Lecture #1

Data Representation and Characteristics

Dr. Charnon Pattiyanon
Assistant Director of IT, Instructor
Department of Artificial Intelligence and Computer Engineering
CMKL University

Evolution of Computer Systems

Mainframes & PCs

1970s – 1980s

Client Server & Internet

1990s – 2000s

Cloud, Mobile & Big Data

2000s – 2010s

Intelligent Technologies

2010s – 2020s

ENABLING TECHNOLOGIES

- Transitors and Silicon Revolution
- Large-Scale Mainframe Computing Adoption
- Emergence of PCs
- Plant Floor Automation

- Widespread PC Adoption
- Broadband Internet
- ERP and Business Process Technologies

- Mobile and Smartphone Ubiquity
- Cloud Computing
- Social Network
- Big Data

- Machine Learning (ML) and Artificial Intelligence (AI)
- Internet-of-Things (IoT) and Distributed Computing
- Blockchain

CUSTOMER VALUE CREATION

Industrial Automation

Business Process Automation

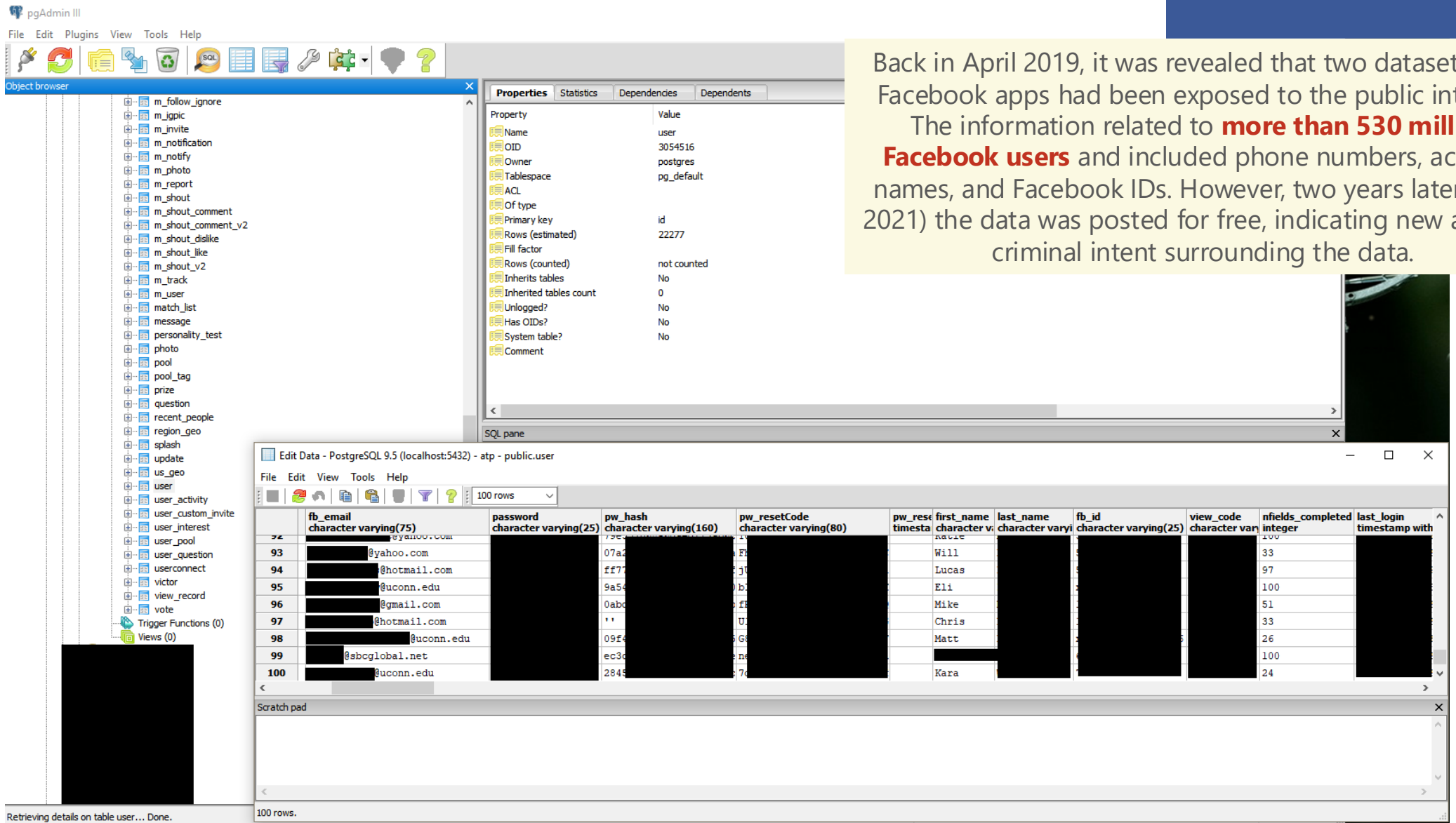
Digital Transformation

Intelligent Enterprise

You may have noticed that **data and information are becoming increasingly important** to the functionality and performance of **modern computer systems**.

Data is one of an organization's most valuable assets, and as such, it is often a primary target for **cybercriminals**.

Cybercrimes on Data and Information



Back in April 2019, it was revealed that two datasets from Facebook apps had been exposed to the public internet. The information related to **more than 530 million Facebook users** and included phone numbers, account names, and Facebook IDs. However, two years later (April 2021) the data was posted for free, indicating new and real criminal intent surrounding the data.

Cybercrimes on Data and Information

Massive Leak Of Stolen Thai PII Data On Dark Web By Cybercriminals

Recently, the Criminal Court in Thailand issued an order to block the website 9near.org. This action was taken after the site threatened to disclose the personal information of **55 million Thai citizens**, allegedly obtained from vaccine registration records. The court further declared that any other websites found distributing data from "9near.org" would also face blocking. This measure follows a request from the **Digital Economy and Society (DES) Ministry**, which is preparing for the likely apprehension of the individual responsible for the hack.

The person running the website, who goes by "**9Near – Hacktivist**", made an announcement on the Breach Forum website, claiming they had accessed personal details of **55 million people from Thailand**. This data includes full names, birthdates, ID card numbers, and phone numbers. Recently, the Rural Doctors Society suggested that this information might have originated from a leak at the **Public Health Ministry's Immunization Centre**.

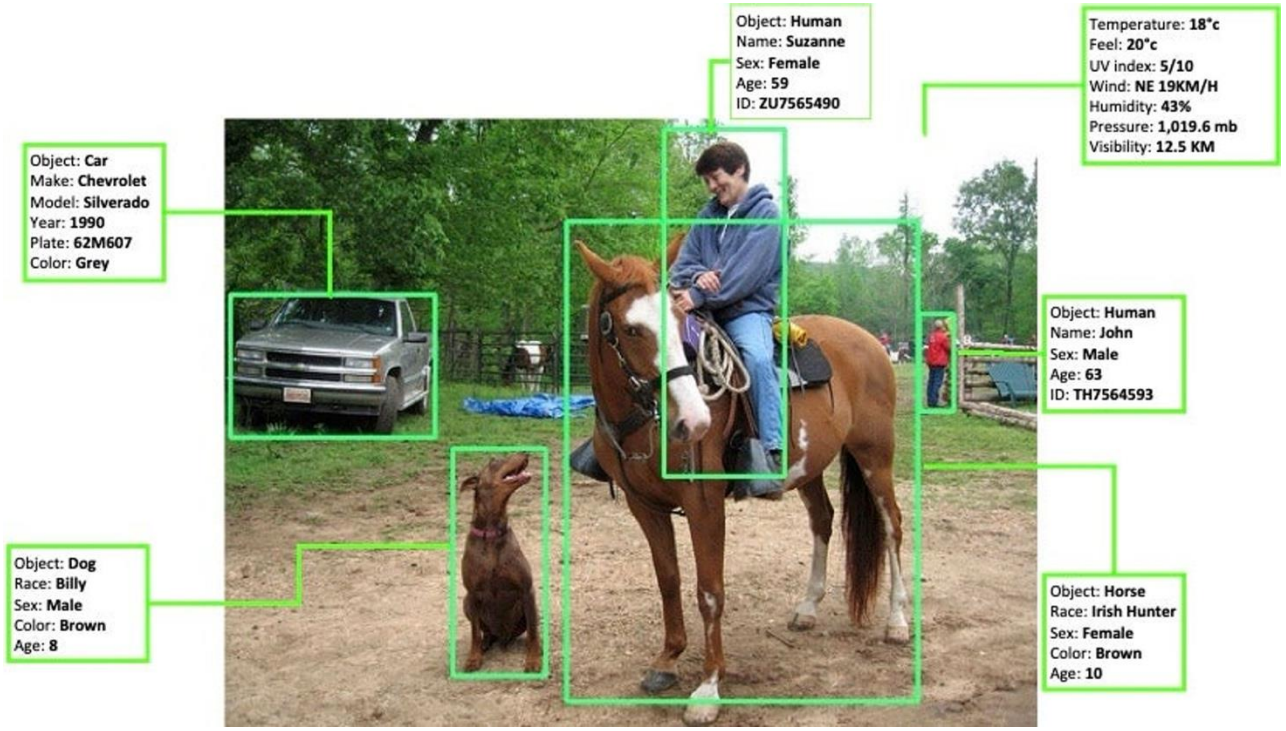
Are these data breaches critical?

Is this incident close enough to you?

What do you think will happen if adversaries gain access to this leaked data?

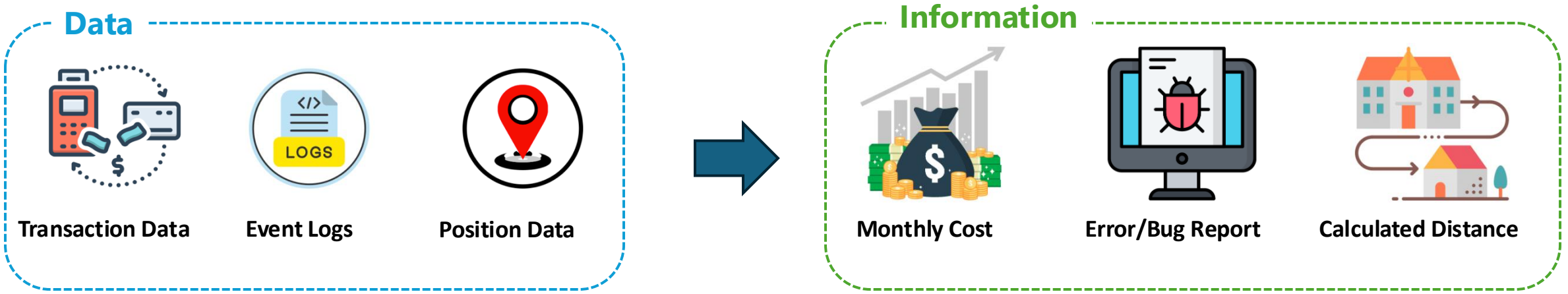
What Are Data and Information?

- **Data** is defined as:
  - Raw or unprocessed records gathered from various knowledge sources.
  - *Basic facts or statistics* that can be used for further analysis or applied to serve **a variety of purposes**.
  - Data is **collected** every time you make a purchase, browse a website, travel, make a call, or post on social media.
  - Data can originate from a wide range of **sources**, including sensors, surveys, experiments, observations, and existing records (such as historical data from financial transactions).



What Are Data and Information?

- **Information** refers to data that has undergone a process of **refinement**, **organization**, **processing**, or **summarization** in such a way that it yields meaning and relevance.
- In an organizational or operational context, information serves as a vital resource that enables decision-makers, such as operators, supervisors, or managers, to interpret and comprehend the current state of affairs.
- Unlike raw data, which may be voluminous and difficult to interpret in its unprocessed form, information has been transformed to highlight patterns, trends, or key indicators that are essential for informed decision-making, situational awareness, and strategic planning.



Types of Data

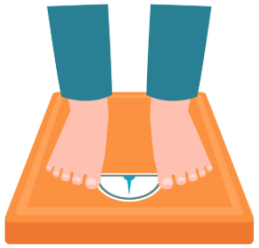
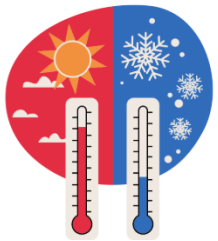
- **Data can be classified based on various perspectives**, such as *value*, *velocity*, *structure*, *sensitivity*, or other relevant characteristics.
- From a purely statistical perspective, **data can be categorized into two major types based on their values**.

Types of Data: Quantitative Data

- **Quantitative (numerical) data** refers to information that can be **expressed**, **measured**, and **compared using numerical values**, such as **integers or real numbers**[1].

Continuous Data

- **Continuous data** is a type of **quantitative data** that can be meaningfully divided into finer levels.
- It is measured on a scale or continuum and can take **almost any numeric value**—either within a finite or infinite range (interval) or as a ratio that compares two or more quantities.


Discrete Data

- **Discrete data** consists of finite, numeric, and countable values that **cannot be subdivided into smaller parts**.
- These values are typically whole numbers and represent **individual units**.

  Examples of discrete variables include counts and binary indicators.


Types of Data: Qualitative Data

- **Qualitative (categorical) data** refers to **non-numerical information**, such as *opinions*, *feelings*, *perceptions*, and *attitudes*. This type of data helps answer questions like "How did it occur?" or "Why did this occur?"[1]

Nominal Data

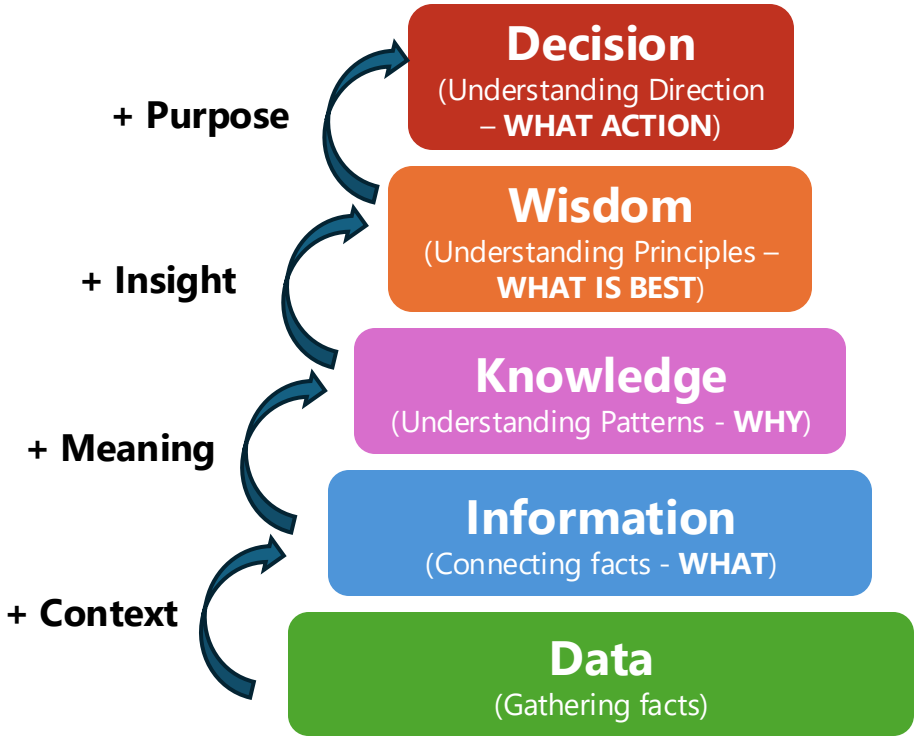
- **Nominal data** is a type of **categorical data** that has no inherent numerical value or order.
- It consists of names, labels, or categories used to **classify** and **organize** information into **distinct groups**

Ordinal Data

- **Ordinal data** is a type of categorical data that has a meaningful order or ranking associated with its values.

Impact of Data: DIKW Model

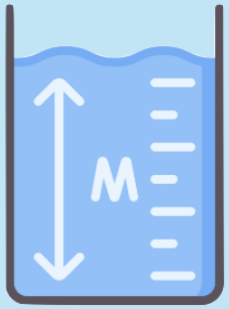
- There are found terms that are typically used to define the impact of data, which are **Data**, **Information**, **Knowledge**, and **Wisdom**.



- **Data** is considered the raw material for wise decision-making because it provides an objective and evidence-based foundation for drawing accurate conclusions.
- By analyzing large volumes of data using methods such as statistical analysis or machine learning algorithms, we can:
  - **connect facts together** – This is **Information**.
  - **uncover hidden patterns and insights** that may not have been immediately apparent. – This is **Knowledge** and **Wisdom**.
- Finally, wisdom emerges when these insights are applied with experience and sound judgment, enabling informed decisions about the **next course of action** and influencing **future strategies**. – This is a **decision**.

Characteristics of Data

- The **five main and innate characteristics of data (5Vs)** are:

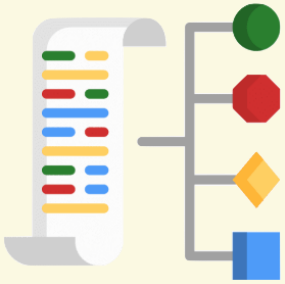


Volume

The total amount of data an organization generates and stores.

Velocity

The speed at which data is generated, transmitted, and processed into usable insights.



Variety

The diversity of data sources and formats. Data may come in **structured**, **semi-structured**, or **unstructured** forms, collected from multiple platforms and technologies.

Veracity

The **accuracy, trustworthiness, and quality** of the data. It considers issues like missing values, inconsistencies, and whether the data holds meaningful value.

Value

The potential **usefulness and insights** an organization can extract from the data. This characteristic relates to the **contextual meaning** and strategic decisions that data can support.

Veracity: Dimensions of Data and Information Quality

- These **dimensions of data and information quality** define the criteria by which we **evaluate their quality**:

Completeness

Completeness measures if the data is sufficient to deliver meaningful inferences and decisions.

Accuracy

Data accuracy is the level to which data represents the real-world scenario and confirms with a verifiable source.

Consistency

Consistency measures if the same information stored and used at multiple instances matches.

Validity

Validity signifies that the value attributes are available for aligning with the specific domain or requirement

Uniqueness

Uniqueness indicates if it is a single recorded instance in the data set used and ensures no duplication or overlaps.

Integrity

Integrity indicates that the attributes are maintained correctly, even as data gets stored and used in diverse systems.

Sensitivity of Data

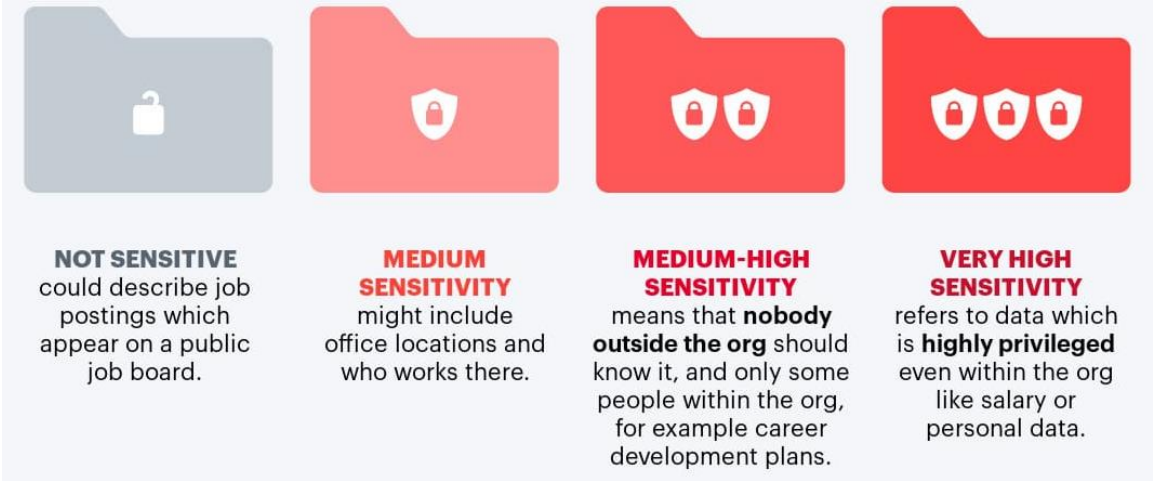
- In addition to those characteristics and quality dimensions, **sensitivity** is another important aspect to consider when **gathering data for analysis or decision-making**.

Full Name e.g., “John Doe”	**Age** e.g., 18 years old
Email Address e.g., “john@mail.com”	**Home Address** e.g., “125 Street Av., CA, USA”
IP Address e.g., “107.118.22.98”	**Social Media Profile** e.g., “@elonmusk”
Phone Number e.g., “+6612 345 6789”	**Salary** e.g., “$26,250/Year”
Citizen ID e.g., “1-2345-67890-12-3”	**Credit Card Number** e.g., “1234 5678 9012 3456”
Date of Birth e.g., “1991-12-26”	**Bank Account Number** e.g., “123-45678-901”
Biometric ID e.g., “3fb69891b552c0…”	**Health/Medical Record** e.g., “2024-12-01 Covid-19”
Affiliation e.g., “CMKL University”	**Geolocation/Position** e.g., “(-77.0364, 38.8951)”

Is this classification true in every case?

How could we identify the need of data protection based on the data attributes?

- (High/Moderate/Low) Sensitivity ?
- Secret/Confidential ?
- Personal or Public ?

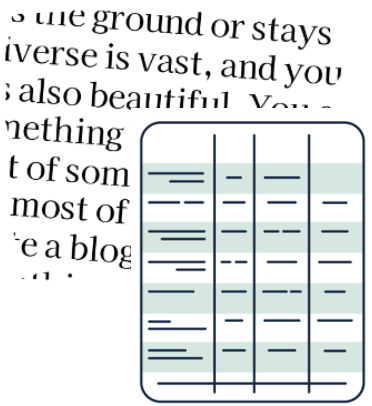


Data Representations

- There are **4 main formats** of data representation that we are currently using in information systems:

Numerical Data

- Computers understand on **1 and 0** (Binary Number System)
- Forming numbers with multiple bits of binaries, e.g., 14 = 1110.
- **Example of Numerical Data:**
  - Age, Phone Number, Passport Number, Credit Card Number

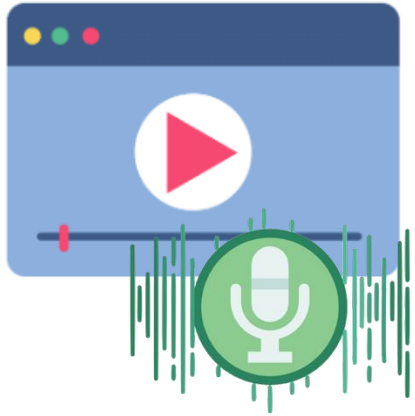


Text/Character Data

- Character data includes letters, symbols, and numerals.
- Character data is encoded as numerals in different schemes, e.g. **ASCII** (7 bits), **UNICODE** (16 bits), **UTF-8** (7 bits).
- **Example of Numerical Data:**
  - Name, Home Address, Log

Image Data

- Image data is represented in the form of **pixels**.
- *RGB* and *CMYK* are well-known schemes where each pixel is represented as a triple or quadruple of color values.
- **E.g.**, (255,255,255) = A White Pixel in RGB scheme.



Video/Audio Data

- Video data is just **a sequence of image data plus audio**.
- Audio data is **a pitch value** of sound at some exact time.
- *Frame-Per-Second (FPS)* is a unit describing how many frames are processed per second.
- **HEVC, MP4, WebM, mJpeg, MKV** are some video encoding schemes.

That is it for today's lecture!

(Let's continue to our lab session)

Homework Assignment

1. Browse to the Github website and download the movie_sample_dataset in the CSV format.
   https://github.com/erajabi/Python_examples/blob/master/movie_sample_dataset.csv
2. Open **Google Collab** and create a new (Jupyter) notebook.
3. Try to upload the dataset into the Google Colab environment.
4. Use a **Pandas** package to read the CSV file and print out the first 10 records, using the following command:

```
# import the pandas library
import pandas as pd

# read a CSV file into a pandas DataFrame
df = pd.read_csv('filename.csv')

# display the first few rows
df.head()
```

End of the Lecture

Please don’t hesitate to raise your hand and ask questions if you're curious about anything!