Data Literacy: Glossary and Index

Table of Contents

[Glossary 4](#_Toc96313745)

[ADVANCED RESEARCH COMPUTING 4](#_Toc96313746)

[Aggregations 4](#_Toc96313747)

[ALGORITHM 4](#_Toc96313748)

[ANALYTICS 4](#_Toc96313749)

[API 4](#_Toc96313750)

[Arithmetic operations 4](#_Toc96313751)

[ARTIFICIAL INTELLIGENCE 4](#_Toc96313752)

[Bar Chart 5](#_Toc96313753)

[BEHAVIOUR(AL) ANALYTICS 5](#_Toc96313754)

[BIG DATA 5](#_Toc96313755)

[Big data isn’t always big. 5](#_Toc96313756)

[Binning 5](#_Toc96313757)

[Box Plot 5](#_Toc96313758)

[Bucketing 5](#_Toc96313759)

[CAUSAL INFERENCE 5](#_Toc96313760)

[CLASSIFICATION 5](#_Toc96313761)

[Cloud (“The Cloud”) 6](#_Toc96313762)

[CLOUD COMPUTING 6](#_Toc96313763)

[Cloud Storage 6](#_Toc96313764)

[CLUSTER (Compute) 6](#_Toc96313765)

[CLUSTERING 6](#_Toc96313766)

[COMPUTING WITH DATA 6](#_Toc96313767)

[Correlation 6](#_Toc96313768)

[CORRELATION MINING 6](#_Toc96313769)

[Critical Thinking 7](#_Toc96313770)

[Crowdsourcing 7](#_Toc96313771)

[CYBERSECURITY 7](#_Toc96313772)

[DATA 7](#_Toc96313773)

[Data (Clean, Dirty) 7](#_Toc96313774)

[DATA AGGREGATION 7](#_Toc96313775)

[Data Analysis [TODO: Seems more like a definition for Data Transformation(?)] 7](#_Toc96313776)

[Data Bias 7](#_Toc96313777)

[DATA CLEANING 7](#_Toc96313778)

[DATA EXPLORATION AND PREPARATION 8](#_Toc96313779)

[DATA INTEGRATION 8](#_Toc96313780)

[Data Journalism 8](#_Toc96313781)

[Data Lake 8](#_Toc96313782)

[Data Literacy 8](#_Toc96313783)

[DATA MINING 8](#_Toc96313784)

[Data Model 8](#_Toc96313785)

[Data Privacy 8](#_Toc96313786)

[Data Quality 9](#_Toc96313787)

[DATA REPRESENTATION 9](#_Toc96313788)

[DATA SCIENCE 9](#_Toc96313789)

[DATA TRANSFORMATION 9](#_Toc96313790)

[Data types 9](#_Toc96313791)

[Data Warehouse 10](#_Toc96313792)

[DATABASE 10](#_Toc96313793)

[Database 2 10](#_Toc96313794)

[Database Model 10](#_Toc96313795)

[Dendrogram 10](#_Toc96313796)

[DESCRIPTIVE ANALYTICS 10](#_Toc96313797)

[DISTRIBUTED FILE SYSTEM 10](#_Toc96313798)

[Entity–Relationship Model 10](#_Toc96313799)

[FEATURE 11](#_Toc96313800)

[FEATURE ENGINEERING 11](#_Toc96313801)

[File Formats 11](#_Toc96313802)

[Filter 11](#_Toc96313803)

[FLEXIBLE SCHEMA 12](#_Toc96313804)

[Foreign key 12](#_Toc96313805)

[GENERATIVE MODELING 12](#_Toc96313806)

[GIS 12](#_Toc96313807)

[GOODNESS OF FIT 12](#_Toc96313808)

[Groupby 12](#_Toc96313809)

[Heatmap 12](#_Toc96313810)

[Histogram 12](#_Toc96313811)

[HTML 13](#_Toc96313812)

[Infographics 13](#_Toc96313813)

[INFORMATION 13](#_Toc96313814)

[INSIGHT 13](#_Toc96313815)

[IOT (INTERNET OF THINGS) 13](#_Toc96313816)

[Join 13](#_Toc96313817)

[KNOWLEDGE DISCOVERY 13](#_Toc96313818)

[Line Plot 13](#_Toc96313819)

[Logical operations 13](#_Toc96313820)

[MACHINE LEARNING 14](#_Toc96313821)

[METADATA 14](#_Toc96313822)

[Model 14](#_Toc96313823)

[MODEL FITTING 14](#_Toc96313824)

[NETWORK 14](#_Toc96313825)

[NETWORK ANALYSIS 14](#_Toc96313826)

[NO-SQL DATABASE 14](#_Toc96313827)

[Null values 14](#_Toc96313828)

[Pair plot 14](#_Toc96313829)

[Pie Chart 14](#_Toc96313830)

[Pivot 14](#_Toc96313831)

[Population 15](#_Toc96313832)

[PREDICTIVE ANALYTICS 15](#_Toc96313833)

[PRESCRIPTIVE ANALYTICS 15](#_Toc96313834)

[Primary key 15](#_Toc96313835)

[Query (Data) 15](#_Toc96313836)

[REGRESSION 15](#_Toc96313837)

[RELATIONSHIP MINING 15](#_Toc96313838)

[Repository (data repo) 15](#_Toc96313839)

[Sample (Representative, Proxy) 15](#_Toc96313840)

[Schema (Database) 16](#_Toc96313841)

[SECURE COMPUTING 16](#_Toc96313842)

[Select 16](#_Toc96313843)

[Skewness 16](#_Toc96313844)

[Sort 16](#_Toc96313845)

[SQL 16](#_Toc96313846)

[Standard Deviation 16](#_Toc96313847)

[Statistics 16](#_Toc96313848)

[Storytelling 16](#_Toc96313849)

[String or character operations 16](#_Toc96313850)

[STRUCTURED DATA 17](#_Toc96313851)

[TABULAR DATA 17](#_Toc96313852)

[TEXT MINING 17](#_Toc96313853)

[Tidy tables 17](#_Toc96313854)

[Time Series 18](#_Toc96313855)

[Transpose 18](#_Toc96313856)

[Tree structure 18](#_Toc96313857)

[Unbalanced Data 18](#_Toc96313858)

[Union 18](#_Toc96313859)

[UNSTRUCTURED DATA 18](#_Toc96313860)

[URL 18](#_Toc96313861)

[Variance 18](#_Toc96313862)

[Violin Plot 18](#_Toc96313863)

[VISUALIZATION 18](#_Toc96313864)

[Visualization 19](#_Toc96313865)

[VISUALIZATION DASHBOARD 19](#_Toc96313866)

[Web scraping 19](#_Toc96313867)

[Word Cloud 19](#_Toc96313868)

[Categories 19](#_Toc96313869)

[Data Inquiry Disciplines 19](#_Toc96313870)

[Data Analysis Techniques 20](#_Toc96313871)

[Data Processing 21](#_Toc96313872)

[Storage and Physical Data Sources 21](#_Toc96313873)

[Types of Data 22](#_Toc96313874)

[Operations 22](#_Toc96313875)

[Lessons Index 23](#_Toc96313876)

[Lessons 01 and 02 23](#_Toc96313877)

[Lesson 03 23](#_Toc96313878)

[Lesson 04 23](#_Toc96313879)

[Lesson 05 24](#_Toc96313880)

[Lesson 06 24](#_Toc96313881)

[Lesson 07 24](#_Toc96313882)

[Lesson 08 25](#_Toc96313883)

[Lesson 09 25](#_Toc96313884)

[Lesson 10 26](#_Toc96313885)

[Lesson 11 26](#_Toc96313886)

[Lesson 12 27](#_Toc96313887)

[Lesson 13 27](#_Toc96313888)

[Lesson 14 27](#_Toc96313889)

[Lesson 15 27](#_Toc96313890)

[References 27](#_Toc96313891)

# 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>

### Aggregations

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

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

### 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 values 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: …. [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>

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

### 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.[15](#fn15) <#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>

### 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. [16](#fn16) <#data-analysis>

### Critical Thinking

The objective analysis and evaluation of an issue to form a judgement [17](#fn17). The process of analysing information in order to make a logical decision about the extent to which you believe something to be true or false [18](#fn18). <#disciplines> <#lesson-04>

### Crowdsourcing

TODO. <#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). [19](#fn19) <#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, Inconsistent, or duplicated.[20](#fn20) <#processing> <#lesson-04> <#lesson-06>

### DATA AGGREGATION

TODO:

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

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

### Data Bias

Whether the (sample) data’s [balance](#balance) in the 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. [21](#fn21) <#processing> <#lesson-06> <#lesson-10> <#lesson-11> <#lesson-12>

### Data Journalism

TODO: <#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%20warehouse) 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 label reserved for data scientists or specialists. Data literacy should be thought of as “the ability of non-specialists to make use of data” and measure “a person’s ability to read, work with, analyze and argue with data”, presumably using simple statistics such as means and percentages [22](#fn22). <#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.[23](#fn23),[24](#fn24) <#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 of 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. [25](#fn25) <#disciplines> <#lesson-04>

### Data Quality

The usefulness and reliability of data are assessed by its quality. High-quality data can be defined as appropriate, consistent, complete and accurate.

TODO: Dimensions that the quality can be assessed on include relevance (does the statistical information matter?), accessibility (can users access the statistical information?), accuracy (is the statistical information representative of the targeted measurement?), timeliness (is the lag between the period of reference and the availability of the statistical information acceptable?), interpretability (is metadata available and complete?), and coherence (is the statistical information consistent over time, between regions and across sub-populations?). [26](#fn26) <#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. [27](#fn27) <#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). [28](#fn28),[29](#fn29) <#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

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

#### Float

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

#### Integer

A whole number.

#### String

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

#### Date / Datetime

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

### 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.[30](#fn30) <#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-11>

[^]: https://www.coursehero.com/file/76896898/databasepdf/

### Database Model

A database model is a type of [data model](#data-model) that determines its logical structure. It fundamentally determines in which manner data can be stored, organized and manipulated. The most popular example of a 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](#X4ecc7f23bff029dea0f81fd81e6de5ccec8a656), it illustrates the arrangement of the clusters produced by the corresponding analyses. <#lesson-09>

### DESCRIPTIVE ANALYTICS

An initial stage of data processing that involves creating a summary of historical data with the goal of answering the question, “What happened?”[31](#fn31),[32](#fn32) <#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.[33](#fn33),[34](#fn34) <#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). For end-users of a database, the ER model is part of the [metadata](#metadata). [^] <#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

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

### Filter

TODO: <#operations> <#lesson-07>

#### TXT

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

#### CSV

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

#### TSV

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

#### JSON

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

#### XML

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

#### YAML

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

### 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.[35](#fn35) <#data-types>

### Foreign key

TODO: <#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. [36](#fn36)

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

### GOODNESS OF FIT

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

### Groupby

TODO: Collect and [aggregations](#aggregation) and [transformations](#data-transformations) summarize…. <#operations> <#lesson-07>

### Heatmap

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

### 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>

### HTML

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

### Infographics

TODO: <#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). [38](#fn38)

### INSIGHT

Actionable [information](#information) gained by interpreting [data analysis](#analytics) results, helping people to make more informed decisions. [39](#fn39)

### 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.[40](#fn40),[41](#fn41) <#storage>

### Join

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

[^]: “SQL Joins – W3 Schools”. [Online]. Available: https://www.w3schools.com/sql/sql\_join.asp

### 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.[42](#fn42) <#disciplines>

### Line Plot

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

### Logical operations

Operations and manipulations according to the rules of logic and boolean algebra. Logical operations evaluate *truth values* (true or false) of a logical statement. Logical statements are constructed with statements including conjunction (*and*), the disjunction (*or*), and the negation (*not*), along 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. [43](#fn43) <#data-types> <#lesson-03> <#lesson-11>

### Model

TODO: <#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.[44](#fn44),[45](#fn45)

### 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. [46](#fn46)

### NO-SQL DATABASE

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

### 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. [47](#fn47) <#lesson-06>

### Pair plot

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

### Pie Chart

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

### 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. [48](#fn48) <#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.[49](#fn49),[50](#fn50)

### PRESCRIPTIVE ANALYTICS

A type of [analytics](#analytics) with the goal of using data to determine the best course of action for a specific scenario. [51](#fn51) <#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: 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>

### REGRESSION

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

### RELATIONSHIP MINING

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

### Repository (data repo)

Sometimes abbreviated as *repo*, it is a synonym for [database](#database). <#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). [52](#fn52) <#data-analysis> <#lesson-04>

### 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. [53](#fn53),[54](#fn54) <#disciplines>

### Select

TODO: <#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 to the right, it is said to be skewed. <#lesson-09>

### Sort

TODO: 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. [28] <#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 it is a measure 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

TODO: <#disciplines> <#lesson-13>

### String or character operations

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

#### Parse

To analyze (a string or text) into logical syntactic components.[55](#fn55)

#### Indexing

Indexing (or selecting) string characters by their position, *zero-th* character, *first* character

#### 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.[56](#fn56),[57](#fn57) <#data-types> <#lesson-03>

### TABULAR DATA

Data items or observations that are organized as rows that contain values under columns that correspond to specific variables or properties. [58](#fn58),[59](#fn59) <#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

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

#### Wide table format

TODO: each holds different attributes in separate columns [pivot](#pivot). <#data-types> <#lesson-03>

### TEXT MINING

A process of analyzing [text](#unstructured-data) to capture key concepts, themes, relationships and trends.[60](#fn60) <#data-analysis>

### Tidy tables

Datasets are commonly messy. Getting data into tidy [61](#fn61) 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

TODO: *(better connected with Video 8(?), but 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. [62](#fn62) <#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 is used to [combine](#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. [63](#fn63),[64](#fn64) <#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 it is a measure of how far a set of numbers is spread out from their average value. <#lesson-05>

### Violin Plot

TODO: …. [visualization](#visualization) …. <#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. [65](#fn65),[66](#fn66) <#data-analysis> [#lesson-05] [#lesson-05] (#lesson-09)(#lesson-09) (#lesson-13)(#lesson-13)

### Visualization

The technique for creating images, diagrams, or animations to communicate a message. Also referred to as *Viz*. <#data-analysis> <#lesson-05> (#lesson-09)(#lesson-09) (#lesson-13)(#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

TODO: Maybe(?)

Web Scraping is when you programmatically extract data from a website

Before Scraping

* Check to see if CSV, JSON, or XML version of an HTML page are available
* Check that you have permission/legality to scrape

If you do scrape

* Be careful to not to overburden the site with your requests
* Test code on small requests
* Save the results of each request, so you don’t have to repeat the request unnecessarily

<#lesson-10>

[^]: CMPT 733

### Word Cloud

TODO: …. [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 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)

[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

[Aggregations](#aggregations)

[Arithmetic operations](#arithmetic-operations)

[Binning](#binning)

[Bucketing](#bucketing)

[DATA TRANSFORMATION](#data-transformation)

[Filter](#filter)

[Groupby](#groupby)

[Join](#join)

[Logical operations](#logical-operations)

[Pivot](#pivot)

[Select](#select)

[Sort](#sort)

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

[Transpose](#transpose)

[Union](#union)

# Lessons Index

## Lessons 01 and 02

[Data Literacy](#data-literacy)

## Lesson 03

[ARTIFICIAL INTELLIGENCE](#artificial-intelligence)

[Crowdsourcing](#crowdsourcing)

[DATA](#data)

[DATA TRANSFORMATION](#data-transformation)

[Data types](#data-types)

[File Formats](#file-formats)

[MACHINE LEARNING](#machine-learning)

[METADATA](#metadata)

[Pivot](#pivot)

[STRUCTURED DATA](#structured-data)

[TABULAR DATA](#tabular-data)

[Tree structure](#tree-structure)

[UNSTRUCTURED DATA](#unstructured-data)

## Lesson 04

[BIG DATA](#big-data)

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

[Critical Thinking](#critical-thinking)

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

[Data Bias](#data-bias)

[Data Privacy](#data-privacy)

[Data Quality](#data-quality)

[Population](#population)

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

[Unbalanced Data](#unbalanced-data)

## Lesson 05

[Correlation](#correlation)

[Histogram](#histogram)

[Standard Deviation](#standard-deviation)

[Statistics](#statistics)

[Variance](#variance)

[Visualization](#visualization-1)

## Lesson 06

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

[DATA INTEGRATION](#data-integration)

[DATA TRANSFORMATION](#data-transformation)

[Join](#join)

[Null values](#null-values)

[Tidy tables](#tidy-tables)

## Lesson 07

[Filter](#filter)

[Groupby](#groupby)

[Query (Data)](#query-data)

[Select](#select)

[Sort](#sort)

[SQL](#sql)

## Lesson 08

[Aggregations](#aggregations)

[Arithmetic operations](#arithmetic-operations)

[Binning](#binning)

[Bucketing](#bucketing)

[DATA TRANSFORMATION](#data-transformation)

[Logical operations](#logical-operations)

[Pivot](#pivot)

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

[Transpose](#transpose)

## Lesson 09

[Bar Chart](#bar-chart)

[Box Plot](#box-plot)

[Correlation](#correlation)

[Data Analysis [TODO: Seems more like a definition for Data Transformation(?)]](#X79060956e1db0d007ea7fa4354ea4b4a78f209c)

[Dendrogram](#dendrogram)

[Heatmap](#heatmap)

[Line Plot](#line-plot)

[Model](#model)

[Pair plot](#pair-plot)

[Pie Chart](#pie-chart)

[Skewness](#skewness)

[Standard Deviation](#standard-deviation)

[Violin Plot](#violin-plot)

[Word Cloud](#word-cloud)

## Lesson 10

[API](#api)

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

[CLUSTER (Compute)](#cluster-compute)

[DATA INTEGRATION](#data-integration)

[DATABASE](#database)

[GIS](#gis)

[HTML](#html)

[Query (Data)](#query-data)

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

[SQL](#sql)

[URL](#url)

[Web scraping](#web-scraping)

## Lesson 11

[Cloud Storage](#cloud-storage)

[DATA INTEGRATION](#data-integration)

[Data Lake](#data-lake)

[Data Model](#data-model)

[Data Warehouse](#data-warehouse)

[DATABASE](#database)

[Database 2](#database-2)

[Database Model](#database-model)

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

[METADATA](#metadata)

[Query (Data)](#query-data)

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

[SQL](#sql)

## Lesson 12

[DATA INTEGRATION](#data-integration)

[DATA TRANSFORMATION](#data-transformation)

[Foreign key](#foreign-key)

[Primary key](#primary-key)

[SQL](#sql)

[Union](#union)

## Lesson 13

[Data Journalism](#data-journalism)

[Infographics](#infographics)

[Storytelling](#storytelling)

[Time Series](#time-series)

## Lesson 14

## Lesson 15

# References

1. “Westgrid: What We Do.” [Online]. Available: <https://www.westgrid.ca/about_westgrid/what_we_do>[↩︎](#fnref1)
2. “Compute Canada. Update January 11, 2017.” [Online]. Available: <https://www.computecanada.ca/wp-content/uploads/2015/02/Update-January-11-2017.pdf>[↩︎](#fnref2)
3. “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. “Who Coined ‘Cloud Computing’?” [Online]. Available: <https://www.technologyreview.com/s/425970/who-coined-cloud-computing>[↩︎](#fnref13)
14. “Cloud computing: a simple introduction.” [Online]. Available: <http://www.explainthatstuff.com/cloud-computing-introduction.html>[↩︎](#fnref14)
15. Donoho, D. (2017). 50 years of data science. *Journal of Computational and Graphical Statistics*, *26*(4), 745-766.[↩︎](#fnref15)
16. “Introduction to Correlation” [Online]. Available: https://www.datascience.com/blog/introduction-to-correlation-learn-data-science-tutorials[↩︎](#fnref16)
17. “Critical Thinking”. Oxford University Press. [Online]. Available: https://www.lexico.com/definition/critical\_thinking[↩︎](#fnref17)
18. “Critical Thinking”. Oxford University Press. [Online]. Available: https://www.oxfordlearnersdictionaries.com/definition/english/critical-thinking[↩︎](#fnref18)
19. “EI-ISAC Cybersecurity Spotlight – CIA Triad” [Online]. Available: https://www.cisecurity.org/spotlight/ei-isac-cybersecurity-spotlight-cia-triad/[↩︎](#fnref19)
20. “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[↩︎](#fnref20)
21. “Mini-glossary: Big data terms you should know.” [Online]. Available: <http://www.techrepublic.com/article/mini-glossary-big-data-terms-you-should-know>[↩︎](#fnref21)
22. “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[↩︎](#fnref22)
23. “A beginner’s guide to big data terminology.” [Online]. Available: <http://dataconomy.com/2016/05/a-beginners-guide-to-big-data-terminology>[↩︎](#fnref23)
24. “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. “Internet Privacy” [Online]. Available: https://www.techopedia.com/definition/24954/internet-privacy[↩︎](#fnref25)
26. “Data, statistical information and statistics: Data quality” [Online]. Available: https://www150.statcan.gc.ca/n1/edu/power-pouvoir/ch3/quality-qualite/5214788-eng.htm.[↩︎](#fnref26)
27. Donoho, D. (2017). 50 years of data science. *Journal of Computational and Graphical Statistics*, *26*(4), 745-766.[↩︎](#fnref27)
28. Donoho, D. (2017). 50 years of data science. *Journal of Computational and Graphical Statistics*, *26*(4), 745-766.[↩︎](#fnref28)
29. “The Data Science Venn Diagram” [Online]. Available: <http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram>[↩︎](#fnref29)
30. “A beginner’s guide to big data terminology.” [Online]. Available: <http://dataconomy.com/2016/05/a-beginners-guide-to-big-data-terminology>[↩︎](#fnref30)
31. “Descriptive analytics definition by Tech Target.” [Online]. Available: <http://whatis.techtarget.com/definition/descriptive-analytics>[↩︎](#fnref31)
32. “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>[↩︎](#fnref32)
33. “Distributed file system definition by Tech Target.” [Online]. Available: <http://searchwindowsserver.techtarget.com/definition/distributed-file-system-DFS>[↩︎](#fnref33)
34. “Distributed file system definition by Techopedia.” [Online]. Available: <https://www.techopedia.com/definition/1825/distributed-file-system-dfs>[↩︎](#fnref34)
35. “Data Modeling Introduction” [Online]. Available: <https://docs.mongodb.com/manual/core/data-modeling-introduction/>[↩︎](#fnref35)
36. Donoho, D. (2017). 50 years of data science. *Journal of Computational and Graphical Statistics*, *26*(4), 745-766.[↩︎](#fnref36)
37. “What is GIS?”. [Online]. Available: https://www.esri.com/en-us/what-is-gis/overview[↩︎](#fnref37)
38. “Data vs. Information vs. Insight” [Online]. Available: https://online.ben.edu/programs/mba/resources/data-vs-information-vs-insight[↩︎](#fnref38)
39. “Data vs. Information vs. Insight” [Online]. Available: https://online.ben.edu/programs/mba/resources/data-vs-information-vs-insight[↩︎](#fnref39)
40. “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>[↩︎](#fnref40)
41. “Internet of things definition by Tech Target.” [Online]. Available: <http://internetofthingsagenda.techtarget.com/definition/Internet-of-Things-IoT>[↩︎](#fnref41)
42. “Overview of the KDD Process” [Online]. Available: http://www2.cs.uregina.ca/~dbd/cs831/notes/kdd/1\_kdd.html[↩︎](#fnref42)
43. Greenberg, J. (2003). Metadata and the world wide web. *Encyclopedia of library and information science*, *3*, 1876-1888.[↩︎](#fnref43)
44. “Computer network definition by Techopedia.” [Online]. Available: <https://www.techopedia.com/definition/25597/computer-network>[↩︎](#fnref44)
45. “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>[↩︎](#fnref45)
46. “Social Network Analysis: An Introduction by Orgnet,LLC.” [Online]. Available: <http://www.orgnet.com/sna.html>[↩︎](#fnref46)
47. “Null (SQL)”. [Online]. Wikipedia, The Free Encyclopedia. Available: https://en.wikipedia.org/w/index.php?title=Null\_(SQL)[↩︎](#fnref47)
48. “Population”. Oxford University Press. [Online]. Available: https://www.lexico.com/definition/population[↩︎](#fnref48)
49. “A beginner’s guide to big data terminology.” [Online]. Available: <http://dataconomy.com/2016/05/a-beginners-guide-to-big-data-terminology>[↩︎](#fnref49)
50. Donoho, D. (2017). 50 years of data science. *Journal of Computational and Graphical Statistics*, *26*(4), 745-766.[↩︎](#fnref50)
51. “A beginner’s guide to big data terminology.” [Online]. Available: <http://dataconomy.com/2016/05/a-beginners-guide-to-big-data-terminology>[↩︎](#fnref51)
52. “Statistics: Power from Data! Glossary” [Online]. Available: https://www150.statcan.gc.ca/n1/edu/power-pouvoir/glossary-glossaire/5214842-eng.htm#Sample[↩︎](#fnref52)
53. “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>[↩︎](#fnref53)
54. “Mini-glossary: Big data terms you should know.” [Online]. Available: <http://www.techrepublic.com/article/mini-glossary-big-data-terms-you-should-know>[↩︎](#fnref54)
55. “Parse”. Oxford University Press. [Online]. Available: https://www.lexico.com/definition/parse[↩︎](#fnref55)
56. “Introduction to Structured Data, by Google.” [Online]. Available: <https://developers.google.com/search/docs/guides/intro-structured-data>[↩︎](#fnref56)
57. “Structured data definition by Webopedia.” [Online]. Available: <http://www.webopedia.com/TERM/S/structured_data.html>[↩︎](#fnref57)
58. “Introduction to Tabular Data” [Online]. Available: <https://papl.cs.brown.edu/2016/intro-tabular-data.html>[↩︎](#fnref58)
59. “Tabular Database” [Online]. Available: <https://www.techopedia.com/definition/26181/tabular-database>[↩︎](#fnref59)
60. “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[↩︎](#fnref60)
61. “Tidy Data”. [Online]. R for Data Science. Available: https://r4ds.had.co.nz/tidy-data.html[↩︎](#fnref61)
62. “Tree” from the National Institute of Standards and Technology. [Online]. Available: https://xlinux.nist.gov/dads/HTML/tree.html[↩︎](#fnref62)
63. “Unstructured data definition by Techopedia.” [Online]. Available: <https://www.techopedia.com/definition/13865/unstructured-data>[↩︎](#fnref63)
64. “Solving the Unstructured Data Challenge.” [Online]. Available: <http://www.cio.com/article/2941015/big-data/solving-the-unstructured-data-challenge.html>[↩︎](#fnref64)
65. “Data visualization definition by Tech Target.” [Online]. Available: <http://searchbusinessanalytics.techtarget.com/definition/data-visualization>[↩︎](#fnref65)
66. “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>[↩︎](#fnref66)