# 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

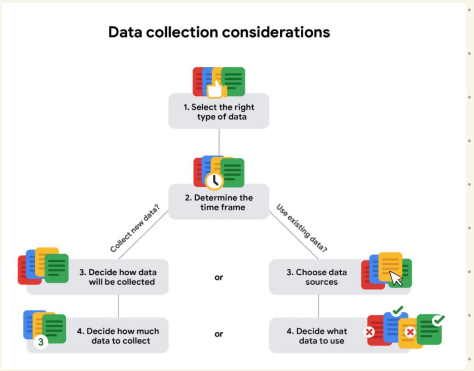
**Third-party data** (not reliable)
- outsources not directly collected
- doesn't have a direct relationship with data.

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

จากการซื้อมาใช้งาน
**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

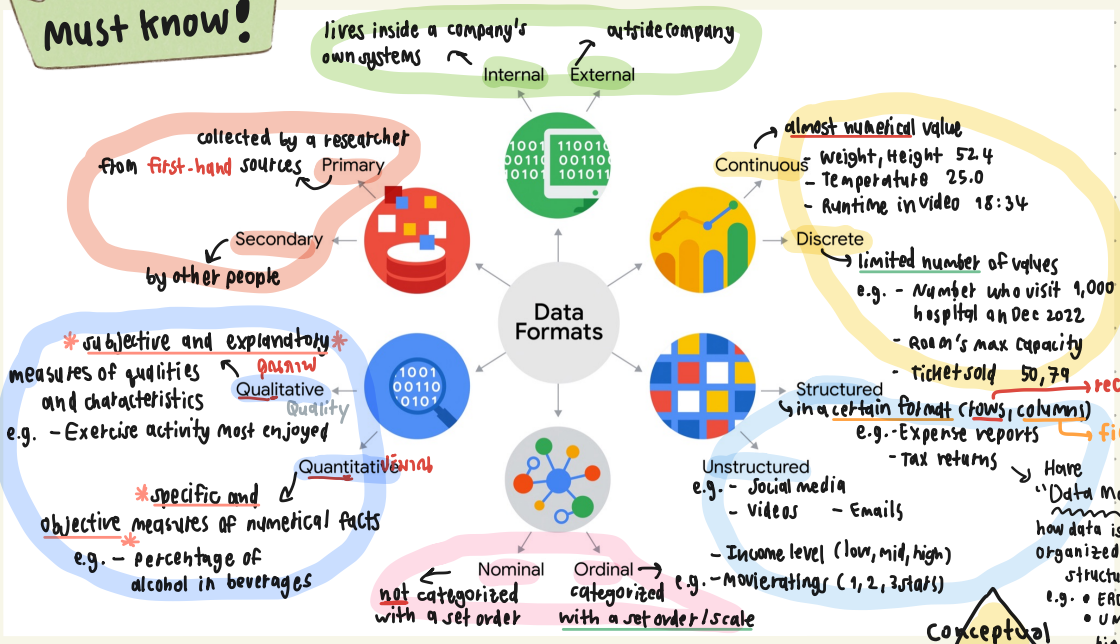


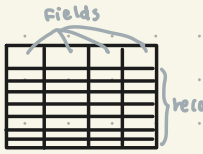

## Data Formats

- Quantitative data → e.g. numbers, range, price
  - Discrete → counted & limited number e.g. price, rating stars (no decimal accepted version)
  - Continuous → unspecified number of possible points (have decimal limited) 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 e.g. 1 2 3 4 5 stars

1

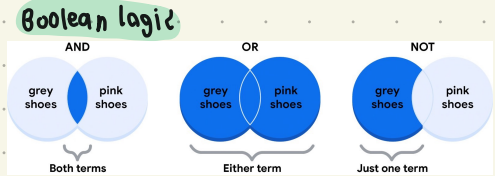
## Must know!




**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)




| id | fname | lname |
|---|---|---|
| 1 | Ariya | pp |
| 2 | Donald | Trump |

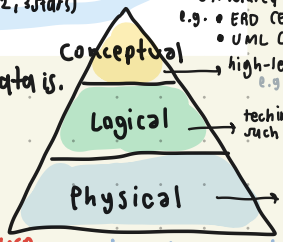
easy to understand & often use

**Wide Data format**
- every data subject has a single row with multiple columns to hold the values

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 |
|---|---|---|
| AR6 | Argentina | 2010 |
| AR6 | Argentina | 2012 |
| AR6 | Argentina | 2011 |




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

**Week2:** Data Bias (4) / Ensuring Data Integrity (Good data / Bad data) / Data Ethics (6)


2

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

- **Data Bias** → type of error that systematically skews results.

Types of Data Bias:

- **Sampling Bias** → a sample isn't representative of the population (small group sampling)
  - Prevent: Random sampling
    - Randomly sample it with fairness e.g. not all random is all female/male
    - Visualization (easy to recognize it)
- **Observer / Experimenter / Research Bias** → Different peoples observe things differently. (A / B)
- **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

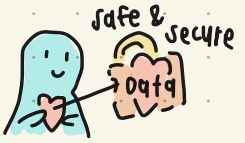
→ Sometimes "Good" can be subjective, therefore to define what's good with "ROCCC"

- "Good" Data Sources

**ROCCC**

- 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 (Validation) Public Datasets

"Bad" Data Sources don't ROCCC




**Ethics!**

- Data Ethics (Do / Don't)
  - ↳ well-founded standards of right and wrong, and data privacy included
    - GDPA (General Data Protection Regulation of the European Region)

** - Aspects of Data Ethics **

Important !!!

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

สิทธิในการตรวจสอบ, แก้ไข, อัพเดท ข้อมูลของตัวเองได้

**Data Anonymization** is the process of **protecting people's private or sensitive data** by **eliminating that kind of information.**

e.g.
- Names
- IP addresses
- Medical Records
- Email addresses
- Photographs

ต่างๆ กัน

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.

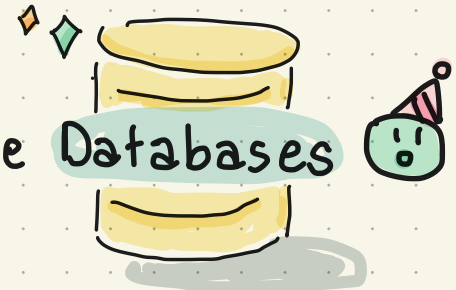
**Personal identifiable information (PII)**

is data that likely to identify a person and make information known about them.

* It's important to keep it safe *

such as
- person's address
- Credit Card information
- Social security number

Next week, we'll play with the **Databases**



# Week 3: Databases, Metadata
Working on large datasets in SQL

- Primary key
- Candidate / Alternate / Secondary key ทั้งหมดนี้เลือกเป็น PK ก็ได้ unique identify each row ได้ ก็คือ candidate key

**Normalization**
is a process of organizing data in a relational database.
e.g. • creating tables,
• establishing relationships between tables

**Database** → is a collection of data stored in a computer system.
- metadata → "data about data"
  → tells what data is about.

**Relational Database** → is a database that contains a series of related tables that can be connected via their relationships.
→ consisted of Table name, Attributes —
- **Primary key**: identifier that references a column in which each value is unique (→ can't be null / blank, only 1 / table *)
- **Foreign key**: reference key, how one table connected to another table (can have ≥ 1)
- **Composite key**: primary key that constructed using multiple columns in a table.

Note:
* เหมือน Inspecting a dataset (พวกแรก) อันนี้