# BIG DATA GLOSSARY

## (Compiled by José Hernández-Orallo, November 2014)

**IMPORTANT: This material is copied from McKinsey Global Institute "Big data: The next frontier for innovation, competition, and productivity" 2011, Copyright © McKinsey & Company 2011 (for any use, refer to the original publication). Entries in blue have been completed or adapted from Wikipedia (for any use, cite Wikipedia).**

**A/B testing.** A technique in which a control group is compared with a variety of test groups in order to determine what treatments (i.e., changes) will improve a given objective variable, e.g., marketing response rate. This technique is also known as split testing or bucket testing. An example application is determining what copy text, layouts, images, or colors will improve conversion rates on an e-commerce Web site. Big data enables huge numbers of tests to be executed and analyzed, ensuring that groups are of sufficient size to detect meaningful (i.e., statistically significant) differences between the control and treatment groups (see statistics). When more than one variable is simultaneously manipulated in the treatment, the multivariate generalization of this technique, which applies statistical modeling, is often called "A/B/N" testing.

**Association rule learning.** A set of techniques for discovering interesting relationships, i.e., "association rules," among variables in large databases. These techniques consist of a variety of algorithms to generate and test possible rules. One application is market basket analysis, in which a retailer can determine which products are frequently bought together and use this information for marketing (a commonly cited example is the discovery that many supermarket shoppers who buy diapers also tend to buy beer). Used for data mining.

**Big Table.** Proprietary distributed database system built on the Google File System. Inspiration for HBase.

**Business intelligence (BI).** A type of application software designed to report, analyze, and present data. BI tools are often used to read data that have been previously stored in a data warehouse or data mart. BI tools can also be used to create standard reports that are generated on a periodic basis, or to display information on real-time management dashboards, i.e., integrated displays of metrics that measure the performance of a system.

**Cassandra**. An open source (free) database management system designed to handle huge amounts of data on a distributed system. This system was originally developed at Facebook and is now managed as a project of the Apache Software foundation.

**Classification**. A set of techniques to identify the categories in which new data points belong, based on a training set containing data points that have already been categorized. One application is the prediction of segment-specific customer behavior (e.g., buying decisions, churn rate, consumption rate) where there is a clear hypothesis or objective outcome. These techniques are often described as supervised learning because of the existence of a training set; they stand in contrast to cluster analysis, a type of unsupervised learning. Used for data mining.

**Cloud computing.** A computing paradigm in which highly scalable computing resources, often configured as a distributed system, are provided as a service through a network.

**Cluster analysis.** A statistical method for classifying objects that splits a diverse group into smaller groups of similar objects, whose characteristics of similarity are not known in advance. An example of cluster analysis is segmenting consumers into self-similar groups for targeted marketing. This is a type of unsupervised learning because training data are not used. This technique is in contrast to classification, a type of supervised learning. Used for data mining.

**Crowdsourcing.** A technique for collecting data submitted by a large group of people or ommunity (i.e., the “crowd”) through an open call, usually through networked media such as the Web. This is a type of mass collaboration and an instance of using Web.

**Dashboard**: real-time user interface that provides at-a-glance views of KPIs (key performance indicators) relevant to a particular objective or business process (e.g. sales, marketing, human resources, or production)

**Data analysis:** process of inspecting, cleaning, transforming, and modeling data with the goal of discovering useful information, suggesting conclusions, and supporting decision-making. Data analysis has multiple facets and approaches, encompassing diverse techniques under a variety of names, in different business, science, and social science domains.

**Data fusion and data integration.** A set of techniques that integrate and analyze data from multiple sources in order to develop insights in ways that are more efficient and potentially more accurate than if they were developed by analyzing a single source of data. Signal processing techniques can be used to implement some types of data fusion. One example of an application is sensor data from the Internet of Things being combined to develop an integrated perspective on the performance of a complex distributed system such as an oil refinery. Data from social media, analyzed by natural language processing, can be combined with real-time sales data, in order to determine what effect a marketing campaign is having on customer sentiment and purchasing behavior.

**Data mart.** Subset of a data warehouse, used to provide data to users usually through business intelligence tools.

**Data mining.** A set of techniques to extract patterns from large datasets by combining methods from statistics and machine learning with database management. These techniques include association rule learning, cluster analysis, classification, and regression. Applications include mining customer data to determine segments most likely to respond to an offer, mining human resources data to identify characteristics of most successful employees, or market basket analysis to model the purchase behavior of customers.

**Data munging or wrangling:** process of manually converting or mapping data from one "raw" form into another format that allows for more convenient consumption of the data with the help of semi-automated tools. This may include further munging, data visualization, data aggregation, training a statistical model, as well as many other potential uses. Data munging as a process typically follows a set of general steps which begin with extracting the data in a raw form from the data source, "munging" the raw data using algorithms (e.g. sorting) or parsing the data into predefined data structures, and finally depositing the resulting content into a data sink for storage and future use.

**Data science:** the study of the generalizable extraction of knowledge from data, yet the key word is science. It incorporates varying elements and builds on techniques and theories from many fields, including signal processing, mathematics, probability models, machine learning, statistical learning, computer programming, data engineering, pattern recognition and learning, visualization, uncertainty modeling, data warehousing, and high performance computing with the goal of extracting meaning from data and creating data products. The subject is not restricted to only big data, although the fact that data is scaling up makes big data an important aspect of data science. Another key ingredient that boosted the practice and applicability of data science is the development of machine learning - a branch of artificial intelligence - which is used to uncover patterns from data and develop practical and usable predictive models. A practitioner of data science is called a data scientist.

**Data warehouse**. Specialized database optimized for reporting, often used for storing large amounts of structured data. Data is uploaded using ETL (extract, transform, and load) tools from operational data stores, and reports are often generated using business intelligence tools.

**Distributed database**: a database in which storage devices are not all attached to a common processing unit such as the CPU, controlled by a distributed database management system (together sometimes called a distributed database system). It may be stored in multiple computers, located in the same physical location; or may be dispersed over a network of interconnected computers. Unlike parallel systems, in which the processors are tightly coupled and constitute a single database system, a distributed database system consists of loosely-coupled sites that share no physical components.

Distributed system. Multiple computers, communicating through a network, used to solve a common computational problem. The problem is divided into multiple tasks, each of which is solved by one or more computers working in parallel. Benefits of distributed systems include higher performance at a lower cost (i.e., because a cluster of lower-end computers can be less expensive than a single higher-end computer), higher reliability (i.e., because of a lack of a single point of failure), and more scalability (i.e., because increasing the power of a distributed system can be accomplished by simply adding more nodes rather than completely replacing a central computer).

Dynamo. Proprietary distributed data storage system developed by Amazon.

Ensemble learning. Using multiple predictive models (each developed using statistics and/or machine learning) to obtain better predictive performance than could be obtained from any of the constituent models. This is a type of supervised learning.

Extract, transform, and load (ETL). Software tools used to extract data from outside sources, transform them to fit operational needs, and load them into a database or data warehouse.

Genetic algorithms. A technique used for optimization that is inspired by the process of natural evolution or “survival of the fittest.” In this technique, potential solutions are encoded as “chromosomes” that can combine and mutate. These individual chromosomes are selected for survival within a modeled “environment” that determines the fitness or performance of each individual in the population. Often described as a type of “evolutionary algorithm,” these algorithms are well-suited for solving nonlinear problems. Examples of applications include improving job scheduling in manufacturing and optimizing the performance of an investment portfolio.

Google File System. Proprietary distributed file system developed by Google; part of the inspiration for Hadoop.

Hadoop. An open source (free) software framework for processing huge datasets on certain kinds of problems on a distributed system. Its development was inspired by Google’s MapReduce and Google File System. It was originally developed at Yahoo! and is now managed as a project of the Apache Software Foundation.

HBase. An open source (free), distributed, non-relational database modeled on Google’s Big Table. It was originally developed by Powerset and is now managed as a project of the Apache Software foundation as part of the Hadoop.

Knowledge Discovery from Databases (KDD): an interdisciplinary subfield of computer science, is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. Now a synonym of data mining.

Machine learning. A subspecialty of computer science (within a field historically called “artificial intelligence”) concerned with the design and development of algorithms that allow computers to evolve behaviors based on empirical data. A major focus of machine learning research is to automatically learn to recognize complex patterns and make intelligent decisions based on data. Natural language processing is an example of machine learning.

MapReduce. A software framework introduced by Google for processing huge datasets on certain kinds of problems on a distributed system. Also implemented in Hadoop.

Mashup. An application that uses and combines data presentation or functionality from two or more sources to create new services. These applications are often made available on the Web, and frequently use data accessed through open application programming interfaces or from open data sources.

Metadata. Data that describes the content and context of data files, e.g., means of creation, purpose, time and date of creation, and author.

Natural language processing (NLP). A set of techniques from a subspecialty of computer science (within a field historically called "artificial intelligence") and linguistics that uses computer algorithms to analyze human (natural) language. Many NLP techniques are types of machine learning. One application of NLP is using sentiment analysis on social media to determine how prospective customers are reacting to a branding campaign.

Network analysis. A set of techniques used to characterize relationships among discrete nodes in a graph or a network. In social network analysis, connections between individuals in a community or organization are analyzed, e.g., how information travels, or who has the most influence over whom. Examples of applications include identifying key opinion leaders to target for marketing, and identifying bottlenecks in enterprise information flows.

Neural networks. Computational models, inspired by the structure and workings of biological neural networks (i.e., the cells and connections within a brain), that find patterns in data. Neural networks are well-suited for finding nonlinear patterns. They can be used for pattern recognition and optimization. Some neural network applications involve supervised learning and others involve unsupervised learning. Examples of applications include identifying high-value customers that are at risk of leaving a particular company and identifying fraudulent insurance claims.

Non-relational database. A database that< does not store data in tables (rows and columns). (In contrast to relational database).

NoSQL: A NoSQL (often interpreted as Not Only SQL) database provides a mechanism for storage and retrieval of data that is modeled in means other than the tabular relations used in relational databases. Motivations for this approach include simplicity of design, horizontal scaling and finer control over availability. The data structure (e.g. key-value, graph, or document) differs from the RDBMS, and therefore some operations are faster in NoSQL and some in RDBMS. NoSQL systems are also called "Not only SQL" to emphasize that they may also support SQL-like query languages.

OLAP (On-line Analytic Processing): approach to answering multi-dimensional analytical (MDA) queries swiftly. OLAP is part of the broader category of business intelligence, which also encompasses relational database, report writing and data mining. Typical applications of OLAP include business reporting for sales, marketing, management reporting, business process management (BPM), budgeting and forecasting, financial reporting and similar areas, with new applications coming up, such as agriculture. The term OLAP was created as a slight modification of the traditional database term Online Transaction Processing ("OLTP")

OLTP (On-line Transactional Processing): is a class of information systems that facilitate and manage transaction-oriented applications, typically for data entry and retrieval transaction processing. Many relational databases have an OLTP use.

Optimization. A portfolio of numerical techniques used to redesign complex systems and processes to improve their performance according to one or more objective measures (e.g., cost, speed, or reliability). Examples of applications include improving operational processes such as scheduling, routing, and floor layout, and making strategic decisions such as product range strategy, linked investment analysis, and R&D portfolio strategy. Genetic algorithms are an example of an optimization technique.

Pattern recognition. A set of machine learning techniques that assign some sort of output value (or *label*) to a given input value (or *instance*) according to a specific algorithm. Classification techniques are an example.

Predictive modeling. A set of techniques in which a mathematical model is created or chosen to best predict the probability of an outcome. An example of an application in customer relationship management is the use of predictive models to estimate the likelihood that a customer will “churn” (i.e., change providers) or the likelihood that a customer can be cross-sold another product. Regression is one example of the many predictive modeling techniques.

R. An open source (free) programming language and software environment for statistical computing and graphics. The R language has become a de facto standard among statisticians for developing statistical software and is widely used for statistical software development and data analysis. R is part of the GNU Project, a collaboration that supports open source projects.

Regression. A set of statistical techniques to determine how the value of the dependent variable changes when one or more independent variables is modified. Often used for forecasting or prediction. Examples of applications include forecasting sales volumes based on various market and economic variables or determining what measurable manufacturing parameters most influence customer satisfaction. Used for data mining.

Relational database. A database made up of a collection of tables (relations), i.e., data are stored in rows and columns. Relational database management systems (RDBMS) store a type of structured data. SQL is the most widely used language for managing relational databases (see item below).

Semi-structured data. Data that do not conform to fixed fields but contain tags and other markers to separate data elements. Examples of semi-structured data include XML or HTML-tagged text. Contrast with structured data and unstructured data.

Sentiment analysis. Application of natural language processing and other analytic techniques to identify and extract subjective information from source text material. Key aspects of these analyses include identifying the feature, aspect, or product about which a sentiment is being expressed, and determining the type, “polarity” (i.e., positive, negative, or neutral) and the degree and strength of the sentiment. Examples of applications include companies applying sentiment analysis to analyze social media (e.g., blogs, microblogs, and social networks) to determine how different customer segments and stakeholders are reacting to their products and actions.

Signal processing. A set of techniques from electrical engineering and applied mathematics originally developed to analyze discrete and continuous signals, i.e., representations of analog physical quantities (even if represented digitally) such as radio signals, sounds, and images. This category includes techniques from signal detection theory, which quantifies the ability to discern between signal and noise. Sample applications include modeling for time series analysis or implementing data fusion to determine a more precise reading by combining data from a set of less precise data sources (i.e., extracting the signal from the noise).

Simulation. Modeling the behavior of complex systems, often used for forecasting, predicting and scenario planning. Monte Carlo simulations, for example, are a class of algorithms that rely on repeated random sampling, i.e., running thousands of simulations, each based on different assumptions. The result is a histogram that gives a probability distribution of outcomes. One application is assessing the likelihood of meeting financial targets given uncertainties about the success of various initiatives.

Spatial analysis. A set of techniques, some applied from statistics, which analyze the topological, geometric, or geographic properties encoded in a data set. Often the data for spatial analysis come from geographic information systems (GIS) that capture data including location information, e.g., addresses or latitude/longitude coordinates. Examples of applications include the incorporation of spatial data into spatial regressions (e.g., how is consumer willingness to purchase a product correlated with location?) or simulations (e.g., how would a manufacturing supply chain network perform with sites in different locations?).

SQL. Originally an acronym for structured query language, SQL is a computer language designed for managing data in relational databases. This technique includes the ability to insert, query, update, and delete data, as well as manage data schema (database structures) and control access to data in the database.

Statistics. The science of the collection, organization, and interpretation of data, including the design of surveys and experiments. Statistical techniques are often used to make judgments about what relationships between variables could have occurred by chance (the “null hypothesis”), and what relationships between variables likely result from some kind of underlying causal relationship (i.e., that are “statistically significant”). Statistical techniques are also used to reduce the likelihood of Type I errors (“false positives”) and Type II errors (“false negatives”). An example of an application is A/B testing to determine what types of marketing material will most increase revenue.

Stream processing. Technologies designed to process large real-time streams of event data. Stream processing enables applications such as algorithmic trading in financial services, RFID event processing applications, fraud detection, process monitoring, and location-based services in telecommunications. Also known as event stream processing.

Structured data. Data that reside in fixed fields. Examples of structured data include relational databases or data in spreadsheets. Contrast with semi-structured data and unstructured data.

Supervised learning. The set of machine learning techniques that infer a function or relationship from a set of training data. Examples include classification and support vector machines.30 This is different from unsupervised learning.

Time series analysis. Set of techniques from both statistics and signal processing for analyzing sequences of data points, representing values at successive times, to extract meaningful characteristics from the data. Examples of time series analysis include the hourly value of a stock market index or the number of patients diagnosed with a given condition every day. Time series forecasting is the use of a model to predict future values of a time series based on known past values of the same or other series. Some of these techniques, e.g., structural modeling, decompose a series into trend, seasonal, and residual components, which can be useful for identifying cyclical patterns in the data. Examples of applications include forecasting sales figures, or predicting the number of people who will be diagnosed with an infectious disease.

Unstructured data. Data that do not reside in fixed fields. Examples include free-form text (e.g., books, articles, body of e-mail messages), untagged audio, image and video data. Contrast with structured data and semi-structured data.

Unsupervised learning. A set of machine learning techniques that finds hidden structure in unlabeled data. Cluster analysis is an example of unsupervised learning (in contrast to supervised learning).

VLDB: a very large database is a database that contains an extremely high number of tuples (database rows), or occupies an extremely large physical filesystem storage space. Since the year 2011, this term is now referred to as big data by industry.

Visualization. Techniques used for creating images, diagrams, or animations to communicate, understand, and improve the results of big data analyses (see the last section of this chapter).