Manifesto on Data Quality

Written by Mark Nadeau in February, 2021

Purpose: This document shares ideas that will help business analysts, data engineers, information systems specialists, and decision-makers recognize and mitigate the risks of data quality.

Abstract about data quality

First, let's talk about the different types of data quality. After all, it's important to appreciate all the flavors in order to really understand how and where trouble arises in the data process lifecycle.

Relevance: This problem occurs is the data isn't relevant – or not relevant enough – to the ultimate decision. Relevance can be measured by the degree to which a sample set of data is representative of a population; it's also measured by the amount of systemic separation or distinction between the targeted data and the network of interest.

Accuracy: This relates to the amount of noise, distortion, or interference introduced when the information is gathered, the quality of any forward error correction that travels with the signal, and the precision with which the data is transformed in every iteration of analysis.

Coherence: This is about the conformance of collected set of data points to some common source, structure, lineage, etc. So this relates to precision and the relationships that connect the data to the other parts of the set.

Timeliness: This characterizes how fresh or old the data is; how much time has elapsed between the event and the descriptive telling of the event. True "real time" data isn't possible, but that's the ideal. I guess this is related to relevance.

Computational Cost: This is a measure of the amount of "extras" that burden the data stream from any overdesign of analysis. For example, a redundant classification scheme could create a data table that's twice the size that it needs to be, which makes the application run poorly and complicates searches and lookups. Also, this could be about optimizing the number of model features and correlation variables that will impact the processing time of a simulation. So you could say that this is really just about timeliness.

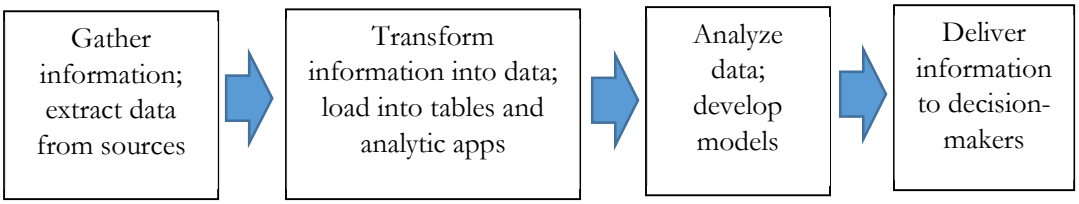
Suitability: This is about the user story -- whether the data is accessible to the right users to optimize the entire organization's total cost of ownership and BI capability. It's also about preventing access to the data for inappropriate motives, such as by malicious agents intent on fraud or cyber-attack. Another way to describe suitability is "relevance to purpose."

Credibility: This is a subjective measure of the confidence that the decision-makers and stakeholders will have in the integrity of the analysis and the data that informed it. I suppose credibility can be influenced by mere suspicion of any other data quality type. Also, it can be influenced by the tone and the style of the presentation of the findings and the visualization of the data models. It's also related to the extent of the decision-maker's knowledge about data analysis, system architecture, and statistical inference.

The Data Mining Lifecycle

Information is converted into data in order to facilitate analysis and modeling, but the output of any analysis is in more nebulous and context-rich format of information. So the analyst's work includes converting information to data and then back into information. And of course something is bound to be lost in the course of those changes!

The data mining process (from a very high level):



Some quality issues that arise when gathering information (and how to avoid them):

- The (seemingly) simple process of defining sample groupings and scoping the study can compromise relevance of the data. The best way to avoid this trouble is to clearly identify objectives, use Mutually Exclusive, Comprehensively Exhaustive categories, use strategic association rules, and define a data quality plan before launching into any discovery work.
- Information is often elicited from subject matter experts or derived from natural language or just from general observation. All sorts of biases can distort the SME's recollection and retelling of the information, and the interviewer may easily distort or reduce the information when translating the telling into a record or a data set.
- The target that's been selected for data gathering may include noise and irrelevant influences. Recognize those noises so that it can be culled from the good parts.
- It's impossible to observe anything at a small scale; something is lost through *decoherence* -- collapsing a wave function through observation! So… don't ever assume that your data provides the whole story.
- Calibration, fitness, and sensitivity of sampling and sensing equipment can influence data accuracy and the ultimate credibility of any of the downstream data. So… ensure appropriate calibration, fitness, and sensitivity of all sampling and sensing equipment!
- When data is transmitted across any device, noise and distortion can be introduced by the transmitter and by the channel. Use correction and amplification techniques as needed.
- Extraction from a data table or a database may ignore metadata and key field relationships, which may render much of the product meaningless. Avoid this trouble by first profiling the structure, content, and relationships of any structured or unstructured data set.
- Users can populate forms with the wrong type of data for a given field. Design forms carefully to control the type, length, and format of user input data.

Some quality issues that arise when transforming and loading data (and how to avoid them):

- Sloppy cross-walking can turn rows into columns and vice versa. Don't pivot inadvertently.
- Extraction from relational databases may be time intensive. Improve this by designing systems that use portioning and parallel processing. And techniques like clustering servers may improve processing speed of platforms.

- Validation during extraction may result in improper rejection of data
- On the ETL platform, a Notification of Change in Data may not be discrete, forcing a reload of the entire dataset; transform errors can be avoided by cleansing, character set crosswalk, encoding freeform values, aggregating, disaggregating repeats, and pivoting
- The user may improperly operate a data visualization interface; design interfaces to be fool-proof and intuitively easy to use.
- Formatting and data quality of the findings may fail to meet the needs of downstream consumers of the presented (or reported) data. So get to know the stakeholder needs!

Some quality issues that arise during analysis and modeling (and how to avoid them):

- The analyst may incorrectly infer that the data is representative of a population. Avoid this ghastly mistake by using a proper research hypothesis and statistical tests.
- Testing a hypothesis on the very same data set that was used to generate the hypothesis will lead to a wildly inaccurate model. Ensure your models will perform in the real world by distinguishing a set of "test" data for verification
- Time scales of the predictive model may be inconsistent with the descriptive data on which it's based. To avoid this, base the analysis on normal distributions of unclustered data; normalize any time-sensitive data like costs and population ratios to a present value.
- When prompted by a software modeling systems, an input distribution for a simulation can be completely irrelevant to the way the system behaves in real life. The analyst using the software should understand the different types of distribution functions (lognormal, Poisson, etc.) and be able to select the one that best mimics the system(s) being studied.

Some quality issues that arise when communicating findings (and how to avoid them):

- An audience can easily misinterpret the graphical visualization of data – trends can be inferred when there are no trends; proportions can be skewed; scales can become exaggerated. Use the right graphic to convey information; properly format charts and tables.
- The audience may fail to appreciate accuracy of the risk likelihood unless confidence intervals are clearly defined and explained. Further, a distinction should be made between types of risk – estimate accuracy, program risks that are being addressed, risks to project completion, etc.
- Costs and benefits of various alternatives can't be fairly compared or weighed unless they've been properly normalized and weighed within some sort of value framework or balanced scorecard. When faced with the task of comparing apples to oranges, decision-makers will happily revert to bias and intuition, in which case the analysis that preceded the decision was a complete waste of time.
- Running reports from production layers of databases creates costs in the operational processing. Data engineers should design the pipeline around dimensional data models to allow information to be quickly and efficiently targeted and accessed for multiple purposes.

A final note from the author: There is no final note! I hope that those who use this document will continue to add examples and solutions. (Already I can think of a hundred more.)