# Glossary

## Data Analytics

---

## Terms and definitions from Course 3

## A

**Access control:** Features such as password protection, user permissions, and encryption that are used to protect a spreadsheet

**Administrative metadata:** Metadata that indicates the technical source of a digital asset

**Agenda**: A list of scheduled appointments

**Analytical thinking:** The process of identifying and defining a problem, then solving it by using data in an organized, step-by-step manner

**Audio file:** Digitized audio storage usually in an MP3, AAC, or other compressed format

## B

**Bad data source:** A data source that is not reliable, original, comprehensive, current, and cited (ROCCC)

**Bias:** A conscious or subconscious preference in favor of or against a person, group of people, or thing

**Boolean data:** A data type with only two possible values, usually true or false

## C

**Confirmation bias:** The tendency to search for or interpret information in a way that confirms pre-existing beliefs

**Consent:** The aspect of data ethics that presumes an individual's right to know how and why their personal data will be used before agreeing to provide it

**Continuous data:** Data that is measured and can have almost any numeric value

**Cookie:** A small file stored on a computer that contains information about its users

**CSV (comma-separated values) file:** A delimited text file that uses a comma to separate values

**Currency:** The aspect of data ethics that presumes individuals should be aware of financial transactions resulting from the use of their personal data and the scale of those transactions

# D

**Data anonymization:** The process of protecting people's private or sensitive data by eliminating identifying information

**Data bias:** When a preference in favor of or against a person, group of people, or thing systematically skews data analysis results in a certain direction

**Data element:** A piece of information in a dataset

**Data ethics:** Well-founded standards of right and wrong that dictate how data is collected, shared, and used

**Data governance:** A process for ensuring the formal management of a company's data assets

**Data interoperability:** The ability to integrate data from multiple sources and a key factor leading to the successful use of open data among companies and governments

**Data model:** A tool for organizing data elements and how they relate to one another

**Data privacy:** Preserving a data subject's information any time a data transaction occurs

**Data security:** Protecting data from unauthorized access or corruption by adopting safety measures

**Data type:** An attribute that describes a piece of data based on its values, its programming language, or the operations it can perform

**Data visualization:** The graphical representation of data

**Descriptive metadata:** Metadata that describes a piece of data and can be used to identify it at a later point in time

**Digital photo:** An electronic or computer-based image usually in BMP or JPG format

**Discrete data:** Data that is counted and has a limited number of values

# E

**Ethics:** Well-founded standards of right and wrong that prescribe what humans ought to do, usually in terms of rights, obligations, benefits to society, fairness, or specific virtues

**Experimenter bias:** The tendency for different people to observe things differently (also called observer bias)

**External data:** Data that lives, and is generated, outside of an organization

# F

**Fairness:** A quality of data analysis that does not create or reinforce bias

**Field:** A single piece of information from a row or column of a spreadsheet; in a data table, typically a column in the table

**First-party data:** Data collected by an individual or group using their own resources

**Foreign key:** A field within a database table that is a primary key in another table (Refer to primary key)

**FROM:** The section of a query that indicates where the selected data comes from

# G

**General Data Protection Regulation of the European Union (GDPR):** Policy-making body in the European Union created to help protect people and their data

**Geolocation:** The geographical location of a person or device by means of digital information

**Good data source:** A data source that is reliable, original, comprehensive, current, and cited (ROCCC)

# H

# I

**Internal data:** Data that lives within a company's own systems

**Interpretation bias:** The tendency to interpret ambiguous situations in a positive or negative way

# J

# K

# L

**Long data:** A dataset in which each row is one time point per subject, so each subject has data in multiple rows

# M

**Mentor:** Someone who shares knowledge, skills, and experience to help another grow both professionally and personally

**Metadata:** Data about data

**Metadata repository:** A database created to store metadata

# N

**Naming conventions:** Consistent guidelines that describe the content, creation date, and version of a file in its name

**Networking:** Building relationships by meeting people both in person and online

**Nominal data:** A type of qualitative data that is categorized without a set order

**Normalized database:** A database in which only related data is stored in each table

**Notebook:** An interactive, editable programming environment for creating data reports and showcasing data skills

# O

**Observer bias:** The tendency for different people to observe things differently (also called experimenter bias)

**Open data:** Data that is available to the public

**Openness:** The aspect of data ethics that promotes the free access, usage, and sharing of data

**Ordinal data:** Qualitative data with a set order or scale

**Ownership**: The aspect of data ethics that presumes individuals own the raw data they provide and have primary control over its usage, processing, and sharing

# P

**Pixel:** In digital imaging, a small area of illumination on a display screen that, when combined with other adjacent areas, forms a digital image

**Population:** In data analytics, all possible data values in a dataset

**Primary key:** An identifier in a database that references a column in which each value is unique (Refer to foreign key)

# Q

# R

**Record:** A collection of related data in a data table, usually synonymous with row

**Redundancy:** When the same piece of data is stored in two or more places

**Relational database:** A database that contains a series of tables that can be connected to form relationships

# S

**Sample:** In data analytics, a segment of a population that is representative of the entire population

**Sampling bias:** Overrepresenting or underrepresenting certain members of a population as a result of working with a sample that is not representative of the population as a whole

**Schema:** A way of describing how something, such as data, is organized

**Second-party data:** Data collected by a group directly from its audience and then sold

**SELECT:** The section of a query that indicates the subset of a dataset

**Social media:** Websites and applications through which users create and share content or participate in social networking

**String data type:** A sequence of characters and punctuation that contains textual information (Refer to Text data type)

**Structural metadata:** Metadata that indicates how a piece of data is organized and whether it is part of one or more than one data collection

**Structured data:** Data organized in a certain format such as rows and columns

# T

**Text data type:** A sequence of characters and punctuation that contains textual information (also called string data type)

**Third-party data:** Data provided from outside sources who didn't collect it directly

**Transaction transparency:** The aspect of data ethics that presumes all data-processing activities and algorithms should be explainable and understood by the individual who provides the data

# U

**Unbiased sampling:** When the sample of the population being measured is representative of the population as a whole

**United States Census Bureau:** An agency in the U.S. Department of Commerce that serves as the nation's leading provider of quality data about its people and economy

**Unstructured data:** Data that is not organized in any easily identifiable manner

# V

**Video file:** A collection of images, audio files, and other data usually encoded in a compressed format such as MP4, MV4, MOV, AVI, or FLV

# W

**WHERE:** The section of a query that specifies criteria that the requested data must meet

**Wide data:** A dataset in which every data subject has a single row with multiple columns to hold the values of various attributes of the subject

**World Health Organization:** An organization whose primary role is to direct and coordinate international health within the United Nations system

X

Y

Z