Data Literacy: Glossary and Index

Table of Contents

[Glossary 6](#_Toc97808520)

[Advanced Research Computing 6](#_Toc97808521)

[Algorithm 6](#_Toc97808522)

[Analytics 6](#_Toc97808523)

[API 7](#_Toc97808524)

[Arithmetic operations 7](#_Toc97808525)

[Artificial Intelligence 7](#_Toc97808526)

[Bar Chart 7](#_Toc97808527)

[Behaviour(al) Analytics 7](#_Toc97808528)

[Big Data 7](#_Toc97808529)

[Binning 7](#_Toc97808530)

[Box Plot 7](#_Toc97808531)

[Bucketing 7](#_Toc97808532)

[Causal Inference 8](#_Toc97808533)

[Chartjunk 8](#_Toc97808534)

[Classification 8](#_Toc97808535)

[Cloud (“The Cloud”) 8](#_Toc97808536)

[Cloud Computing 8](#_Toc97808537)

[Cloud Storage 8](#_Toc97808538)

[Cluster (Compute) 8](#_Toc97808539)

[Clustering 8](#_Toc97808540)

[Colour Map 8](#_Toc97808541)

[Computing with Data 9](#_Toc97808542)

[Correlation 9](#_Toc97808543)

[Correlation Mining 9](#_Toc97808544)

[Critical Thinking 9](#_Toc97808545)

[Crowdsourcing 9](#_Toc97808546)

[Cybersecurity 9](#_Toc97808547)

[Data 9](#_Toc97808548)

[Data (Clean, Dirty) 9](#_Toc97808549)

[Data Aggregation 9](#_Toc97808550)

[Data Analysis 10](#_Toc97808551)

[Data Bias 10](#_Toc97808552)

[Data Cleaning 10](#_Toc97808553)

[Data Exploration and Preparation 10](#_Toc97808554)

[Data Integration 10](#_Toc97808555)

[Data Journalism 10](#_Toc97808556)

[Data Lake 10](#_Toc97808557)

[Data Literacy 10](#_Toc97808558)

[Data Mining 10](#_Toc97808559)

[Data Model 11](#_Toc97808560)

[Data Privacy 11](#_Toc97808561)

[Data Quality 11](#_Toc97808562)

[Data Representation 11](#_Toc97808563)

[Data Science 11](#_Toc97808564)

[Data Transformation 11](#_Toc97808565)

[Data types 11](#_Toc97808566)

[Data Warehouse 12](#_Toc97808567)

[Data-to-ink ratio 12](#_Toc97808568)

[Database 12](#_Toc97808569)

[Database Model 12](#_Toc97808570)

[Dendrogram 12](#_Toc97808571)

[Descriptive Analytics 12](#_Toc97808572)

[Distributed File System 12](#_Toc97808573)

[Entity–Relationship Model 12](#_Toc97808574)

[Feature 13](#_Toc97808575)

[Feature Engineering 13](#_Toc97808576)

[File Formats 13](#_Toc97808577)

[Filter 14](#_Toc97808578)

[Flexible Schema 14](#_Toc97808579)

[Foreign key 14](#_Toc97808580)

[Generative Modeling 14](#_Toc97808581)

[GIS 14](#_Toc97808582)

[Goodness of Fit 14](#_Toc97808583)

[Group by 14](#_Toc97808584)

[Heatmap 14](#_Toc97808585)

[Histogram 14](#_Toc97808586)

[HTML 15](#_Toc97808587)

[Infographics 15](#_Toc97808588)

[Information 15](#_Toc97808589)

[Insight 15](#_Toc97808590)

[IOT (Internet of Things) 15](#_Toc97808591)

[Join 15](#_Toc97808592)

[Knowledge Discovery 15](#_Toc97808593)

[Lie factor 15](#_Toc97808594)

[Line Plot 15](#_Toc97808595)

[Logical operations 16](#_Toc97808596)

[Machine Learning 16](#_Toc97808597)

[Metadata 16](#_Toc97808598)

[Model 16](#_Toc97808599)

[Model Fitting 16](#_Toc97808600)

[Network 16](#_Toc97808601)

[Network Analysis 16](#_Toc97808602)

[No-SQL Database 16](#_Toc97808603)

[Null values 17](#_Toc97808604)

[Pair plot 17](#_Toc97808605)

[Pictographs Plot 17](#_Toc97808606)

[Pie Chart 17](#_Toc97808607)

[Pivot 17](#_Toc97808608)

[Population 17](#_Toc97808609)

[Predictive Analytics 17](#_Toc97808610)

[Prescriptive Analytics 17](#_Toc97808611)

[Primary key 17](#_Toc97808612)

[Query (Data) 17](#_Toc97808613)

[Regression 18](#_Toc97808614)

[Relationship Mining 18](#_Toc97808615)

[Repository (data repo) 18](#_Toc97808616)

[Sample (Representative, Proxy) 18](#_Toc97808617)

[Scatter Plot 18](#_Toc97808618)

[Schema (Database) 18](#_Toc97808619)

[Secure Computing 18](#_Toc97808620)

[Select 18](#_Toc97808621)

[Skewness 18](#_Toc97808622)

[Sort 18](#_Toc97808623)

[SQL 19](#_Toc97808624)

[Standard Deviation 19](#_Toc97808625)

[Statistics 19](#_Toc97808626)

[Storytelling 19](#_Toc97808627)

[String or character operations 19](#_Toc97808628)

[Structured Data 19](#_Toc97808629)

[Tabular Data 19](#_Toc97808630)

[Text Mining 20](#_Toc97808631)

[Tidy tables 20](#_Toc97808632)

[Time Series 20](#_Toc97808633)

[Transpose 20](#_Toc97808634)

[Tree structure 20](#_Toc97808635)

[Unbalanced Data 20](#_Toc97808636)

[Union 20](#_Toc97808637)

[Unstructured Data 21](#_Toc97808638)

[URL 21](#_Toc97808639)

[Variance 21](#_Toc97808640)

[Violin Plot 21](#_Toc97808641)

[Visualization 21](#_Toc97808642)

[Visualization Dashboard 21](#_Toc97808643)

[Web scraping 21](#_Toc97808644)

[Word Cloud 21](#_Toc97808645)

[Categories 22](#_Toc97808646)

[Data Inquiry Disciplines 22](#_Toc97808647)

[Data Analysis Techniques 22](#_Toc97808648)

[Data Processing 23](#_Toc97808649)

[Storage and Physical Data Sources 24](#_Toc97808650)

[Types of Data 24](#_Toc97808651)

[Operations 25](#_Toc97808652)

[Lessons Index 25](#_Toc97808653)

[Lessons 01 and 02 25](#_Toc97808654)

[Data Literacy 25](#_Toc97808655)

[Lesson 03 26](#_Toc97808656)

[Artificial Intelligence 26](#_Toc97808657)

[Crowdsourcing 26](#_Toc97808658)

[Data 26](#_Toc97808659)

[Data Transformation 26](#_Toc97808660)

[Data types 26](#_Toc97808661)

[File Formats 26](#_Toc97808662)

[Machine Learning 27](#_Toc97808663)

[Metadata 27](#_Toc97808664)

[Pivot 27](#_Toc97808665)

[Structured Data 28](#_Toc97808666)

[Tabular Data 28](#_Toc97808667)

[Tree structure 28](#_Toc97808668)

[Unstructured Data 28](#_Toc97808669)

[Lesson 04 28](#_Toc97808670)

[Big Data 28](#_Toc97808671)

[Critical Thinking 28](#_Toc97808672)

[Data (Clean, Dirty) 29](#_Toc97808673)

[Data Bias 29](#_Toc97808674)

[Data Privacy 29](#_Toc97808675)

[Data Quality 29](#_Toc97808676)

[Population 29](#_Toc97808677)

[Sample (Representative, Proxy) 29](#_Toc97808678)

[Unbalanced Data 29](#_Toc97808679)

[Lesson 05 29](#_Toc97808680)

[Correlation 29](#_Toc97808681)

[Histogram 30](#_Toc97808682)

[Standard Deviation 30](#_Toc97808683)

[Statistics 30](#_Toc97808684)

[Variance 30](#_Toc97808685)

[Visualization 30](#_Toc97808686)

[Lesson 06 30](#_Toc97808687)

[Data (Clean, Dirty) 30](#_Toc97808688)

[Data Integration 30](#_Toc97808689)

[Data Transformation 30](#_Toc97808690)

[Join 31](#_Toc97808691)

[Null values 31](#_Toc97808692)

[Tidy tables 31](#_Toc97808693)

[Lesson 07 31](#_Toc97808694)

[Filter 31](#_Toc97808695)

[Group by 31](#_Toc97808696)

[Query (Data) 31](#_Toc97808697)

[Select 31](#_Toc97808698)

[Sort 31](#_Toc97808699)

[SQL 32](#_Toc97808700)

[Lesson 08 32](#_Toc97808701)

[Arithmetic operations 32](#_Toc97808702)

[Binning 32](#_Toc97808703)

[Bucketing 32](#_Toc97808704)

[Data Aggregation 32](#_Toc97808705)

[Data Transformation 32](#_Toc97808706)

[Join 32](#_Toc97808707)

[Logical operations 32](#_Toc97808708)

[Pivot 32](#_Toc97808709)

[String or character operations 33](#_Toc97808710)

[Time Series 33](#_Toc97808711)

[Transpose 33](#_Toc97808712)

[Lesson 09 33](#_Toc97808713)

[Bar Chart 33](#_Toc97808714)

[Box Plot 33](#_Toc97808715)

[Correlation 33](#_Toc97808716)

[Data Analysis 33](#_Toc97808717)

[Dendrogram 34](#_Toc97808718)

[Heatmap 34](#_Toc97808719)

[Histogram 34](#_Toc97808720)

[Line Plot 34](#_Toc97808721)

[Model 34](#_Toc97808722)

[Pair plot 34](#_Toc97808723)

[Pie Chart 34](#_Toc97808724)

[Scatter Plot 34](#_Toc97808725)

[Skewness 35](#_Toc97808726)

[Standard Deviation 35](#_Toc97808727)

[Violin Plot 35](#_Toc97808728)

[Visualization 35](#_Toc97808729)

[Word Cloud 35](#_Toc97808730)

[Lesson 10 35](#_Toc97808731)

[API 35](#_Toc97808732)

[Cloud (“The Cloud”) 35](#_Toc97808733)

[Cluster (Compute) 35](#_Toc97808734)

[Data Integration 36](#_Toc97808735)

[Database 36](#_Toc97808736)

[GIS 36](#_Toc97808737)

[HTML 36](#_Toc97808738)

[Query (Data) 36](#_Toc97808739)

[Repository (data repo) 36](#_Toc97808740)

[SQL 36](#_Toc97808741)

[URL 36](#_Toc97808742)

[Web scraping 36](#_Toc97808743)

[Lesson 11 37](#_Toc97808744)

[Cloud Storage 37](#_Toc97808745)

[Data Integration 37](#_Toc97808746)

[Data Lake 37](#_Toc97808747)

[Data Model 37](#_Toc97808748)

[Data Warehouse 37](#_Toc97808749)

[Database 37](#_Toc97808750)

[Database Model 37](#_Toc97808751)

[Entity–Relationship Model 37](#_Toc97808752)

[Metadata 37](#_Toc97808753)

[Query (Data) 38](#_Toc97808754)

[Schema (Database) 38](#_Toc97808755)

[SQL 38](#_Toc97808756)

[Lesson 12 38](#_Toc97808757)

[Data Integration 38](#_Toc97808758)

[Data Transformation 38](#_Toc97808759)

[Foreign key 38](#_Toc97808760)

[Join 38](#_Toc97808761)

[Primary key 38](#_Toc97808762)

[SQL 38](#_Toc97808763)

[Union 39](#_Toc97808764)

[Lesson 13 39](#_Toc97808765)

[Bar Chart 39](#_Toc97808766)

[Data Journalism 39](#_Toc97808767)

[Histogram 39](#_Toc97808768)

[Infographics 39](#_Toc97808769)

[Line Plot 39](#_Toc97808770)

[Pie Chart 39](#_Toc97808771)

[Scatter Plot 39](#_Toc97808772)

[Storytelling 40](#_Toc97808773)

[Time Series 40](#_Toc97808774)

[Visualization 40](#_Toc97808775)

[Lesson 14 40](#_Toc97808776)

[Bar Chart 40](#_Toc97808777)

[Chartjunk 40](#_Toc97808778)

[Colour Map 40](#_Toc97808779)

[Data-to-ink ratio 40](#_Toc97808780)

[Histogram 40](#_Toc97808781)

[Lie factor 41](#_Toc97808782)

[Line Plot 41](#_Toc97808783)

[Pictographs Plot 41](#_Toc97808784)

[Pie Chart 41](#_Toc97808785)

[Scatter Plot 41](#_Toc97808786)

[Lesson 15 41](#_Toc97808787)

[References 41](#_Toc97808788)

# 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>

### API

API is an acronym for Application Programming Interface. Through the API, a Program (like a website or code) accesses an Application (a database or another application) through a common language, protocol (or Interface) through which the two systems interact. <#lesson-10>

### Arithmetic operations

Includes basic statistics: such as *sums*, *counts*, *means* (or *averages*), *medians*, *percentiles*, or [*standard deviations*](#standard-deviation). <#operations> <#lesson-08>

### 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>

### Bar Chart

A type of data [visualization](#visualization) that uses an array of bars (or rectangles) that share a common baseline, but vary in height to represent a set of data values. A horizontal version is also common, giving room for longer text labels. Multiple bars for a single point can be grouped or stacked. <#lesson-09> <#lesson-13> <#lesson-14>

### 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> <#lesson-04>

### Binning

An operation on a range of numbers in which the entire range is divided into *bins* (or *intervals*) of a *bin width* (or *bin size*, *step size*). Another name for binning is *bucketing* <#operations> <#lesson-08>

### Box Plot

A type of data [visualization](#visualization) that shows how values of a variable are distributed. Line markers along an axis show how far the lower and upper quartiles are from the [median](#arithmetic-operations) giving an idea of central tendency, dispersion, and skew. See also [histogram](#histogram) and [violin plot](#violin-plot). <#lesson-09>

### Bucketing

Another name for bucketing is [binning](#binning). <#operations> <#lesson-08>

### 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>

### Chartjunk

Elements of a graphic that distract or are not needed to communicate information ([data](#data)). Elements, such as, unnecessary text, grid lines and frames, ornamental icons, pictures or backgrounds, gradient [colours](#colour) or shading, gratuitous 3D effects and gimmick fonts. See also [data-to-ink-ratio](#data-to-ink-ratio) and [lie-factor](#lie-factor). [12](#fn12) <#lesson-14>

### Classification

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

### Cloud (“The Cloud”)

Any resource stored on the internet (instead of locally on a PC) is said to be stored on the cloud. Cloud storage is convenient in that clients do not have to manage the storage and infrastructure; they just access it from other internet-connected devices. <#disciplines> <#lesson-10>

### 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. [13](#fn13),[14](#fn14) <#disciplines>

### Cloud Storage

Stores data on the Internet with a [cloud computing](#cloud-computing) provider who manages and operates data storage as a service. [15](#fn15) <#storage> <#lesson-11>

### Cluster (Compute)

A [network](#network) of computers (or compute nodes) that work on tasks together in parallel. A cluster is one way to address [Big Data](#big-data) needs. The concept is different from [clustering](#clustering). <#lesson-10>

### 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>

### Colour Map

[Visual encoding](#visualization) of different levels of a variable by using a fixed scale (or map) of colours. Synonyms: Colour palette, Colour scheme. Colour (Hue), Colou (Shade) [16](#fn16) <#lesson-14>

### 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.[17](#fn17) <#disciplines>

### Correlation

Correlation is a relationship between two variables in a dataset, indicating that their values tend to change proportionally to each other. While it is a measure of association, it does not imply [causation](#causal-inference). <#data-analysis> <#lesson-05> <#lesson-09>

### Correlation Mining

As a measure of association between two variables, correlation can be used for [prediction](#predictive-analytics) and indicate the presence of a [causal relationship](#causal-inference). It does, however, not sufficiently imply causation. [18](#fn18) <#data-analysis>

### Critical Thinking

A thought process to determine whether we believe something to be true or false by logically analyzing and evaluating evidence and other information. [19](#fn19) <#disciplines> <#lesson-04>

### Crowdsourcing

A collaborative approach for collection and maintenance of data where people use websites to share information of common interest. The result can be large, high-quality datasets. <#lesson-03>

### 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). [20](#fn20) <#disciplines>

### 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 (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, misplaced, inconsistent, or duplicated. [21](#fn21) <#processing> <#lesson-04> <#lesson-06>

### Data Aggregation

A [group-by](#group-by) operation that reduces data. Values of a variable (or column) that are observed for a group of points (or rows of a [table](#tabular-data)) are combined into a single value. [22](#fn22) <#operations>, <#lesson-08>

### Data Analysis

Data analysis is a process of inspecting, [cleansing](#clean-dirty), [transforming](#data-transformation), and modelling data to discover useful information, informing conclusions, and supporting [decision-making](#critical-thinking). <#analytics> <#lesson-09>

### Data Bias

Whether the (sample) data’s [balance](#balance) is representative [sample](#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 Cleaning

Careful removal of erroneous or unreliable data points. Also see [clean and dirty data](#clean-dirty). <#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. [23](#fn23) <#processing> <#lesson-06> <#lesson-10> <#lesson-11> <#lesson-12>

### Data Journalism

Refers to the field that opens up when combining traditional investigative skills and [storytelling](#storytelling), with the modern availability and scale of digital information. [24](#fn24) <#disciplines> <#lesson-13>

### Data Lake

[Cloud storage](#cloud-storage) that provides flat access to structured and unstructured datasets or objects using [URL](#url)s as identifiers. De-central alternative to [data warehouse](#data-warehouse). [25](#fn25) <#lesson-11>

### Data Literacy

Data literacy is the intersection of verbal, numerical and graphical literacy. It is not a label reserved for data scientists or specialists. Data literacy should be considered “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 [26](#fn26). <#disciplines> <#lesson-01>

### 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.[27](#fn27),[28](#fn28) <#disciplines>

### Data Model

A data model captures how elements of data relate to each other to describe properties of structured real-world entities. [29](#fn29) <#storage> <#lesson-11>

### 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 (relevant and timely), consistent (coherent), complete (interpretable and accessible) and accurate. [31](#fn31) <#processing> <#lesson-04>

### Data Representation

The form in which [data](#data) is stored, processed and transmitted such that its [information content](#information) and [context](#metadata) are retained as much as possible. Choices of #[format](#file-formats) are influenced by hardware, software or other constraints around processing and analysis resources. [32](#fn32) <#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). [33](#fn33),[34](#fn34) <#disciplines>

### Data Transformation

To convert data from [one form into another](#data-representation) using [arithmetic operations](#arithmetic-operations) or more complex [algorithms](#algorithm), possibly turning [unstructured data](#unstructured-data) into [structured form](#structured-data) to support [data analysis](#data-analysis). <#processing> <#operations> <#lesson-03> <#lesson-06> <#lesson-08> <#lesson-12>

### Data types

For [data](#data) to be interpretable to a computer, it will be encoded among basic data types. <#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](#data-representation) of a date (may include time) e.g. YYYYMMDD or YYYYMMDDhhmmss.

### Data Warehouse

System for reporting and data analysis central to business intelligence that integrates data from one or more sources. <#storage> <#lesson-11>

### Data-to-ink ratio

A measure of conciseness of a [visualization](#visualization) that discourages the use of unnecessary decorations, but seeks to display all of the relevant data as clearly as possible. Related topics are [chartjunk](#chartjunk) and [lie-factor](#lie-factor). [35](#fn35) <#lesson-14>

### 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.[36](#fn36) [37](#fn37) <#storage> <#lesson-10> <#lesson-11>

### Database Model

A [data model](#data-model) that determines the logical structure or type of a [database](#database). The most popular database model is the relational model, implemented using a [table-based](#tabular-data) format. [38](#fn38) <#storage> <#lesson-11>

### Dendrogram

A [diagram](#visualization) that represents a tree, for instance, to show the nested grouping of points resulting from hierarchical [clustering](#clustering). [39](#fn39) <#lesson-09>

### 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?”[40](#fn40),[41](#fn41) <#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.[42](#fn42),[43](#fn43) <#storage>

### Entity–Relationship Model

Describes types of entities (things of interest) and specifies relationships that can exist between entities (instances of those entity types). The ER model is part of the [metadata](#metadata) for end-users of a database. [44](#fn44) <#storage> <#lesson-11>

### 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.

### File Formats

A computer stores files in various formats. The file format, usually specified in the file extension, determines how the computer should read and process the information within. For example the extension .jpeg for images, .docx for Word. [Tabular data](#tabular-data) file formats for [structured](#structured-data) data include the following. <#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 a 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 [structure](#structured-data). YAML is a compact format commonly used to share configurations, settings and unsophisticated data files.

### Filter

An [operation](#operations) that reduces a dataset by selecting all data points that satisfy some query condition or rule. See also [SQL](#sql). <#lesson-07>

### 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.[45](#fn45) <#data-types>

### Foreign key

Specific column in a [data table](#tabular-data) that can be used to identify related rows in another table. Basis for [join](#join) operations in [SQL](#sql) [queries](#query). <#lesson-12>

### 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. [46](#fn46)

### GIS

A geographic information system (GIS) manages different types of data so that it can be [integrated](#join) based on where things are, using spatial location attributes or geocodes. [47](#fn47) <#storage> <#lesson-10>

### Goodness of Fit

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

### Group by

*Group by* is a [transformation](#data-transformations) of a [table](#tabular-data) in which rows with the same value are collected *by*, and chosen columns are [aggregated](#data-aggregation) or summarized ([arithmetic](#arithmetic-operations), [string](#string) or [logical](#logical-operations)). <#operations> <#lesson-07>

### Heatmap

Data [visualization](#visualization) technique that shows the magnitude of a phenomenon using a [colour map](#colour-map) in a two-dimensional data plot. Based on the colour variation, visual cues show whether the phenomenon is clustered or varies over space. [48](#fn48) <#lesson-09>

### Histogram

[Visualization](#visualization) that groups data values into ranges or bins, using a [bar plot](#bar-chart) to indicate how value counts progress across bins, giving an idea of the distribution of values. See also [violin plot](#violin-plot). <#data-analysis> <#lesson-05> <#lesson-09> <#lesson-13> <#lesson-14>

### HTML

HyperText Markup Language is a coding language for designing documents (webpages) displayed in a web browser. <#lesson-10>

### Infographics

A type of presentation for data [storytelling](#storytelling) suitable for large display in poster form, visually grouping messages by themes, using data [vis](#visualization) and concise language. <#lesson-13>

### 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). [49](#fn49)

### Insight

Actionable [information](#information) gained by interpreting [data analysis](#analytics) results, helping people to make more informed decisions. [50](#fn50)

### 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.[51](#fn51),[52](#fn52) <#storage>

### Join

A *join* operation is used to [combine](#data-transformation) rows from two or more tables, based on a related column between them. [53](#fn53) <#query> <#operations> <#lesson-06> <#lesson-08> <#lesson-12>

### 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.[54](#fn54) <#disciplines>

### Lie factor

Lie factor measures the distortion of a graphic by comparing the size of the effect shown in the graphic and the size of the effect shown in the data. Related topics are [chartjunk](#chartjunk) and [data-to-ink-ratio](#data-to-ink-ratio). [55](#fn55) <#lesson-14>

### Line Plot

Type of [visualization](#visualization) that represents a sequence of values as points connected by line segments to show short- and long-term trends. <#lesson-09> <#lesson-13> <#lesson-14>

### Logical operations

Operations and manipulations according to the rules of logic and boolean algebra. Logical operations evaluate a logical statement’s *truth values* (true or false). Logical statements are constructed with components including conjunction (*and*), the disjunction (*or*), and the negation (*not*), in any combination with comparators *greater than* (>), *less than* (<), and *equals to* (=). <#operations> <#lesson-08>

### 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 analysis](#data-analysis). <#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. [56](#fn56) <#data-types> <#lesson-03> <#lesson-11>

### Model

A model simplifies or substitutes knowledge, information or understanding about a topic. Physical models (or replicas) like organisms, the solar system, structures are common. To a [data scientist](#data-science), models are abstractions, for example, theories, hypotheses, predictions, [algorithms](#algorithm), mathematical formulas or computer simulations. Techniques for construction models include (but are not limited to) [artificial intelligence](#artificial-intelligence), [machine learning](#machine-learning), [statistical](#statistics), [analytics](#analytics), [data mining](#data-mining) or [Big Data](#big-data) techniques. [57](#fn57) <#data-analysis> <#lesson-09>

### 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.[58](#fn58),[59](#fn59)

### 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. [60](#fn60)

### No-SQL Database

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

### Null values

Null or NULL is a special marker to indicate that a value does not exist. A null value indicates a lack of a value, which is not the same thing as a value of zero. [61](#fn61) <#lesson-06>

### Pair plot

[Plot](#visualization) pairwise relationships in a dataset. <#lesson-09>

### Pictographs Plot

Like a [scatter plot](#scatter-plot) except using small thumbnail images in the position of the data points. [visualization](#visualization) <#lesson-14>

### Pie Chart

[Visualization](#visualization) type that shows how differently proportioned parts add up to a whole using differently angled wedges or slices of a circle or pie. <#lesson-09> <#lesson-13> <#lesson-14>

### Pivot

Pivoting is an extension of the [transpose](#transpose) operation. The pivot operation converts between [wide-](#wide-table-format) and [narrow-](#narrow-table-format) table formats. <#operations> <#lesson-03> <#lesson-08>

### Population

A collection of items under consideration. [62](#fn62) <#data-analysis> <#lesson-04>

### 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), modelling, [machine learning](#machine-learning) and statistics.[63](#fn63),[64](#fn64)

### Prescriptive Analytics

A type of [analytics](#analytics) with the goal of using data to determine the best course of action for a specific scenario. [65](#fn65) <#data-analysis>

### Primary key

Specific attributes or table columns in a [relational](#database-model) data [table](#tabular-data) that uniquely identify a data point or row. Most commonly, a single attribute: a unique ID. [66](#fn66) <#query> <#lesson-12>

### Query (Data)

Request for data from one or more [tables](#tabular). [SQL](#sql) is commonly used to express a query, but also plain text search is a type of query. <#lesson-07> <#lesson-10> <#lesson-11>

### 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>

### Repository (data repo)

Sometimes abbreviated as *repo*, it is often used as a synonym for a [database](#database). Generally, it means any place in which things may be stored. <#storage> <#lesson-10>

### 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). [67](#fn67) <#data-analysis> <#lesson-04>

### Scatter Plot

A data point in a scatter plot [represents](#visualization) values of two variables with their horizontal and vertical position in a rectangular frame of reference. The scattering of multiple data points can show trends, groupings, linear, and non-linear relationships. See also [data analysis](#data-analysis). <#lesson-09> <#lesson-13> <#lesson-14>

### Schema (Database)

Description of tabular data structure giving name and data type of the variable associated with each table column. See also [database](#database). <#storage> <#lesson-11>

### 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. [68](#fn68),[69](#fn69) <#disciplines>

### Select

A type of [query](#query) that requests certain columns, either by selecting existing columns or by performing [operation](#operations) to create new columns. See also [SQL](#sql). <#lesson-07>

### Skewness

Skewness refers to a distortion or asymmetry that deviates from the symmetrical bell curve, or normal distribution, in a set of data. If the curve is shifted to the left or the right, it is said to be skewed. <#lesson-09>

### Sort

A type of data [transformation](#data-transformations) that orders rows by some chosen column. While simple, sorting is a powerful [operation](#operations) that supports many tasks, including ranking, extreme values, or duplicate detection. <#lesson-07>

### SQL

SQL (pronounced “ess-que-el” or “sequel”) stands for Structured [Query](#query) Language. 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. [70](#fn70) [71](#fn71) <#processing> <#lesson-07> <#lesson-10> <#lesson-11> <#lesson-12>

### Standard Deviation

Standard deviation is the square root of [variance](#variance). It is a measure of dispersion, meaning of how far a set of numbers is spread out from their average value. <#lesson-05> <#lesson-09>

### Statistics

Statistics is the discipline concerned with developing and studying methods for collecting, organizing, analyzing, interpreting, and presenting empirical data. Common statistical metrics are mean (average), median, mode and percentiles. <#data-analysis> <#lesson-05> <#arithmetic-operations>

### Storytelling

The ability to tell or write stories and narratives to communicate insights and actions drawn data, optionally aided by effective [visualizations](#visualization). [72](#fn72) <#disciplines> <#lesson-13>

### String or character operations

Manipulations with [string data types](#data-types) include the following. <#operations> <#lesson-08>

#### Parse

To analyze (a string or text) into logical syntactic components.[73](#fn73)

#### Indexing

Indexing (or selecting) string characters by their position, *zeroth* character, the *first* character, etc.

#### Concatenation

Combining items (strings) end-to-end.

### 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.[74](#fn74),[75](#fn75) <#data-types> <#lesson-03>

### Tabular Data

Data items or observations organized as rows, containing entries, under columns, that correspond to separate kinds of values or attributes. The intersection of a row and a column is a cell, which corresponds to individual data values. [76](#fn76),[77](#fn77) <#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(s)](#feature), property(ies), variable(s), key(s) (as in key-value), factor(s), dimension(s)
* Synonyms: pairs, key-value, attribute-value

#### Narrow table format

Narrow, stacked or long, is presented with one column containing all data values and another column containing the kind of values. Related topic [pivot](#pivot). <#data-types> <#lesson-03>

#### Wide table format

Wide, or unstacked tabular data presents each [variable/attribute/feature](#feature) in a separate column. Related topic, [pivot](#pivot). <#data-types> <#lesson-03>

### Text Mining

A process of analyzing [text](#unstructured-data) to capture key concepts, themes, relationships and trends.[78](#fn78) <#data-analysis>

### Tidy tables

Datasets are commonly messy. Getting data into tidy [79](#fn79) format requires upfront work, but that work pays off in the long term. There are three interrelated rules which make a dataset tidy:

1. Each variable must have its own column.
2. Each observation must have its own row.
3. Each value must have its own cell.

<#lesson-06>

### Time Series

A sequence of values associated with successive times. <#lesson-08> <#lesson-13>

### Transpose

A mathematical operation on a table where a new table is obtained by interchanging each row and the corresponding column. <#operations> <#lesson-08>

### 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. [80](#fn80) <#data-types> <#data-types> <#lesson-03>

### Unbalanced Data

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

### Union

A *union* operation that [combines](#data-transformation) tables with the same number and [types](#data-types) of columns. This operation is like vertical [concatenation](#concatenation). <#query> <#operations> <#lesson-12>

### Unstructured Data

Data that is not organized in a pre-defined way under a single data model. Examples include text, images, audio, or video. [81](#fn81),[82](#fn82) <#data-types>, <#lesson-03>

### URL

The URL (short for Uniform Resource Locator) is the location of a webpage or file (including pictures, videos, or CSV) on the Internet. Files (or resources) stored on the cloud have unique URLs. <#lesson-10>

### Variance

Measure of dispersion, indicating how far the values of a variable are spread out around its [sample](#sample) mean. Low variance indicates nearly constant values throughout the [population](#population). Also see [standard deviation](#standard-deviation). [83](#fn83) <#operations> <#lesson-05>

### Violin Plot

Similar to a [histogram](#histogram), this plot type shows how values of a variable are distributed in a dataset, indicating more often occurring values with thickness that extends symmetrically around a central axis. The name refers to the typical look that most datasets have in this [visualization](#visualization) method. [box plots](#box-plot) <#lesson-09>

### 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. [84](#fn84), [85](#fn85), [86](#fn86) <#data-analysis> <#lesson-05> <#lesson-09> <#lesson-13>

### 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>

### Web scraping

The practice of extracting data from webpages by automatically requesting and processing the HTML source code of these pages. Alternatively, obtaining data via a structured file format or direct database or API access is usually preferred, whenever possible. Web search engines, such as Google or Bing, are based on getting their data via web scraping. <#lesson-10>

### Word Cloud

Text summary method that displays important words from a document indicating varying degrees of relevance by varying the font size of the words. The most important words stand out at a glance, giving a rough idea of the topic of a text. [visualization](#visualization) <#lesson-09>

# 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)

[Cloud (“The Cloud”)](#cloud-the-cloud)

[Cloud Computing](#cloud-computing)

[Computing with Data](#computing-with-data)

[Critical Thinking](#critical-thinking)

[Cybersecurity](#cybersecurity)

[Data Journalism](#data-journalism)

[Data Literacy](#data-literacy)

[Data Mining](#data-mining)

[Data Privacy](#data-privacy)

[Data Science](#data-science)

[Knowledge Discovery](#knowledge-discovery)

[Machine Learning](#machine-learning)

[Secure Computing](#secure-computing)

[Storytelling](#storytelling)

## Data Analysis Techniques

[Behaviour(al) Analytics](#behavioural-analytics)

[Causal Inference](#causal-inference)

[Classification](#classification)

[Clustering](#clustering)

[Correlation](#correlation)

[Correlation Mining](#correlation-mining)

[Data Bias](#data-bias)

[Data Transformation](#data-transformation)

[Descriptive Analytics](#descriptive-analytics)

[Histogram](#histogram)

[Machine Learning](#machine-learning)

[Model](#model)

[Population](#population)

[Prescriptive Analytics](#prescriptive-analytics)

[Regression](#regression)

[Relationship Mining](#relationship-mining)

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

[Scatter Plot](#scatter-plot)

[Statistics](#statistics)

[Text Mining](#text-mining)

[Visualization](#visualization)

[Visualization Dashboard](#visualization-dashboard)

## Data Processing

[Algorithm](#algorithm)

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

[Data Cleaning](#data-cleaning)

[Data Exploration and Preparation](#data-exploration-and-preparation)

[Data Integration](#data-integration)

[Data Quality](#data-quality)

[Data Representation](#data-representation)

[Data Transformation](#data-transformation)

[Goodness of Fit](#goodness-of-fit)

[SQL](#sql)

## Storage and Physical Data Sources

[Cloud Storage](#cloud-storage)

[Data Model](#data-model)

[Data Warehouse](#data-warehouse)

[Database](#database)

[Database Model](#database-model)

[Distributed File System](#distributed-file-system)

[Entity–Relationship Model](#entityrelationship-model)

[GIS](#gis)

[IOT (Internet of Things)](#iot-internet-of-things)

[No-SQL Database](#no-sql-database)

[Repository (data repo)](#repository-data-repo)

[Schema (Database)](#schema-database)

## Types of Data

[Data](#data)

[Data types](#data-types)

[File Formats](#file-formats)

[Flexible Schema](#flexible-schema)

[Metadata](#metadata)

[String or character operations](#string-or-character-operations)

[Structured Data](#structured-data)

[Tabular Data](#tabular-data)

[Tree structure](#tree-structure)

[Union](#union)

[Unstructured Data](#unstructured-data)

## Operations

[Arithmetic operations](#arithmetic-operations)

[Binning](#binning)

[Bucketing](#bucketing)

[Data Aggregation](#data-aggregation)

[Data Transformation](#data-transformation)

[Filter](#filter)

[Group by](#group-by)

[Join](#join)

[Logical operations](#logical-operations)

[Pivot](#pivot)

[Select](#select)

[Sort](#sort)

[String or character operations](#string-or-character-operations)

[Transpose](#transpose)

[Union](#union)

[Variance](#variance)

# Lessons Index

## Lessons 01 and 02

### Data Literacy

Data literacy is the intersection of verbal, numerical and graphical literacy. It is not a label reserved for data scientists or specialists. Data literacy should be considered “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 [26](#fn26). <#disciplines> <#lesson-01>

## Lesson 03

### 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>

### Crowdsourcing

A collaborative approach for collection and maintenance of data where people use websites to share information of common interest. The result can be large, high-quality datasets. <#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 Transformation

To convert data from [one form into another](#data-representation) using [arithmetic operations](#arithmetic-operations) or more complex [algorithms](#algorithm), possibly turning [unstructured data](#unstructured-data) into [structured form](#structured-data) to support [data analysis](#data-analysis). <#processing> <#operations> <#lesson-03> <#lesson-06> <#lesson-08> <#lesson-12>

### Data types

For [data](#data) to be interpretable to a computer, it will be encoded among basic data types. <#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](#data-representation) of a date (may include time) e.g. YYYYMMDD or YYYYMMDDhhmmss.

### File Formats

A computer stores files in various formats. The file format, usually specified in the file extension, determines how the computer should read and process the information within. For example the extension .jpeg for images, .docx for Word. [Tabular data](#tabular-data) file formats for [structured](#structured-data) data include the following. <#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 a 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 [structure](#structured-data). YAML is a compact format commonly used to share configurations, settings and unsophisticated data files.

### 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 analysis](#data-analysis). <#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. [56](#fn56) <#data-types> <#lesson-03> <#lesson-11>

### Pivot

Pivoting is an extension of the [transpose](#transpose) operation. The pivot operation converts between [wide-](#wide-table-format) and [narrow-](#narrow-table-format) table formats. <#operations> <#lesson-03> <#lesson-08>

### 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.[74](#fn74),[75](#fn75) <#data-types> <#lesson-03>

### Tabular Data

Data items or observations organized as rows, containing entries, under columns, that correspond to separate kinds of values or attributes. The intersection of a row and a column is a cell, which corresponds to individual data values. [76](#fn76),[77](#fn77) <#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(s)](#feature), property(ies), variable(s), key(s) (as in key-value), factor(s), dimension(s)
* Synonyms: pairs, key-value, attribute-value

#### Narrow table format

Narrow, stacked or long, is presented with one column containing all data values and another column containing the kind of values. Related topic [pivot](#pivot). <#data-types> <#lesson-03>

#### Wide table format

Wide, or unstacked tabular data presents each [variable/attribute/feature](#feature) in a separate column. Related topic, [pivot](#pivot). <#data-types> <#lesson-03>

### 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. [80](#fn80) <#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. [81](#fn81),[82](#fn82) <#data-types>, <#lesson-03>

## Lesson 04

### 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> <#lesson-04>

### Critical Thinking

A thought process to determine whether we believe something to be true or false by logically analyzing and evaluating evidence and other information. [19](#fn19) <#disciplines> <#lesson-04>

### 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, misplaced, inconsistent, or duplicated. [21](#fn21) <#processing> <#lesson-04> <#lesson-06>

### Data Bias

Whether the (sample) data’s [balance](#balance) is representative [sample](#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 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 (relevant and timely), consistent (coherent), complete (interpretable and accessible) and accurate. [31](#fn31) <#processing> <#lesson-04>

### Population

A collection of items under consideration. [62](#fn62) <#data-analysis> <#lesson-04>

### 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). [67](#fn67) <#data-analysis> <#lesson-04>

### Unbalanced Data

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

## Lesson 05

### Correlation

Correlation is a relationship between two variables in a dataset, indicating that their values tend to change proportionally to each other. While it is a measure of association, it does not imply [causation](#causal-inference). <#data-analysis> <#lesson-05> <#lesson-09>

### Histogram

[Visualization](#visualization) that groups data values into ranges or bins, using a [bar plot](#bar-chart) to indicate how value counts progress across bins, giving an idea of the distribution of values. See also [violin plot](#violin-plot). <#data-analysis> <#lesson-05> <#lesson-09> <#lesson-13> <#lesson-14>

### Standard Deviation

Standard deviation is the square root of [variance](#variance). It is a measure of dispersion, meaning of how far a set of numbers is spread out from their average value. <#lesson-05> <#lesson-09>

### Statistics

Statistics is the discipline concerned with developing and studying methods for collecting, organizing, analyzing, interpreting, and presenting empirical data. Common statistical metrics are mean (average), median, mode and percentiles. <#data-analysis> <#lesson-05> <#arithmetic-operations>

### Variance

Measure of dispersion, indicating how far the values of a variable are spread out around its [sample](#sample) mean. Low variance indicates nearly constant values throughout the [population](#population). Also see [standard deviation](#standard-deviation). [83](#fn83) <#operations> <#lesson-05>

### 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. [84](#fn84), [85](#fn85), [86](#fn86) <#data-analysis> <#lesson-05> <#lesson-09> <#lesson-13>

## Lesson 06

### 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, misplaced, inconsistent, or duplicated. [21](#fn21) <#processing> <#lesson-04> <#lesson-06>

### Data Integration

The process of combining information from different data sources in preparation for data processing. [23](#fn23) <#processing> <#lesson-06> <#lesson-10> <#lesson-11> <#lesson-12>

### Data Transformation

To convert data from [one form into another](#data-representation) using [arithmetic operations](#arithmetic-operations) or more complex [algorithms](#algorithm), possibly turning [unstructured data](#unstructured-data) into [structured form](#structured-data) to support [data analysis](#data-analysis). <#processing> <#operations> <#lesson-03> <#lesson-06> <#lesson-08> <#lesson-12>

### Join

A *join* operation is used to [combine](#data-transformation) rows from two or more tables, based on a related column between them. [53](#fn53) <#query> <#operations> <#lesson-06> <#lesson-08> <#lesson-12>

### Null values

Null or NULL is a special marker to indicate that a value does not exist. A null value indicates a lack of a value, which is not the same thing as a value of zero. [61](#fn61) <#lesson-06>

### Tidy tables

Datasets are commonly messy. Getting data into tidy [79](#fn79) format requires upfront work, but that work pays off in the long term. There are three interrelated rules which make a dataset tidy:

1. Each variable must have its own column.
2. Each observation must have its own row.
3. Each value must have its own cell.

<#lesson-06>

## Lesson 07

### Filter

An [operation](#operations) that reduces a dataset by selecting all data points that satisfy some query condition or rule. See also [SQL](#sql). <#lesson-07>

### Group by

*Group by* is a [transformation](#data-transformations) of a [table](#tabular-data) in which rows with the same value are collected *by*, and chosen columns are [aggregated](#data-aggregation) or summarized ([arithmetic](#arithmetic-operations), [string](#string) or [logical](#logical-operations)). <#operations> <#lesson-07>

### Query (Data)

Request for data from one or more [tables](#tabular). [SQL](#sql) is commonly used to express a query, but also plain text search is a type of query. <#lesson-07> <#lesson-10> <#lesson-11>

### Select

A type of [query](#query) that requests certain columns, either by selecting existing columns or by performing [operation](#operations) to create new columns. See also [SQL](#sql). <#lesson-07>

### Sort

A type of data [transformation](#data-transformations) that orders rows by some chosen column. While simple, sorting is a powerful [operation](#operations) that supports many tasks, including ranking, extreme values, or duplicate detection. <#lesson-07>

### SQL

SQL (pronounced “ess-que-el” or “sequel”) stands for Structured [Query](#query) Language. 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. [70](#fn70) [71](#fn71) <#processing> <#lesson-07> <#lesson-10> <#lesson-11> <#lesson-12>

## Lesson 08

### Arithmetic operations

Includes basic statistics: such as *sums*, *counts*, *means* (or *averages*), *medians*, *percentiles*, or [*standard deviations*](#standard-deviation). <#operations> <#lesson-08>

### Binning

An operation on a range of numbers in which the entire range is divided into *bins* (or *intervals*) of a *bin width* (or *bin size*, *step size*). Another name for binning is *bucketing* <#operations> <#lesson-08>

### Bucketing

Another name for bucketing is [binning](#binning). <#operations> <#lesson-08>

### Data Aggregation

A [group-by](#group-by) operation that reduces data. Values of a variable (or column) that are observed for a group of points (or rows of a [table](#tabular-data)) are combined into a single value. [22](#fn22) <#operations>, <#lesson-08>

### Data Transformation

To convert data from [one form into another](#data-representation) using [arithmetic operations](#arithmetic-operations) or more complex [algorithms](#algorithm), possibly turning [unstructured data](#unstructured-data) into [structured form](#structured-data) to support [data analysis](#data-analysis). <#processing> <#operations> <#lesson-03> <#lesson-06> <#lesson-08> <#lesson-12>

### Join

A *join* operation is used to [combine](#data-transformation) rows from two or more tables, based on a related column between them. [53](#fn53) <#query> <#operations> <#lesson-06> <#lesson-08> <#lesson-12>

### Logical operations

Operations and manipulations according to the rules of logic and boolean algebra. Logical operations evaluate a logical statement’s *truth values* (true or false). Logical statements are constructed with components including conjunction (*and*), the disjunction (*or*), and the negation (*not*), in any combination with comparators *greater than* (>), *less than* (<), and *equals to* (=). <#operations> <#lesson-08>

### Pivot

Pivoting is an extension of the [transpose](#transpose) operation. The pivot operation converts between [wide-](#wide-table-format) and [narrow-](#narrow-table-format) table formats. <#operations> <#lesson-03> <#lesson-08>

### String or character operations

Manipulations with [string data types](#data-types) include the following. <#operations> <#lesson-08>

#### Parse

To analyze (a string or text) into logical syntactic components.[73](#fn73)

#### Indexing

Indexing (or selecting) string characters by their position, *zeroth* character, the *first* character, etc.

#### Concatenation

Combining items (strings) end-to-end.

### Time Series

A sequence of values associated with successive times. <#lesson-08> <#lesson-13>

### Transpose

A mathematical operation on a table where a new table is obtained by interchanging each row and the corresponding column. <#operations> <#lesson-08>

## Lesson 09

### Bar Chart

A type of data [visualization](#visualization) that uses an array of bars (or rectangles) that share a common baseline, but vary in height to represent a set of data values. A horizontal version is also common, giving room for longer text labels. Multiple bars for a single point can be grouped or stacked. <#lesson-09> <#lesson-13> <#lesson-14>

### Box Plot

A type of data [visualization](#visualization) that shows how values of a variable are distributed. Line markers along an axis show how far the lower and upper quartiles are from the [median](#arithmetic-operations) giving an idea of central tendency, dispersion, and skew. See also [histogram](#histogram) and [violin plot](#violin-plot). <#lesson-09>

### Correlation

Correlation is a relationship between two variables in a dataset, indicating that their values tend to change proportionally to each other. While it is a measure of association, it does not imply [causation](#causal-inference). <#data-analysis> <#lesson-05> <#lesson-09>

### Data Analysis

Data analysis is a process of inspecting, [cleansing](#clean-dirty), [transforming](#data-transformation), and modelling data to discover useful information, informing conclusions, and supporting [decision-making](#critical-thinking). <#analytics> <#lesson-09>

### Dendrogram

A [diagram](#visualization) that represents a tree, for instance, to show the nested grouping of points resulting from hierarchical [clustering](#clustering). [39](#fn39) <#lesson-09>

### Heatmap

Data [visualization](#visualization) technique that shows the magnitude of a phenomenon using a [colour map](#colour-map) in a two-dimensional data plot. Based on the colour variation, visual cues show whether the phenomenon is clustered or varies over space. [48](#fn48) <#lesson-09>

### Histogram

[Visualization](#visualization) that groups data values into ranges or bins, using a [bar plot](#bar-chart) to indicate how value counts progress across bins, giving an idea of the distribution of values. See also [violin plot](#violin-plot). <#data-analysis> <#lesson-05> <#lesson-09> <#lesson-13> <#lesson-14>

### Line Plot

Type of [visualization](#visualization) that represents a sequence of values as points connected by line segments to show short- and long-term trends. <#lesson-09> <#lesson-13> <#lesson-14>

### Model

A model simplifies or substitutes knowledge, information or understanding about a topic. Physical models (or replicas) like organisms, the solar system, structures are common. To a [data scientist](#data-science), models are abstractions, for example, theories, hypotheses, predictions, [algorithms](#algorithm), mathematical formulas or computer simulations. Techniques for construction models include (but are not limited to) [artificial intelligence](#artificial-intelligence), [machine learning](#machine-learning), [statistical](#statistics), [analytics](#analytics), [data mining](#data-mining) or [Big Data](#big-data) techniques. [57](#fn57) <#data-analysis> <#lesson-09>

### Pair plot

[Plot](#visualization) pairwise relationships in a dataset. <#lesson-09>

### Pie Chart

[Visualization](#visualization) type that shows how differently proportioned parts add up to a whole using differently angled wedges or slices of a circle or pie. <#lesson-09> <#lesson-13> <#lesson-14>

### Scatter Plot

A data point in a scatter plot [represents](#visualization) values of two variables with their horizontal and vertical position in a rectangular frame of reference. The scattering of multiple data points can show trends, groupings, linear, and non-linear relationships. See also [data analysis](#data-analysis). <#lesson-09> <#lesson-13> <#lesson-14>

### Skewness

Skewness refers to a distortion or asymmetry that deviates from the symmetrical bell curve, or normal distribution, in a set of data. If the curve is shifted to the left or the right, it is said to be skewed. <#lesson-09>

### Standard Deviation

Standard deviation is the square root of [variance](#variance). It is a measure of dispersion, meaning of how far a set of numbers is spread out from their average value. <#lesson-05> <#lesson-09>

### Violin Plot

Similar to a [histogram](#histogram), this plot type shows how values of a variable are distributed in a dataset, indicating more often occurring values with thickness that extends symmetrically around a central axis. The name refers to the typical look that most datasets have in this [visualization](#visualization) method. [box plots](#box-plot) <#lesson-09>

### 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. [84](#fn84), [85](#fn85), [86](#fn86) <#data-analysis> <#lesson-05> <#lesson-09> <#lesson-13>

### Word Cloud

Text summary method that displays important words from a document indicating varying degrees of relevance by varying the font size of the words. The most important words stand out at a glance, giving a rough idea of the topic of a text. [visualization](#visualization) <#lesson-09>

## Lesson 10

### API

API is an acronym for Application Programming Interface. Through the API, a Program (like a website or code) accesses an Application (a database or another application) through a common language, protocol (or Interface) through which the two systems interact. <#lesson-10>

### Cloud (“The Cloud”)

Any resource stored on the internet (instead of locally on a PC) is said to be stored on the cloud. Cloud storage is convenient in that clients do not have to manage the storage and infrastructure; they just access it from other internet-connected devices. <#disciplines> <#lesson-10>

### Cluster (Compute)

A [network](#network) of computers (or compute nodes) that work on tasks together in parallel. A cluster is one way to address [Big Data](#big-data) needs. The concept is different from [clustering](#clustering). <#lesson-10>

### Data Integration

The process of combining information from different data sources in preparation for data processing. [23](#fn23) <#processing> <#lesson-06> <#lesson-10> <#lesson-11> <#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.[36](#fn36) [37](#fn37) <#storage> <#lesson-10> <#lesson-11>

### GIS

A geographic information system (GIS) manages different types of data so that it can be [integrated](#join) based on where things are, using spatial location attributes or geocodes. [47](#fn47) <#storage> <#lesson-10>

### HTML

HyperText Markup Language is a coding language for designing documents (webpages) displayed in a web browser. <#lesson-10>

### Query (Data)

Request for data from one or more [tables](#tabular). [SQL](#sql) is commonly used to express a query, but also plain text search is a type of query. <#lesson-07> <#lesson-10> <#lesson-11>

### Repository (data repo)

Sometimes abbreviated as *repo*, it is often used as a synonym for a [database](#database). Generally, it means any place in which things may be stored. <#storage> <#lesson-10>

### SQL

SQL (pronounced “ess-que-el” or “sequel”) stands for Structured [Query](#query) Language. 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. [70](#fn70) [71](#fn71) <#processing> <#lesson-07> <#lesson-10> <#lesson-11> <#lesson-12>

### URL

The URL (short for Uniform Resource Locator) is the location of a webpage or file (including pictures, videos, or CSV) on the Internet. Files (or resources) stored on the cloud have unique URLs. <#lesson-10>

### Web scraping

The practice of extracting data from webpages by automatically requesting and processing the HTML source code of these pages. Alternatively, obtaining data via a structured file format or direct database or API access is usually preferred, whenever possible. Web search engines, such as Google or Bing, are based on getting their data via web scraping. <#lesson-10>

## Lesson 11

### Cloud Storage

Stores data on the Internet with a [cloud computing](#cloud-computing) provider who manages and operates data storage as a service. [15](#fn15) <#storage> <#lesson-11>

### Data Integration

The process of combining information from different data sources in preparation for data processing. [23](#fn23) <#processing> <#lesson-06> <#lesson-10> <#lesson-11> <#lesson-12>

### Data Lake

[Cloud storage](#cloud-storage) that provides flat access to structured and unstructured datasets or objects using [URL](#url)s as identifiers. De-central alternative to [data warehouse](#data-warehouse). [25](#fn25) <#lesson-11>

### Data Model

A data model captures how elements of data relate to each other to describe properties of structured real-world entities. [29](#fn29) <#storage> <#lesson-11>

### Data Warehouse

System for reporting and data analysis central to business intelligence that integrates data from one or more sources. <#storage> <#lesson-11>

### 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.[36](#fn36) [37](#fn37) <#storage> <#lesson-10> <#lesson-11>

### Database Model

A [data model](#data-model) that determines the logical structure or type of a [database](#database). The most popular database model is the relational model, implemented using a [table-based](#tabular-data) format. [38](#fn38) <#storage> <#lesson-11>

### Entity–Relationship Model

Describes types of entities (things of interest) and specifies relationships that can exist between entities (instances of those entity types). The ER model is part of the [metadata](#metadata) for end-users of a database. [44](#fn44) <#storage> <#lesson-11>

### 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. [56](#fn56) <#data-types> <#lesson-03> <#lesson-11>

### Query (Data)

Request for data from one or more [tables](#tabular). [SQL](#sql) is commonly used to express a query, but also plain text search is a type of query. <#lesson-07> <#lesson-10> <#lesson-11>

### Schema (Database)

Description of tabular data structure giving name and data type of the variable associated with each table column. See also [database](#database). <#storage> <#lesson-11>

### SQL

SQL (pronounced “ess-que-el” or “sequel”) stands for Structured [Query](#query) Language. 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. [70](#fn70) [71](#fn71) <#processing> <#lesson-07> <#lesson-10> <#lesson-11> <#lesson-12>

## Lesson 12

### Data Integration

The process of combining information from different data sources in preparation for data processing. [23](#fn23) <#processing> <#lesson-06> <#lesson-10> <#lesson-11> <#lesson-12>

### Data Transformation

To convert data from [one form into another](#data-representation) using [arithmetic operations](#arithmetic-operations) or more complex [algorithms](#algorithm), possibly turning [unstructured data](#unstructured-data) into [structured form](#structured-data) to support [data analysis](#data-analysis). <#processing> <#operations> <#lesson-03> <#lesson-06> <#lesson-08> <#lesson-12>

### Foreign key

Specific column in a [data table](#tabular-data) that can be used to identify related rows in another table. Basis for [join](#join) operations in [SQL](#sql) [queries](#query). <#lesson-12>

### Join

A *join* operation is used to [combine](#data-transformation) rows from two or more tables, based on a related column between them. [53](#fn53) <#query> <#operations> <#lesson-06> <#lesson-08> <#lesson-12>

### Primary key

Specific attributes or table columns in a [relational](#database-model) data [table](#tabular-data) that uniquely identify a data point or row. Most commonly, a single attribute: a unique ID. [66](#fn66) <#query> <#lesson-12>

### SQL

SQL (pronounced “ess-que-el” or “sequel”) stands for Structured [Query](#query) Language. 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. [70](#fn70) [71](#fn71) <#processing> <#lesson-07> <#lesson-10> <#lesson-11> <#lesson-12>

### Union

A *union* operation that [combines](#data-transformation) tables with the same number and [types](#data-types) of columns. This operation is like vertical [concatenation](#concatenation). <#query> <#operations> <#lesson-12>

## Lesson 13

### Bar Chart

A type of data [visualization](#visualization) that uses an array of bars (or rectangles) that share a common baseline, but vary in height to represent a set of data values. A horizontal version is also common, giving room for longer text labels. Multiple bars for a single point can be grouped or stacked. <#lesson-09> <#lesson-13> <#lesson-14>

### Data Journalism

Refers to the field that opens up when combining traditional investigative skills and [storytelling](#storytelling), with the modern availability and scale of digital information. [24](#fn24) <#disciplines> <#lesson-13>

### Histogram

[Visualization](#visualization) that groups data values into ranges or bins, using a [bar plot](#bar-chart) to indicate how value counts progress across bins, giving an idea of the distribution of values. See also [violin plot](#violin-plot). <#data-analysis> <#lesson-05> <#lesson-09> <#lesson-13> <#lesson-14>

### Infographics

A type of presentation for data [storytelling](#storytelling) suitable for large display in poster form, visually grouping messages by themes, using data [vis](#visualization) and concise language. <#lesson-13>

### Line Plot

Type of [visualization](#visualization) that represents a sequence of values as points connected by line segments to show short- and long-term trends. <#lesson-09> <#lesson-13> <#lesson-14>

### Pie Chart

[Visualization](#visualization) type that shows how differently proportioned parts add up to a whole using differently angled wedges or slices of a circle or pie. <#lesson-09> <#lesson-13> <#lesson-14>

### Scatter Plot

A data point in a scatter plot [represents](#visualization) values of two variables with their horizontal and vertical position in a rectangular frame of reference. The scattering of multiple data points can show trends, groupings, linear, and non-linear relationships. See also [data analysis](#data-analysis). <#lesson-09> <#lesson-13> <#lesson-14>

### Storytelling

The ability to tell or write stories and narratives to communicate insights and actions drawn data, optionally aided by effective [visualizations](#visualization). [72](#fn72) <#disciplines> <#lesson-13>

### Time Series

A sequence of values associated with successive times. <#lesson-08> <#lesson-13>

### 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. [84](#fn84), [85](#fn85), [86](#fn86) <#data-analysis> <#lesson-05> <#lesson-09> <#lesson-13>

## Lesson 14

### Bar Chart

A type of data [visualization](#visualization) that uses an array of bars (or rectangles) that share a common baseline, but vary in height to represent a set of data values. A horizontal version is also common, giving room for longer text labels. Multiple bars for a single point can be grouped or stacked. <#lesson-09> <#lesson-13> <#lesson-14>

### Chartjunk

Elements of a graphic that distract or are not needed to communicate information ([data](#data)). Elements, such as, unnecessary text, grid lines and frames, ornamental icons, pictures or backgrounds, gradient [colours](#colour) or shading, gratuitous 3D effects and gimmick fonts. See also [data-to-ink-ratio](#data-to-ink-ratio) and [lie-factor](#lie-factor). [12](#fn12) <#lesson-14>

### Colour Map

[Visual encoding](#visualization) of different levels of a variable by using a fixed scale (or map) of colours. Synonyms: Colour palette, Colour scheme. Colour (Hue), Colou (Shade) [16](#fn16) <#lesson-14>

### Data-to-ink ratio

A measure of conciseness of a [visualization](#visualization) that discourages the use of unnecessary decorations, but seeks to display all of the relevant data as clearly as possible. Related topics are [chartjunk](#chartjunk) and [lie-factor](#lie-factor). [35](#fn35) <#lesson-14>

### Histogram

[Visualization](#visualization) that groups data values into ranges or bins, using a [bar plot](#bar-chart) to indicate how value counts progress across bins, giving an idea of the distribution of values. See also [violin plot](#violin-plot). <#data-analysis> <#lesson-05> <#lesson-09> <#lesson-13> <#lesson-14>

### Lie factor

Lie factor measures the distortion of a graphic by comparing the size of the effect shown in the graphic and the size of the effect shown in the data. Related topics are [chartjunk](#chartjunk) and [data-to-ink-ratio](#data-to-ink-ratio). [55](#fn55) <#lesson-14>

### Line Plot

Type of [visualization](#visualization) that represents a sequence of values as points connected by line segments to show short- and long-term trends. <#lesson-09> <#lesson-13> <#lesson-14>

### Pictographs Plot

Like a [scatter plot](#scatter-plot) except using small thumbnail images in the position of the data points. [visualization](#visualization) <#lesson-14>

### Pie Chart

[Visualization](#visualization) type that shows how differently proportioned parts add up to a whole using differently angled wedges or slices of a circle or pie. <#lesson-09> <#lesson-13> <#lesson-14>

### Scatter Plot

A data point in a scatter plot [represents](#visualization) values of two variables with their horizontal and vertical position in a rectangular frame of reference. The scattering of multiple data points can show trends, groupings, linear, and non-linear relationships. See also [data analysis](#data-analysis). <#lesson-09> <#lesson-13> <#lesson-14>

## Lesson 15

# References

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. “What is GIS?”. [Online]. Available: https://www.esri.com/en-us/what-is-gis/overview[↩︎](#fnref12)
13. “Who Coined ‘Cloud Computing’?” [Online]. Available: <https://www.technologyreview.com/s/425970/who-coined-cloud-computing>[↩︎](#fnref13)
14. “Cloud computing: a simple introduction.” [Online]. Available: <http://www.explainthatstuff.com/cloud-computing-introduction.html>[↩︎](#fnref14)
15. “What is Cloud Storage? | AWS” . [Online]. Available: https://aws.amazon.com/what-is-cloud-storage/[↩︎](#fnref15)
16. “Choosing Colormaps in Matplotlib”. [Online]. Available: https://matplotlib.org/stable/tutorials/colors/colormaps.html[↩︎](#fnref16)
17. Donoho, D. (2017). 50 years of data science. *Journal of Computational and Graphical Statistics*, *26*(4), 745-766.[↩︎](#fnref17)
18. “Introduction to Correlation” [Online]. Available: https://www.datascience.com/blog/introduction-to-correlation-learn-data-science-tutorials[↩︎](#fnref18)
19. “Critical Thinking”. Oxford University Press. [Online]. Available: https://www.lexico.com/definition/critical\_thinking[↩︎](#fnref19)
20. “EI-ISAC Cybersecurity Spotlight – CIA Triad” [Online]. Available: https://www.cisecurity.org/spotlight/ei-isac-cybersecurity-spotlight-cia-triad/[↩︎](#fnref20)
21. “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[↩︎](#fnref21)
22. “Aggregation”. Oxford University Press. [Online]. Available: https://www.lexico.com/definition/aggregation[↩︎](#fnref22)
23. “Mini-glossary: Big data terms you should know.” [Online]. Available: <http://www.techrepublic.com/article/mini-glossary-big-data-terms-you-should-know>[↩︎](#fnref23)
24. Bradshaw, P. (2012). What is Data Journalism? In *The Data Journalism Handbook*. European Journalism Centre. Available: https://datajournalism.com/read/handbook/one[↩︎](#fnref24)
25. “What Is a Data Lake?” Definition from SearchDataManagement. [Online]. Available: https://searchdatamanagement.techtarget.com/definition/data-lake[↩︎](#fnref25)
26. “Data Literacy: What It Is and How to Measure It in the Public Service”. Statistics Canada. [Online]. Available: https://www150.statcan.gc.ca/n1/pub/11-633-x/11-633-x2019003-eng.htm[↩︎](#fnref26)
27. “A beginner’s guide to big data terminology.” [Online]. Available: <http://dataconomy.com/2016/05/a-beginners-guide-to-big-data-terminology>[↩︎](#fnref27)
28. “Mini-glossary: Big data terms you should know.” [Online]. Available: <http://www.techrepublic.com/article/mini-glossary-big-data-terms-you-should-know>[↩︎](#fnref28)
29. [Online]. Available: https://en.wikipedia.org/wiki/Data\_model[↩︎](#fnref29)
30. “Internet Privacy” [Online]. Available: https://www.techopedia.com/definition/24954/internet-privacy[↩︎](#fnref30)
31. “Data, statistical information and statistics: Data quality” [Online]. Available: https://www150.statcan.gc.ca/n1/edu/power-pouvoir/ch3/quality-qualite/5214788-eng.htm.[↩︎](#fnref31)
32. Donoho, D. (2017). 50 years of data science. *Journal of Computational and Graphical Statistics*, *26*(4), 745-766.[↩︎](#fnref32)
33. Donoho, D. (2017). 50 years of data science. *Journal of Computational and Graphical Statistics*, *26*(4), 745-766.[↩︎](#fnref33)
34. “The Data Science Venn Diagram” [Online]. Available: <http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram>[↩︎](#fnref34)
35. “What is GIS?”. [Online]. Available: https://www.esri.com/en-us/what-is-gis/overview[↩︎](#fnref35)
36. “A beginner’s guide to big data terminology.” [Online]. Available: <http://dataconomy.com/2016/05/a-beginners-guide-to-big-data-terminology>[↩︎](#fnref36)
37. “What is a Database?” [Online]. Available: https://www.oracle.com/database/what-is-database/[↩︎](#fnref37)
38. Database model - Wikipedia. [Online]. Available: https://en.wikipedia.org/wiki/Database\_model[↩︎](#fnref38)
39. How Do You Explain A Dendrogram? – chetumenu.com. [Online]. Available: https://chetumenu.com/how-do-you-explain-a-dendrogram/[↩︎](#fnref39)
40. “Descriptive analytics definition by Tech Target.” [Online]. Available: <http://whatis.techtarget.com/definition/descriptive-analytics>[↩︎](#fnref40)
41. “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>[↩︎](#fnref41)
42. “Distributed file system definition by Tech Target.” [Online]. Available: <http://searchwindowsserver.techtarget.com/definition/distributed-file-system-DFS>[↩︎](#fnref42)
43. “Distributed file system definition by Techopedia.” [Online]. Available: <https://www.techopedia.com/definition/1825/distributed-file-system-dfs>[↩︎](#fnref43)
44. Entity–relationship model wiki | TheReaderWiki. [Online]. Available: https://thereaderwiki.com/en/Entity\_relationship\_diagram[↩︎](#fnref44)
45. “Data Modeling Introduction” [Online]. Available: <https://docs.mongodb.com/manual/core/data-modeling-introduction/>[↩︎](#fnref45)
46. Donoho, D. (2017). 50 years of data science. *Journal of Computational and Graphical Statistics*, *26*(4), 745-766.[↩︎](#fnref46)
47. “What is GIS?”. [Online]. Available: https://www.esri.com/en-us/what-is-gis/overview[↩︎](#fnref47)
48. Heatmap – Biomedcourses. https://www.biomedcourses.com/2021/11/24/heatmap/[↩︎](#fnref48)
49. “Data vs. Information vs. Insight” [Online]. Available: https://online.ben.edu/programs/mba/resources/data-vs-information-vs-insight[↩︎](#fnref49)
50. “Data vs. Information vs. Insight” [Online]. Available: https://online.ben.edu/programs/mba/resources/data-vs-information-vs-insight[↩︎](#fnref50)
51. “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>[↩︎](#fnref51)
52. “Internet of things definition by Tech Target.” [Online]. Available: <http://internetofthingsagenda.techtarget.com/definition/Internet-of-Things-IoT>[↩︎](#fnref52)
53. “SQL Joins – W3 Schools”. [Online]. Available: https://www.w3schools.com/sql/sql\_join.asp[↩︎](#fnref53)
54. “Overview of the KDD Process” [Online]. Available: http://www2.cs.uregina.ca/~dbd/cs831/notes/kdd/1\_kdd.html[↩︎](#fnref54)
55. “What is GIS?”. [Online]. Available: https://www.esri.com/en-us/what-is-gis/overview[↩︎](#fnref55)
56. Greenberg, J. (2003). Metadata and the world wide web. *Encyclopedia of library and information science*, *3*, 1876-1888.[↩︎](#fnref56)
57. Pease, Craig M. and Bull, James J. (2006). Models are the Building Blocks of Science. In *Scientific Decision-making* [Course book] (The University of Texas at Austin, Bio301D)\_. Available: http://utw10426.utweb.utexas.edu/Topics/Models/Text.html[↩︎](#fnref57)
58. “Computer network definition by Techopedia.” [Online]. Available: <https://www.techopedia.com/definition/25597/computer-network>[↩︎](#fnref58)
59. “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>[↩︎](#fnref59)
60. “Social Network Analysis: An Introduction by Orgnet,LLC.” [Online]. Available: <http://www.orgnet.com/sna.html>[↩︎](#fnref60)
61. “Null (SQL)”. [Online]. Wikipedia, The Free Encyclopedia. Available: https://en.wikipedia.org/w/index.php?title=Null\_(SQL)[↩︎](#fnref61)
62. “Population”. Oxford University Press. [Online]. Available: https://www.lexico.com/definition/population[↩︎](#fnref62)
63. “A beginner’s guide to big data terminology.” [Online]. Available: <http://dataconomy.com/2016/05/a-beginners-guide-to-big-data-terminology>[↩︎](#fnref63)
64. Donoho, D. (2017). 50 years of data science. *Journal of Computational and Graphical Statistics*, *26*(4), 745-766.[↩︎](#fnref64)
65. “A beginner’s guide to big data terminology.” [Online]. Available: <http://dataconomy.com/2016/05/a-beginners-guide-to-big-data-terminology>[↩︎](#fnref65)
66. Models - Micron. [Online]. Available: https://kimtoo.gitbook.io/workspace/models[↩︎](#fnref66)
67. “Statistics: Power from Data! Glossary” [Online]. Available: https://www150.statcan.gc.ca/n1/edu/power-pouvoir/glossary-glossaire/5214842-eng.htm#Sample[↩︎](#fnref67)
68. “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>[↩︎](#fnref68)
69. “Mini-glossary: Big data terms you should know.” [Online]. Available: <http://www.techrepublic.com/article/mini-glossary-big-data-terms-you-should-know>[↩︎](#fnref69)
70. Donoho, D. (2017). 50 years of data science. *Journal of Computational and Graphical Statistics*, *26*(4), 745-766.[↩︎](#fnref70)
71. “SQL Tutorial” [Online]. Available: https://www.w3schools.com/sql/[↩︎](#fnref71)
72. “Data Storytelling: How to Effectively Tell a Story with Data”. [Blog].Harvard Business School. Available: https://online.hbs.edu/blog/post/data-storytelling[↩︎](#fnref72)
73. “Parse”. Oxford University Press. [Online]. Available: https://www.lexico.com/definition/parse[↩︎](#fnref73)
74. “Introduction to Structured Data, by Google.” [Online]. Available: <https://developers.google.com/search/docs/guides/intro-structured-data>[↩︎](#fnref74)
75. “Structured data definition by Webopedia.” [Online]. Available: <http://www.webopedia.com/TERM/S/structured_data.html>[↩︎](#fnref75)
76. “Introduction to Tabular Data” [Online]. Available: <https://papl.cs.brown.edu/2016/intro-tabular-data.html>[↩︎](#fnref76)
77. “Tabular Database” [Online]. Available: <https://www.techopedia.com/definition/26181/tabular-database>[↩︎](#fnref77)
78. “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[↩︎](#fnref78)
79. “Tidy Data”. [Online]. R for Data Science. Available: https://r4ds.had.co.nz/tidy-data.html[↩︎](#fnref79)
80. “Tree” from the National Institute of Standards and Technology. [Online]. Available: https://xlinux.nist.gov/dads/HTML/tree.html[↩︎](#fnref80)
81. “Unstructured data definition by Techopedia.” [Online]. Available: <https://www.techopedia.com/definition/13865/unstructured-data>[↩︎](#fnref81)
82. “Solving the Unstructured Data Challenge.” [Online]. Available: <http://www.cio.com/article/2941015/big-data/solving-the-unstructured-data-challenge.html>[↩︎](#fnref82)
83. Variance - Wikipedia. [Online]. Available: https://en.wikipedia.org/wiki/Variance[↩︎](#fnref83)
84. “Data visualization definition by Tech Target.” [Online]. Available: <http://searchbusinessanalytics.techtarget.com/definition/data-visualization>[↩︎](#fnref84)
85. “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>[↩︎](#fnref85)
86. Tufte, E. (1983). Theory of Data Graphics. *The Visual Display of Quantitative Information*, pp 57 and 91-123. Graphics Press. Available: https://archive.org/details/visualdisplayofq0000tuft[↩︎](#fnref86)