

percent. The associations we discovered between the product lines may be of value in the overall design of the website and in the marketing strategies this website may want to take. For example, the analysis found that customers who purchased Item_00 had a stronger propensity to purchase Item_02 than Item_04 or Item_05 and Item_01. One marketing strategy might include notifying via e-mail all customers who have purchased Item_00 about a discounted offer for Item_02. Clustering and segmentation can also be used to do a similar type of analysis on other data components, such as the “key-words” in your log referrer file and those visitor sessions which resulted in actual sales.

Through this process you can select other clusters generated by the SOM network and extract rules to describe other segments which can provide conditions in the form of IF/THEN rules describing the relationships between website visitor attributes and the products you are selling. Through this clustering and rule-generation process you can select other products or dollar ranges to gain an insight about why certain items are purchased by some visitors while not by others. Still another form of clustering involves finding associations in time, which can be used to discover relationships between events. This type of clustering in time is sequencing, and can be used to uncover relationships in how a visitor moves through your website or if they purchase one item what is the probability they will buy another.

Figure 9-11 is a view of multiple product clusters that represent the html pages for various product lines. As you can see, some of the clusters overlap, while others do not. It is important to focus on those

Figure 9-11
A view of
the distances
between the
key clusters
uncovered.

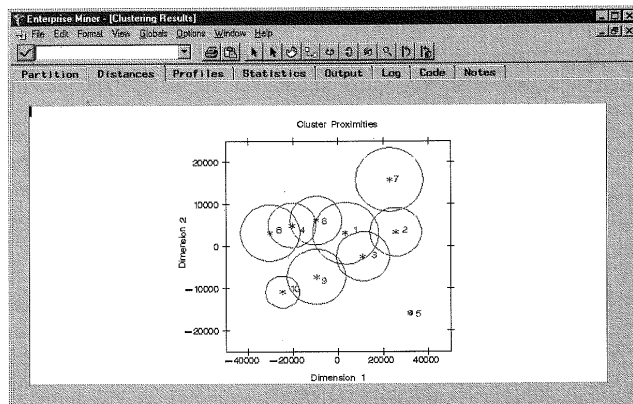
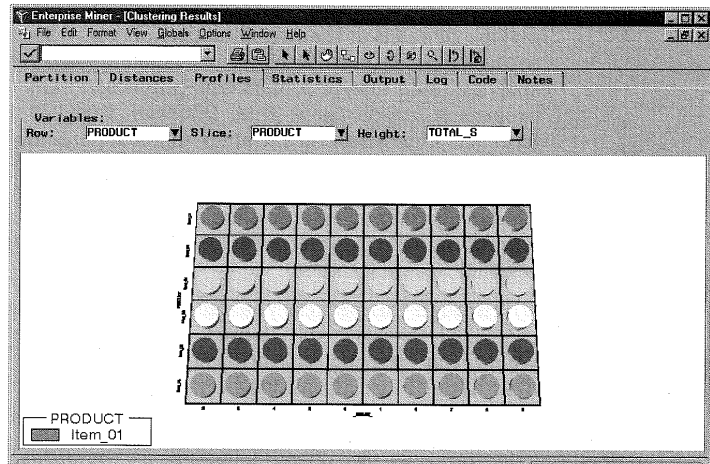


Figure 9-12
*Clustering by
 product Item_01
 and total num-
 ber of sales.*



that show some closed distances or overlapping, for they are prime candidates for cross-selling. For example, cluster 4 (Product A) overlaps with cluster 6 (Product M) and cluster 8 (Product Z). This may indicate a sequencing order for navigating through this website, but it may also point out an interest by those visitors who purchased Product A to also purchase Products M and Z. Figure 9-12 is another view of the same data using a different format, by color, with each row being a different product and the height represented by total sales.

Segmentation

As is the case of any database marketing program, website segmentation enables you to adjust your retailing strategies and goals to fit the reality of the situation. Through segmentation you can narrow your focus on distinct sub-populations of your website visitors. Instead of marketing products to all customers and visitors in the same manner, you want to customize your communications and narrow your focus on each subgroup. You also want to market them on the basis of your “relations” with them and focus on providing them the products and services they prefer.

You can segment your website customers into distinct classes, such as High, Average, or Low, and modify your marketing messages accordingly. For High value clients, you want to concentrate on

retaining them; as such, your marketing efforts will be to reward them. Conversely, with Low value clients, your strategy might be to educate them on your products and services, thus your marketing efforts will be geared toward informing them. In both instances, segmentation of your web customer data can drive the content of how your e-mail campaigns are geared to these separate and distinct customers: one to reward, the other to inform.

Through the mining of your website data and the discovery of the demographic features of your visitors your goal will most likely be to provide them with the type of information, product, and services they are most likely to want. You can also design your website to be more appealing to your visitors by knowing their lifestyles and the type of activities and products they prefer. For example, a website visited by consumers with kids should probably be designed differently from one visited by single males. One of the benefits of e-commerce is the flexibility to narrow your marketing messages based on lifestyle and preferences of your customers. Through the segmentation process you can begin to provide value-enhanced information and services based on the type of content your customers are most likely to be interested in.

To do this type of customer segmentation a data mining tool incorporating machine-learning and statistical algorithms, such as C5.0, ID3, CART, or CHAID, can be used to generate a graphical decision tree. A decision tree is a method of segmenting a data set on the basis of a desired output, such as the total number of purchases made by visitors. For example, a decision tree can be generated in order to explore the differences in the number of sales by such factors as your customers' income, which you can obtain from such household data providers as Polk or Experian. In the decision tree shown in Figure 9-13, an analysis of 10,000 online sales was conducted in order to discover what demographic and transactional factors impacted these online sales. The dependent variable, that is, the output we are segmenting the data on, is the total number of sales made at this website, which you can see varies between 0 (19.4%), 1 (39.4%), 2 (19.8%), 3-4 (20.6%), and 5+ (0.8%).

In this instance there were no significant differences in the total Number of Sales based on customer income levels. However, a common data mining technique is to sift through various customer attributes until you discover some meaningful "splits" in your data. This sifting process can be done with this type of data mining tool, providing you the flexibility to partition your data in various ways. You can split your data on the basis of the number of children a visitor

Figure 9-13
The propensity to purchase is very evenly distributed along income levels.

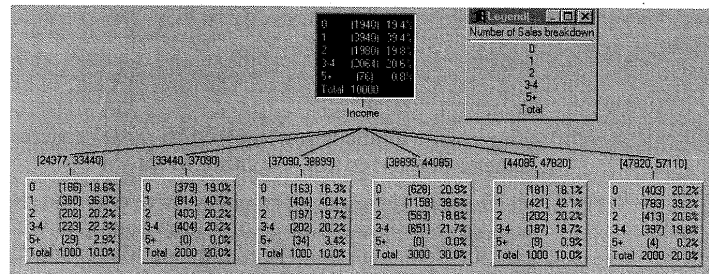
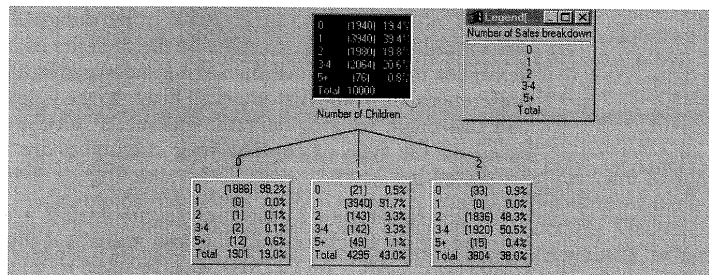


Figure 9-14
The number of children has a decided difference on the total number of sales.

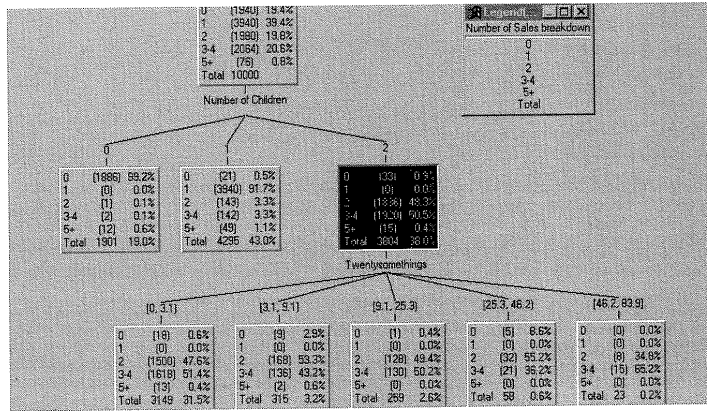


has, the search engine they used, the keywords they entered, their household income, or any other data attributes you have included in your data set. In Figure 9-14, for example, the decision tree revealed an interesting split: the number of children impacts the total number of sales.

Using this drill-down technique and this partitioning type of data mining tool, you can further focus on a particular profitable market segment in your website customer database. For example, in the decision tree in Figure 9-15, we drilled down on those individuals with two children in order to find that those residing in “Twentysomething” consumer class ZIP codes tended to make multiple purchases much more commonly than those of the general population.

Although the distributions to the “Twentysomethings” consumer class is very even there appears to be a higher number of sales of three or four items to visitors from ZIP codes where the demographic breakdown is between 46.2 and 83.9 percent for that consumer class.

Figure 9-15
The segmentation
of sales to
the consumer
class "Twenty-
somethings"



Using this type of segmentation tool we can easily explore and change customer attributes and see the impact and the resulting "splits" on the basis of other factors, such as the value of their cars, as shown in Figure 9-16.

The tree in Figure 9-16 can also be expressed in the form of IF/THEN rules, which can be easily generated in the following format:

RULE_1

IF Vehicle Value = High, Medium or Unknown

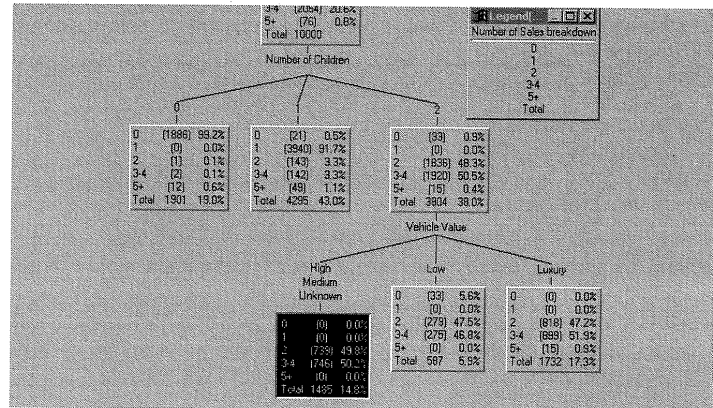
AND Number of Children = 2

THEN

Number of Sales = 0	0.0%	
Number of Sales = 1	0.0%	
Number of Sales = 2	49.8%	<= High Probability
Number of Sales = 3-4	50.2%	<= High Probability
Number of Sales = 5+	0.0%	

These rules, generated via decision tree analyses, can enable web designers, analysts, administrators, and marketing managers to understand and surmise in concise statements thousands of online transactions. It is important to note that multiple rules can be generated for

Figure 9-16
A decision tree
converging on
product sales
based on multi-
ple consumer
attributes.



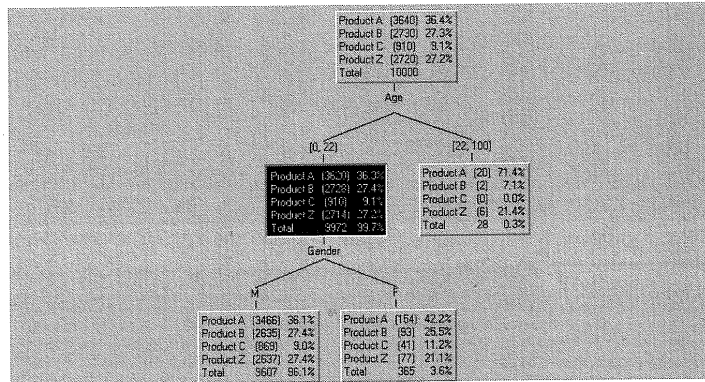
each individual page (HTM) in your website, such as the following, which predicts the expected dollar amount in total sales:

IF	Income	=	[24377, 33441)
AND	Search Engine	=	Infoseek
AND	Domain	=	NET
AND	Vehicle Value	=	Luxury
AND	Middle American	=	[0, 10)
AND	Home Value	=	[43023, 70808)
THEN	Total Dollar Sales: Average \$342.918		

Note that this type of rule is based on observed behavior consisting of past online transactions, which have been appended with demographic and psychographic information prior to being exposed to the machine-learning algorithm for segmentation. You can construct this type of segmentation analysis on the basis of product lines, specific sale items, or expected dollar sales. You can even do it on a visitor's propensity to click through on a banner, promo, ad, or any other expected online behavior at your website, as shown in Figure 9-17.

The segmentation of your website data can be done on the basis of age group, gender, income levels, etc. Most data mining tools give you the flexibility to partition your data on the basis of what you want to explore, or they can do their analysis on "automatic" mode with the core algorithm doing the statistical testing and partitioning on its own.

Figure 9-17
This tree is segmenting visitors under their propensities to purchase multiple products.



You should try to keep the decision trees you construct at a fairly high level, that is, avoid “bushy” trees with too many branches. The construction of overly complex trees with more than three or four levels can lead to a problem when you try to apply the results to new unseen data. A common problem is that of *overfitting*. This is a situation in which an overly bushy tree fits the data it is constructed with in such detail that it has a problem in generalizing when exposed to new cases or samples it has not seen before.

The core of the analysis can also be changed at any time. For example, rather than segmenting your data on the basis of projected “Number of Sales,” you can change your analysis to mine something like “Total Amount of Sales” so that a dollar amount can be projected for each of your website customers, rather than the total number of online sales. Figures 9-18 and 9-19 illustrate such a segmentation.

In a large e-commerce site with thousands of session records and hundreds of pages, a data mining analysis such as this can assist it in determining the customer and transactional factors impacting its online sales. For example, at what point does it differentiate between “sporadic” and “occasional” website browsers, or “high” and “premium” customers? Does it do it on the basis of two or three purchases over the last six months, or five and eight transactions during the last thirty days? In short, how does it quantify customer loyalty, value, and profitability? Techniques and tools such as these are designed to autonomously segment a data set based on statistically significant intervals and attributes. These types of analysis can “interrogate” the data from a website and uncover market segments in an unbiased

Figure 9-18
Income levels of customers can impact the amount of dollars they spend at your website.

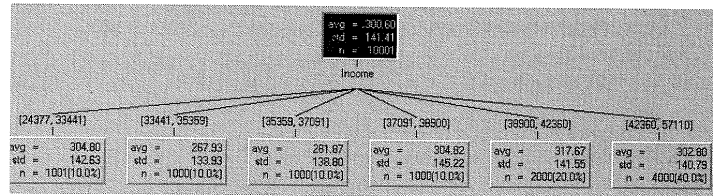
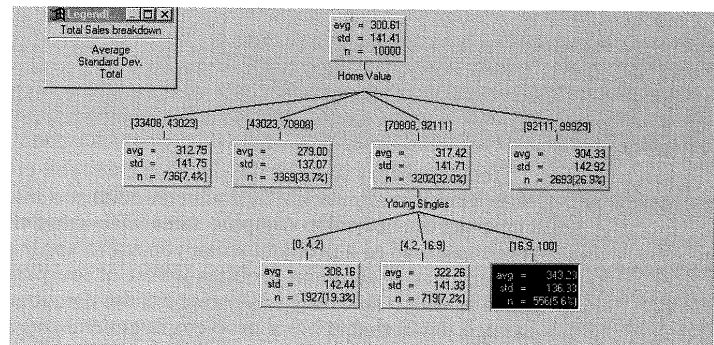


Figure 9-19
A market segment is identified with a higher than average Total Sales amount.



manner and in the process answer some of the key business questions for the retailer about its visitors and customers.

Data mining technology can identify the factors and number of transactions that differentiate browsers from buyers. As demonstrated, this type of data mining analysis can segment your website customers into various classes, from which you can further calibrate their value and loyalty—enabling you to formulate unique ads and marketing strategies. Segmentation can generate rules that are especially effective not only in discovering patterns in your data but also in describing their findings in the form of concise statements that web, marketing, and business professionals can easily understand and use. The bottom line is that the data can be segmented and viewed via any attribute you desire. For example, in the tree in Figure 9-20 we wanted to view the relationship between Total Sales on the basis of the search engine used by the visitor, so we “forced” the split to get the view shown.

One of the advantages of segmentation is that it can be used to discover the analogous features of your website customers. After all, one of the most sought after “ores” of data mining are customer composites, which in these cases happen to be your profitable and loyal web-

Figure 9-20
Infoseekers tend to spend more at this website than visitors from other search engines.

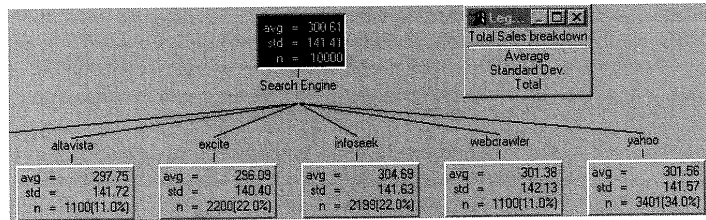
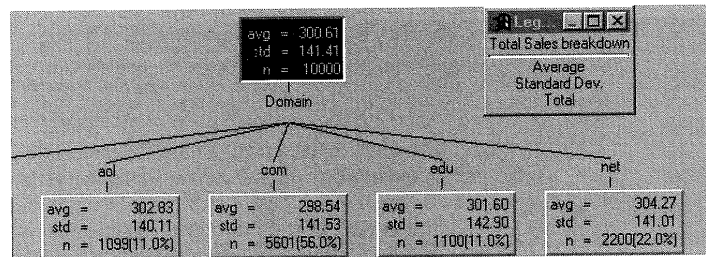


Figure 9-21
Visitors with NET domains tend to average higher sales at this website.



site visitors. Using this type of analysis a key question in your mind should be: who are my most profitable visitors and what are their features? These features may be in the form of their domain, keywords, age, income level, or other demographic attribute. Finally, segmentation of your website customers can be seen as your analysis of their lifetime value (LTV): i.e., the duration of their relationship with your website and the amount of money they have spent during that time (Figure 9-21).

Classification and Prediction

Another objective you may have for data mining your website is to be able to model the behavior of your visitors and assign a predictive score to them. This is commonly accomplished via data mining the patterns of priorly observed customer behavior. For example, credit card fraud detection neural network systems look at historical patterns of fraudulent schemes in order to detect new ones among thousands and even millions of daily transactions. Prediction via neural network modeling is commonly done in the credit card business sector as well as in banking and the insurance industry via such financial service applications as risk-analysis, portfolio management, and loan

approval systems. They all use some form of neural network technology in an attempt to maximize their profits and minimize their costs. A similar situation arises with your website, in which you can model past online customer behavior in order to predict or anticipate how a new visitor is likely to behave.

For example, you may like to know whether a new visitor to your website is going to make a high number of purchases or spend a large amount of money. Using a neural network tool we can begin to train a model to recognize and predict the expected purchases for each website customer. To construct a predictive model you need to have a sample of customer sessions of both profitable and unprofitable customers. In this instance a neural network is being trained to recognize the features and actions of profitable customers, what they look like, where they come from, how they search, how long they stay, and more importantly, what they are likely to purchase.

As with a decision tree tool, a neural network looks at all the available attributes in your data in order to converge on a few important clues that communicate how visitors are likely to behave and respond to your marketing efforts and/or products or services. For classification we use a data mining tool incorporating a neural network to construct a model, which begins with a zero accuracy rate, in an attempt to predict the expected amount of purchases visitors are likely to make at a website (Figure 9-22).

At first a neural network cannot distinguish between customers who made purchases and those who did not and thus it cannot predict Total Sales. However, it gradually learns to distinguish between the examples as it recycles positive and negative samples of shoppers. Soon it begins to learn the attributes and behavior of each, and improves its accuracy in predicting Total Sales in just a few minutes of training time (Figure 9-23).

Eventually this neural network tool achieves an accuracy rate of 80 percent in predicting the total number of purchases a visitor is likely to make at this website. This particular tool, like almost all other data mining packages, allows the user to easily “split” their data into training and testing sets. The training data set is used to construct the neural network model, while the testing data set is used to evaluate the accuracy of the model. Once the user is satisfied with the accuracy of the model, C code can be generated from most of these tools which can be used to “score” new visitors to a website so that predictions can be made about their overall propensity to make purchases and their potential profitability.

Figure 9-22
A neural network begins with a zero predicted accuracy rate.

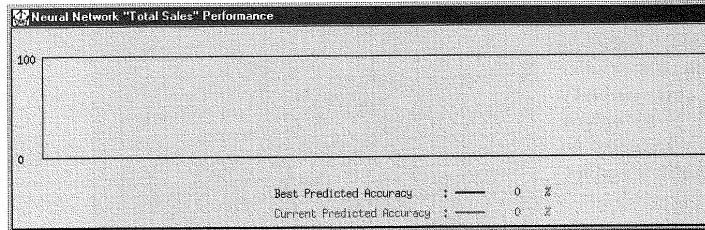
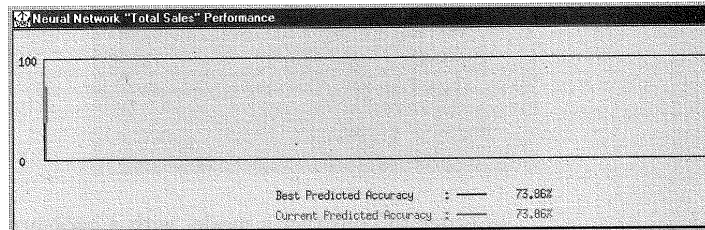


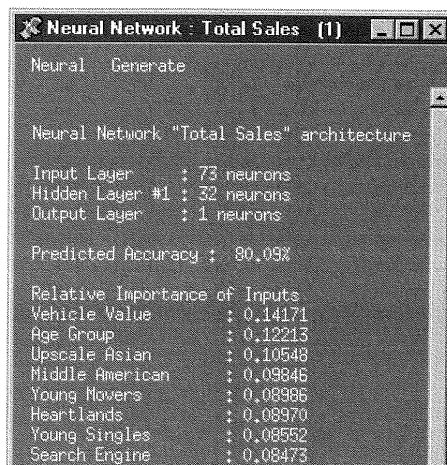
Figure 9-23
A training session generally takes a few minutes to learn.



Once a model has been constructed to predict the potential amount of purchases a visitor is likely to make, the data mining tool can be used to generate a sensitivity analysis report. This report describes the predicted accuracy of the test data set, which the model had not seen during its construction, so we know how accurate we can expect its predictions to be on unseen new visitor transactions. The sensitivity analysis report also prioritizes the customer attributes most important in predicting a visitor's potential amount of purchases. For example, Figure 9-24 shows that for this model the three top factors were the website visitor's vehicle value, their age group, and their membership of the "Upscale Asian" consumer class.

A neural network can be used to predict other factors, such as which ads and banners visitors are likely to click through or how many purchases a visitor is likely to make. One way an electronic retailer can use a predictive model is to score all new visitors on their propensity to purchase its products or services. Having daily scored all its new visitors it can next prioritize by scores these "potential buyers" and make specific offers via e-mail, starting of course with those with the highest probability of becoming loyal customers based on their characteristics.

Figure 9-24
The total amount of sales a visitor is likely to make is impacted by multiple factors.



In order to construct a model with a back-propagation neural network you will need to “train” it with as many negative and positive samples as you can get. It is interesting to note that the factors affecting the outcome you are mining can vary in terms of the input variables values and their range. Ensure that your training data set contains examples of all conditions likely to occur once you deploy your model. One robust feature of neural networks is that they are able to perform even with spotty or missing values, so that, for example, if you don’t have an input value, such as Age_Group, the model will still try to predict your desired output of Total_Sales.

It is important to construct small models rather than complex ones. For example, you may want to construct a separate model for every product on your website. This way you have the ability to generate propensity-to-purchase scores for every product or service you are selling, rather than a very general behavioral model. Of course for some retailers this will be impossible, such as a bookseller; however, you may want to construct models along product lines, such as a non-fiction model and a fiction model.

Models can be constructed to predict various factors including, for example, one that tries to predict the total number of purchases a new visitor is likely to make from your website. Using the same inputs or data we can construct this second model by simply changing the output from Total Sales to Number of Sales (Figure 9-25).

Figure 9-25
The data is again split and a new neural network model is created.

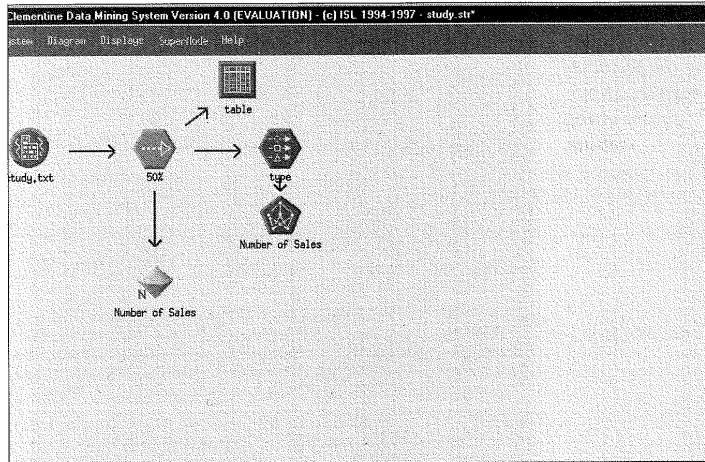
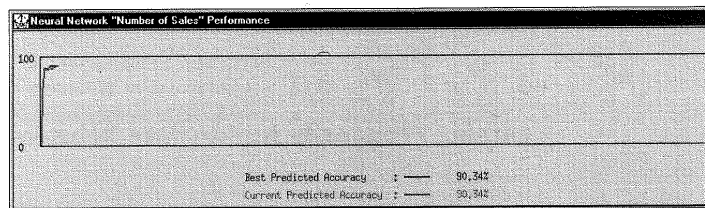


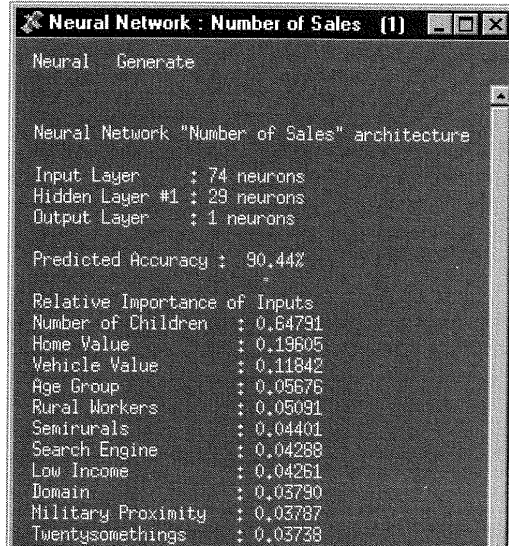
Figure 9-26
This model tends to be more accurate in predicting the number of sales.



As with the prior model, the neural network starts to learn the patterns of these accounts by comparing those that represent a high number of purchases against those with a low number of sales. Gradually every attribute in the data is evaluated and the model begins to increase its accuracy, eventually becoming able to predict the number of purchases a visitor is likely to make at a rate of 90 percent (Figure 9-26).

Interestingly the factors impacting the prediction of total number of purchases are different from those attributes impacting the predicted total dollar sales. Even though these two models are being used to mine the data from the same website, different factors impact the values we are trying to predict. For example, as shown in Figure 9-27, for this second model the most important factor is not the value of the visitor's auto, but is instead—by far, the number of children a visitor has.

Figure 9-27
*Factors impact-
 ing the predicted
 number of pur-
 chases at this
 website.*



Learning Propensity to Purchase Behavior

Through the analysis of your website's transactional server data—which has been commingled with household and demographic information—you can compile a profile of your potential future new clients. You can also begin to model and predict their expected behavior, and for a retailing website, their propensity to make online purchases. The process is based on an analysis of consumer features and behavior, and of course in the leveraging of (AI) technology.

For example, suppose Joe Blow buys "Product Z" and pays \$46 via his Visa card.

1. *Observed Behavior.* The number of sales made to this visitor is recorded, as is the

dollar amount, which is compiled and merged with other data from your web server, your data warehouse, and external household and demographic information, all of which will become a single string of data to be analyzed with thousands of other customer transactions by a data mining tool.

2. *Neural Network.* The data mining tool incorporating a back-propagation neural network analyzes the entire customer transactional process and the visitor attributes and behavior from a training data set, from which a signature gradually evolves. As the neural network trains itself on the relationships it gradually learns the

patterns distinguishing high sales customers from low or no sale visitors.

3. **Predictive Model.** This is the formula or set of weights written in code which classifies the observed behavior. It is the codified end product of the data mining process, which can be deployed at the server level to predict and classify new website visitors.
4. **New Visitors.** The behavior and attributes of new visitors are matched against the predictive model in order to ascertain their expected profitability and propensity to purchase your products and services. You may want to split your visitors into different classes based on their predicted outcome.
5. **Score.** The predictive model can be used to generate a "score" on each visitor and customer in order to classify and predict their online behavior and overall value to your website. You may want to prioritize your visitors based on their predicted profitability scores and probability to click through on ads or make a purchase of certain products and services.
6. **Marketing Action.** On the basis of the predicted classification or behavioral score of your new visitors you are able to present

the appropriate ads, banners, and marketing messages. You may want to also make a targeted e-mail marketing campaign on the basis of what products or services they are likely to want to purchase.

Thus at the conclusion of this process, your marketing decision is: E-mail all the Joe Blow look-alikes about Product Z.

Figure 9-28 *The Store Clerk can be an AI*

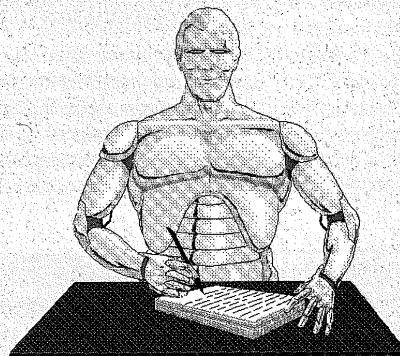
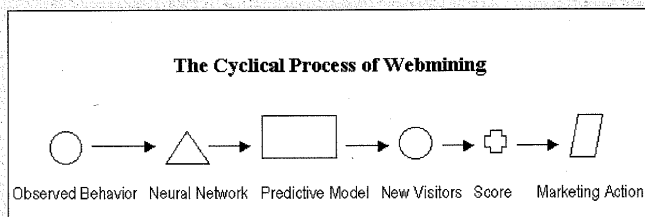


Figure 9-29
E-mining is the process of evolving to meet your customer's demands and needs.

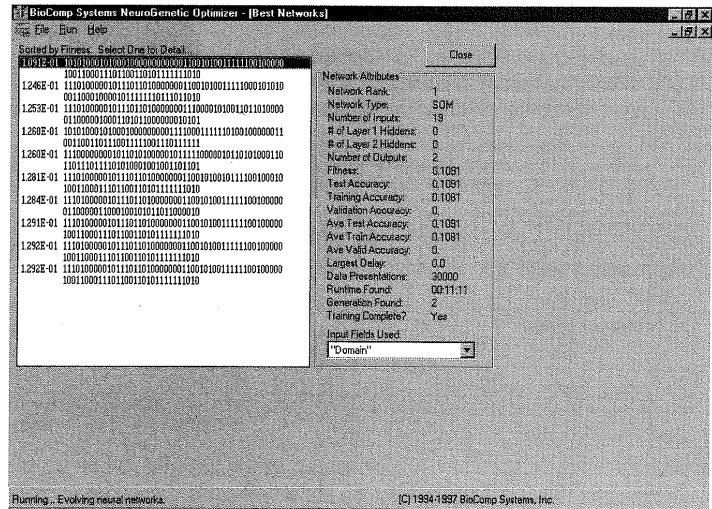


Optimization

One of the new ways data mining technologies such as genetic algorithms are being used by some forward-thinking physical retailers is in the overall design of their stores. Some retailers are beginning to use both neural networks and genetic algorithms to assist them in how they design their storefronts, how they arrange their shelves, how they position their products, and how they manage their entire inventory. Used primarily to assist them in this logistic problem, they are also using these technologies to evaluate the consumption patterns of their customers in order to maximize the efficiency of how and what inventory they deliver in order to minimize their overhead. Some retailers in fact are using this combination of nets and GAs to plan their entire stores, from the height of their shelves to the sizing of their parking lots (Figure 9–30).

As we have seen earlier, neural networks are excellent tools for analyzing patterns in data—large amounts of data. When combined with genetic algorithms their effectiveness improves, because GAs are primarily engines of optimization. When used in tandem, the GAs can work at the front end of a neural network to fine tune and recycle a number of architecture settings for a neural network, trying out mul-

Figure 9–30
A genetic algorithm tries thousands of combinations in tuning a neural network.



multiple models and gradually converging on the ideal solution to very complex problems.

You need to be aware about a couple of things regarding this data mining approach. First, it is very computationally intensive, meaning that it usually takes time to arrive at an optimum solution. Convergence on a model can often take several hours and, at times, several days. If you'll recall, a GA goes through several processes in its search for a solution including that of reproducing, crossover, and mutation in its random search for the optimum answer to problems. In this instance it is looking for the combined factors which affect the output you are trying to discover, such as "Total Number of Sales" or "Total Amount of Purchases."

This optimization process will recycle, mutate, and evolve until it reaches an optimum result as set by the user or to the tool default settings. As with a physical store, a website can also use a neural network and a genetic algorithm to increase its effectiveness by optimizing its overall design. The optimization of the storefront is done through a process of trial and error in order to improve overall online sales for the website. It is a process of evolutionary retailing. The particular data mining tool shown in Figure 9-31 provides the user with a wide range of instruments with which to observe the process of how the

Figure 9-31
*Multiple views
of the process of
evolutionary
retailing.*

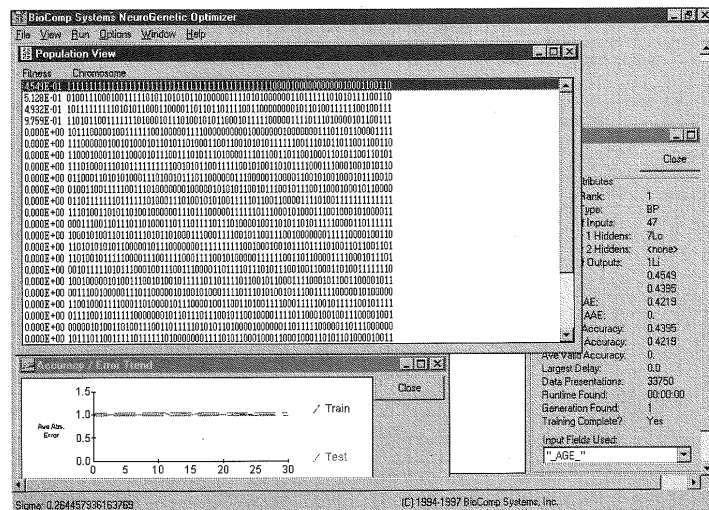
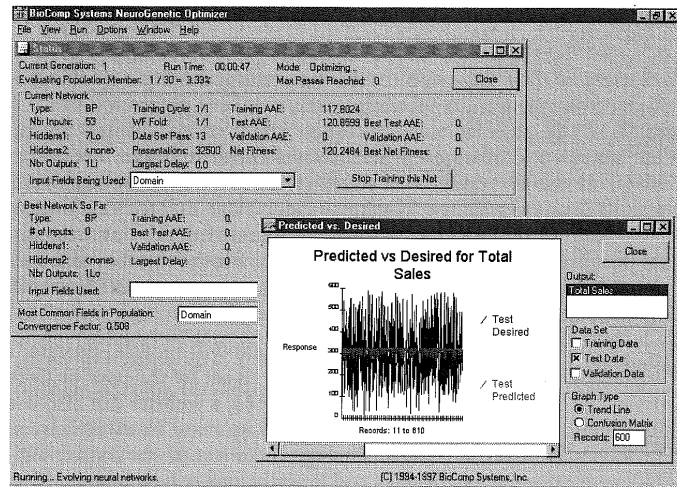


Figure 9-32
A back-propagation network can be optimized by a GA to predict your online sales.



predicted outcome is eventually converged on—through a process of training, testing, and adjustment.

Similar to the optimization problem of physical storefronts, websites can also benefit from the deployment of GAs to assist them in the overall design of their products and services in their virtual stores. Just as physical retailers are using these data mining technologies to assist them in the design of their stores and the management of their inventory, so can web teams use GAs and nets to assist them in mining for the most effective website design for e-commerce. For example, what banners or offers and incentives should be paired up for each of your various product lines? What is the most revenue-effective sequence to arrange your pages? Which offers tend to maximize clickthroughs or ad views, or what product combinations tend to be optimized for improved sales?

As we have said from the beginning of the book, data mining technology is about perception and learning, which is what neural networks and genetic algorithms are designed to emulate. This evolutionary type of computing technologies can assist the designer and administrator of a website in optimizing their virtual storefront to ensure they are getting their biggest bang for their buck. Through the use of these data mining tools you can leverage the pattern recognition capabilities of the network to maximize your marketing and retailing efforts on the Web, with the GA being used to optimize the effectiveness of the network.

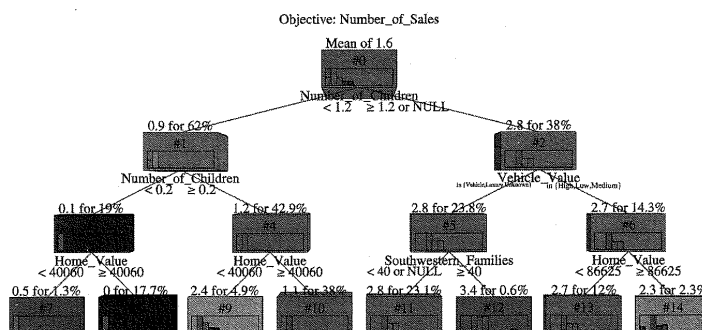
Visualization

Having the ability to get close to your data is very important. The ability to use one of the most powerful instruments of perception—the human eye—is critical in the mining of your website data. A data mining analysis using a machine-learning algorithm, which generates a decision tree, can offer you an insightful view of your website customer database. The tree illustrated in Figure 9–33 provides a segmented view of factors affecting the number of sales at this website.

This decision tree in Figure 9–33 revealed several interesting market segments impacting the number of sales at this website. It discovered that the number of children and additional household and demographic variables impacted the number of sales, which averaged 1.6. The most significant rules discovered were as follows:

IF	Number of Children is greater than or equal to 1.2
THEN	Number of Sales was 2.8 for 38% of the sample
IF	Number of Children is greater than or equal to 1.2
AND	Vehicle Value is Luxury or Unknown
THEN	Number of Sales was 2.8 for 23.8% of the sample
IF	Number of Children is greater than or equal to 1.2
AND	Vehicle Value is Luxury or Unknown
AND	Southwestern Families ZIP code is greater than or equal to 40%
THEN	Number of Sales was 3.4 for 0.6% of the sample

Figure 9–33
This view highlights the factors resulting in higher sales levels.



The identified consumer group of Southwestern Families is a small market, 2.8 percent of all households in the United States, with a small share (1.8 percent) of disposable income. They are a young population reflecting an emphasis on children and family. The average family size for this group is 3.8 which is 20 percent higher than the U.S. average. The spending of this consumer group is family-oriented, primarily baby products and children's apparel. Located in urbanized areas and smaller cities, 95 percent of these consumers can be found in New Mexico, Texas, Arizona, Colorado, and California, with the highest concentrations being found in communities like Wharton, TX, San Bernardino, CA, and Santa Rosa, NM.

Views of your website customers can be performed via multiple factors—for example, as shown in Figure 9–34, you can segment your website visitors on the basis of the total number of purchases by the types of products sold. Note that in Figure 9–34 for consumers of certain “Home Essential” products, the average number of purchases is extremely high: 101.7 orders, for a sample size of 4.8 percent of all customers. This three-dimensional decision tree generated from an analysis of a retailer website also found other products in combination which point to consumers with higher than average sales. This particular tool allows for the data to be viewed both in a decision tree format or as in a geographic map format. The map in Figure 9–35 is a view of the same website data displayed according to geographic state.

Figure 9–34
This decision tree is segmenting total sales on the basis of type of products purchased.

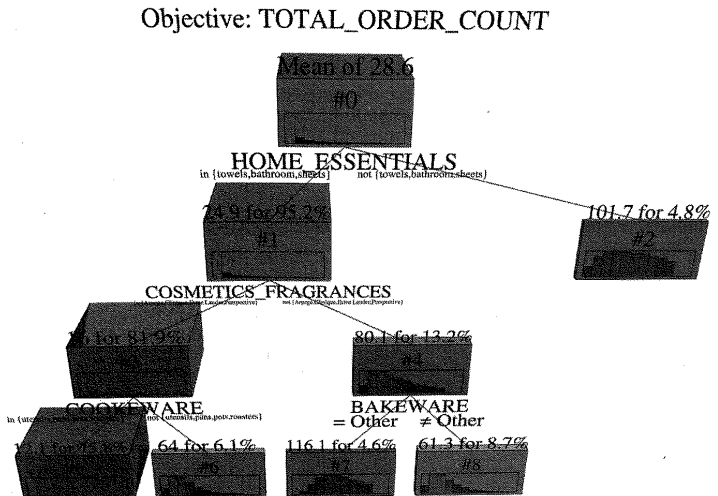


Figure 9-35
A map view of a
website cus-
tomer database.

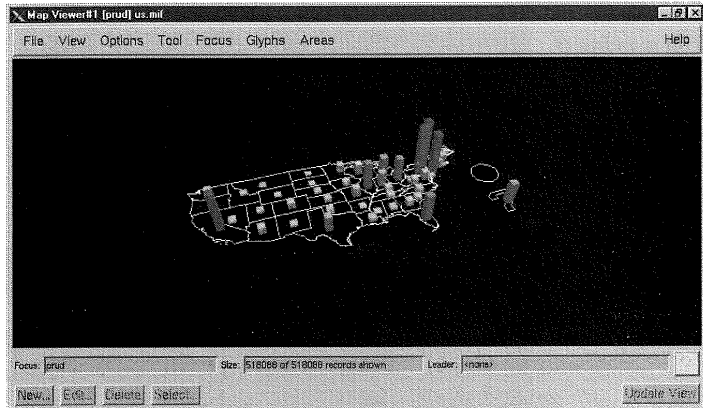
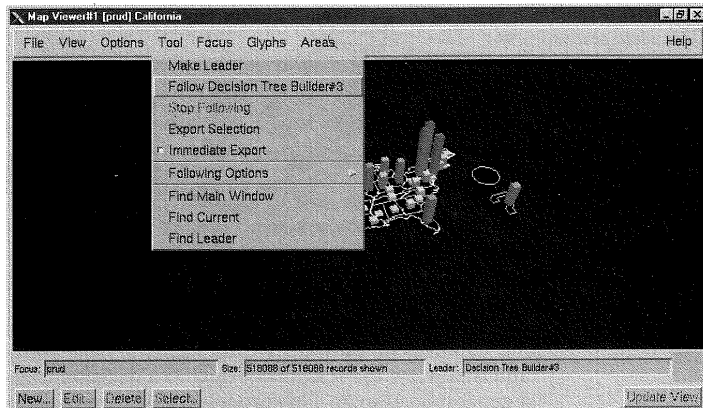


Figure 9-36
The view of the
map can be
made to follow
that of the deci-
sion tree.



Using a linking feature within the data mining software, we can view specific market segments discovered via the decision tree component (Figure 9-36). The map in Figure 9-37 reflects a view of *only* those customers who were identified in the decision tree as having a higher than average number of sales, the consumers of HOME ESSENTIAL products. This market segment, identified by the previous decision tree, can be further visually analyzed using this data mining tool. Simply by selecting the state of California, additional insight can be gained at the county level, as shown in Figure 9-38.

Figure 9-37
This is a view of website customers who made a high number of purchases.

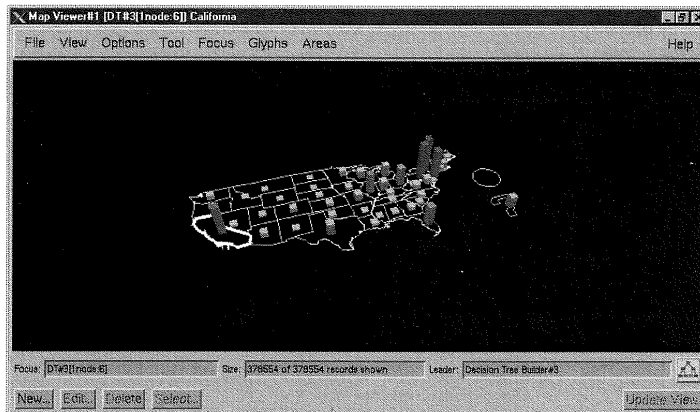
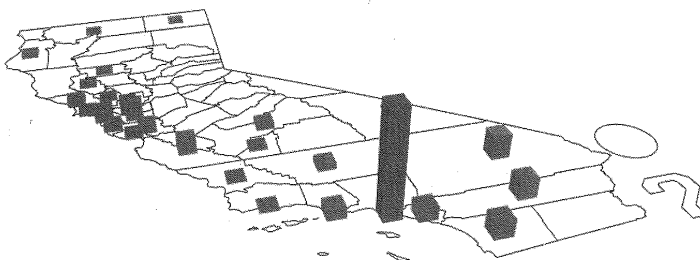


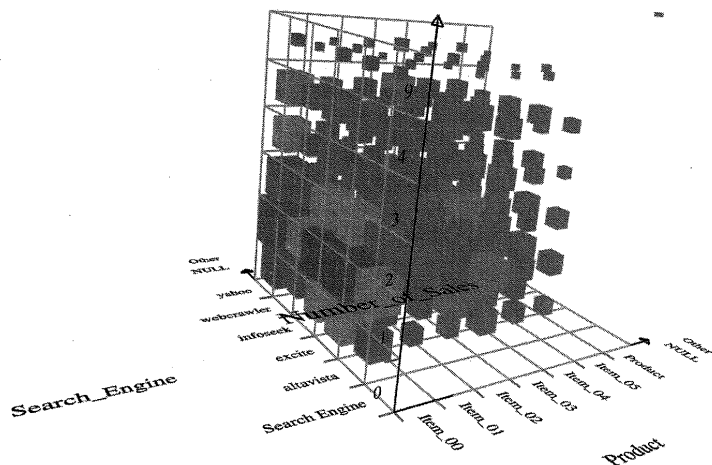
Figure 9-38
This view shows a concentration of customers who made a high number of purchases are mostly located around San Diego.



Using this visualization tool we can continue to drill down to the ZIP code level. This illustrates how an analysis using a machine-learning component can be used to segment your website customer data and then view the results not only in a decision tree format but also in a geographical map. This process transforms the segmentation of your stateless data into a geographical format so that you can visually see your online customers on a map. This data mining tool also allows you to drill down further to a more refined level so that you can literally see what states and cities your website visitors are coming from.

The same process can be applied to a cross distribution viewer so that you can do some exploratory analysis by which you compare the total number of purchases against other variables, such as the search engine a visitor used, or the type of product they purchased. This graphical output can be linked to either the decision tree or the geographical map components. Figure 9-39 is a view of the example

Figure 9-39
A view of product types by number of sales and search engine origins.



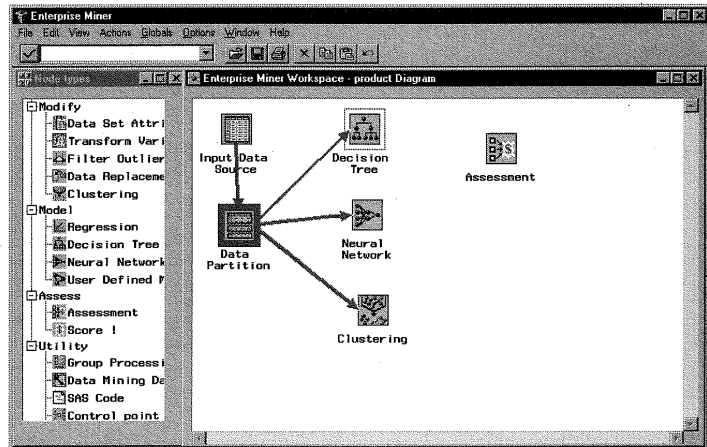
website's customer database by the dimensions of search engine, product lines, and number of sales in a three-dimensional cube.

The darker shading in this three-dimensional viewer represents a higher than average rate of sales. For example, the red cube on the lower left shows a high number of Yahoo! visitors are purchasing Item_00. Using this type of tool also allows the user to click on any object in order to drill down for additional information and to extract the SQL syntax, which generated it. A visual tool such as this allows the Web administrator to interact with the data from their website at a very close level. Visualization coupled with data mining algorithms and geographical mapping features enable the user to discover relations and explore them further in an interactive, three-dimensional environment.

It's an Integrated Iterative Process

Data mining your website is an iterative process very similar to an organic method of learning and adapting. It is a process of observation, mapping, modeling, and codifying your website visitors' online behavior. Every interaction at your website represents an opportunity for your company to learn something about your current customers and potential clients. Data mining software and techniques provide you a biological feed-forward method of observing your visitors'

Figure 9-40
*An integrated
 e-mining
 workspace.*



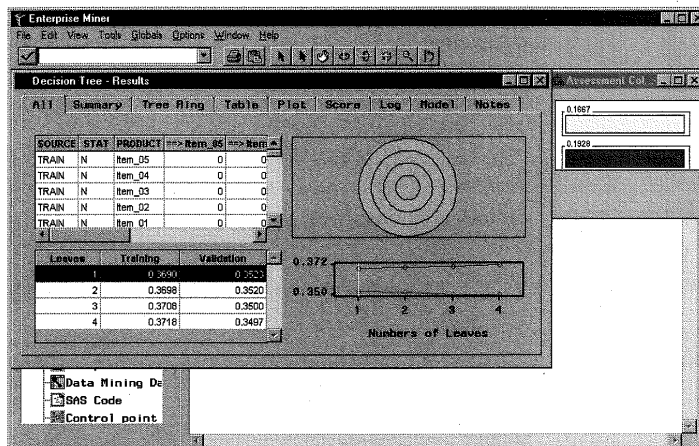
actions and reacting to those actions by providing them what they want in a customized and personalized manner.

As we have seen, there are various formats by which data mining techniques and tools can generate their results and benefits to a retailing website and the teams that are responsible for their maintenance. We have seen the clustering of data with inherent similarities. We have seen decision trees and IF/THEN rules, which can segment your data into significant subsets of website customer types. We have seen modeling via a neural network by itself and coupled with a genetic algorithm for optimization. We have seen how all of these results can be visualized in three-dimensional cubes and geographical maps.

Data mining provides web administrators and designers the ability to extend their customer service capabilities and enhance their perceptions about their markets and visitors. This type of inductive analysis is not like a query reporting statement in which the analyst poses a question. A data mining analysis is more along the lines of using an algorithm or network to discover a hidden signature in the data accumulating in your server. A signature or patterns which can be used to predict future visitor behavior, which a firm can use to:

- Design their website to maximize sales
- Anticipate demand on online inventory
- Attract more advertising to their site
- Predict website purchasing patterns

Figure 9-41
Multiple views
of a website
database.



- View evolving market developments
- Identify new website customers
- Be more competitive online

The e-commerce uses for data mining are many. The technology and techniques can help you discover the attributes of current customers in order to find new visitors who fit that profile. Conversely, data mining can be used to discover unprofitable customers. Data mining can produce an enterprise model that can be used for “what if” scenarios for website market segmentation. The tools can also be used to increase website visitor satisfaction and improve customer service by focusing on your online client relationship. Data mining can also be used to find quality problems, monitor improvements, and measure the success of cross-sales opportunities in your website. For example, Figure 9-41 shows how it is possible to generate a decision tree in order to segment and view the features of your website visitors and their attributes:

Data mining can be used to fine tune market demand with your product supply chain, reducing inventory carrying costs and cyclical inventory ups and downs, on the basis of purchasing patterns discovered through your analysis. It can mean the answer to such questions as:

- What characteristics make a new website visitor more responsive to a special offer?
- What affinity programs will generate the most interest and revenue sales?

- What kinds of online customers generate more transactions and revenue?
- How should your marketing and advertising be timed to maximize yield?

Figure 9-42
You can model on prior behavior in order to anticipate new visitors' propensity to buy.

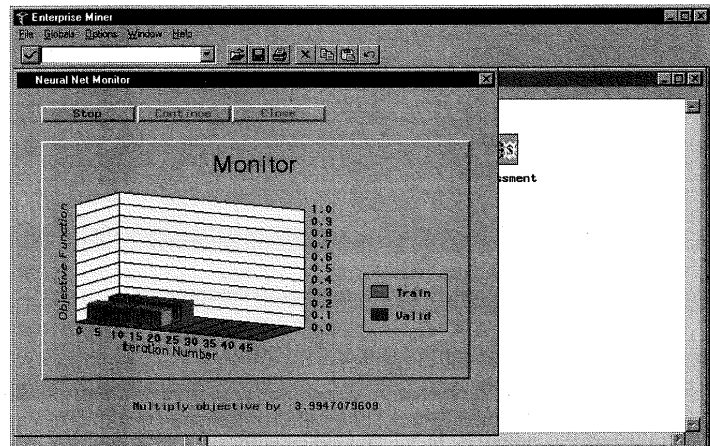


Figure 9-43
You can monitor, adjust, and refine your predictive models.

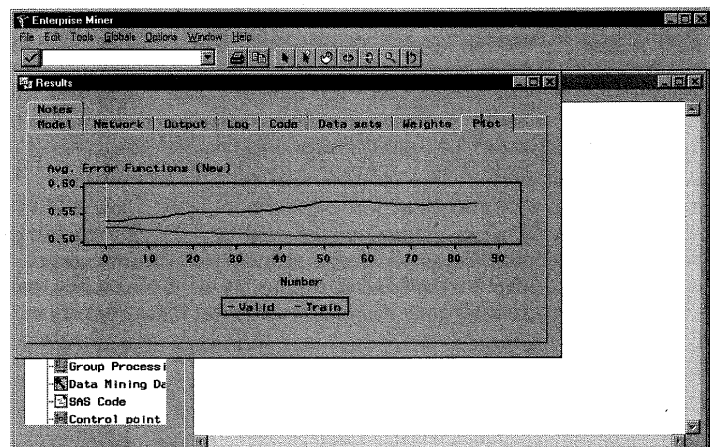
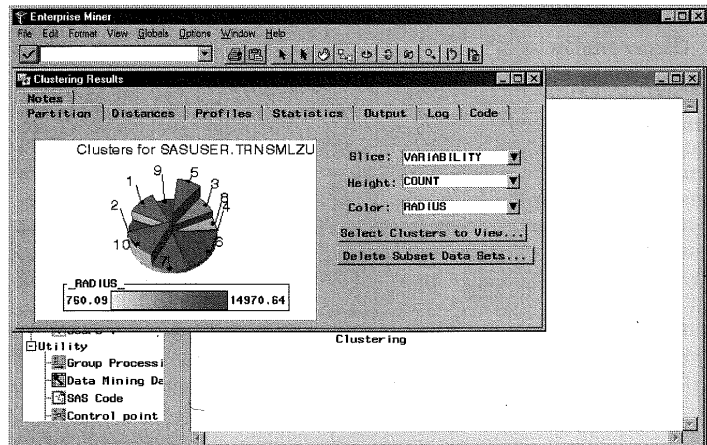


Figure 9-44
Multiple clusters
can be extracted
and viewed from
your data.



A neural network can be used to train on the patterns of past website customers in order to predict the future behavior of new visitors (Figure 9-42).

A key advantage to data mining your website is that it will readily point out patterns of traffic, which will help you to improve the overall design of your site. For large commercial sites the discovery of common paths taken by thousands of visitors may point to some overlooked opportunities for cross-selling of related products. A data set from your website can be split into training, testing, and validating subsets, which can be used to measure and compare the accuracy of your predictive model (Figure 9-43). Using the same data from your website you can create clusters, market segments, and predictive models, as shown in Figure 9-44.

The demographics you discover, whether via ZIP code registration forms or other methods, will also assist you and your design team in the planning of your site. By knowing who visits and makes purchases at your website, you are able to know the type of products and services they are most receptive to purchasing. Figure 9-45 is a graphical view of the multiple clusters discovered from an analysis as they relate to such factors as search engines, gender, age groups, etc.

In the end, the benefits of data mining technologies lie in the ability to integrate all of these processes so that clustering can be linked to segmentation and segmentation be linked to a neural network, say,

Figure 9-45
Cluster profiles
can be viewed and
altered on
the fly by select-
ing the various
attributes.

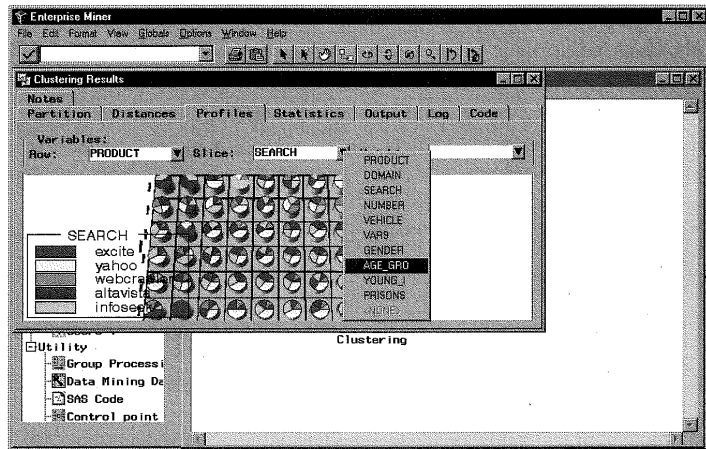
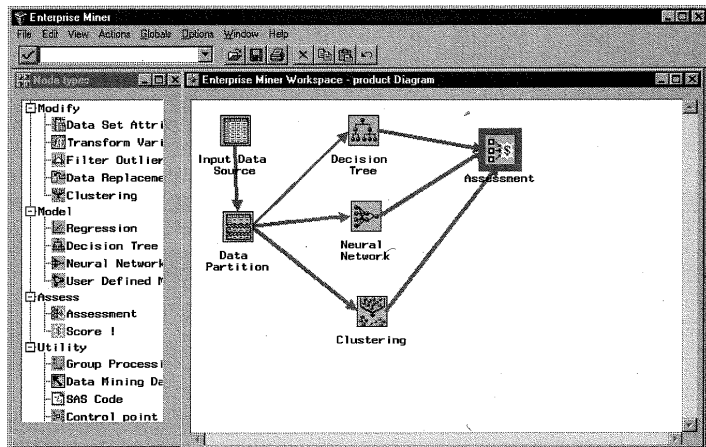


Figure 9-46
Clustering, seg-
mentation, and
classification
can be done in
parallel.



for a specific market segment. Where a neural network can be optimized by a genetic algorithm, with its results ported over to a three-dimensional viewer that can generate a geographic map, and so on. Having the ability to segment, cluster, and search for meaningful patterns in your data can be accomplished in parallel, allowing you to

gain as much insight as possible about your website visitors and customers (Figure 9-46).

Data Mining Future

The data mining industry is a relatively new one. Most companies are recent startups. There are over a hundred companies providing an assortment of tools and services, consulting and seminars, and even specialized hardware. Surprisingly, almost all of them are concentrating their efforts exclusively in the area of datamarts and data warehouses; no company is looking at the Web as a potential market for mining (with the possible exception of Aptex, a spin-off from HNC). This is astonishing, in light of the explosive growth of electronic commerce and the Web as a marketing and retailing channel. As websites increasingly becomes the first point of contact between a company and its current and future customers the mining of this transactional data will grow in importance. So far most of these efforts have involved traffic reports, collaborative filtering, and the aggregate pooling of cookies. Few companies are looking at data mining technology to analyze the customer information being generated from their website.

Soon the core data mining technology will begin to be incorporated directly into web server systems in order to directly complement e-commerce activities. There is also considerable emerging research in the fields of databases, machine learning, applied statistics, visualization, and other fields relevant to data mining and knowledge discovery. Despite all the activities both in the market and in R&D laboratories, the field of data mining is still in its infancy. The data mining market is still in the early adoption phase, with most products and vendors still not fully matured. Many of the current commercial products do not yet have the robustness, scalability, and functionality that customers expect from other, more mature technologies. Various data-mining algorithms impose serious restrictions on their application. The process of data mining is far from automated, and is more an art than a science, and therefore quite difficult to use effectively. The field of data mining websites is also an interdisciplinary one, requiring cross-fertilization of several fields. It introduces new complexity involving logs, cookies, HTTP, CGI, and evolving new technologies from the field of electronic commerce like Java. Having said all of that, the benefits of e-mining can be staggering.

The Personalization of Self

In the new web-centric economy, customer expectations and demands will require that websites provide a new unprecedented level of customer service and product customization. It also means that consumers will grow to expect a new level of self-service and total customer control. It will be the age of the consumer who will require services and products via the Web tailor-made to their exact specifications, which demand that the retailer of the future be friendly, responsive, and above all, very knowledgeable about the demands of each of their customers. Data mining can assist the retailer of the future with this type of personalization.

Already we are beginning to see this with the personalization of certain search engine portals. By providing these portals with some personal information about yourself, such as the ZIP code you reside in, they are able to provide you targeted information about your city, state, time zone, latitude, and longitude. These portals can personalize the information you want so that you can customize your data feeds for:

Stock Portfolio	Headline News
Sports Scores	Sports Ticker
Home and Living Links	Favorite Links
Horoscope	Cartoons
Editorial Columns	Lottery
Search the Web	Weather
Sunrise/Sunset	Moon
Tides	TV Listings
Movie Listings	Video Releases
Upcoming Local Events	Reminders
Notepad	Holidays
Religious Holidays	U.S. Holidays/Events
Sports Reminders	Weather Reminders

However, the preference information you provide these portals can be mined and new and unique profiles about your likes and dislikes can be developed in order to personalize what information, ads,

banners, links, and offers are presented to you. A future benefit of data mining is that it will filter information and prevent data overload, which is already at this early phase of the Web developing into a major problem. Data mining may well end up serving the function of content sifting and information filtering.

Right Turns, Hand Grenades, and Integrity

Consumer studies have found that 90 percent of the time, upon entering a store like a supermarket, shoppers turn right rather than left or straight. The question today of course is how do they behave in a virtual store, such as your website? Given thousands of visits, how do customers behave when they hit your site: do they go right or left? These are some of the questions technologies such as data mining are attempting to comprehend and synthesize about online customer behavior. This is the type of insight web teams and designers, marketers, and business units need to know about their visitors. They need to know also what these visitors want when they do turn left or right. Data mining in combination with other tools and technologies is all an effort to understand this type of online customer behavior.

Data mining can be compared to a hand grenade: you don't have to be perfect to be on target. As you mine your data you most likely will see incremental improvements, but this is OK. Soon you will see that data mining will provide you a more effective way of targeting your sales and marketing efforts.

Remember that the information that you capture and mine to enhance your relationship with your customers is the most valuable asset your website can have in a digitized networked economy. It provides your website the ability of distinguishing yourself from your competitors. It allows you to know what your customers want and what they like—and as such it gives you the leverage to provide it to them in your services and products. As we said in the beginning of the book, your customer information should be protected and guarded. Your integrity and customer relationship are riding on it.

There is an old joke about two campers who are surprised by a grizzly bear late at night. One of the campers stops to put his shoes on. The other one asks, "What are you doing? You know you can't outrun a bear!" To which the first camper replies, "I don't have to outrun the bear. All I have to do is outrun you!" Sometimes, being better than your competitor is good enough. Use data mining to serve your customers and to provide them the precise, personalized, prompt attention they cannot get anywhere else.

Appendix

Privacy Consortiums, Standards, and Legislation

With recent concerns over the gathering of personal information, privacy consortiums, software standards, and online legislation have evolved in an effort to regulate the Web. The following are a few of these privacy initiatives, which in the capturing and analysis of customer information you should be aware of.

TRUSTe

TRUSTe is a nonprofit web organization founded to promote online privacy. The TRUSTe “trustmark” is an online branded symbol that signifies a website that has made a commitment not to disclose its private practices by posting a privacy statement that is backed by TRUSTe’s assurance process. All TRUSTe licensees display the trustmark either on their website with a link to the site’s privacy statement or directly on the privacy statement. At a minimum, each privacy statement discloses:

- What type of information your site gathers
- How the information will be used
- Who the information will be shared with (if anyone)

TRUSTe member sites agree to disclose information management practices in privacy statements, display the trustmark, adhere to stated privacy practices, and cooperate with all reviews. The TRUSTe trustmarks labeling system tells the website visitors that the site they are shopping in adheres to a three-tier policy:

No Exchange. A website bearing the TRUSTe No Exchange trustmark lets website visitors know that no personal information is being collected or tracked.

Relational Exchange. This *trustmark* tells visitors that the personal information being collected from the website will be retained exclusively by that site and will not be shared, transferred, or sold to third parties.

Third Party Exchange. This trustmark alerts the Web visitor that this site will exchange their personal information with other parties. A privacy statement informs the visitor on what is being done with their personal information.

Direct Marketing Association

Not surprisingly, the DMA has also opted for self-regulation rather than government intervention. The DMA, however, has taken the initiative to issue guidelines on the collection of personal information alerting marketers to furnish individuals the opportunity to prohibit the disclosure of their personal information. The DMA has also issued specific guidelines regarding the collection of data from and about children, recommending that marketers be sensitive to parents' concerns and consent over the solicitation of children's name, address, and other personal information.

Internet Engineering Task Force

The IETF created a Request for Comments *RFC 2109 State Management Mechanism* in an attempt to set some cookie standards. The proposal sought to establish some industry guidelines on the use and control of how cookies are controlled by future versions of browsers. The proposal wanted to give web browsers the ability to block, reject, and delete cookies. New browser versions from Netscape and Microsoft did not implement the proposal, however.

Platform for Privacy Preference Project

Formed by the World Wide Web Consortium (W3C), the P3P seeks to ensure that there are standard privacy practices and protocols for

the collection of personal information. It tries to give web visitors the right to informed consent. The P3P specifications will enable websites to express their privacy practices and users to exercise preferences over those practices. P3P products will allow users to be informed of site practices, to delegate decisions to their computer when possible, and allow users to tailor their relationship to specific sites. P3P gives users the ability to make informed decisions regarding their personal information.

The P3P seeks to establish protocols for providing the user the choice of whether personal data will be passed from the browser to the server and how that information will be used. It is also seeking to establish standards on cookies and the Open Profile Standard (OPS). P3P will support future digital certificate and digital signature capabilities, as they become available. P3P can be incorporated into browsers, servers, or proxy servers that sit between a client and server.

Open Profile Standard

The idea behind the Open Profiling Standard is that user information could be maintained in partitioned files, initially on users' hard disks. Separate areas would hold demographic information, web-specific information (such as e-mail addresses), and information specific to websites, such as content preferences or agent profiling keys. Users would choose which partitions a website's server could look behind and when and if websites could trade information, with the local resident profile as broker. Digital certificates from VeriSign would guarantee the user a measure of security on what was released, and the site a measure of accuracy that apparent users would be who they said they were.

The proposal, backed by more than sixty Internet vendors and the Federal Trade Commission, aims to form a universal system so that Internet users do not have to enter the same information at every site they visit. It calls for personal data such as age, gender, mailing address, hobbies, and phone number to be stored in a personal profile file that resides on the users' hard disks. With the user's permission, that profile can then be accessed by websites, who can tailor content for individuals based on individual preferences.

OPS builds upon two primary pieces of technology: vCards, or electronic business cards, that can pass information across multiple

platforms, and digital certificates that ensure that parties on the Web are who they claim to be. The standard also calls for encryption so that personal information can only be viewed by those for whom it's intended. The OPS is a standard protocol proposed by Firefly, Netscape, and VeriSign which seeks to establish a way for the exchange, encryption, and storage of profile information from the browser to the server, the idea being that the website visitor would have final word on what profile information is passed to the requesting server. The OPS initiative proposes the following:

Permission Management: Protection of profile information, with constraints on access and modification.

Standard Attribute: A common standard for profile information, including identification, demographic, numeric, personal, contact, currency, and preference data.

Transaction Logging: Record of exchange of profile information between browser and server.

The personal preset profile would be completed once, much like other browser settings, when the software is installed and would be controlled by the user—passing the profile information to those servers it wants to. This would alleviate the need to complete multiple registration forms at various sites.

Federal Trade Commission

The concerns over privacy in the Web have led to hearings by the FTC involving advocacy groups, online marketers, and legislators. After several hearings on the subject of privacy the FTC recommended the Web industry practice self-regulation rather than having Congress take regulatory action.

European Data Protection Directive

Perhaps the greatest impact on the protection of online privacy for American web visitors and customers may not come from U.S. legislators at all but may instead come about because of a European initiative known as the "European Data Protection Directive." Under this new directive any country that trades personal information with the U.K., France, Germany, Spain, Italy, or any other of the ten

EU states will be required to comply with Europe's new strict standards for privacy protection. Under the directive European citizens are guaranteed:

- the right of have access to their data
- the right to have inaccurate data rectified
- the right to know where the data originated
- the right to recourse against unlawful processing
- the right to withhold permission to use their data in direct marketing campaigns

To enforce the directive, every EU country will have a privacy commissioner, to whom every individual will have the right to appeal should their rights be violated. EU states will not be allowed to send personal information to countries that do not maintain similar privacy laws—including the United States. This could mean that certain personal information gleaned from cookie, log, and CGI or Java registration forms by commercial U.S. websites from European individuals will have to comply with the European directive.

Glossary

Algorithm. A computational process that takes a value or set of values as input and produces an output value.

Analytical model. A process for analyzing a database or table, such as a graph, a decision tree, a neural network, or a cluster. A model may be descriptive or predictive.

Anomalous data. Dirty or spotty data that result from errors or that represent unusual events.

API. Application Program Interface. Some data mining tools provide APIs for using models in scoring new accounts or cases in production systems.

ASCII. Acronym for the American Standard Code for Information Interchange. The American National Standards Institute established this standard set of character codes for the transfer of text between systems.

Association. A rule or condition which describes how often events occur together. For example, if a person buys wine he will also buy cheese 15 percent of the time and French bread 20 percent of the time.

Back-propagation neural network. An architecture network design that uses an input, an output, and a hidden layer. During training, information is propagated back through the network and used to update the connection weights.

C5.0. A machine-learning algorithm that provides a set of rules that you can apply to a new unseen data set to predict which records will have a given outcome, it was developed by J. Ross Quinlan. C5.0 is the successor to the C4.5 and ID3 (Iterative Dichotomiser) algorithms.

CART. Classification and Regression Trees. A statistical algorithm technique used for segmentation of a database. It creates binary trees and segments in two-way splits and was developed by L. Briemen in 1984.

CBR. Case-Based Reasoning. A cyclic process of solving a new problem: when a problem is successfully solved, the experience is retained in order to solve similar problems in the future. When it fails, the reason for the failure is identified and remembered in order to avoid the same mistake in the future.

CHAID. Chi-Square Automatic Interaction Detection. A statistical algorithm that segments a data set by using chi-square tests to create multi-way splits. The chi-square test measures the statistical association between categorical variables. It was developed by J. A. Hartigan in 1975.

Churn. A term for customer attrition or for customers joining a client plan and then quitting within a short period of time.

CGI. Common Gateway Interface. A standard for external gateway programs to interface with a web server. It is executed in real-time and it generally outputs dynamic information from a database query.

Classification. The process of dividing a data set 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, where distance is measured with respect to a variable you are trying to predict.

Clustering. The process of dividing a data set into a heterogeneous population into a number of more homogeneous subgroups or clusters. Unlike classification, clustering does not rely on predefined classes.

Collaborative filtering. A system that supports exploratory searching by providing users with information derived from the experiences of previous users. Collaborative filtering attempts to find information, products, and services of interest to their users, often using some scoring function.

Continuous value. Continuous value data has a value in an interval of real numbers. The value does not have to be an integer and is the opposite of discrete or categorical.

Cookies. Cookies are a standard mechanism that allows a website server to deliver an identification tag to a client browser, request that the client store the tag and, in certain circumstances, return the tag to

the server. Cookies allow websites to maintain information on a particular user across HTTP connections.

Data mining. The iterative process of extracting hidden predictive patterns or profiles from large databases, using AI technologies as well as statistical and marketing techniques.

Data validation. The process of ensuring that all values in a database are consistent and correctly recorded.

Data warehouse. A decision support system for storing and delivering massive quantities of data that is nonvolatile, cumulative, integrated, time-variant, and subject oriented, used in support of management's decision-making process.

Decision tree. A graphical representation of the relationships between a dependent variable (output) and a set of independent variables (inputs). Usually in the form of a tree-shaped structure that represents a set of decisions. The tree may be binary or multibranch depending on the algorithm used to segment the data. Each node represents a test of decision.

Dependent variable. The field that you wish to analyze or predict, such as *Will Buy* or *Total Predicted Sales Amount*. The output or desired result from your data mining analysis.

Dimension. In a flat or relational database, each field in a record represents a dimension. In a multidimensional table, a dimension is a set of similar entities. For example, a multidimensional sales database might include the dimensions Product, Time, and City.

Expert system. A program that contains both declarative (facts) and procedural (actions) knowledge to emulate the reasoning processes of human experts in a particular domain.

Exploratory analysis. The use of graphical and descriptive statistical techniques to explore the hidden relationships and structure of a dataset.

Forms. A technique by which a user can input information on a web page and send it to the requesting site. There are innumerable uses for this technique for capturing customer information.

Fuzzy logic. A synonym for continuous logic, a deductive system predicated on the notion that truth is a multi-valued continuous quantity. A mathematical measurement of degree of membership, not necessarily 0 or 1.

Genetic algorithms. A set of techniques that use the concepts of natural evolution—processes such as genetic combination, mutation, and natural selection—to find the optimal solution to a problem.

HTML. Hypertext Markup Language.

HTTP. Hypertext Transfer Protocol. The communications protocol used to transport hypertext over the Internet and over private intranets.

Inductive data analysis. A data mining analysis that provides insights into trends, behaviors, or events that have already occurred and infers generalizations from examples in the data.

Kohonen neural network. See *Self-Organizing Map*.

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.

Machine learning. A branch of AI that deals with the design and application of learning algorithms.

Market basket analysis. A data mining analysis typically done by retailers who treat the purchase of a number of items (the shopping basket) as a single transaction with the intent of finding trends across large numbers of transactions and exploiting unique buying patterns.

MPP. Massively Parallel Processing. A computer configuration that is able to use CPUs simultaneously.

Multidimensional database. A database designed for on-line analytical processing. Structured as a multidimensional hypercube with one axis per dimension.

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.

Neural network. A nonlinear predictive model that learns through training and resembles a biological neural network in structure.

Nonlinear model. An analytical model that does not assume linear relationships in the coefficients of the variables being studied.

OLAP. Online 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 be significant.

Overfitting. Opposite of generalization, a tendency to assign importance to random variations in a data set.

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.

Prediction. A structure and process for predicting the values of specified variables in a dataset. A model resulting from data mining designed to predict future trends, behaviors, or events based on historical data.

Rule generated. The extraction of IF/THEN condition from data based on statistical significance.

Segmentation. The subdivision of a population according to variables which are good discriminators.

Self-organizing map. A neural network architecture that uses unsupervised learning for cluster analysis. It was developed by Teuvo Kohonen and is also known as a Kohonen network.

Sequencing. The analysis of a sequence of measurements made at specified time intervals. Time is usually the dominating dimension of the data.

SMP. Symmetric Multi-Processor. A multiprocessor hardware system design in which memory is shared among the processors.

SQL. Structured Query Language. A standard interactive and programming language for getting information from and updating a database. Although SQL is both an ANSI and an ISO standard, many database products support SQL with proprietary extensions to the standard language.

TCP/IP. Transmission Control Protocol/Internet Protocol. The set of communication protocols used by all computers on the Internet. While TCP/IP has always been used as the standard protocol for UNIX, versions of TCP/IP are now available for all major operating systems.

Visualization. The visual interpretation of complex relationships in multidimensional data.

Index

A

AC2, 106
 access logs, 194–195, 210–211
 Accrued, 6
 Accrue Insight, 58
 accuracy, 137–138
 ACORN Consumer Classification System, 157–158, 233–238
 Acxiom, 229–232, 230–232
 Adaptive Probabilistic Concept Modeling (APCM), 250
 address fields, 94
 AdForce, 258–259
 advertising
 management software, 258–259
 predicting clickthroughs of, 328
 targeting, 74–77, 252–253
 affluent families category, 233–234
 agent logs, 194, 198
 server configuration for, 199–200
 agent software, 130, 277
 Agentware Content Server, 250
 AID, 103
 algorithms. *See also* genetic algorithms (GAs); machine-learning algorithms
 association rule, 14–15
 decision tree, 14–15
 decision tree generation with, 102–104
 defined, 355
 determining most appropriate, 104–105
 genetic, 107–114
 K-nearest neighbor statistical, 310, 358
 machine-learning, 63–64, 98–107
 pattern recognition in, 83–84
 segmentation with, 98–99, 104–105
 Alice, 106
 analysis
 deductive, 84–85
 drill up, 42
 exploratory, 357
 inductive, 85–86
 link, 66–67
 market basket, 310, 358
 regression, 48
 sensitivity, 140, 327
 traffic, 6, 9
 Analyst Notebook, 67
 analytical models, 355. *See also* models/modeling
 Andromedia, 6
 anomalous data, 355
 antihuser field, 196
 Apache server configuration, 199–200
 APCM (Adaptive Probabilistic Concept Modeling), 250
 API (Application Program Interface), 355
 Applied Geographic Solutions, 228–229
 Aptex, 9, 62, 249–250
 Aquas Bazaar Analyzer, 58
 Aria, 58–59
 artificial intelligence (AI), 1, 83–86.
 See also neural networks
 case-based reasoning, 114–115
 deductive analysis in, 84–85
 fuzzy logic, 115–117
 inductive analysis in, 85–86
 OptiMark retailing system, 296
 Art Technology Group (ATG), 248–249
 ASCII, 355
 association, 13–15
 defined, 355
 in e-mining, 310–318
 rules/graphs, 14–15
 when to use, 122
 Astra, 59
 SiteManager, 65
 attributes, 45–46, 80
 computing entropy of, 99–100

of cookies, 206–207
 determining inclusion/exclusion of, 132–133
 discretizing, 92–93, 133
 identifying, 63–64
 auctions, online, 282, 295
 Autoclass, 44
 Autonomy, 250

B

back-propagation neural networks, 10–11. *See also* neural networks
 defined, 355
 modeling with, 328–329
 for nonlinear data, 144–145
 optimizing with genetic algorithms, 334
 structure and operation of, 88–90
 training, 91–92
 BackWeb, 129
 balancing data, 97
 banners
 data collection on, 126
 predicting clickthroughs of, 328
 baselines, 121–123
Beyond Regression (Werbo), 88
 biological-based business, 1
 BrainMaker, 98
 brand building sites, 124–125
 brand loyalty
 choice and, 36
 measuring, 45–46
 Briemen, L., 103
 BroadVision One-To-One, 250–252
 browsers, differentiating from buyers, 324–325
 browsing, personalizing, 9
 Business Miner, 106
 business-to-business transactions
 data providers on, 244
 in e-retailing, 44
 buttons, 127

C

- C4.5
 - in machine-learning algorithms, 98–99
 - when to use, 104
- C5.0, 103, 106
 - defined, 355
 - in machine-learning algorithms, 98–99
- CACI, 233–238
- CART (Classification and Regression Tree), 63, 77, 106
 - decision tree generation with, 102–103
 - defined, 356
 - in machine-learning algorithms, 98–99
 - when to use, 104
- case-based reasoning (CBR), 114–115, 270
 - customer service with, 284–286
 - defined, 356
- CBR Express, 115
- C code, 11, 69–70
 - weighting in, 69–70
- CERN server configuration, 201
- CGI. *See* Common Gateway Interface (CGI)
- CHAID (Chi-squared Automatic Interaction Detection), 63, 77
 - conservatism of, 103
 - for decision tree generation, 103
 - defined, 356
 - in machine-learning algorithms, 98–99
 - when to use, 104
- channels, 129
- chat, 130
- Chorus, 242
- chromosomes, 107
- churn
 - defined, 356
 - models, 119–121
- city dwellers category, 236–237
- Claritas, 48–49, 238–239
- classification, 10–11, 30–32
 - defined, 356
 - in e-mining, 325–331
 - with GAs, 112–113
 - with neural networks, 97
 - when to use, 122
- Clementine, 49, 96–97, 174–178
 - algorithms in, 103
 - data preparation for, 174
 - disadvantages of, 177–178
 - linking clustering to rule generation in, 175–177
 - linking supervised and unsupervised learning in, 178
 - machine-learning in, 106
 - modeling with, 174–175
- clickstream view, 247–248
- clickthroughs, 5, 75
 - data collection on, 126
 - predicting, 328
- clustering, 15, 28–32
 - compared with segmentation, 29
 - defined, 356
 - in e-mining, 310–318
 - with SOMs, 95–97
 - when to use, 147
- Cognos Scenario, 106
- collaborative filtering, 9, 115, 131.
 - See also* agent software
 - defined, 356
 - Firefly, 255
 - LikeMinds, 259–260
- comma-delimited files, 35
- CommentURL, 207
- Commerce Server, 217
- committees, classifier, 112–113
- Common Gateway Interface (CGI), 221–223
 - cookie creation with, 211–213
 - data component assembly with, 223–225
 - defined, 356
 - form processing with, 218–221
 - privacy issues and, 352–353
 - retrieving cookie values with, 212
 - virtual cookies, 210–211
- Common Log Format, 195–197
 - redundant data in, 226
- competitive intelligence, 47
 - conditions, 4
 - transactional, 12–13
- connectivity, of data mining tools, 140
- Consumer InfoBase, 230
- continuous values, 356
- cookies, 203–213
 - aggregate, 9, 210, 217–218
 - applications of, 207–210
 - assigning ID numbers with, 34–35
 - attributes of, 206–207
 - comments in, 206–207
 - configuring, 210–211
 - defined, 203, 356–357
 - future of, 206–207
 - globally unique identification, 215
 - information stored in typical, 205–206
 - limitations of, 210
 - linking to other data, 207–210
 - maximum number of, 205
 - merging data from, 22
 - operation of, 204–205, 208–209
- OPS and, 213–218
 - privacy issues and, 352–353
 - sample .txt file, 205
 - virtual, 210–211
- coupons, online, for market research, 76–77
- credit history profiles, 280
- credit information, 232, 239–242, 246–247
 - types of, 279–280
- CRISP-DM (Cross-Industry Standard Process for Data Mining), 154
- cross-bar associative networks, 87–88
- CrossGraphs, 67
- crossover, 109–110
- customers
 - ACORN classification of, 157–158
 - anticipating behavior of, 7–8
 - calculating purchasing propensity with neural networks, 94
 - database construction, 224–225
 - dialog with for marketing, 72
 - differentiating browsers from buyers, 324–325
 - lifetime value of, 45–46, 77–81
 - loyalty of, 45–46
 - obtaining information from, 21–22
 - OPS standard for data on, 213–218
 - permission from to use data, 22, 352
 - power of in e-commerce, 294–295
 - predicting purchases by, 326–331
 - preventing attrition of, 119–121 and privacy issues, 20–21
 - profiling, 33–35, 279–282, 330
 - and relational marketing, 19–20
 - VALS/iVALS categories, 151–152
 - visualizing, 335–339
 - well-informed, 268–269
- customer service, 285–287
- cybergraphics, 259–260

D

- DART, 252–253
- Darwin, 184
- data
 - adding outside, 8, 34–35, 80–81, 305–309
 - anomalous, 355
 - assembling components of, 223–225
 - balancing, 97
 - brand, 124–125
 - CGI and, 221–223

- classification/prediction with, 30–32
- clustering, 15, 28–32
- components of, 193–226
- conversions, 133
- cookie, 203–213
- creating customer, 124
- customer permission for use of, 22
- determining required, 74–75
- dirty, 226
- evaluating, 134–135
- explosion of available, 43–45
- from forms, 218–221
- importance of familiarity with, 132
- linking to other company databases, 134
- location of, 225–226
- marketing, 228–247
- missing, 133–134
- modeling, 30–32
- operational, 133
- OPS and, 213–218
- portability and cost of, 131–132
- preparing, 132–133
- preparing for neural network training, 92–93
- preprocessing, 140
- and privacy issues, 20–21
- providers of external, 227–265, 243
- random sampling, 133–134
- representing, 226
- scaling, 97
- selecting, 123–132
- selecting for neural network training, 91–92
- selection checklist, 123–124
- server log files, 194–198
- from single website hit, 202–203
- skewed, 97, 144–145
- transactional, 131, 330
- validation of, 357
- visualization of, 15–16, 65–68, 333–339
- webographics, 247–264
- data analysis. *See also* data mining
 - defined, 42
 - differences from other analysis, 42
 - sample of, 26–32
- Database Mining Marksman, 159–161
- Data Desk, 49
- DataEngine, 116, 178–181
- data interpretation systems, 53–54
- Data Junction, 134
- data mining. *See also* data mining tools; e-mining
 - architecture for, 35
 - artificial intelligence in, 83–86
 - assessing ROI from, 24
 - association with, 13–15
 - bookseller example of, 25–32
 - classification/prediction with, 10–11
 - clustering with, 15, 28–32, 95–97, 310–318
 - compared with OLAP, 51–54
 - compared with statistics, 48–50
 - compared with website analysis tools, 54–65
 - data needed for, 22
 - data preparation for, 6–7, 22–23
 - defined, 4–5, 357
 - determining frequency for, 24, 153–154
 - formats of, 6, 35
 - history of, 7–8
 - identifying objective for, 119–123
 - importance of, 8–10, 39–40
 - integration of, 24–25
 - optimization with, 17–19, 113–114, 332–334
 - output formats for, 69–71
 - outsourcing, 144
 - partitioning single transaction with, 77–81
 - planning for, 74–75, 119–123
 - presenting results of, 148–149
 - for relational marketing, 19–20
 - scheduling, 38–39
 - segmentation with, 11–13, 29–30, 98–99, 104–105, 312–325
 - separate servers for, 23
 - as simplifier, 78
 - solutions with, 68–71
 - steps in, 22–25, 119–154
 - traditional applications of, 46–47
 - unbiased/nonlinear nature of, 44–45
 - uses of, 4–7
 - using multiple techniques in, 17–19
 - using multiple tools in, 117
 - validating results of, 23–24, 148
 - visualization with, 15–16, 65–68, 335–339
 - vs. manual analysis, 44–45
 - vs. query/user-driven analysis, 5–6
 - why and who questions in, 71
- Data Mining: Trends, Technology, and Implementation Imperatives*, 47
- data mining tools, 6, 154–191
- accuracy of, 138
- built-for-business features of, 53
- connectivity of, 141
- containing GAs, 113
- efficiency of, 143–144
- evaluation copies of, 145–146
- formats available in, 139
- import/export in, 141
- list of leading, 189–192
- memory management of, 141–142
- multiple paradigm, 156
- with neural networks, 97–98
- outsourcing *vs.*, 144
- paradigms supported by, 143
- performance of, 142
- platform/price considerations in, 144
- preprocessing required by, 140
- robustness of, 142–143, 156
- scalability of, 137–138, 156
- selecting, 137–146
- selection criteria, 154–159
- single-paradigm, 159–172
- solutions provided by, 139–140
- suites, 173–183
- data preparation, 6–7, 22–23, 132–134, 140
 - for neural network training, 92–93
 - tools for, 135
- Data Surveyor, 106
- data warehouses, 228, 264–265
 - adding web data to, 307–309
 - customer profiling with, 305–309
 - defined, 357
 - integrating website data with, 149–154
 - integrating with customer information, 37
 - OLAP and, 51–52
 - processing applications *vs.* subject applications in, 264–265
- date fields, 93–94
- DecisionCentre, 187–188
- Decisionhouse, 106, 181–183
- Decision Series, 184
- decision trees, 12–13
 - advantages of, 103–104
 - binary *vs.* multiple branch, 99–101
 - defined, 357
 - machine-learning and, 63–64
 - overfitting, 323
- deductive analysis, 84–85
- degrees of membership, 116
- delivery systems, 9
- demographic/household data
 - in assigning keys, 34–35
 - customer profiling with, 305–309
 - enhancing LTV profiles with, 80–81

demographic/household data
(continued)
 providers, cautions for use of, 243
 providers of, 48–49, 227–265
 dependent variables, 4
 defined, 357
 derived ratios, 135–136
 DIAMONDS, 67
 digital certificates, 352
 dimensions
 defined, 357
 in OLAP, 51
 Direct Marketing Association
 (DMA), 350
 Direct Marketing Technology
 Inc., 241
 discard attribute, 207
 discounts, to encourage providing
 customer information, 21–22
 discretizing, 92–93, 133
 distribution networks, 268
 DNS (Domain Name System), 195
 domain attribute, 207
 Domain Name System (DNS), 195
 DoubleClick, 9, 252–253
 downtown residents category, 238
 drill up analysis, 42
 Dunn & Bradstreet, 244
 Dynamo Product Suite, 248

E

electronic commerce, 8–10. *See also*
 e-retailing
 the individual in, 72–73
 service as goal of, 73–74
 e-mail
 data collection with, 129
 marketing campaigns, 23–24, 274
 soliciting feedback via, 281
 Emerging Technologies Research
 Group, 43
 e-mining, 299–347
 classification/prediction in,
 326–331
 clustering in, 310–318
 customer profiling in, 305–309
 data capture in, 301–303
 data selection in, 305
 data sets, 309–310
 future of, 344–345
 goal of, 300
 as integrated iterative process,
 339–345
 optimization in, 332–334
 personalization in, 345–346
 prediction of consumer behavior
 in, 345–346
 process of, 300–301

sample website session, 303–304
 segmentation in, 319–325
 visualization in, 335–339
 employment profiles, 280
 ENDS (Equifax National Decision
 Systems), 239
 Engage Technologies, 253–254
 Enterprise Miner, 49, 185
 Equifax, 232, 239–241
 National Decision Systems
 (ENDS), 239
 e-retailing, 267–297
 attracting customers, 287–289
 business-to-business, 44
 buyer/seller relationship in,
 272–274
 consumer behavior in, 294–295
 consumer profiling in, 279–282
 customer interaction in, 268,
 269–272
 customer service and, 284–286
 data mining for, 290–293
 distribution and, 268
 effects of, 294–295
 effects of on buying/selling,
 269–272
 growth of, 43–44, 294–295, 297
 incentives in, 288
 integrating ordering and inven-
 tory in, 288–289
 keyword searches in, 275
 marketplace in, 275–277
 market segments in, 281–282
 modeling behavior in, 270–272
 personalization in, 282–283
 potential of, 270
 product as message in, 269–272
 product customization and,
 277–279
 products for, 294
 propensity to buy, deterioration
 of, 274
 real-time buying in, 282–283
 relationship with data
 mining, 268
 sale segmentation, 272–274
 security and, 282, 291
 storefront organization in,
 287–289
 targeted marketing in, 275–277
 error logs, 197
 European Data Protection Directive,
 352–353
 Evolver, 113
 Experian, 8, 232, 241–242
 expert systems, 84–85
 defined, 357
 exploratory analysis, 357
 extended log format, 198

F

Facets, 246
 factory and farm communities cate-
 gory, 237
 Fair Isaac, 8, 48
 Falcon, 160
 Federal Trade Commission
 (FTC), 352
 feedback
 customer loyalty and, 73
 soliciting via e-mail, 281
 FICO, 48
 file editors, data inspection
 with, 133
 file penetration, 243
 filtering
 collaborative, 9, 255, 259–260
 in customer profiling, 280–281
 Firefly, 9, 255
 Navigator Catalog, 255
 Passport, 255
 FIRM, 103
 fitness function, 107–108
 flat-file format, 35
 formats
 available in data mining
 tools, 138–139
 solution, 135–136
 forms
 customer data with, 218–221
 defined, 357
 information in, 34
 sample of, 220
 server processing of, 218–220
 Forrester Research, Inc.,
 43–44, 294
 fuzzy logic, 115–117
 defined, 357
 in OptiMark, 296

G

gain criterion, 77
 GainSmarts, 185
 Gartner Group, 43
 genetic algorithms (GAs), 9
 applications of, 113–114
 classification/prediction with,
 10–11
 crossover in, 109–110
 data mining tools using, 113
 defined, 358
 development of, 107
 discovery of, 85–86
 example of binary, 110–112
 genomes/chromosomes in,
 107–108
 mutation in, 110

- in NGO, 168–170
 - optimization with, 17–19, 112
 - optimizing back-propagation networks with, 334
 - optimizing electronic storefront design with, 332–334
 - output of, 107–108
 - prediction with, 112–113
 - product shelf display/location optimization with, 113–114
 - reproduction in, 108–109
 - resource requirements of, 142, 333
 - schemata in, 108
 - when to use, 137
 - genomes, 107
 - GET command, 196
 - globally unique identification cookies (GUIDs), 215
 - graphs. *See* visualization
 - graph theory, 66
 - GroupLens, 262–263
 - GT/Catalog, 256
 - GuestTrack, 256
 - GUIDs (globally unique identification cookies), 215
 - Gustos Software, 257
 - GVA, 67
- H**
- Hartigan, J. S., 103
 - HEAD command, 196
 - High Tech Connect, 244, 246
 - Hinton, G. E., 88
 - Hit List, 6, 59–60
 - HNC Database Mining Marksman, 159–161
 - Holland, John, 85–86, 107, 108
 - Hopfield, John, 85–86, 87–88
 - Hopfield Model, 87–88
 - host field, 195, 203
 - HTTP request field, 196
 - HyperText Transfer Protocol (HTTP), 204
 - defined, 358
- I**
- identification field, 195–196
 - ID3 (Interactive Dichotomizer), 63–64, 85–86
 - decision trees from, 100–101
 - development of, 99
 - in machine-learning algorithms, 98–99
 - when to use, 104
 - IDIS, 106
 - IF/THEN rules, 12–13
 - compared with decision trees, 103–104
 - as data mining output, 69
 - in e-mining, 321–323
 - generating decision trees from, 101–102
 - with machine-learning algorithms, 99, 101–102
 - IMAGIX-4D, 67
 - IMGIS, 258–259
 - import/export in data mining tools, 140–141
 - incentives
 - in e-retailing, 288
 - for providing customer information, 21–22
 - incompatibility, principle of, 116
 - IND, 106
 - independent variables, 4
 - individual identity profiles, 280
 - inductive analysis, 85–86
 - defined, 358
 - Information America, 244
 - information gain, 101, 104
 - information visualization. *See* visualization
 - INSPECT, 98
 - Intelligent Cross-Sell, 289
 - Intelligent Miner, 185–186
 - IntelliWeb, 261–262
 - Interactive Dichotomizer (ID3). *See* ID3 (Interactive Dichotomizer)
 - Intermind, 129
 - International Data Corporation (IDC), 43, 44, 56–57, 293
 - Internet, growth of, 43–44
 - Internet Engineering Task Force (IETF), 350
 - Internet Profiles Corporation (I/PRO), 257
 - Intersé, 60
 - intervals, 45–46, 80
 - identifying, 63–64
 - InText, 259
 - inventory, 9
 - controls, 33
 - IP addresses, 195
 - iVALS, 151–152
- J**
- Java
 - forms, 223
 - privacy issues and, 352–353
 - virtual cookies, 210–211
 - JavaScript
 - data component assembly with, 223–225
 - forms, 223
- K**
- KATE-Tools, 106
 - keys, visitor, 34–35
 - keywords
 - and product lines, 13–14
 - tracking with web analysis tools, 57
 - k-nearest neighbor statistical technique, 310, 358
 - KnowledgeSEEKER, 163–167
 - algorithms in, 103, 106–107
 - decision tree creation with, 164
 - decision tree settings in, 165–166
 - decision tree validation with, 165
 - output formats available from, 166–167
 - KnowX, 244, 245
 - Kohonen, Tuevo, 95
 - Kohonen networks. *See* Self-Organizing Maps (SOMs)
- L**
- Learning Intelligent software, 262–263
 - Le Cun, Y., 88
 - Lexis-Nexis, 244
 - Lifestyle Selector, 242
 - lifetime value (LTV), 45–46
 - appending psychographic data to, 80–81
 - determining, 77–81
 - LikeMinds, 259–260
 - linear models, 358
 - linear regression, 358
 - link analysis
 - defined, 66
 - drawbacks to, 67
 - systems, 66–67
 - LISR, 85
 - log file analyzers, 57. *See also* website analysis tools
 - LogFormat string configuration, 210
 - logistic regression, 358
 - loyalty
 - choice and, 36
 - customer, 45–46
 - feedback to foster, 73
 - measuring brand, 45–46
- M**
- machine-based learning, defined, 358
 - machine-learning algorithms
 - appropriate data for, 134
 - automatic segmentation with, 104–105

- machine-learning algorithms
 - (*continued*)
 - clustering ZIP codes with, 94
 - data mining tools with, 106–107
 - decision tree generation and, 63–64
 - discovery of, 85–86
 - learning websites with, 105–106
 - prioritization tools in, 139
 - for segmentation in e-mining, 319–324
 - segmenting data with, 98–99
 - for skewed data, 144–145
 - supervised learning in, 98–99
 - when to use, 137
- MAID, 103
- Marimba, 129
- market basket analysis, 310
 - defined, 358
- marketing. *See also* e-retailing
 - data providers, 228–247
 - strategy, 33
- market share, 47
- massively parallel processing systems (MPPs), 138
 - defined, 358
- MatchLogic, 260–261
- MATLAB, 49
 - NN Toolbox, 98
- Max-Age attribute, 207
- McKenna, Regis, 72
- McLuhan, Marshall, 269
- measurements, 139–140
 - in OLAP, 51
- memory management, 141
- META Group, 46–47
- Metaphor Mixer, 65
- MicroMass, 261–262
- Microsoft Personalization System (MPS), 215–217
- Microsoft Site Server, 289
- Microsoft User Properties Database Object (UPD), 217
- MicroVision, 239, 241
- MineSet, 186
- Minsky, Marvin, 85
- Model 1, 186
- ModelQuest, 98
- models/modeling, 30–32
 - analytical, defined, 355
 - checking accuracy of, 153–154
 - constructing, 145–147
 - linear, defined, 358
 - measuring accuracy of, 147
 - nonlinear, defined, 359
 - predictive, 10–11
 - in preplanning, 120
 - training, 326–331
- MOSAIC system, 228–229, 241
- multidimensional databases, 358
- mutation, 110
- N**
- name/value pairs, 220
- National Center for Supercomputing Applications (NCSA), 195
- server configuration, 200
- nearest neighbor technique, 310
 - defined, 358
- NeoVista, 8, 184
 - Solutions, 113
- net.Analysis Pro, 60
- NETdelivery, 129
- NetIntellect, 6, 60
- NetLine, 62, 257
- NETMAP, 67
- NetPerceptions, 262–263
- Netscape server configuration, 201
- NetStats Pro, 60
- NetTracker, 6, 60, 62
- neural networks
 - appropriate data for, 134
 - back-propagation, 10–11, 88–90
 - C code, 11
 - classification/prediction with, 10–11, 30–32, 97
 - current applications of, 88
 - in data mining tools, 97–98
 - data preparation for, 92–93, 97–98
 - defined, 359
 - determining purchasing propensity with, 94
 - feedback, 85–86
 - Hopfield's work on, 87–88
 - inputting date and time fields into, 93–94
 - objectives in using, 95
 - predicting purchases with, 326–331
 - SOMs, 15, 95–97
 - structure of, 86–87
 - supervised learning in, 90–91
 - training, 86–87, 91–92, 95
 - when to use, 98, 137
- NeuralWorks Predict, 98
- NeuralWorks Professional II/PLUS, 98
- NeuroGenetic Optimizer (NGO), 98, 113, 168–170
- noise, 142
- nonlinear models, 359
- nonresidential neighborhoods category, 238
- O**
- ODBC drivers, 35
- OMEGA Predictive Modeling System, 113
- The One-to-One Future* (Peppers, Rogers), 72–73
- online analytical processing (OLAP)
 - defined, 359
- online analytical processing (OLAP), 51–52
 - compared with data mining, 52–54
 - current tools for, 52
- online auctions, 282, 295
- Open Profiling Standard, 351–352
- Open Profiling System (OPS), 213–218
- operational data, 133
- OptiMark, 117, 296
 - customer profile in, 308–309
- optimization, 17–19
 - of electronic storefront design, 332–334
 - of neural networks with GAs, 112
 - of product shelf display/location with GAs, 113–114
- Orchestrate, 186
- outliers, 359
- output, 4
 - agent and referrer log, 199
 - formats for, 69–71
 - of genetic algorithms, 107–108
 - KnowledgeSEEKER, 166–167
- outsourcing, 144
- overfitting, 323
 - defined, 359
- P**
- packet sniffing, 57, 58
- Papert, Seymour, 85
- parallelism, 138
- parallel processing, 359. *See also* massively parallel processing systems (MPPs)
- Parker, D, 88
- Path attribute, 207
- pattern recognition, 5, 41–42
- Pattern Recognition
 - Workbench/Model 1, 187
- PSYCLE, 49
- Pearson's correlation, 50
- performance, of data mining tools, 141–142
- Perl, 221
- personalization, 345–346
 - of browsing, 9
 - customer control over profiles and, 37–38
 - determining extent of, 38
 - as future, 36–39

- of portals, 1–4
- with server software, 215–217
- planning
 - baselines, 121–123
 - determining data requirements, 74–75
 - identifying business objective in, 119–123
 - people involved in, 120–121
- Platform for Privacy Preference Project (P3P), 350–351
- Polk, 242, 244, 246
- polynomial networks, 10–11
- population of solutions, 107
- pop ups, 128
- portals, personalization of, 1–4, 346
- Port attribute, 207
- POST command, 196
- Preclass, 106
- prediction, 10–11, 30–32
 - defined, 359
 - in e-mining, 326–331
 - with GAs, 112
 - with neural networks, 97
- Preference Server, 259
- preprocessing, 139–140
- PressVu, 60–61
- Priceline.com, 295
- pricing, 33
- principle of incompatibility, 116
- prioritization, 139
- privacy issues, 20–21, 75, 131–132
 - policy on, 21
 - standards and consortiums in, 349–353
- PRIZM, 238–239
- process-modeling tools, 120
- process-models, 154
- products
 - availability of, 33
 - customization of, 277–279
 - for e-retailing, 294
 - lines and keywords, 13–14
 - as message in e-retailing, 269–272
 - shelf display/location optimization, 113–114
 - website visitors' relationship to, 13–15
- Profiler, 106
- ProfitMax, 160
- Proforma, 98
- Prolog, 85
- proof-of-concept software, 146
- PRW, 98
- psychographics, 72. *See also* demographic/household data
 - defined, 151
 - iVALS categories in, 151–152
 - VALS categories in, 151
- P3P (Platform for Privacy Preference Project), 350–351
- public record profiles, 280
- push, 129
- P\$YCLE, 239
- Q**
 - quantitative display, 67–68. *See also* visualization
 - query generation packages, 51, 53–54
 - Quinlan, J. Ross, 85–86, 99, 101
- R**
 - random sampling, 133–134
 - referrer logs, 194, 198
 - for keyword data, 203
 - server configuration for, 199–200
 - regression analysis, 48
 - relational marketing, 19–20
 - advantages of, 72
 - benefits of data mining for, 74
 - defined, 20
 - determining customer value and, 77–81
 - dialog in, 72
 - individual-based, 72–73
 - online coupons for research in, 76–77
 - reproduction, 108–109
 - retail design with GAs, 9, 332–334
 - retirement styles category, 235–236
 - Revenue Projection Model, 247
 - robustness, 142–143
 - rule generation, 359. *See also* IF/THEN rules
 - rules. *See* IF/THEN rules
 - Rumelhart, D. E., 88
- S**
 - SAS, 48, 49, 50
 - Enterprise Miner, 106
 - scalability, 137
 - scaling data, 97
 - Schank, Roger, 115
 - schemata, 108
 - Scorecard Builder, 182–183
 - search engines, personalization of, 346
 - Secure attribute, 207
 - security, 75, 131–132. *See also* privacy issues
 - and e-retailing, 282, 291
 - policy on, 21
 - segmentation, 11–13
 - analysis, 29–30
 - automatic, 104–105
 - compared with clustering, 29
 - customer composites with, 325
 - defined, 359
 - in e-mining, 318–325
 - with machine-learning algorithms, 98–99
 - sale, 272–274
 - SelectCast, 160, 249–250
 - SelectResponse, 249–250
 - self-organization, 28
 - self-organizing maps (SOMs), 15
 - calibrating clusters with, 96–97
 - for clustering in e-mining, 310–318
 - compared with back-propagation networks, 96
 - defined, 359
 - resource requirements of, 142
 - structure of, 95–96
 - when to use, 137
 - sensitivity analysis, 32, 140, 327
 - Sentry, 247
 - sequencing, 122–123
 - defined, 359
 - in e-mining, 310–318
 - when to use, 147
 - server log files, 194–198
 - agent logs, 194, 198
 - anthuser field, 196
 - error log, 197
 - extended log format, 198
 - host field, 195
 - HTTP request field, 196
 - identification field, 195–196
 - information in, 34
 - referrer logs, 194, 198–200, 203
 - status code field, 196–197
 - time stamp, 196
 - transfer logs, 194–195, 210–211
 - transfer volume field, 197
 - as website analysis tool, 55
 - servers
 - agent and referrer log output from, 199
 - analysis of, 5–6
 - Apache configurations, 199–200
 - CERN configuration, 201
 - configuring, 199–202
 - data mining technology incorporated in, 345
 - NCSA configurations, 200
 - Netscape configuration, 201
 - website configuration, 202
 - sessions, determining number of, 203
 - shipping resource allocation, 33
 - signatures, identifying, 5
 - Silhouette, 246
 - skewed data, 97

- solutions
 - formatting, 135–136
 - integrating, 149–153
 - population of, 107
 - validating, 148
 - SOMs. *See* self-organizing maps (SOMs)
 - S-Plus, 49
 - SpotFire, 65
 - SPSS, 49, 50
 - Neural Connection, 98
 - SQL (structured query language)
 - defined, 359–360
 - syntax, 70
 - SRI International, 151
 - S/S-Plus, 49
 - StarRater, 257
 - statelessness, 204
 - statistics, 48–49
 - compared with data mining, 50
 - descriptive, 48
 - popular tools in, 49
 - regression analysis, 48
 - traditional, 48
 - STATlab, 49
 - status code field, 196–197
 - StoryServer4, 263–264
 - strategic planning, 46–47
 - structured query language (SQL), 70, 359–360
 - SuperQuery, 106
 - supervised learning
 - data preparation for, 147
 - defined, 53
 - in machine-learning algorithms, 98–99
 - for neural network training, 90–91
 - targeting marketing with, 74–77
 - supply chain, 341
 - SurfReport, 61, 62
 - symmetric multiprocessing systems (SMPs), 138
 - defined, 359
- T**
- TCP/IP, 360
 - Temple-MVV, 67
 - THAID, 103
 - 4Thought, 98, 161–163
 - 3Dstats, 58
 - time fields, 93–94
 - time stamps, 196
 - TotalList, 246
 - traffic analysis, 6, 9
 - traffic monitoring, 57. *See also* website analysis tools
 - Trajecta, 98
 - transactional conditions, 12–13
 - transactional data, 131
 - profiling customers with, 330
 - transfer logs, 194–195
 - writing cookies to, 210–211
 - transfer volume field, 197
 - Trans Union, 232, 246–247
 - TREEDISC, 103
 - TrueCount, 260–261
 - TRUSTe, 349–350
- U**
- UltraEdit-32, 134
 - unsupervised learning
 - data preparation for, 147
 - defined, 53
 - in SOMs, 96
 - up and coming singles category, 235
 - UPD (Microsoft User Properties Database Object), 217
 - upscale households category, 234–235
- V**
- Values and Lifestyles (VALS) categories, 151
 - variables, 4
 - vCards, 351–352
 - Version attribute, 207
 - Vignette, 263–264
 - virtual reality markup language (VRML), 58
 - Visible Decisions (VDI), 65
 - Visual-ID, 65
 - visualization, 15–16, 65–68
 - defined, 360
 - in e-mining, 335–339
 - link analysis systems, 66–67
 - packages, 65–66
 - quantitative display, 67–68
 - VRML (virtual reality markup language), 58
- W**
- WalMart, 7–8, 44
 - WATSON, 67
 - webcasting, 129
 - WebInterests!, 259
 - webographics, 247–264
 - website analysis tools, 54–61
 - abilities of, 55–56
 - compared with data mining tools, 61–65
 - keyword tracking with, 57
 - list of current, 58–61
 - log file analyzers, 57
 - traffic analysis/monitoring, 57
 - websites
 - banners, 126
 - brand building, 124–125
 - buttons, 127
 - chat, 130
 - content of and data acquisition from, 73–74
 - cost of, 56–57
 - e-mail data collection on, 130
 - fine-tuning design of, 342–344
 - integrating data from with data warehouses, 149–154
 - measuring efficiency of, 289
 - optimization of, 17–19
 - pop ups, 128
 - push, 129
 - sample e-mining session of, 303–304
 - targeting ads on, 74–77
 - WebSite server configuration, 202
 - website visitors
 - determining purchasers, 10–11
 - determining type of, 11–13
 - ID numbers/keys for, 34–35
 - relationship of to products, 13–15
 - WebTrends, 6, 61
 - Werbo, Paul, 88
 - Williams, R. J., 88
 - WizWhy, 106, 171–172
 - World Wide Web
 - data mining tool evaluation copies on, 145–146
 - future of data mining on, 345
 - growth of, 43–44
 - as hyper-competitive market, 46
- X**
- XAID, 103
- Y**
- young mobile adults category, 236
- Z**
- Zadeh, Lotfi, 115–116
 - ZIP codes. *See also* demographic/household data
 - ACORN Consumer Classification System, 157–159
 - importance of in data analysis, 224
 - successful suburbanites, 157–158
 - training neural networks on, 94
 - wealthy seaboard suburbs, 157–158

Data Mining Your Website

J E S U S M E N A

"E-commerce promises the delivery of products and services for markets of one. Data mining is the enabler for making this happen. If you want to know who your customers are, what they want, and how to keep them, then you need to read this book."

—Andrew Braunberg, Senior Editor, *Data Mining News*

Turn Web data into knowledge about your customers.

This exciting book will help companies create, capture, enhance, and analyze one of their most valuable new sources of marketing information—usage and transactional data from a website. A company's website is a primary point of contact with its customers and a medium in which visitors' actions are messages about who they are and what they want.

Data Mining Your Website will teach you the tools, techniques, and technologies you'll need to profile current and potential customers and predict online interests and behavior. You'll learn how to extract insights into online buying patterns from the huge pools of information your website generates and how to apply this knowledge to design a website that better attracts, engages, and retains online customers.

Data Mining Your Website explains how data mining is a foundation for the new field of web-based, interactive retailing, marketing, and advertising. This innovative book will help web developers and marketers, webmasters, and data management professionals harness powerful new tools and processes.

- The first book to apply data mining specifically to e-commerce
- Learn effective methods for gathering, managing, and mining web customer information
- Use data mining to profile customers and create personalized e-commerce programs
- Understand the "10 Steps to Mining Your Web Data"

Jesus Mena is the CEO of WebMiner (www.webminer.com), an e-mining consulting company. He has worked with data mining and artificial intelligence technologies for over ten years and has mined banking, catalog, financial, government, retailing, telecom, and web databases.



Digital Press

An imprint of Butterworth-Heinemann

<http://www.bh.com/digitalpress>

