# Types of Data

CPE 232: Data Models

*Dr. Sansiri Tarnpradab*

*Department of Computer Engineering, KMUTT*

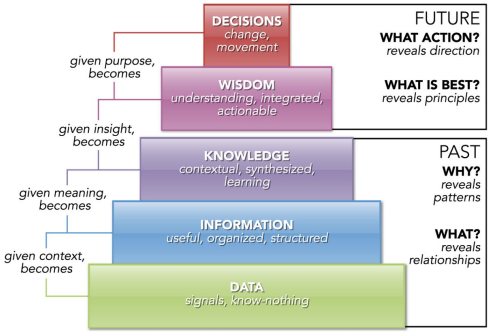
# Outline

- Review
- Structured Data
- Semi-structured Data
- Unstructured Data

# Review

Decision

DIKW



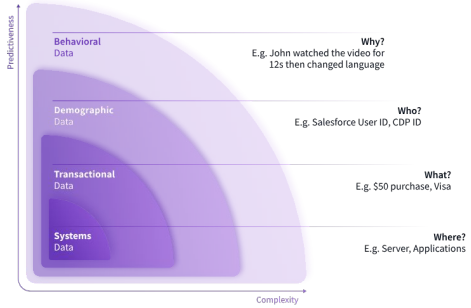

Demographic (Who)

Transactional (What)

Interaction (What, Who)

**Types**

Behavioral (Why)




**Data**

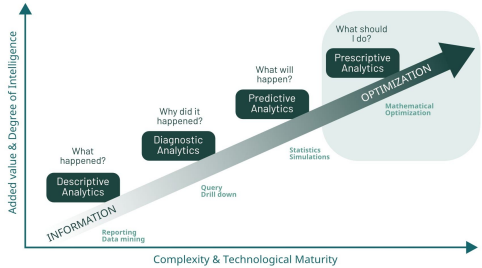
**Analytics**

Descriptive (What happened?)

Diagnostic (Why it happened?)

Predictive (When will it happen?)

Prescriptive (What should you do?)




**Modeling**

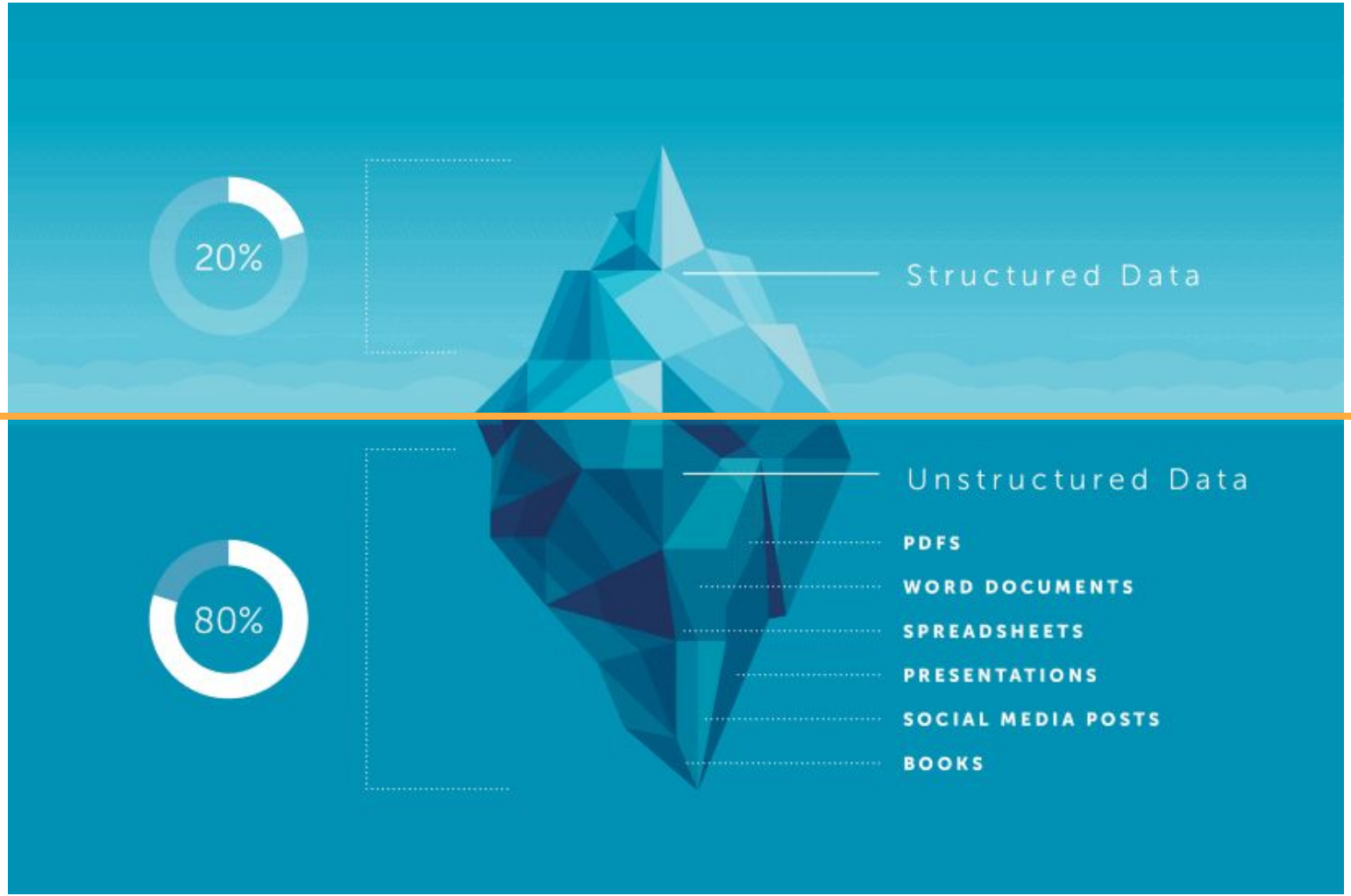
Create an abstract representation

Machine Learning

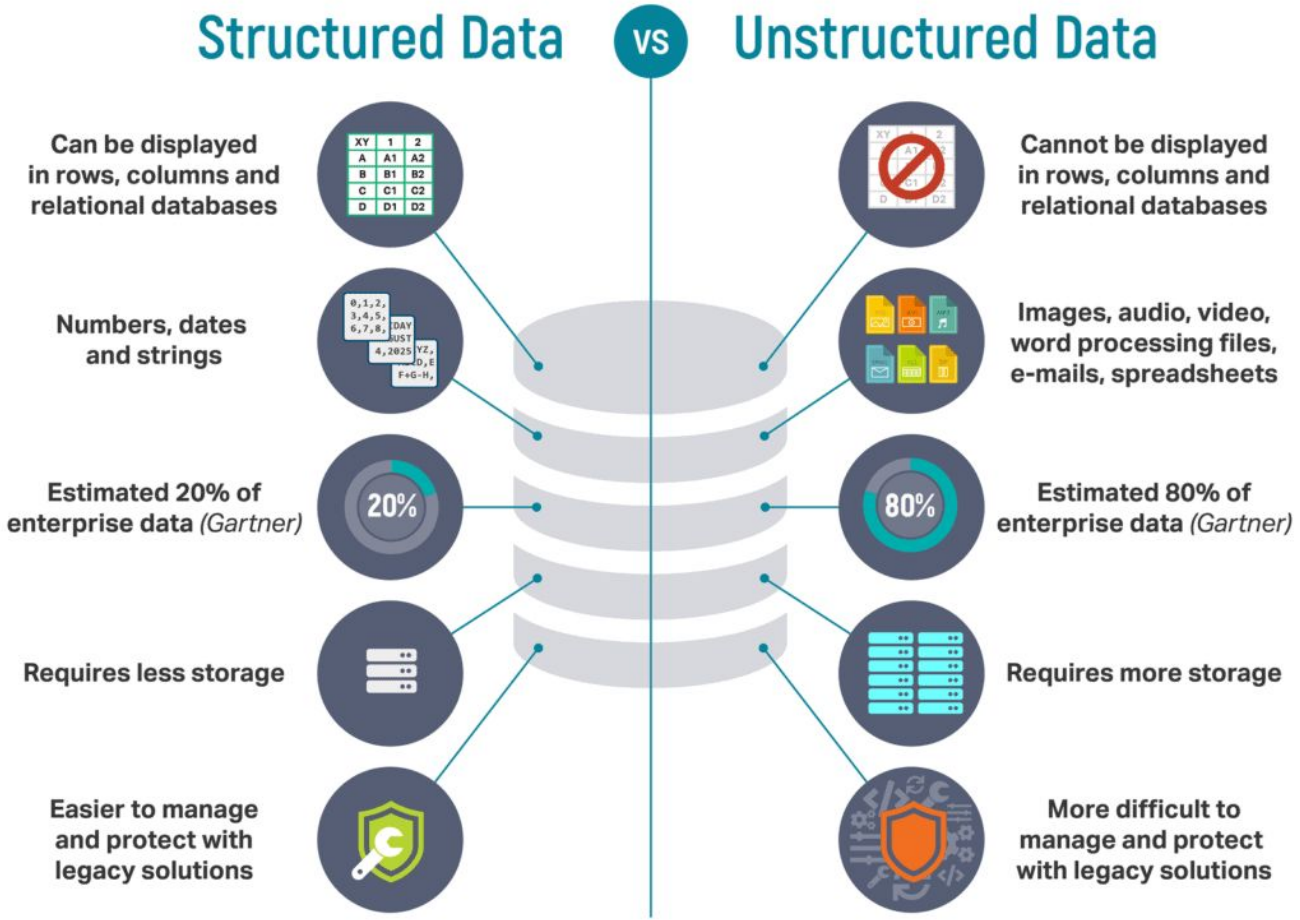
Visualization

# Different Types of Data

## (Format-wise)



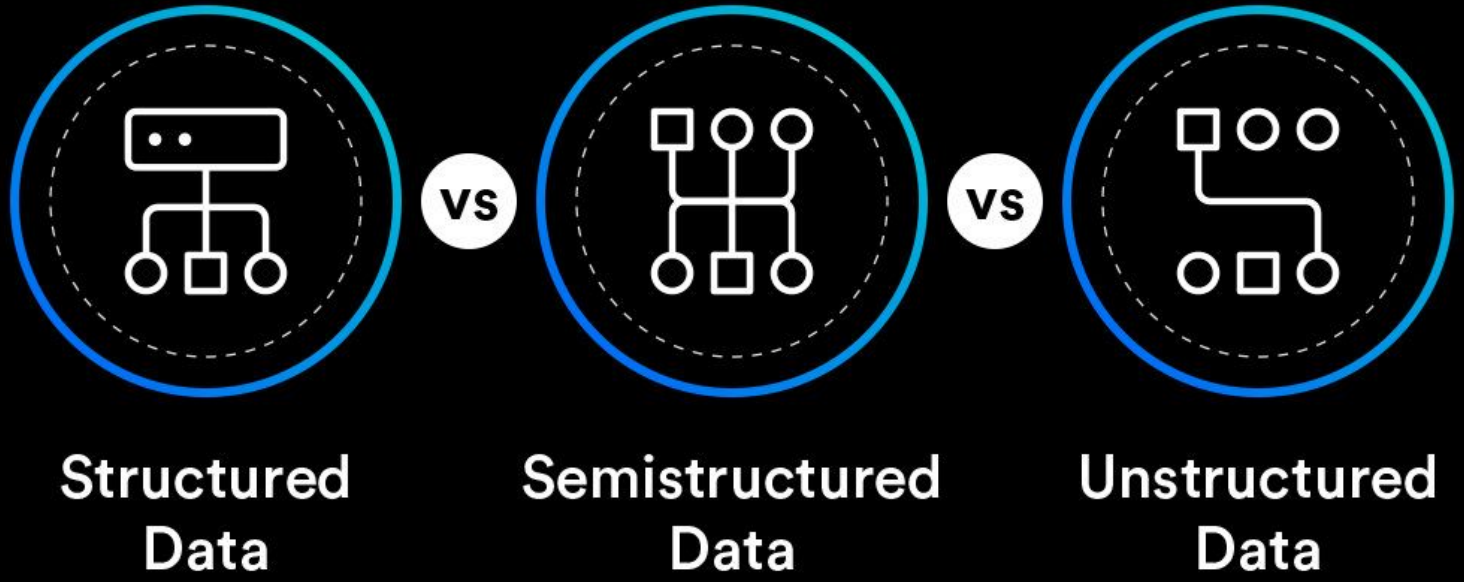

Ref: https://k21academy.com/microsoft-azure/dp-900/structured-data-vs-unstructured-data-vs-semi-structured-data/








Ref: https://k21academy.com/microsoft-azure/dp-900/structured-data-vs-unstructured-data-vs-semi-structured-data/






Ref: https://automationhero.ai/blog/whats-the-difference-between-structured-semi-structured-and-unstructured-data/

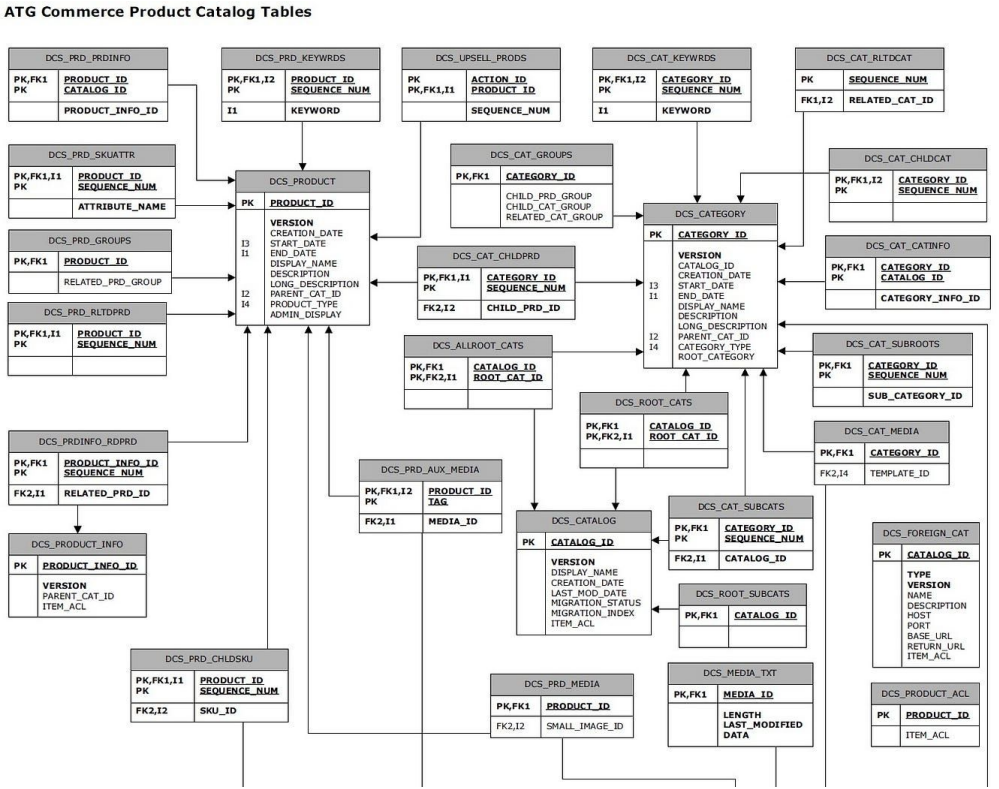
**Structured Data**

**Semi-structured Data**

**Unstructured Data**

# Structured Data

- Organized and predefined
- Tables
  - Rows
  - Columns
- Relational databases
- ER diagram
- SQL
  - Domain-specific language
  - For managing data in relational databases




*Ref: https://betterprogramming.pub/what-is-an-entity-relationship-diagram-d5db69a87971*

# Structured Data: Tabular Data

- Table
- Records are represented in rows (Horizontal)
- Attributes are represented in columns (Vertical)

Attributes

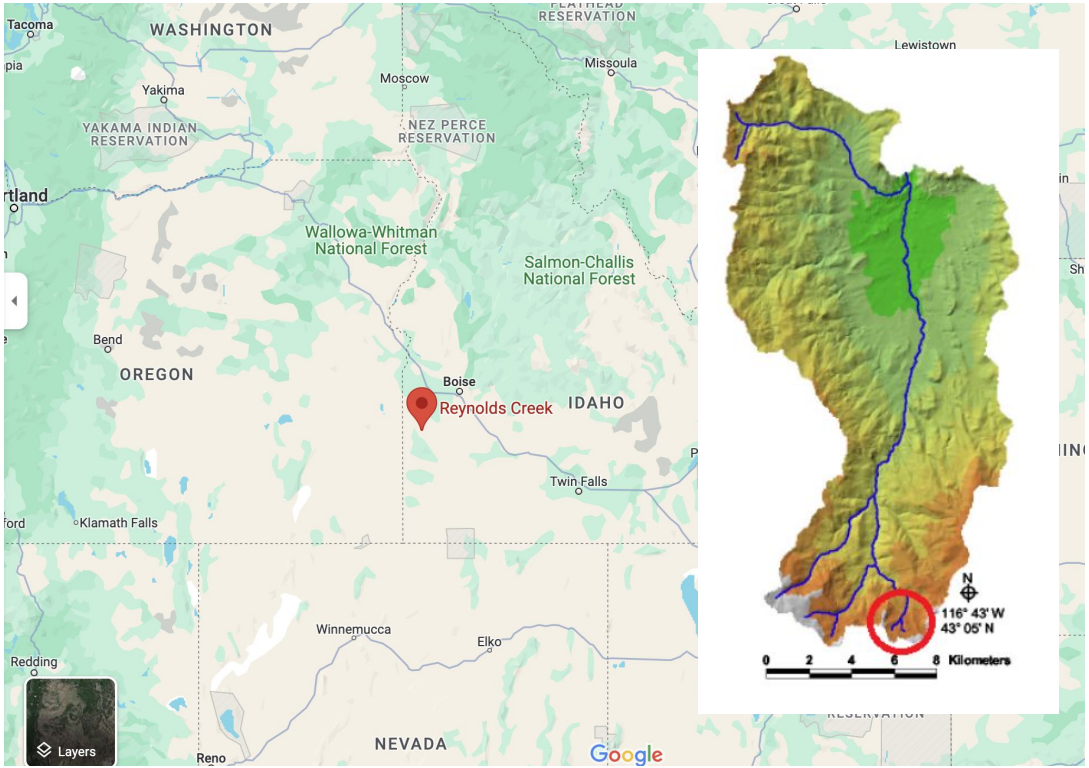
Records

| ID | Name | Email | Status |
|---|---|---|---|
| 01 | Autumn Larissa | au_larissa@kmutt.ac.th | Enrolled |
| 02 | Clover Aaron | cl_aaron@kmutt.ac.th | Enrolled |
| 03 | Coretta Russell | co_russel@kmutt.ac.th | Enrolled |
| 04 | Emmanuhel Hamlet | em_hamlet@kmutt.ac.th | Enrolled |

# Example: USDA Data

USDA Agricultural Research Service

**RCEW**: Reynolds Creek Experimental Watershed, Idaho



*References:*
- USA Map: https://www.ebay.ca/itm/363637325123
-USDA: https://chucktownfloods.cofc.edu/data-sources/usda-agricultural-research-service-gis-server/
- RCEW water stations: Tarnpradab, Sansiri, et al. "Neural networks for prediction of stream flow based on snow accumulation." 2014 IEEE Symposium on Computational Intelligence for Engineering Solutions (CIES). IEEE, 2014.

# Data & Attributes

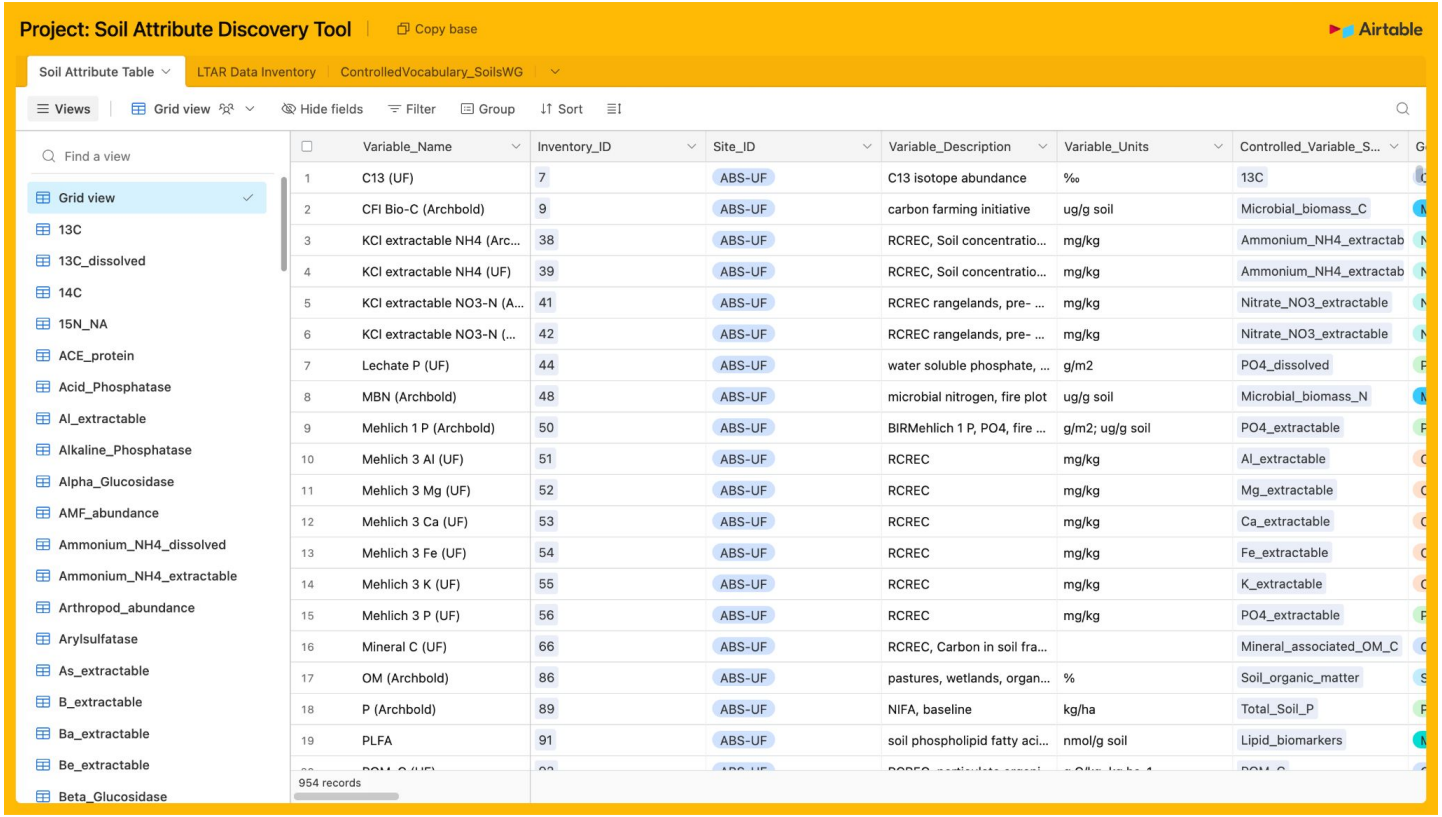
For instance…

- MM/DD/YYYY
- precipitation
- temperature
- moisture
- soil moisture
- soil temperature
- daily streamflow
- snow
- snow water equivalent (SWE)
- geographic


Ref: https://www.reynoldscreekczotour.com/about

# Soil Attributes

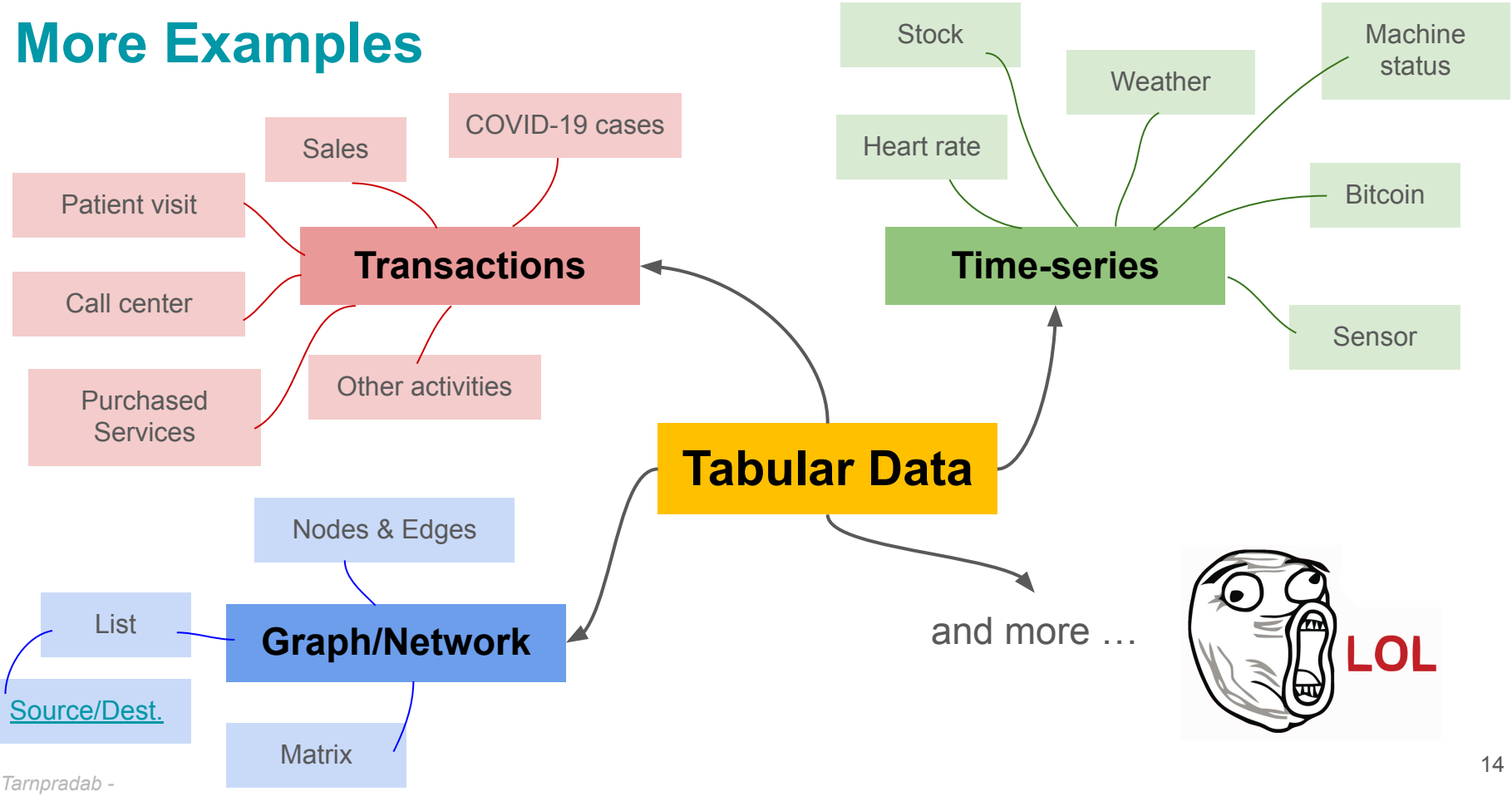
[Access Data]




| | Variable_Name | Inventory_ID | Site_ID | Variable_Description | Variable_Units | Controlled_Variable_S... |
|---|---|---|---|---|---|---|
| 1 | C13 (UF) | 7 | ABS-UF | C13 isotope abundance | ‰ | 13C |
| 2 | CFI Bio-C (Archbold) | 9 | ABS-UF | carbon farming initiative | ug/g soil | Microbial_biomass_C |
| 3 | KCl extractable NH4 (Arc... | 38 | ABS-UF | RCREC, Soil concentratio... | mg/kg | Ammonium_NH4_extractab |
| 4 | KCl extractable NH4 (UF) | 39 | ABS-UF | RCREC, Soil concentratio... | mg/kg | Ammonium_NH4_extractab |
| 5 | KCl extractable NO3-N (A... | 41 | ABS-UF | RCREC rangelands, pre- ... | mg/kg | Nitrate_NO3_extractable |
| 6 | KCl extractable NO3-N (... | 42 | ABS-UF | RCREC rangelands, pre- ... | mg/kg | Nitrate_NO3_extractable |
| 7 | Lechate P (UF) | 44 | ABS-UF | water soluble phosphate, ... | g/m2 | PO4_dissolved |
| 8 | MBN (Archbold) | 48 | ABS-UF | microbial nitrogen, fire plot | ug/g soil | Microbial_biomass_N |
| 9 | Mehlich 1 P (Archbold) | 50 | ABS-UF | BIRMehlich 1 P, PO4, fire ... | g/m2; ug/g soil | PO4_extractable |
| 10 | Mehlich 3 Al (UF) | 51 | ABS-UF | RCREC | mg/kg | Al_extractable |
| 11 | Mehlich 3 Mg (UF) | 52 | ABS-UF | RCREC | mg/kg | Mg_extractable |
| 12 | Mehlich 3 Ca (UF) | 53 | ABS-UF | RCREC | mg/kg | Ca_extractable |
| 13 | Mehlich 3 Fe (UF) | 54 | ABS-UF | RCREC | mg/kg | Fe_extractable |
| 14 | Mehlich 3 K (UF) | 55 | ABS-UF | RCREC | mg/kg | K_extractable |
| 15 | Mehlich 3 P (UF) | 56 | ABS-UF | RCREC | mg/kg | PO4_extractable |
| 16 | Mineral C (UF) | 66 | ABS-UF | RCREC, Carbon in soil fra... | | Mineral_associated_OM_C |
| 17 | OM (Archbold) | 86 | ABS-UF | pastures, wetlands, organ... | % | Soil_organic_matter |
| 18 | P (Archbold) | 89 | ABS-UF | NIFA, baseline | kg/ha | Total_Soil_P |
| 19 | PLFA | 91 | ABS-UF | soil phospholipid fatty aci... | nmol/g soil | Lipid_biomarkers |

Ref: https://airtable.com/appsb2TI92hnqjjSL/shrBUmoc3aLA2b4Rx/tblrfVUqxVbbYDDfo/viwd4RcFbjodKYiMu?blocks=hide

# More Examples



# SQL and Structured Data: A Perfect Match

- SQL
  - Structured Query Language
  - Designed for managing and querying data in relational databases
  - Optimized for performing operations on rows and columns efficiently

| ID | Name | Email | Status |
|---|---|---|---|
| 01 | Autumn Larissa | au_larissa@kmutt.ac.th | Enrolled |
| 02 | Clover Aaron | cl_aaron@kmutt.ac.th | Enrolled |
| 03 | Coretta Russell | co_russel@kmutt.ac.th | Enrolled |
| 04 | Emmanuhel Hamlet | em_hamlet@kmutt.ac.th | Enrolled |
| 05 | Liana Rivers | li_rivers@kmutt.ac.th | Withdrawn |

**Student records** (columns: ID, Name, Email, Status)

**SQL query example:**

```
SELECT Name, Email
FROM Students
WHERE Status = 'Enrolled';
```

# Structured Data: CSV

- Row-based
- Comma-separated value

| Pros | Cons |
| --- | --- |
| • Very common<br>• Easy to write and debug<br>• Lightweight | • Not effective as the size grows<br>• Takes a longer time to read<br>• Inefficient for large datasets (inefficient query esp. when only a subset of the columns are needed) |

# Structured Data: Parquet

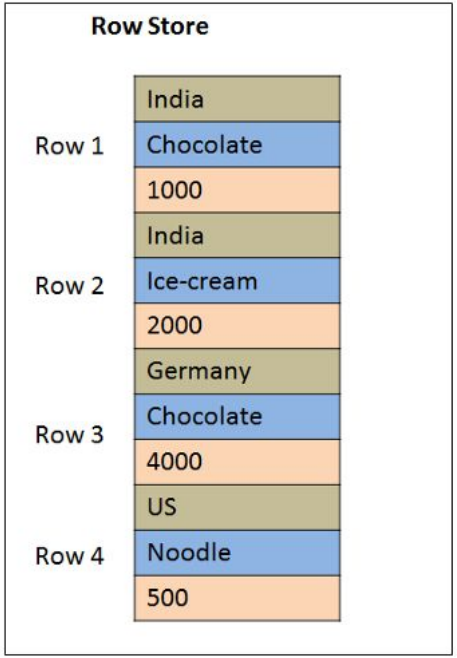
- Column-based
- Highly optimized for big data processing

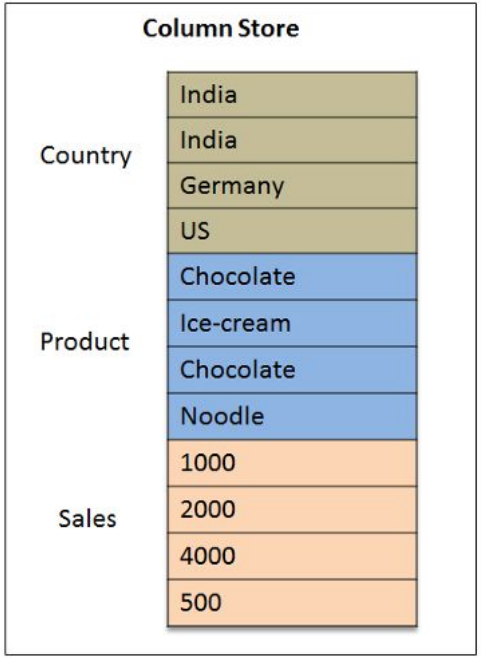
| Pros | Cons |
| --- | --- |
| • Efficient compression<br>• Fast query performance | • Slower write performance<br>• Not optimized for update operations<br>  ○ Modifying data can be cumbersome<br>  ○ Requires rewriting entire files |

# CSV vs Parquet

**Table**

| | Country | Product | Sales |
|---|---|---|---|
| Row 1 | India | Chocolate | 1000 |
| Row 2 | India | Ice-cream | 2000 |
| Row 3 | Germany | Chocolate | 4000 |
| Row 4 | US | Noodle | 500 |



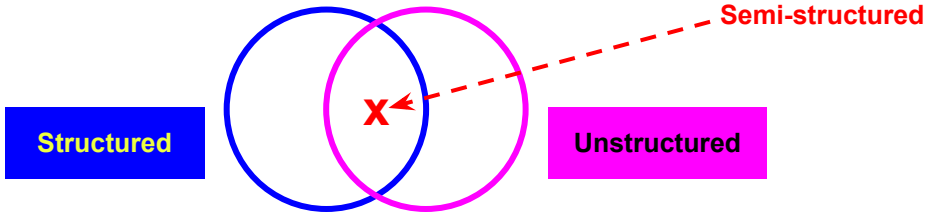

Ref img

**Structured Data**

**Semi-structured Data**

**Unstructured Data**

# Semi-structured Data




- Does not follow the format of a tabular data model or relational databases
- Does not have a fixed schema
- Organizations using semi-structured data may need help predicting where the information of interest is located
- Examples:
  - HTML
  - XML
  - JSON

# Semi-structured Data: HTML

- Hypertext Markup Language
- A standard markup language for creating web pages
- Tags and attributes that structure and present content on the web
- Examples:
  - Web scraping results
  - Online articles
  - Blog posts
  - Forum threads

```
<!DOCTYPE html>
<html>
<body>

<h1>My First Heading</h1>

<p>My first paragraph.</p>

</body>
</html>
```

# Semi-structured Data: XML

- eXtensible Markup Language
- A markup language
- Tags to define elements and attributes
- Used to represent and distribute data structures which can be often difficult to create using more standard tabular formats
- Allows for custom data structures

```
<data>
    <student name="John">
        <email>john@mail.com</email>
        <grade>A</grade>
        <age>16</age>
    </student>
    <student name="Alice">
        <email>alice@mail.com</email>
        <grade>B</grade>
        <age>17</age>
    </student>
    <student name="Bob">
        <email>bob@mail.com</email>
        <grade>C</grade>
        <age>16</age>
    </student>
    <student name="Hannah">
        <email>hannah@mail.com</email>
        <grade>A</grade>
        <age>17</age>
    </student>
</data>
```

# Semi-structured Data: JSON

- JavaScript Object Notation
- For storing and transmitting structured data
- Suitable for representing complex objects, arrays, and various data types
- Widely used in web APIs, configuration files, and data storage
- Examples:
  - User profiles
  - Social media posts
  - Product catalogs


```
{
    "statuses": [
        {
            "created_at": "Wed Jul 27 08:07:42 +0000 2016",
            "id": 7.5821225317067e+17,
            "id_str": "758212253170671616",
            "text": "RT @Skrip_Shit: Hidup itu emang penuh kekecewaan. Salah satunya
            "truncated": false,
            "entities": {
                "hashtags": [],
                "symbols": [],
                "user_mentions": [
                    {
                        "screen_name": "Skrip_Shit",
                        "name": "IG: MahasiswaSejati",
                        "id": 394951639,
                        "id_str": "394951639",
                        "indices": [
                            3,
                            14
                        ]
                    }
```


Ref: https://stackoverflow.com/questions/38609558/how-do-i-edit-created-at-from-twitter-json-api

# SQL and Semi-structured Data: To Some Extent

PostgreSQL and MySQL provide native support for **JSON** and **XML** data types.

**PostgreSQL (JSON)**

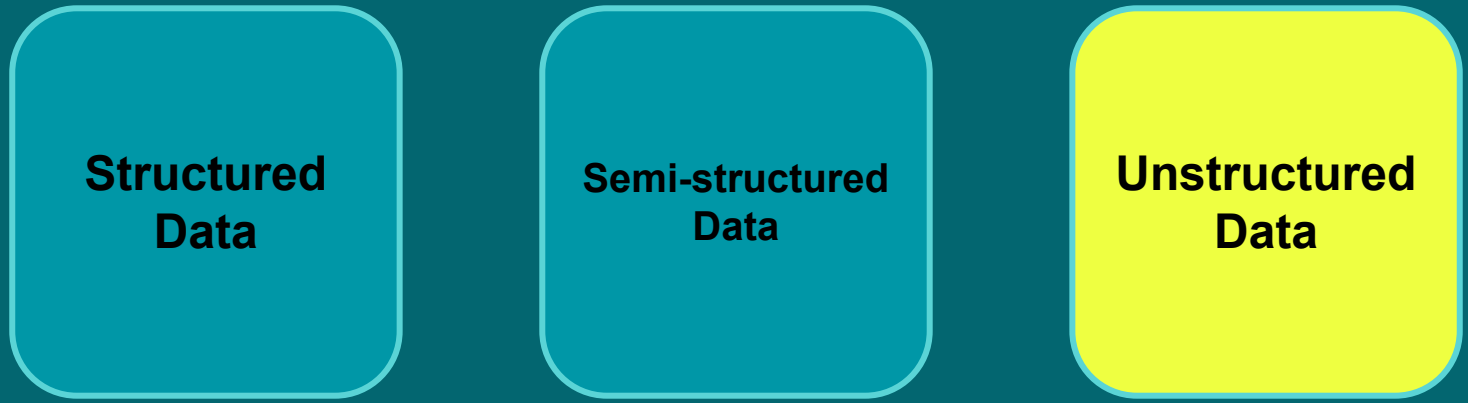
```
SELECT json_data->>'name' AS name
FROM users
WHERE json_data->>'age' = '25';
```

**MySQL (JSON)**

```
SELECT JSON_EXTRACT(json_column, '$.name') AS name
FROM users
WHERE JSON_EXTRACT(json_column, '$.age') = '25';
```

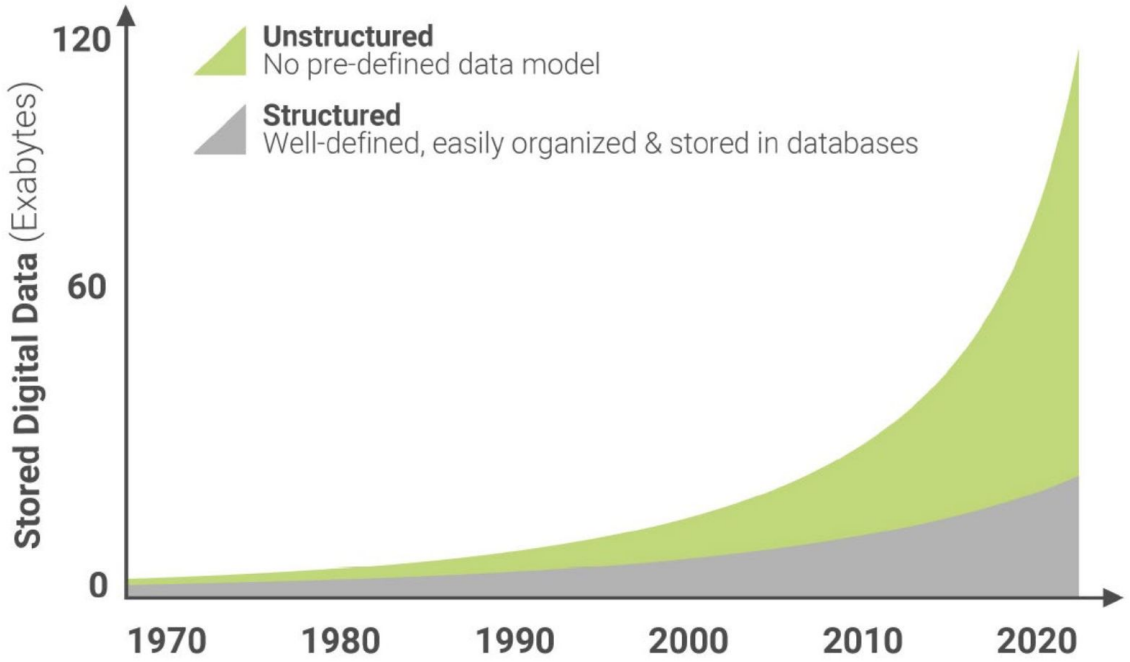
**SQL Server (XML)**

```
SELECT xml_column.value('(/user/name)[1]', 'VARCHAR(100)') AS name
FROM users
WHERE xml_column.value('(/user/age)[1]', 'INT') = 25;
```


Structured Data
Semi-structured Data
Unstructured Data

# Unstructured Data

- Format-free
- Not organized into any particular format
- Examples:
  - Text
  - Audio
  - Image
  - Video




Ref: https://www.datanami.com/2019/01/14/from-oscar-to-ai-mining-visual-assets-for-fun-and-profit/

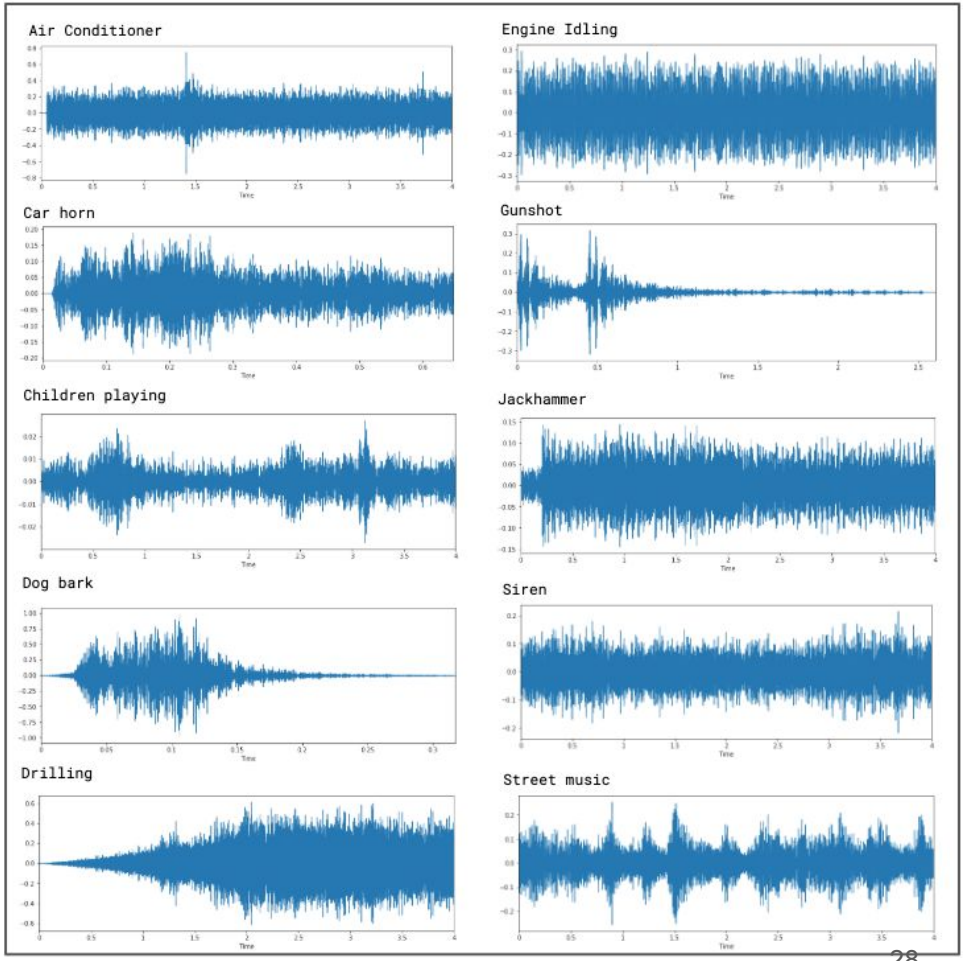
# Unstructured Data: Text

- Varies in length, language, and formatting
- Lacks a predefined structure
- Examples:
  - Emails
  - Chat logs
  - Surveys
  - Call center
  - Voice transcriptions
  - Meeting minutes
  - Social media (Facebook, X, etc)
  - Customer/Member transactions
  - Reviews


*IMDb Dataset - From 1888 to 2023*
*Ref: https://www.kaggle.com/datasets/komalkhetlani/imdb-dataset*

# Unstructured Data: Audio

- A sound wave → a continuous signal
- MP3 (.mp3), WAV (.wav), FLAC (.flac)
- Examples:
  - Speech data (Spoken words in different languages, accents, and dialects)
  - Different sounds (Animal sounds, sounds of objects, etc.)
  - Music data (music or song recordings)
  - Other digitally recorded human sounds (e.g. coughs, sneezes, or snores
  - Far-flung speech or other background noises

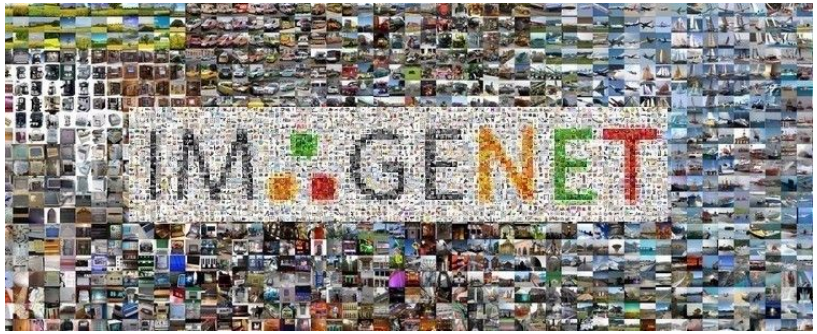



Ref: https://mikesmales.medium.com/sound-classification-using-deep-learning-8bc2aa1990b7

# Unstructured Data: Image

- A visual representation in form of a function `f(x,y)`
- `f` is related to the brightness (or color) at point `(x,y)`

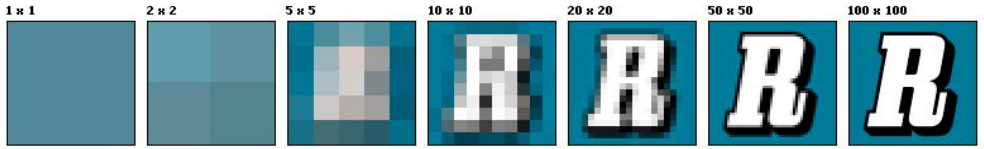
Ref: https://en.wikipedia.org/wiki/MNIST_database


Ref: https://cv.gluon.ai/build/examples_datasets/imagenet.html

# Digital image

- Discrete samples `f[x,y]` representing continuous image `f(x,y)`
  - `f[x,y]` is a 2D array
  - Pixel (picture element)




Ref: https://en.wikipedia.org/wiki/Image_resolution

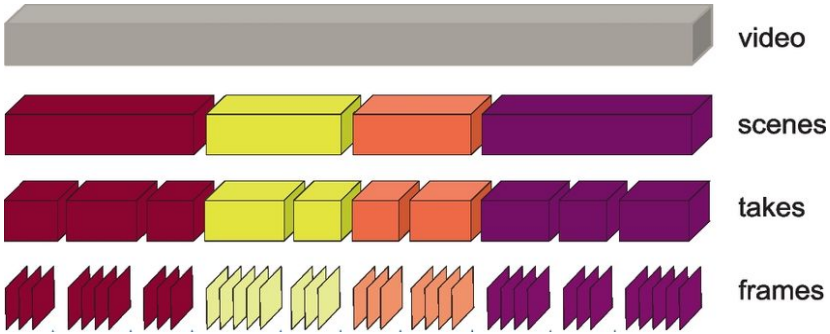
- Image resolution
  - The level of detail of an image
  - High-resolution displays have a higher pixel density
  - More pixels in the same amount of physical space

- Color components
  - Red: `R[x,y]`
  - Green: `G[x,y]`
  - Blue: `B[x,y]`
  - Monochrome: `R[x,y] = G[x,y] = B[x,y]`

Ref: https://www.slideserve.com/seamus/digital-image-processing-lecture-14-color-image-processing

# Unstructured Data: Video

- Audio-visual information in any digital or analog format
- MP4 (.mp4), AVI (.avi), MOV (.mov)
- Sequences of images (i.e. video frames)




Ref:
https://www.researchgate.net/publication/370808978_Survey_on_Video_Colorization_Concepts_Methods_and_Applications/figures?lo=1&utm_source=google&utm_medium=organic

*Ref GIF: Big Buck Bunny*

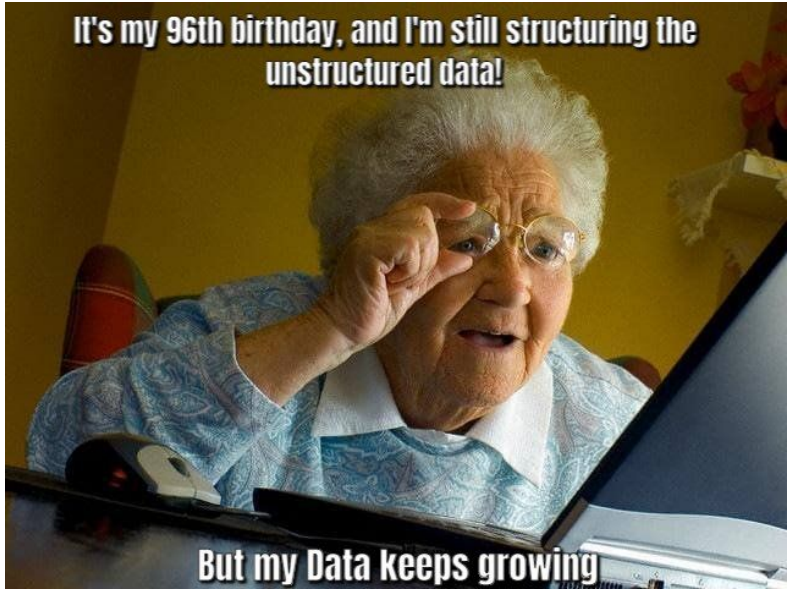
# SQL and Unstructured Data: Not Directly

- SQL databases are designed for tabular, relational data
- Inefficient for storing and querying unstructured data
- Workaround
  - Store unstructured data as BLOBs (Binary Large Objects) or TEXT columns
  - The actual query of that data needs to be done with other tools

| Feature | Structured Data | Semi-structured Data | Unstructured Data |
|---|---|---|---|
| **Format** | Predefined | Flexible | Native |
| **Organization** | Easy | Medium | Difficult |
| **Analysis** | Easy | Medium | Difficult |
| **Examples** | Tables, Spreadsheets, Databases | HTML, XML, JSON | Text, Audio, Video, Images |
| **Pros** | Efficient retrieval & query<br>Facilitates consistency & accuracy | More context than traditional structured data<br>More flexible & scalable | Provide rich and diverse information, leading to more in-depth insights and a better understanding of complex patterns and relationships |
| **Cons** | Not suitable for modern data sources which lack a well-defined structure | More challenging to extract<br>Inconsistencies could compromise data quality | More challenging to extract<br>Scalable storage concerns due to complexity of unstructured data |

# In Summary

- Structured Data
  - Tabular: rows & columns
  - Spreadsheets, CSV, Parquet
- Semi-structured Data
  - More flexible: tags & attributes
  - HTML, XML, JSON
- Unstructured Data
  - Format-free
  - Text, Audio, Image, Video




Ref: https://twitter.com/DataDynamicsInc/status/1522877966090661888

# Q & A