LEVEL 1: Data Literacy

Day 2

COURSE OUTLINE

- Definition of Data
- Types of Data
- Data Representation
- Sources of Data
- Data Cleaning

LEARNING OBJECTIVES

- Understand the formal definition of data and its role in analytics and machine learning
- Identify and differentiate between major data types used in ML
- Describe how data can be represented, structured, and encoded for computational use
- Recognize common sources of data across real-world contexts and industries
- Explain the importance of data cleaning and evaluate basic strategies for improving data quality

Data Literacy

Introduction

What is data?

Data Literacy

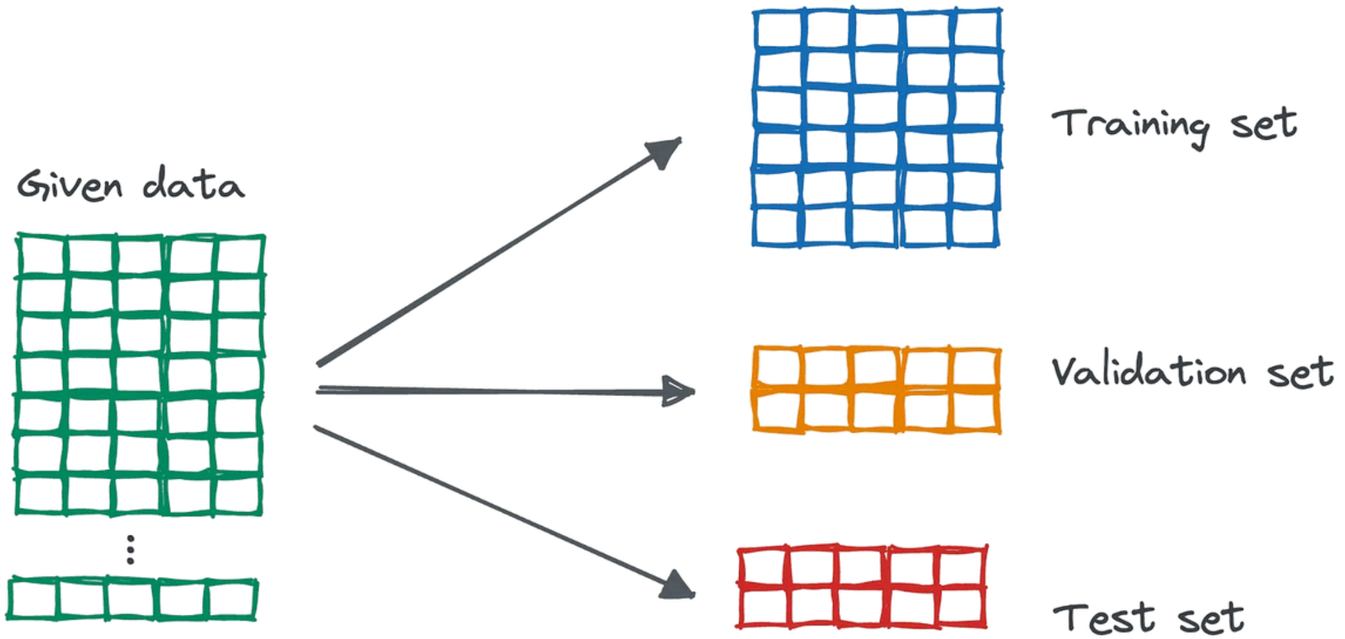
- Data is a collection of raw facts and observations.

Data Literacy

- Data is the fuel of Machine Learning models.

Data Literacy

- Data used to train, validate, and test a machine-learning model.

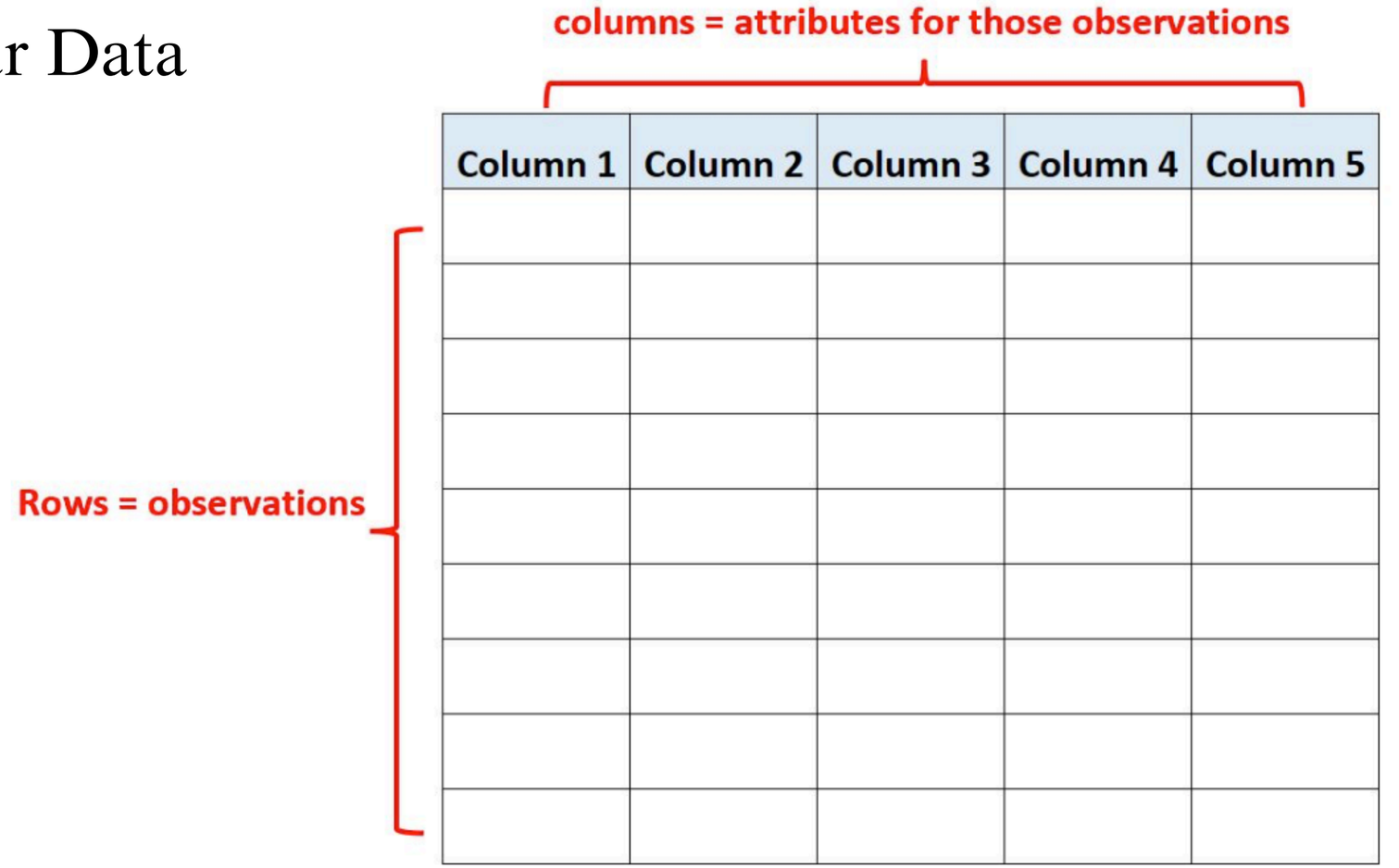


Data Literacy

Types of Data

Types of Data

- Tabular Data



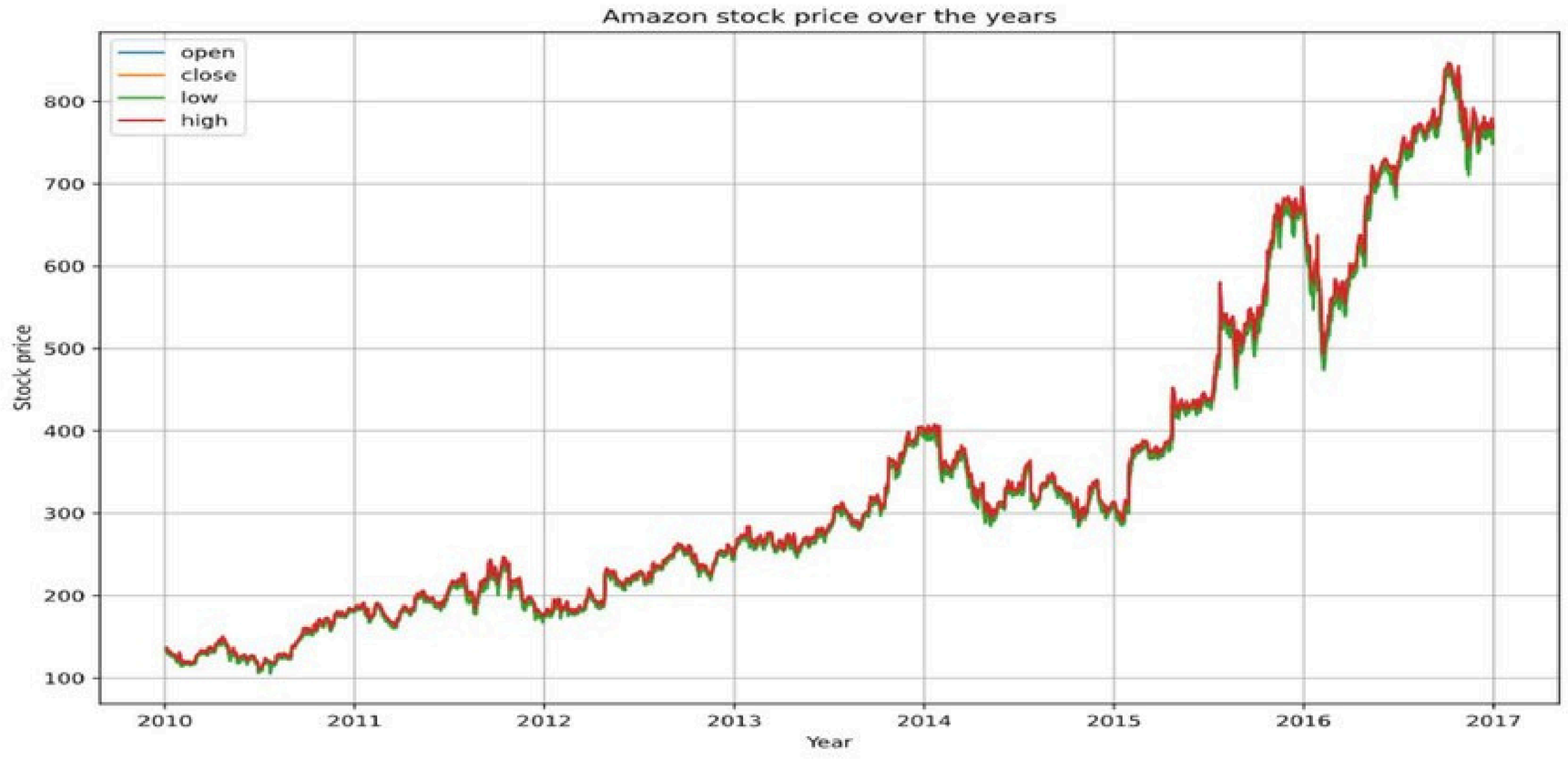
Types of Data

- Tabular Data
  - It is the most common data format in ML.
  - Works well with many traditional ML algorithms.

Size (m^2)	Rooms	Type	Parking	Price (SAR)
120	3	Apartment	1	950,000
180	4	Villa	2	1,350,000
140	3	Apartment	1	980,000
220	5	Villa	3	1,850,000
160	4	Townhouse	2	1,200,000

Types of Data

- Time-Series Data



Types of Data

- Time-Series Data

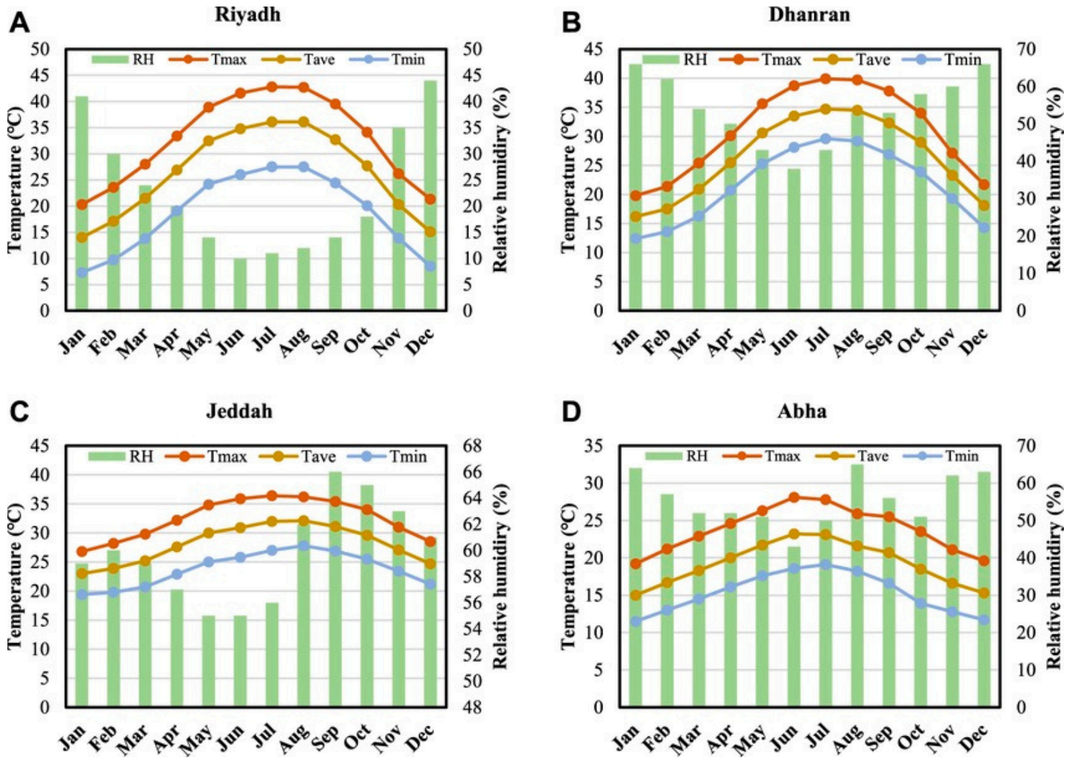
- Data collected over time in sequence.
- Order matters as past affects the future.

Date	Open Price (SAR)	Close Price (SAR)	High	Low	Volume
Jan 1, 2025	128.5	130.1	131.2	127.8	1.2M
Jan 2, 2025	130.15	129.3	131	128.5	1.0M
Jan 3, 2025	129.4	132.6	133.1	129.1	1.5M
Jan 4, 2025	132.7	131.8	134.2	131.1	1.3M
Jan 5, 2025	131.9	135.5	136	131.4	1.6M

Types of Data

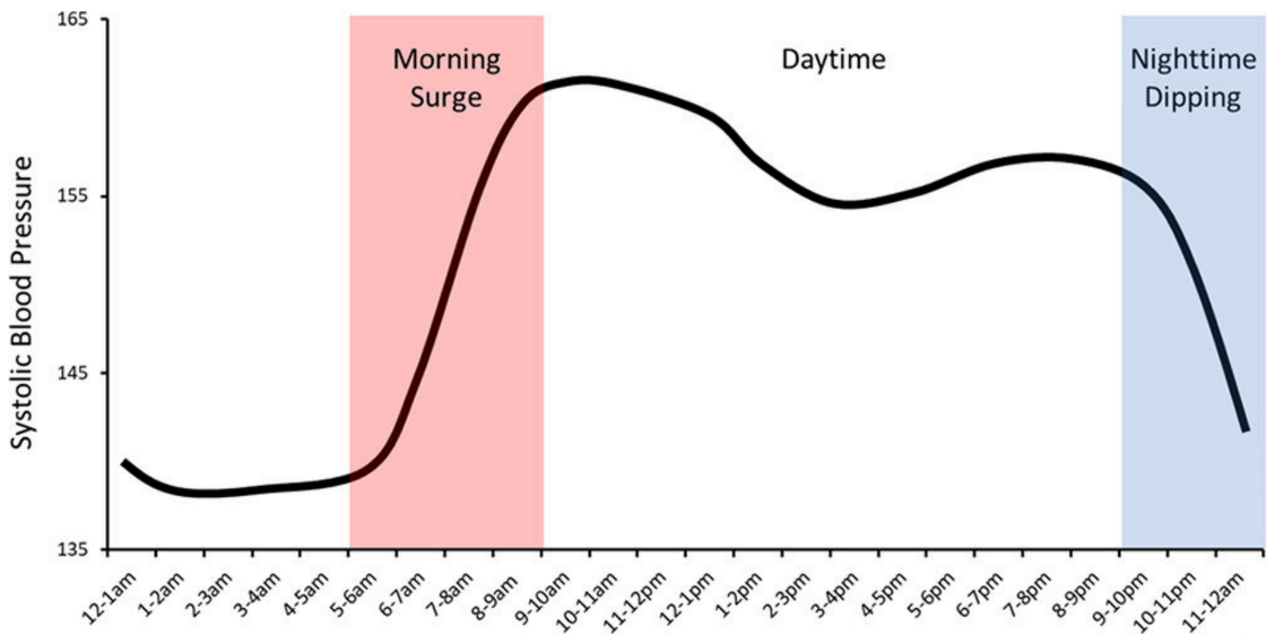
- Time-Series Data
  - Stock-market

  - Daily temperature

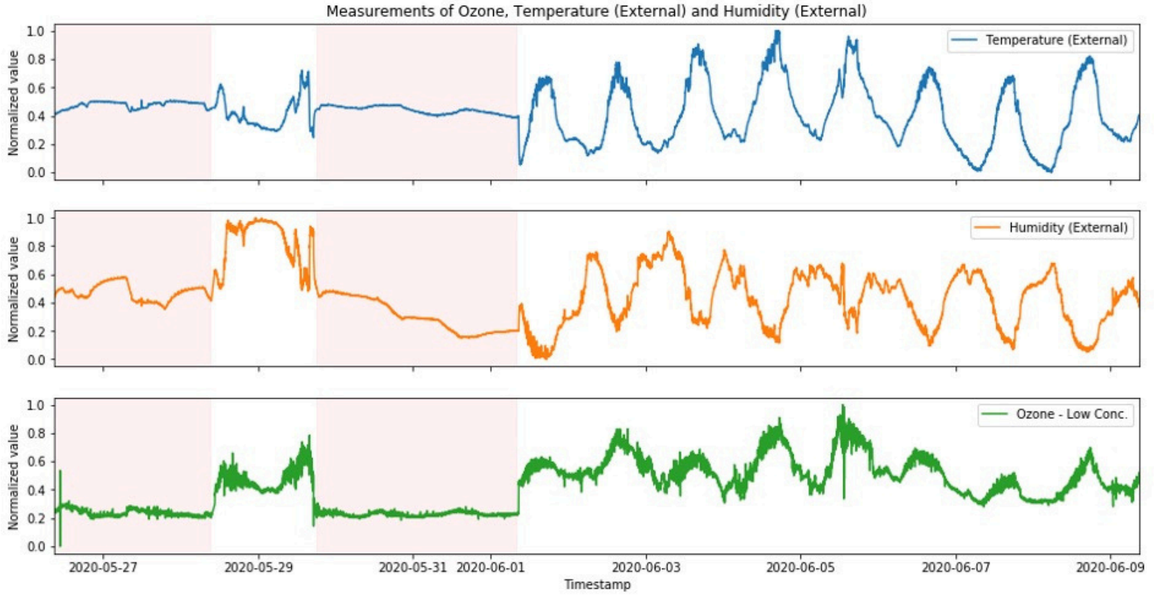


Types of Data

- Time-Series Data
  - Blood pressure



  - Sensors Data



Types of Data

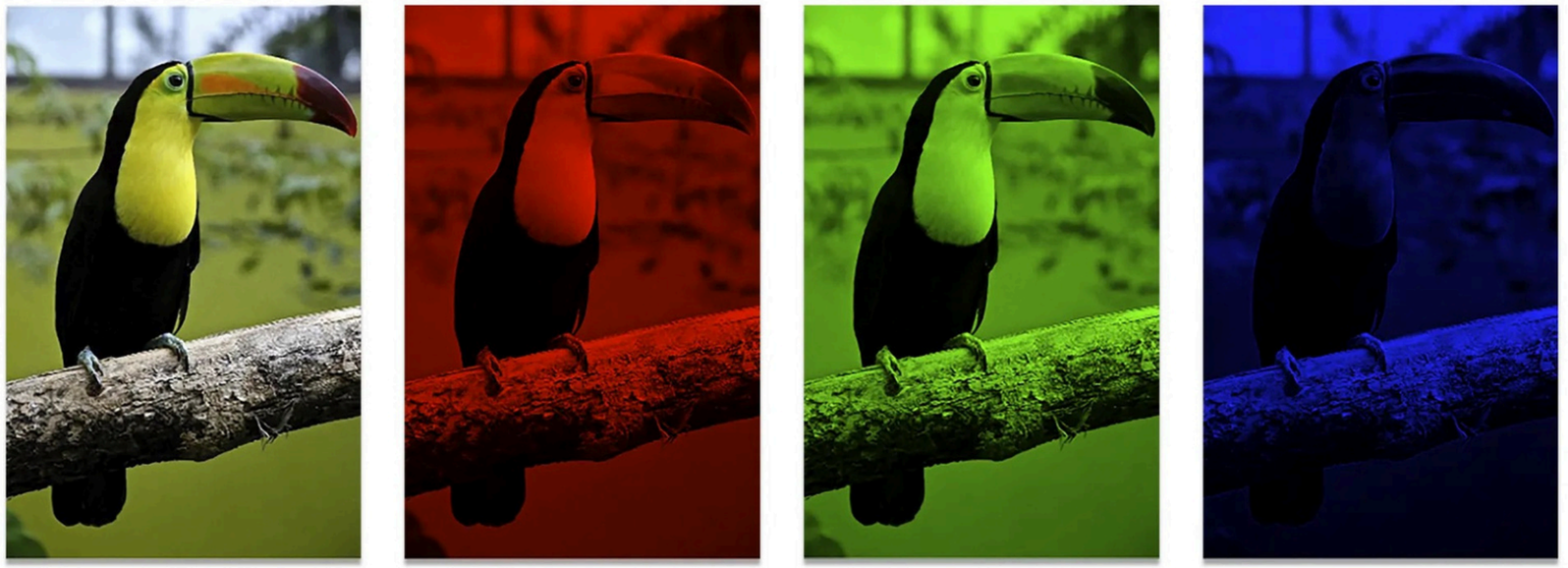
- Image Data

Types of Data

- Image Data

Images are composed of pixels, each encoding specific color and intensity information.

- Colored Images

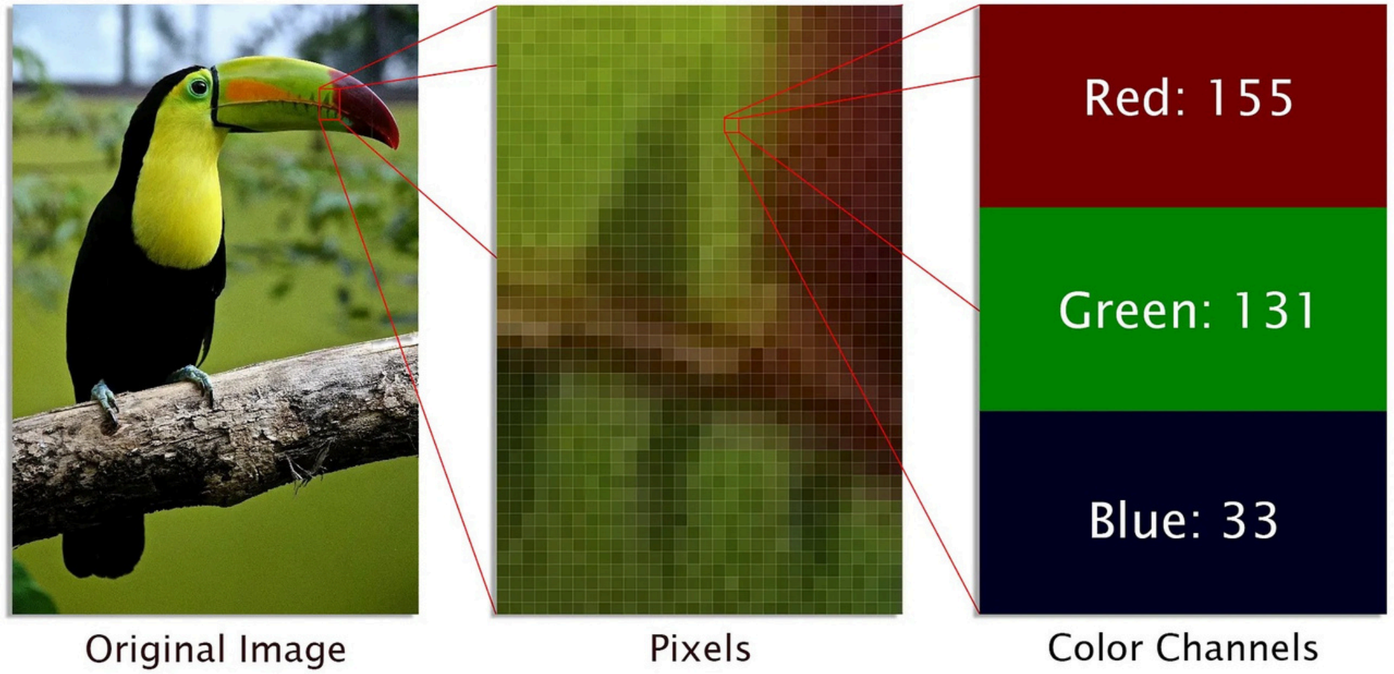


Types of Data

- Image Data

Images are composed of pixels, each encoding specific color and intensity information.

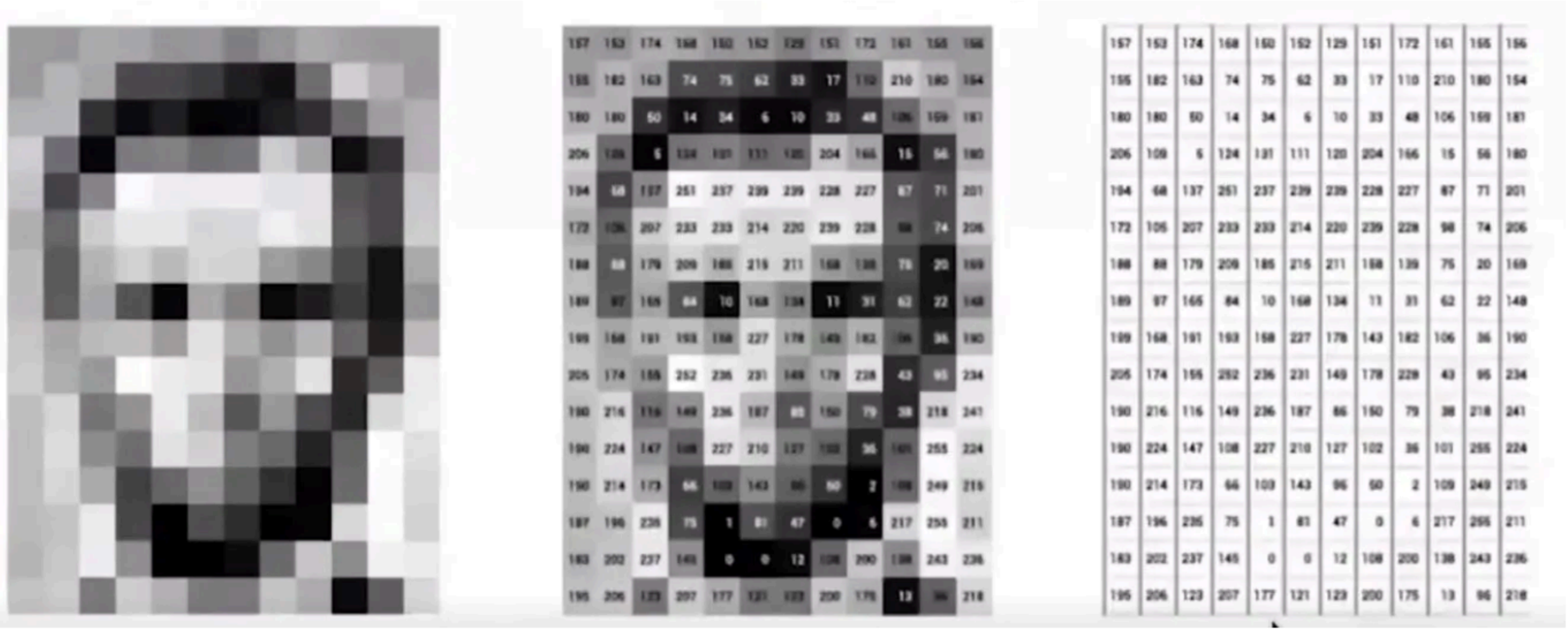
- Colored Images



Types of Data

- Image Data

- Black\white
- Grey images

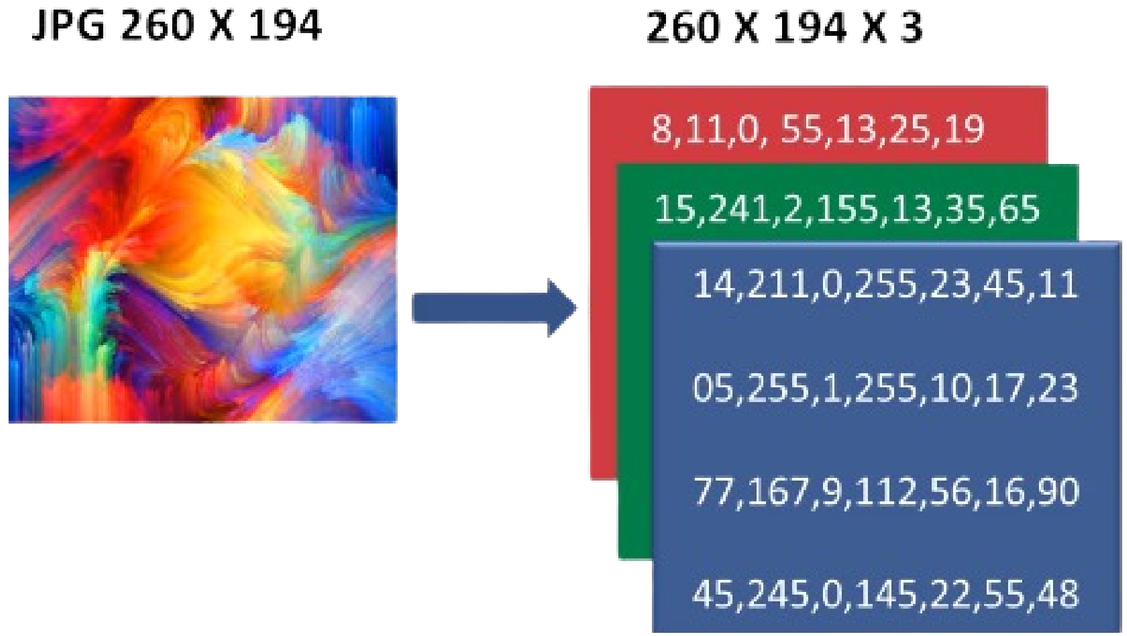


Types of Data

Tabular = structured rows & columns

Size (m^2)	Rooms	Type	Parking	Price (SAR)
120	3	Apartment	1	950,000
180	4	Villa	2	1,350,000
140	3	Apartment	1	980,000
220	5	Villa	3	1,850,000
160	4	Townhouse	2	1,200,000

Image = unstructured pixel grid

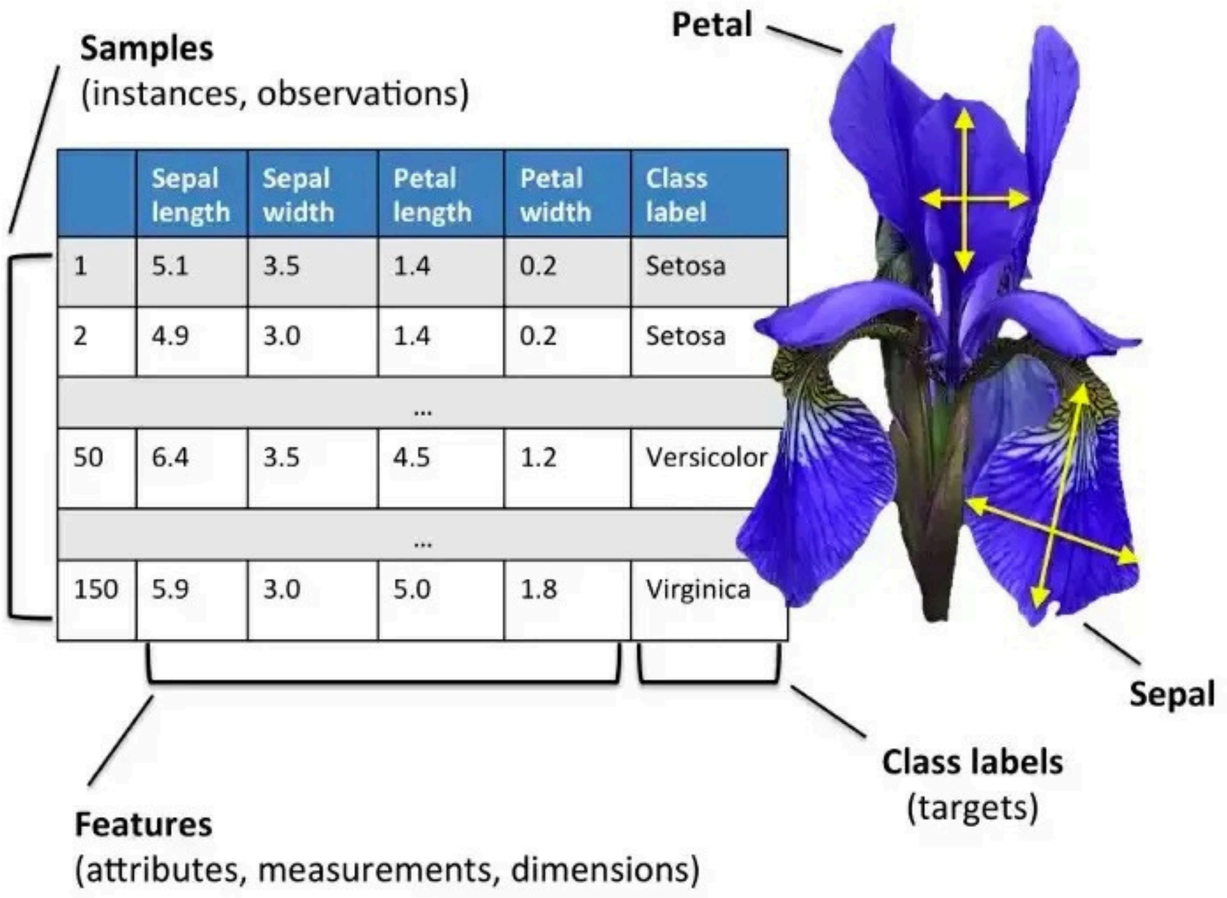


Types of Data

- How to classify these flowers?

Types of Data

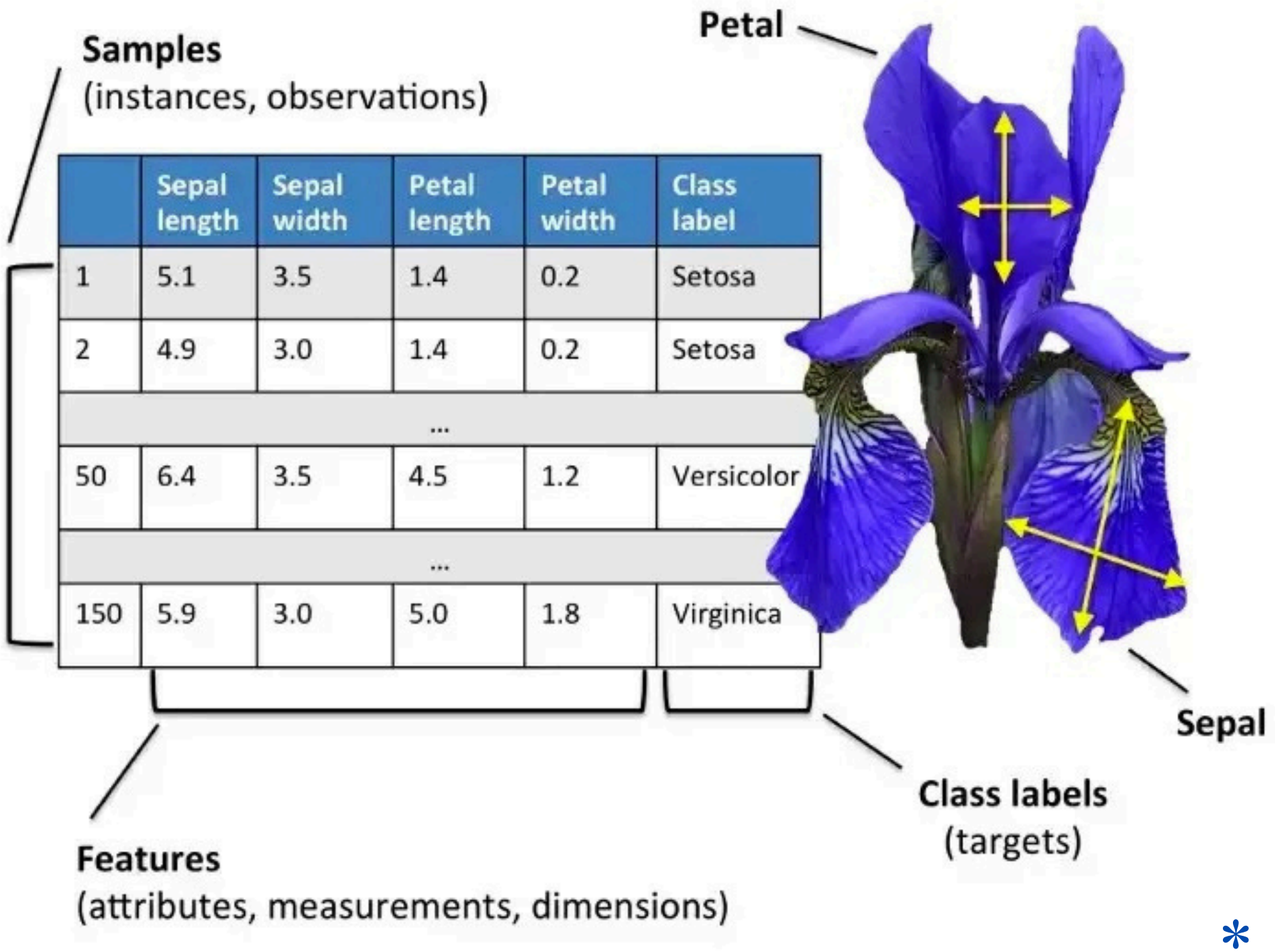
- Traditional tabular data



	Sepal length	Sepal width	Petal length	Petal width	Class label
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
...					
50	6.4	3.5	4.5	1.2	Versicolor
...					
150	5.9	3.0	5.0	1.8	Virginica

Types of Data

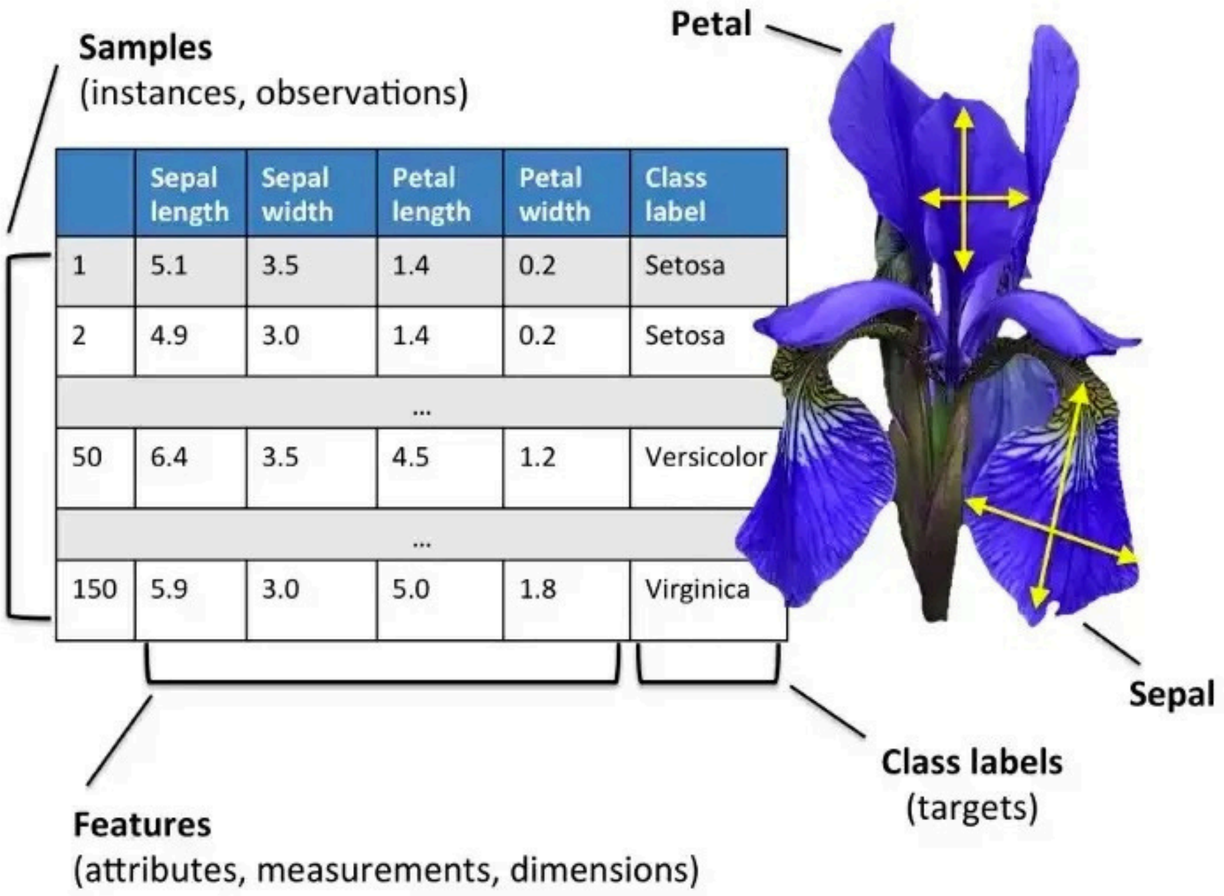
- Traditional tabular data



	Sepal length	Sepal width	Petal length	Petal width	Class label
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
			...		
50	6.4	3.5	4.5	1.2	Versicolor
			...		
150	5.9	3.0	5.0	1.8	Virginica

What's the problem?

Types of Data



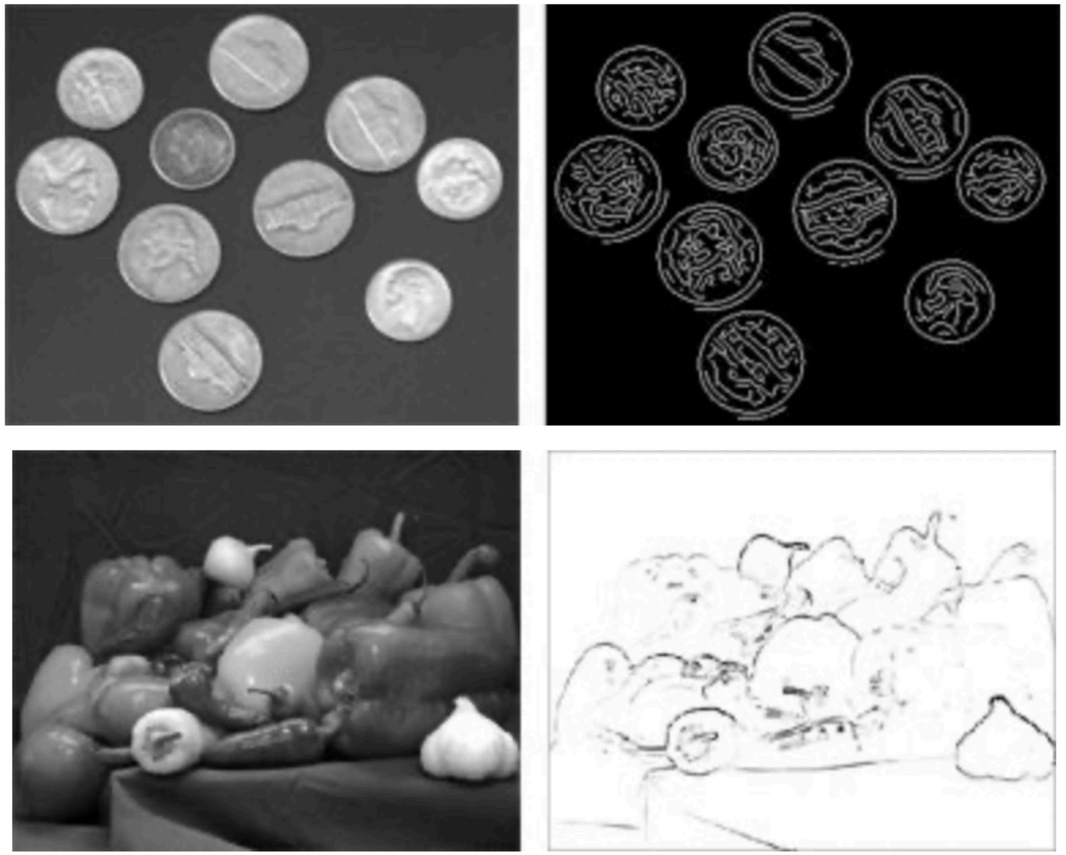
	Sepal length	Sepal width	Petal length	Petal width	Class label
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
...					
50	6.4	3.5	4.5	1.2	Versicolor
...					
150	5.9	3.0	5.0	1.8	Virginica

- Human-designed features capture only simple, linear patterns.

Types of Data

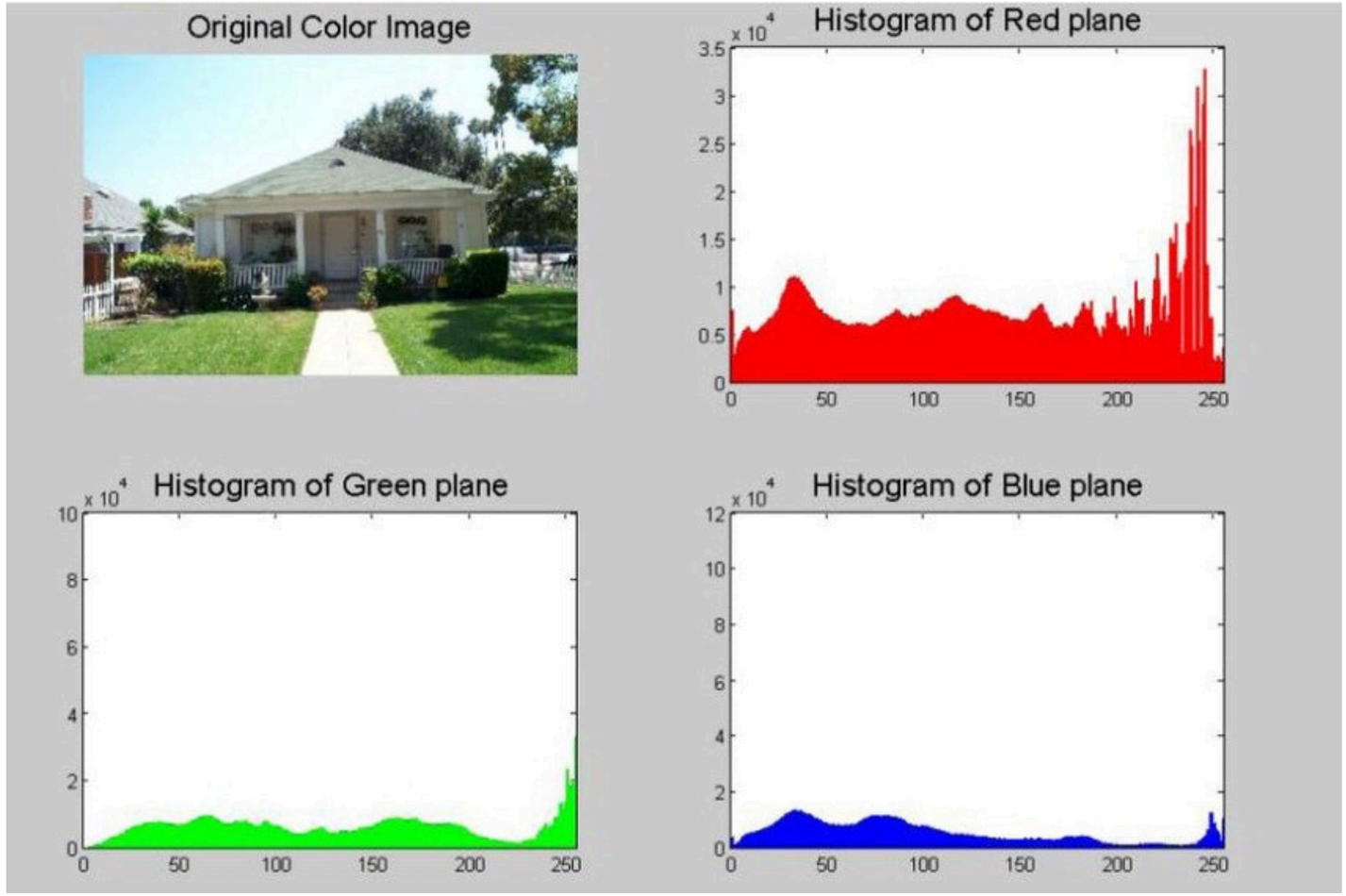
- In Machine Learning, images are first transformed into manually engineered features.

- Edges



Types of Data

- Color histograms



Types of Data

512 × 512 pixels

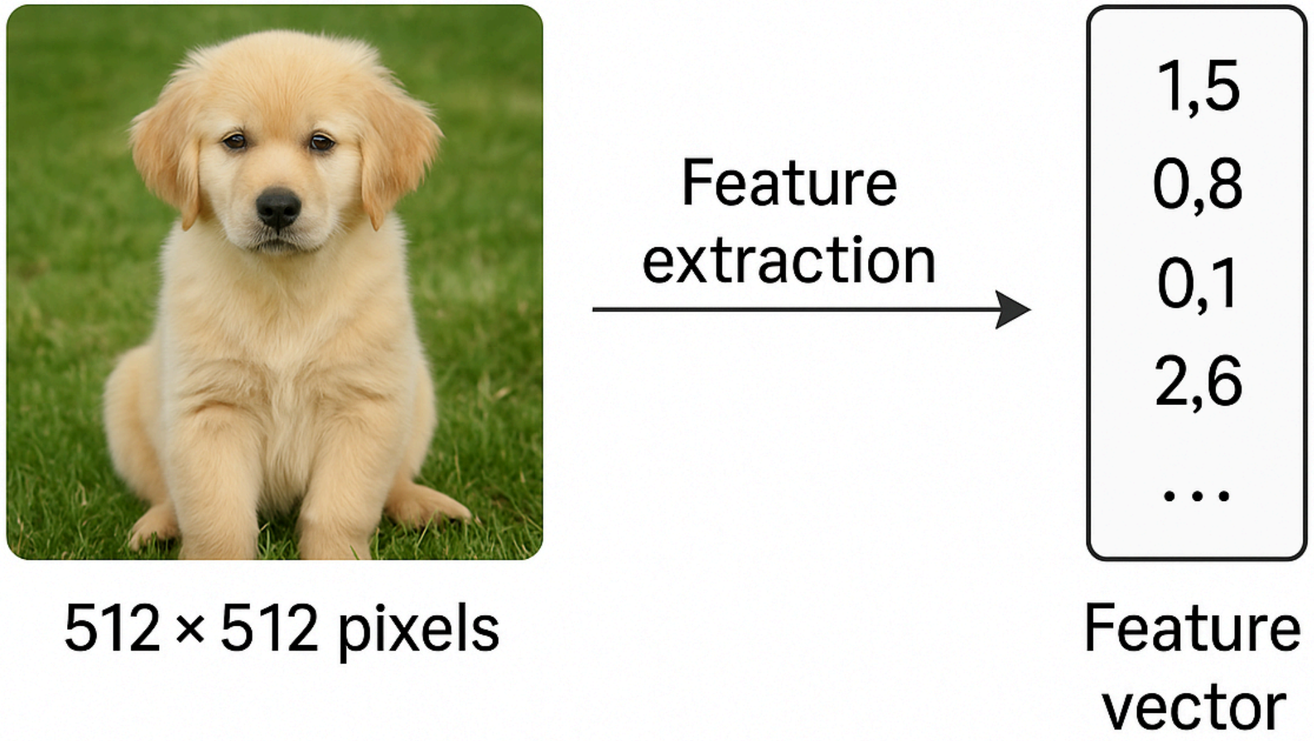
Feature extraction

1,5
0,8
0,1
2,6
...

Feature vector

What's the problem here?

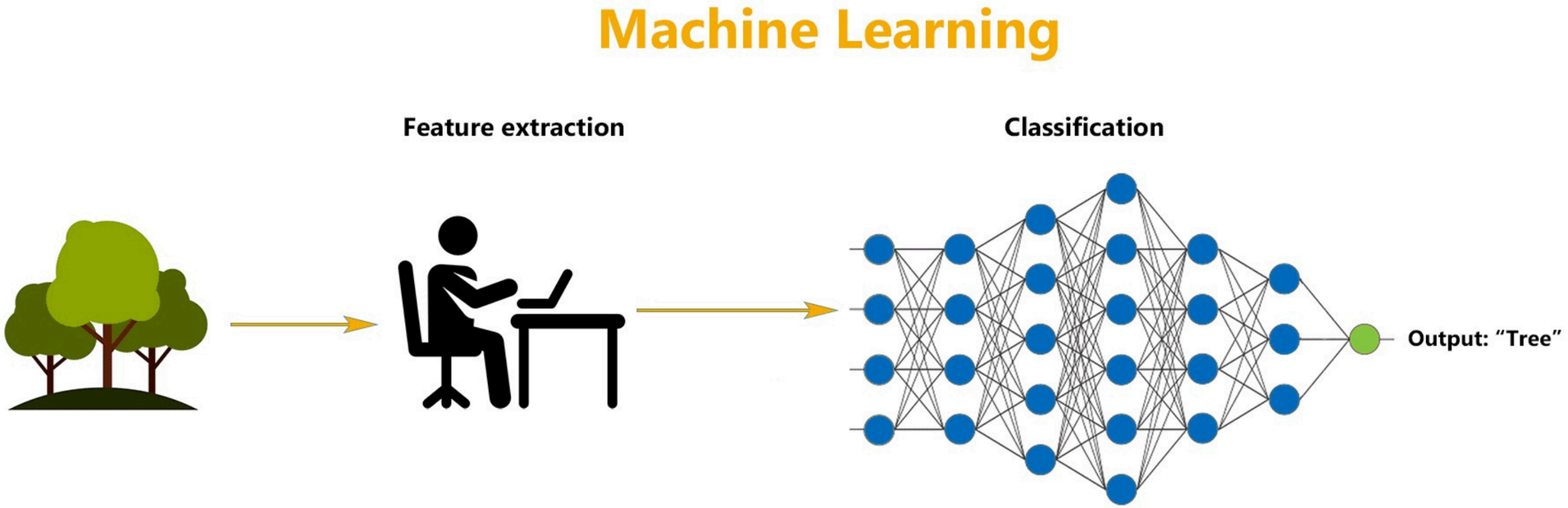
Types of Data



- Most raw data information is lost when reduced to a few features.

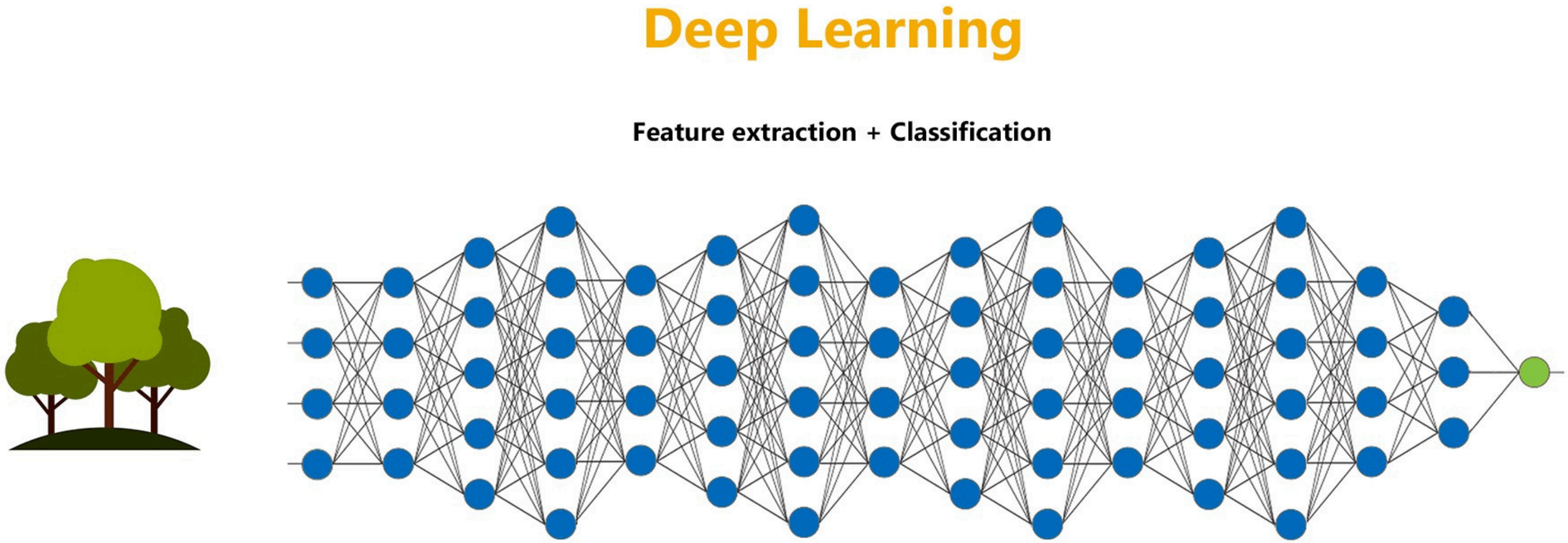
Types of Data

- Traditional ML needs manually extracted features, while deep learning learns features directly from images automatically.



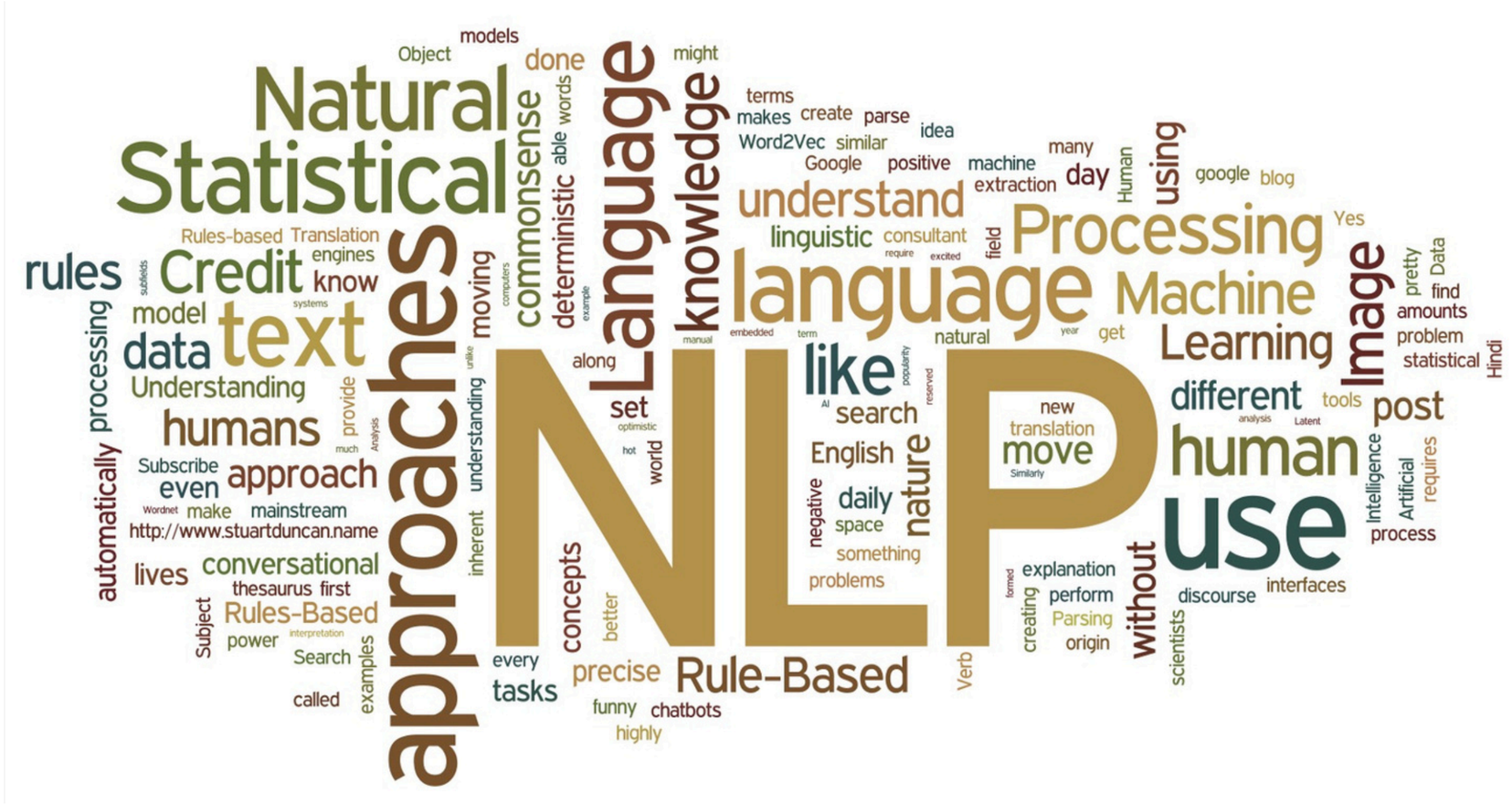
Types of Data

- Traditional ML needs manually extracted features, while deep learning learns features directly from images automatically.



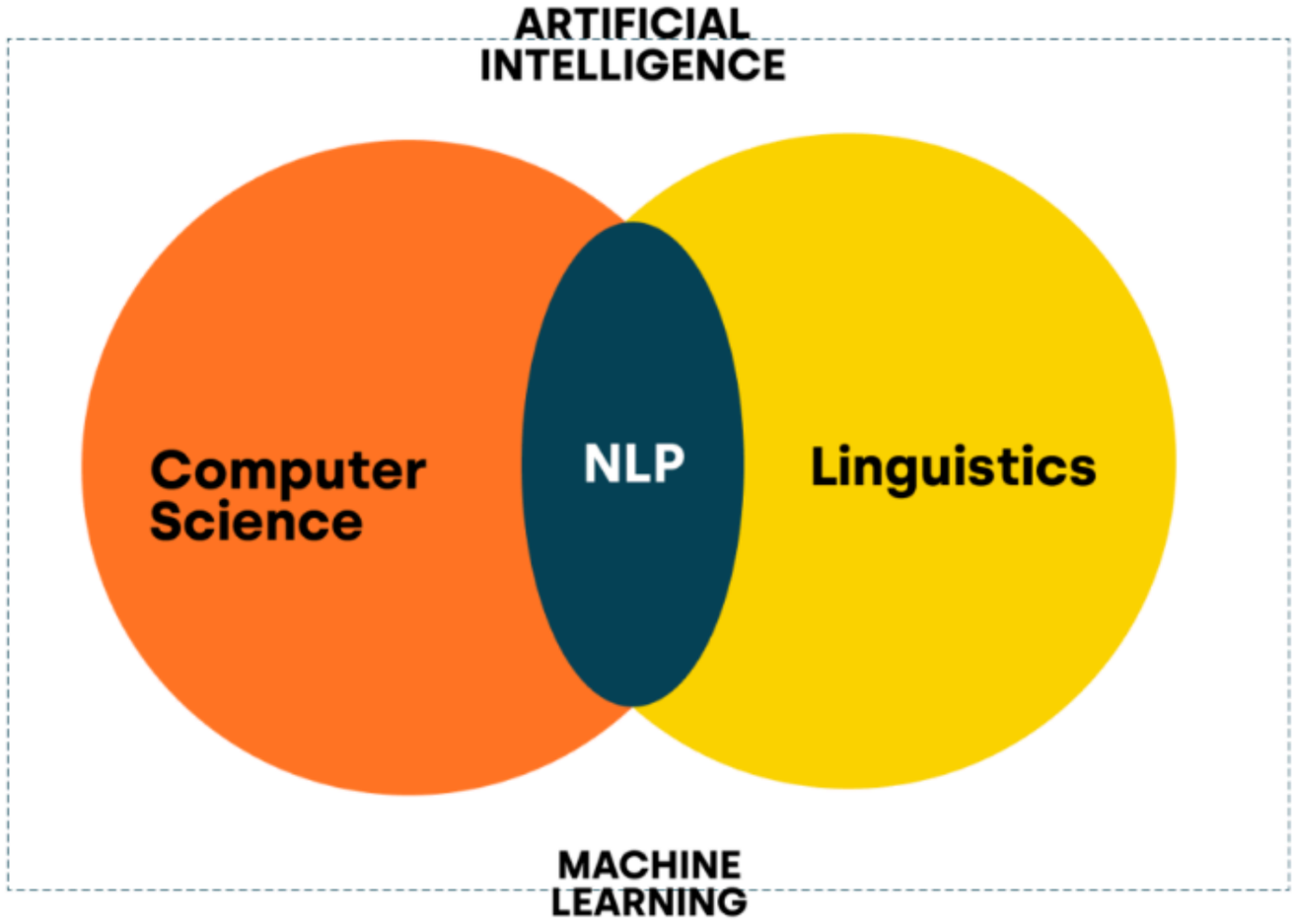
Types of Data

- Textual Data



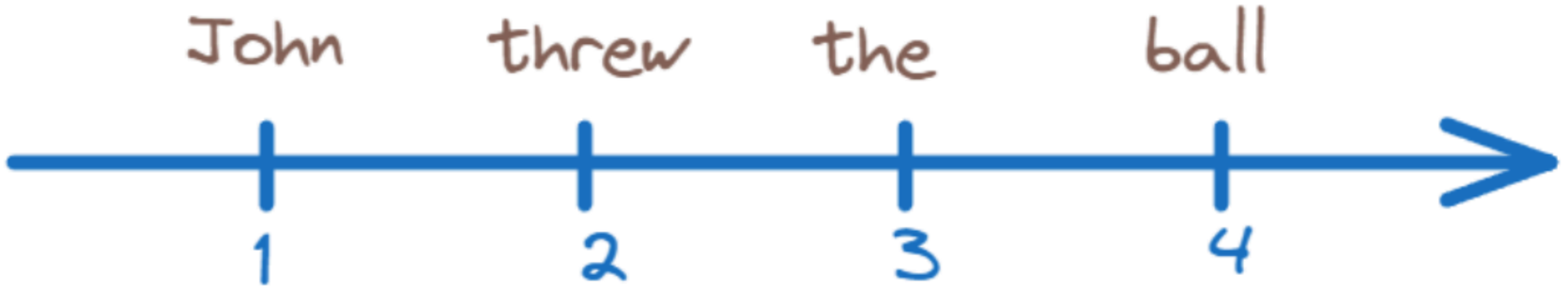
Types of Data

- Textual Data



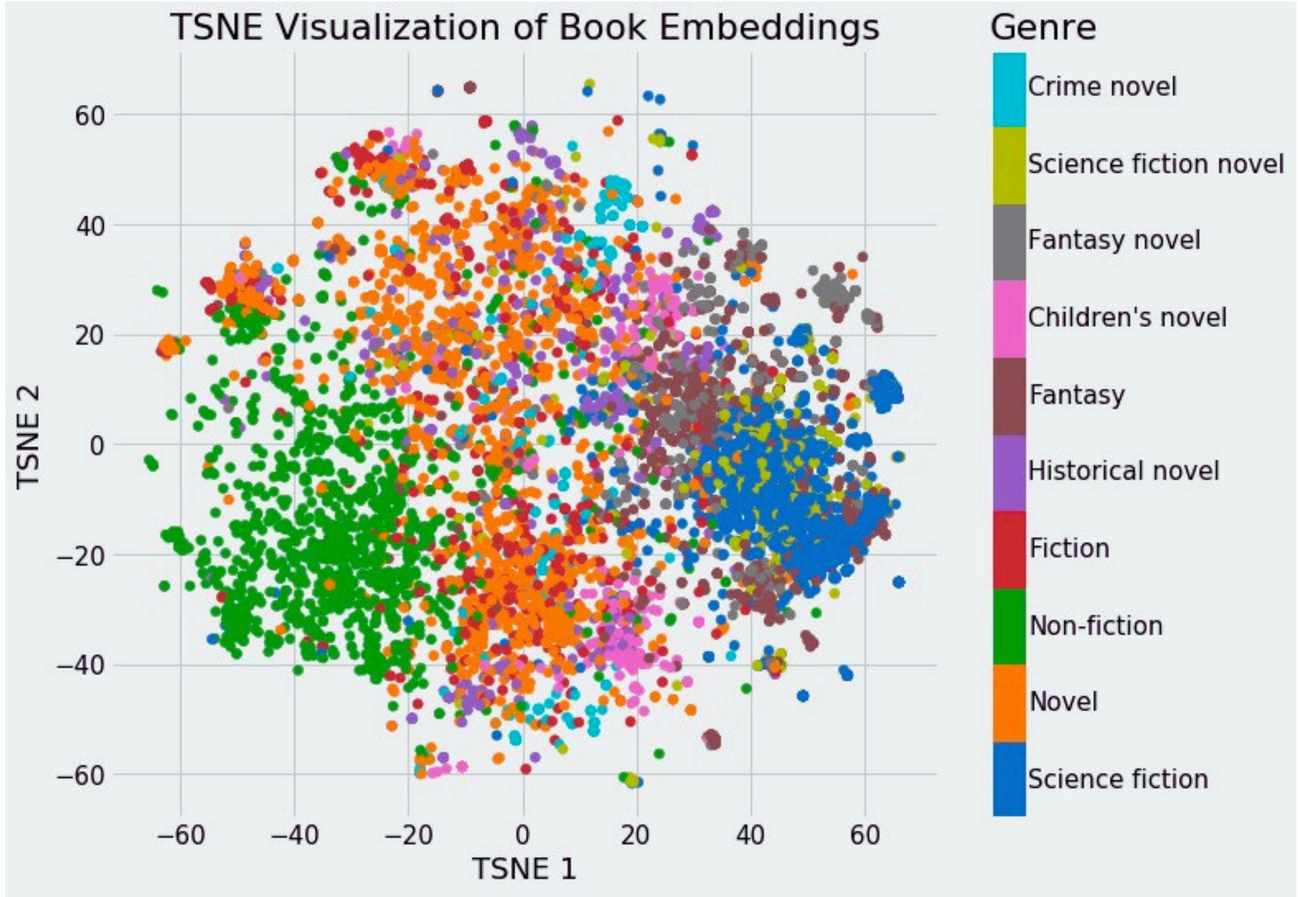
Types of Data

- **Sequential:** words have order and structure.



Types of Data

- **High-dimensional:** vocabulary size can be thousands/millions.



Types of Data

- **Ambiguous:** Same word may mean different things in different sentences.

food

due to its tartness, it is often combined with sweeter juices, such as **apple** or grape

organization

apple is rumored to be working on a smartwatch, which maybe be called an "iwatch."

organization

A cliq app was released for **apple**'s ios devices in august.

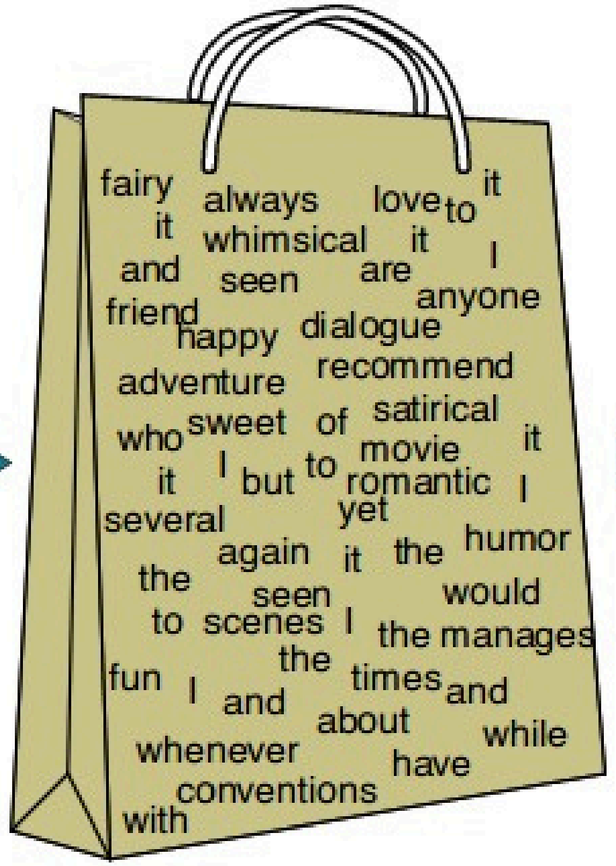
word	semantic classes
apple	food, organization

Types of Data

- Text Representation

- Bag of words

I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet!



it	6
I	5
the	4
to	3
and	3
seen	2
yet	1
would	1
whimsical	1
times	1
sweet	1
satirical	1
adventure	1
genre	1
fairy	1
humor	1
have	1
great	1
…	…

Types of Data

- Text Representation

- Bag of words

Document	NLP	is	fun	very	powerful
Doc1: NLP is fun	?	?	?	?	?
Doc2: NLP is very fun	?	?	?	?	?
Doc3: NLP is powerful	?	?	?	?	?

Types of Data

- Text Representation

- Bag of words

Document	NLP	is	fun	very	powerful
Doc1: NLP is fun	1	1	1	0	0
Doc2: NLP is very fun	1	1	1	1	0
Doc3: NLP is powerful	1	1	0	0	1

Types of Data

- Text Representation
  - TF-IDF

- Term Frequency (TF)
- Measures how often a word appears in a document

$$TF_{ij} = \frac{\text{Number of times the term appears in the document}}{\text{Total number of terms in the document}}$$

Types of Data

- Text Representation
  - TF-IDF

- Inverse Document Frequency (IDF)
- Measures how rare or common a word is across the corpus

$$IDF_i = \log\left(\frac{\text{Total number of documents}}{\text{Number of documents containing term } i}\right)$$

Types of Data

- Text Representation
  - TF-IDF

$$TF_{ij} = \frac{\text{Number of times the term appears in the document}}{\text{Total number of terms in the document}}$$

$$*$$

$$IDF_i = \log\left(\frac{\text{Total number of documents}}{\text{Number of documents containing term } i}\right)$$

- High IDF ⇒ word is rare but valuable
- Low IDF ⇒ word is common (e.g., “the”, “a”, “and”)

Types of Data

- Text Representation
  - TF-IDF
    - Term Frequency (TF)
    - Measures how often a word appears in a document

$$TF_{ij} = \frac{\text{Number of times term appears in the document}}{\text{Total Number of terms in the document}}$$

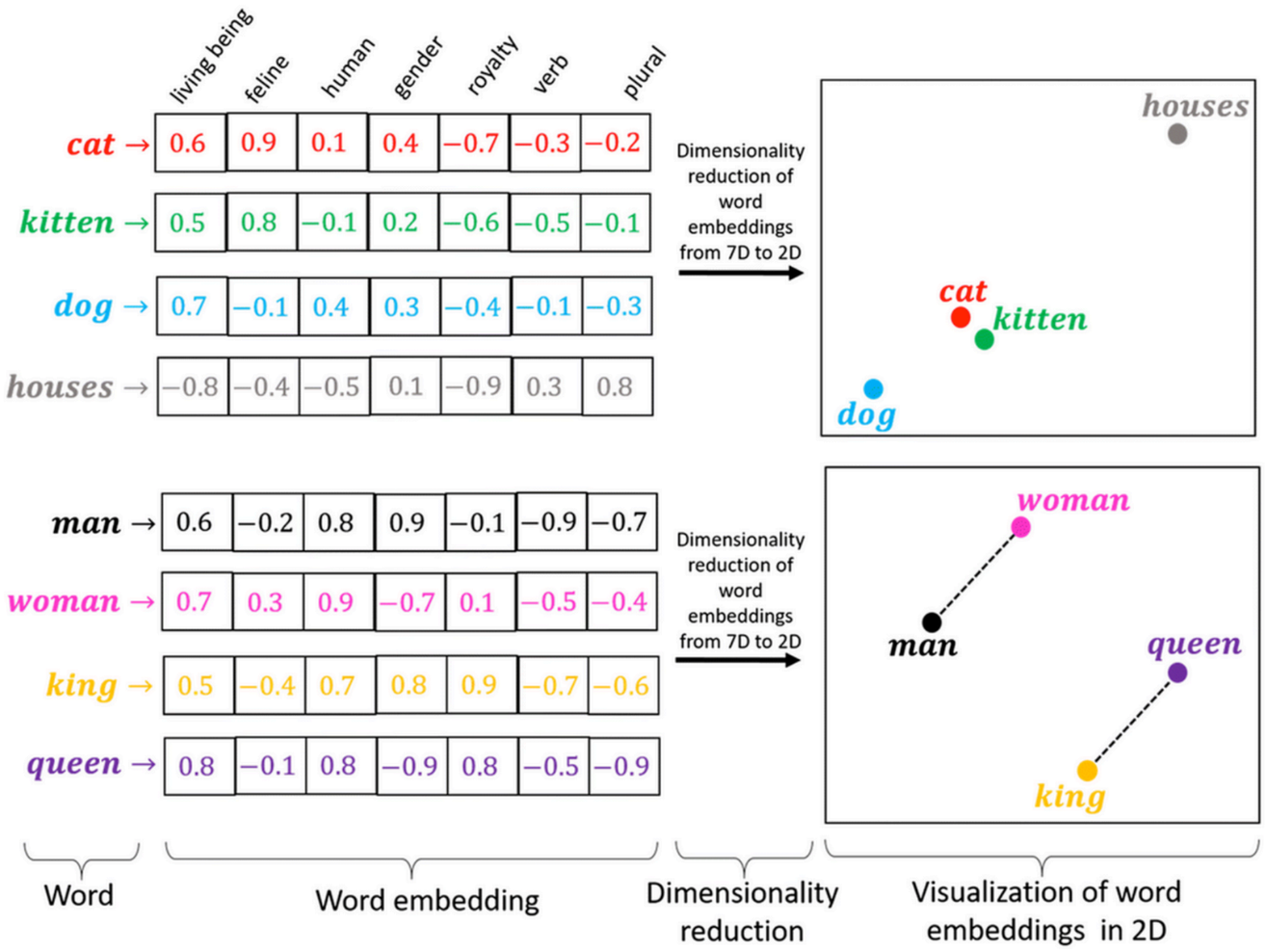
    - Inverse Document Frequency (IDF)
    - Measures how rare or common a word is across the corpus

$$IDF_i = \log\left(\frac{\text{Total number of documents}}{\text{Number of documents with term } i \text{ in it}}\right)$$

- High IDF ⇒ word is rare but valuable
- Low IDF ⇒ word is common (e.g., “the”, “a”, “and”)

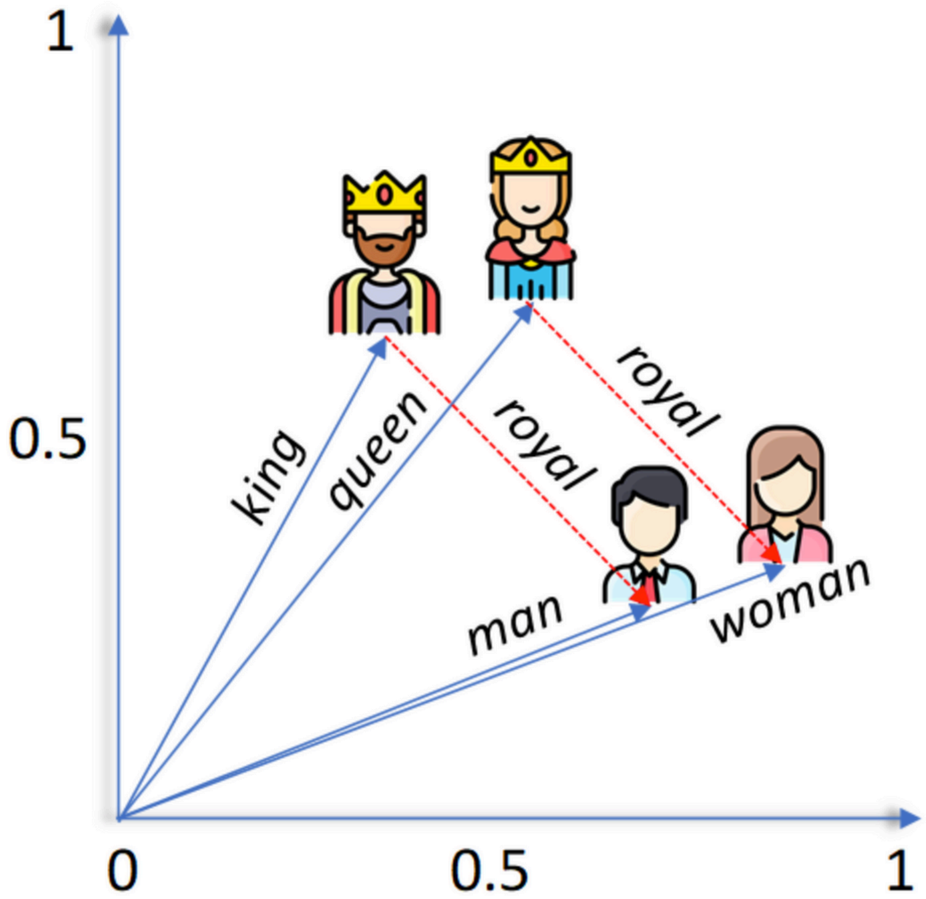
Types of Data

- Word Embedding



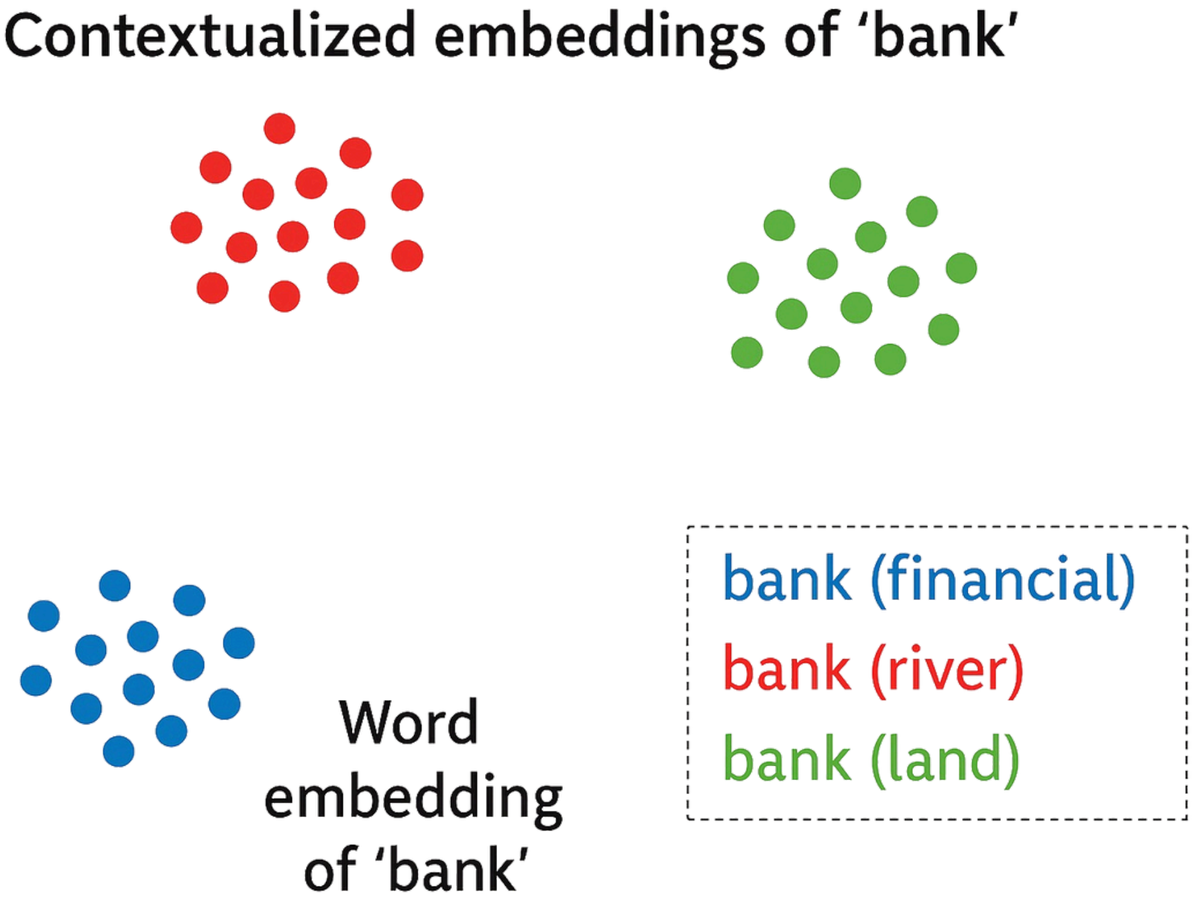
Types of Data

- Word Embedding



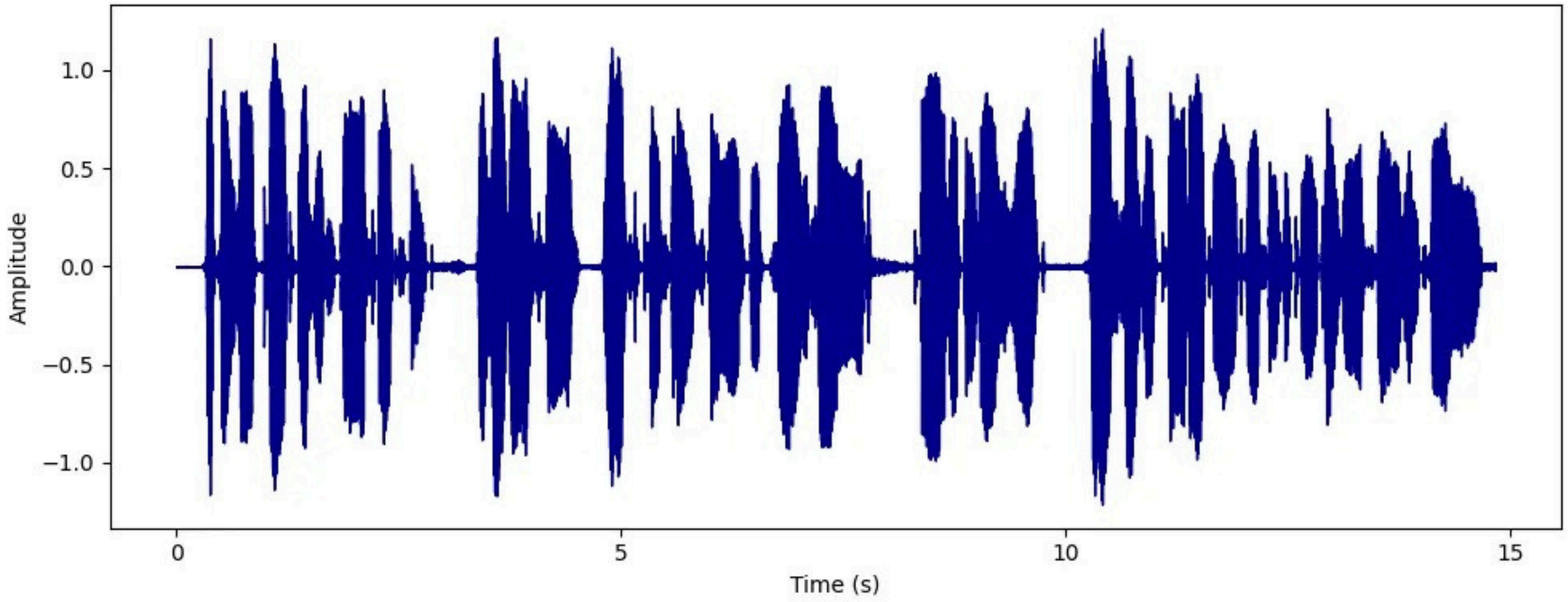
Types of Data

- Contextualized Embeddings



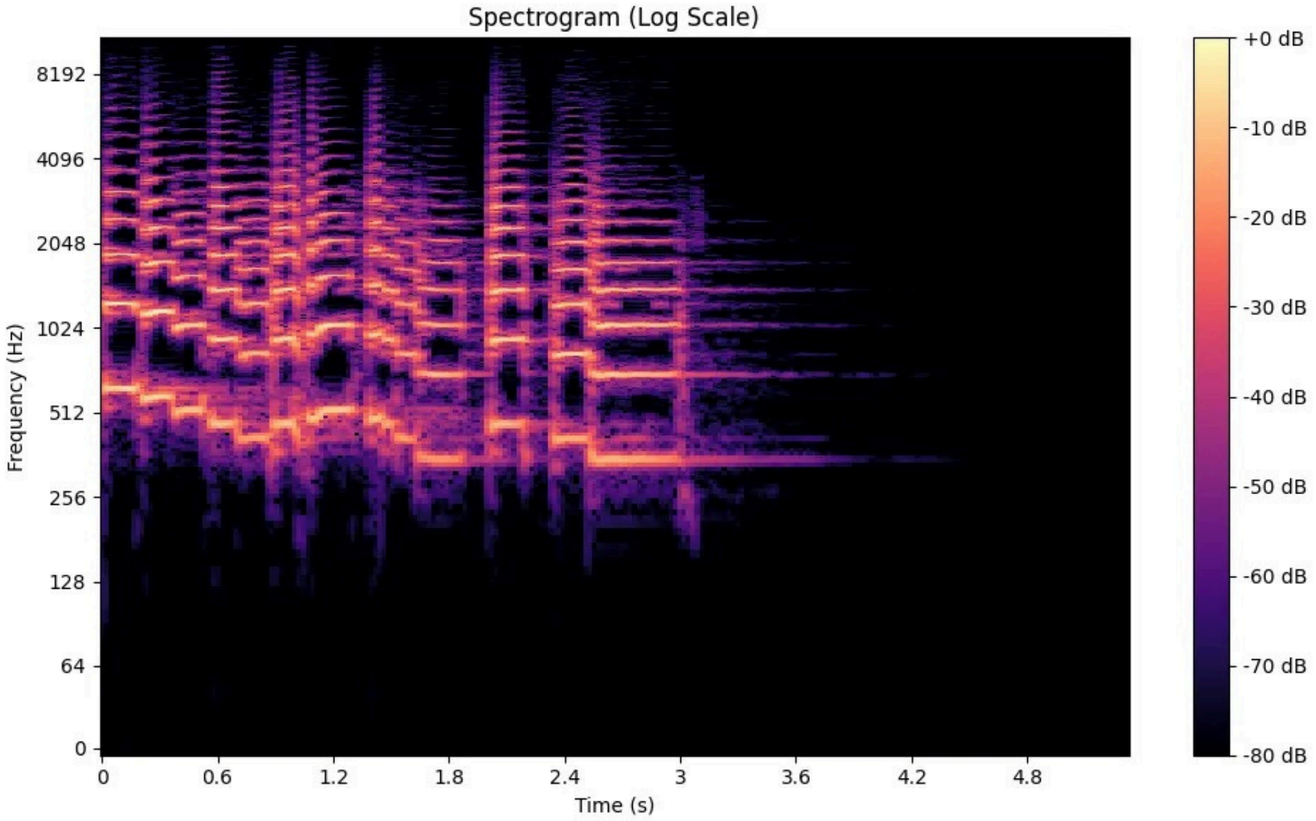
Types of Data

- Audio data



Types of Data

- Audio data



Data Literacy

Sources of Data

Sources of Data

kaggle

OPEN DATA PLATFORM
منصـــة البيـــانات المفتـــوحة

OpenML

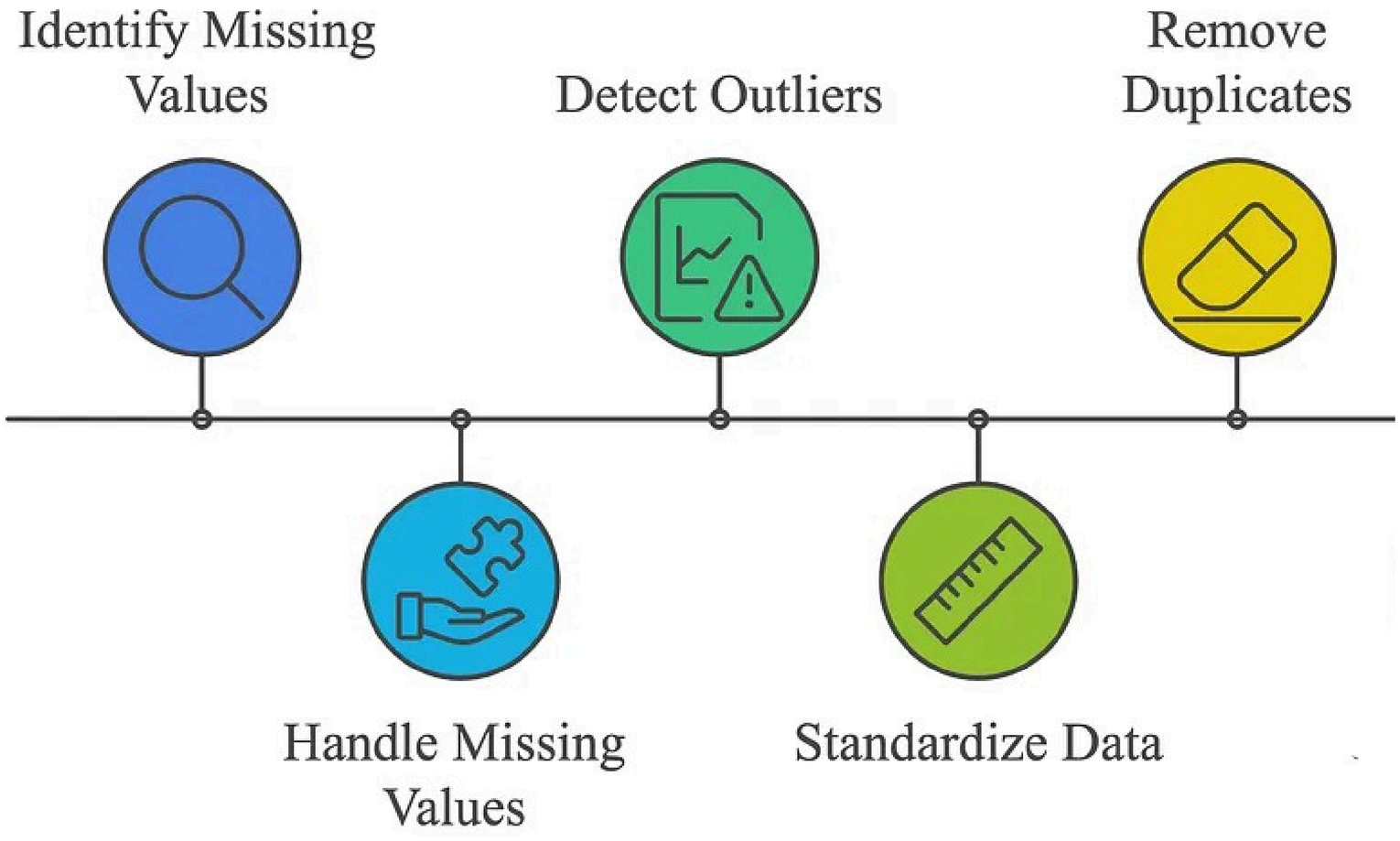
Data Literacy

Data Cleaning

Data Cleaning

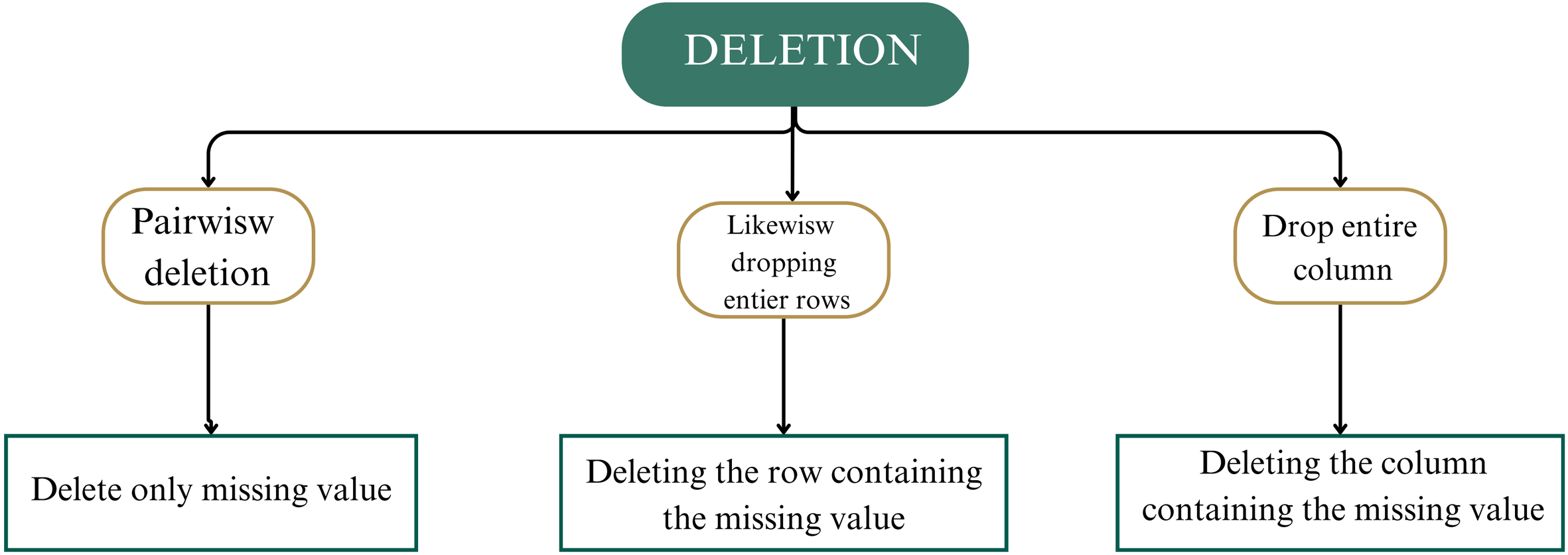
- **Data cleaning** is the process of fixing, correcting, and preparing raw data so it becomes accurate, consistent, and usable for analysis or machine learning.

Data Cleaning



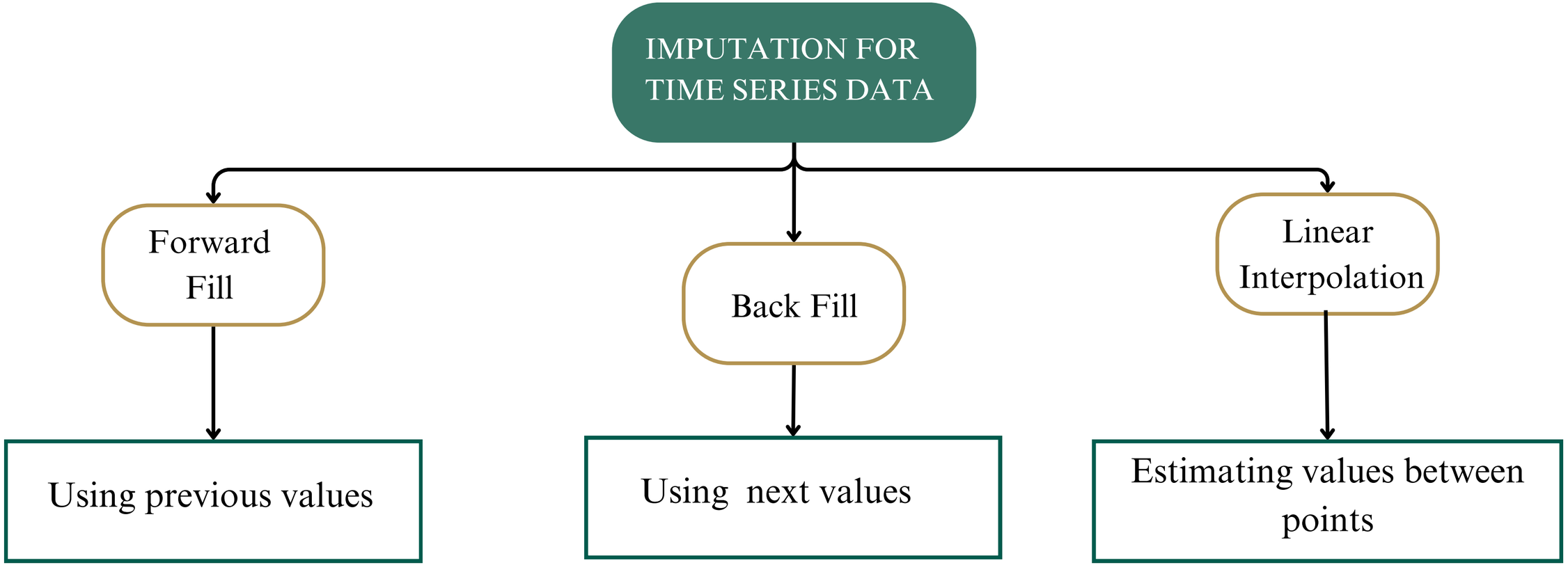
Data Cleaning

- Missing Values



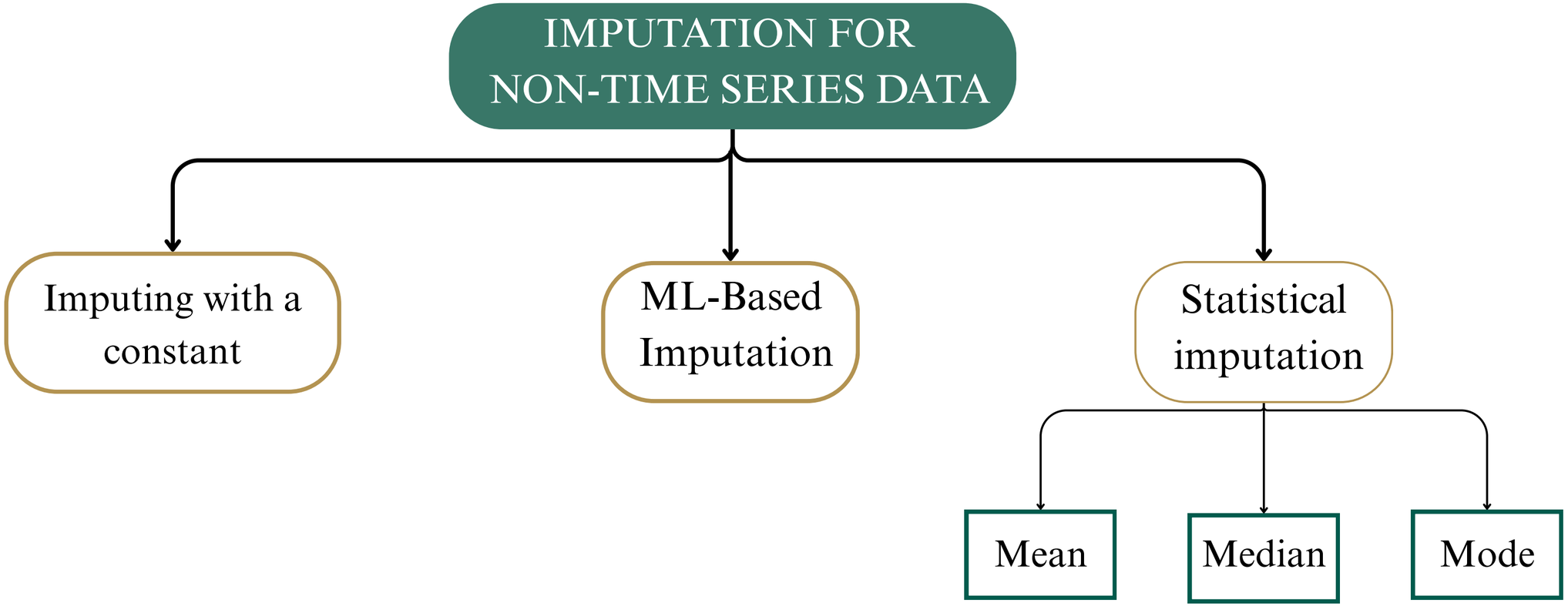
Data Cleaning

- Missing Values



Data Cleaning

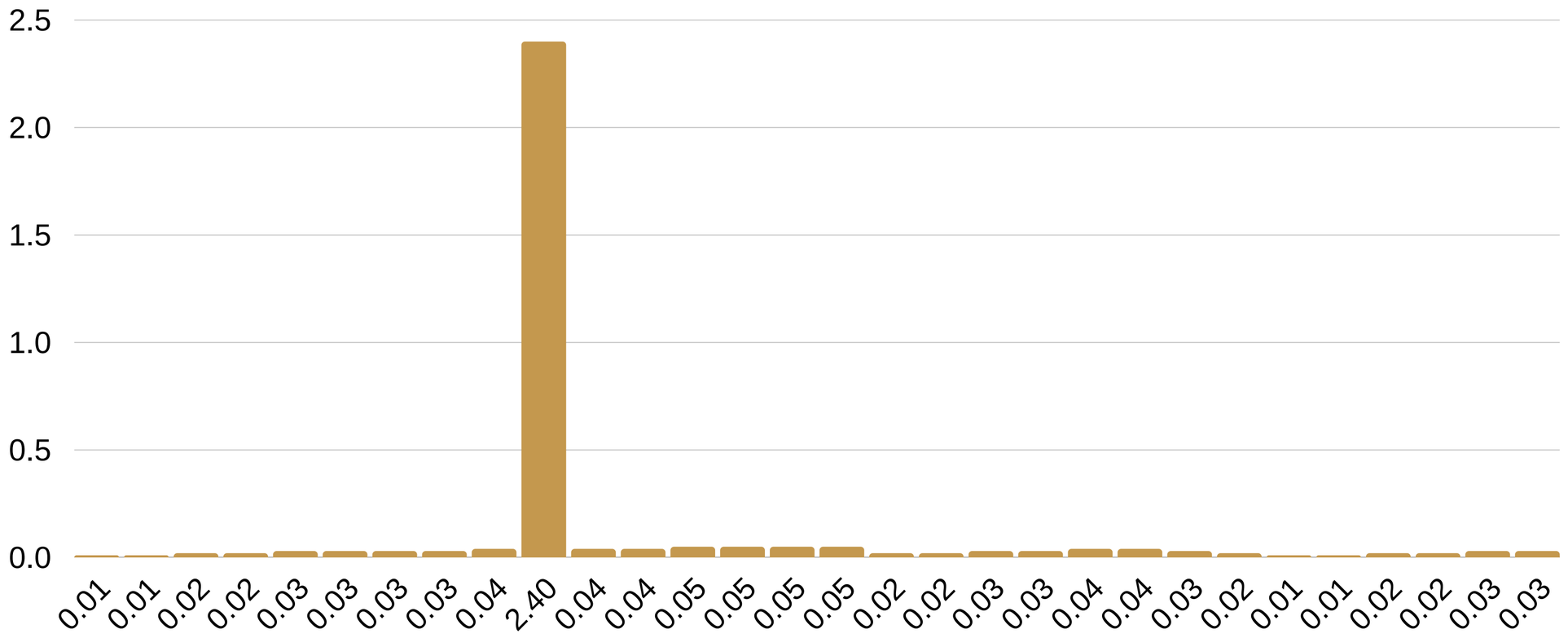
- Missing Values



Does the type of imputation matter ?

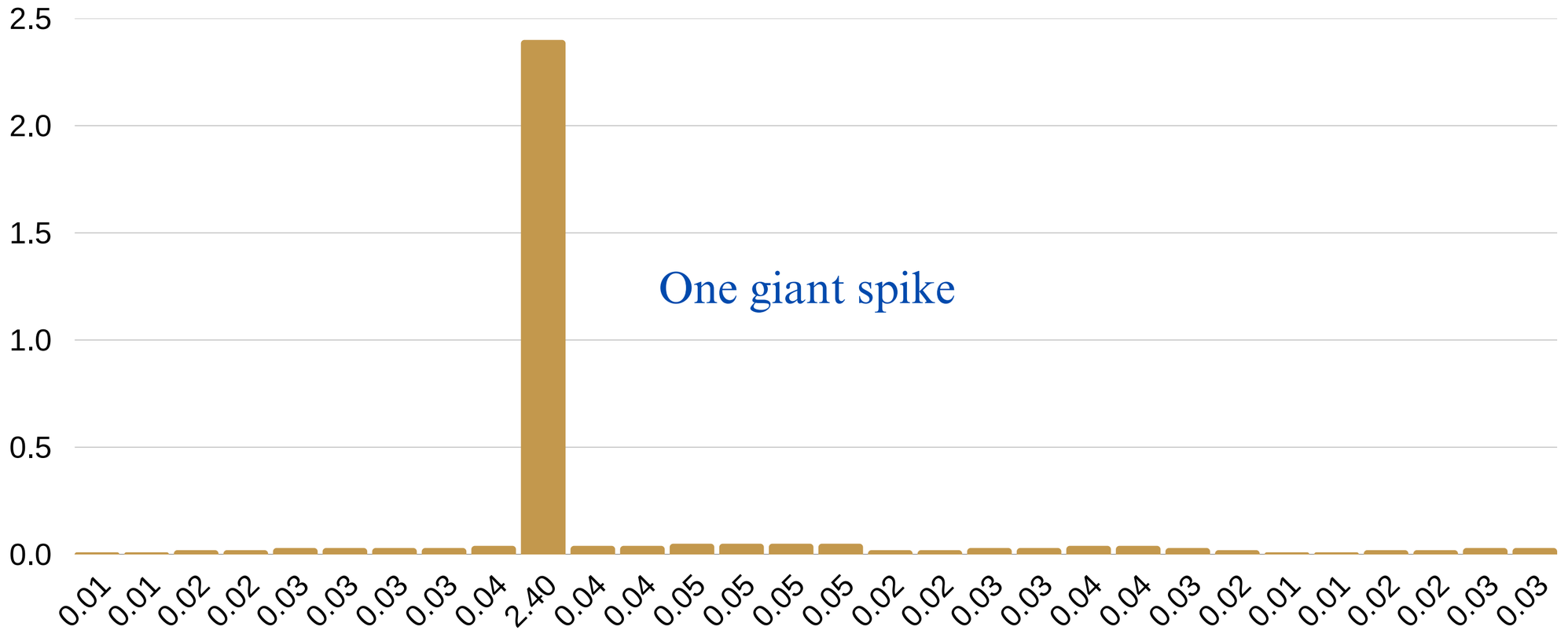
Data Cleaning

What's wrong with this histogram?



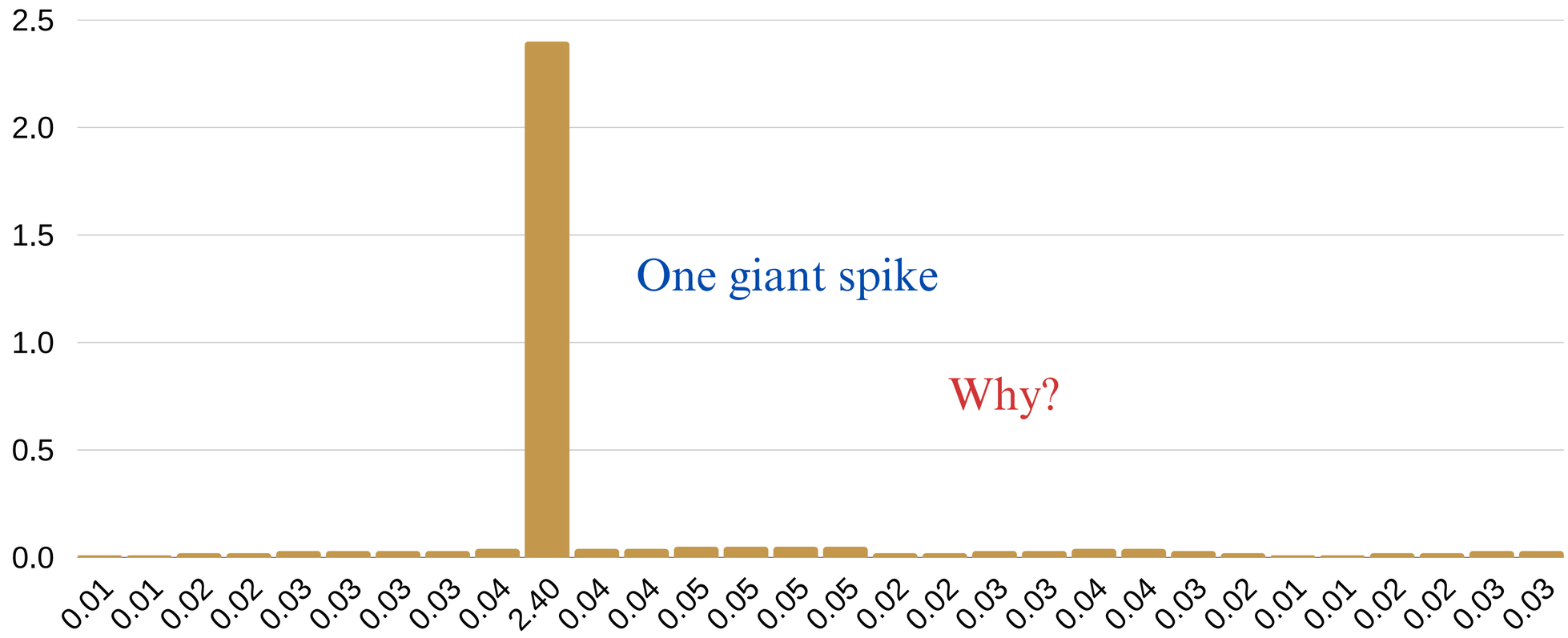
Data Cleaning

What's wrong with this histogram?



Data Cleaning

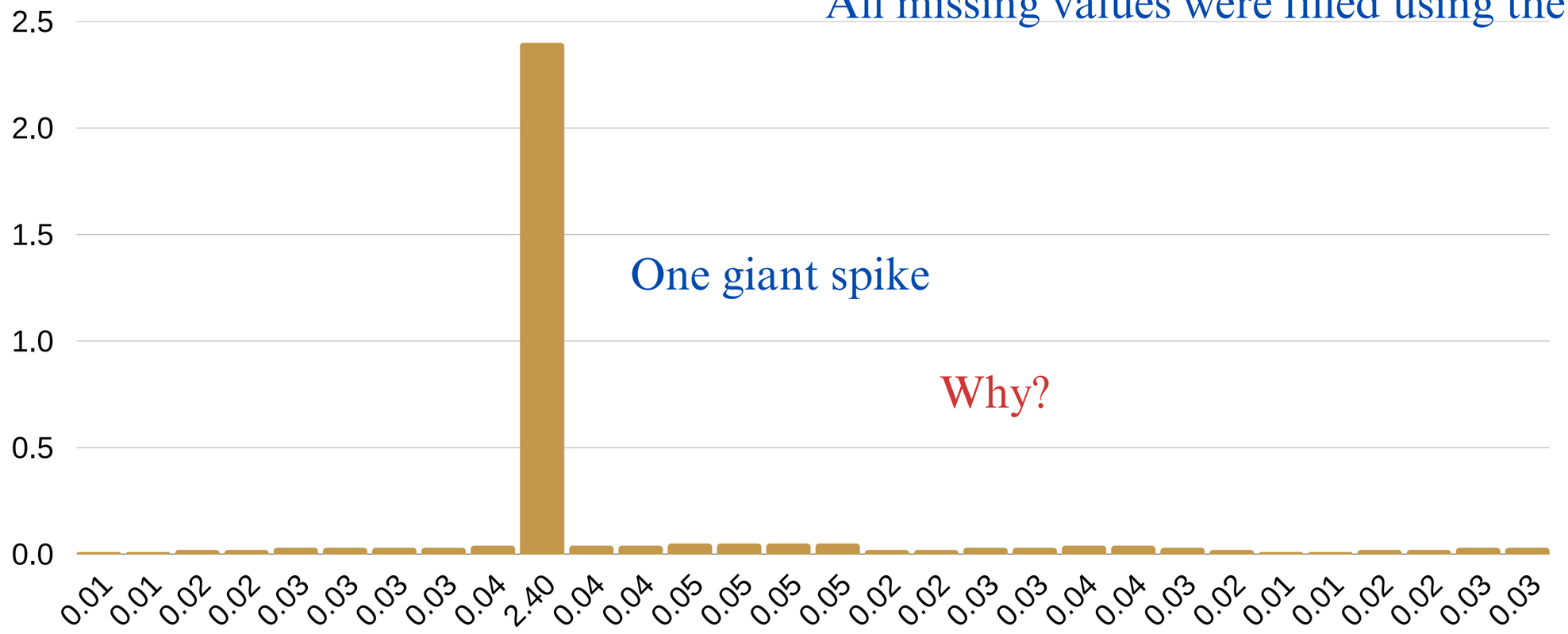
What's wrong with this histogram?



Data Cleaning

What's wrong with this histogram?

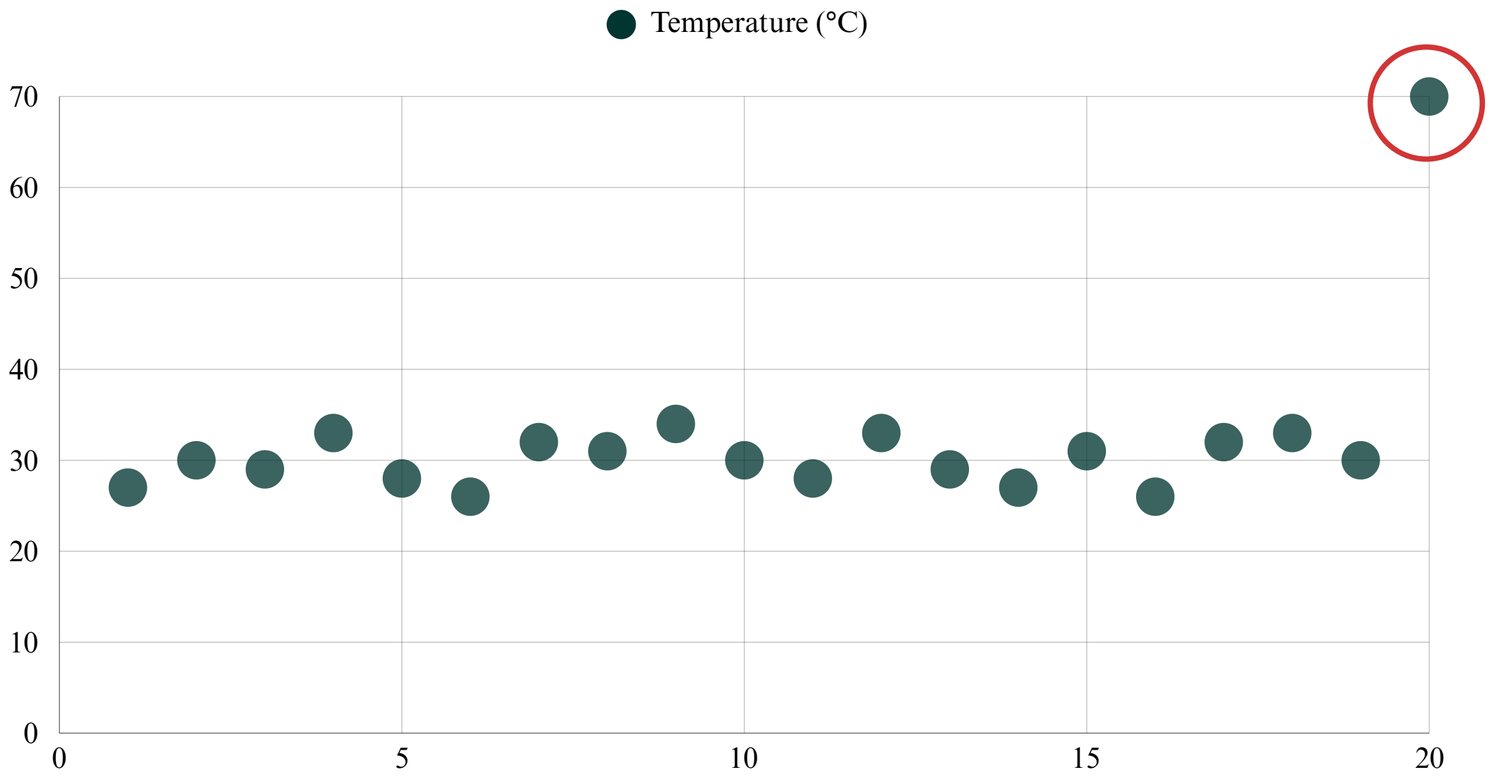
All missing values were filled using the mean



Data Cleaning

- Outliars

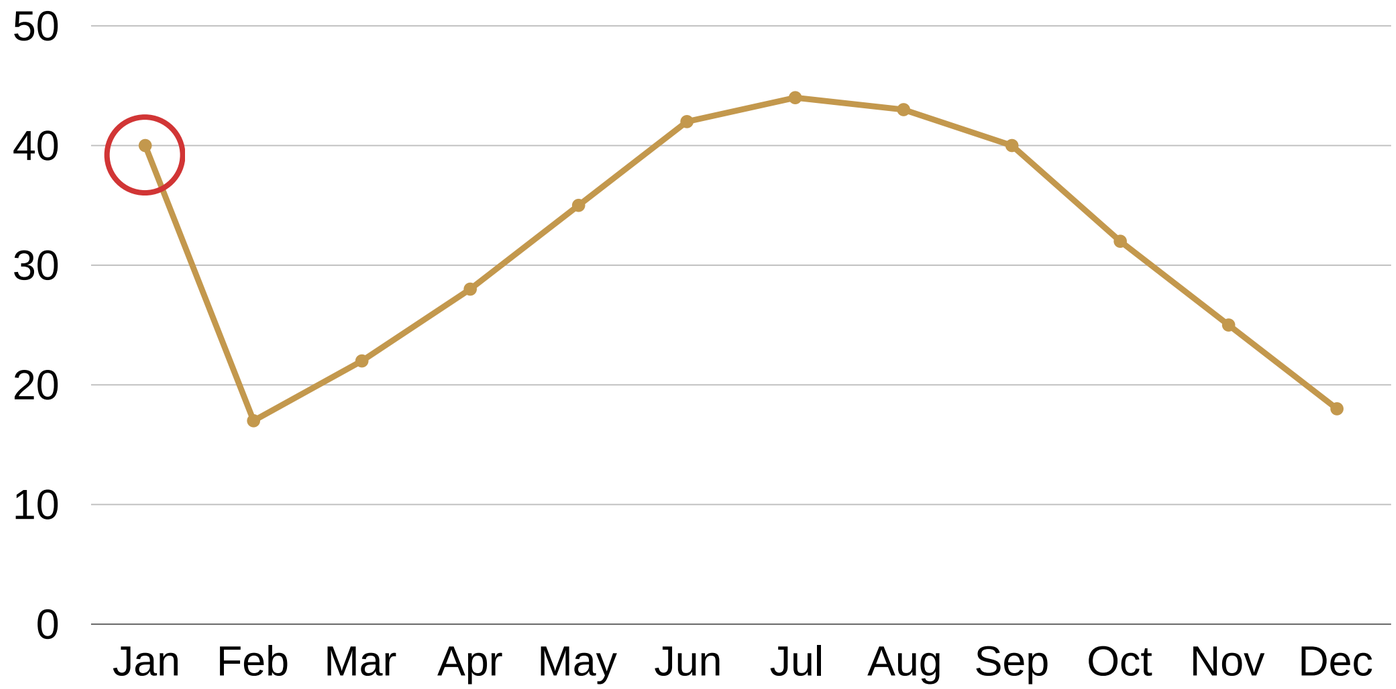
- Point Outliers



90% of temperatures are 25–35°C , one value is 70°C.

Data Cleaning

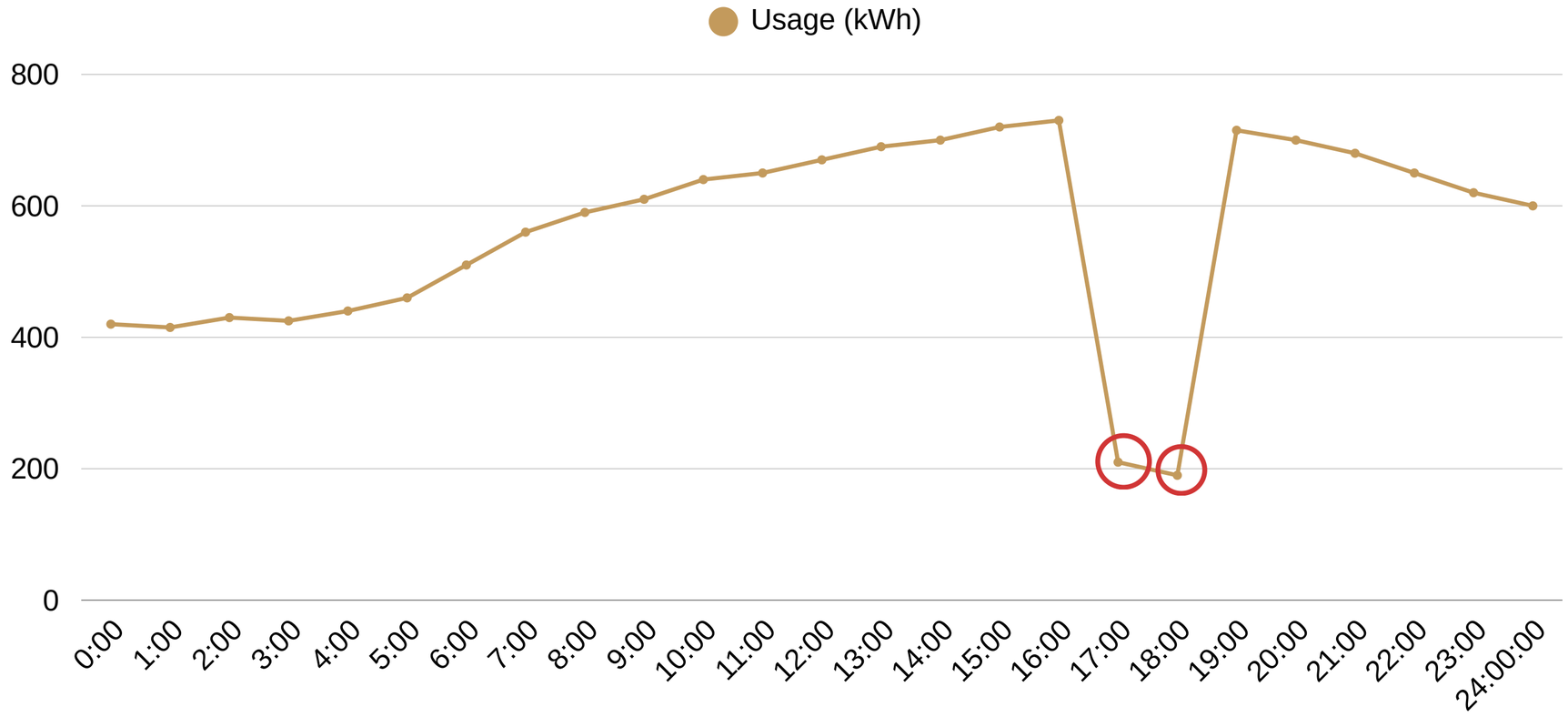
- Contextual Outliers



40°C normal in Riyadh summer, but 40°C in January becomes an outlier

Data Cleaning

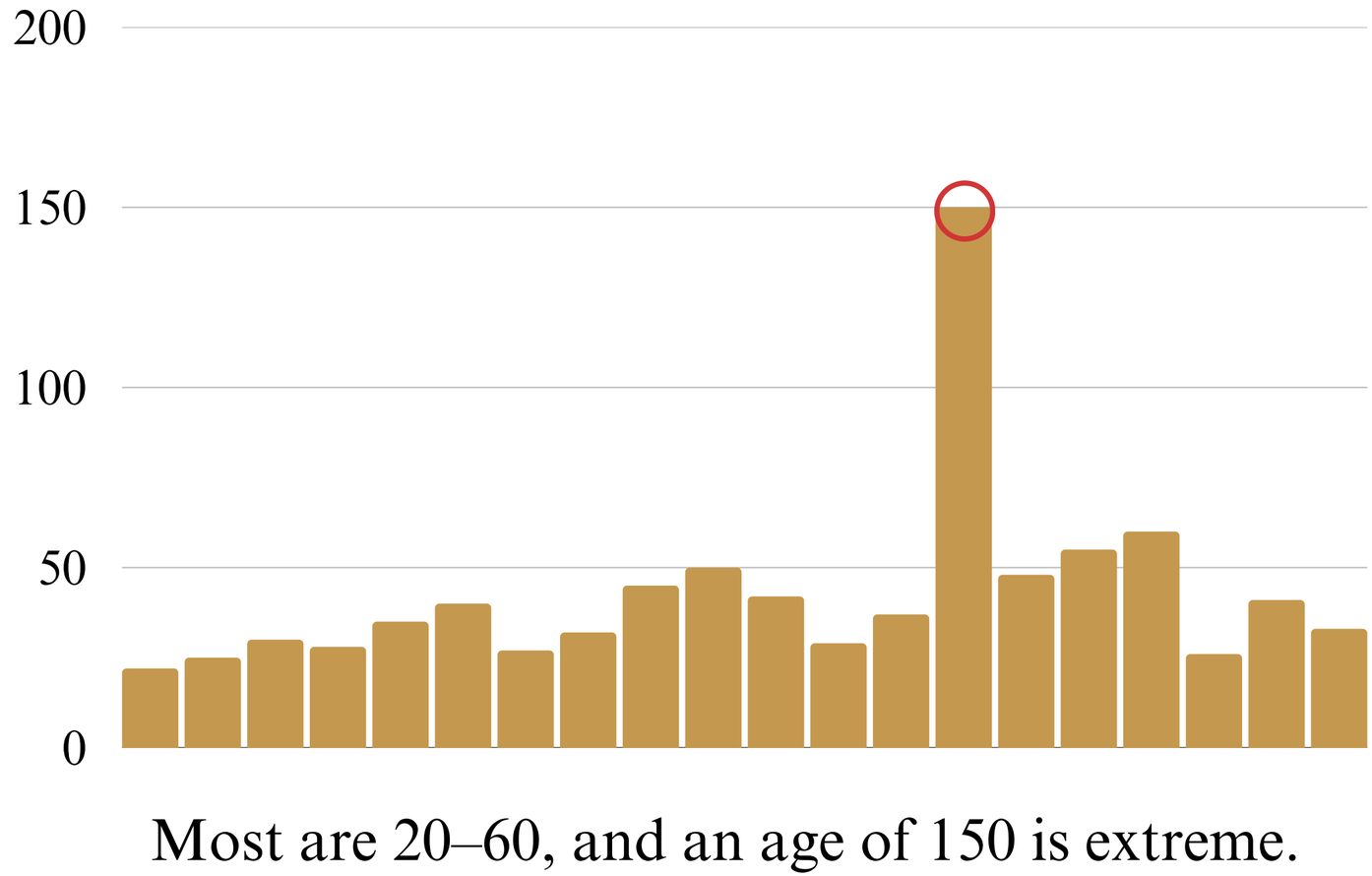
- Collective Outliers



Sudden drop in electricity usage for 2 hours, unusual as a chunk.

Data Cleaning

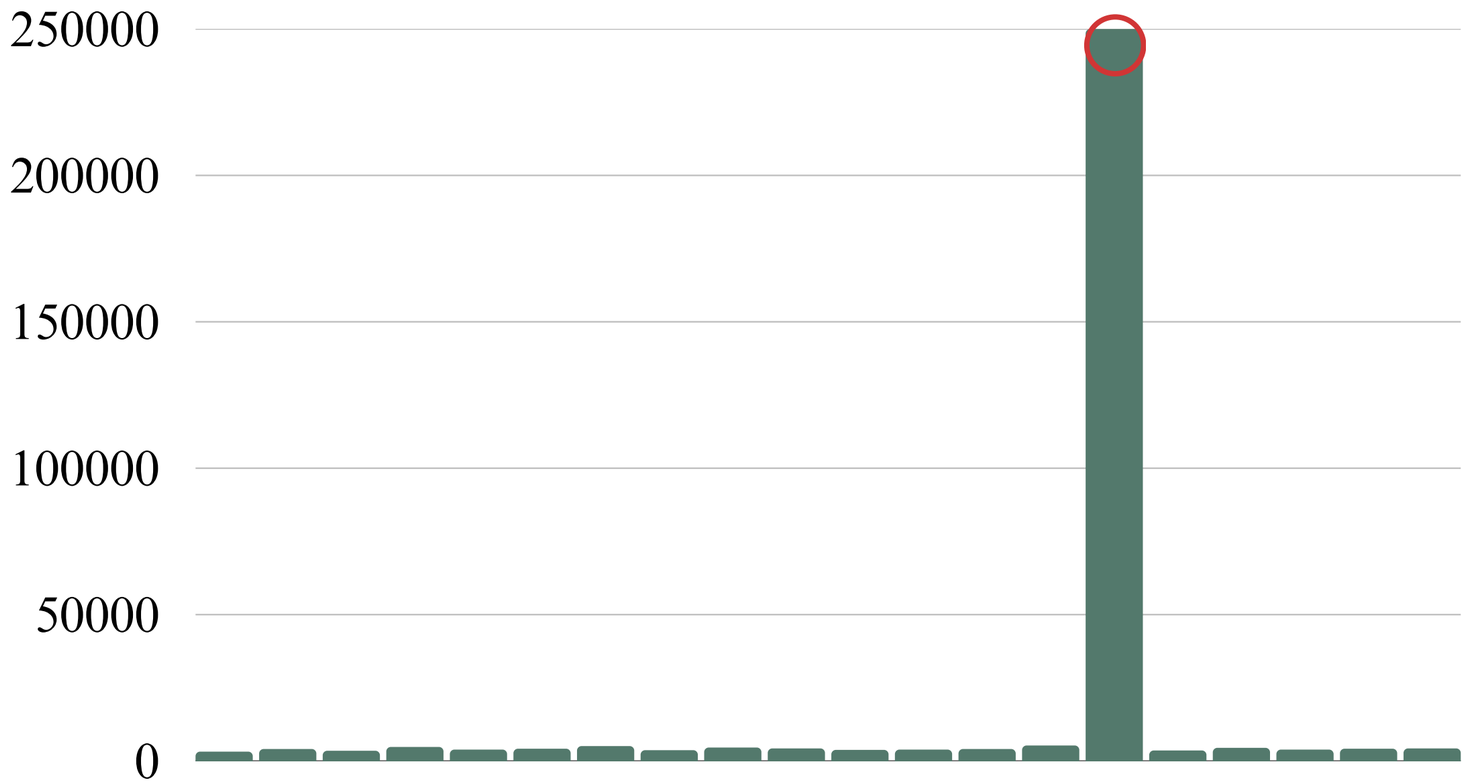
- Extreme Value Outliers



Most are 20–60, and an age of 150 is extreme.

Data Cleaning

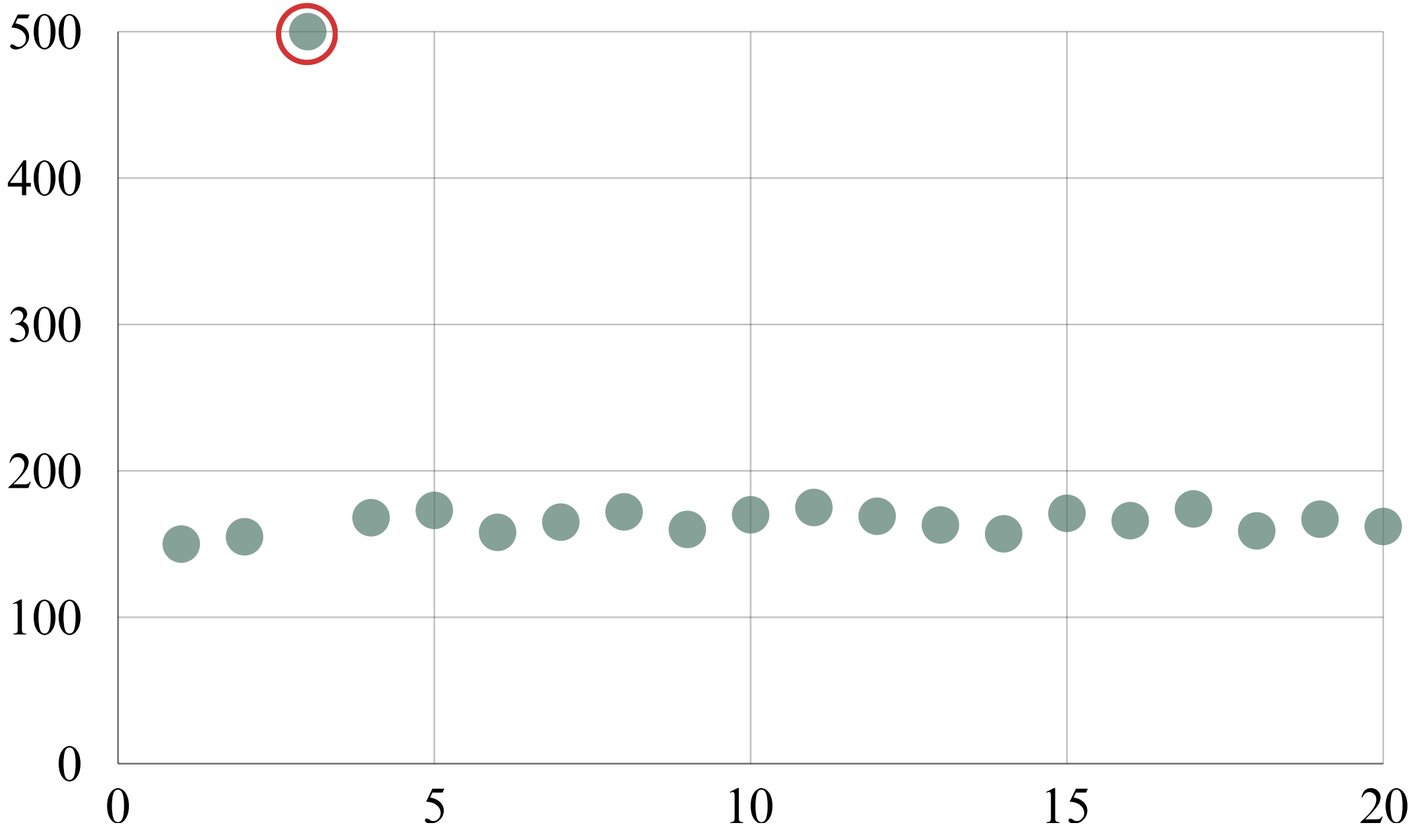
- Natural Outliers



Millionaire salary inside middle-class salary data.

Data Cleaning

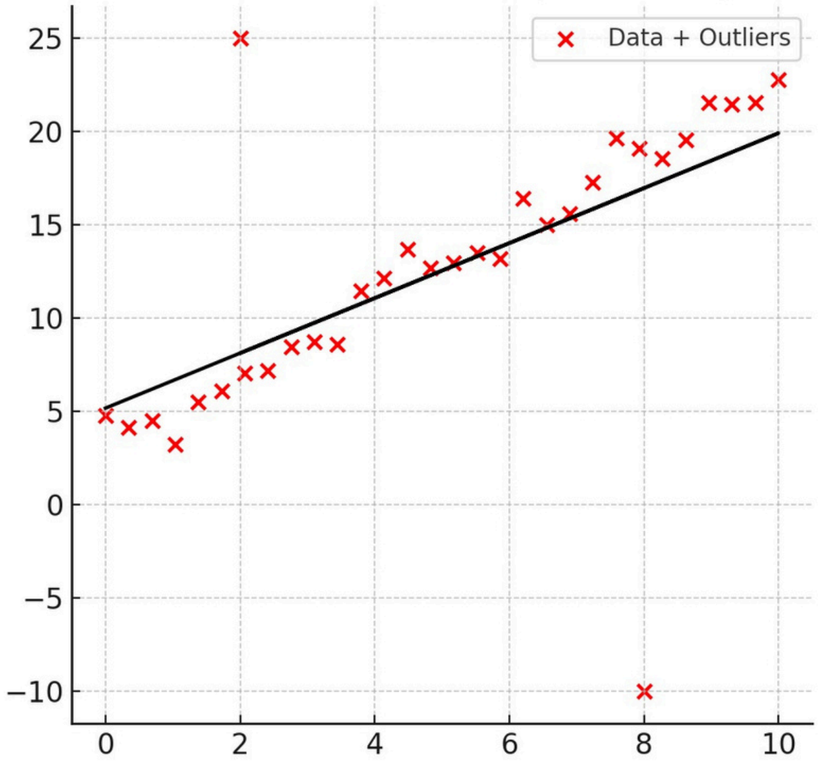
- Error Outlier

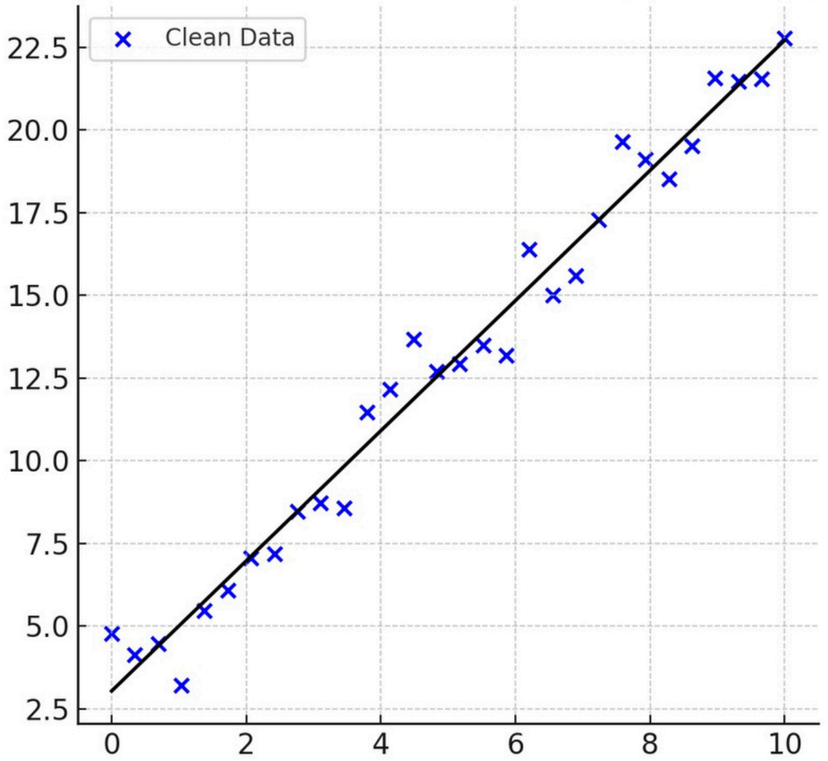


Height recorded as 500 cm instead of 150 cm

Data Cleaning

How outliers affect the behavior of a machine learning model.





Data Literacy

Hands-On Activity

Hands-On Activity

[Egypt Traffic Violations Dataset (2024)]

Hands-On Activity

HR Analytics

Materials

Day 2 Materials

Workshop Materials day 2

Quiz part 1

[quiz part 1]

quiz part 1

Quiz part 2

quiz part 2

quiz part 2


Q
A

Thank You