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

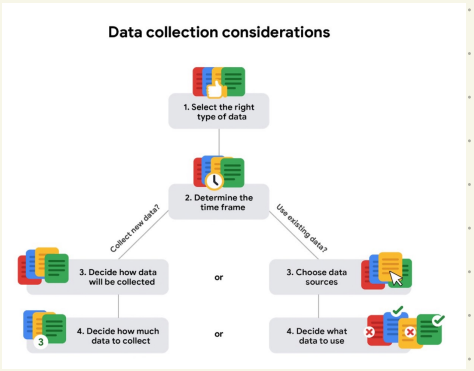
Outside Org.
- **Third-party data** (not reliable)
  - outsources not directly collected
  - doesn't have a direct relationship with data.

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

จากองค์กรหรือคนที่ทราบ
- **Second-party data** เช่น สพ.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



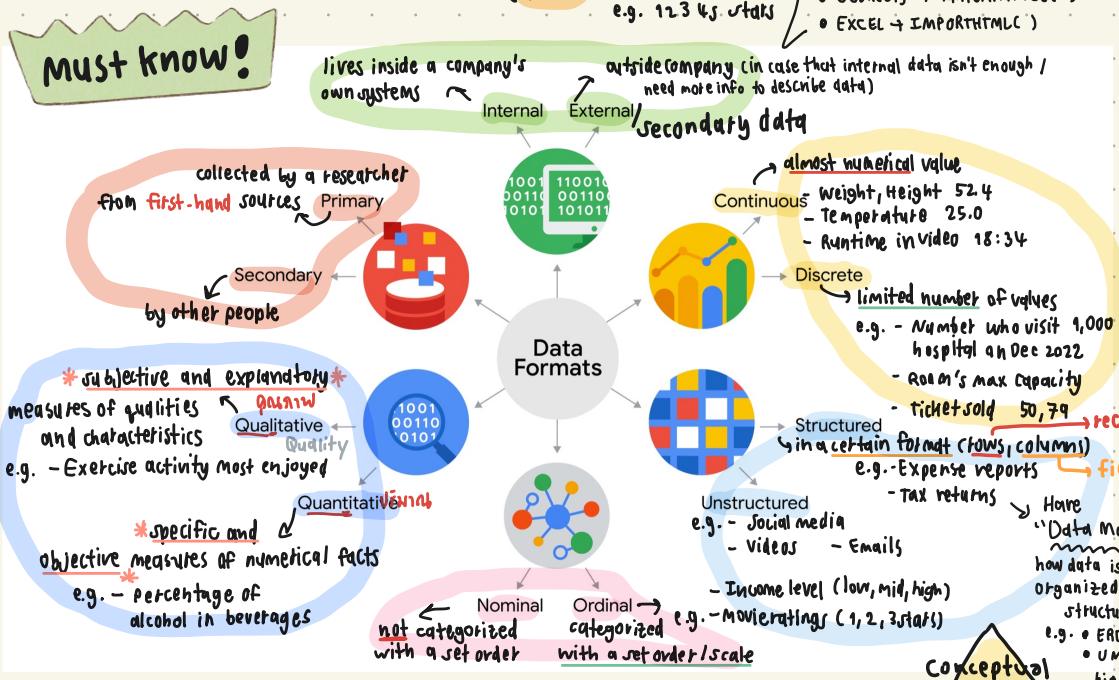
1. Select the right type of data
2. Determine the time frame
   - Need answer immediately → Historical Data

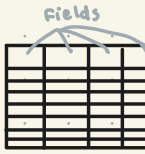
Data Formats

- 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

From External source → a spreadsheet
- GGSheets → IMPORTRANGE( )
- EXCEL → IMPORTHTML( )

Must know!

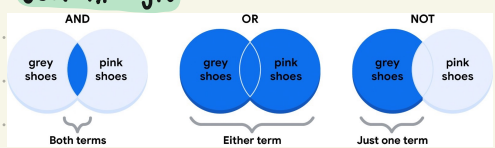




Data Types → a specific kind of data attribute that tells what kind of value data is.

In spreadsheet:
- Number (1, 3.14, 3.50, 11.)
- Text (String) ("Chan") → not used in calculation
- Boolean (TRUE/FALSE)

Boolean logic



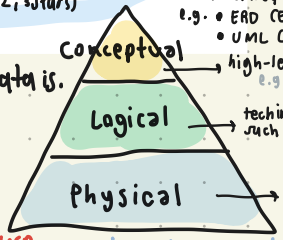
id	fname	lname
1	Ariya	pp
2	Donald	Trump

easy to understand & often use

- **Wide Data format**
  - every data subject has a single row with multiple columns to hold the values
- **Long Data format**
  - each row is one time point per subject, each subject will have data in multiple rows

what to use depends on your work

country	countryname	year
AR6	Argentina	2010
AR6	Argentina	2012
AR6	Argentina	2011



- 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

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.
- **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
  - 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

ROCCC

"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 (masking, Hashing, Blanking) that kind of information.

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

ต่างๆ กัน

Open Data only if..

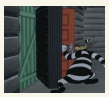
For data to be considered open, it has to: (Data can be considered open when meet 3 of these standards)
1. Be available and accessible to the public as a complete dataset.
2. Be provided under terms that allow it to be reused and distributed.
3. Allow universal participation, so that everyone can use, reuse, and redistribute the data.

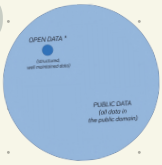
Open data is the information that has been published on government-sanctioned portals the data is structured, open-licensed, and well maintained.

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

Public data is the data that exists everywhere else.
It's freely available, but not really accessible on the web.
The data is unstructured.





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

- Primary key
- Candidate / Alternate / Secondary key ทั้งหมดนี้เกิดจาก เลือกเป็น PK แล้ว 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

3

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

- Important **metadata** → "data about data"
  - → tells what data is about.
  - → such as Title and Description, Tags and Categories, who created it and when
  - from Email, Photo, Spreadsheet, Websites, Books, ...

Metadata Repository
↳ a database specifically created to store metadata to make it easier to use, help track who uses metadata.

Data Governance → Helps company manage data
is a process to ensure the formal management of a company's data assets.
by Metadata Analyst/Specialist.

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.

- Comma-separated values (CSV) (delineated by character / space / tab)
  - → helps examining a small subset of a large dataset
  - → distinguishes values from one another.

Google BigQuery
- 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 —
- **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.

Naming Conventions
country_code ← "US", "US", "us", "Us" } not the same

- 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 ' '

Week 4: Protecting Data

Organizing Data

Benefits of organizing data:
- Easier to find & use.
- Avoid making mistakes during your analysis.
- Protect your data.

* Avoid spaces and special characters in file names *
- naming conventions (meaningful, include date & version, short & sweet) → YYYY_MM_DD_V02
- foldering (Organize files into folders) → also separate unfinished files from finished one in the folder by creating "archives"
- metadata to describe data
- automatically back up your files

4

Data security → is protecting data from unauthorized access / corruption by adopting safety measures. ("Access Control")

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** on GitHub, LinkedIn
- 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