Data Literacy: Glossary and Index

Table of Contents

[Glossary 2](#_Toc96131026)

[ADVANCED RESEARCH COMPUTING 2](#_Toc96131027)

[ALGORITHM 3](#_Toc96131028)

[ANALYTICS 3](#_Toc96131029)

[ARTIFICIAL INTELLIGENCE 3](#_Toc96131030)

[BEHAVIOUR(AL) ANALYTICS 3](#_Toc96131031)

[BIG DATA 3](#_Toc96131032)

[Big data isn’t always big. 3](#_Toc96131033)

[CAUSAL INFERENCE 3](#_Toc96131034)

[CLASSIFICATION 3](#_Toc96131035)

[CLOUD COMPUTING 3](#_Toc96131036)

[CLUSTER 4](#_Toc96131037)

[CLUSTERING 4](#_Toc96131038)

[COMPUTING WITH DATA 4](#_Toc96131039)

[CORRELATION MINING 4](#_Toc96131040)

[CYBERSECURITY 4](#_Toc96131041)

[Critical Thinking 4](#_Toc96131042)

[Crowdsourcing 4](#_Toc96131043)

[DATA 4](#_Toc96131044)

[DATA AGGREGATION 4](#_Toc96131045)

[DATA CLEANING 4](#_Toc96131046)

[DATA EXPLORATION AND PREPARATION 5](#_Toc96131047)

[DATA INTEGRATION 5](#_Toc96131048)

[DATA MINING 5](#_Toc96131049)

[DATA REPRESENTATION 5](#_Toc96131050)

[DATA SCIENCE 5](#_Toc96131051)

[DATA TRANSFORMATION 5](#_Toc96131052)

[DATABASE 5](#_Toc96131053)

[DESCRIPTIVE ANALYTICS 5](#_Toc96131054)

[DISTRIBUTED FILE SYSTEM 5](#_Toc96131055)

[Data (Clean, Dirty) 6](#_Toc96131056)

[Data Bias 6](#_Toc96131057)

[Data Literacy 6](#_Toc96131058)

[Data Privacy 6](#_Toc96131059)

[Data Quality 6](#_Toc96131060)

[Data types 6](#_Toc96131061)

[FEATURE 7](#_Toc96131062)

[FEATURE ENGINEERING 7](#_Toc96131063)

[FLEXIBLE SCHEMA 7](#_Toc96131064)

[File Formats 7](#_Toc96131065)

[GENERATIVE MODELING 8](#_Toc96131066)

[GOODNESS OF FIT 8](#_Toc96131067)

[INFORMATION 8](#_Toc96131068)

[INSIGHT 8](#_Toc96131069)

[IOT (INTERNET OF THINGS) 8](#_Toc96131070)

[KNOWLEDGE DISCOVERY 8](#_Toc96131071)

[MACHINE LEARNING 8](#_Toc96131072)

[METADATA 9](#_Toc96131073)

[MODEL FITTING 9](#_Toc96131074)

[NETWORK 9](#_Toc96131075)

[NETWORK ANALYSIS 9](#_Toc96131076)

[NO-SQL DATABASE 9](#_Toc96131077)

[PREDICTIVE ANALYTICS 9](#_Toc96131078)

[PRESCRIPTIVE ANALYTICS 9](#_Toc96131079)

[Population 9](#_Toc96131080)

[REGRESSION 9](#_Toc96131081)

[RELATIONSHIP MINING 9](#_Toc96131082)

[SECURE COMPUTING 9](#_Toc96131083)

[SQL 10](#_Toc96131084)

[STRUCTURED DATA 10](#_Toc96131085)

[Sample (Representative, Proxy) 10](#_Toc96131086)

[TABULAR DATA 10](#_Toc96131087)

[TEXT MINING 10](#_Toc96131088)

[Tree structure 10](#_Toc96131089)

[UNSTRUCTURED DATA 10](#_Toc96131090)

[Unbalanced Data 11](#_Toc96131091)

[VISUALIZATION 11](#_Toc96131092)

[VISUALIZATION DASHBOARD 11](#_Toc96131093)

[Categories 11](#_Toc96131094)

[Data Inquiry Disciplines 11](#_Toc96131095)

[Data Analysis Techniques 12](#_Toc96131096)

[Data Processing 12](#_Toc96131097)

[Storage and Physical Data Sources 13](#_Toc96131098)

[Types of Data 13](#_Toc96131099)

[Lessons Index 13](#_Toc96131100)

[Lessons 01 and 02 13](#_Toc96131101)

[Lesson 03 13](#_Toc96131102)

[Lesson 04 14](#_Toc96131103)

[References 14](#_Toc96131104)

# Glossary

### ADVANCED RESEARCH COMPUTING

Advanced research computing (ARC) provides massive computational horsepower and [storage](#storage-and-physical-data-sources) in a [cloud environment](#cloud-computing) to handle problems and data that are too complex for a single desktop computer. [1](#fn1),[2](#fn2) <#disciplines>

### ALGORITHM

A sequence of instructions telling a computer how to answer a specific question. [3](#fn3),[4](#fn4) <#processing>

### ANALYTICS

The process of using [statistical models](#machine-learning) and software to [transform data](#data-transformation) into useful [information](#information) and to draw conclusions towards effective decision making.[5](#fn5),[6](#fn6) <#disciplines>

### ARTIFICIAL INTELLIGENCE

A branch of computer science that allows machines to acquire and apply knowledge to handle new inputs and analyze patterns to solve diverse problems. [7](#fn7) <#disciplines> <#lesson-03>

### BEHAVIOUR(AL) ANALYTICS

A type of business [analytics](#analytics) that examines behavioural data about people to understand how and why individuals act the way they do and to make more accurate predictions for future behaviour.[8](#fn8),[9](#fn9) <#data-analysis>

### BIG DATA

Refers to the massive amounts of [data](#data) generated around the world that is too large, complex or varied for traditional processing software. Its potential to be analyzed for valuable information is enabled by technology such as [advanced research computing](#advanced-research-computing). [10](#fn10),[11](#fn11) <#disciplines>

### Big data isn’t always big.

Many describe it with “[The Four V’s](http://www.ibmbigdatahub.com/infographic/four-vs-big-data)” and there exist other definitions as well, including the [5 Vs](https://bicorner.com/2015/04/17/why-only-one-of-the-5-vs-of-big-data-really-matters/)’ or [7 V’s](http://dataconomy.com/seven-vs-big-data/). <#disciplines> <#lesson-04>

* Volume: includes the amount of data.
* Velocity: is data that arrives quickly and constantly, for example, streaming data.
* Variety: includes many differently-structured (or less-structured) input data sets.
* Veracity: is some data that might be incorrect or of unknown correctness.

### CAUSAL INFERENCE

Determines whether observations made in one variable are the reason for an effect observed in another variable, possibly occurring at a later time. See also [correlation mining](#correlation-mining) and [predictive analytics](#predictive-analytics). <#data-analysis>

### CLASSIFICATION

An approach in [machine learning](#machine-learning) where a program is [trained](#model-fitting) with labeled [data](#data) to determine which category a new observation belongs to. See also [clustering](#clustering). <#data-analysis>

### CLOUD COMPUTING

Access to data, [storage](#distributed-file-system), applications, and other computing resources made available to many users on-demand over the [Internet](#network) to improve [scalable computing power](#advanced-research-computing) and reliability. [12](#fn12),[13](#fn13) <#disciplines>

### CLUSTER

A [network](#network) of computers (or compute nodes) that work on tasks together in parallel. The concept is different from [clustering](#clustering).

### CLUSTERING

A [machine learning](#machine-learning) technique that groups similar data points together to uncover hidden structure. Unlike [classification](#classification), clustering does not require labelled training data and is a type of unsupervised learning. <#data-analysis>

### COMPUTING WITH DATA

Learning from data using computing tools and programming languages, such as Python or R with large ecosystems of libraries providing a convenient level of abstraction and statistical methods that can be organized into data processing pipelines.[14](#fn14) <#disciplines>

### CORRELATION MINING

As a measure of association between two variables, correlation can be used for [prediction](#predictive-analytics) and indicate the presence of [causal relationship](#causal-inference). It does, however, not sufficiently imply causation. [15](#fn15) <#data-analysis>

### CYBERSECURITY

A model in information security designed to govern and evaluate how an organization handles data when it is stored, transmitted or processed. This model emphasizes that data should not be accessed without authorization, should not be altered or compromised without authorization, and should be accessible upon legitimate request. See [Secure Computing](#secure-computing). [16](#fn16) <#disciplines>

### Critical Thinking

The objective analysis and evaluation of an issue to form a judgement [^43]. The process of analysing information in order to make a logical decision about the extent to which you believe something to be true or false [^44]. <#disciplines> <#lesson-04>

### Crowdsourcing

TODO. <#lesson-03>

### DATA

Collection of examples, observations, measurements, facts, points, or other items of information that can be represented in [structured](#structured-data) or [unstructured](#unstructured-data) form. <#data-types> <#lesson-03>

### DATA AGGREGATION

TODO:

### DATA CLEANING

Careful removal of erroneous or unreliable data points. <#processing>

### DATA EXPLORATION AND PREPARATION

Exploratory data analysis (EDA) is a formative step in the [creation of models](#model-fitting). [Views of the data](#visualization) are used to learn about patterns or relationships among variables. This includes [data cleaning](#data-cleaning) and manipulation for further analysis. <#processing>

### DATA INTEGRATION

The process of combining information from different data sources in preparation for data processing. [17](#fn17) <#processing>

### DATA MINING

An analytical process where large datasets are explored or “mined” in search of meaningful patterns, relationships or insights. The process can include statistics, [machine learning](#machine-learning) or other forms of artificial intelligence.[18](#fn18),[19](#fn19) <#disciplines>

### DATA REPRESENTATION

The form in which [data](#data) are stored, processed and transmitted such that its [information content](#information) and [context](#metadata) are retained as much as possible. Choices of form are influenced by hardware, software or other constraints around processing and analysis resources. [20](#fn20) <#processing>

### DATA SCIENCE

A multidisciplinary activity combining programming skills, math and statistical analysis, and sector-specific expertise to extract [insights](#insight) from [data](#data). Often performed in stages: 1. [Data Exploration and Preparation](#data-exploration-and-preparation), 2. [Data Representation](#data-representation) and [Transformation](#data-transformation), 3. [Computing with Data](#computing-with-data), 4. [Machine Learning](#machine-learning), 5. [Data Visualization and Presentation](#visualization). [21](#fn21),[22](#fn22) <#disciplines>

### DATA TRANSFORMATION

The process of converting data from [one form into another](#data-representation) to gain better [insight](#insight). For instance, converting [unstructured data](#unstructured-data) into [structured form](#structured-data) that can be analyzed further. <#processing> <#lesson-03> <#lesson-03> <#lesson-06> <#lesson-08> <#lesson-12>

### DATABASE

An organized collection of [data](#data) that allows easy access, management, updating and analysis of data. Commonly used databases are MySQL, PostgreSQL, as well as various [NoSQL](#no-sql-database) options.[23](#fn23) <#storage>

### DESCRIPTIVE ANALYTICS

An initial stage of data processing that involves creating a summary of historical data with the goal of answering the question, “What happened?”[24](#fn24),[25](#fn25) <#data-analysis>

### DISTRIBUTED FILE SYSTEM

A mechanism that stores files on servers and allows clients, with permission, to store and process files as if they were stored on their own computer.[26](#fn26),[27](#fn27) <#storage>

### Data (Clean, Dirty)

Data that contains erroneous or unreliable data points is called dirty data. Dirty data must be cleaned to correct and adjusted before it is usable, usually a tedious task. Causes of dirty data include being Outdated, insecure, incomplete, incomplete (missing), Inaccurate data, Misplaced data, Inconsistent, or duplicated.[28](#fn28) <#processing> <#lesson-04> <#lesson-06>

### Data Bias

Whether the (sample) data’s [balance](#balance) in the is representative [#sample] or not. If the balance is representative, then there is No Bias. When the balance is not representative, then the dataset is biased. <#data-analysis> <#lesson-04>

### Data Literacy

Data literacy is the intersection of verbal, numerical and graphical literacy. It is not label reserved for data scientists or specialists. Data literacy should be thought of as “the ability of non-specialists to make use of data” and measure “a person’s ability to read, work with, analyze and argue with data”, presumably using simple statistics such as means and percentages [29](#fn29). <#disciplines> <#lesson-01>

### Data Privacy

Refers to the vast range of technologies, protocols, and concepts that give individual users or other parties more privacy protections in their online use. Online privacy takes many forms, including mandatory privacy statements on websites, data sharing controls, data transparency initiatives and more. [30](#fn30) <#disciplines> <#lesson-04>

### Data Quality

The usefulness and reliability of data are assessed by its quality. High-quality data can be defined as appropriate, consistent, complete and accurate. TODO: Dimensions that the quality can be assessed on include relevance (does the statistical information matter?), accessibility (can users access the statistical information?), accuracy (is the statistical information representative of the targeted measurement?), timeliness (is the lag between the period of reference and the availability of the statistical information acceptable?), interpretability (is metadata available and complete?), and coherence (is the statistical information consistent over time, between regions and across sub-populations?). [^45] <#processing> <#lesson-04>

### Data types

TODO intro. <#data-types> <#lesson-03>

#### Float

Any number, whole or fraction, with unknown decimal places.

#### Integer

A whole number.

#### String

Multiple characters together are called strings. e.g ‘aeiou’ is a string of the characters ‘a’, ‘e’, ‘i’, ‘o’, ‘u’.

#### Date / Datetime

A representation of a date (may include time) e.g. YYYYMMDD or YYYYMMDDhhmmss.

### FEATURE

A key property that characterizes a data point (representing some real-world object) in the context of a [machine learning](#machine-learning) problem or other type of [analysis](#analytics).

### FEATURE ENGINEERING

A human-driven process of finding the most important [features](#feature) to develop [predictive](#predictive-analytics) models. Formerly, as “art” to the “science” of [machine learning](#machine-learning), [big data](#big-data) increasingly moves towards automated representation learning.

### FLEXIBLE SCHEMA

Unlike [SQL](#sql) databases with a [tabular](#tabular-data) [structure](#structured-data), objects or documents stored in a flexible schema can be different from one another.[31](#fn31) <#data-types>

### File Formats

TODO intro <#data-types> <#lesson-03>

#### TXT

A generic file format which may hold [structured](#structured-data) data. The structure is preserved with a delimiter or separator; comma, semicolon or tab indentations are common separators.

#### CSV

A text file that uses commas as a delimiter. Often the file extension is .csv but the generic .txt is also used. The structured in the text file are *comma*-separated values.

#### TSV

A text file that uses tab indentation as a delimiter. Often the file extension is .tsv but the generic .txt is also used. The structured in the text file are *tab*-separated values.

#### JSON

*JavaScript Object Notation* is a human-readable file format that stores [structured](#structured-data) data as attribute-value pairs. JSON is a language-independent data format and is a common format for use with web applications,

#### XML

*eXtensible Markup Language* is human-readable file format that stores [structured](#structured-data) data using *tags*, *elements* and *attributes*. The XML format is commonly used across the Internet.

#### YAML

*Yet Another Markup Language* is a human-readable file format that stores data using indents, hyphens (-) and colon punctuation (:) to impose [structured](#structured-data). YAML is a compact format and is commonly used to share configurations, settings and unsophisticated data files.

### GENERATIVE MODELING

[Fitting](#model-fitting) a [model](#machine-learning) that can generate synthetic data beyond [given observations](#data). For example, a language model trained on a collection of text can be used to suggest possible words to modify or continue a given phrase. [32](#fn32)

### GOODNESS OF FIT

Summarizes how well the values observed in the data agree with those values expected by the model. <#processing>

### INFORMATION

Meaning encoded in [data](#data) that answers questions to better understand a concept by interpreting data within the context of its problem setting or domain. See also data analysis [algorithms](#algorithm) and information [visualization](#visualization). [33](#fn33)

### INSIGHT

Actionable [information](#information) gained by interpreting [data analysis](#analytics) results, helping people to make more informed decisions. [34](#fn34)

### IOT (INTERNET OF THINGS)

A system of [connected](#network), “smart” objects, including smartphones, wearables and smart-appliances, that collect and exchange information without requiring human interaction. The IoT is a huge generator of data.[35](#fn35),[36](#fn36) <#storage>

### KNOWLEDGE DISCOVERY

Aims to extract [insight](#insight) from data in [databases](#database). It involves a [number of steps](#data-science) including the evaluation and possibly [interpretation](#visualization) of [patterns](#data-mining) to gather [insight](#insight) and knowledge.[37](#fn37) <#disciplines>

### MACHINE LEARNING

A part of [artificial intelligence](#artificial-intelligence) that enables machines to learn from experience to perform certain tasks by using [algorithmic](#algorithm) models that are [trained](#model-fitting) to imitate patterns present in [data](#data) to support [prediction](#predictive-analytics), [data generation](#generative-modeling) and other forms of [data analytics](#analytics). <#disciplines> <#lesson-03>

### METADATA

Also called “data about data,” it provides a [structured description](#structured-data) and context for a data point - document, image, or file - to help organize, find and understand the data. [38](#fn38) <#data-types> <#lesson-03>

### MODEL FITTING

Optimizes [how well](#goodness-of-fit) a [machine learning model](#machine-learning) can make predictions for previously unseen test data.

### NETWORK

A system where multiple computing devices are connected to each other to exchange information and resources through a data link. For example, the Internet.[39](#fn39),[40](#fn40)

### NETWORK ANALYSIS

Mapping and measuring the relationships between people, groups, organizations, computers and other connected entities. It is used to simplify complex relationships, to make them easier to analyze. [41](#fn41)

### NO-SQL DATABASE

A type of database designed to handle [large](#advanced-research-computing) volumes of data that [may not have a structure](#flexible-schema). <#storage>

### PREDICTIVE ANALYTICS

A process for analyzing current [data](#data) to determine future events or other unknowns. Related to [descriptive](#descriptive-analytics) and [prescriptive analytics](#prescriptive-analytics), it draws on techniques from [data mining](#data-mining), modeling, [machine learning](#machine-learning) and statistics.[42](#fn42),[43](#fn43)

### PRESCRIPTIVE ANALYTICS

A type of [analytics](#analytics) with the goal of using data to determine the best course of action for a specific scenario. [44](#fn44) <#data-analysis>

### Population

A collection of items under consideration. [45](#fn45) <#data-analysis> <#lesson-04>

### REGRESSION

A type of [statistical model](#machine-learning) that [predicts](#predictive-analytics) numerical values (instead of [class labels](#classification)). <#data-analysis>

### RELATIONSHIP MINING

Relationship mining examines associations between two or more variables in a dataset, for example, by [correlation mining](#correlation-mining) and [causal mining](#causal-inference). <#data-analysis>

### SECURE COMPUTING

Efforts to ensure privacy and to protect data, devices and computing systems from harm such as hacking, damage, and malpractice, and mitigate service disruptions. [46](#fn46),[47](#fn47) <#disciplines>

### SQL

SQL is a language that works with structured, [tabular data](#tabular-data) allowing to query and construct such data. It forms a standard for many [database](#database) systems. [28] <#processing>

### STRUCTURED DATA

[Data](#data) that is organized into clearly defined fields associated with variables or attributes, such as dates, words, or numbers that are recorded for each observation or item. Items are often represented as rows of [tabular spreadsheets](#tabular-data) that can be stored in a [database](#database) for easy processing and analysis.[48](#fn48),[49](#fn49) <#data-types> <#lesson-03>

### Sample (Representative, Proxy)

A subset of the units of a [population](#population). The sample serves as a proxy, or representative sample subset of the true nature of the [population](#population). [50](#fn50) <#data-analysis> <#lesson-04>

### TABULAR DATA

Data items or observations that are organized as rows that contain values under columns that correspond to specific variables or properties. [51](#fn51),[52](#fn52) <#data-types> <#lesson-03> - Synonyms: row(s), record(s), example(s), observation(s), measurement(s), fact(s), point(s) and other items of information - Synonyms: column(s) also referred to as field(s), header(s), attribute(s), [feature](#feature)(s), property(ies), variable(s), key(s) (as in key-value), factor(s), dimension(s) - Synonyms: pairs, key-value, attribute-value

#### Narrow table format

TODO: one column contains all the attributes and another column contains the lists of values. <#data-types> <#lesson-03>

#### Wide table format

TODO: each holds different attributes in separate columns. <#data-types> <#lesson-03>

### TEXT MINING

A process of analyzing [text](#unstructured-data) to capture key concepts, themes, relationships and trends.[53](#fn53) <#data-analysis>

### Tree structure

A *tree* is a widely used data type that simulates a hierarchical tree structure, with a root value, branches, leaves, and subtrees of children with a parent node. [54](#fn54) <#data-types> <#data-types> <#lesson-03>

### UNSTRUCTURED DATA

Data that is not organized in a pre-defined way under a single data model. Examples include text, images, audio, or video. [55](#fn55),[56](#fn56) <#data-types>, <#lesson-03>

### Unbalanced Data

Whether majority or minorities groups exist in a dataset; imbalance is common and sometimes expected. <#lesson-04>

### VISUALIZATION

Visual representation of data including plots, charts, maps and infographics to support people with tasks such as sensemaking, effective communication of information, and improved pattern detection. For example, flight operations using a radar screen that displays sensor data in real-time. [57](#fn57),[58](#fn58) <#data-analysis>

### VISUALIZATION DASHBOARD

Multiple interactive graphical views of [data](#data) that would be too complex if presented in a [single visualization](#visualization), for instance, to support complex narratives and [insights](#insight) by showing key performance indicators (KPIs) for monitoring of a business or other processes. <#data-analysis>

# Categories

The categories offer an alternative grouping that helps to find related terms.

## Data Inquiry Disciplines

[ADVANCED RESEARCH COMPUTING](#advanced-research-computing)

[ANALYTICS](#analytics)

[ARTIFICIAL INTELLIGENCE](#artificial-intelligence)

[BIG DATA](#big-data)

[Big data isn’t always big.](#big-data-isnt-always-big.)

[CLOUD COMPUTING](#cloud-computing)

[COMPUTING WITH DATA](#computing-with-data)

[CYBERSECURITY](#cybersecurity)

[Critical Thinking](#critical-thinking)

[DATA MINING](#data-mining)

[DATA SCIENCE](#data-science)

[Data Literacy](#data-literacy)

[Data Privacy](#data-privacy)

[KNOWLEDGE DISCOVERY](#knowledge-discovery)

[MACHINE LEARNING](#machine-learning)

[SECURE COMPUTING](#secure-computing)

## Data Analysis Techniques

[BEHAVIOUR(AL) ANALYTICS](#behavioural-analytics)

[CAUSAL INFERENCE](#causal-inference)

[CLASSIFICATION](#classification)

[CLUSTERING](#clustering)

[CORRELATION MINING](#correlation-mining)

[DESCRIPTIVE ANALYTICS](#descriptive-analytics)

[Data Bias](#data-bias)

[PRESCRIPTIVE ANALYTICS](#prescriptive-analytics)

[Population](#population)

[REGRESSION](#regression)

[RELATIONSHIP MINING](#relationship-mining)

[Sample (Representative, Proxy)](#sample-representative-proxy)

[TEXT MINING](#text-mining)

[VISUALIZATION](#visualization)

[VISUALIZATION DASHBOARD](#visualization-dashboard)

## Data Processing

[ALGORITHM](#algorithm)

[DATA CLEANING](#data-cleaning)

[DATA EXPLORATION AND PREPARATION](#data-exploration-and-preparation)

[DATA INTEGRATION](#data-integration)

[DATA REPRESENTATION](#data-representation)

[DATA TRANSFORMATION](#data-transformation)

[Data (Clean, Dirty)](#data-clean-dirty)

[Data Quality](#data-quality)

[GOODNESS OF FIT](#goodness-of-fit)

[SQL](#sql)

## Storage and Physical Data Sources

[DATABASE](#database)

[DISTRIBUTED FILE SYSTEM](#distributed-file-system)

[IOT (INTERNET OF THINGS)](#iot-internet-of-things)

[NO-SQL DATABASE](#no-sql-database)

## Types of Data

[DATA](#data)

[Data types](#data-types)

[FLEXIBLE SCHEMA](#flexible-schema)

[File Formats](#file-formats)

[METADATA](#metadata)

[STRUCTURED DATA](#structured-data)

[TABULAR DATA](#tabular-data)

[Tree structure](#tree-structure)

[UNSTRUCTURED DATA](#unstructured-data)

# Lessons Index

## Lessons 01 and 02

[Data Literacy](#data-literacy)

## Lesson 03

[ARTIFICIAL INTELLIGENCE](#artificial-intelligence)

[Crowdsourcing](#crowdsourcing)

[DATA](#data)

[DATA TRANSFORMATION](#data-transformation)

[Data types](#data-types)

[File Formats](#file-formats)

[MACHINE LEARNING](#machine-learning)

[METADATA](#metadata)

[STRUCTURED DATA](#structured-data)

[TABULAR DATA](#tabular-data)

[Tree structure](#tree-structure)

[UNSTRUCTURED DATA](#unstructured-data)

## Lesson 04

[Big data isn’t always big.](#big-data-isnt-always-big.)

[Critical Thinking](#critical-thinking)

[Data (Clean, Dirty)](#data-clean-dirty)

[Data Bias](#data-bias)

[Data Privacy](#data-privacy)

[Data Quality](#data-quality)

[Population](#population)

[Sample (Representative, Proxy)](#sample-representative-proxy)

[Unbalanced Data](#unbalanced-data)

# References

[^43] “Critical Thinking”. [Online]. Available at: https://www.lexico.com/definition/critical\_thinking

[^44] “Critical Thinking”. [Online]. Available at: https://www.oxfordlearnersdictionaries.com/definition/english/critical-thinking

[^45] “Data, statistical information and statistics: Data quality” [Online]. Available: https://www150.statcan.gc.ca/n1/edu/power-pouvoir/ch3/quality-qualite/5214788-eng.htm.

1. “Westgrid: What We Do.” [Online]. Available: <https://www.westgrid.ca/about_westgrid/what_we_do>[↩︎](#fnref1)
2. “Compute Canada. Update January 11, 2017.” [Online]. Available: <https://www.computecanada.ca/wp-content/uploads/2015/02/Update-January-11-2017.pdf>[↩︎](#fnref2)
3. “A beginner’s guide to big data terminology.” [Online]. Available: <http://dataconomy.com/2016/05/a-beginners-guide-to-big-data-terminology>[↩︎](#fnref3)
4. “Mini-glossary: Big data terms you should know.” [Online]. Available: <http://www.techrepublic.com/article/mini-glossary-big-data-terms-you-should-know>[↩︎](#fnref4)
5. “A beginner’s guide to big data terminology.” [Online]. Available: <http://dataconomy.com/2016/05/a-beginners-guide-to-big-data-terminology>[↩︎](#fnref5)
6. “Mini-glossary: Big data terms you should know.” [Online]. Available: <http://www.techrepublic.com/article/mini-glossary-big-data-terms-you-should-know>[↩︎](#fnref6)
7. “Artificial Intelligence” [Online]. Available: https://www.techopedia.com/definition/190/artificial-intelligence-ai[↩︎](#fnref7)
8. “Behavioral analytics, definition by Techopedia.” [Online]. Available: <https://www.techopedia.com/definition/30308/behavioral-analytics>[↩︎](#fnref8)
9. “Detecting advanced threats with user behavior analytics.” [Online]. Available: <http://www.networkworld.com/article/2904356/security0/detecting-advanced-threats-with-user-behavior-analytics.html>[↩︎](#fnref9)
10. “Big data definition by Tech Target.” [Online]. Available: <http://searchcloudcomputing.techtarget.com/definition/big-data-Big-Data>[↩︎](#fnref10)
11. “What Is Big Data?” [Online]. Available: <https://www.forbes.com/sites/lisaarthur/2013/08/15/what-is-big-data/#4fb6509b5c85>[↩︎](#fnref11)
12. “Who Coined ‘Cloud Computing’?” [Online]. Available: <https://www.technologyreview.com/s/425970/who-coined-cloud-computing>[↩︎](#fnref12)
13. “Cloud computing: a simple introduction.” [Online]. Available: <http://www.explainthatstuff.com/cloud-computing-introduction.html>[↩︎](#fnref13)
14. Donoho, D. (2017). 50 years of data science. *Journal of Computational and Graphical Statistics*, *26*(4), 745-766.[↩︎](#fnref14)
15. “Introduction to Correlation” [Online]. Available: https://www.datascience.com/blog/introduction-to-correlation-learn-data-science-tutorials[↩︎](#fnref15)
16. “EI-ISAC Cybersecurity Spotlight – CIA Triad” [Online]. Available: https://www.cisecurity.org/spotlight/ei-isac-cybersecurity-spotlight-cia-triad/[↩︎](#fnref16)
17. “Mini-glossary: Big data terms you should know.” [Online]. Available: <http://www.techrepublic.com/article/mini-glossary-big-data-terms-you-should-know>[↩︎](#fnref17)
18. “A beginner’s guide to big data terminology.” [Online]. Available: <http://dataconomy.com/2016/05/a-beginners-guide-to-big-data-terminology>[↩︎](#fnref18)
19. “Mini-glossary: Big data terms you should know.” [Online]. Available: <http://www.techrepublic.com/article/mini-glossary-big-data-terms-you-should-know>[↩︎](#fnref19)
20. Donoho, D. (2017). 50 years of data science. *Journal of Computational and Graphical Statistics*, *26*(4), 745-766.[↩︎](#fnref20)
21. Donoho, D. (2017). 50 years of data science. *Journal of Computational and Graphical Statistics*, *26*(4), 745-766.[↩︎](#fnref21)
22. “The Data Science Venn Diagram” [Online]. Available: <http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram>[↩︎](#fnref22)
23. “A beginner’s guide to big data terminology.” [Online]. Available: <http://dataconomy.com/2016/05/a-beginners-guide-to-big-data-terminology>[↩︎](#fnref23)
24. “Descriptive analytics definition by Tech Target.” [Online]. Available: <http://whatis.techtarget.com/definition/descriptive-analytics>[↩︎](#fnref24)
25. “Four Types of Big Data Analytics and Examples of Their Use.” [Online]. Available: <http://www.ingrammicroadvisor.com/data-center/four-types-of-big-data-analytics-and-examples-of-their-use>[↩︎](#fnref25)
26. “Distributed file system definition by Tech Target.” [Online]. Available: <http://searchwindowsserver.techtarget.com/definition/distributed-file-system-DFS>[↩︎](#fnref26)
27. “Distributed file system definition by Techopedia.” [Online]. Available: <https://www.techopedia.com/definition/1825/distributed-file-system-dfs>[↩︎](#fnref27)
28. “The 7 Most Common Types of Dirty Data (and how to clean them)”. [Online]. Available: https://www.ringlead.com/blog/the-7-most-common-types-of-dirty-data-and-how-to-clean-them[↩︎](#fnref28)
29. “Data Literacy: What It Is and How to Measure It in the Public Service”. Statistics Canada. [Online]. Available at: https://www150.statcan.gc.ca/n1/pub/11-633-x/11-633-x2019003-eng.htm[↩︎](#fnref29)
30. “Internet Privacy.” [Online]. Available: https://www.techopedia.com/definition/24954/internet-privacy[↩︎](#fnref30)
31. “Data Modeling Introduction” [Online]. Available: <https://docs.mongodb.com/manual/core/data-modeling-introduction/>[↩︎](#fnref31)
32. Donoho, D. (2017). 50 years of data science. *Journal of Computational and Graphical Statistics*, *26*(4), 745-766.[↩︎](#fnref32)
33. “Data vs. Information vs. Insight” [Online]. Available: https://online.ben.edu/programs/mba/resources/data-vs-information-vs-insight[↩︎](#fnref33)
34. “Data vs. Information vs. Insight” [Online]. Available: https://online.ben.edu/programs/mba/resources/data-vs-information-vs-insight[↩︎](#fnref34)
35. “A Simple Explanation Of ‘The Internet Of Things’.” [Online]. Available: <https://www.forbes.com/sites/jacobmorgan/2014/05/13/simple-explanation-internet-things-that-anyone-can-understand/#34e4f89b1d09>[↩︎](#fnref35)
36. “Internet of things definition by Tech Target.” [Online]. Available: <http://internetofthingsagenda.techtarget.com/definition/Internet-of-Things-IoT>[↩︎](#fnref36)
37. “Overview of the KDD Process” [Online]. Available: http://www2.cs.uregina.ca/~dbd/cs831/notes/kdd/1\_kdd.html[↩︎](#fnref37)
38. Greenberg, J. (2003). Metadata and the world wide web. *Encyclopedia of library and information science*, *3*, 1876-1888.[↩︎](#fnref38)
39. “Computer network definition by Techopedia.” [Online]. Available: <https://www.techopedia.com/definition/25597/computer-network>[↩︎](#fnref39)
40. “What is computer networking?” [Online]. Available: <https://www.lifewire.com/what-is-computer-networking-816249) [www.lifewire.com/what-is-computer-networking-816249.](https://www.lifewire.com/what-is-computer-networking-816249>[↩︎](#fnref40)
41. “Social Network Analysis: An Introduction by Orgnet,LLC.” [Online]. Available: <http://www.orgnet.com/sna.html>[↩︎](#fnref41)
42. “A beginner’s guide to big data terminology.” [Online]. Available: <http://dataconomy.com/2016/05/a-beginners-guide-to-big-data-terminology>[↩︎](#fnref42)
43. Donoho, D. (2017). 50 years of data science. *Journal of Computational and Graphical Statistics*, *26*(4), 745-766.[↩︎](#fnref43)
44. “A beginner’s guide to big data terminology.” [Online]. Available: <http://dataconomy.com/2016/05/a-beginners-guide-to-big-data-terminology>[↩︎](#fnref44)
45. “Population”. [Online]. Available at: https://www.lexico.com/definition/population[↩︎](#fnref45)
46. “The myth of secure computing.” [Online]. Available: <https://hbr.org/2003/06/the-myth-of-secure-computing) [hbr.org/2003/06/the-myth-of-secure-computing.](https://hbr.org/2003/06/the-myth-of-secure-computing>[↩︎](#fnref46)
47. “Mini-glossary: Big data terms you should know.” [Online]. Available: <http://www.techrepublic.com/article/mini-glossary-big-data-terms-you-should-know>[↩︎](#fnref47)
48. “Introduction to Structured Data, by Google.” [Online]. Available: <https://developers.google.com/search/docs/guides/intro-structured-data>[↩︎](#fnref48)
49. “Structured data definition by Webopedia.” [Online]. Available: <http://www.webopedia.com/TERM/S/structured_data.html>[↩︎](#fnref49)
50. “Statistics: Power from Data! Glossary” [Online]. Available: https://www150.statcan.gc.ca/n1/edu/power-pouvoir/glossary-glossaire/5214842-eng.htm#Sample[↩︎](#fnref50)
51. “Introduction to Tabular Data” [Online]. Available: <https://papl.cs.brown.edu/2016/intro-tabular-data.html>[↩︎](#fnref51)
52. “Tabular Database” [Online]. Available: <https://www.techopedia.com/definition/26181/tabular-database>[↩︎](#fnref52)
53. “About Text Mining” [Online]. Available: https://www.ibm.com/support/knowledgecenter/en/SS3RA7\_15.0.0/com.ibm.spss.ta.help/tm\_intro\_tm\_defined.htm[↩︎](#fnref53)
54. “Tree” from the National Institute of Standards and Technology. [Online]. Available at: https://xlinux.nist.gov/dads/HTML/tree.html[↩︎](#fnref54)
55. “Unstructured data definition by Techopedia.” [Online]. Available: <https://www.techopedia.com/definition/13865/unstructured-data>[↩︎](#fnref55)
56. “Solving the Unstructured Data Challenge.” [Online]. Available: <http://www.cio.com/article/2941015/big-data/solving-the-unstructured-data-challenge.html>[↩︎](#fnref56)
57. “Data visualization definition by Tech Target.” [Online]. Available: <http://searchbusinessanalytics.techtarget.com/definition/data-visualization>[↩︎](#fnref57)
58. “Data visualization for human perception.” [Online]. Available: <https://www.interaction-design.org/literature/book/the-encyclopedia-of-human-computer-interaction-2nd-ed/data-visualization-for-human-perception>[↩︎](#fnref58)