What data to collect

Mittwoch, 9. Juli 2025 10:43

How Data is collected
+ Interviews
+ Observations
+ Forms
+ Questionnaires
+ Survey
+ Cookies

Data collection considerations
+ how the data will be collected
+ choose data sources
+ decide what data to use
+ how much data to collect
+ select the right data type
+ determine a time frame

First-party data
+ data collected by individual or group using own resources

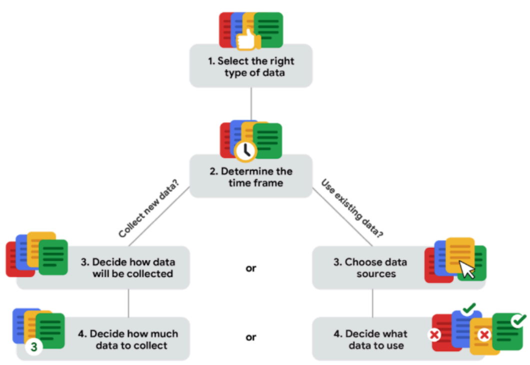
Second-party data
+ data collected by group directly from its audience and then sold

Third-party data
+ data collected from outside sources who did not collect it directly

Population
+ all possible data values in a certain dataset

Sample
+ part of a population that is representative of the population

Data collection considerations


1. Select the right type of data
2. Determine the time frame
Collect new data?
Use existing data?
3. Decide how data will be collected
or
3. Choose data sources
3
4. Decide how much data to collect
or
4. Decide what data to use

Datei auswählen Keine ausgewählt

Data Formats & Structures

Mittwoch, 9. Juli 2025 10:43

Qualitative Data
L> can not be listed by numbers
L> description, genre, movie title

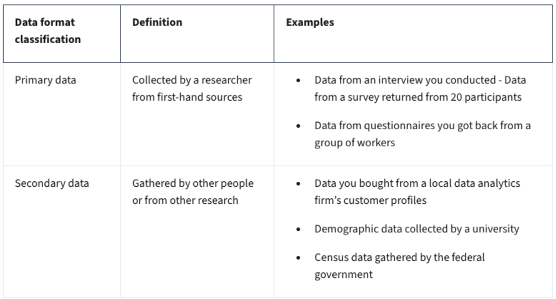
Quantitative Data
L> can be counted or expressed by numbers
<u>Discrete Data</u>
L> data that is counted and has a limited number of values
L> movie's budget and box office revenue
L> these numbers are limited - can be nothing between 1 & 2 cents
L> same with stars rating - only full stars allowed to be discrete
<u>Continuous Data</u>
L> is measured and can have almost any numeric value
L> for example measured using a timer
L> Run time of a movie 'The Data Analyst' - 110.0356 minutes
<u>Nominal Data</u>
L> qualitative data categorised **without** a set of order
L> doesn't have a sequence
L> Asking if someone watched a movie - yes, no, not sure
L> that's nominal - don't have a particular order
<u>Ordinal Data</u>
L> qualitative data **with** a set order or scale
L> ranking a movie from 1 - 5, some say 3, some 4, some 2
L> these rankings are in order of how a person likes a movie

Internal Data
L> data that lives within a company's own system
L> more reliable and easy to collect

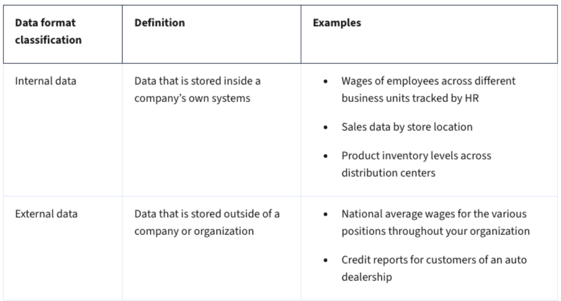
External Data
L> data that lives and is generated outside of an organisation
L> valuable when analysis depends on as many resources as possible

Structured Data
L> data organised in a certain format such as rows and columns#
L> spreadsheets and relational databases are two examples
L> they can store data in a structured way
L> makes data easier searchable and more analysis ready

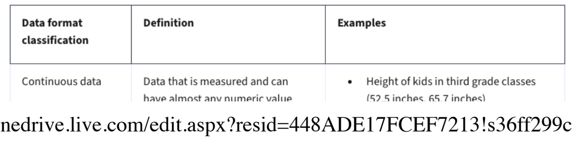
Unstructured Data
L> not organised in any easily identifiable manner
L> audio and video files are examples
L> might have internal structure, but doesn't fit into rows/columns



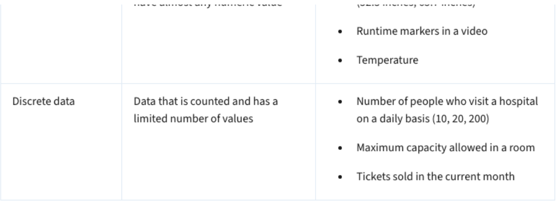
Data format classification	Definition	Examples
Primary data	Collected by a researcher from first-hand sources	• Data from an interview you conducted - Data from a survey returned from 20 participants • Data from questionnaires you got back from a group of workers
Secondary data	Gathered by other people or from other research	• Data you bought from a local data analytics firm's customer profiles • Demographic data collected by a university • Census data gathered by the federal government



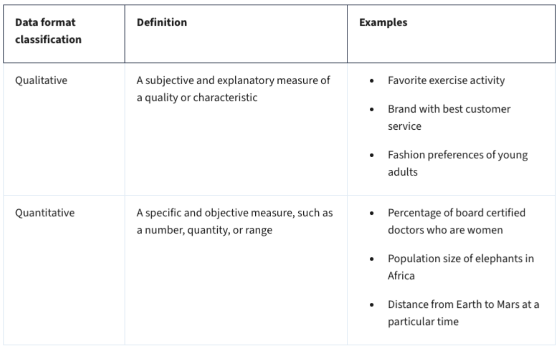
Data format classification	Definition	Examples
Internal data	Data that is stored inside a company's own systems	• Wages of employees across different business units tracked by HR • Sales data by store location • Product inventory levels across distribution centers
External data	Data that is stored outside of a company or organization	• National average wages for the various positions throughout your organization • Credit reports for customers of an auto dealership



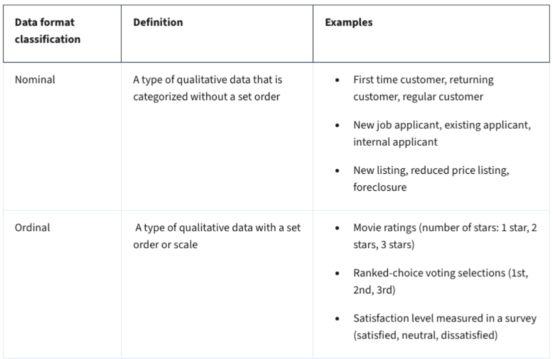
Data format classification	Definition	Examples
Continuous data	Data that is measured and can have almost any numeric value	• Height of kids in third grade classes (52.5 inches, 65.7 inches)



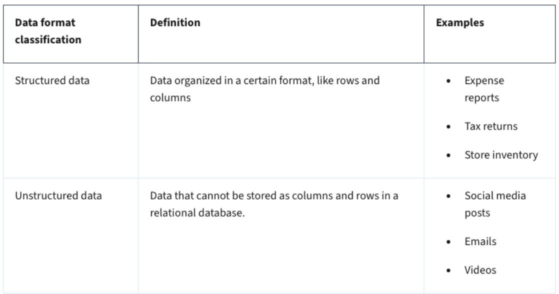
	have almost any numeric value	(32.5 inches, 33.7 inches) • Runtime markers in a video • Temperature
Discrete data	Data that is counted and has a limited number of values	• Number of people who visit a hospital on a daily basis (10, 20, 200) • Maximum capacity allowed in a room • Tickets sold in the current month



Data format classification	Definition	Examples
Qualitative	A subjective and explanatory measure of a quality or characteristic	• Favorite exercise activity • Brand with best customer service • Fashion preferences of young adults
Quantitative	A specific and objective measure, such as a number, quantity, or range	• Percentage of board certified doctors who are women • Population size of elephants in Africa • Distance from Earth to Mars at a particular time



Data format classification	Definition	Examples
Nominal	A type of qualitative data that is categorized without a set order	• First time customer, returning customer, regular customer • New job applicant, existing applicant, internal applicant • New listing, reduced price listing, foreclosure
Ordinal	A type of qualitative data with a set order or scale	• Movie ratings (number of stars: 1 star, 2 stars, 3 stars) • Ranked-choice voting selections (1st, 2nd, 3rd) • Satisfaction level measured in a survey (satisfied, neutral, dissatisfied)



Data format classification	Definition	Examples
Structured data	Data organized in a certain format, like rows and columns	• Expense reports • Tax returns • Store inventory
Unstructured data	Data that cannot be stored as columns and rows in a relational database.	• Social media posts • Emails • Videos

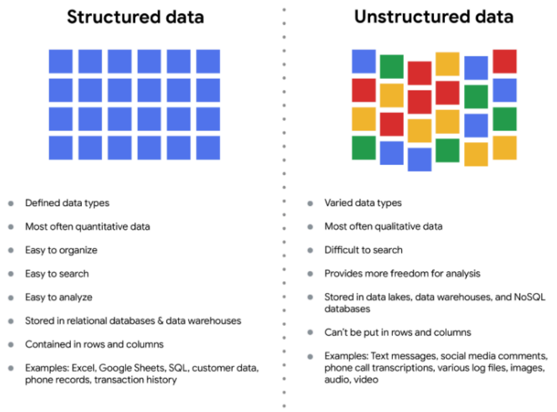
Data Models

Mittwoch, 9. Juli 2025 10:43

Structured data works best in a **Data Model**
L> a model that is used for organising data elements
L> and how they relate to each other
L> providing some kind of map of how data is organised

What are **Data Elements**
L> pieces of information, such as people's names, account numbers,
or addresses

Sources of structured data
L> spreadsheets
L> databases that store datasets

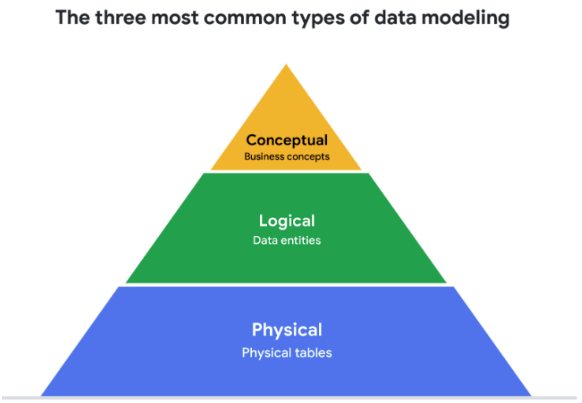


Structured data

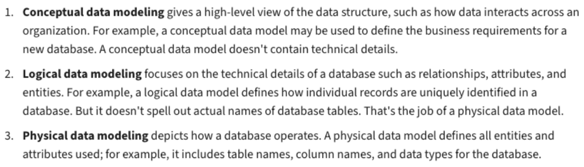
- Defined data types
- Most often quantitative data
- Easy to organize
- Easy to search
- Easy to analyze
- Stored in relational databases & data warehouses
- Contained in rows and columns
- Examples: Excel, Google Sheets, SQL, customer data, phone records, transaction history

Unstructured data

- Varied data types
- Most often qualitative data
- Difficult to search
- Provides more freedom for analysis
- Stored in data lakes, data warehouses, and NoSQL databases
- Can't be put in rows and columns
- Examples: Text messages, social media comments, phone call transcriptions, various log files, images, audio, video



The three most common types of data modeling



1. **Conceptual data modeling** gives a high-level view of the data structure, such as how data interacts across an organization. For example, a conceptual data model may be used to define the business requirements for a new database. A conceptual data model doesn't contain technical details.
2. **Logical data modeling** focuses on the technical details of a database such as relationships, attributes, and entities. For example, a logical data model defines how individual records are uniquely identified in a database. But it doesn't spell out actual names of database tables. That's the job of a physical data model.
3. **Physical data modeling** depicts how a database operates. A physical data model defines all entities and attributes used; for example, it includes table names, column names, and data types for the database.

2 Data-Modelling Techniques *(Approaches)*
+ Entity Relationship Diagram (ERD)
L> visual ways to understand relationship between entities in model
+ Unified Modelling Language Diagram (UML)
L> very detailed diagrams describing the structure of system
L> such as entities, attributes, operations, and their relationships

Datei auswählen Keine ausgewählt

Data Types

Mittwoch, 9. Juli 2025 10:43

Data Type
L> specific kind of data attribute
L> tells what kind of value the data is

Data Types in Spreadsheets
L> number
L> text or string
L> boolean

Text/String Data Type
L> sequence of characters and punctuation
L> contains textual information
L> phone numbers, street addresses
L> treated like text, not like numbers

Boolean Data Type
L> A data type with only two possible values
L> such as TRUE or FALSE
L> in example it is handled by over or under the number 50
L> we can add more words than T/F, it will still be Boolean

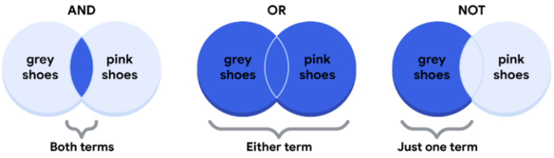
Datei auswählen Keine ausgewählt

Boolean Logic

Mittwoch, 9. Juli 2025 10:43

Usage
L> wide range of data analysis tasks
L> writing queries for searches
L> checking for conditions when writing programming code

Boolean Logic Example



AND - Operator
L> can make conditions add up to become true

OR - Operator
L> either one condition is enough, both conditions are also ok

NOT - Operator
L> IF (Colour="Grey) AND (Colour=**NOT** "Pink") then buy them
L> this example explains how it can be used as a negative condition

Very special is the **power of multiple conditions!**

Alternative Terms
+ rows share the meaning with records
+ columns share the meaning with fields

End

Datei auswählen Keine ausgewählt

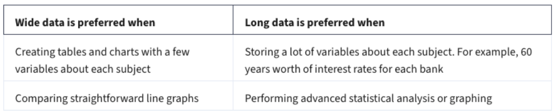
Wide vs Long Data

Mittwoch, 9. Juli 2025 10:43

Wide Data
L> every data subject has a single row
L> multiple columns hold the values
L> helpful for comparing specific attributes across different subjects

Long Data
L> each row represents one observation per subject
L> each subject will be represented by multiple rows
L> useful for comparing changes over time
L> useful for making comparisons across subjects

For our example
L> before each year of population had an own column
L> you could compare the numbers next to each other
L> countries were not repeating
L> this was wide data
L> now all years are in one column
L> you can compare the numbers beneath each other
L> countries are repeating
L> this is long data



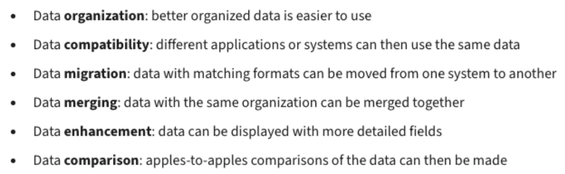
Wide data is preferred when	Long data is preferred when
Creating tables and charts with a few variables about each subject	Storing a lot of variables about each subject. For example, 60 years worth of interest rates for each bank
Comparing straightforward line graphs	Performing advanced statistical analysis or graphing

Datei auswählen Keine ausgewählt

Transforming Data

Mittwoch, 9. Juli 2025 10:43

Goals for data transformation



- Data **organization**: better organized data is easier to use
- Data **compatibility**: different applications or systems can then use the same data
- Data **migration**: data with matching formats can be moved from one system to another
- Data **merging**: data with the same organization can be merged together
- Data **enhancement**: data can be displayed with more detailed fields
- Data **comparison**: apples-to-apples comparisons of the data can then be made

Data Merging
L> converting a database into another one *(compatibility)*

Datei auswählen Keine ausgewählt

Glossary

Mittwoch, 9. Juli 2025 10:43

Glossary terms from module 2

Terms and definitions for Course 3, Module 1

Agenda: A list of scheduled appointments

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

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

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

Data element: A piece of information in a dataset

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

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

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

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

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

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

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

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

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

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

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

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)

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

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

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

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

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

Datei auswählen Keine ausgewählt