

Information Quality for Business Intelligence and Data Mining: Assuring Quality for Strategic Information Uses

*Larry P. English, President and Principal
Information Impact International, Inc.*
© 2005 Information Impact International

During the summer of 2005, scientists confirmed that there is a real pattern of global warming when they discovered and resolved an information quality problem in the data capture of surface temperatures. Satellites collecting data at the equator had reported temperatures that over time were relatively stable or showed a possible cooling trend. However, the satellites collecting that data had drifted off course and were reporting as daytime temperatures readings that were actually taken at night. Corrections to the data confirmed that there is warming at the equator that is consistent with surface warming around the globe.

This example illustrates how important trend findings can be obscured, misidentified or interpreted incorrectly if there are information quality problems anywhere in the information value chain.

Introduction

Information problems in information definition, data content, data preparation and information presentation can cause business intelligence processes to fail.

Here, I identify some of the critical information quality (IQ) problems in collection, preparing and presenting information for business intelligence and data mining along with IQ principles for mitigation or prevention.

Information Quality Issues for Data Mining and Statistical Analysis

Problems that hamper effective statistical data analysis stem from many sources of error introduction. First, data may not be clearly or accurately defined, causing a mismatch in the definition and the actual facts collected. Data can be captured inaccurately, or samples can be biased in record selection. Information quality decay causes data to become inaccurate when the characteristic of a real-world object changes. For example, if the price of an item changes, updated price values must be captured to assure the integrity of the analysis.

It is vital for the analyst to have or to conduct an information quality assessment to assure accuracy – not just validity – and completeness of data early in data preparation to allow time for any correction initiatives and preparation for mining.

Data preparation failure occurs when data is transformed in a way that is not able to be analyzed correctly by the data mining tools.

Finally, presentation graphics or display may not clearly convey the significance in the discovered patterns. Some examples:

❑ **Clear, correct, complete information definition:** An example of poor data definition comes from a survey taken by university students about student cell phone use. One attribute, “Included Minutes per Month,” was defined as: “Monthly allowed calling minutes that is written in the contract between the cell phone service provider and the customer.” The sample of data

collected included “400,” “5000,” “9999,” “500 plus night,” “Unlimited,” “625 (5000 n&w),” “500/free night weekend,” “50,000” and “-” (missing data). Due to the lack of clarity of the definition and absence of any data formatting, this data had to be manipulated and transformed for proper analysis.

IQ principle: Define data with business subject matter experts. Develop a consensus standard for values or data format. Provide training to information producers. Assess IQ for conformance to standards.

❑ **Measurement or data collection errors:** The global warming study described at the beginning of this article represents one kind of measurement error. Others include incorrect calibration of the temperature measurement device or improper placement of the measurement device.

IQ principle: Verify the calibration of measurement devices periodically. Assure consistent placement to capture data at a time, place or conditions that enable the identification of meaningful trends, e.g., taking surface temperature at the same location, at the same time of day by all satellites.

❑ **Sampling bias:** Data must be representative of the population being studied. If there is undiscovered sample bias and the population is not proportionately represented, the discovered trends will not be representative.

IQ principle: If sampling is made at information collection, assure that items or objects are selected by statistical sampling techniques, so that each object has equal likelihood of being selected. If data is being sampled, the data set must have a representative sample to the real-world collection of objects or events it represents. Record samples must be made using the same statistical sampling. If there are different strata in the population, assure a proportionate representation of each stratum, such as among the different classifications of frequent flyers and non-frequent flyers.

This having been said, there are some cases where you need to develop a “biased” sample that includes a higher proportion of outliers in order to predict rare events, such as fraudulent transactions. The modeling tools can better predict and can correct for the over-sampled cases or objects. The IQ principle here is that if the purpose of the analysis is to detect a rare class, use a training set in which the rare case is *over-represented*. If you specify correct prior probabilities of the rare case, the data mining predictions (posterior probabilities) will be correctly adjusted no matter what the proportions in the training set. If no prior probabilities are specified, the estimated posterior probabilities for the rare case will be too high, and the data mining predictions will be biased.

❑ **Missing values:** Data sources often contain observations that have missing values for one or more variables. Missing values can result from data collection errors, incomplete customer responses, actual system and measurement failures, or from a revision of the data collection scope over time, such as tracking new variables that were not included in the previous data collection schema. If an observation contains a missing value, then by default that observation is not used for modeling methods like neural network or linear regression. However, rejecting all incomplete observations may ignore useful or important information still contained in the non-missing variables. Rejecting all incomplete observations may also bias the sample, since observations that have missing values may have other things in common as well.

IQ principle: How should we treat missing data values? While there is no single correct answer, there are guidelines.

The first and best choice is to go back to the original real-world object and collect the data if it is knowable, such as the birth date of a person, and if the time of collection does not conflict

with the time of collection of the other data, such as temperature on a different day from the other data points. For events, such as measurements at a point in time, there must be a reliable recording of the event data to capture it with accuracy. Estimating the “best” missing value replacement technique requires assumptions about the true (missing) data. For example, if a variable’s data distribution follows a normal population response, you may replace a missing value with the mean of the variable. Be aware that replacing missing values with the mean, median, or another measure of central tendency is simple, *but* it can greatly affect a variable’s sample distribution. Use these replacement statistics carefully and only when the effect is minimal.

Another imputation technique replaces missing values with the mean of all other responses given by that data source, such as the exit poll responses at a specific precinct. This assumes that the input from that specific data source conforms to a normal distribution. Another technique studies the data to see if the missing values occur in only a few variables. If those variables are determined to be insignificant, the variables can be rejected from the analysis. However, the observations can still be used by the modeling nodes.

At a point there may be too much missing data for acceptable statistical analysis, and you may have to discard such attributes from the data. Another strategy is to use a modeling technique like decision trees, which automatically handle missing values. Finally, you may want to create a missing value indicator attribute and use it as candidate predictors in the model. The presence or absence of a value itself can be predictive.

❑ **Inaccurate values:** In most cases, inaccurate data can cause processes to fail. The higher the frequency the more severe the failure. Some errors in variables such as salary amount or age, can tolerate some precision error without significant trend discovery failure.

IQ principle: As with missing data, the best form of correction is to return to the real-world object to re-measure or discover the correct value.

❑ **Value synonyms:** Where data does not have a standardized value set, there may be different data values that represent the same characteristic, such as unit-of-measure synonyms “12,” “Doz,” and “Dz.” This causes the problem of dilution of patterns involving unit of measure of one dozen items in an order unit. If there were a relatively normal distribution of values among the three synonyms, the frequency of occurrence of unit of measure of one dozen will represent only about one third of all items with a real unit of measure of one dozen.

IQ principle: Identify and standardize the synonyms to a single value. This cannot be done arbitrarily; you must involve the business subject-matter experts. The real solution requires this to be standardized in the source processes and databases.

❑ **Overloaded variable values:** Often, data that is not controlled contains values that do not represent the characteristic the variable was designed for. Knowledge workers, in the absence of a well-designed database may have to “force” new facts into existing data elements. These overloaded fields create problems in trend correlation because they represent a different characteristic about the object or event that may bias the correlation of the original characteristic. For example, a Gender Code data element contained supposed “valid values” of “male,” “female,” “initials,” “ambiguous,” and “unknown.” The last three values did not represent gender; they represented why a gender-assignment routine was not able to determine gender by looking at the first name of the person.

IQ principle: For overloaded variables that represent multiple characteristics important to trend identification, break them out into separate variables and assure you have the correct definition.¹ However, if the overloaded values are mutually exclusive, this will introduce missing

data in both variables. In the Gender Code example above, if you were not able to contact the persons, you would have to provide a value of “unknown” and determine the impact of the missing data on your trend analysis.

❑ **Currency:** Currency represents the age of the data. Different trend analyses may require different ages of information. Identifying meaningful patterns requires having data of a common time period. For example, insurance policies have changes in business rules over time. You would not take policies in force 10 years ago and analyze them against the features of the comparable policies being sold today.

IQ principle: Understand the currency of all data required for a given model and assure data selection fits the age requirements.

❑ **Concurrency:** Concurrency is the timing difference of equivalence of data in one data store to another data store based on movement of data from one store to another. Records should be equivalent in content once the records reach a downstream data store. Data that is extracted from different databases may reach a given data set at different times. For example, orders reported today in the order fulfillment database will not be found in the historical order ODS (Operational Data Store) until tomorrow because they are extracted and loaded nightly. Shipments are not loaded until the end of the week. Returns are processed and loaded only after end of month. This causes problems in bringing data together to study patterns when the time periods of the transactions are different. To handle concurrency issues you must assure that the data extracted from multiple data sets represents a single time period.

IQ principle: Establish extract schedules (or extract transactions based on dates) from the various databases that will assure that transactions represent events or objects at a single point in time or time period. Solve the root causes by minimizing unnecessary redundant databases and information float. Eliminate the need for moving data to another database if that data can support all processes across the life cycle, such as persons or organizations that may be in a state of “prospect,” “active customer,” “preferred customer” or “inactive customer.”

Maintain appropriate date and time stamps and relationships of events to assure correlation of returns to the orders for which they are returned.

❑ **Outliers and anomalies:** Outliers are values that do not fit the expected set or range of valid values. Outliers may be errors or they may be accurate values but are beyond the realm of reasonability. A client of mine bought an item from a home improvement store for one penny (\$0.01). This item price is an outlier and is an inaccurate price. With the cell phone study mentioned above, the 50,000 “Included Minutes per Month” was both an outlier and an error, inasmuch as 50,000 minutes represents 34 days, 17 hours and 20 minutes, impossible to cram into even a 31-day month. It apparently was an arbitrary number given for “unlimited minutes.”

Outliers, even when correct, cause problems in mining, for they can often skew the statistical analysis. Figure 1 shows the bias in the frequency distribution of the “Included minutes” with the 50,000 minute outlier.

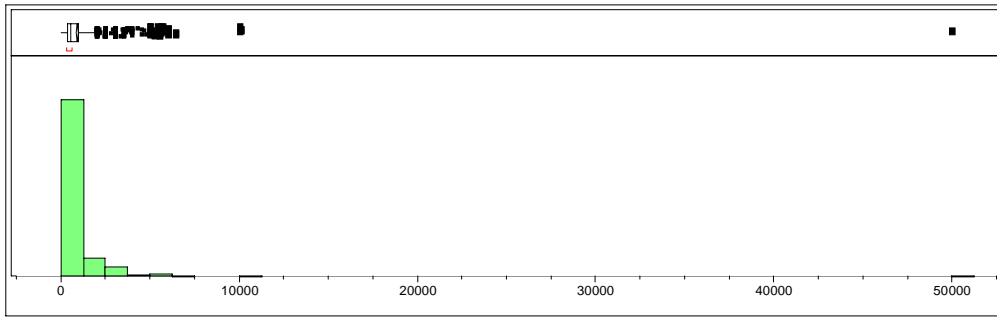


Figure 1 Outlier Bias in "Included Minutes" in Cell Phone Plans²

IQ principle: Outliers that are in fact errors should be corrected to the valid value. The 50,000 minute value, created to estimate a plan with “unlimited” minutes, had the negative side effect of skewing the valid minute “maximums.” To handle this problem you might (1) drop the outlier(s), or (2) estimate a realistic maximum, derived from analysis of actual minutes used by customers with “unlimited” minute plans. See Figure 2.

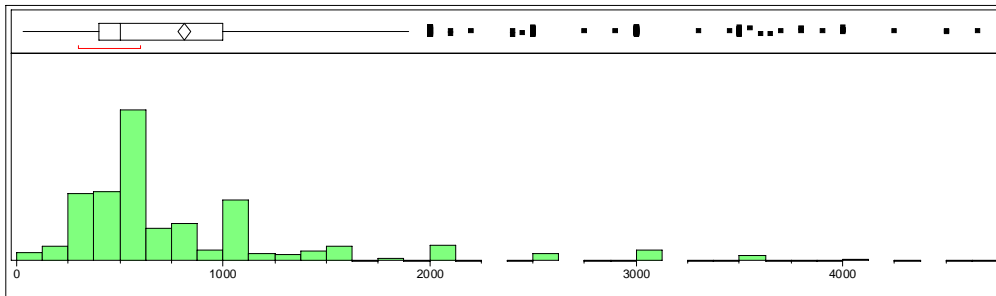


Figure 2 "Included Minutes" Without the 50,000 Minute Outlier³

The data (Figure 2) still looks skewed, with the right half of the chart falling beyond three standard deviations, so you might explore methods to “normalize” the data, such as using a logarithmic scale or by standardizing the data from 0 to 1 (or -1 to 1). See the discussion about mapping attribute data to numeric values, below.

Mapping categorical data to numeric values. Categorical data or attribute codes are not easily interpreted as to the relative relationship in trend analysis. In order to communicate to a statistical analysis tool, the codes or data must be analyzed to its relative position on a continuum. For example, a mining tool cannot differentiate between “Gold,” “Platinum,” or “Executive Platinum” frequent flyers by the text. This prevents such data from being useful in trend analysis.

IQ principle: Because most modeling techniques cannot interpret alphabetic codes, categorical or alphabetic information must be translated to ordered numeric values that can be interpreted for correlation. For example the “Included Minutes” values in the cell phone study above need to be translated to numbers between 0 and 1 that represents the relative degree of time between them because the modeling tool cannot interpret “unlimited” and “50,000” to be

the same, nor “625 (5000 n&w)” and “500/free night weekend,” to be roughly the same. For example, the mapping might look something like this:

<i>Actual Data Value</i>	<i>Numerical “Equivalent”</i>
200	0.1
400	0.15
500 plus night	0.4
500 / free night & weekend	0.5
625 (5,000 night & weekend)	0.5
10,000	0.8
Unlimited	1.0

Figure 3 Mapping Categorical Data to Numerical Values for Analysis

❑ **Modeling errors (correlated attributes):** Some data elements may be redundant in that they tell you the same information about an object, such as “birth date” and “age,” “gender” and “personal title,” or “frequent flyer status” and “miles flown.” When multiple correlated attributes are included they will skew the analysis.

IQ principle: Identify pairs of attributes that have a direct or closely direct (or indirect) correlation and eliminate one of the attributes, generally the one derivable, such as “age” versus “birth date.”

❑ **Data preparation errors (enhancement):** Often, trends and patterns cannot be determined with much precision without having external data representing real factors that influence behavior. Data preparation includes not only internal data, but also external data. For example, unexpected cold spells or warm spells influence behavior of purchases. Interest rate fluctuations, political events and other external factors can provide critical variables useful for predicting behavior.

IQ principle: You must understand your predictive model and determine whether internally known data are sufficient or whether external data can provide variables that are required to correctly interpret customer or product behavior.

Once you determine that external data is needed, you must acquire it, *and* you must assess its quality. Find out what IQ processes the information provider uses to assure quality of their information.

Conclusion

Assuring the quality of information for effective data mining and business intelligence begins long before the extraction and preparation of the data for mining. It begins with clear, accurate and complete definition of the data itself (the information product specification), and with error-proofing and controlling the processes that capture the data (for completeness, accuracy and precision), and with maintaining the quality as it may be subject to change.

Without first assuring the quality of the information definition, the data content the data preparation and the information presentation, the business intelligence conclusions will be “neither good business nor intelligent.”⁴

The Author

Larry English, President and Principal of Information Impact International, is an internationally recognized authority in information management and information quality. He has consulted in 29 countries on five continents. His TIQM[®] methodology applies quality principles to information quality management and has been implemented in many organizations worldwide. English was featured as one of the “21 Voices for the 21st Century” in the American Society for Quality’s *Quality Progress* journal. His book, *Improving Data Warehouse and Business Information Quality*, was hailed as “the Information Quality Bible for the Information Age” by Masaaki Imai, the creator of the Kaizen quality system. It has been translated into Japanese by the first information services organization to win the Deming Prize for Quality. English writes the “Plain English about Information Quality” column in *DM Review* magazine. Now in its 10th year, the column is consistently one of the magazine’s most popular. English serves as an editorial adviser for *DM Review* and the *Quality Assurance Journal*. He chairs The Information Quality Conferences in the United States and London and is co-founder of the International Association for Information and Data Quality (IAIDQ). English can be reached at Larry.English@infoimpact.com.

¹ Larry English, *Improving Data Warehouse and Business Information Quality*, New York: Wiley & Sons, 1999, p. 252.

² Weiwei Chen, “IQ in Data Mining,” analysis of cell phone use, using SAS JMP.

³ *Ibid.*

⁴ **sascom** Special Report: *BI and Beyond*.