Glossary

Knowledge is understood as *information and experience internalised or assimilated* through a process, commonly referred to as "learning". It provides the "learner" with the capacity to make effective decisions autonomously. Knowledge can be explicit, in which case it can be cost-effectively externalised to be communicated and embedded in tangible products, including books, standard procedures and intangible products such as patents, design and software. But it can also be tacit, based on an "amalgam of information and experience", which is too costly to codify and thus to externalise.

Information is often seen as the *meaning* resulting from the interpretation of facts as conveyed through *data* or other sources such as words. This meaning is reflected in the structure or organisation of the underlying source, including its hidden relationships and patterns of correlations, which can be revealed through *data analytics*. Information is therefore always context-dependent: it depends on the capacity to extract meaning from the information source; this capacity depending on available data analytic techniques and technologies as well as the skills and (pre-)knowledge of the data analyst.

Data are understood as the *representation of facts* stored or transmitted as qualified or quantified symbols. Data have no inherent meaning; however, they can be domain-specific. In contrast to knowledge and information, data are assumed to have an "objective existence", and they can be measured, namely in bits and bytes (see Table below). Data are typically gained from information when that information is *encoded* so it can be stored or communicated. Data can also be the result of *datafication*, a portmanteau for "data" and "quantification", where a phenomenon or object is transformed into quantified symbols. Datafication should not be confused with *digitisation*, which refers to the process of encoding information into *binary digits* (i.e. bits) so it can be processed by computers. Data that have not been digitised cannot be processed by computers.

Big data initially referred to data for which the i) *volume* became an issue in terms of data management and processing. Further definitions highlighted other important characteristics of "big data", such as ii) *velocity,* or the speed at which data are generated, accessed, processed and analysed (referring to real-time data), and iii) *variety* (referring to *unstructured* data and the capacity to link diverse data sets). These three properties – volume, velocity and variety – are therefore often considered to be the three main characteristics, and are commonly referred to as the three Vs, of big data. There is a major limitation with definitions based on the 3Vs, however: they are in continuous flux, as they describe technical properties that depend on the evolving state of the art in data storage and processing. Furthermore, these definitions misleadingly suggest that data are the main source of value. While it is true in the case of volume, what is behind variety and velocity is primarily *data analytics* – that is, the capacity to analyse unstructured diverse data in (close to) real time. Furthermore the term "big data" does not suggest how the data are used what type of innovation they can enable, or a how they relate to other concepts such as (e.g.) *open data*, *linked data*, and *data mashups*.

Units for measuring the volume of data

Unit	Size	What it means
Bit (B)	1 or 0	Short for "binary digit", after the binary code (1 or 0) computers uses to store and process data.
Byte (B)	8 bits	Enough information to create a number or an English letter in computer code. It is the basic unit of computing.
Kilobyte (KB)	1 000 B	From "thousand" in Greek. One page of typed text is 2 KB.
Megabyte (MB)	1 000 KB	From "large" in Greek. The complete works of Shakespeare total 5MB. A typical pop song is about 4 MB.
Gigabyte (GB)	1 000 MB	From "giant" in Greek. A two-hour film can be compressed into 1-2GB.
Terabyte (TB)	1 000 GB	From "monster" in Greek. All the catalogued books in the US Library of Congress total around 15 TB.
Petabyte (PB)	1 000 TB	All letters delivered by America's postal service in 2011 will amount to around 5 PB; Google processes around 1 PB every hour.
Exabyte (EB)	1 000 PB	Equivalent to 10 billion copies of *The Economist.*
Zettabyte (ZB)	1 000 EB	The total amount of information in existence in 2011 was around 1.2 ZB.
Yottabyte (YB)	1 000 ZB	Currently too big to imagine.

Note: The prefixes are set by an intergovernmental group, the International Bureau of Weights and Measures. Yotta and Zetta were added in 1991; terms for larger amounts have yet to be established.

Source: Adopted from *The Economist (2010), "Data, data everywhere", The Economist, 25 February,* www.economist.com/node/15557443.

Structured data are data based on a predefined *data model* (i.e. an abstract representation of "real world" objects and phenomenon). Such models can be explicit, as in the case of a structured query language (SQL) database, where the data model is reflected in the structure of the database's tables. The data model can also be implicit, as in the case of *semi-structured data* (e.g. structured web content), where the underlying model can be made explicit at relatively low cost. In contrast, **unstructured data** are data that have no predefined data model and where such a model cannot be cost-effectively extracted. Typical examples include text-heavy data sets such as text documents and e-mails, as well as multimedia content such as videos, images and audio streams. The difference between structured, semi-structured, and unstructured data is becoming less important since with rising computing capacities, *data analytics* are increasingly able to automatically extract some structures embedded in unstructured data, including multimedia content.

Linked data typically refers to structured data that are published so that they can be interlinked. Data linkage is a means to contextualise data and thus enable the extraction of further information, which is greater than the sum of the information from the isolated **data silos**. The concept of linked data is closely related to the concept of open data, for which the full benefits can only be achieved if the isolated open data sets can be interlinked. Open standards play an important role in an interlinked data ecosystem.

Metadata are data about entities, including (**primary) data**. Metadata provide the necessary context without which the primary data cannot be accessed, linked, or fully understood. Metadata can be i) descriptive (based on attributes used to search and find an entity), ii) structural (describing the structure and organisation of an entity such as databases), and iii) administrative (providing information to help manage a resource). The concept of metadata is closely related to the concept of linked data, since metadata and primary data are by definition linked.

Personal data are defined by the OECD *Guidelines Governing the Protection of Privacy and Transborder Flows of Personal Data* as "any information relating to an identified or identifiable individual (data subject)". Any data that are not related to an identified or identifiable individual are therefore "non-personal" data. However, *data analytics* has made it easier to relate seemingly non-personal data to an identified or identifiable individual, thus blurring the boundaries between non-personal and personal data (see Chapter 5). It should be noted that the definition of personal data applied here does not distinguish between data (as inherently meaningless representation of facts) and information (as the *meaning* resulting from the interpretation of data). In other words, personal data and personal information are used as synonyms in this report.

Data can be **volunteered** when they are explicitly shared (by a data subject). Examples include creating a social network profile and entering credit card information for online purchases. They can be **observed** when it is captured by recording activities. In contrast to volunteered data where the data subject is actively and purposefully sharing its data, the role of the observed data subject is passive. Examples of observed data include location data of cellular mobile phones, and web usage behaviour. And finally, information can be **inferred** as the result of *data analytics*. Examples include credit scores calculated based on an individual's financial history. It is interesting to note that personal information can be "inferred" from several pieces of seemingly "anonymous" or "non-personal" data.

Public sector (government) data, in respect to the OECD *Recommendation of the Council for Enhanced Access and More Effective Use of Public Sector Information* (PSI), are data generated, created, collected, processed, preserved, maintained, disseminated, or funded by or for the government or public institutions (see Chapter 10). They are: i) dynamic and continually generated, ii) often directly produced by the public sector, or iii) associated with the functioning of the public sector (e.g. meteorological data, geo-spatial data, business statistics), and iv) often readily useable in commercial applications with relatively little transformation, as well as being the basis of extensive elaboration. Public sector data are a subset of PSI, which includes not only data but also *digital content,* such as text documents and multimedia files. The terms "public sector data" and "government data" are used as synonyms. The often used term "open government data" refers to public sector data made available as *open data*.

Open data does not describe a specific type of data. The key characteristic is the attribute "open", which specifies how access to data is *managed*, namely on *non-discriminatory terms* or "access on equal terms" as stated in the OECD *Recommendation of the Council on Principles and Guidelines for Access to Research Data from Public Funding*. In other words, data become "open" when access is not limited based on users' identity or intended use of the data (see Chapter 4). "Openness" should not be understood as a binary attribute but rather as a *continuum,* ranging from i) *closed* (with access only by e.g. the data controller or data subject), to ii) *commons* with possible restriction to a community (e.g. of researchers), to iii) *(unlimited) access granted to the public* as the highest degree of openness. Three key factors affect the degree of openness:

- technological design (including e.g. availability, machine readability and interoperability)
- intellectual property rights (IPRs) (including copyright as well as other IPRs applicable to databases and trade secrets)

- pricing, with marginal cost pricing being recommended by the OECD (2006) Council Recommendation on Access to Research Data from Public Funding and the OECD *Recommendation of the Council on Enhanced Access and More Effective Use of PSI*.

Data analytics refers to the set of techniques and tools used to extract information from data by revealing the context in which the data are embedded, their organisation and their structure. In the case of visual analytics the emphasis lies on data visualisation including (interactive) data exploration. Data analytics reveals the signal from the noise and with that the data's manifold hidden relations (patterns) including correlations, and interactions between facts, entities, and concepts. A number of terms are used (in this volume as synonyms) to refer to data analytics, some of which may include aspects that go beyond data analysis:

- **Data (text) mining** and **knowledge discovery** typically refer to data analysis but include aspects such as data pre-processing (cleaning), as well as model and inference considerations.
- **Profiling** is often used to describe the construction of profiles and the classification of entities in specific profiles.
- **Business intelligence**, a term that refers to the analysis of business-related data as often stored in databases (data warehouses) and mainly used for business reporting and monitoring purposes.
- **Machine** or **statistical learning** is a subfield in computer science, and more specifically artificial intelligence (AI), concerned with the design, development and use of data analytic algorithms that allow computers to "learn" – that is, to improve performance with every data set analysed.

From:

Data-Driven Innovation

Big Data for Growth and Well-Being

Access the complete publication at:
https://doi.org/10.1787/9789264229358-en

Please cite this chapter as:

OECD (2015), "Glossary", in *Data-Driven Innovation: Big Data for Growth and Well-Being*, OECD Publishing, Paris.

DOI: https://doi.org/10.1787/9789264229358-15-en