighbor**

When trying to solve new problems, many people often look to similar problems that have already been solved to find a solution for their new problem. K-nearest neighbor is a classification technique that uses this same method. It decides which class to assign your case to by examining a number, the “k” in k-nearest neighbor, of the most similar cases or neighbors.

This model has two main steps. The first step is to determine the closest neighbors to your test case. To determine the closest neighbors to your test case the “k” is looked at to determine which other cases to compare to your test case. Once you determine which cases you are comparing to, you can proceed with the next step which is determining which of the neighbors to classify your test case with. This step involves comparing attributes of your test case to the neighbors' attributes to determine where to classify your test case. Each of the neighbors, also known as classes, may be very different from each other leading to biased classifications. One class may have 20 samples in it, while another class may have 2 samples in it. To compensate for these

differences, complex algorithms scour each class looking at the samples of cases and also how far away the class is from the test case. Once the algorithms put your case with a neighbor, you can look at the other cases to see how they were solved to help you on your current test case.

## **10. How Data Mining Works**

Web sites can track your buying habits along with the prices that you pay. Visiting a site often and paying full price for items can be documented. Therefore, you probably will continue to pay full price for any items you buy. However if the site determines that you are a bargain shopper and you only buy their items at a discounted rate, it may start offering items at a discounted rate. Pricing will be determined by who you are and who the shopping site determines you are, meaning your previous history of buying patterns. What this means is that, when you go to your favorite shopping site and discover that an item you want costs \$17.00. Your friend may go to the same site and view the same exact item and discover that it costs him/her \$20.00. Data mining allows sites to look into your previous history, hoping to further entice you to buy their product according to your experiences. (Berger)

Data mining isn't some magic wand that sits in your data warehouse sending unknown patterns and correlations to you. It's a very complex process. Data mining assists business analysts with finding patterns and relationships within the data. Two Crows, a leading data mining firm, produced a data mining tutorial called Introduction to Data Mining and Knowledge Discovery. This tutorial states that there are seven steps to the data mining process for knowledge discovery. Knowledge Discovery is the discovery of understandable knowledge.

- Define business problem
- Build data mining database
- Explore data
- Prepare data for modeling
- Build model
- Evaluate model
- Deploy model and results

Over the next few pages I will go into detail explaining each of these stages and how they get accomplished.

### **10.1. Define Business Problem**

The first of the seven steps is to “Define the business problem”. The prerequisite to data mining is understanding your data and your business. Without this knowledge you will not be able to identify the problem’s you are trying to solve, prepare the data for mining, or correctly interpret your results. To make data mining work for you, you need to write a clear statement of your objectives. You need to know what you want to do, whether to increase response time from a mailing campaign or get more members to your organization. Also it is a good idea to know who you will use the output to accomplish your problem. Depending on your objectives, you will end up building a very unique model.

### **10.2. Build a Data Mining Database**

The reason you need to understand the data, is because the data to be mined needs to be put into a database. Therefore, it is important to know what data is related to your objectives. Depending on the amount and/or complexity of the data a spreadsheet or a flat file may be used. Generally to be successful, is not a good idea to use your company’s data warehouse for this. It is a good idea to build a separate data mart, where you get to pick and choose what to put in. If you build a separate data mart data miners

will understand the data better being used in the data mining process because they control what enters the mart. This data mart will also be customized, including only the data you need to solve your business problem.

Mining the data in your data warehouse will make you an active participant in your data mining project and help you to have a better understanding. A single trial model of a data mart may require several passes through your data warehouse. There may be several iterations of preparing the data mart as you learn something from the model suggesting you modify the data. In addition, you may want to add data from outside your organization, possibly through surveys. There are several tasks in building your data mining database.

### **10.2.1. Data Collection and Description**

The first task is the data collection and data description. Here you will identify the sources of the data you will be mining. This phase will be necessary if some of the data that you need has never been collected before. These reports list the properties of the different data sets and describe the contents of each file or database table. Some of the elements included in these reports are:

- Source of data
- Owner
- Cost
- Privacy requirements
- Data types
- Range of values
- Number of missing values

These reports help the knowledge process by giving you a very detailed understanding of where you are getting your data and what kind of data it is.

### **10.2.2. Selection**

The next step in building the data mining database is the selection process. Here is where you select the subset of data to mine. The selection process needs to be done by somebody who understands the business and what problem is trying to be solved. As a result this person will know what data needs to be used to come up with a solution to the problem. Before putting data into your data mart you need to eliminate irrelevant or unneeded data. Some reasons for the elimination of data would be cost, restrictions on the data use, or quality problems.

### **10.2.3. Quality Assessment and Cleansing**

The most important task in building your data mining database is the data quality assessment and data cleansing. If you want to build good models, it is important to have good data. This follows the principle garbage in, garbage out. A data quality assessment looks at characteristics of your data that will affect the quality of the model. Types of data quality problems include incorrect values in a field or that a value for a field doesn't exist. Incorrect values may arise when a person enters their state into the city field for a survey. Many incorrect values often occur by mistake if not concentrating when inputting information. Although incorrect values may be hard to detect it is important to find and fix them to help increase the accuracy to your data mining solution. A reason for missing values is if customers leave fields blank when filling out a survey. If this problem persists, you may want to pay special attention to it. Perhaps missing values is a story within itself. What if all poor people leave their income blank? This may require that you build a separate variable to identify the missing values. Another approach to dealing with missing values is calculating a substitute value. If a database consists of

60% males and 40% females, you may want to fill in missing gender values with male 60% of the time and female 40% of the time. (Introduction to Data Mining and Knowledge Discovery)

#### **10.2.4. Integration and Consolidation**

The data you need for your data mart may be located in several different places, such as different databases or other data warehouses and data marts. Data integration and consolidation involves combining data from different sources into a single mining database and fixing differences in data values from different sources. There are often large differences in the way data is defined and used in different databases. A common problem is using data from different countries where they have different monetary systems. If databases are used from the U.S. and Mexico, it will be necessary to convert the dollars and pesos into one or the other so that they can be used together.

#### **10.2.5. Metadata Construction**

The information in the reports of the data collection and description is the basis for the metadata infrastructure. These reports will provide information that will be used when the physical databases or data marts are created. This information will also be used by the analysts to understand the data and build the models.

#### **10.2.6. Load the Data Mining Database**

As previously discussed, it is a good idea to load the data to be used in its own database or data mart. Depending on the size of your data, you may need a Database Management System as opposed to a flat file. Having already collected, cleansed and integrated the data it is now time to actually load it into a database, such as Microsoft

Access. Depending on the complexity of the data and database design, this step may require a DBMS professional.

### **10.2.7 Maintain the Data Mining Database**

Once you have created your database to use in your model, it is necessary to monitor your database for accuracy. Another important step is to backup your database, so hours of hard work aren't lost on a hardware malfunction. Also to improve performance or reclaim disk storage, your database may periodically require re-organization. This could include getting rid of data no longer in use by the data mining process.

### **10.3. Explore the Data**

In order to build a successful model, you need to know and understand your data. Gathering a variety of numerical summaries and looking at the description of the data is a good way to do analysis. Building graphs and visualization tools are good examples to use because of how effective they can be in analyzing your data. Graphs and visualization tools are effective because key patterns, relationships, exceptional values, and missing values often stand out. With these summaries and descriptions you want to identify the most important fields in predicting an outcome and determine which derived values may be useful for your approach. If you have a large data mart, then exploring this data can be very time consuming. A good user-friendly interface and a powerful computer will extensively cut down on your exploration time.

### **10.4. Prepare Data for Modeling**

This is the final data preparation step before building models. First, you want to select which variables to use from your data set to build your model with. It would be

great if you could feed all your variables to your data mining tool and let it find the best variables and make predictions off of those, but it doesn't work this way. With each new variable you add, the time it takes to build a model is increased significantly.

Next, you may want to take a subset or sample of your data in which to build models with. Building a model with all of your data usually requires buying a bigger computer with more processing power. Generally, the results you receive from a model built off all your data or a model built with a properly selected random sample of your data will be the same. Building more models off a sample of your data will result in a more accurate and robust representation of your problem, than using one model built off all your data.

It is also necessary to transform variables in compliance with the algorithm you use for building your model. Each model may require you to represent your data a little bit differently. For example, one model may require that you represent income in a range such as High, Medium, and Low. On the other hand a different model may require that you represent income simply as an integer.

## **10.5. Data Mining Model Building**

When it is time to build a model you can't just pick one of them and go on, it isn't that easy. It is necessary to do exploration on the different models to decide which one will be most useful in solving your business problem. You may find that when searching for a model it is necessary to make modifications to your data. The process of building predictive models requires a well-defined training and validation practice in order to insure the most accurate models. This practice known as supervised learning, is training your model on a portion of the data and then testing and validating it on the remainder of

the data. Supervised learning involves setting a target output, and then comparing your actual output to your target and computing the difference.

Training and testing your data mining model requires the data to be split into two groups: one for model training and one for model testing. This process can be implemented on historical data of promotional mailings that you have done. The first thing to do is take part of the data and build your model to create a test database. Your test database is generally between 5% and 33% of your database.(Introduction to Data Mining and Knowledge Discovery) This test database is then compared to the other section of data that you didn't use for training. The reason for the comparisons is to see how accurate your model is and determine what kind of changes to be made to it, if any. A model is completely built when the cycle of training and testing are complete. The accuracy rate you generate is a good estimate on how your model will perform on future databases (future promotional mailings) that are similar to your training and test databases.

## **10.6. Evaluate the Results**

When evaluating your results you can use a couple of different approaches: accuracy, lift, or profit. Accuracy seems like it would be the best approach to use, but it often doesn't tell the whole story. It's nice to know that a model predicted 75% accuracy, but what about the other 25%. You are unable to determine what the flaws and errors were in the 25%, so future mistakes wouldn't happen. Lift measures the improvement achieved by a predictive model compared to not using a model. Lift is a change ratio that can show the increase or decrease of responses to a mailing that you get when using a model. It helps determine if using models is worth the time and effort that goes in to

them. For example, instead of getting a 5% response from a random 20% of a population, you now get a 25% response from 20% of a mined population. Depending on whether you choose to maximize lift, profit, or return on investment you will choose a different percentage of a population to send a promotional mailing to.

## **10.7. Deploy the Model and Results**

Once a data mining model has been built it can be used in a couple of different ways. One way is for an analyst to recommend actions based on his/her viewing of the model and its results. An analyst can look at the clusters the model has identified and rules that defined the model and determine what type of a market would increase profit for their company.

Another way is to apply the model to different data sets or databases. The model could be used to flag certain records depending on its attributes. Also the model could be used to assign a probability to an action, such as responding to a marketing campaign.

Often models are part of a business process of predicting risk analysis or fraud detection. In these cases the predictive model is incorporated into an application. A loan officer could incorporate a predictive model into a car loan application to assist them in evaluating the applicant. The model would communicate with the officer flagging them when an applicant's information matches previous applicants who didn't pay back the loan.

When utilizing a complex application, data mining is often a critical part of the final product. Knowledge discovered with a data mining tool can be combined with knowledge from experts to evaluate databases. In a fraud detection system, known

patterns of fraud may be combined with newly discovered patterns to assist fraud investigators detecting fraudulent claims.

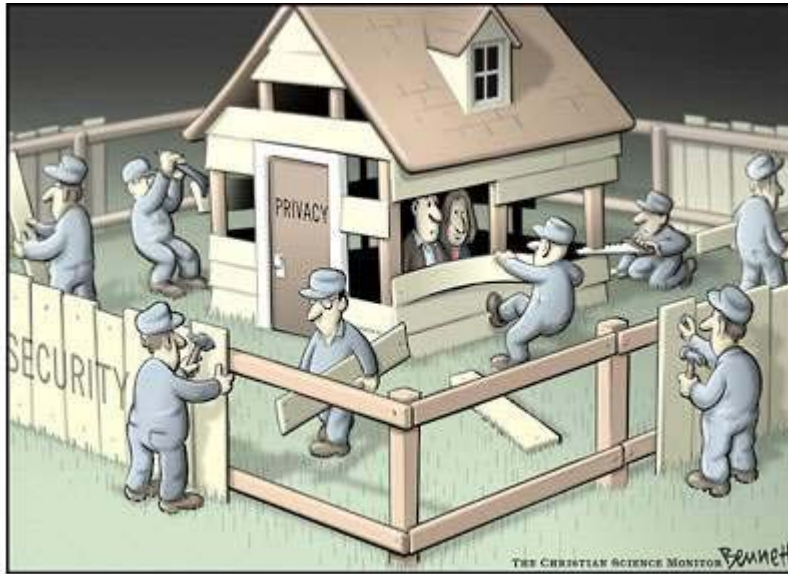
## **11. Mining the U.S. Voters**

With the current election year, data mining is becoming widely used by politicians. Politicians first stop is at a comprehensive database of U.S. voters. There are only a few of these databases, and the Democrats and Republicans each own one. Both parties comb through these databases gathering information on who the voters are and how to campaign towards them.

Each party's database has the name of every one of the 168 million or so registered voters in the country. Each voter record also contains their phone number, addresses, voting history, income range, and so on up to as much as several hundred points of data on each voter. The information for the databases was gathered from state voter registration rolls, census reports, consumer data mining companies, and direct marketing vendors.

Politicians use data mining to their advantage in three ways. First, by locating likely voters with a greater accuracy, campaign dollars can be spent more wisely and efficient. Second, data mining provides ways of discovering and turning out new voters. Finally, it allows politicians to cluster voters thus allowing them to send out individualized campaign messages to each cluster. With data mining, politicians can better understand what each voter wants, and then send out a message that only that voter wants to hear. Data mining has the potential to really boost campaigns, and increase victories for those who utilize it. (Gertner)

Figure 3: Privacy Invasion?



## 12. Disadvantages with Data Mining

Along with the advantages, data mining definitely has its disadvantages as well. Data mining has the capability to help companies become highly successful by giving them predictive information that managers can act upon. As you know, data mining works by scouring databases looking for previously unknown patterns in data. Well, the data in these databases can be personal information about individual people that they don't even know the companies have. Personal privacy is a major concern currently in this country. It is a concern that people's personal information can be obtained and then used in some way to harm them, such as identity theft. Therefore for data mining to be efficient a sacrifice may have to be made, your personal and private information.

You may wonder how companies get your personal information when you think that you have never come in contact with them. Companies can get your personal information a number of ways:

- Surveys

- Questionnaires
- Applications
- Previous transactions
- Buying it

Once you give your personal information to a company, you may think it is safe with them, but it's not. Companies have been known to sell the personal information they have about consumers to other organizations for a "nice" price. It is often the case, especially with credit card companies, that they will sell your information not worrying about the havoc it could cause on the victims.

You may have noticed before when filling in applications on the Internet that companies assure you that your information will be stored in a secure and protected database. Although companies have a lot of personal information about us available online, they do not have sufficient security systems in place to protect that information.

With insufficient security systems in place, the door opens for hackers to break into the systems and steal information. For example, the Ford Motor Company had to inform 13,000 consumers that their personal information, including Social Security number and payment history, had been accessed by hackers who broke into their database. (Chhay) This illustrates just how insufficient companies are at protecting customer's personal information. Companies are making profits from the customer's personal information, but they don't want to spend money to get a secure system to protect that same information.

Information obtained through data mining for marketing purposes is intended to be used in an ethical way. There's a chance that unethical businesses will use data mining in order to take advantage of vulnerable people or discriminate against a certain group. Since data mining can be used for a number of reasons, we need to look at

different groups that data mining will affect, whether beneficial or detrimental. These groups are:

- Consumers
- Organizations
- Government

## **12.1. Consumers**

Consumers can benefit from data mining by having organizations customize their product and service it to fit their individual needs. As a result the consumer's privacy may be lost in the deal. Data mining is a major way that companies can invade consumer's privacy. It maybe surprising how much organizations know about us, such as: birthday, home and work address, home and work telephone, number of children, medical conditions, or favorite types of entertainment. What organizations do with this information is what has consumers worried.

Data mining that allows organizations to identify its best customers, could also be used by deceitful organizations to attack vulnerable consumers such as the elderly, the unrefined, or the sick. The deceitful organizations would offer these consumers inferior deals knowing that they are more likely to fall into their trap. Suppose that John Doe does a survey and through data mining someone predicts that a loved one of his has a terminal disease. Maybe through this prediction, John Doe starts receiving letters from a fictitious company about a cure for this disease but it will cost him a lot of money. Since John Doe wants to save his loved one, he may decide to send money to this fictitious company for their cure.

Insurance companies can use data mining to discriminate against a certain market of people. They can use it to find out what types of people are more likely to develop

cancer. With this newly discovered information, insurance companies can decide not to sell insurance to these people or sell it at a much higher premium. Also, insurance companies could use data mining to find out what types of people are more prone to smoke and then not sell insurance to them. The predictive knowledge that data mining can discover is mind boggling, but how this knowledge is used is the problem!!!

The process of data mining isn't a perfect error-free process, mistakes can often occur. Consumers who have good credentials, such as credit history, can also be affected. An unfortunate incident would be if a bank rejects a customer with a good credit history, because their profile got mixed up with a customer who has the same name but a bad credit history. The customer with the good credit didn't do anything to deserve this, but will just have to deal with this unfortunate mistake. As a result this mistake from data mining could be a huge headache to all parties involved. Maybe this along with privacy issues is telling us that our nation is not quite ready for data mining to be implemented in today's market.

## **12.2. Organizations**

Data mining affects organizations differently than consumers. Data mining certainly benefits organizations more than it does consumers. If organizations better understood their market, they could focus their marketing and advertising to just their target market, thus reducing costs. With data mining, organizations can achieve this goal. It will help companies discover new patterns, enabling them to better serve their customers. It would also help these companies minimize their risk and increase profits. Since companies minimize their risk, they claim to pass on the benefits to everyone by offering lower premiums or interest rates but how do we know.

Data mining helps advertising companies to understand which customers are more likely to buy their product or service. The companies can find out if they need to gear their advertising more towards a different age, sex, or race. Then advertising can just be geared at the customers who are likely to purchase their product. With this new and improved target market, organizations can save substantial amounts of money by not sending out their advertisements to consumers who won't respond. Any organization that sends out mass advertisements to a general market would benefit from data mining. (Edelstein, Building Profitable Customer Relationships With Data Mining)

Data mining will benefit organizations, as shown above, but what if they can't utilize data mining tools. Since data mining requires the use of personal data, consumers may ask the government to step in and protect them. Consumers may ask for privacy rules which impose restrictions on data collection. With identity theft on the rise, consumers may find comfort knowing they are protected from anybody accessing their private information. Although the government may be asked to step in, they may not want to because of how data mining affects or benefits them.

### **12.3. Government**

The government finds themselves in a dilemma when dealing with the issue of data mining. On the one hand they want to protect the American people's right to privacy, but on the other they want to employ data mining because of its benefits. The government may be the only help consumers have in stopping organizations from using their personal information. Many people may not like the fact that organizations are benefiting from using your information without approval. In dealing with this issue, the

government may be asked to enforce laws stopping organizations from using personal data that they weren't given the permission to use.

Just like organizations, the government can benefit from data mining. The government is always looking to tighten up our security system. The government wants access to the public's personal data so that it can better protect them and protect this country from terrorists. With the help of data mining, the government may be able to determine which individuals are more likely to perform terrorist acts before the act is committed. The government needs to examine both sides of their data mining dilemma and decide if and how data mining can work in our country.

### **12.3.1. Congress Involvement in Data Mining Projects**

Congress has decided to cut a Pentagon office developing the terrorist tracking technology because of an outcry over privacy implications. Privacy advocates feared that if such powerful technology was created, government agents could use it on any database. This incident which took place in late February of 2004 shows just how involved the government currently is with data mining. The government is still financing research tools that can mine millions of public and private records for information about terrorists, but is cutting back.

Some of the data mining projects from Adm. John Poindexter's Total Information Awareness which haven't been cut were transferred to U.S. intelligence offices. "Poindexter's goal was to predict terrorist acts by looking for telltale patterns of activity in passport applications, visas, work permits, driver's licenses, car rentals, airline ticket purchases, and arrests, as well as credit transactions and education, medical and housing records."(Anti-terror record mining research continues) His projects, with the help of the

Advance Research and Development Activity included developing software that can extract information from databases as well as text, voices, other audio, video, graphs, images, maps, equations and chemical formulas. The reason Congress cut some of Poindexter's projects was because the research on transactions could put innocent Americans under suspicion for terrorist activities.

Although it is not known for sure, which of Poindexter's programs were transferred it is believed that some of the surviving programs are 18 data mining projects known as Evidence Extraction and Link Discovery. One of the rules that Congress is enforcing on the surviving projects is that research can only be used against non-U.S. citizens in this country, not against Americans on U.S. soil. Poindexter claims that his research will not only connect the dots that enable the U.S. to predict and pre-empt terrorist attacks it will decide which dots to connect. (Anti-terror record mining research continues)

#### **12.4. Analysis of Ethical Issues**

After seeing the different viewpoints, data mining arguably has much higher benefits for organizations and the government compared to consumers. So does this lead us to support data mining or reject it? If we choose to support data mining then this would be unfair to consumers because their right of privacy could be violated.

Organizations would be obtaining or even buying personal and private information just to better their company. If we choose to reject data mining, then it would be unfair to the businesses and government that plan on using it in ethical way, such as finding cures for diseases. Restricting data mining could affect data mining and businesses profits, thus causing them to increase prices to consumers.

### **13. Data Mining Software**

So a company decides that it wants to use data mining to help understand its data, thus helping to make more educated decisions. There are numerous data mining packages on the market today. Some software packages are tailored to a specific market, such as the marketers predicting response to a campaign (DMM and GainSmarts), and some are designed to use specific models such as neural networks (BrainMaker and NeuroSolutions) or decision trees (AC2 and Decisionhouse). (Software Suites for Data Mining and Knowledge Discovery)

Some software packages, such as DBMiner can pretty much do it all. DBMiner uses multiple models to help create market clusters, product associations, and much more. DBMiner helps companies find predictive information for its toughest questions. DBMiner's customers include Microsoft, Boeing, Hewlett Packard, and Target. Data mining software is implemented on top of databases, so it is important to get the software that is compatible with your type of PC. (DBMiner) Examples of other data mining software packages include GhostMiner, Teradata Warehouse Miner, and BioComp i-Suite. (Software Suites for Data Mining and Knowledge Discovery)

A key element to remember when buying mining software is the price. It may come as a surprise, as it did to me just how expensive the software is. Depending on how complex you want to get, the software can cost in the tens of thousands of dollars. Even though this price seems too expensive to fit into a company's budget, it has great potential to boost company profits, just as it has for Anheuser-Busch.

## 14. Employment in Data Mining

Included is an ad for a job opening involving data mining at @RISK Inc. located in Pennsylvania. (Jobs)

**From: Jan Crafton**

**Date: 5 Jan 2004**

**Subject: Berwyn, PA: Statistician/Advanced Modeling Specialist at @Risk**

@RISK Inc., ([www.atRisk.com](http://www.atRisk.com)) a rapidly growing corporation that specializes in predictive solutions and data mining within the field of Customer Relationship Management (CRM), is seeking a statistician/data modeler. Candidate must be proficient in both classical statistical modeling techniques as well as neural network and machine learning techniques.

We are searching for individuals that thrive in a fast paced organization that is focused on the quest for practical applications. Candidates must possess a strong background and some experience in an applied, business environment. A minimum of a master's degree in the field of engineering, statistics, mathematics or computer science as well as experience with modern machine learning procedures for classification analysis is required. Competitive salary and benefits package. Send resume in confidence to Jan Crafton at [jcrafton@atrisk.com](mailto:jcrafton@atrisk.com) or to:

Jan Crafton  
@RISK, Inc.  
1205 Westlakes Drive  
Suite 180  
Berwyn, PA 19312

As you can see applicants interested in data mining need to be sharp and well educated.

There are several other job openings posted on the KDnuggets website, which is a website devoted to data mining. Companies hiring data miners require extensive qualifications, including a mix of the following:

- Minimum of a Masters or a PhD in either Computer Science, Mathematics, Business, Physics, or a related field
- Work experience in either data modeling, data warehousing, machine learning
- Experienced with Java/C++ and SQL
- Excellent writing and oral communications skills
- Strong technical background

As you can see by the qualifications, the work in data mining isn't learned overnight.

Individuals need to be highly educated and have several years of work experience in related fields.

## **15. Conclusion**

Data mining is a powerful tool for extracting relationships and patterns from data that can provide answers to critical business questions. Many companies currently have enormous amounts of data containing valuable information on how to run their business. Data mining offers great promise in helping organizations uncover valuable patterns hidden in their databases. Data mining can provide a powerful competitive advantage to companies willing to provide the financial and technical resources as well as the willingness to learn their business problem and the data itself.

Data mining has major benefits for organizations and the government. However, the major flaw with data mining is privacy invasion to the consumer. Results from data mining can yield major benefits from increasing profits to reducing costs. Data mining has the potential to revamp organizations but at what cost to the individual?

## Works Cited

- “Anti-terror record mining research continues.” CNN.com. 23 Feb 2004.  
[www.cnn.com/2004/LAW/02/03/terror.privacy.ap/index.html](http://www.cnn.com/2004/LAW/02/03/terror.privacy.ap/index.html)
- Berger, Sandy. “Consumers Win, Lose With Online Data Mining Trend.” AARP. 16  
Jan. 2004 [www.aarp.org/computers-features/Articles/a2002-07-10-  
computers\\_features\\_datamining.html](http://www.aarp.org/computers-features/Articles/a2002-07-10-computers_features_datamining.html)
- Chhay, Heng. “Data Mining.” 21 Jan. 2004.  
[http://ceserv.engr.scu.edu/StudentWebPages/hchhay/hchhay\\_FinalPaper.htm](http://ceserv.engr.scu.edu/StudentWebPages/hchhay/hchhay_FinalPaper.htm)
- “Data Mining: What is Data Mining” 15 Jan. 2004  
[www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/datamining  
.htm](http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/datamining.htm)
- DBMiner 25, Feb. 2004. [www.dbminer.com](http://www.dbminer.com)
- Edelstein, Herb. “Building Profitable Customer Relationships With Data Mining.” Two  
Crows Corporation. 4 Jan. 2004. [www.twocrows.com/crm-dm.pdf](http://www.twocrows.com/crm-dm.pdf)
- Edelstein, Herb. “Data Mining in Depth: Using Data Mining to Find Terrorists.” DM  
Review. 16 Jan. 2004 [www.dmreview.com/master.dfm?NavID=55&EdID=6655](http://www.dmreview.com/master.dfm?NavID=55&EdID=6655)
- Euliano, Neil R. and W. Curt Lefebvre and Jose C. Principe Neural and Adaptive  
Systems. John Wiley & Sons, Inc. 2000.
- Gertner, Jon. “Mining the US Election Process.” New York Times. 25 Feb 2004.  
[www.kdnuggets.com/news/2004/n04/17i.html](http://www.kdnuggets.com/news/2004/n04/17i.html)
- “Introduction to Data Mining and Knowledge Discovery,” Third Edition Two Crows  
Corporation. 25 Jan. 2004 [www.twocrows.com](http://www.twocrows.com)
- “Jobs.” 26, Jan. 2004. [www.kdnuggets.com/nes/2004/n01/18i.html](http://www.kdnuggets.com/nes/2004/n01/18i.html)

Kelleher, Kevin. "66,207,896 bottles of beer on the wall." CNN.com. 25, Feb. 2004.

[www.cnn.com/2004/TECH/ptech/02/25/bus2.feat.beer.network/index.html](http://www.cnn.com/2004/TECH/ptech/02/25/bus2.feat.beer.network/index.html)

Manley, Denis. "Data Mining." Enterprise Systems. 29, Jan. 2004

[http://www.comp.dit.ie/dmanley/Enterprise%20Systems%20FT228\\_3/notes%202000\\_2001/datawarehouses%20and%20data%20mining/2](http://www.comp.dit.ie/dmanley/Enterprise%20Systems%20FT228_3/notes%202000_2001/datawarehouses%20and%20data%20mining/2)

"Neural networks: Advanced tutorial." PMSI. 5 Feb. 2004 [www.pmsi.fr/neurin2a.htm](http://www.pmsi.fr/neurin2a.htm)

Pigg, Susan. "Diapers, drinking and data" The Toronto Star. 3 Feb. 2004. [www.hope-tindall.com/peter/2002\\_aug\\_16.htm](http://www.hope-tindall.com/peter/2002_aug_16.htm)

"Software Suites for Data Mining and Knowledge Discovery." KDnuggets. 26, March 2004. [www.kdnuggets.com/software/suites.html](http://www.kdnuggets.com/software/suites.html)