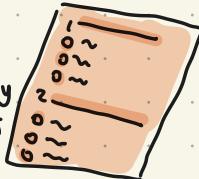


# Week 1: Data Exploration

## Data Formats & Data Structures

Data can be collected by

1. Interviews
2. Observations → mostly used by scientists
3. forms
4. Questionnaires
5. Surveys
6. Cookies → track people online's activities and interests



in owner's name

First-party data

↳ data collected by an individual or group using their own resources.

in a third party's name

Second-party data

↳ data collected by a group directly from its audience and then sold.

outside org.  
Third-party data (not reliable)  
↳ outsources not directly collected  
↳ doesn't have a direct relationship with data.

### Data Collection Considerations

③ How the data will be collected?

④ choose data sources

④ Decide what data to use

④ How much data to collect?

① Select the right data type

② Determine the time frame

Population → entire sample → some groups

Need answer immediately → Historical Data

Discrete → counted & limited number

e.g. price, rating, stats (no decimal accepted version)

Continuous → unspecified number of possible points

3.41.....

Nominal → categorized without a set order

(no sequence)

e.g. Yes/No

Ordinal → within an order

e.g. 1, 2, 3, 4, 5

from External source → 3. Spreadsheet

o Google Sheets → IMPORTRANGE()

o Excel → IMPORTHTML()

### Must know!

lives inside a company's own systems

outside company (in case that internal data isn't enough / need more info to describe data)

Internal / External / Secondary data

collected by a researcher from first-hand sources

Primary

Secondary

by other people

\* subjective and explanatory \*  
measures of qualities and characteristics  
e.g. - Exercise activity most enjoyed

\* specific and objective measures of numerical facts  
e.g. - Percentage of alcohol in beverages

0101 110010  
00110 010101  
010101

almost numerical value  
Weight, Height 52.4  
- Temperature 25.0  
- Runtime in video 18:34

Continuous

limited number of values  
e.g. - Number who visit a hospital in Dec 2022  
- Room's max capacity

- tickets sold 50, 79  
records

Discrete

in a certain format (rows, columns)  
e.g. - Expense reports  
- TAX returns

Structured

e.g. - Social media  
- Videos

Unstructured

e.g. - Emails

Nominal

not categorized with a set order

Ordinal

categorized with a set order / scale

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

1001  
00110  
0101

Week 2: Data Bias (4)  
Ensuring Data Integrity < Good data (😊) Bad data (😢)

- Bias → a preference in favor of or against a person, group of people / things.



# 2

Types of Data Bias

- **Data Bias** → type of error that systematically skews results.
  - **Sampling Bias** → a sample isn't representative of the population (small group sampling)  
**Prevent:**
    - Randomly sample it with fairness e.g. not all random is all female/male
    - Visualization easy to recognize it's
  - **Observer / Experimenter / Research Bias** → Different peoples observe things differently.  
R<sup>A</sup> / R<sup>B</sup>
  - **Interpretation Bias** → Interpret ambiguous things differently (they might have diff backgrounds)
  - **Confirmation Bias** → search for interpret info in a way that confirms pre-existing beliefs.

All affects way we make sense of data

### • "Good" Data Sources

Reliable: ☺

Original: with original sources

Comprehensive: research all aspects

Current: to present up-to-date

Cited: more credible e.g. Academic Research, Government data, vetted public datasets



### "Bad" Data Sources don't ROCCC



### • Data Ethics (Do / Don't)

↳ well-founded standards of right and wrong, and data privacy included

o GDPR (General Data Protection Regulation of the European Region)

### \*\* o Aspects of Data Ethics \*\*

- 1) Ownership: individuals own the raw data they provide and have primary control over its usage.
- 2) Transaction transparency: all data-processing activities and algorithms should be completely explainable.
- 3) Consent: an individual's right to know explicitly details about how and why their data will be used.
- 4) Currency: individuals should be aware of financial transactions.
- 5) Privacy (personal): preserving a data subject's information ("Data Protection")
- 6) Openness: free access, usage, sharing of their data e.g. Kaggle datasets

สิ่งที่ต้องคำนึงถึง  
ตรวจสอบ, แก้ไข,  
รักษาความปลอดภัย  
และตรวจสอบได้

- Names
- IP addresses
- Medical Records
- Email addresses
- Photographs

Open Data  
only if...

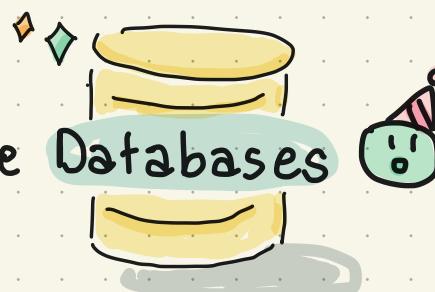
For data to be considered open, it has to: (Data can be considered open when meet 3 of these standards)  
① Be available and accessible to the public as a complete dataset.

② Be provided under terms that allow it to be reused and distributed.

③ Allow universal participation, so that everyone can use, reuse, and redistribute the data.

ดีกว่ากัน

Next week,  
we'll play with the Databases



### Week 3: Databases, Metadata

Working on large datasets in SQL

- Primary key
- Candidate / Alternate / Secondary key → unique identifier in the PK is unique identifying each row in the candidate key

**Database** → is a collection of data stored in a computer system.

Important metadata

→ "data about data" ↗ **Metadata Repository**

→ tells what data is about. ↗ a database specifically created

such as Title and Description, Tags and Categories, who created it and when from Email, Photo, Spreadsheet, Websites, Books, ...

3 Types of metadata  
① Descriptive: metadata that describes a piece of data such as ISBN behind book.  
② Structural: metadata that indicates how a piece of data is organized such as table of contents.  
③ Administrative: metadata that indicates the technical source of a digital asset such date & time photo taken.

delineated by character / space / tab

④ Comma-Separated Values (CSV) ↗ helps examining a small subset of a large dataset  
↳ distinguishes values from one another.



Sandbox (up to 12 projects, free)

Free trial (with Google Cloud, may have cost for upgrade)



**Relational Database**

is a database that contains a series of related tables that can be connected via their relationships.

consisted of Table name, Attributes - ↗ can't be NULL / blank, only 1/table \*

"us"  
"US"  
"us"  
"Us"  
country\_code ↗ not the same

**Primary key**: identifier that references a column in which each value is unique

**Foreign key**: reference key, how one table can have ≥ 1 connected to another table

**Composite key**: primary key that constructed using multiple columns in a table.

Naming Conventions

- **SQL Dialects**, some SQL dialects are case-sensitive such as Big Query, however most of them is case-insensitive like MySQL, PostgreSQL, SQL Server.
- Column names should be all lowercase. → when create a new column use "snake\_case" format e.g. as table\_name
- Table names — " — in CamelCase.
- To treat the text as string use "" or ''

\* Avoid spaces and special characters in file names \*

naming conventions (meaningful), include date & version, short & sweet)

YYYYMMDDV02

foldeing (organize files into folders)

metadata to describe data

automatically back up your files

also separate unfinished files from finished one in the folder by creating "archives"

### Week 4: Protecting Data



Benefits of organizing data:

- Easier to find & use.
- Avoid making mistakes during your analysis.
- protect your data.

"Access Control"

Data security → is protecting data from unauthorized access / corruption by adopting safety measures.

Security measures:

1) Encryption

2) Tokenization: replaces the data elements you want to protect with randomly generated data or "Token". The original data is stored in a separated location and mapped to the tokens. (Tokenized data and Token Mapping)

(optional)

Week 5: Online Presence

GitHub (x) on LinkedIn (in)

Making connections

Allows recruiter to see yourself → make sure the post content is appropriate

both online & offline

Networking → professional relationship building.

→ meet people with same interests.

→ mentors

5



You did a great job!  
Keep on !!

4