Data Literacy: Glossary and Index

Table of Contents

[Glossary 8](#_Toc97749266)

[ADVANCED RESEARCH COMPUTING 8](#_Toc97749267)

[Aggregation (Data) 8](#_Toc97749268)

[ALGORITHM 8](#_Toc97749269)

[ANALYTICS 8](#_Toc97749270)

[API 8](#_Toc97749271)

[Arithmetic operations 9](#_Toc97749272)

[ARTIFICIAL INTELLIGENCE 9](#_Toc97749273)

[Bar Chart 9](#_Toc97749274)

[BEHAVIOUR(AL) ANALYTICS 9](#_Toc97749275)

[BIG DATA 9](#_Toc97749276)

[Big data isn’t always big. 9](#_Toc97749277)

[Binning 9](#_Toc97749278)

[Box Plot 9](#_Toc97749279)

[Bucketing 9](#_Toc97749280)

[CAUSAL INFERENCE 10](#_Toc97749281)

[Chartjunk 10](#_Toc97749282)

[CLASSIFICATION 10](#_Toc97749283)

[Cloud (“The Cloud”) 10](#_Toc97749284)

[CLOUD COMPUTING 10](#_Toc97749285)

[Cloud Storage 10](#_Toc97749286)

[CLUSTER (Compute) 10](#_Toc97749287)

[CLUSTERING 10](#_Toc97749288)

[Colour Map 11](#_Toc97749289)

[COMPUTING WITH DATA 11](#_Toc97749290)

[Correlation 11](#_Toc97749291)

[CORRELATION MINING 11](#_Toc97749292)

[Critical Thinking 11](#_Toc97749293)

[Crowdsourcing 11](#_Toc97749294)

[CYBERSECURITY 11](#_Toc97749295)

[DATA 11](#_Toc97749296)

[Data (Clean, Dirty) 12](#_Toc97749297)

[DATA AGGREGATION 12](#_Toc97749298)

[Data Analysis [TODO: Seems more like a definition for Data Transformation(?)] 12](#_Toc97749299)

[Data Bias 12](#_Toc97749300)

[DATA CLEANING 12](#_Toc97749301)

[DATA EXPLORATION AND PREPARATION 12](#_Toc97749302)

[DATA INTEGRATION 12](#_Toc97749303)

[Data Journalism 12](#_Toc97749304)

[Data Lake 12](#_Toc97749305)

[Data Literacy 13](#_Toc97749306)

[DATA MINING 13](#_Toc97749307)

[Data Model 13](#_Toc97749308)

[Data Privacy 13](#_Toc97749309)

[Data Quality 13](#_Toc97749310)

[DATA REPRESENTATION 13](#_Toc97749311)

[DATA SCIENCE 13](#_Toc97749312)

[DATA TRANSFORMATION 14](#_Toc97749313)

[Data types 14](#_Toc97749314)

[Data Warehouse 14](#_Toc97749315)

[Data-to-ink ratio 14](#_Toc97749316)

[DATABASE 14](#_Toc97749317)

[Database 2 14](#_Toc97749318)

[Database Model 15](#_Toc97749319)

[Dendrogram 15](#_Toc97749320)

[DESCRIPTIVE ANALYTICS 15](#_Toc97749321)

[DISTRIBUTED FILE SYSTEM 15](#_Toc97749322)

[Entity–Relationship Model 15](#_Toc97749323)

[FEATURE 15](#_Toc97749324)

[FEATURE ENGINEERING 15](#_Toc97749325)

[File Formats 16](#_Toc97749326)

[Filter 16](#_Toc97749327)

[FLEXIBLE SCHEMA 16](#_Toc97749328)

[Foreign key 16](#_Toc97749329)

[GENERATIVE MODELING 17](#_Toc97749330)

[GIS 17](#_Toc97749331)

[GOODNESS OF FIT 17](#_Toc97749332)

[Group by 17](#_Toc97749333)

[Heatmap 17](#_Toc97749334)

[Histogram 17](#_Toc97749335)

[HTML 17](#_Toc97749336)

[Infographics 18](#_Toc97749337)

[INFORMATION 18](#_Toc97749338)

[INSIGHT 18](#_Toc97749339)

[IOT (INTERNET OF THINGS) 18](#_Toc97749340)

[Join 18](#_Toc97749341)

[KNOWLEDGE DISCOVERY 18](#_Toc97749342)

[Lie factor 18](#_Toc97749343)

[Line Plot 18](#_Toc97749344)

[Logical operations 18](#_Toc97749345)

[MACHINE LEARNING 19](#_Toc97749346)

[METADATA 19](#_Toc97749347)

[Model 19](#_Toc97749348)

[MODEL FITTING 19](#_Toc97749349)

[NETWORK 19](#_Toc97749350)

[NETWORK ANALYSIS 19](#_Toc97749351)

[NO-SQL DATABASE 19](#_Toc97749352)

[Null values 19](#_Toc97749353)

[Pair plot 19](#_Toc97749354)

[Pictographs Plot 20](#_Toc97749355)

[Pie Chart 20](#_Toc97749356)

[Pivot 20](#_Toc97749357)

[Population 20](#_Toc97749358)

[PREDICTIVE ANALYTICS 20](#_Toc97749359)

[PRESCRIPTIVE ANALYTICS 20](#_Toc97749360)

[Primary key 20](#_Toc97749361)

[Query (Data) 20](#_Toc97749362)

[REGRESSION 20](#_Toc97749363)

[RELATIONSHIP MINING 21](#_Toc97749364)

[Repository (data repo) 21](#_Toc97749365)

[Sample (Representative, Proxy) 21](#_Toc97749366)

[Scatter Plot 21](#_Toc97749367)

[Schema (Database) 21](#_Toc97749368)

[SECURE COMPUTING 21](#_Toc97749369)

[Select 21](#_Toc97749370)

[Skewness 21](#_Toc97749371)

[Sort 21](#_Toc97749372)

[SQL 21](#_Toc97749373)

[Standard Deviation 22](#_Toc97749374)

[Statistics 22](#_Toc97749375)

[Storytelling 22](#_Toc97749376)

[String or character operations 22](#_Toc97749377)

[STRUCTURED DATA 22](#_Toc97749378)

[TABULAR DATA 22](#_Toc97749379)

[TEXT MINING 23](#_Toc97749380)

[Tidy tables 23](#_Toc97749381)

[Time Series 23](#_Toc97749382)

[Transpose 23](#_Toc97749383)

[Tree structure 23](#_Toc97749384)

[Unbalanced Data 23](#_Toc97749385)

[Union 23](#_Toc97749386)

[UNSTRUCTURED DATA 24](#_Toc97749387)

[URL 24](#_Toc97749388)

[Variance 24](#_Toc97749389)

[Violin Plot 24](#_Toc97749390)

[VISUALIZATION 24](#_Toc97749391)

[Visualization 24](#_Toc97749392)

[VISUALIZATION DASHBOARD 24](#_Toc97749393)

[Web scraping 24](#_Toc97749394)

[Word Cloud 25](#_Toc97749395)

[Categories 25](#_Toc97749396)

[Data Inquiry Disciplines 25](#_Toc97749397)

[Data Analysis Techniques 26](#_Toc97749398)

[Data Processing 26](#_Toc97749399)

[Storage and Physical Data Sources 27](#_Toc97749400)

[Types of Data 27](#_Toc97749401)

[Operations 28](#_Toc97749402)

[Lessons Index 29](#_Toc97749403)

[Lessons 01 and 02 29](#_Toc97749404)

[Data Literacy 29](#_Toc97749405)

[Lesson 03 29](#_Toc97749406)

[ARTIFICIAL INTELLIGENCE 29](#_Toc97749407)

[Crowdsourcing 29](#_Toc97749408)

[DATA 29](#_Toc97749409)

[DATA TRANSFORMATION 29](#_Toc97749410)

[Data types 29](#_Toc97749411)

[File Formats 30](#_Toc97749412)

[MACHINE LEARNING 30](#_Toc97749413)

[METADATA 31](#_Toc97749414)

[Pivot 31](#_Toc97749415)

[STRUCTURED DATA 31](#_Toc97749416)

[TABULAR DATA 31](#_Toc97749417)

[Tree structure 31](#_Toc97749418)

[UNSTRUCTURED DATA 31](#_Toc97749419)

[Lesson 04 32](#_Toc97749420)

[BIG DATA 32](#_Toc97749421)

[Big data isn’t always big. 32](#_Toc97749422)

[Critical Thinking 32](#_Toc97749423)

[Data (Clean, Dirty) 32](#_Toc97749424)

[Data Bias 32](#_Toc97749425)

[Data Privacy 32](#_Toc97749426)

[Data Quality 33](#_Toc97749427)

[Population 33](#_Toc97749428)

[Sample (Representative, Proxy) 33](#_Toc97749429)

[Unbalanced Data 33](#_Toc97749430)

[Lesson 05 33](#_Toc97749431)

[Correlation 33](#_Toc97749432)

[Histogram 33](#_Toc97749433)

[Standard Deviation 33](#_Toc97749434)

[Statistics 34](#_Toc97749435)

[Variance 34](#_Toc97749436)

[VISUALIZATION 34](#_Toc97749437)

[Visualization 34](#_Toc97749438)

[Lesson 06 34](#_Toc97749439)

[Data (Clean, Dirty) 34](#_Toc97749440)

[DATA INTEGRATION 34](#_Toc97749441)

[DATA TRANSFORMATION 34](#_Toc97749442)

[Join 35](#_Toc97749443)

[Null values 35](#_Toc97749444)

[Tidy tables 35](#_Toc97749445)

[DATA TRANSFORMATION 35](#_Toc97749446)

[Data (Clean, Dirty) 35](#_Toc97749447)

[Lesson 07 35](#_Toc97749448)

[Filter 35](#_Toc97749449)

[Group by 35](#_Toc97749450)

[Query (Data) 35](#_Toc97749451)

[Select 36](#_Toc97749452)

[Sort 36](#_Toc97749453)

[SQL 36](#_Toc97749454)

[Lesson 08 36](#_Toc97749455)

[Aggregation (Data) 36](#_Toc97749456)

[Arithmetic operations 36](#_Toc97749457)

[Binning 36](#_Toc97749458)

[Bucketing 36](#_Toc97749459)

[DATA TRANSFORMATION 36](#_Toc97749460)

[Join 37](#_Toc97749461)

[Logical operations 37](#_Toc97749462)

[Pivot 37](#_Toc97749463)

[String or character operations 37](#_Toc97749464)

[Transpose 37](#_Toc97749465)

[DATA TRANSFORMATION 37](#_Toc97749466)

[Pivot 37](#_Toc97749467)

[DATA TRANSFORMATION 38](#_Toc97749468)

[Join 38](#_Toc97749469)

[DATA TRANSFORMATION 38](#_Toc97749470)

[Lesson 09 38](#_Toc97749471)

[Bar Chart 38](#_Toc97749472)

[Box Plot 38](#_Toc97749473)

[Correlation 38](#_Toc97749474)

[Data Analysis [TODO: Seems more like a definition for Data Transformation(?)] 38](#_Toc97749475)

[Dendrogram 39](#_Toc97749476)

[Heatmap 39](#_Toc97749477)

[Histogram 39](#_Toc97749478)

[Line Plot 39](#_Toc97749479)

[Model 39](#_Toc97749480)

[Pair plot 39](#_Toc97749481)

[Pie Chart 39](#_Toc97749482)

[Scatter Plot 40](#_Toc97749483)

[Skewness 40](#_Toc97749484)

[Standard Deviation 40](#_Toc97749485)

[Violin Plot 40](#_Toc97749486)

[VISUALIZATION 40](#_Toc97749487)

[Visualization 40](#_Toc97749488)

[Word Cloud 40](#_Toc97749489)

[Correlation 40](#_Toc97749490)

[Histogram 41](#_Toc97749491)

[Standard Deviation 41](#_Toc97749492)

[VISUALIZATION 41](#_Toc97749493)

[Visualization 41](#_Toc97749494)

[Lesson 10 41](#_Toc97749495)

[API 41](#_Toc97749496)

[Cloud (“The Cloud”) 41](#_Toc97749497)

[CLUSTER (Compute) 42](#_Toc97749498)

[DATA INTEGRATION 42](#_Toc97749499)

[DATABASE 42](#_Toc97749500)

[Database 2 42](#_Toc97749501)

[GIS 42](#_Toc97749502)

[HTML 42](#_Toc97749503)

[Query (Data) 42](#_Toc97749504)

[Repository (data repo) 43](#_Toc97749505)

[SQL 43](#_Toc97749506)

[URL 43](#_Toc97749507)

[Web scraping 43](#_Toc97749508)

[DATA INTEGRATION 43](#_Toc97749509)

[Query (Data) 43](#_Toc97749510)

[SQL 43](#_Toc97749511)

[Lesson 11 44](#_Toc97749512)

[Cloud Storage 44](#_Toc97749513)

[DATA INTEGRATION 44](#_Toc97749514)

[Data Lake 44](#_Toc97749515)

[Data Model 44](#_Toc97749516)

[Data Warehouse 44](#_Toc97749517)

[DATABASE 44](#_Toc97749518)

[Database 2 45](#_Toc97749519)

[Database Model 45](#_Toc97749520)

[Entity–Relationship Model 45](#_Toc97749521)

[METADATA 45](#_Toc97749522)

[Query (Data) 45](#_Toc97749523)

[Schema (Database) 45](#_Toc97749524)

[SQL 46](#_Toc97749525)

[METADATA 46](#_Toc97749526)

[DATA INTEGRATION 46](#_Toc97749527)

[Query (Data) 46](#_Toc97749528)

[SQL 46](#_Toc97749529)

[DATA INTEGRATION 46](#_Toc97749530)

[DATABASE 46](#_Toc97749531)

[Database 2 46](#_Toc97749532)

[Query (Data) 47](#_Toc97749533)

[SQL 47](#_Toc97749534)

[DATA INTEGRATION 47](#_Toc97749535)

[Query (Data) 47](#_Toc97749536)

[SQL 47](#_Toc97749537)

[Lesson 12 47](#_Toc97749538)

[DATA INTEGRATION 47](#_Toc97749539)

[DATA TRANSFORMATION 47](#_Toc97749540)

[Foreign key 48](#_Toc97749541)

[Join 48](#_Toc97749542)

[Primary key 48](#_Toc97749543)

[SQL 48](#_Toc97749544)

[Union 48](#_Toc97749545)

[DATA TRANSFORMATION 48](#_Toc97749546)

[DATA INTEGRATION 48](#_Toc97749547)

[DATA TRANSFORMATION 48](#_Toc97749548)

[Join 48](#_Toc97749549)

[DATA TRANSFORMATION 49](#_Toc97749550)

[SQL 49](#_Toc97749551)

[DATA TRANSFORMATION 49](#_Toc97749552)

[Join 49](#_Toc97749553)

[DATA TRANSFORMATION 49](#_Toc97749554)

[DATA TRANSFORMATION 49](#_Toc97749555)

[Join 49](#_Toc97749556)

[DATA TRANSFORMATION 49](#_Toc97749557)

[DATA INTEGRATION 49](#_Toc97749558)

[SQL 50](#_Toc97749559)

[DATA INTEGRATION 50](#_Toc97749560)

[SQL 50](#_Toc97749561)

[DATA INTEGRATION 50](#_Toc97749562)

[SQL 50](#_Toc97749563)

[DATA INTEGRATION 50](#_Toc97749564)

[SQL 50](#_Toc97749565)

[DATA INTEGRATION 50](#_Toc97749566)

[SQL 50](#_Toc97749567)

[DATA INTEGRATION 51](#_Toc97749568)

[SQL 51](#_Toc97749569)

[Lesson 13 51](#_Toc97749570)

[Bar Chart 51](#_Toc97749571)

[Data Journalism 51](#_Toc97749572)

[Histogram 51](#_Toc97749573)

[Infographics 51](#_Toc97749574)

[Line Plot 51](#_Toc97749575)

[Pie Chart 51](#_Toc97749576)

[Scatter Plot 51](#_Toc97749577)

[Storytelling 52](#_Toc97749578)

[Time Series 52](#_Toc97749579)

[VISUALIZATION 52](#_Toc97749580)

[Visualization 52](#_Toc97749581)

[Histogram 52](#_Toc97749582)

[VISUALIZATION 52](#_Toc97749583)

[Visualization 52](#_Toc97749584)

[Bar Chart 53](#_Toc97749585)

[Histogram 53](#_Toc97749586)

[Line Plot 53](#_Toc97749587)

[Pie Chart 53](#_Toc97749588)

[Scatter Plot 53](#_Toc97749589)

[VISUALIZATION 53](#_Toc97749590)

[Visualization 53](#_Toc97749591)

[Histogram 53](#_Toc97749592)

[VISUALIZATION 54](#_Toc97749593)

[Visualization 54](#_Toc97749594)

[Lesson 14 54](#_Toc97749595)

[Bar Chart 54](#_Toc97749596)

[Chartjunk 54](#_Toc97749597)

[Colour Map 54](#_Toc97749598)

[Data-to-ink ratio 54](#_Toc97749599)

[Histogram 54](#_Toc97749600)

[Lie factor 55](#_Toc97749601)

[Line Plot 55](#_Toc97749602)

[Pictographs Plot 55](#_Toc97749603)

[Pie Chart 55](#_Toc97749604)

[Scatter Plot 55](#_Toc97749605)

[Histogram 55](#_Toc97749606)

[Bar Chart 55](#_Toc97749607)

[Histogram 55](#_Toc97749608)

[Line Plot 55](#_Toc97749609)

[Pie Chart 56](#_Toc97749610)

[Scatter Plot 56](#_Toc97749611)

[Histogram 56](#_Toc97749612)

[Bar Chart 56](#_Toc97749613)

[Histogram 56](#_Toc97749614)

[Line Plot 56](#_Toc97749615)

[Pie Chart 56](#_Toc97749616)

[Scatter Plot 56](#_Toc97749617)

[Histogram 56](#_Toc97749618)

[Bar Chart 57](#_Toc97749619)

[Histogram 57](#_Toc97749620)

[Line Plot 57](#_Toc97749621)

[Pie Chart 57](#_Toc97749622)

[Scatter Plot 57](#_Toc97749623)

[Histogram 57](#_Toc97749624)

[Lesson 15 57](#_Toc97749625)

[References 57](#_Toc97749626)

# 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>

### Aggregation (Data)

A cluster of things that have come or been brought together. [3](#fn3) <#operations> <#lesson-08>

TODO: same as entry “Data Aggregation” (?)

### ALGORITHM

A sequence of instructions telling a computer how to answer a specific question. [4](#fn4),[5](#fn5) <#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.[6](#fn6),[7](#fn7) <#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. [8](#fn8) <#disciplines> <#lesson-03>

### Bar Chart

TODO: Bar Chart ….vertical, horizontal, grouped, stacked. [visualization](#visualization) …. <#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.[9](#fn9),[10](#fn10) <#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). [11](#fn11),[12](#fn12) <#disciplines> <#lesson-04>

### Big data isn’t always big.

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

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

### 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

TODO: Box Plot .. [visualization](#visualization) …. <#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 graphics 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. Alsa see [data-to-ink-ratio](#data-to-ink-ratio) and [lie-factor](#lie-factor). [13](#fn13) <#lesson-14>

### CLASSIFICATION

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

### Cloud (“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. [14](#fn14),[15](#fn15) <#disciplines>

### Cloud Storage

Cloud storage is a [cloud](#cloud) computing model that stores data on the Internet through a cloud computing provider who manages and operates data storage as a service. It’s delivered on demand with just-in-time capacity and costs, and eliminates buying and managing your own data storage infrastructure. This gives you agility, global scale and durability, with “anytime, anywhere” data access. [^] <#storage> <#lesson-11>

[^]: “What is Cloud Storage? | AWS” . [Online]. Available: https://aws.amazon.com/what-is-cloud-storage/

### 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

TODO: Colour … Synonyms: Color map, Color palette, Color scheme. Color(Hue), Color(Shade) [visualization](#visualization) …. <#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.[16](#fn16) <#disciplines>

### Correlation

Correlation or dependence is any statistical relationship between two random variables or bivariate data, whether causal or not. In the broadest sense, correlation is any statistical association, though it accurately refers to the degree to which a pair of variables are linearly related. Correlation is not equivalent to [causation](#causal-inference). <#data-analysis> <#lesson-05> <#lesson-09>

[^]: Sarvjit Kaur on LinkedIn: #dataskience #comparison # …. https://www.linkedin.com/posts/sarvjit-kaur\_dataskience-comparison-correlation-activity-6884786381676249088-fUlX

### CORRELATION MINING

As a measure of association between two variables, correlation can be used for [prediction](#predictive-analytics) and indicate the presence of [causal relationship](#causal-inference). It does, however, not sufficiently imply causation. [17](#fn17) <#data-analysis>

### Critical Thinking

The objective analysis and evaluation of an issue to form a judgement [18](#fn18). The process of analyzing information in order to make a logical decision about the extent to which you believe something to be true or false [19](#fn19). <#disciplines> <#lesson-04>

[^]: From 2 different Oxford dictionary sources. See footnotes [20](#fn20) and [21](#fn21).

### Crowdsourcing

TODO. Crowdsourcing <#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). [22](#fn22) <#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 data, misplaced data, Inconsistenti, or duplicated. [23](#fn23) <#processing> <#lesson-04> <#lesson-06>

### DATA AGGREGATION

TODO: DATA AGGREGATION. And entry for “[Aggregation](#aggregation)” already exists.

### Data Analysis [TODO: Seems more like a definition for Data Transformation(?)]

Data analysis is a process of inspecting, cleansing, transforming, and modelling data to discover useful information, informing conclusions, and supporting decision-making. <#lesson-09>

TODO: Change “Data Analysis” to “Analytics”. However, an entry for “[Analytics](#analytics)” already exists. Compare and consolidate.

### 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. [24](#fn24) <#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. [25](#fn25) <#disciplines> <#lesson-13>

### Data Lake

A data lake is a storage repository that holds a vast amount of raw data in its native format until it is needed for analytics applications. While a traditional [data warehouse](#data-warehouse) stores data in hierarchical dimensions and tables, a data lake uses a flat architecture to store data, primarily in files or object storage. [^] <#storage> <#lesson-11>

[^]: What Is a Data Lake? Definition from SearchDataManagement. https://searchdatamanagement.techtarget.com/definition/data-lake

### 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 (or datamodel) is an abstract model that organizes elements of data and standardizes how they relate to one another and to attributes real-world entities. [^] <#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. [29](#fn29) <#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 accesible) and accurate. [30](#fn30) <#processing> <#lesson-04>

### DATA REPRESENTATION

The form in which [data](#data) are stored, processed and transmitted such that its [information content](#information) and [context](#metadata) are retained as much as possible. Choices of form are influenced by hardware, software or other constraints around processing and analysis resources. [31](#fn31) <#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). [32](#fn32),[33](#fn33) <#disciplines>

### DATA TRANSFORMATION

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

In computing, a data warehouse, also known as an enterprise data warehouse, is a system used for reporting and data analysis and is considered a core component of business intelligence. DWs are central repositories of integrated data from one or more disparate sources. [^] <#storage> <#lesson-11>

[^]: Data Warehouse Interview Questions | TestDome. https://www.testdome.com/d/data-warehouse-interview-questions/1203

### 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). [34](#fn34) <#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.[35](#fn35) <#storage> <#lesson-10> <#lesson-11>

### Database 2

A database is an organized collection of [structured](#structured) information, or data typically stored electronically in a computer system. <#storage> <#lesson-10> <#lesson-11>

[^]: https://www.oracle.com/database/what-is-database/

### Database Model

A database model is a type of [data model](#data-model) that determines its logical structure. It fundamentally determines how data can be stored, organized and manipulated. The most popular database model is the relational model, which uses a table-based format. [^] <#storage> <#lesson-11>

[^]: Database model - Wikipedia. https://en.wikipedia.org/wiki/Database\_model

### Dendrogram

A dendrogram is a [diagram](#visualization) representing a tree. This diagrammatic representation is frequently used in different contexts: in hierarchical [clustering](#clustering), it illustrates the arrangement of the clusters produced by the corresponding analyses. <#lesson-09>

[^]: How Do You Explain A Dendrogram? – chetumenu.com. https://chetumenu.com/how-do-you-explain-a-dendrogram/

### 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?”[36](#fn36),[37](#fn37) <#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.[38](#fn38),[39](#fn39) <#storage>

### Entity–Relationship Model

An entity–relationship model (or ER model) describes interrelated things of interest in a specific domain of knowledge. A basic ER model is composed of entity types (which classify the 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. [^] <#storage> <#lesson-11>

[^]: Entity–relationship model wiki | TheReaderWiki. https://thereaderwiki.com/en/Entity\_relationship\_diagram

### 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

TODO: Filter …not restricted to [SQL](#sql)… <#operations> <#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.[40](#fn40) <#data-types>

### Foreign key

TODO: Foreign key, skip? …not restricted to [SQL](#sql)… <#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. [41](#fn41)

### GIS

A geographic information system (GIS) is a system that creates, manages, analyzes, and maps all types of data. GIS connects data to a map, integrating location data (where things are) with all types of descriptive information (what things are like there). This provides a foundation for mapping and analysis that is used in science and almost every industry. GIS helps users understand patterns, relationships, and geographic context. The benefits include improved communication and efficiency as well as better management and decision making. [42](#fn42) <#storage> <#lesson-10>

[^]: From Esri/ArcGIS. See footnote [43](#fn43).

### 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 unchosen columns are [aggregated](#aggregation) or summarized ([arithmetic](#arithmetic-operations), [string](#string) or [logical](#logical-operations)). <#operations> <#lesson-07>

### Heatmap

A heat map (or heatmap) is a data [visualization](#visualization) technique that shows the magnitude of a phenomenon as color in two dimensions. The colour variation may be by [hue or intensity](#colour), giving obvious visual cues to the reader about how the phenomenon is clustered or varies over space. <#lesson-09>

[^]: Heatmap – Biomedcourses. https://www.biomedcourses.com/2021/11/24/heatmap/ [^]: Implementation of Hierarchical clustering using Python. https://hands-on.cloud/implementation-of-hierarchical-clustering-using-python/

### Histogram

A histogram is a graphical representation that organizes a group of data points into user-specified ranges. Similar in appearance to a bar graph, the histogram condenses a data series into an easily interpreted <#visualization> by taking many data points and grouping them into logical, ordered ranges or bins. <#data-analysis> <#lesson-05> <#lesson-09> <#lesson-13> <#lesson-14>

[^]: What is Histogram in Statistics? | Line Served. https://lineserved.com/new/what-is-histogram-in-statistics/

### HTML

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

### Infographics

TODO: Infographics … <#visualization> <#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). [44](#fn44)

### INSIGHT

Actionable [information](#information) gained by interpreting [data analysis](#analytics) results, helping people to make more informed decisions. [45](#fn45)

### 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.[46](#fn46),[47](#fn47) <#storage>

### Join

A *join* operation is used to [combine](#data-transformation) rows from two or more tables, based on a related column between them. [48](#fn48) <#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.[49](#fn49) <#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). [50](#fn50) <#lesson-14>

### Line Plot

TODO: Line Plot …. [visualization](#visualization) …. <#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 analytics](#analytics). <#disciplines> <#lesson-03>

### METADATA

Also called “data about data” it provides a [structured description](#structured-data) and context for a data point - document, image, or file - to help organize, find and understand the data. [51](#fn51) <#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. [52](#fn52) <#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.[53](#fn53),[54](#fn54)

### 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. [55](#fn55)

### 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. [56](#fn56) <#lesson-06>

### Pair plot

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

### Pictographs Plot

TODO: Pictographs Plot …. [visualization](#visualization) …. <#lesson-14>

### Pie Chart

TODO: Pie Chart …. [visualization](#visualization) …. <#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. [57](#fn57) <#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), modeling, [machine learning](#machine-learning) and statistics.[58](#fn58),[59](#fn59)

### PRESCRIPTIVE ANALYTICS

A type of [analytics](#analytics) with the goal of using data to determine the best course of action for a specific scenario. [60](#fn60) <#data-analysis>

### Primary key

In the [relational model](#database-model) of [databases](#database), a primary key is a specific choice of a minimal set of attributes that uniquely specify a tuple in a relation. Informally, a primary key is “which attributes uniquely identify a record,” and in simple cases constitute a single attribute: a unique ID. [^] <#query> <#lesson-12>

[^]: Models - Micron. https://kimtoo.gitbook.io/workspace/models

### Query (Data)

TODO: Query (data) check. Queries are not restricted to SQL A query is a request for data or information from a [table](#tabular) or combination of tables. This data may be generated as results returned by Structured Query Language ([SQL](#sql)) or as pictorials, graphs or complex results, e.g., trend analyses from data-mining tools. <#lesson-07> <#lesson-10> <#lesson-11>

[^]: What is the definition of a query in a database …. https://answerthirst.com/what-is-the-definition-of-a-query-in-a-database/

### 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 often used as 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). [61](#fn61) <#data-analysis> <#lesson-04>

### Scatter Plot

TODO: Scatter Plot …. [visualization](#visualization) …. <#lesson-09> <#lesson-13> <#lesson-14>

### Schema (Database)

The [database](#database) schema is its structure described in a formal language supported by the database management system (DBMS). The term “schema” refers to the organization of data as a blueprint of how the database is constructed. [^] <#storage> <#lesson-11>

[^]: Database schema - CodeDocs. https://codedocs.org/what-is/database-schema

### 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. [62](#fn62),[63](#fn63) <#disciplines>

### Select

TODO: Select …not restricted to [SQL](#sql)… <#operations> <#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

TODO: Sort ….. skip? Also called *rank*… [transformations](#data-transformations)…. <#operations> <#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. [64](#fn64) <#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>

### Storytelling

The ability to tell or write stories and narratives to communicate insights and actions drawn data, optionally aided by effective [visualizations](#visualization). [65](#fn65) <#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.[66](#fn66)

#### 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.[67](#fn67),[68](#fn68) <#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. [69](#fn69),[70](#fn70) <#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 presetns each **Error! Hyperlink reference not valid.** 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.[71](#fn71) <#data-analysis>

### Tidy tables

Datasets are commonly messy. Getting data into tidy [72](#fn72) 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.

TODO: Time series *(better connected with Video 8(?), but first mentioned in 13)* <#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. [73](#fn73) <#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. [74](#fn74),[75](#fn75) <#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

In probability theory and statistics, variance is the expectation of the squared deviation of a random variable from its [population](#population) mean or [sample](#sample) mean. Variance is a measure of dispersion, meaning how far a set of numbers is spread out from their average value. Also see [standard deviation](#standard-deviation). <#operations> <#lesson-05>

[^]: Variance - Wikipedia. https://en.wikipedia.org/wiki/Variance

### 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. [76](#fn76),[77](#fn77) <#data-analysis> <#lesson-05> <#lesson-09> <#lesson-13>

### Visualization

Sometimes referred to as *Viz*. The technique for creating images, diagrams, or animations to communicate messages. *Explanatory*, or informative, visualizations are used to communicate to an audience. *Exploratory* visuals are used to investigate and uncover unknown patterns in data. <#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 degree of relevance by varying the font size of the words. The most important words stand out at a glance, giving an 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)

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

[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)

[DESCRIPTIVE ANALYTICS](#descriptive-analytics)

[Histogram](#histogram)

[Model](#model)

[Population](#population)

[PRESCRIPTIVE ANALYTICS](#prescriptive-analytics)

[REGRESSION](#regression)

[RELATIONSHIP MINING](#relationship-mining)

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

[Statistics](#statistics)

[TEXT MINING](#text-mining)

[VISUALIZATION](#visualization)

[Visualization](#visualization-1)

[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 Lake](#data-lake)

[Data Model](#data-model)

[Data Warehouse](#data-warehouse)

[DATABASE](#database)

[Database 2](#database-2)

[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

[Aggregation (Data)](#aggregation-data)

[Arithmetic operations](#arithmetic-operations)

[Binning](#binning)

[Bucketing](#bucketing)

[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. [8](#fn8) <#disciplines> <#lesson-03>

### Crowdsourcing

TODO. Crowdsourcing <#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

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

### METADATA

Also called “data about data” it provides a [structured description](#structured-data) and context for a data point - document, image, or file - to help organize, find and understand the data. [51](#fn51) <#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.[67](#fn67),[68](#fn68) <#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. [69](#fn69),[70](#fn70) <#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 presetns each **Error! Hyperlink reference not valid.** 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. [73](#fn73) <#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. [74](#fn74),[75](#fn75) <#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). [11](#fn11),[12](#fn12) <#disciplines> <#lesson-04>

### Big data isn’t always big.

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

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

### Critical Thinking

The objective analysis and evaluation of an issue to form a judgement [18](#fn18). The process of analyzing information in order to make a logical decision about the extent to which you believe something to be true or false [19](#fn19). <#disciplines> <#lesson-04>

[^]: From 2 different Oxford dictionary sources. See footnotes [20](#fn20) and [21](#fn21).

### Data (Clean, Dirty)

Data that contains erroneous or unreliable data points is called dirty data. Dirty data must be cleaned to correct and adjusted before it is usable, usually a tedious task. Causes of dirty data include being outdated, insecure, incomplete, incomplete (missing), inaccurate data, misplaced data, Inconsistenti, or duplicated. [23](#fn23) <#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. [29](#fn29) <#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 accesible) and accurate. [30](#fn30) <#processing> <#lesson-04>

### Population

A collection of items under consideration. [57](#fn57) <#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). [61](#fn61) <#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 or dependence is any statistical relationship between two random variables or bivariate data, whether causal or not. In the broadest sense, correlation is any statistical association, though it accurately refers to the degree to which a pair of variables are linearly related. Correlation is not equivalent to [causation](#causal-inference). <#data-analysis> <#lesson-05> <#lesson-09>

[^]: Sarvjit Kaur on LinkedIn: #dataskience #comparison # …. https://www.linkedin.com/posts/sarvjit-kaur\_dataskience-comparison-correlation-activity-6884786381676249088-fUlX

### Histogram

A histogram is a graphical representation that organizes a group of data points into user-specified ranges. Similar in appearance to a bar graph, the histogram condenses a data series into an easily interpreted <#visualization> by taking many data points and grouping them into logical, ordered ranges or bins. <#data-analysis> <#lesson-05> <#lesson-09> <#lesson-13> <#lesson-14>

[^]: What is Histogram in Statistics? | Line Served. https://lineserved.com/new/what-is-histogram-in-statistics/

### 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>

### Variance

In probability theory and statistics, variance is the expectation of the squared deviation of a random variable from its [population](#population) mean or [sample](#sample) mean. Variance is a measure of dispersion, meaning how far a set of numbers is spread out from their average value. Also see [standard deviation](#standard-deviation). <#operations> <#lesson-05>

[^]: Variance - Wikipedia. https://en.wikipedia.org/wiki/Variance

### 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. [76](#fn76),[77](#fn77) <#data-analysis> <#lesson-05> <#lesson-09> <#lesson-13>

### Visualization

Sometimes referred to as *Viz*. The technique for creating images, diagrams, or animations to communicate messages. *Explanatory*, or informative, visualizations are used to communicate to an audience. *Exploratory* visuals are used to investigate and uncover unknown patterns in data. <#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 data, misplaced data, Inconsistenti, or duplicated. [23](#fn23) <#processing> <#lesson-04> <#lesson-06>

### DATA INTEGRATION

The process of combining information from different data sources in preparation for data processing. [24](#fn24) <#processing> <#lesson-06> <#lesson-10> <#lesson-11> <#lesson-12>

### DATA TRANSFORMATION

The process of converting data from [one form into another](#data-representation) to gain better [insight](#insight). For instance, converting [unstructured data](#unstructured-data) into [structured form](#structured-data) that can be analyzed further. <#processing> <#operations> <#lesson-03> <#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. [48](#fn48) <#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. [56](#fn56) <#lesson-06>

### Tidy tables

Datasets are commonly messy. Getting data into tidy [72](#fn72) 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>

### DATA TRANSFORMATION

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

### Data (Clean, Dirty)

Data that contains erroneous or unreliable data points is called dirty data. Dirty data must be cleaned to correct and adjusted before it is usable, usually a tedious task. Causes of dirty data include being outdated, insecure, incomplete, incomplete (missing), inaccurate data, misplaced data, Inconsistenti, or duplicated. [23](#fn23) <#processing> <#lesson-04> <#lesson-06>

## Lesson 07

### Filter

TODO: Filter …not restricted to [SQL](#sql)… <#operations> <#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 unchosen columns are [aggregated](#aggregation) or summarized ([arithmetic](#arithmetic-operations), [string](#string) or [logical](#logical-operations)). <#operations> <#lesson-07>

### Query (Data)

TODO: Query (data) check. Queries are not restricted to SQL A query is a request for data or information from a [table](#tabular) or combination of tables. This data may be generated as results returned by Structured Query Language ([SQL](#sql)) or as pictorials, graphs or complex results, e.g., trend analyses from data-mining tools. <#lesson-07> <#lesson-10> <#lesson-11>

[^]: What is the definition of a query in a database …. https://answerthirst.com/what-is-the-definition-of-a-query-in-a-database/

### Select

TODO: Select …not restricted to [SQL](#sql)… <#operations> <#lesson-07>

### Sort

TODO: Sort ….. skip? Also called *rank*… [transformations](#data-transformations)…. <#operations> <#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. [64](#fn64) <#processing> <#lesson-07> <#lesson-10> <#lesson-11> <#lesson-12>

## Lesson 08

### Aggregation (Data)

A cluster of things that have come or been brought together. [3](#fn3) <#operations> <#lesson-08>

TODO: same as entry “Data Aggregation” (?)

### 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 TRANSFORMATION

The process of converting data from [one form into another](#data-representation) to gain better [insight](#insight). For instance, converting [unstructured data](#unstructured-data) into [structured form](#structured-data) that can be analyzed further. <#processing> <#operations> <#lesson-03> <#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. [48](#fn48) <#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.[66](#fn66)

#### Indexing

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

#### Concatenation

Combining items (strings) end-to-end.

### Transpose

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

### DATA TRANSFORMATION

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

### 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>

### DATA TRANSFORMATION

The process of converting data from [one form into another](#data-representation) to gain better [insight](#insight). For instance, converting [unstructured data](#unstructured-data) into [structured form](#structured-data) that can be analyzed further. <#processing> <#operations> <#lesson-03> <#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. [48](#fn48) <#query> <#operations> <#lesson-06> <#lesson-08> <#lesson-12>

### DATA TRANSFORMATION

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

## Lesson 09

### Bar Chart

TODO: Bar Chart ….vertical, horizontal, grouped, stacked. [visualization](#visualization) …. <#lesson-09> <#lesson-13> <#lesson-14>

### Box Plot

TODO: Box Plot .. [visualization](#visualization) …. <#lesson-09>

### Correlation

Correlation or dependence is any statistical relationship between two random variables or bivariate data, whether causal or not. In the broadest sense, correlation is any statistical association, though it accurately refers to the degree to which a pair of variables are linearly related. Correlation is not equivalent to [causation](#causal-inference). <#data-analysis> <#lesson-05> <#lesson-09>

[^]: Sarvjit Kaur on LinkedIn: #dataskience #comparison # …. https://www.linkedin.com/posts/sarvjit-kaur\_dataskience-comparison-correlation-activity-6884786381676249088-fUlX

### Data Analysis [TODO: Seems more like a definition for Data Transformation(?)]

Data analysis is a process of inspecting, cleansing, transforming, and modelling data to discover useful information, informing conclusions, and supporting decision-making. <#lesson-09>

TODO: Change “Data Analysis” to “Analytics”. However, an entry for “[Analytics](#analytics)” already exists. Compare and consolidate.

### Dendrogram

A dendrogram is a [diagram](#visualization) representing a tree. This diagrammatic representation is frequently used in different contexts: in hierarchical [clustering](#clustering), it illustrates the arrangement of the clusters produced by the corresponding analyses. <#lesson-09>

[^]: How Do You Explain A Dendrogram? – chetumenu.com. https://chetumenu.com/how-do-you-explain-a-dendrogram/

### Heatmap

A heat map (or heatmap) is a data [visualization](#visualization) technique that shows the magnitude of a phenomenon as color in two dimensions. The colour variation may be by [hue or intensity](#colour), giving obvious visual cues to the reader about how the phenomenon is clustered or varies over space. <#lesson-09>

[^]: Heatmap – Biomedcourses. https://www.biomedcourses.com/2021/11/24/heatmap/ [^]: Implementation of Hierarchical clustering using Python. https://hands-on.cloud/implementation-of-hierarchical-clustering-using-python/

### Histogram

A histogram is a graphical representation that organizes a group of data points into user-specified ranges. Similar in appearance to a bar graph, the histogram condenses a data series into an easily interpreted <#visualization> by taking many data points and grouping them into logical, ordered ranges or bins. <#data-analysis> <#lesson-05> <#lesson-09> <#lesson-13> <#lesson-14>

[^]: What is Histogram in Statistics? | Line Served. https://lineserved.com/new/what-is-histogram-in-statistics/

### Line Plot

TODO: Line Plot …. [visualization](#visualization) …. <#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. [52](#fn52) <#data-analysis> <#lesson-09>

### Pair plot

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

### Pie Chart

TODO: Pie Chart …. [visualization](#visualization) …. <#lesson-09> <#lesson-13> <#lesson-14>

### Scatter Plot

TODO: Scatter Plot …. [visualization](#visualization) …. <#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. [76](#fn76),[77](#fn77) <#data-analysis> <#lesson-05> <#lesson-09> <#lesson-13>

### Visualization

Sometimes referred to as *Viz*. The technique for creating images, diagrams, or animations to communicate messages. *Explanatory*, or informative, visualizations are used to communicate to an audience. *Exploratory* visuals are used to investigate and uncover unknown patterns in data. <#data-analysis> <#lesson-05> <#lesson-09> <#lesson-13>

### Word Cloud

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

### Correlation

Correlation or dependence is any statistical relationship between two random variables or bivariate data, whether causal or not. In the broadest sense, correlation is any statistical association, though it accurately refers to the degree to which a pair of variables are linearly related. Correlation is not equivalent to [causation](#causal-inference). <#data-analysis> <#lesson-05> <#lesson-09>

[^]: Sarvjit Kaur on LinkedIn: #dataskience #comparison # …. https://www.linkedin.com/posts/sarvjit-kaur\_dataskience-comparison-correlation-activity-6884786381676249088-fUlX

### Histogram

A histogram is a graphical representation that organizes a group of data points into user-specified ranges. Similar in appearance to a bar graph, the histogram condenses a data series into an easily interpreted <#visualization> by taking many data points and grouping them into logical, ordered ranges or bins. <#data-analysis> <#lesson-05> <#lesson-09> <#lesson-13> <#lesson-14>

[^]: What is Histogram in Statistics? | Line Served. https://lineserved.com/new/what-is-histogram-in-statistics/

### 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>

### 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. [76](#fn76),[77](#fn77) <#data-analysis> <#lesson-05> <#lesson-09> <#lesson-13>

### Visualization

Sometimes referred to as *Viz*. The technique for creating images, diagrams, or animations to communicate messages. *Explanatory*, or informative, visualizations are used to communicate to an audience. *Exploratory* visuals are used to investigate and uncover unknown patterns in data. <#data-analysis> <#lesson-05> <#lesson-09> <#lesson-13>

## 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. [24](#fn24) <#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.[35](#fn35) <#storage> <#lesson-10> <#lesson-11>

### Database 2

A database is an organized collection of [structured](#structured) information, or data typically stored electronically in a computer system. <#storage> <#lesson-10> <#lesson-11>

[^]: https://www.oracle.com/database/what-is-database/

### GIS

A geographic information system (GIS) is a system that creates, manages, analyzes, and maps all types of data. GIS connects data to a map, integrating location data (where things are) with all types of descriptive information (what things are like there). This provides a foundation for mapping and analysis that is used in science and almost every industry. GIS helps users understand patterns, relationships, and geographic context. The benefits include improved communication and efficiency as well as better management and decision making. [42](#fn42) <#storage> <#lesson-10>

[^]: From Esri/ArcGIS. See footnote [43](#fn43).

### HTML

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

### Query (Data)

TODO: Query (data) check. Queries are not restricted to SQL A query is a request for data or information from a [table](#tabular) or combination of tables. This data may be generated as results returned by Structured Query Language ([SQL](#sql)) or as pictorials, graphs or complex results, e.g., trend analyses from data-mining tools. <#lesson-07> <#lesson-10> <#lesson-11>

[^]: What is the definition of a query in a database …. https://answerthirst.com/what-is-the-definition-of-a-query-in-a-database/

### Repository (data repo)

Sometimes abbreviated as *repo*, it often used as 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. [64](#fn64) <#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>

### DATA INTEGRATION

The process of combining information from different data sources in preparation for data processing. [24](#fn24) <#processing> <#lesson-06> <#lesson-10> <#lesson-11> <#lesson-12>

### Query (Data)

TODO: Query (data) check. Queries are not restricted to SQL A query is a request for data or information from a [table](#tabular) or combination of tables. This data may be generated as results returned by Structured Query Language ([SQL](#sql)) or as pictorials, graphs or complex results, e.g., trend analyses from data-mining tools. <#lesson-07> <#lesson-10> <#lesson-11>

[^]: What is the definition of a query in a database …. https://answerthirst.com/what-is-the-definition-of-a-query-in-a-database/

### 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. [64](#fn64) <#processing> <#lesson-07> <#lesson-10> <#lesson-11> <#lesson-12>

## Lesson 11

### Cloud Storage

Cloud storage is a [cloud](#cloud) computing model that stores data on the Internet through a cloud computing provider who manages and operates data storage as a service. It’s delivered on demand with just-in-time capacity and costs, and eliminates buying and managing your own data storage infrastructure. This gives you agility, global scale and durability, with “anytime, anywhere” data access. [^] <#storage> <#lesson-11>

[^]: “What is Cloud Storage? | AWS” . [Online]. Available: https://aws.amazon.com/what-is-cloud-storage/

### DATA INTEGRATION

The process of combining information from different data sources in preparation for data processing. [24](#fn24) <#processing> <#lesson-06> <#lesson-10> <#lesson-11> <#lesson-12>

### Data Lake

A data lake is a storage repository that holds a vast amount of raw data in its native format until it is needed for analytics applications. While a traditional [data warehouse](#data-warehouse) stores data in hierarchical dimensions and tables, a data lake uses a flat architecture to store data, primarily in files or object storage. [^] <#storage> <#lesson-11>

[^]: What Is a Data Lake? Definition from SearchDataManagement. https://searchdatamanagement.techtarget.com/definition/data-lake

### Data Model

A data model (or datamodel) is an abstract model that organizes elements of data and standardizes how they relate to one another and to attributes real-world entities. [^] <#storage> <#lesson-11>

### Data Warehouse

In computing, a data warehouse, also known as an enterprise data warehouse, is a system used for reporting and data analysis and is considered a core component of business intelligence. DWs are central repositories of integrated data from one or more disparate sources. [^] <#storage> <#lesson-11>

[^]: Data Warehouse Interview Questions | TestDome. https://www.testdome.com/d/data-warehouse-interview-questions/1203

### 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.[35](#fn35) <#storage> <#lesson-10> <#lesson-11>

### Database 2

A database is an organized collection of [structured](#structured) information, or data typically stored electronically in a computer system. <#storage> <#lesson-10> <#lesson-11>

[^]: https://www.oracle.com/database/what-is-database/

### Database Model

A database model is a type of [data model](#data-model) that determines its logical structure. It fundamentally determines how data can be stored, organized and manipulated. The most popular database model is the relational model, which uses a table-based format. [^] <#storage> <#lesson-11>

[^]: Database model - Wikipedia. https://en.wikipedia.org/wiki/Database\_model

### Entity–Relationship Model

An entity–relationship model (or ER model) describes interrelated things of interest in a specific domain of knowledge. A basic ER model is composed of entity types (which classify the 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. [^] <#storage> <#lesson-11>

[^]: Entity–relationship model wiki | TheReaderWiki. https://thereaderwiki.com/en/Entity\_relationship\_diagram

### 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. [51](#fn51) <#data-types> <#lesson-03> <#lesson-11>

### Query (Data)

TODO: Query (data) check. Queries are not restricted to SQL A query is a request for data or information from a [table](#tabular) or combination of tables. This data may be generated as results returned by Structured Query Language ([SQL](#sql)) or as pictorials, graphs or complex results, e.g., trend analyses from data-mining tools. <#lesson-07> <#lesson-10> <#lesson-11>

[^]: What is the definition of a query in a database …. https://answerthirst.com/what-is-the-definition-of-a-query-in-a-database/

### Schema (Database)

The [database](#database) schema is its structure described in a formal language supported by the database management system (DBMS). The term “schema” refers to the organization of data as a blueprint of how the database is constructed. [^] <#storage> <#lesson-11>

[^]: Database schema - CodeDocs. https://codedocs.org/what-is/database-schema

### 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. [64](#fn64) <#processing> <#lesson-07> <#lesson-10> <#lesson-11> <#lesson-12>

### 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. [51](#fn51) <#data-types> <#lesson-03> <#lesson-11>

### DATA INTEGRATION

The process of combining information from different data sources in preparation for data processing. [24](#fn24) <#processing> <#lesson-06> <#lesson-10> <#lesson-11> <#lesson-12>

### Query (Data)

TODO: Query (data) check. Queries are not restricted to SQL A query is a request for data or information from a [table](#tabular) or combination of tables. This data may be generated as results returned by Structured Query Language ([SQL](#sql)) or as pictorials, graphs or complex results, e.g., trend analyses from data-mining tools. <#lesson-07> <#lesson-10> <#lesson-11>

[^]: What is the definition of a query in a database …. https://answerthirst.com/what-is-the-definition-of-a-query-in-a-database/

### 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. [64](#fn64) <#processing> <#lesson-07> <#lesson-10> <#lesson-11> <#lesson-12>

### DATA INTEGRATION

The process of combining information from different data sources in preparation for data processing. [24](#fn24) <#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.[35](#fn35) <#storage> <#lesson-10> <#lesson-11>

### Database 2

A database is an organized collection of [structured](#structured) information, or data typically stored electronically in a computer system. <#storage> <#lesson-10> <#lesson-11>

[^]: https://www.oracle.com/database/what-is-database/

### Query (Data)

TODO: Query (data) check. Queries are not restricted to SQL A query is a request for data or information from a [table](#tabular) or combination of tables. This data may be generated as results returned by Structured Query Language ([SQL](#sql)) or as pictorials, graphs or complex results, e.g., trend analyses from data-mining tools. <#lesson-07> <#lesson-10> <#lesson-11>

[^]: What is the definition of a query in a database …. https://answerthirst.com/what-is-the-definition-of-a-query-in-a-database/

### 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. [64](#fn64) <#processing> <#lesson-07> <#lesson-10> <#lesson-11> <#lesson-12>

### DATA INTEGRATION

The process of combining information from different data sources in preparation for data processing. [24](#fn24) <#processing> <#lesson-06> <#lesson-10> <#lesson-11> <#lesson-12>

### Query (Data)

TODO: Query (data) check. Queries are not restricted to SQL A query is a request for data or information from a [table](#tabular) or combination of tables. This data may be generated as results returned by Structured Query Language ([SQL](#sql)) or as pictorials, graphs or complex results, e.g., trend analyses from data-mining tools. <#lesson-07> <#lesson-10> <#lesson-11>

[^]: What is the definition of a query in a database …. https://answerthirst.com/what-is-the-definition-of-a-query-in-a-database/

### 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. [64](#fn64) <#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. [24](#fn24) <#processing> <#lesson-06> <#lesson-10> <#lesson-11> <#lesson-12>

### DATA TRANSFORMATION

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

### Foreign key

TODO: Foreign key, skip? …not restricted to [SQL](#sql)… <#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. [48](#fn48) <#query> <#operations> <#lesson-06> <#lesson-08> <#lesson-12>

### Primary key

In the [relational model](#database-model) of [databases](#database), a primary key is a specific choice of a minimal set of attributes that uniquely specify a tuple in a relation. Informally, a primary key is “which attributes uniquely identify a record,” and in simple cases constitute a single attribute: a unique ID. [^] <#query> <#lesson-12>

[^]: Models - Micron. https://kimtoo.gitbook.io/workspace/models

### 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. [64](#fn64) <#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>

### DATA TRANSFORMATION

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

### DATA INTEGRATION

The process of combining information from different data sources in preparation for data processing. [24](#fn24) <#processing> <#lesson-06> <#lesson-10> <#lesson-11> <#lesson-12>

### DATA TRANSFORMATION

The process of converting data from [one form into another](#data-representation) to gain better [insight](#insight). For instance, converting [unstructured data](#unstructured-data) into [structured form](#structured-data) that can be analyzed further. <#processing> <#operations> <#lesson-03> <#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. [48](#fn48) <#query> <#operations> <#lesson-06> <#lesson-08> <#lesson-12>

### DATA TRANSFORMATION

The process of converting data from [one form into another](#data-representation) to gain better [insight](#insight). For instance, converting [unstructured data](#unstructured-data) into [structured form](#structured-data) that can be analyzed further. <#processing> <#operations> <#lesson-03> <#lesson-03> <#lesson-06> <#lesson-08> <#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. [64](#fn64) <#processing> <#lesson-07> <#lesson-10> <#lesson-11> <#lesson-12>

### DATA TRANSFORMATION

The process of converting data from [one form into another](#data-representation) to gain better [insight](#insight). For instance, converting [unstructured data](#unstructured-data) into [structured form](#structured-data) that can be analyzed further. <#processing> <#operations> <#lesson-03> <#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. [48](#fn48) <#query> <#operations> <#lesson-06> <#lesson-08> <#lesson-12>

### DATA TRANSFORMATION

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

### DATA TRANSFORMATION

The process of converting data from [one form into another](#data-representation) to gain better [insight](#insight). For instance, converting [unstructured data](#unstructured-data) into [structured form](#structured-data) that can be analyzed further. <#processing> <#operations> <#lesson-03> <#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. [48](#fn48) <#query> <#operations> <#lesson-06> <#lesson-08> <#lesson-12>

### DATA TRANSFORMATION

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

### DATA INTEGRATION

The process of combining information from different data sources in preparation for data processing. [24](#fn24) <#processing> <#lesson-06> <#lesson-10> <#lesson-11> <#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. [64](#fn64) <#processing> <#lesson-07> <#lesson-10> <#lesson-11> <#lesson-12>

### DATA INTEGRATION

The process of combining information from different data sources in preparation for data processing. [24](#fn24) <#processing> <#lesson-06> <#lesson-10> <#lesson-11> <#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. [64](#fn64) <#processing> <#lesson-07> <#lesson-10> <#lesson-11> <#lesson-12>

### DATA INTEGRATION

The process of combining information from different data sources in preparation for data processing. [24](#fn24) <#processing> <#lesson-06> <#lesson-10> <#lesson-11> <#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. [64](#fn64) <#processing> <#lesson-07> <#lesson-10> <#lesson-11> <#lesson-12>

### DATA INTEGRATION

The process of combining information from different data sources in preparation for data processing. [24](#fn24) <#processing> <#lesson-06> <#lesson-10> <#lesson-11> <#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. [64](#fn64) <#processing> <#lesson-07> <#lesson-10> <#lesson-11> <#lesson-12>

### DATA INTEGRATION

The process of combining information from different data sources in preparation for data processing. [24](#fn24) <#processing> <#lesson-06> <#lesson-10> <#lesson-11> <#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. [64](#fn64) <#processing> <#lesson-07> <#lesson-10> <#lesson-11> <#lesson-12>

### DATA INTEGRATION

The process of combining information from different data sources in preparation for data processing. [24](#fn24) <#processing> <#lesson-06> <#lesson-10> <#lesson-11> <#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. [64](#fn64) <#processing> <#lesson-07> <#lesson-10> <#lesson-11> <#lesson-12>

## Lesson 13

### Bar Chart

TODO: Bar Chart ….vertical, horizontal, grouped, stacked. [visualization](#visualization) …. <#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. [25](#fn25) <#disciplines> <#lesson-13>

### Histogram

A histogram is a graphical representation that organizes a group of data points into user-specified ranges. Similar in appearance to a bar graph, the histogram condenses a data series into an easily interpreted <#visualization> by taking many data points and grouping them into logical, ordered ranges or bins. <#data-analysis> <#lesson-05> <#lesson-09> <#lesson-13> <#lesson-14>

[^]: What is Histogram in Statistics? | Line Served. https://lineserved.com/new/what-is-histogram-in-statistics/

### Infographics

TODO: Infographics … <#visualization> <#lesson-13>

### Line Plot

TODO: Line Plot …. [visualization](#visualization) …. <#lesson-09> <#lesson-13> <#lesson-14>

### Pie Chart

TODO: Pie Chart …. [visualization](#visualization) …. <#lesson-09> <#lesson-13> <#lesson-14>

### Scatter Plot

TODO: Scatter Plot …. [visualization](#visualization) …. <#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). [65](#fn65) <#disciplines> <#lesson-13>

### Time Series

A sequence of values associated with successive times.

TODO: Time series *(better connected with Video 8(?), but first mentioned in 13)* <#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. [76](#fn76),[77](#fn77) <#data-analysis> <#lesson-05> <#lesson-09> <#lesson-13>

### Visualization

Sometimes referred to as *Viz*. The technique for creating images, diagrams, or animations to communicate messages. *Explanatory*, or informative, visualizations are used to communicate to an audience. *Exploratory* visuals are used to investigate and uncover unknown patterns in data. <#data-analysis> <#lesson-05> <#lesson-09> <#lesson-13>

### Histogram

A histogram is a graphical representation that organizes a group of data points into user-specified ranges. Similar in appearance to a bar graph, the histogram condenses a data series into an easily interpreted <#visualization> by taking many data points and grouping them into logical, ordered ranges or bins. <#data-analysis> <#lesson-05> <#lesson-09> <#lesson-13> <#lesson-14>

[^]: What is Histogram in Statistics? | Line Served. https://lineserved.com/new/what-is-histogram-in-statistics/

### 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. [76](#fn76),[77](#fn77) <#data-analysis> <#lesson-05> <#lesson-09> <#lesson-13>

### Visualization

Sometimes referred to as *Viz*. The technique for creating images, diagrams, or animations to communicate messages. *Explanatory*, or informative, visualizations are used to communicate to an audience. *Exploratory* visuals are used to investigate and uncover unknown patterns in data. <#data-analysis> <#lesson-05> <#lesson-09> <#lesson-13>

### Bar Chart

TODO: Bar Chart ….vertical, horizontal, grouped, stacked. [visualization](#visualization) …. <#lesson-09> <#lesson-13> <#lesson-14>

### Histogram

A histogram is a graphical representation that organizes a group of data points into user-specified ranges. Similar in appearance to a bar graph, the histogram condenses a data series into an easily interpreted <#visualization> by taking many data points and grouping them into logical, ordered ranges or bins. <#data-analysis> <#lesson-05> <#lesson-09> <#lesson-13> <#lesson-14>

[^]: What is Histogram in Statistics? | Line Served. https://lineserved.com/new/what-is-histogram-in-statistics/

### Line Plot

TODO: Line Plot …. [visualization](#visualization) …. <#lesson-09> <#lesson-13> <#lesson-14>

### Pie Chart

TODO: Pie Chart …. [visualization](#visualization) …. <#lesson-09> <#lesson-13> <#lesson-14>

### Scatter Plot

TODO: Scatter Plot …. [visualization](#visualization) …. <#lesson-09> <#lesson-13> <#lesson-14>

### 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. [76](#fn76),[77](#fn77) <#data-analysis> <#lesson-05> <#lesson-09> <#lesson-13>

### Visualization

Sometimes referred to as *Viz*. The technique for creating images, diagrams, or animations to communicate messages. *Explanatory*, or informative, visualizations are used to communicate to an audience. *Exploratory* visuals are used to investigate and uncover unknown patterns in data. <#data-analysis> <#lesson-05> <#lesson-09> <#lesson-13>

### Histogram

A histogram is a graphical representation that organizes a group of data points into user-specified ranges. Similar in appearance to a bar graph, the histogram condenses a data series into an easily interpreted <#visualization> by taking many data points and grouping them into logical, ordered ranges or bins. <#data-analysis> <#lesson-05> <#lesson-09> <#lesson-13> <#lesson-14>

[^]: What is Histogram in Statistics? | Line Served. https://lineserved.com/new/what-is-histogram-in-statistics/

### 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. [76](#fn76),[77](#fn77) <#data-analysis> <#lesson-05> <#lesson-09> <#lesson-13>

### Visualization

Sometimes referred to as *Viz*. The technique for creating images, diagrams, or animations to communicate messages. *Explanatory*, or informative, visualizations are used to communicate to an audience. *Exploratory* visuals are used to investigate and uncover unknown patterns in data. <#data-analysis> <#lesson-05> <#lesson-09> <#lesson-13>

## Lesson 14

### Bar Chart

TODO: Bar Chart ….vertical, horizontal, grouped, stacked. [visualization](#visualization) …. <#lesson-09> <#lesson-13> <#lesson-14>

### Chartjunk

Elements of a graphics 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. Alsa see [data-to-ink-ratio](#data-to-ink-ratio) and [lie-factor](#lie-factor). [13](#fn13) <#lesson-14>

### Colour Map

TODO: Colour … Synonyms: Color map, Color palette, Color scheme. Color(Hue), Color(Shade) [visualization](#visualization) …. <#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). [34](#fn34) <#lesson-14>

### Histogram

A histogram is a graphical representation that organizes a group of data points into user-specified ranges. Similar in appearance to a bar graph, the histogram condenses a data series into an easily interpreted <#visualization> by taking many data points and grouping them into logical, ordered ranges or bins. <#data-analysis> <#lesson-05> <#lesson-09> <#lesson-13> <#lesson-14>

[^]: What is Histogram in Statistics? | Line Served. https://lineserved.com/new/what-is-histogram-in-statistics/

### 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). [50](#fn50) <#lesson-14>

### Line Plot

TODO: Line Plot …. [visualization](#visualization) …. <#lesson-09> <#lesson-13> <#lesson-14>

### Pictographs Plot

TODO: Pictographs Plot …. [visualization](#visualization) …. <#lesson-14>

### Pie Chart

TODO: Pie Chart …. [visualization](#visualization) …. <#lesson-09> <#lesson-13> <#lesson-14>

### Scatter Plot

TODO: Scatter Plot …. [visualization](#visualization) …. <#lesson-09> <#lesson-13> <#lesson-14>

### Histogram

A histogram is a graphical representation that organizes a group of data points into user-specified ranges. Similar in appearance to a bar graph, the histogram condenses a data series into an easily interpreted <#visualization> by taking many data points and grouping them into logical, ordered ranges or bins. <#data-analysis> <#lesson-05> <#lesson-09> <#lesson-13> <#lesson-14>

[^]: What is Histogram in Statistics? | Line Served. https://lineserved.com/new/what-is-histogram-in-statistics/

### Bar Chart

TODO: Bar Chart ….vertical, horizontal, grouped, stacked. [visualization](#visualization) …. <#lesson-09> <#lesson-13> <#lesson-14>

### Histogram

A histogram is a graphical representation that organizes a group of data points into user-specified ranges. Similar in appearance to a bar graph, the histogram condenses a data series into an easily interpreted <#visualization> by taking many data points and grouping them into logical, ordered ranges or bins. <#data-analysis> <#lesson-05> <#lesson-09> <#lesson-13> <#lesson-14>

[^]: What is Histogram in Statistics? | Line Served. https://lineserved.com/new/what-is-histogram-in-statistics/

### Line Plot

TODO: Line Plot …. [visualization](#visualization) …. <#lesson-09> <#lesson-13> <#lesson-14>

### Pie Chart

TODO: Pie Chart …. [visualization](#visualization) …. <#lesson-09> <#lesson-13> <#lesson-14>

### Scatter Plot

TODO: Scatter Plot …. [visualization](#visualization) …. <#lesson-09> <#lesson-13> <#lesson-14>

### Histogram

A histogram is a graphical representation that organizes a group of data points into user-specified ranges. Similar in appearance to a bar graph, the histogram condenses a data series into an easily interpreted <#visualization> by taking many data points and grouping them into logical, ordered ranges or bins. <#data-analysis> <#lesson-05> <#lesson-09> <#lesson-13> <#lesson-14>

[^]: What is Histogram in Statistics? | Line Served. https://lineserved.com/new/what-is-histogram-in-statistics/

### Bar Chart

TODO: Bar Chart ….vertical, horizontal, grouped, stacked. [visualization](#visualization) …. <#lesson-09> <#lesson-13> <#lesson-14>

### Histogram

A histogram is a graphical representation that organizes a group of data points into user-specified ranges. Similar in appearance to a bar graph, the histogram condenses a data series into an easily interpreted <#visualization> by taking many data points and grouping them into logical, ordered ranges or bins. <#data-analysis> <#lesson-05> <#lesson-09> <#lesson-13> <#lesson-14>

[^]: What is Histogram in Statistics? | Line Served. https://lineserved.com/new/what-is-histogram-in-statistics/

### Line Plot

TODO: Line Plot …. [visualization](#visualization) …. <#lesson-09> <#lesson-13> <#lesson-14>

### Pie Chart

TODO: Pie Chart …. [visualization](#visualization) …. <#lesson-09> <#lesson-13> <#lesson-14>

### Scatter Plot

TODO: Scatter Plot …. [visualization](#visualization) …. <#lesson-09> <#lesson-13> <#lesson-14>

### Histogram

A histogram is a graphical representation that organizes a group of data points into user-specified ranges. Similar in appearance to a bar graph, the histogram condenses a data series into an easily interpreted <#visualization> by taking many data points and grouping them into logical, ordered ranges or bins. <#data-analysis> <#lesson-05> <#lesson-09> <#lesson-13> <#lesson-14>

[^]: What is Histogram in Statistics? | Line Served. https://lineserved.com/new/what-is-histogram-in-statistics/

### Bar Chart

TODO: Bar Chart ….vertical, horizontal, grouped, stacked. [visualization](#visualization) …. <#lesson-09> <#lesson-13> <#lesson-14>

### Histogram

A histogram is a graphical representation that organizes a group of data points into user-specified ranges. Similar in appearance to a bar graph, the histogram condenses a data series into an easily interpreted <#visualization> by taking many data points and grouping them into logical, ordered ranges or bins. <#data-analysis> <#lesson-05> <#lesson-09> <#lesson-13> <#lesson-14>

[^]: What is Histogram in Statistics? | Line Served. https://lineserved.com/new/what-is-histogram-in-statistics/

### Line Plot

TODO: Line Plot …. [visualization](#visualization) …. <#lesson-09> <#lesson-13> <#lesson-14>

### Pie Chart

TODO: Pie Chart …. [visualization](#visualization) …. <#lesson-09> <#lesson-13> <#lesson-14>

### Scatter Plot

TODO: Scatter Plot …. [visualization](#visualization) …. <#lesson-09> <#lesson-13> <#lesson-14>

### Histogram

A histogram is a graphical representation that organizes a group of data points into user-specified ranges. Similar in appearance to a bar graph, the histogram condenses a data series into an easily interpreted <#visualization> by taking many data points and grouping them into logical, ordered ranges or bins. <#data-analysis> <#lesson-05> <#lesson-09> <#lesson-13> <#lesson-14>

[^]: What is Histogram in Statistics? | Line Served. https://lineserved.com/new/what-is-histogram-in-statistics/

## 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. “Aggregations”. Oxford University Press. [Online]. Available: https://www.lexico.com/definition/aggregation[↩︎](#fnref3)
4. “A beginner’s guide to big data terminology.” [Online]. Available: <http://dataconomy.com/2016/05/a-beginners-guide-to-big-data-terminology>[↩︎](#fnref4)
5. “Mini-glossary: Big data terms you should know.” [Online]. Available: <http://www.techrepublic.com/article/mini-glossary-big-data-terms-you-should-know>[↩︎](#fnref5)
6. “A beginner’s guide to big data terminology.” [Online]. Available: <http://dataconomy.com/2016/05/a-beginners-guide-to-big-data-terminology>[↩︎](#fnref6)
7. “Mini-glossary: Big data terms you should know.” [Online]. Available: <http://www.techrepublic.com/article/mini-glossary-big-data-terms-you-should-know>[↩︎](#fnref7)
8. “Artificial Intelligence” [Online]. Available: https://www.techopedia.com/definition/190/artificial-intelligence-ai[↩︎](#fnref8)
9. “Behavioral analytics, definition by Techopedia.” [Online]. Available: <https://www.techopedia.com/definition/30308/behavioral-analytics>[↩︎](#fnref9)
10. “Detecting advanced threats with user behavior analytics.” [Online]. Available: <http://www.networkworld.com/article/2904356/security0/detecting-advanced-threats-with-user-behavior-analytics.html>[↩︎](#fnref10)
11. “Big data definition by Tech Target.” [Online]. Available: <http://searchcloudcomputing.techtarget.com/definition/big-data-Big-Data>[↩︎](#fnref11)
12. “What Is Big Data?” [Online]. Available: <https://www.forbes.com/sites/lisaarthur/2013/08/15/what-is-big-data/#4fb6509b5c85>[↩︎](#fnref12)
13. “What is GIS?”. [Online]. Available: https://www.esri.com/en-us/what-is-gis/overview[↩︎](#fnref13)
14. “Who Coined ‘Cloud Computing’?” [Online]. Available: <https://www.technologyreview.com/s/425970/who-coined-cloud-computing>[↩︎](#fnref14)
15. “Cloud computing: a simple introduction.” [Online]. Available: <http://www.explainthatstuff.com/cloud-computing-introduction.html>[↩︎](#fnref15)
16. Donoho, D. (2017). 50 years of data science. *Journal of Computational and Graphical Statistics*, *26*(4), 745-766.[↩︎](#fnref16)
17. “Introduction to Correlation” [Online]. Available: https://www.datascience.com/blog/introduction-to-correlation-learn-data-science-tutorials[↩︎](#fnref17)
18. “Critical Thinking”. Oxford University Press. [Online]. Available: https://www.lexico.com/definition/critical\_thinking[↩︎](#fnref18)
19. “Critical Thinking”. Oxford University Press. [Online]. Available: https://www.oxfordlearnersdictionaries.com/definition/english/critical-thinking[↩︎](#fnref19)
20. “Critical Thinking”. Oxford University Press. [Online]. Available: https://www.lexico.com/definition/critical\_thinking[↩︎](#fnref20)
21. “Critical Thinking”. Oxford University Press. [Online]. Available: https://www.oxfordlearnersdictionaries.com/definition/english/critical-thinking[↩︎](#fnref21)
22. “EI-ISAC Cybersecurity Spotlight – CIA Triad” [Online]. Available: https://www.cisecurity.org/spotlight/ei-isac-cybersecurity-spotlight-cia-triad/[↩︎](#fnref22)
23. “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[↩︎](#fnref23)
24. “Mini-glossary: Big data terms you should know.” [Online]. Available: <http://www.techrepublic.com/article/mini-glossary-big-data-terms-you-should-know>[↩︎](#fnref24)
25. Bradshaw, P. (2012). What is Data Journalism? In *The Data Journalism Handbook*. European Journalism Centre. Available: https://datajournalism.com/read/handbook/one[↩︎](#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. “Internet Privacy” [Online]. Available: https://www.techopedia.com/definition/24954/internet-privacy[↩︎](#fnref29)
30. “Data, statistical information and statistics: Data quality” [Online]. Available: https://www150.statcan.gc.ca/n1/edu/power-pouvoir/ch3/quality-qualite/5214788-eng.htm.[↩︎](#fnref30)
31. Donoho, D. (2017). 50 years of data science. *Journal of Computational and Graphical Statistics*, *26*(4), 745-766.[↩︎](#fnref31)
32. Donoho, D. (2017). 50 years of data science. *Journal of Computational and Graphical Statistics*, *26*(4), 745-766.[↩︎](#fnref32)
33. “The Data Science Venn Diagram” [Online]. Available: <http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram>[↩︎](#fnref33)
34. “What is GIS?”. [Online]. Available: https://www.esri.com/en-us/what-is-gis/overview[↩︎](#fnref34)
35. “A beginner’s guide to big data terminology.” [Online]. Available: <http://dataconomy.com/2016/05/a-beginners-guide-to-big-data-terminology>[↩︎](#fnref35)
36. “Descriptive analytics definition by Tech Target.” [Online]. Available: <http://whatis.techtarget.com/definition/descriptive-analytics>[↩︎](#fnref36)
37. “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>[↩︎](#fnref37)
38. “Distributed file system definition by Tech Target.” [Online]. Available: <http://searchwindowsserver.techtarget.com/definition/distributed-file-system-DFS>[↩︎](#fnref38)
39. “Distributed file system definition by Techopedia.” [Online]. Available: <https://www.techopedia.com/definition/1825/distributed-file-system-dfs>[↩︎](#fnref39)
40. “Data Modeling Introduction” [Online]. Available: <https://docs.mongodb.com/manual/core/data-modeling-introduction/>[↩︎](#fnref40)
41. Donoho, D. (2017). 50 years of data science. *Journal of Computational and Graphical Statistics*, *26*(4), 745-766.[↩︎](#fnref41)
42. “What is GIS?”. [Online]. Available: https://www.esri.com/en-us/what-is-gis/overview[↩︎](#fnref42)
43. “What is GIS?”. [Online]. Available: https://www.esri.com/en-us/what-is-gis/overview[↩︎](#fnref43)
44. “Data vs. Information vs. Insight” [Online]. Available: https://online.ben.edu/programs/mba/resources/data-vs-information-vs-insight[↩︎](#fnref44)
45. “Data vs. Information vs. Insight” [Online]. Available: https://online.ben.edu/programs/mba/resources/data-vs-information-vs-insight[↩︎](#fnref45)
46. “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>[↩︎](#fnref46)
47. “Internet of things definition by Tech Target.” [Online]. Available: <http://internetofthingsagenda.techtarget.com/definition/Internet-of-Things-IoT>[↩︎](#fnref47)
48. “SQL Joins – W3 Schools”. [Online]. Available: https://www.w3schools.com/sql/sql\_join.asp[↩︎](#fnref48)
49. “Overview of the KDD Process” [Online]. Available: http://www2.cs.uregina.ca/~dbd/cs831/notes/kdd/1\_kdd.html[↩︎](#fnref49)
50. “What is GIS?”. [Online]. Available: https://www.esri.com/en-us/what-is-gis/overview[↩︎](#fnref50)
51. Greenberg, J. (2003). Metadata and the world wide web. *Encyclopedia of library and information science*, *3*, 1876-1888.[↩︎](#fnref51)
52. 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[↩︎](#fnref52)
53. “Computer network definition by Techopedia.” [Online]. Available: <https://www.techopedia.com/definition/25597/computer-network>[↩︎](#fnref53)
54. “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>[↩︎](#fnref54)
55. “Social Network Analysis: An Introduction by Orgnet,LLC.” [Online]. Available: <http://www.orgnet.com/sna.html>[↩︎](#fnref55)
56. “Null (SQL)”. [Online]. Wikipedia, The Free Encyclopedia. Available: https://en.wikipedia.org/w/index.php?title=Null\_(SQL)[↩︎](#fnref56)
57. “Population”. Oxford University Press. [Online]. Available: https://www.lexico.com/definition/population[↩︎](#fnref57)
58. “A beginner’s guide to big data terminology.” [Online]. Available: <http://dataconomy.com/2016/05/a-beginners-guide-to-big-data-terminology>[↩︎](#fnref58)
59. Donoho, D. (2017). 50 years of data science. *Journal of Computational and Graphical Statistics*, *26*(4), 745-766.[↩︎](#fnref59)
60. “A beginner’s guide to big data terminology.” [Online]. Available: <http://dataconomy.com/2016/05/a-beginners-guide-to-big-data-terminology>[↩︎](#fnref60)
61. “Statistics: Power from Data! Glossary” [Online]. Available: https://www150.statcan.gc.ca/n1/edu/power-pouvoir/glossary-glossaire/5214842-eng.htm#Sample[↩︎](#fnref61)
62. “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>[↩︎](#fnref62)
63. “Mini-glossary: Big data terms you should know.” [Online]. Available: <http://www.techrepublic.com/article/mini-glossary-big-data-terms-you-should-know>[↩︎](#fnref63)
64. Donoho, D. (2017). 50 years of data science. *Journal of Computational and Graphical Statistics*, *26*(4), 745-766.[↩︎](#fnref64)
65. “Data Storytelling: How to Effectively Tell a Story with Data”. [Blog].Harvard Business School. Available: https://online.hbs.edu/blog/post/data-storytelling[↩︎](#fnref65)
66. “Parse”. Oxford University Press. [Online]. Available: https://www.lexico.com/definition/parse[↩︎](#fnref66)
67. “Introduction to Structured Data, by Google.” [Online]. Available: <https://developers.google.com/search/docs/guides/intro-structured-data>[↩︎](#fnref67)
68. “Structured data definition by Webopedia.” [Online]. Available: <http://www.webopedia.com/TERM/S/structured_data.html>[↩︎](#fnref68)
69. “Introduction to Tabular Data” [Online]. Available: <https://papl.cs.brown.edu/2016/intro-tabular-data.html>[↩︎](#fnref69)
70. “Tabular Database” [Online]. Available: <https://www.techopedia.com/definition/26181/tabular-database>[↩︎](#fnref70)
71. “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[↩︎](#fnref71)
72. “Tidy Data”. [Online]. R for Data Science. Available: https://r4ds.had.co.nz/tidy-data.html[↩︎](#fnref72)
73. “Tree” from the National Institute of Standards and Technology. [Online]. Available: https://xlinux.nist.gov/dads/HTML/tree.html[↩︎](#fnref73)
74. “Unstructured data definition by Techopedia.” [Online]. Available: <https://www.techopedia.com/definition/13865/unstructured-data>[↩︎](#fnref74)
75. “Solving the Unstructured Data Challenge.” [Online]. Available: <http://www.cio.com/article/2941015/big-data/solving-the-unstructured-data-challenge.html>[↩︎](#fnref75)
76. “Data visualization definition by Tech Target.” [Online]. Available: <http://searchbusinessanalytics.techtarget.com/definition/data-visualization>[↩︎](#fnref76)
77. “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>[↩︎](#fnref77)