Week 1 : Data Exploration

< Protecting data
Ethical data analysis

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

🐱 **Third-party data** (not reliable)
↳ outsources not directly collected

จาก owner ข้อมูลตรง
First-party data
↳ data collected by an individual or group using their own resources.

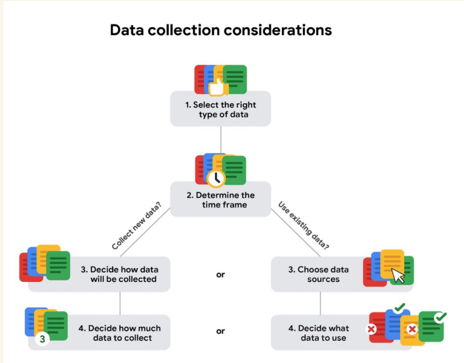
จากองค์กรอื่นที่ขาย
Second-party data เช่น รพ.เก็บข้อมูลสุขภาพจากปชช.
↳ data collected by a group directly from its audience and then sold.

Data Collection Considerations

③ { • How the data will be collected?
• Choose data sources

④ { • Decide what data to use
• How much data to collect? < Population → entire / Sample → some groups

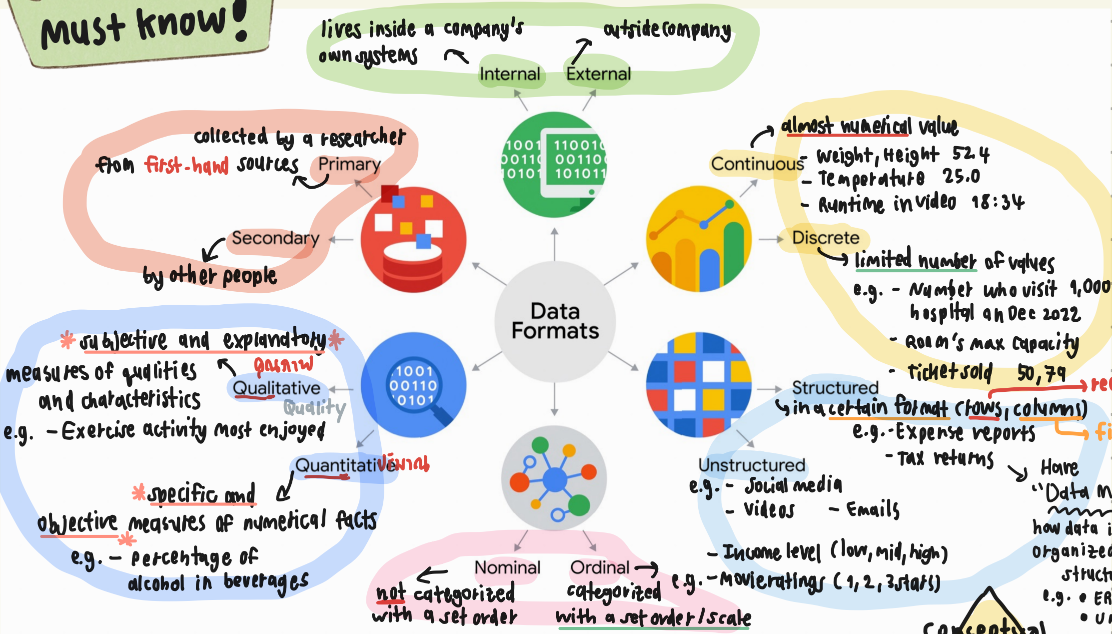
① • Select the right data type
② • Determine the time frame — Need answer immediately → Historical Data

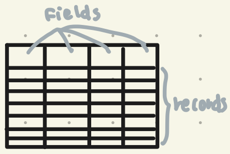


Data Formats

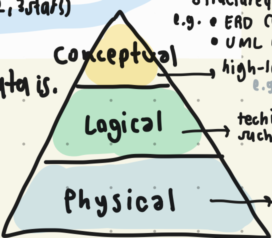
- **Quantitative data** → e.g. numbers, range, price
  - **Discrete** → counted & limited number
    e.g. price, rating stars (no decimal accepted version)
    ↖ have decimal limited
  - **Continuous** → unspecified number of possible points (3.41.....)
- **Qualitative data** → cannot be counted, measured
  e.g. text, description, heading
  - **Nominal** → categorized without a set order (no sequence)
    e.g. Yes/No
  - **Ordinal** → within an order (1 2 3 4 5)
    e.g. 1 2 3 4 5 stars

Must know!





e.g. • ERD (Entity Relationship Diagram)
• UML (Unified Modeling Language)



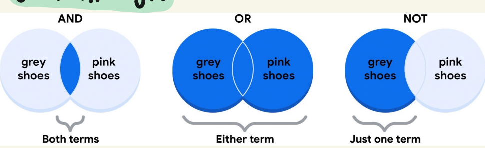
- Conceptual → high-level view of the data structure (e.g. how data interacts across an org.)
- Logical → technical details of a database such as relationships, attributes, and entities
- Physical → depicts how a database operates, define all entities and attributes used (e.g. table names / columns)

Data Types

→ a specific kind of data attribute that tells what kind of value data is.

In Spreadsheet
- Number (1, 3.14, 3.50, 11.)
- Text / String ("Chan") → not used in calculation
- Boolean (True/False)

Boolean logic



easy to understand & often use

Wide Data format

↳ every data subject has a single row with multiple columns to hold the values

id	fname	lname
1	Ariya	PP
2	Donald	Trump

what to use depends on your work

Long Data format

↳ each row is one time point per subject, each subject will have data in multiple rows

Country	countryname	Year
ARG	Argentina	2010
ARG	Argentina	2012
ARG	Argentina	2011

can be converted called "Data Transformation"
process of changing data's format, structure, or values