Glossary

a priori algorithm A data mining tool used in association analysis. It reduces the number of computations by dropping all combinations of attributes with less than a specified level of support.

addition rule Probability rule that states the probabilities of mutually exculsive events can be added together. More generally, the probability of either of two events occuring is $P(A \text{ È } B) = P(A) + P(B) – P(A \cap B)$, or the sum of the probabilities minus the probability of both events occuring together.

agglomerative A clustering approach that starts with the smallest level of items and combines them together to create groups. the opposite of divisive clustering.

aggregation Summarizing over rows of data. SQL aggregate functions include Sum, Count, and Average. Often used with a GROUP BY clause to compute subtotals. OLAP cube browsing displays aggregate totals interactively.

Akaike information criterion (AIC) A goodness of fit measure of forecast error based on the squared difference between observed and predicted Y values. In time series analysis, it measures the error from the autoregressive component and becomes smaller as the AR variance decreases. It is used to compare various models. In general, AIC= 2k + n*ln(residual sum of squares/n), where k is the number of parameters and is the likelihood value. Also see Schwarz criterian.

antecedent The left side of a rule in the statement of an association rule. Typically read as indicating that purchases of the items on the left side lead to purchases of items on the right (consequent) side.

arrangements The number of ways of arranging a set of items. Used for determining probabilities in terms of relative frequency. The number of ways of arranging n items if there are no duplicates is simply n! (n factorial).

association rule A relationship determined from the data that indicates which events or purchase items occur together. Typically written: antecedent -> consequent, the left side implies or indicates that the right side event happens with some estimated level of

attribute relationship Hierarchical probability dimensions can be created with a hyper cube browser. In Microsoft's system, the attributes within that hierarchy should be connected via defined relationships to specify the various levels. For example, a common date hierarchy runs: Date -> Month -> Quarter -> Year.

auto regression In time series analysis an auto regressive relationship is a relationship between the variable in the current time versus lagged time periods: $Y_t = a_0 + a_1 Y_{t-1} + a_2 Y_{t-2} + \ldots + a_p Y_{t-p} + \varepsilon_t$.

auto regressive integrated moving average (ARIMA) A time series estimation technique from Box and Jenkins that estimates coefficients for the auto regression (AR) and moving average (MA) components. The integrated (I) element constitutes the removal of the trend through differencing.

autocorrelation function (ACF) A chart used in time series analysis to help determine the number of lags to use for the moving average component of an ARIMA model. It plots the auto correlation values for each lag value. If the ACF cuts off at a specific lag while the PACF dies down, the cutoff point is a good estimate for the maximum MA lag.

auto-regressive tree with cross prediction (ARTxp) Microsoft's primary time series analysis tool that uses a decision tree approach and cross correlations to predict time series values. The decision tree component identifies break points within the data that have different effects on the time series. An ARIMA model is mixed with the decision tree to predict the series farther out in time.

Bayes' theorem In simple form: $P(A \mid B) = P(B \mid A)/P(B)$. When B is a compound event consisting of many similar events, the denominator is written $\sum P(B \mid A=i) P(A=i)$. The theorem is commonly used to find values when events are sequenced or some data is unavailable. Knowing only the probabilities $P(B \mid A=i)$, it is possible to compute the reverse $P(A \mid B)$. The rule is easiest to see with a decision tree or contingency table.

Bayesian information criterion See Schwarz criterion.

Bayesian probability A way of looking at probability and statistics where probability is subjective and Bayes' theorem is used to update the estimate of probabilities. Rewrite Baye's Theorem as: P(A | B) = P(A) P(B | A) / P(B). P(A) is the initial or a priori estimate and P(A|B) is the posterior probability after the information from event B has been incorporated.

BETWEEN A portion of a SQL statement used to specify a lower and upper bound in a WHERE clause. Commonly used for dates, such as OrderDate BETWEEN 01-Jan-2010 AND 31-Dec-2010.

Boolean algebra Mathematical analysis often applied to query conditions that focuses on the role of AND, OR, and NOT connectors. Complex conditions require thought and testing to ensure all conditions are defined correctly.

Bootstrap A process of expanding limited data to more cases enabling small data sets to be used for more complex analysis. The initial data is treated as a distribution and new samples are generated by randomly drawing data from that sample distribution. It is a common practice for small samples, but it is always better to obtain more data if possible. The small sample might not accurately represent the true data.

Business Intelligence Development Studio Microsoft's client tool used to define analyses for data mining and to create data cubes for browsing. It runs in conjunction with Visual Studio and can be installed from the SQL Server installation process.

calculation Computations on data—commonly performed within a query to operate on data contained within one row of a table.

categorical attribute A discrete measure, often written as text categories, such as gender. When tools require numeric data, CASE or IF statements can be used to assign numbers to each category value. Clustering is difficult with categorical or nominal values because distance measures are arbitrary.

causality A relationship specified in a model where one event causes a second event to happen. The second event always occurs as a result of the first one. In comparison to correlation, correlation is an observed relationship that two variables appear to be related, but it does not mean that causality exists. Commonly, the two variables might both depend on a third variable.

central limit theorem The fundamental theorem in statistics which states that the distribution of the average for any random variable approaches the normal distribution with a sufficiently large number of observations.

chaotic A series or events that exhibit strong variations. In particular, from chaos theory, small changes in independent variables can result in large, possibly discontinuous jumps in the dependent variable.

classification The act of placing data into classes or categories. Typically based on attribute values and usually created via a logistic-type regression, neural network, or decision tree. For example, classify customers on the basis of payment history.

classification matrix A Microsoft tool to view the accuracy of classification tools. It compares the actual number (column) to the predicted number (row) of items placed into each category based on the current model. Cells on the main diagonally contain counts of the correctly predicted classes.

CLUSTER_COUNT An option parameter in the Microsoft clustering tool that enables the analyst to specify the number of clusters to be estimated. A value of zero (0) tells the routine to heuristically find the appropriate number of clusters.

clustering A set of data mining techniques that attempt to define groups or clusters of data based on attribute values. The goal is to identify groups that have small distances from other members of the group relative to larger distances to other groups.

CLUSTERING_METHOD An option parameter in the Microsoft clustering tool that enables the analyst to specify the particular clustering method to be used. The choices are 1=Scalable EM (default), 2=Non-scalable EM, 3=Scalable K-means, and 4=Non-scalable K-means. The non-scalable options can be used only with small numbers of observations, so the main choices are 1 and 3 between the EM and K-means algorithms.

column In a database table, a column represents an attribute or dimension of the data object. For instance, Name and Phone would be typical columns in a Customer or Employee table.

combination The number of arrangements of a set of items when k items are pulled from a set of n. The specific ordering of the items does not matter. C(n, k) = n! / (n-k)! k! This term matches that of the binominal distribution. The Excel function is Combin(n, k).

combinatorial search A clustering search method that tries all combinations of points to determine the best clusters. It is an expensive and time-consuming approach but can obtain the most accurate values. Most K-means cluster methods use combinatorial searches for at least part of the solution.

comma-separated values (CSV) A relatively standard method of storing data in flat files for transfer to other systems. Data for one observation is stored in a single row in the file. Within the row, data for the columns are separated by commas—although most tools have options to specify the delimiters and separators.

conditional probability The probability of some event happening given that another event has already occurred. Written P(A | B) and read as the probability of A given B. It is easiest to see using a contingency table or decision tree. Mathematically, $P(A \mid B) = P(A \cap B) / P(B)$.

confidence Used in association analysis, confidence is a measure of the probability that a rule is true. In probability terms for a rule A -> B, confidence is the conditional probability P(B | A).

consequent The right-side of an association rule, or implication event that is being predicted. A -> B makes B the consequent event based on the antecedent A. Association analysis estimates the probability of the rule.

contingency table A two-dimensional table used to display the count of the observations for two types of events (rows and columns). The values within a given cell are used to compute the joint probabilities P(A and B). The margin totals show the probabilities of each specified event.

continuous data Continuity means data can take any value and there are no gaps or jumps between values. Measures such as weight, height, or distance are common forms of continuous data. Even though a measuring device lacks infinite resolution, the underlying data could take on any value. Measures such as time could be continuous, but are sometimes handled as discrete data—depending on the desired model and the measurement method. If time is measured in seconds or fractions of a second, it is likely to be continuous. If it is measured in months or years, it might be handled as a discrete variable. The point is that the choice of continuous or discrete data is dependent on the problem and the model.

correlation An observed relationship between variables. If variable Y increases when X increases, it represents a positive correlation. Correlation between two variables is measured with the simple correlation coefficient. Regression techniques are used to measure correlation across several variables. Correlation is a statistical observation, it does not imply that causation exists. However, it can sometimes be used to test the validity of a theoretical model that explains causation.

correlation coefficient A measure of the correspondence or association between two variables. It essentially computes the variation of the two variables compared to the standalone variance of each variable.

critical value In hypothesis testing, the critical value is the point at which the null hypothesis is rejected. The value is found from the distribution function or tables. With standard normal data, common critical values are 1.96 for a two-tailed test at 5 percent error, and 2.58 for a two-tailed test at 1 percent error. In Excel, NormSInv(probability/2) and TInv(probability, degrees of freedom) provide the critical values for the standard normal and T distribution.

cross correlation A relationship between two time series. One series might cause changes in second (positive or negative), or the two series might be related to a third series. If one series is cross correlated with a second and the second one is easier to predict, the original series can be forecast.

cross join In a query, a cross join matches every row in the first table with every row in the second table. If the first table contains

m rows and the second contains n rows, the resulting query will contain m * n rows. Across join is created in SQL when no JOIN condition is listed. Cross joins should be avoided except for instances with a small number of rows.

cross validation Testing a model against multiple subsets of the data. Commonly achieved by splitting a model into k non-overlapping sets and estimating the same model on k-1 sets—rotating the dropped set until all k combinations have been estimated.

cross-support A problem that can arise in association analysis, when a market basket contains one item from a high-frequency set and one from a low-frequency set. If item A appears in almost all baskets, any other item that appears could be a random event, yet the computed confidence values will make it appear to be important.

cube browser A software tool to enable managers and analysts to explore data summaries commonly computed within a hyper cube. Most cube browsers highlight summaries in tables using two dimensions (rows and columns). Users can drill down to see details within the subtotals, or roll up the totals to see summaries at a higher level within a hierarchy. Options exist to slice the cube to display rows that meet specified conditions.

cumulative distribution function (cdf) The function $F(x)$ that returns $P(X \leq x)$. For discrete data, it is the sum of the probabilities up to x. For continuous distributions, it is the integral of the probability density function.

curse of dimensionality Many data mining tools are hard to solve when the number of dimensions or attributes is large. Clustering with K-means and association analysis are two main examples, but similar problems arise with most tools. Often, the only solution is to reduce the number of dimensions; such as examining purchases by categories (such as soda and crackers) instead of detailed items (such as Coke, Diet Coke, Pepsi, and so on).

Cyclical variation In time series, variations or patterns that depend on the economic cycle. For example, high points of the cycle represent higher personal income which can lead to greater sales. Requires knowledge of the business cycle to estimate—typically with a series data for gross domestic product or income.

data The fundamental element to be analyzed. Typically a measure of either an attribute or a fact. Data can be numeric or categorical. In most systems, a row of data represents a single observation with columns representing measures of various attributes and facts.

data associations Events or situations that tend to happen together. Market basket analysis is a classic situation, where the goal is to identify items purchased together. But the association concept is general and can be useful for any events.

data definition A set of commands that are used to define data, such as CREATE TABLE. Graphical interfaces are often easier to use, but the data definition commands are useful for creating new tables with a program.

data manipulation A set of commands used to alter the data. The most common commands are UPDATE, INSERT, and DELETE.

data mining The process of using analytical tools to scan large data sets for patterns and provide insight to analysts and managers. The process emphasizes exploration of the data. Some people differentiate between data mining, business intelligence, and business analytics; but the three terms represent aspects of the same concepts. Machine or statistical learning can also be components, where data is used to train a system to identify categories and patterns so these can be used to make future decisions. Data mining typically includes tools to analyze data as well as to search and report data.

Data Source The connection string that defines a link to a source of data in Microsoft Business Intelligence Studio. Each analysis begins by defining at least one data source. Multiple data sources can be created to connect to different databases.

Data Source View In Microsoft Business Intelligence Studio, it defines the tables and named queries that can be used in data analysis. At least one data source must be created first to define the connection to a database. Views are created using SQL syntax, but tables can be pulled from multiple data sources.

data type A type of data that can be held by a column. Each DBMS has predefined system domains (integer, float, string, etc.). Some systems support user-defined domains that are named combinations of other data types.

data warehouse A copy of transaction data stored for high-speed searching, summarizing, and analysis. Special tools, typically including many complex indexes, are often used to store the data. Bulk uploads are generally used to update the data.

database A collection of data stored in a standardized format, designed to be shared by multiple users. A collection of tables for a particular business situation.

database management system (DBMS) A tool to efficiently store and retrieve large amounts of data. Many DBMSs are based on the relational data model which stores data for entities in tables. Rows of data represent a single instance of the entity (such as customer), and the columns identify attributes of the object, such as name and phone number.

decision tree A method of examining data and a tool to classify data into a tree. Each node of the tree contains a conditional statement and represents a split point for the data. For example, one node might test the gender of participants, resulting in three branches from that point: Female, Male, and Other. Decision tree tools attempt to build trees that have significant split points where the contribution of each branch to the goal is different.

dendrogram A graphical display of clusters created with hierarchical clustering. Often used in chemistry, the chart shows various levels of clusters. The bottom level contains the most number of clusters, the top contains the fewest clusters.

dependent variable A variable or attribute that responds to changes in the values of the independent variables.

DESC The modifier in the SQL SELECT … ORDER BY statement that specifies a descending sort (e.g., Z … A). ASC can be used for ascending, but it is the default, so it is not necessary.

digital dashboard Also called digital cockpit or executive information system. A graphical way to display selected data items using gauges, icons, and color coding. The key performance indicators are selected by managers to highlight changing data that affects goals and progress critical to making decisions. The tools include links to details to explore the underlying data.

dimension One attribute or characteristic of an object in a hyper cube. Determining relevant dimensions is an important step in designing a hyper cube and configuring data analysis.

discrete data Data that takes on specific values, but possibly an infinite number. For example, the set of integers is discrete. In many cases, the choice of discrete or continuous data depends on the problem and the model being used.

discretizing The process of converting continuous data into discrete categories. Some tools require discrete data, so categories or bins can be defined by specifying ranges of data. For instance, all people less than 18 versus people 18 or older. The ranges can be defined based on external factors or clusters can be used to find groups of ranges.

distance Most important in cluster analysis, a distance measure is used to determine how far one point is from another, or from the center of a proposed cluster. For numerical attributes, distance can be a common Euclidean measure, $(x\text{-}y)^2$. For multiple attributes, the individual squared differences are summed—giving equal weight to each measure. If attributes have highly different scales, the data might be scaled first. Distance between categorical data is often arbitrary, but is sometimes reduced to zero or one.

divisive clustering A top-down approach to clustering where the top node contains all of the data elements. At each level, the algorithm divides the existing cluster into two new clusters. Typically, the algorithm takes the point that is farthest away from the existing center and then determines which points are closer to the new point than to the existing center.

drill down Expanding a point of data to explore the details. A method associated with data hierarchies to examine data at a lower, more-detailed level within the hierarchy. The opposite of roll up.

drill through An option within many SQL Server analytical tools that enables users to select a result and obtain more detailed data for the item.

dummy variable A variable that is assigned discrete values (often zero and one) to represent various events or characteristics. For example, a variable Fall could be defined as 1 for the fall

months and 0 for others; then used to estimate seasonal variations. Be careful when adding multiple dummy variables to a problem because they can lead to exact multicollinearity. For example, dummy variables for Fall, Winter, Spring, and Summer would cover all of the cases for a year and all four cannot be used if an intercept term is also used.

eigenvalue In the mathematics of linear algebra, it is a scalar value λ such that **A X** = λ **X**, where **A** is a square matrix and **X** is a vector of real numbers. Sometimes written as **A X** – λ I **X** =0, where I is the identity matrix. Eigenvalues and the corresponding eigenvectors are used to decompose a matrix into simpler elements. In decision statistics, the result is known as principal components.

elasticity Percent change in dependent variable (Y) divided by percent change in independent variable (X). If the slope is known, E = dY/dX (X/Y). A convenient way to display change data without the dimensions so values are comparable regardless of the underlying data.

enterprise resource planning (ERP) An integrated computer system running on top of a DBMS. It is designed to collect and organize data from all operations in an organization. Existing systems are strong in accounting, purchasing, and HRM.

Euclidean distance The most common distance measure: d = sqrt(sum($(p - q_i)^2$)). Using the squared terms, the distance measure is always non-negative. It is the standard measure taught in basic geometry.

expectation maximization (EM) One of the two main algorithms to identify clusters in data. K-means is the other. Responsibility functions are defined for each potential cluster based on the relative pdfs. The algorithm computes the responsibility values for each point to assign numbers for both potential clusters. The algorithm creates a soft assignment by allowing a point to belong to more than one cluster—as defined by the responsibility values. The clusters are defined in terms of the means and standard deviations.

expected value For discrete data, $\sum$ p(x) x. For continuous data, $\int$x p(x). The mean of the distribution, or the average that would be expected after a sufficiently large number of trials. Typically written E(X).

experiment A set of events or trials defined on a sample space. Clinical experiments often involve controlled environments where effects of external factors are minimized. Social or business experiments typically measure external variables and estimate the impact of those variables as well as the control variables.

extraction, transformation, and loading (ETL) The process of transferring data from an existing database into a data warehouse. The three steps generally need to be automated so that data can be cleaned and inserted into the data warehouse automatically on a defined schedule. Data warehouse products typically include tools to automate the ETL tasks.

fact An attribute of the data that can be measured and used within a hyper cube to compute summaries. Facts are specified by managers to define concepts of interest.

factorial Given a positive integer n, its factorial or n! is n*(n-1)*(n-2)*…(1). The Excel function is Fact(n). The gamma function is sometimes interpreted as a factorial function for real and complex numbers. Γ(N) = (n – 1)!

forecasting The process of analyzing data to predict values for future or hypothetical situations. Forecasting is often based on models where parameters are estimated from existing data, or on time series analysis which is used to predict future values based on trends and seasonal variations.

FROM The SQL SELECT clause that signifies the tables from which the query will retrieve data. Used in conjunction with the JOIN or INNER JOIN statement.

gap statistic A method of heuristically selecting the number of means (clusters) to use when clustering data. The number of clusters begins at K=1 and the total within-cluster variance is computed. As K is increased, this value drops. Plotting the total value against K often reveals a break point, presumably indicating the natural number of clusters to be used. Various gap statistics have been defined to formalize this change and highlight the number of clusters to choose for the K-means process.

general multiplication rule The method of multiplying probabilities when events might be interdependent: P(A ∩ B) = P(A | B) *P(B).

goodness of fit A method of testing how closely two distributions match each other. The distribution values are split into J categories. The number of observed observations within each category are counted and compared to the number of observations expected to fall within each category. The statistic $X = \sum (O-E)^2/E$ has a chi-square distribution with J-1 degrees of freedom. If the sum is too high, the null hypothesis of equal distributions is rejected.

GROUP BY A SQL SELECT clause that computes an aggregate value for each item in a group. For example, SELECT Department, SUM(Salary) FROM Employee GROUP BY Department; computes and lists the total employee salaries for each department.

HAVING A SQL clause used with the GROUP BY statement. It restricts the output to only those groups that meet the specified condition.

hierarchical clustering A cluster approach that begins with a single cluster and repeatedly divides clusters to compare the results at various numbers of clusters. A dendrogram is often used to show the results for multiple cluster levels.

hierarchy A set of levels that are used to explore data summaries. Natural hierarchies include dates (year, quarter, month, day) and location (continent, nation, state, city). Hierarchies can also be defined for specific circumstances, such as product groupings or employee/managerial levels.

hybrid (HOLAP) A method of storing data for hyper cubes. The base data is stored in relational tables and aggregated totals are stored in a data warehouse. In general, performance is similar to that of the ROLAP model. Compare it to the MOLAP approach.

hyper cube A method of summarizing data, used both to store data for high-speed retrieval and to browse summarized data. A hyper cube represents subtotals across multiple dimensions, where each dimension is one side of the cube. For example, a cross-tabulation is a two-dimensional cube that contains subtotals for dimensions by rows and columns. A three-dimensional cube would add a depth dimension. Hyper cubes can have any number of dimensions, but most tools focus on displaying values for two dimensions at a time in a table.

hypothesis testing The statistical process of evaluating data results. A null hypothesis is defined for a neutral state (such as a coefficient equal to zero). An error rate is specified for a Type I error (the probability of rejecting the null hypothesis if it is true)—often set at 5 or 1 percent. The test statistic is defined. When the experiment is conducted, the null hypothesis is rejected if the observations are improbable given the null hypothesis.

identity An option in SQL Server where the DBMS automatically generates sequential numbers to insert into a primary key column in a database table.

immediate if function (IIF) A DMX function used in SQL Server cubes to return values based on a condition. Useful for setting colors or creating categorical data. For example, color properties can be set using IIF([Measures].[Discount] > 2000, 255, 0). The function has three parameters: 1) the condition, 2) the value to return if the condition is true, and 3) the value to return if the condition is false.

Independent events Events are independent if they are not directly affected by each other and their probabilities do not influence each other. $P(A \cap B) = P(A) * P(B)$.

independent variable A variable or attribute that is usually controllable and changes in values affect the dependent variable.

index A small file that is sorted and provide fast access to data through key values. The sorted dimension can be searched quickly with B-Tree tools and the index points to the matching data. Multiple indexes dramatically reduce data retrieval times, but they slow down data inserts and updates. Hence, data mining tools often use a data warehouse which contains a heavily indexed copy of transaction data.

information Data that has been organized and put into a meaningful context. Information is used to make decisions. For example, information can be the answer to a question, or the results of an analysis that leads to a decision.

information measure From Shannon, sometimes called Shannon's entropy: $H(X) = E[I(X)]$ where $I(X) = \log(1/p) = -\log(p)$. It is a measure of the surprise value of data. It is highest for uniformly random data—because

there is no way to predict which value might arise.

interestingness A big question in association analysis and data mining in general. Correlations and associations can be found statistically, but the results might not be interesting or useful. Dozens of measures of interestingness exist—usually related to the surprise value of the information—but ultimately, decision makers need to evaluate results for potential value.

itemsets Combinations of attributes or products. A simple itemset consists of a single item, but association analysis and other tools often consider multiple combinations of items.

JOIN When data is retrieved from more than one table, the tables must be joined by some column or columns of data. See INNER JOIN and LEFT JOIN.

joint events Two or more events occurring together.

joint probability The probability of two events occurring together: P(A and B) or P(A ∩ B). Using the general multiplication rule, P(A ∩ B) = P(A | B) *P(B).

just-in-time A manufacturing technique that relies on minimal inventories, instead relying on vendors and subcontractors to provide components just in time to be assembled. The method requires detailed communication with vendors.

key performance indicator (KPI) Variables that represent critical data to decision makers. The definition of a KPI typically includes the attribute measure, a trend over time, and a function to convert the measure into a graphical view—either by scaling or by categorizing the data.

K-means One of the two main algorithms to identify clusters in data. Expectation maximization is the other. The algorithm begins with a target of identify K-clusters. The goal is to find the best way to split the data to assign each point to a single cluster. In raw form, the process compares each point to the K clusters computing the distance to minimize the variance within each cluster. It can be a slow algorithm for large data sets.

Knowledge A higher level of understanding, including rules, patterns, and decisions. In an ideal system, data leads to information which leads to knowledge.

lift A measure of impact of a rule or result. In association analysis, lift is often defined as P(B|A)/P(B). This ratio measures the probability of item B being purchased with the rule (A already chosen) versus without the rule—B by itself.

linear regression A data mining tool that is a classic statistical research method. Regression has a dependent variable and several independent variables and determines coefficients that fit the best line to the data points. The process estimates coefficients of the equation: $Y = b0 + b1\ X1 + b2\ X2 + \ldots + bkXk$. In effect, the coefficients identify the strength of the variables in how they affect the dependent variable. The dependent variable must be continuous.

Logical Primary Key Used within Microsoft's Data Source View, a named query should be assigned a logical primary key which acts similarly to a primary key in database design. The selected columns uniquely identify each row within the query.

logistic regression A data mining tool similar to linear regression but the dependent variable is categorical or discrete. It is typically used for classification problems. The method estimates a function which determines the probability of each Y-outcome.

margin totals In a contingency table, the totals of the observations or probabilities computed for a given row or column observation. Often written in the margin, the total represents the probability of the specified event occurring regardless of the value of the secondary event.

market basket A collection of items purchased at the same time. Association analysis can identify specific items that are commonly purchased together.

maximum likelihood estimator (MLE) A method of estimating coefficients for a variety of tools for evaluating dimensions. It is often a choice parameter in how models are estimated. It is one of the more robust estimation methods, but sometimes can be slow.

mean absolute deviation (MAD) A method of measuring distance. The traditional Euclidean measure uses a squared difference, MAD uses the absolute value of the difference. Euclidean places a stronger emphasis on outliers than does MAD.

measure An attribute that evaluates some fact or dimension of the problem. In most systems, it must be a numeric attribute that can be subtotaled or averaged.

metadata Literally, data about data. The explanation or documentation of data items. Simple metadata includes the data type and name. More complex metadata includes descriptions, source information, ownership, and security conditions.

minimum confidence In association analysis, the cutoff level for the confidence measure used to determine whether a rule should be displayed. It is usually a secondary measure and the levels can be changed interactively to increase or decrease the number of rules displayed.

minimum support In association analysis, the cutoff level for evaluating potential rules. Itemsets that fall below the specified level are dropped from further consideration. Setting the level too high can result in no rules that meet the condition. Changing the level typically requires reanalyzing the data.

mixture model The mixture model is the underlying method used in the EM clustering approach. Each cluster is assumed to have some unknown distribution and a given point can be assigned to multiple clusters by a linear combination of the probability functions. The linear coefficients essentially determine the percentage of weight assigned to each cluster.

model A simplification of reality, and an attempt to describe the interrelationship and causality between variables. Models typically are built from theory. Estimates based on models have more power and validity than basic statistical observations.

moving average In time series analysis, the estimation of the coefficients for the lag effects of the error terms. An average is computed across each specified interval, such as MA(3) which can be computed as (Y0+Y1+Y2)/3, and then shifted forward one time period to compute (Y1+Y2+Y3)/3, and so on. In the ARIMA model, $Y_t = \mu + \psi_0\varepsilon_t + \psi_1\varepsilon_{t-1} + \psi_2\varepsilon_{t-2} + \psi_3\varepsilon_{t-3} \ldots$.

multicollinearity A problem that arises when many attributes or dimensions attempt to measure the same thing. Perfect multicollinearity arises when a collection of attributes can be written as linear combinations of each other. Most commonly encountered in regression analysis with too many similar attributes. In most cases, the easiest solution is to drop some of the attributes from the analysis.

multidimensional expressions (MDX) A language initially created by Microsoft to define data cubes and configure data analysis. Several vendors support MDX. Most tools generate MDX automatically, so tools can be used without knowing the detailed syntax. It is useful for performing calculations.

multidimensional OLAP (MOLAP) A method of storing data in a data warehouse for hyper cubes. The data is cleaned and aggregations are pre-computed where possible. Joins and indexes are prebuilt, leading to some duplication. It is often the fastest method to retrieve data. Compare to ROLAP.

multiplication rule If two events A and B are independent, the joint probability can be computed with a simple multiplication of the two separate probabilities: $P(A \cap B) = P(A) * P(B)$. For example, with a fair die and random throws, the probability of obtaining any specific number is 1/6, so the probability of throwing "snake eyes" or two ones is (1/6)(1/6) or (1/36).

mutually exclusive Two events that cannot happen together and have no common outcomes. $P(A \cap B) = 0$. Commonly used when creating discrete, non-overlapping categories.

naïve Bayes A data mining tool based on Bayes' theorem. The goal is to determine which attributes have the strongest effect on a dependent variable. The method works with minimal supervision and is robust so it works well as an initial perspective on the data. The dependent and independent variables need to be discrete data.

named calculation Created within a table in a Microsoft data source view, a named calculation computes row-by-row values. It is similar to adding a computed column within an SQL query. By operating within the data source view, the named calculation can include data from multiple underlying sources.

named query Created within a Microsoft data source view, a named query can combine data from multiple sources and perform calculations equivalent to those within an SQL query. By operating within the data source view, the named calculation can include data from multiple underlying sources.

neural network A collection of artificial neurons loosely designed to mimic the way the human brain operates. Especially useful for tasks that involve pattern recognition. The technique is one of the main machine learning methods and can run with minimal supervision. Effectively, the technique estimates a nonlinear relationship between input and output variables.

nominal dimension A categorical attribute. In clustering, any distance measure of a categorical attribute is nominal because it is an arbitrary assignment.

normalization The process of creating a well-behaved set of tables to efficiently store data, minimize redundancy, and ensure data integrity. See first, second, and third normal form.

NOT The SQL negation operator. Used in the WHERE clause to reverse the truth value of a statement.

one-to-many A common relationship among database tables. For example, a customer can place many orders, but orders come from one customer. As part of the design normalization process, many-to-many relationships are split into two one-to-many relationships.

online analytical processing (OLAP) A computer system designed to help managers retrieve and analyze data. The systems are optimized to rapidly integrate and retrieve data. The storage system is generally incompatible with transaction processing, so it is stored in a data warehouse. A hyper cube browser is a common way of examining data summaries.

online transaction processing (OLTP) A computer system designed to handle daily transactions. It is optimized to record and protect multiple transactions. Because it is generally not compatible with managerial retrieval of data, data is extracted from these systems into a data warehouse.

ORDER BY The clause in the SQL SELECT statement that lists the columns to sort the output. The modifiers ASC and DESC are used to specify ascending and descending sort orders.

order of operations From mathematics, calculations are performed in a standard sequence. For example, multiplication is performed before addition. The order can be altered through the use of parentheses. The order becomes critical when computing values using a hyper cube. For example, dividing or multiplying data columns using a calculation on a cube rarely does what you might expect. Subtotals are computed first followed by cube calculations. Computations must often be made within queries or named calculations to be executed at the row level instead of the cube level.

ordinal measure A ranking such as 1, 2, 3. In clustering, distance is commonly defined by converting to a centered percentage: $v = (i - 1/2) / M$, where M is the highest value.

Orthogonal In geometric terms, it means perpendicular lines. In statistics, two orthogonal components are linearly independent. In principal components, the goal is to find orthogonal factors to describe the data with a smaller number of dimensions.

over fitting A classic problem with data mining and statistical testing in general. Given a set of observations, you could repeatedly build models for that data so that the sample data can be exactly explained by the model. But, the model could fail miserably at predicting the underlying population events because the model is tailored too closely to the specific sample. A simple example might consist of three data points. A quadratic curve could exactly fit those three points, but predicted values from the model might be terrible. The standard solution is to obtain large data sets and withhold part of the data to use to test the model.

ParallelPeriod A useful function within DMX, it is used to retrieve data from prior time periods for the same data attribute. For example, ParallelPeriod ([Calendar].[Year]. [Year], 1, [Calendar].[Year].CurrentMember) retrieves data from the prior year. The function works for the currently aggregated data, such as Week, Day, Month, or Quarter.

parameter A variable in a model or distribution that has specific meaning. The

parameters are estimated from the sample data to define the exact shape of the distribution. For example, the normal or Gaussian distribution has two parameters: μ and σ that represent the mean and standard deviation of the distribution and tailor it to the specific data.

partial autocorrelation function (PACF) A chart used in time series analysis to help determine the number of lags to use for the moving average component of an ARIMA model. It plots the partial auto correlation values for each lag value. If the PACF cuts off at a specific lag while the ACF dies down, the cutoff point is a good estimate for the maximum AR lag term. It is the partial autocorrelation because it estimates the autocorrelation between Yt and Yt+k, with the intervening terms removed.

permutation The number of ways of arranging a set of items when some of them are not included. For example, the number of ways of selecting 3 items from 20 is 6840. $P(n, k) = n!/(n-k)!$ In Excel Permut(n, k). With permutations, each ordering is considered to be different (A, B, C is different from B, A, C). If ordering does not matter, use the formula for combinations.

perspective A defined view of the data in a Microsoft hyper cube. Multiple perspectives can be defined on any cube to limit the data available to one group of users.

Poisson distribution A distribution for discrete data often used to estimate the number of events occurring during a fixed period of time. $P(X = k) = (e^{-\alpha}\alpha^k)/k!$ where the parameter α would be the average number of arrivals expected during the time period and the arrival times are independent of the last event.

posterior distribution In a Bayesian approach with subjective probabilities, it is a resulting distribution that was improved through information obtained in an experiment.

prediction A forecast based on a model estimated from observations, which requires forecast estimates of the independent variables. Predictions can also be made from time series analyses that are formed based on trends and seasonal variations.

PredictTimeSeries Microsoft function A function in DMX that is used to compute predicted values from a time series. The model must already be built and run. The function takes two parameters: the name of the column and the number of periods to be forecast. The text has examples for a simple forecast and a forecast with the standard deviation. The basic format is: SELECT FLATTENED PredictTimeSeries([Column Name], 12) As Forecast FROM [Model Name].

primary key A column or set of columns that identify a particular row in a table.

principal components analysis (PCA) A method to identify the primary orthogonal factors that identify a set of data. The factors are listed in descending order of the percentage of variation explained by each factor. The goal is to describe the data with a smaller number of dimensions.

prior distribution In a Bayesian approach with subjective probabilities, it is the initial probability distribution. The probability and distribution are improved through the addition of information.

probability (1) The relative frequency of some event occurring. (2) A subjective belief about the chance of some event occurring. The first definition is the most common; the second is the foundation of the Bayesian approach. Some basic rules must hold for probabilities: (a) $0 \le p \le 1$, (b) P(A or B) = P(A) + P(B) if A and B mutually exclusive, and (c) the sum of the probability of all events must be one.

probability density function (pdf) For continuous data, the probability of any specific point x is zero, so the density function is defined in terms of the cumulative probability $P(X \le x)$. The cumulative probability function is the integral of the pdf: $F(x) = \int f(x)dx$.

probability distribution For discrete data, the listing of the event x and its associated probability function p(x).

probability function $P(X = x_i)$ for discrete data—the assignment of a probability number to each event. Equivalent to the probability mass function or probability density function for continuous data.

probability mass function See probability density function

query system A DBMS tool that enables users to create queries to retrieve data from a database. SQL is a standard query system found on many DBMSs.

random events The inability to specify events with complete certainty.

random sample A sample of observations selected from a population using some method to randomly choose the observations. All of statistical theory is based on the assumption that random chance is involved in a selection process. If sample data is selected without randomness, the results will be biased by the selection method and could be completely inaccurate.

random variable A function that assigns a number to every possible outcome in the sample space.

recommendation engine An automated process that provides recommendations of similar products to customers. Amazon in books and NetFlix in movies emphasize recommendations to increase sales and rentals.

relational OLAP (ROLAP) Data for OLAP and hyper cubes is stored in relational database tables. The approach is often simpler to configure than MOLAP, but it requires computing subtotals on the fly for all queries. Most systems that use ROLAP include tools to pre-build some subtotals, such as Oracle's materialized views.

relative frequency The most common expression of probability. The number of times an event can arise divided by the total number observations. Straightforward for common games of chance such as dice. The number 3 appears once on a die of 6 sides, so the relative frequency for observing the number 3 should be 1/6.

relative risk A measure of interestingness. It is used in Microsoft's association analysis as a method to compare potential rules. In probability terms, the risk = P(B|A) / P(B|~A). The ratio of the probability that B is selected given A is in the basket, versus B selected when A is not in the basket. The effect is similar to the formula for lift—attempting to measure the gain in the probability of purchasing B when A is present versus not present.

responsibilities The relative probability density functions, such as g0(/g0+g1), used in the expectation maximization clustering algorithm. The responsibility functions identify the weighting assigned to each point by each cluster.

roll up The process of aggregating data to a higher level in a hierarchy. The opposite of drill down in the process of browsing a hyper cube.

root mean square error (RMSE) A measure of error in an estimated model. It is computed as it is written: RMSE = $\sqrt{(\sum e^2 / n)}$

row-by-row calculations Using queries, simple calculations can be made using data on a single row at a time. Standard arithmetical operations (+, -, *, /) are supported. These calculations are performed before any aggregation operations. A few newer systems include support for Lag and Lead operators that can use data from rows above or below the current row.

sample mean The average of the observed values in a sample. Mean = sum(x)/n. The unbiased measure of the central tendency of the sample data.

sample space The set of all possible outcomes of an experiment.

sample variance The sum-of-squared deviation of the observed values in a sample. Variance = $\text{sum}(x - \text{mean})^2/(n-1)$.

Schwarz criterion or Bayesian information criterion (BIC) A goodness of fit measure of forecast error based on the squared difference between observed and predicted Y values. It is used for model selection. BIC = -2 ln(L) + k ln(n) where k is the number of estimated parameters, n is the number of observations and L is the likelihood function. Usually simplified to BIC=residual sum of squares/error variance + k ln(n). Also see Akaike information criteria.

seasonal ARIMA (SARIMA) A variation of the time series ARIMA method where the lags for auto-regression (AR) and moving average (MA) are defined in terms of multiples of the seasonality. For example, monthly data has a seasonality of 12, so the AR term (P) would be 1 or 2 to indicate 12 or 24 months. Similarly, differencing to resolve the trend occurs at the seasonal level, so a difference of 1 is accomplished by subtracting values that are 12 months apart.

seasonal auto-regressive (SAR) The auto-regressive lag structure of a seasonal model but the lag terms are specified in multiples of the seasonality. With monthly data, the seasonal factor is 12, so SAR(1) refers to a 12-month lag: $Y_t = a_1 Y_{t-12}$

seasonal moving average (SMA) The moving average lag structure of a seasonal ARIMA model where the lag terms are specified in terms of the seasonality. Moving average is based on the error (observed – predicted) values. With monthly data, the seasonal factor is 12, so SMA(1) refers to a 12-month lag: $e_t = b_1 e_{t-12}$.

seasonality Time series data often exhibits a seasonal pattern or correlations across an interval of time that corresponds to an annual period. For instance, sales typically increase at the end of the year holiday shopping season or unemployment increases in the summer months when students graduate from school.

seasonally adjusted Time series data is sometimes adjusted by removing seasonal patterns to make it easier to identify trends—particularly with monthly data. For example, sales for November and December might always be higher than September and October, but does that increase represent a trend or the normal seasonal pattern? Government economic data is often seasonally adjusted and it is important to be careful when using government data for time series analysis. Most analyses work better with raw or unadjusted data so the tools can determine the seasonal effect.

SELECT The primary data retrieval command for SQL. The main components are SELECT … FROM … INNER JOIN … WHERE.

Shannon entropy See information measure.

Simpson's paradox Also attributed to Yule, the paradox states that aggregate relationships across groups can be reversed when groups are combined. For instance, it is possible that in every department (subgroup), the percentage of men is less than the percentage of women; yet in the overall combined group, the percentage of men can be greater than the percentage of women.

skewed support Occurs in market basket analysis when the bulk of the items have few sales and a handful of items are sold in almost every basket. It leads to issues of cross-support errors. Because some items are in almost every basket, anything else might appear statistically useful—even though the purchase of low-support items could be completely random.

snowflake A design approach for OLAP data and hyper cubes. It extends the star design by enabling connections to tables through multiple links.

spurious correlation A combination of data or events that appears to be related but can easily occur by random chance. Because data mining tests so many extreme cases, it is helpful to estimate the random chance of critical events happening.

SQL A structured query language supported by most major database management systems. The most common command is of the form: SELECT column list FROM table list JOIN how tables are related WHERE condition ORDER BY columns.

SQL Server Analysis Services (SSAS) A collection of data mining tools provided by Microsoft that are integrated with the SQL Server database management system. The services are typically installed on a server and analyses are created using Visual Studio Business Intelligence tools on a client computer. Tools include, decision tree, naïve Bayes, regression, neural network, and time series analysis.

SQL Server Business Intelligence (BI) See SQL Server Analysis Services (SSAS).

standard deviation The square root of the variance. It is defined in the same units as the original data. From common distributions, most sample data will lie within +/- 2 standard deviations of the mean.

star design A design approach for OLAP data and hyper cubes. A fact table at the center contains measure attributes. Tables containing dimension attributes are connected directly to the fact table. Only one level of connection is supported with a star design, compared to extended connections supported in the snowflake design.

statistic A function of a random sample. Common statistics include the sample mean and variance.

subjective probability The Bayesian method of looking at probability. Probability values are updated based on new information using the Bayesian rule. The relative frequency approach is probably easier to understand initially, but the subjective approach is useful for many business problems.

support In association analysis a measure of the number of times an itemset occurs. The number of times a specified set occurs divided by the total number of observations. In probability terms, the relative frequency or an estimate of P(A) or P(A ∩ B).

table A collection of data for one class or entity. It consists of columns for each attribute and a row of data for each specific entity or object.

time series Data that is measured over time. The time period must be specified, and generally must be at fixed intervals (such as year, quarter, month, week, or day). A single time series uses data from one attribute that is consistently measured over time.

trend A pattern in time series data over time that exists outside of seasonal and cyclical factors.

uniform distribution A probability distribution (or pdf for continuous data) that uniformly allocates the data across a fixed range. All observations are equally likely to arise. For discrete data, $p(x) = 1/n$. For continuous data, $f(x) = 1/(b-a)$ where a and b are the lower and upper bounds. It is a straight horizontal line.

unsupervised learning A general data mining classification where the tool performs with minimal input by an analyst. For example, neural networks operate relatively unsupervised. After the dependent variable and potential independent variables are selected, the tool determines a model that fits the data. In comparison, standard regression analysis requires experience and knowledge to guide the selection of the final model.

variance The second moment about the mean. Or $E[(X - \text{mean})^2]$. The squared-deviation exhibited within the distribution. A measure of the dispersion of the probability distribution.

view A saved query. You can build new queries that retrieve data from the view. A view is saved as an SQL statement—not as the actual data.

Weka An open source set of data mining software written in Java and available free from The University of Waikato in New Zealand (http://www.cs.waikato.ac.nz/ml/weka). The set contains many standard analytic tools and reads standard comma-separated-values files.

WHERE The SQL clause that restricts the rows that will be used in the query. It can also refer to data in subqueries.

wisdom A level above knowledge. Wisdom represents intelligence, or the ability to analyze, learn, adapt to changing conditions, and create knowledge.