Third Session

Alireza Moradi

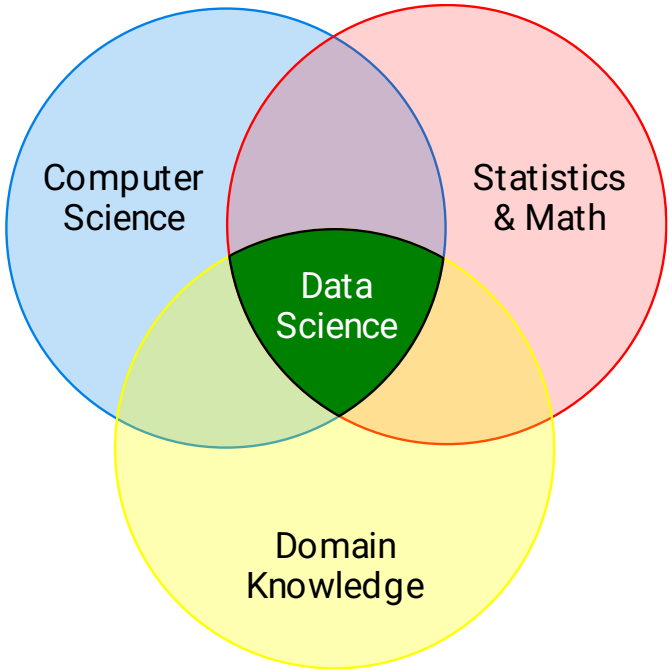
Data Science

- Data Science is the science of **extracting knowledge and insights from data**.

- Data science combines math and statistics, specialized programming, advanced analytics, and artificial intelligence (AI) with specific subject matter expertise to uncover actionable insights hidden in an organization's data. These insights can be used to guide decision making and strategic planning.

- Domain is a well-focused subject area.

- Domain knowledge is the understanding of a specific industry, discipline or activity.

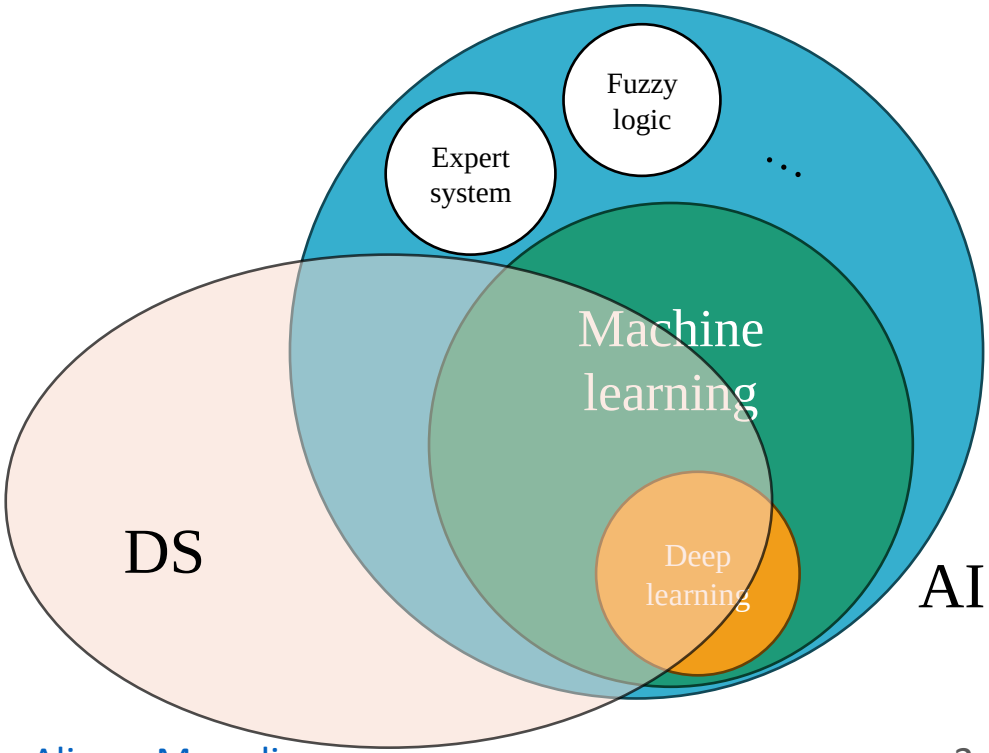


DS vs. AI

- AI aims to create intelligent systems, while Data Science focuses on using data to extract insights and solve problems.

What is the usual output?

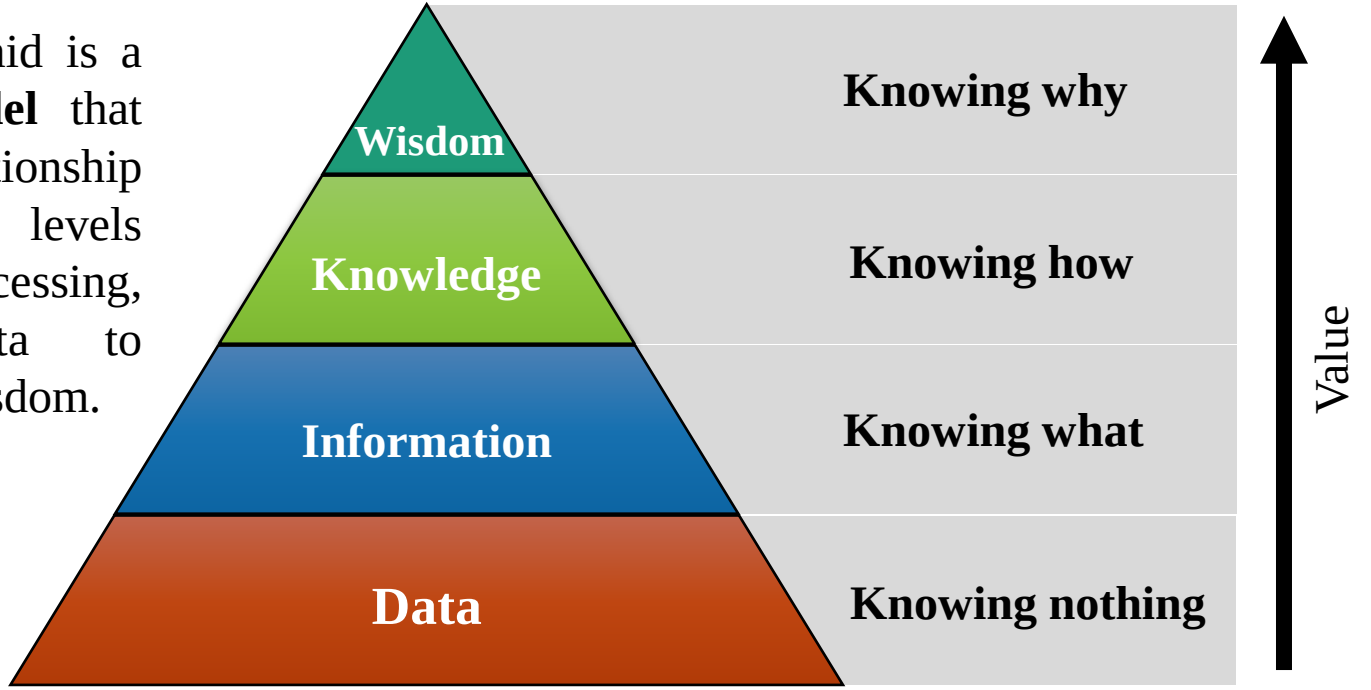
AI project => Software
DS project => Slides, Reports



DIKW pyramid

- DIKW stands for Data, Information, Knowledge, and Wisdom.

- The DIKW pyramid is a **hierarchical model** that represents the relationship between different levels of information processing, from raw data to knowledge and wisdom.



DIKW pyramid; Continue

- The DIKW pyramid is often used in the context of data science to illustrate the **transformation of data into actionable insights**.

1. **Data**: This is the raw, unprocessed facts and figures that are collected from various sources.
2. **Information**: Data becomes information when it is organized, processed, and interpreted in a meaningful way.
3. **Knowledge**: Knowledge is the understanding gained from information, through analysis, interpretation, and synthesis.
4. **Wisdom**: representing the ability to apply knowledge and experience to make sound judgments and decisions.

DIKW pyramid; Continue

- Example:

(1) Imagine you are given a dataset containing a list of temperatures and dates for a city. (2) By organizing the raw temperature data, you can identify the average temperature for each month. (3) By analyzing the monthly temperature averages, you can determine that Augusts are hotter than average, which could be useful for local businesses or government agencies to plan for heatwaves. (4) Based on the knowledge gained from temperature analysis, the city decides to invest in public awareness campaigns to mitigate the impacts of heatwaves.

- By leveraging advanced analytics techniques and technologies, organizations can extract valuable insights from their data and use them to improve their operations, products, and services.

Digital data

- Remind the description of Artificial intelligence:
  “Artificial intelligence is the development and study of computing systems that address a problem associated with some form of intelligence.”

- Computing systems are systems that input, output, process, and store data and information.

- Almost all computing system that we are using now are **digital computers**, so we should know different forms of data in a digital computer.

- If data is "digital", it simply means that it is **discrete data**, and not continuous. In fact, in most contexts, the words "discrete" and "digital" are interchangeable.

- Other computing systems?

Types of data

- We have different types of data:

    1. Structured data (Relational)
    2. Unstructured data (Binary data)
        a) Text data
        b) Image data
        c) Audio data

Object – attribute – value triple

- **Object**: Physical or Abstract items.
  Example for physical => Ball, Car, Student
  Example for abstract => Love, Hate, Mortgage

- **Attribute**: The property or feature of the object.
  Example: color, size, score

- **Value**: Specifies the attribute's assignment(Boolean, Numeric, or String).
  Example: Red, Large, 20, yes

Structured data(Relational)

- Object – attribute – value triple are usually shown in a table form.
- Here we have n object and p attribute.

	Attribute 1	Attribute 2	…	Attribute (p-1)	Attribute p
Object 1	$value_{1,1}$	$value_{1,2}$	…	$value_{1,(p-1)}$	$value_{1,p}$
Object 2	$value_{2,1}$	$value_{2,2}$	…	$value_{2,(p-1)}$	$value_{2,p}$
…	…	…	…	…	…
Object n	$value_{n,1}$	$value_{n,2}$	…	$value_{n,(p-1)}$	$value_{n,p}$

Structured data(Relational); Continue

- Let’s see an example:

Attribute, Dimension, Feature, Variable

	Name	Age	Female?	Score
Student 01	Ali	21	No	17
Student 02	Ahmad	21	No	19
Student 03	Zahra	22	Yes	15
Student 04	Sorosh	19	Yes	20
Student 05	Mehdi	20	No	19
Student 06	Negin	22	Yes	20

Data Object, Sample, Instance, Data point, Object, Record, index

Structured data(Relational); Continue

- As we saw, structured data conforms to a **data model** or schema and is often stored in tabular form.

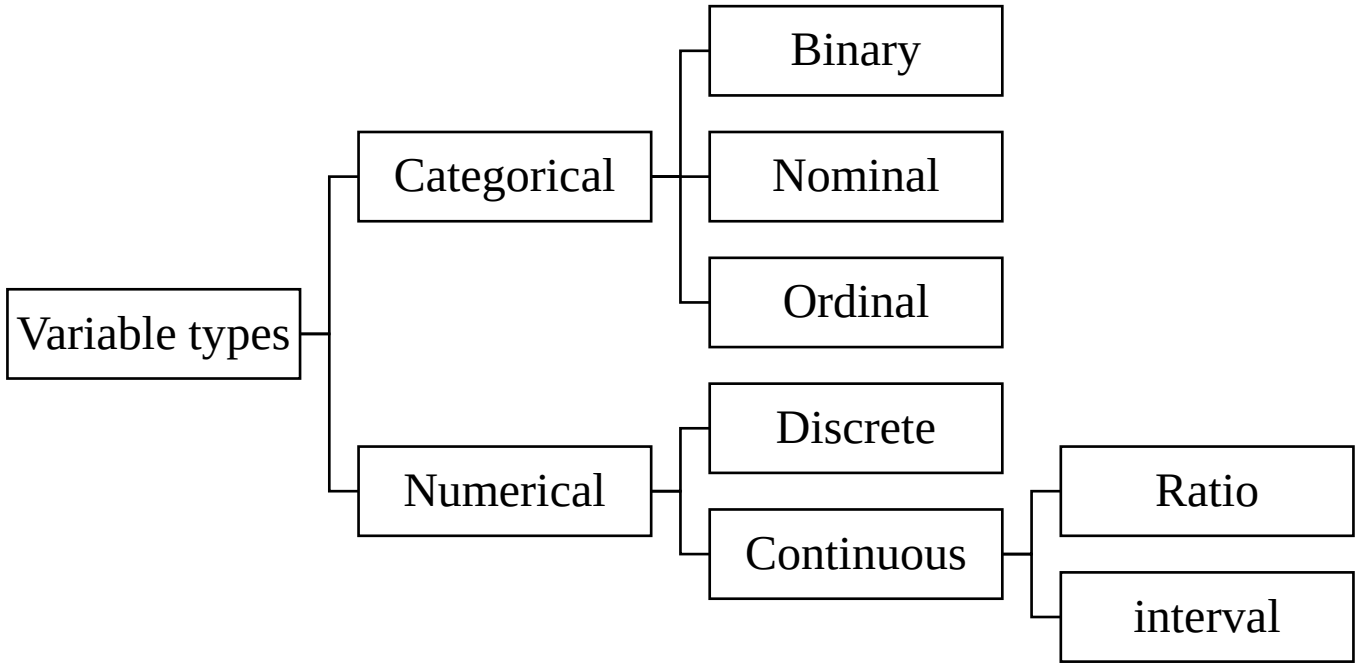
- What do we mean by relational?
  Structured data is used to capture relationships between different entities and is therefore most often stored in a relational database.

Student ID	Name	Course
4010034	Ali	AI
4010035	Matin	DS
4010037	Maryam	AI

Course	Faculty	Time	Level
AI	C.E.	11-13	Intermediate
DS	I.E.	15-17	Advanced

Different variables

- Qualitative => Categorical
- Quantitative => Numerical



Categorical variable; Binary

- A **categorical variable** is a variable that can take on one of a limited, and usually fixed, number of possible values

---

- A binary variable is a variable that has two possible outcomes.

- It can be the outcome of an experiment ("success" or "failure") the response to a yes–no question ("yes" or "no") presence or absence of some feature ("is present" or "is not present").

- Examples: Gender

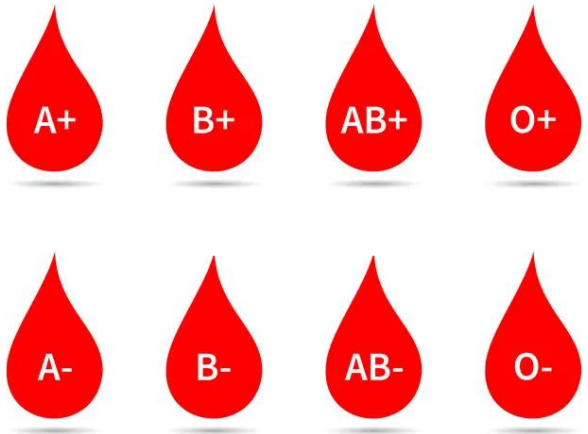
- **Preferred form is 0 and 1**, so that a digital computer can easily understand.

Categorical variable; Nominal

- A purely nominal variable is one that simply allows you to assign categories, but you **cannot clearly order the categories**.

- Binary variables is a special form of nominal variables. (**2 categories that can not be ordered**)

- Examples: blood type, zip code, race, eye color, political party



Categorical variable; Ordinal

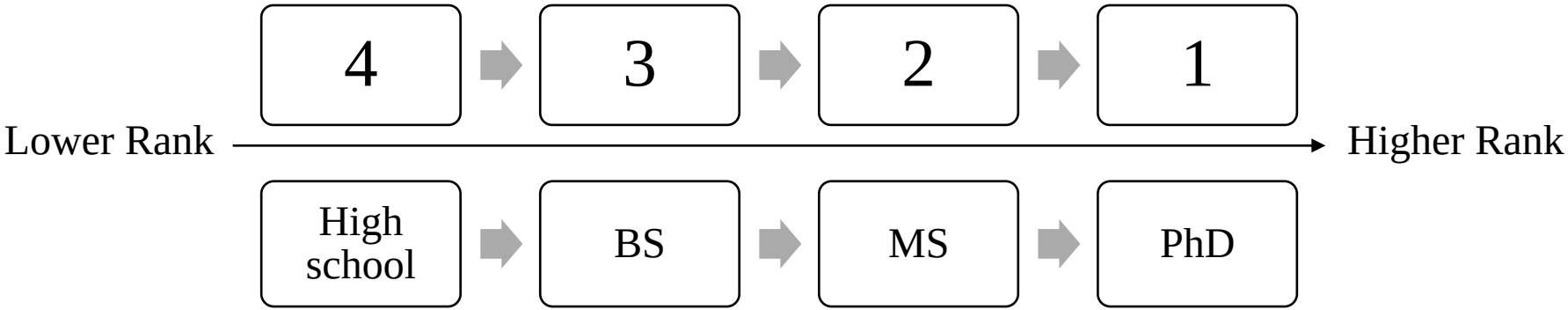
- An ordinal variable is like a categorical variable. The difference between the two is that **there is a clear ordering of the categories**.

- Order => Rank => sequence

- Examples:
  - ✓ Socio-economic status (“low income", "middle income”, ”high income”)
  - ✓ Education level (“high school”, ”BS”, ”MS”, ”PhD”)
  - ✓ Income level (“less than 50K”, “50K-100K”, “over 100K”)
  - ✓ Satisfaction rating (“extremely dislike”, “dislike”, “neutral”, “like”, “extremely like”).

Categorical variable; Ordinal

- Ordinal data provide sequence, and it is possible to assign numbers to the data.
- No numeric operations can be performed.



Numerical variable

- A numerical variable is a quantifiable characteristic whose values are numbers (except numbers which are codes standing up for categories).

- Numeric operations can be performed.

- There are two sub-groups; Discrete variables and continuous variables

- **Discrete variables** can only take on a limited number of values (e.g., only Integer numbers) while **continuous variables** can take on any value and any value between two values (e.g., out to an infinite number of decimal places).

➢ Measuring => Continuous
➢ Counting => Discrete

Numerical variable; Continue

- **Discrete variables** are numeric variables that have a countable number of values between any two values.

- Example: the number of customer complaints or the number of defects.

- Continuous variables have two sub-groups; (1) **Ratio** data has a defined zero point, whereas (2) **interval data** lacks the absolute zero point.

- Example for Ratio
✓ Height
✓ Age
✓ Temperature in Kelvin

- Example for interval
✓ Year
✓ Time
✓ Temperature in Centigrade

Levels of measurement

- Levels of measurement, also called scales of measurement, tell you **how precisely variables are recorded**.

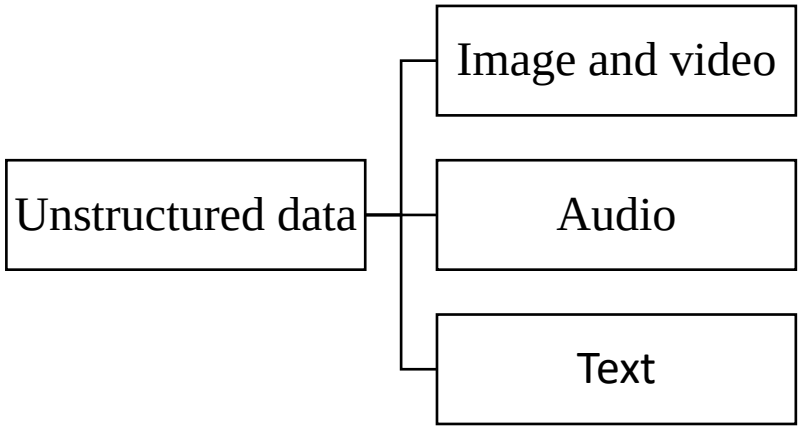
Level	Name	Description	example
1	Nominal	The data can only be categorized.	blood type
2	Ordinal	the data can be categorized and ranked.	Education level
3	Interval	the data can be categorized, ranked, and evenly spaced.	Temperature in Centigrade
4	Ratio	the data can be categorized, ranked, evenly spaced, and has a natural zero.	Temperature in Kelvin

Unstructured data

- Data that **does not conform to a data model** or data schema is known as unstructured data.

- It is estimated that unstructured data makes up 80% of the data within any given enterprise.

- Unstructured data has a **faster growth rate** than structured data.



Unstructured data; Image

- A digital image consists of many small spots of color. These spots are called **pixels**.

- When displayed on a monitor or printed on paper, pixels are so small and so closely packed together that the collective effect on the human eye is a continuous pattern of colors.

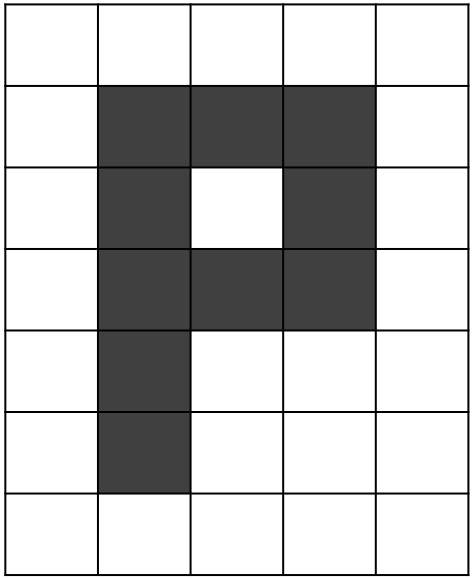
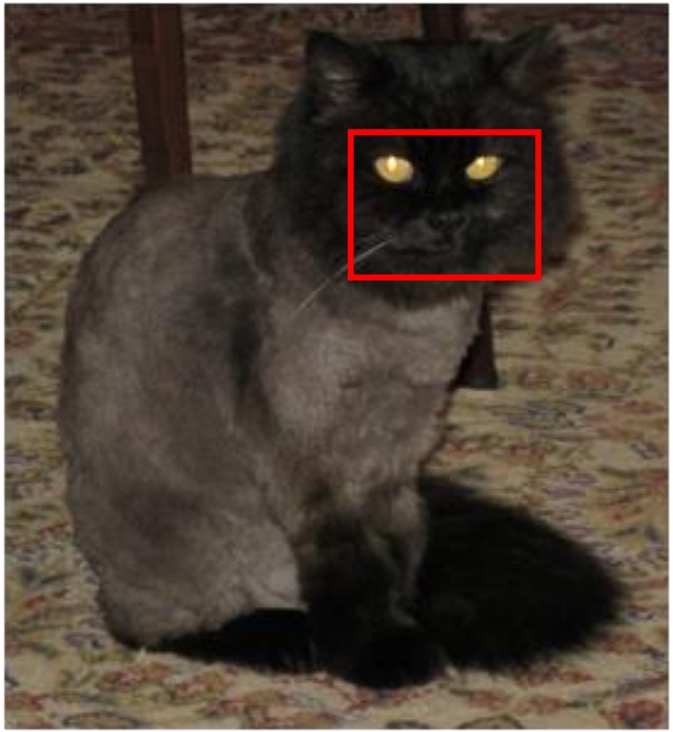


Image of letter P

0	0	0	0	0
0	**1**	**1**	**1**	0
0	**1**	0	**1**	0
0	**1**	**1**	**1**	0
0	**1**	0	0	0
0	**1**	0	0	0
0	0	0	0	0

Image representation

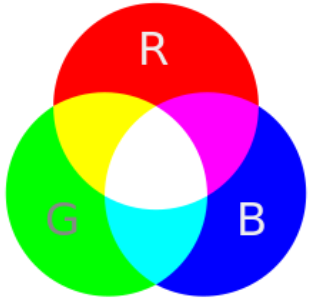
Unstructured data; Image



Unstructured data; Image

- Computers represent the color of a pixel using a trick long known to artists: **mixing primary colors.** The most popular system is RGB.

- **RGB (red, green and blue) system:**
  Red, green and blue can be combined in various proportions to obtain any color in the visible spectrum.



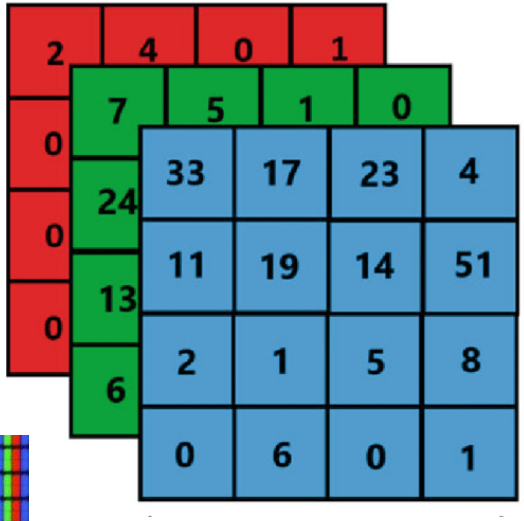
- The RGB system uses 8 bits each for red, green and blue colors. So, each color has values ranging from 0 to 255. This translates into **16,777,216 possible colors** to be precise.

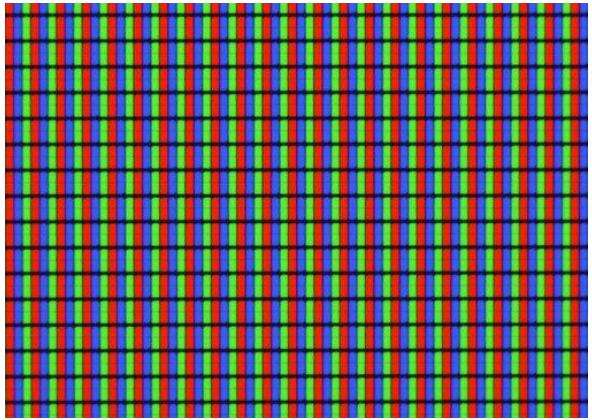
1	0	0	1	1	1	0	0

= 156

Unstructured data; Image

- The 3 colors are also called **channels**.
- So, we store 3 matrices for storing an image.
  $R_{(height \times width)}$, $G_{(height \times width)}$, $B_{(height \times width)}$

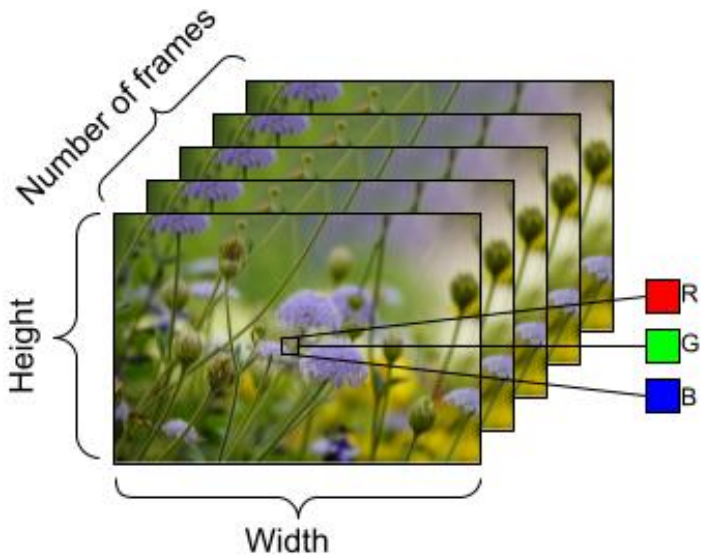




Unstructured data; Video

- Digital video is comprised of a series of images, called **frames**, quickly changing from one to the next.

- This rapidly progressing sequence of images produces the illusion of motion that we perceive and call video.

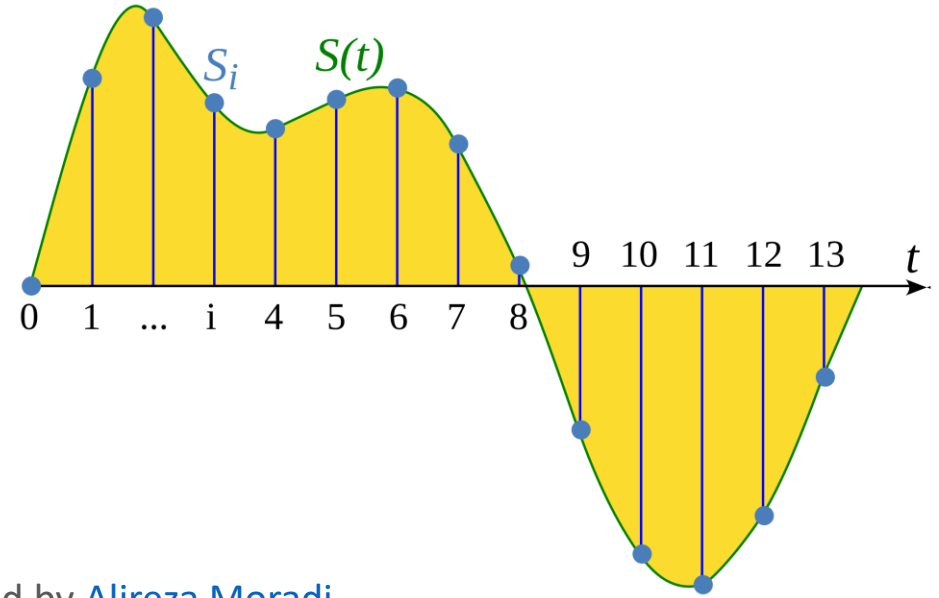


Unstructured data; Audio

- By nature, **a sound wave is a continuous signal**, meaning it contains an infinite number of signal values in each time. To be processed, stored, and transmitted by digital devices, the continuous sound wave needs to be converted into a series of discrete values, known as a digital representation.

- The **sampling rate** (also called **sampling frequency**) is the number of samples taken in one second and is measured in **hertz (Hz).**

- MP3 files => 44100 Hz



Unstructured data; Audio

- Sound is made by changes in air pressure at frequencies that are audible to humans. The **amplitude** of a sound describes the sound pressure level at any given instant and is measured in decibels (dB).

- We perceive the amplitude as loudness. To give you an example, a normal speaking voice is under 60 dB, and a rock concert can be at around 125 dB, pushing the limits of human hearing.

- The **bit depth** of the sample determines with how much precision this amplitude value can be described.

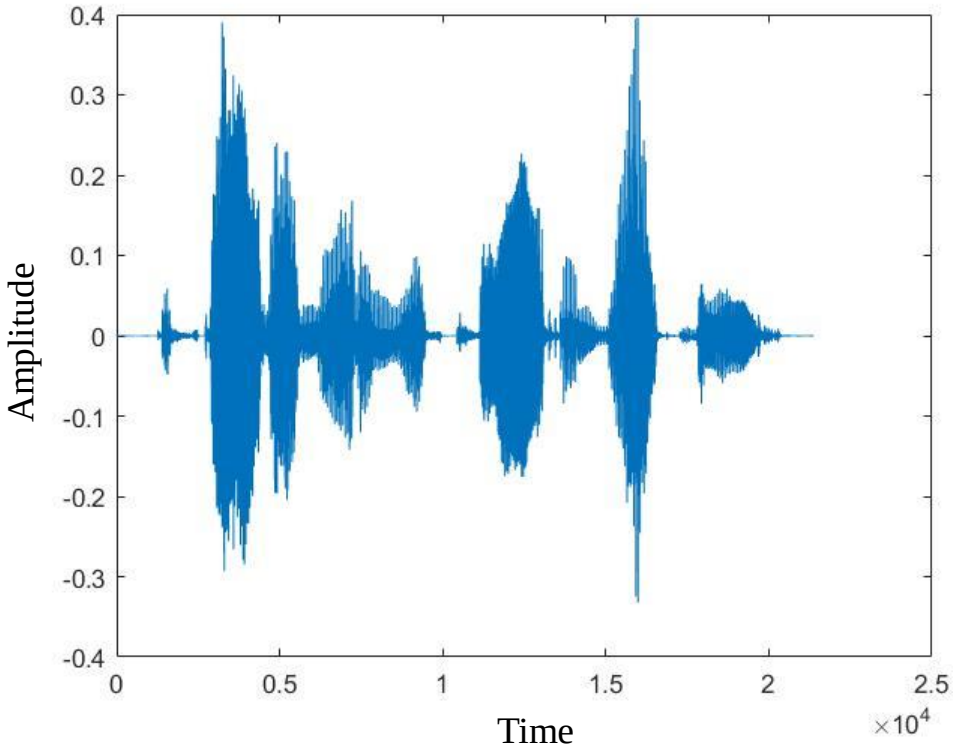
- The most common audio bit depths are 16-bit and 24-bit.

Unstructured data; Audio

- The **time domain representation of sound** involves capturing and analyzing these pressure variations at different time intervals by sampling the sound wave at discrete points in time.

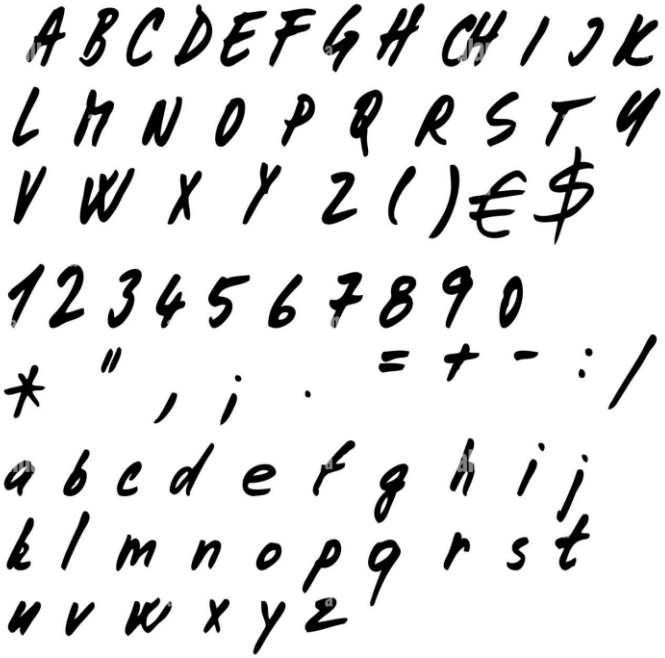
✓ As you see, audio is **a sequence indexed in time**.

✓ In a sequence order is important.



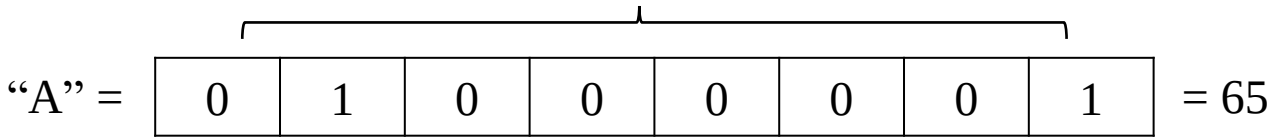
Unstructured data; Text

- In all computers, alpha-numeric data and special characters (i.e., letters, numbers and symbols) are each assigned a specific binary value, called a **character code**.



- **ASCII** (American Standard for Information Interchange) is a coding system for representing characters.

- Example:



Unstructured data; Text

- **Tokenization**, in the realm of Natural Language Processing (NLP) and machine learning, refers to the process of converting a sequence of text into smaller parts, known as **tokens**.

- These tokens can be as small as characters or as long as words. Many language models use word-sized tokens as the basic unit of text processing.

- **Text is also a sequence**. For example, “I love that orange cat” is: