Aaron Cordova: ten principles for good data

Inspired by Dieter Ram's *10 principles for good design*, Aaron Cordova created the following 10 principles for good data.

These principles inform the products Koverse builds to enable organizations to take full advantage of all of their data.

Koverse CTO and cofounder Aaron Cordova

Good data varies in the level of structure.

The structure of data does not determine its usefulness - unstructured text and imagery, semi-structured and nested data, and highly structured records can contain equally valuable information. Technologies such as flexible schemas, natural language processing, and computer vision help unlock the information in these data types.

Good data is as much data as possible.

We no longer need to try to predict the usefulness of data before storing it. Data storage and processing technology has advanced to the point where it is now possible to store data before particular use cases are identified, full of potential, ready to be combined and analyzed with other sources.

Good data is co-located.

The volume of available data is increasing rapidly. As data volume grows, it becomes harder to move quickly. Not only are we now moving computation to the data, but various data sets should be stored close together on physical media so that combining different data sets and asking questions across multiple data sets becomes possible.

Good data is widely accessible.

To be effective, data must be available to the right people at the right time. Not only must the proper access be granted to various groups of decision makers, but the data must be retrievable quickly, at the speed at which decisions must be made. Organizing the data via profiling and indexing makes it possible to ensure the right data can be queried, searched, and delivered in time.

Good data can be traced to its source.

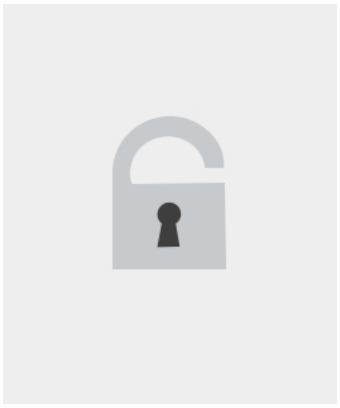
As data is combined, transformed, and summarized, the actions and relationships between source data and derived data sets must be recorded. Decisions that can be made from a data set are only as good as the sources and methods from which it was created. Being able to trace the lineage of data back to original sources is essential for making decisions with high confidence.

Good data should have an original version.

Often assumptions are made as to structure and semantics of a data set. Ideally these assumptions are informed by the data itself, but when it is discovered that an assumption is invalid, it must be possible to go back to the earliest form of the data and start over.

Good data is distributed.

The ability to work with data should only be limited by available resources and not by artificial technical hurdles. By distributing data onto multiple servers and using software designed to coordinate work across these servers, the time it takes to process data becomes a function of the number of servers available, and organizations can increase the number of servers to meet mission need. Many algorithms have already been parallelized to work in these environments and more work is being done all the time.



Good data is protected.

Before data owners are comfortable contributing data to an analytical system, sufficient control must be guaranteed such that those responsible for protecting data can be assured that access is granted only to groups that are authorized to read the data. Powerful security features make it possible to overcome obstacles to bring data together to provide insight without violating the confidentiality of the data.

Good data is understandable.

Billions of records and mounds of text are manageable from a storage and processing perspective, but at some point it should be made to guide thinking and actions within an organization. As raw data is frequently overwhelming and low level, this involves transforming raw data into higher level, often smaller data that can be readily applied to a decision process. These transformations include aggregation, summarization, training statistical models, and visualization to name a few.

Good data flows.

Gathering and organizing data can take so long that by the time it is in the form to support a decision it is already too old. Often this is not due to hardware limitations but is the result of the data flow being a manual, labor intensive process. Good data is frequently updated, and the derivatives of the data are updated in time to be relevant to today's decisions.