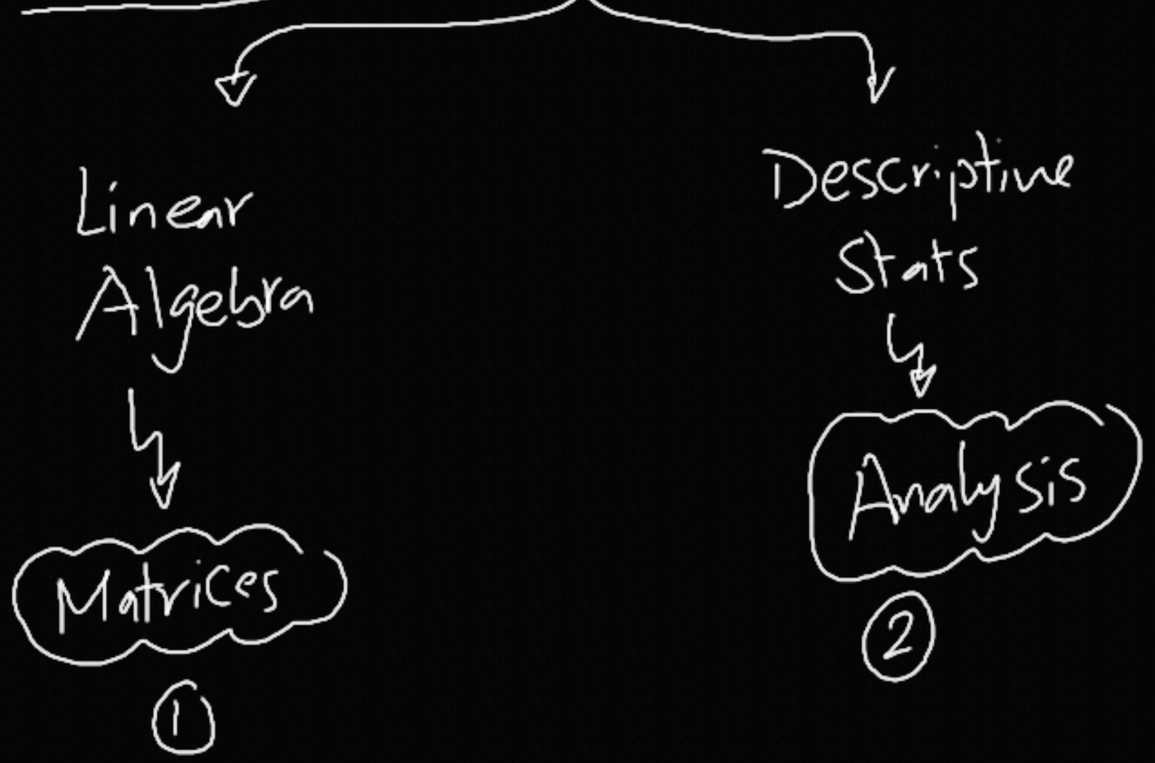
IBM Data Science

1. Intro
2. Python ~> web scraping / API → yfinance / weather
3. Data Analysis (python) → Data → Analysis & Vis / prediction
4. Data Analysis (SQL) → Data sources (Database)
5. Machine Learning (Python)

Math for Data Science

Data Analysis (Math)



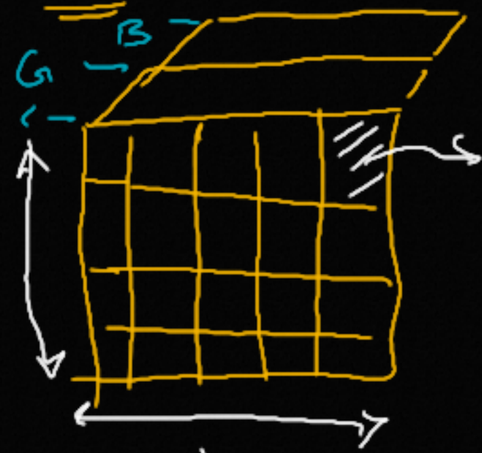
① Linear Algebra (Matrices)

① What is a Matrix?

Array

⇝ one data structure Contains a Collection of numbers

Ex: Images



pixel ⇝ Intensity of Color

grayscale

(0 ⟶ 255)

Pure Black

pure White

(1000, 1000)

1MP

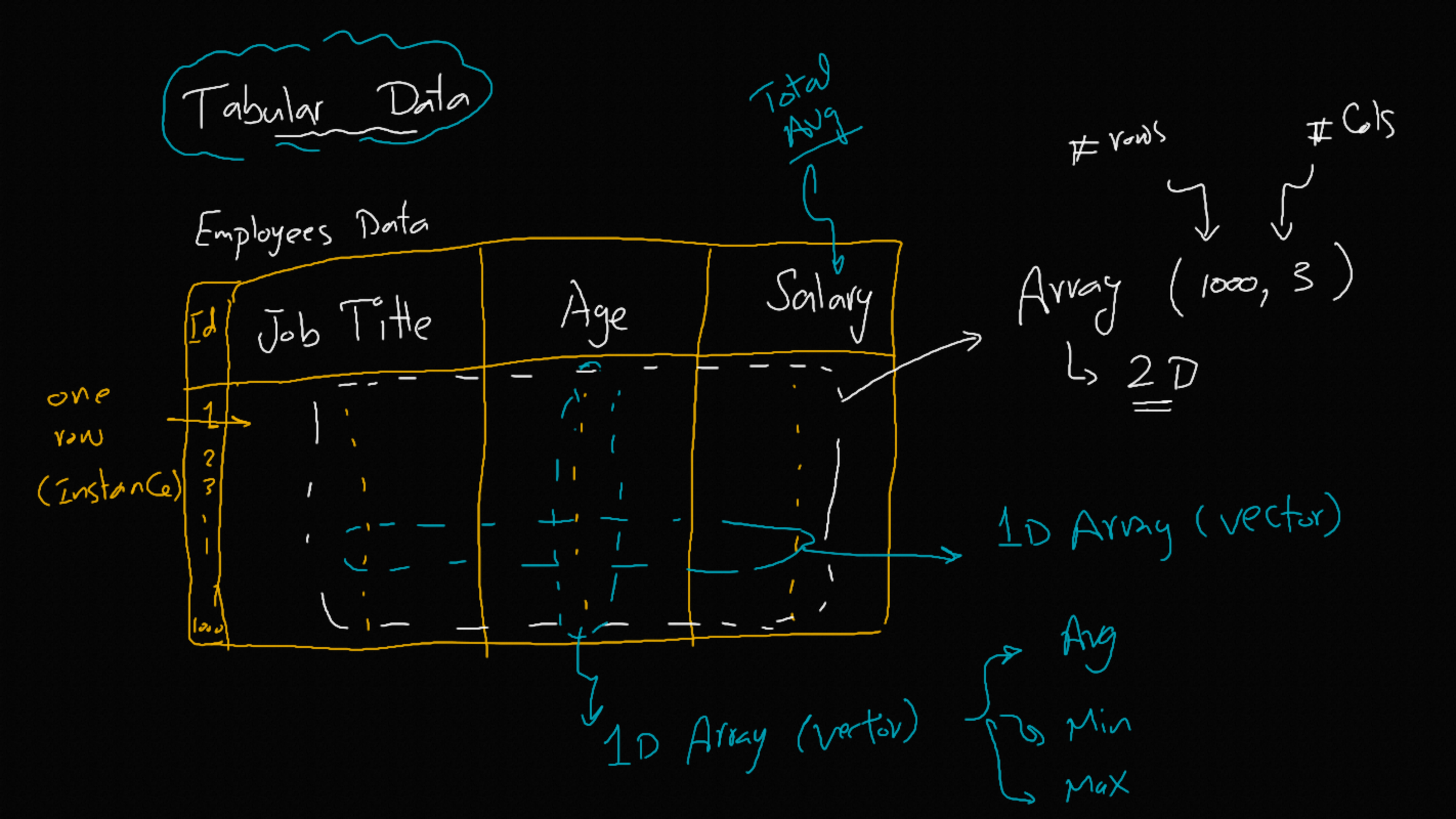
Computer Vision

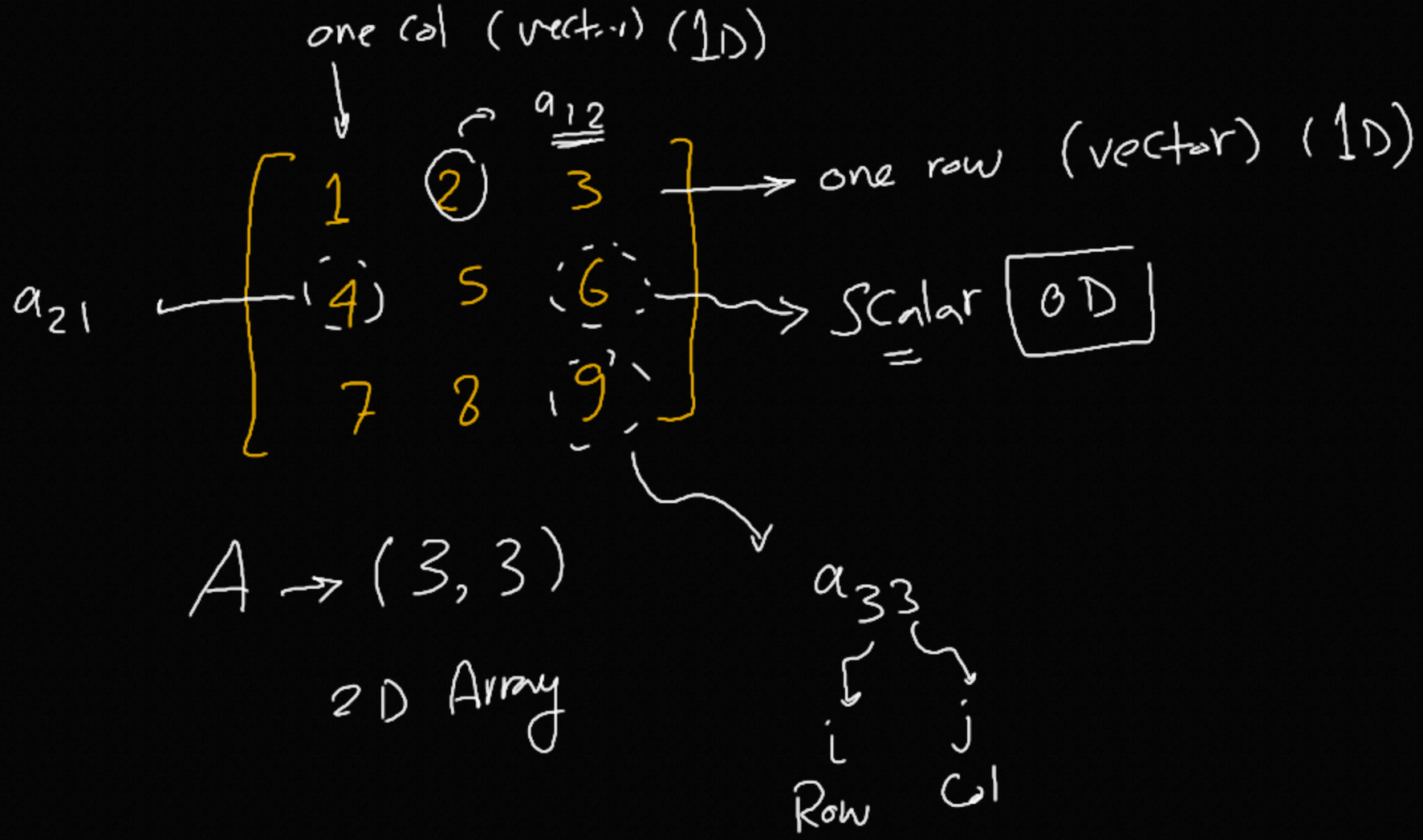
1000

1000

3 Colors

- R [0 → 255]
- G [0 → 255]
- B [0 → 255]


Tabular Data
Total Avg
rows
Cols
Employees Data
Id
Job Title
Age
Salary
Array (1000, 3)
↳ 2D
one row (Instance)
1
2
3
1000
1D Array (vector)
1D Array (vector)
Avg
Min
Max


one col (vector) (1D)
a_{12}
1 2 3
4 5 6
7 8 9
one row (vector) (1D)
a_{21}
Scalar
0D
A → (3, 3)
2D Array
a_{33}
i
j
Row
Col

Math operations

① Element-l by Element operations

$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \pm \begin{bmatrix} 5 & 6 \\ 7 & 8 \end{bmatrix}$$

2x2 ⟷ 2x2

Same Size

Same Size

Salary Bonus

6000 + 500 =

Total → Salary + Bonus

6500

→ → →

+ − * /

② Dot product

Dot product

$$A_{r1} \cdot B_{c1} \longrightarrow C_{11}$$

$$A \begin{bmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{bmatrix} \odot B \begin{bmatrix} 5 & 6 & 7 \\ 9 & 8 & 1 \end{bmatrix} = C \begin{bmatrix} (1\times5+4\times9) & (1\times6+4\times8) & (1\times7+4\times1) \\ (2\times5+5\times9) & (2\times6+5\times8) & (2\times7+5\times1) \\ (3\times5+6\times9) & (3\times6+6\times8) & (3\times7+6\times1) \end{bmatrix}$$

$3 \times \textcircled{2}$ $\quad \textcircled{2} \times 3$

$=$

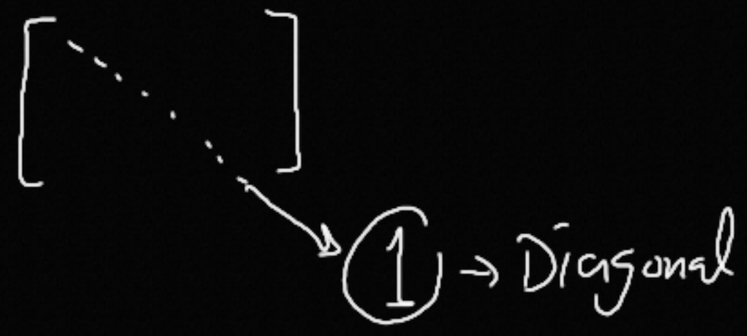
Condition

$=$

Result Dimension

$5 \rightarrow \frac{1}{5} \longrightarrow 5 \times \frac{1}{5} = 1$

- $A \xrightarrow{\text{Inverse}} A^{-1} \rightarrow A A^{-1} = I$



- $A \xrightarrow{\text{Transpose}} A^{T}$

$$\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} \longrightarrow \begin{bmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{bmatrix}$$

Job Title - - - -

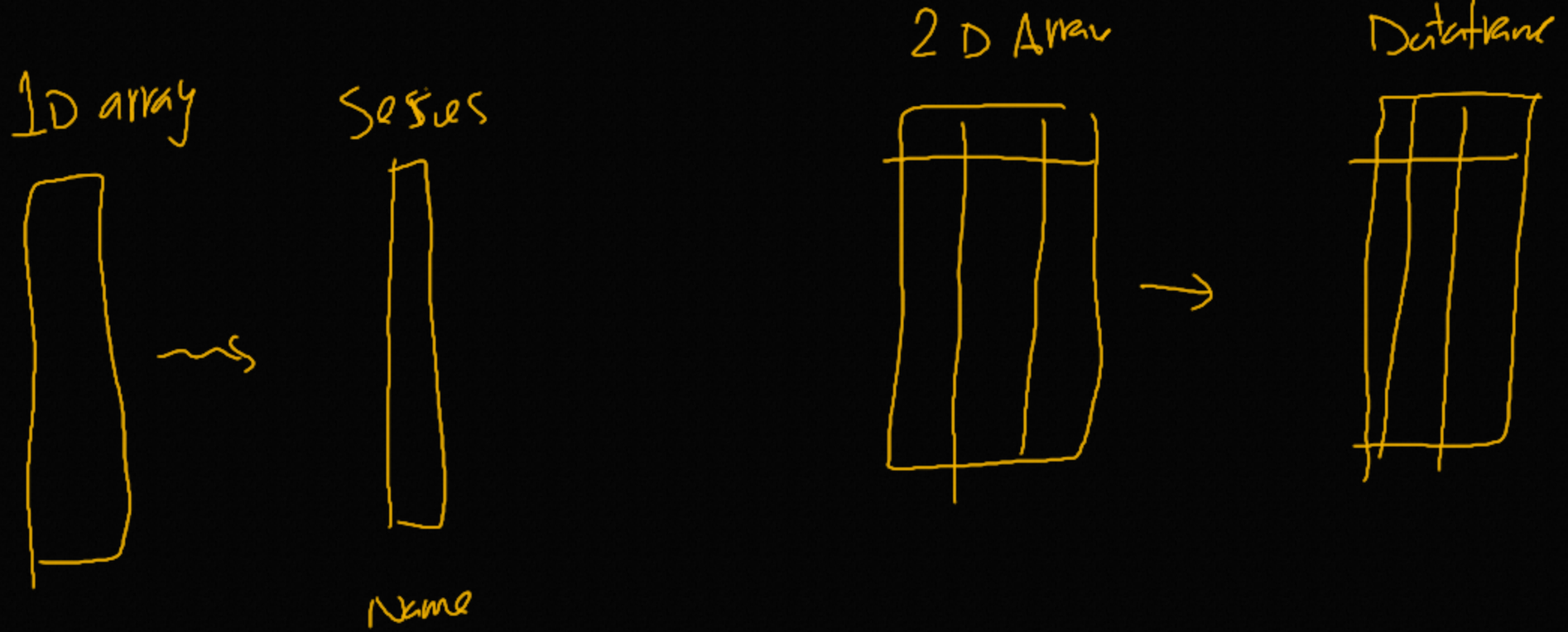
Age - - - -

Salary - - - -

Python ⟶ No array data Structure

↳

Create New Data Structure


1D array
Series
Name
2 D Array
Dataframe

Statistics

→ A branch of Math related to Data

→ It's all About Data

- Collection
- Analysis
- presenting
- Interpreting

→ To get Information

Salaries → Software Testing

All interested group

→ Target → 500,000 ⟿ Population

Conditions

# Exp	Salary
(0 → 3)	Junior
(3 → 6)	Mid
(> 6)	Senior

Salary
Avg. 8000
Avg 13000
Avg 22000

✓ Random
✓ Representitive

Sample

25,000

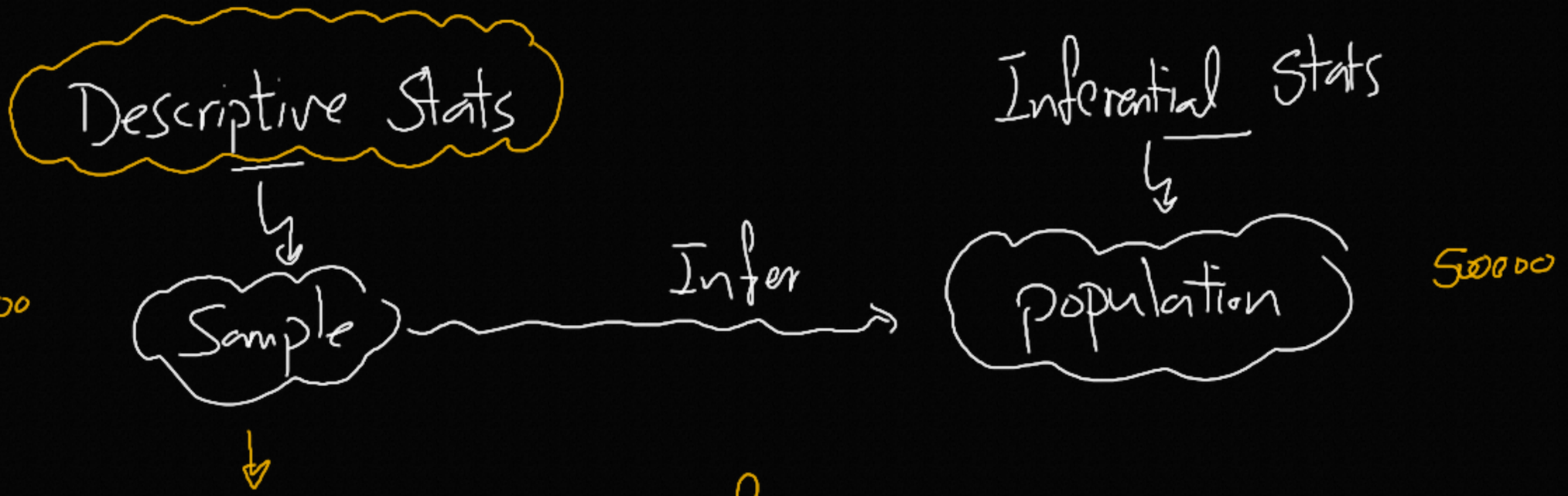
7000

specific group
(partial)

Describe

Infer

Data Analysis



Descriptive Stats

Data Types

- Numerical (quantitative)
  - Continuous
    - Salary, Temp, height, Distance, ----
  - Discrete
    - #Students
- Categorical (qualitative)
  - Nominal → No order
    - Country, Color, Job title
  - ordinal → order / level
    - B, B+, A, A+
    - higher →

multivariate

→ Bivariate

Job title | Age | Exp | Salary → Bivariate

1000 Employers

Univ. Univ. Univ.

$$\text{Mean} = \frac{\text{Total Salary}}{1000} = 1000$$

→ One feature → Continuous

Numerical univ. Analysis

Measures of Center

Centeral Tendency

	Avg.	Min
Co.1 →	9000	-
Co.2 →	15000	-

} Compare

- Why Center? → One representative Number
  - ↳ for Comparison

Mean

9000, 10000, 11000

10000

Median

Middle Nb. (Arranged)

Mode

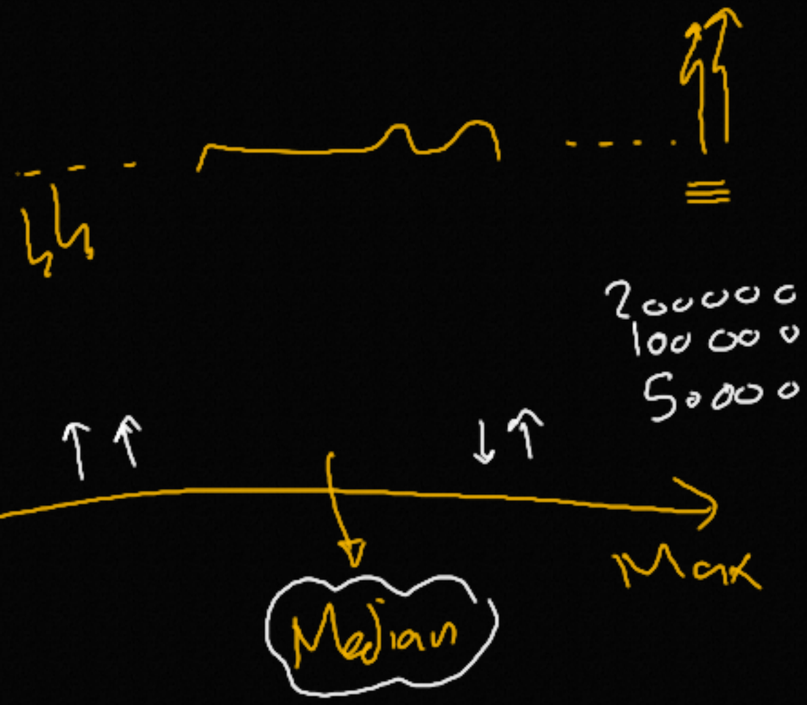
Most freq.

Center

↳ ① Mean → sensitive ↗ dt data
↘ outliers

$\frac{\sum 1000 \text{ Emp} \cdots}{1000}$ ← 5 Emp Salary ↑↑

Mean → ↑↑



② Median → sensitivity ↓↓

③ Mode → Eg 5000 ← Most freq
SA 1000
w 50

8000 ←
↑
Mode

Country
Eg
SA
Dubi
USA

odd No.

Even No

Mean $\frac{A+B}{2}$

Center → One Number

↳ Tell Me More

Five Num. Summary

* 5 No. Summary :

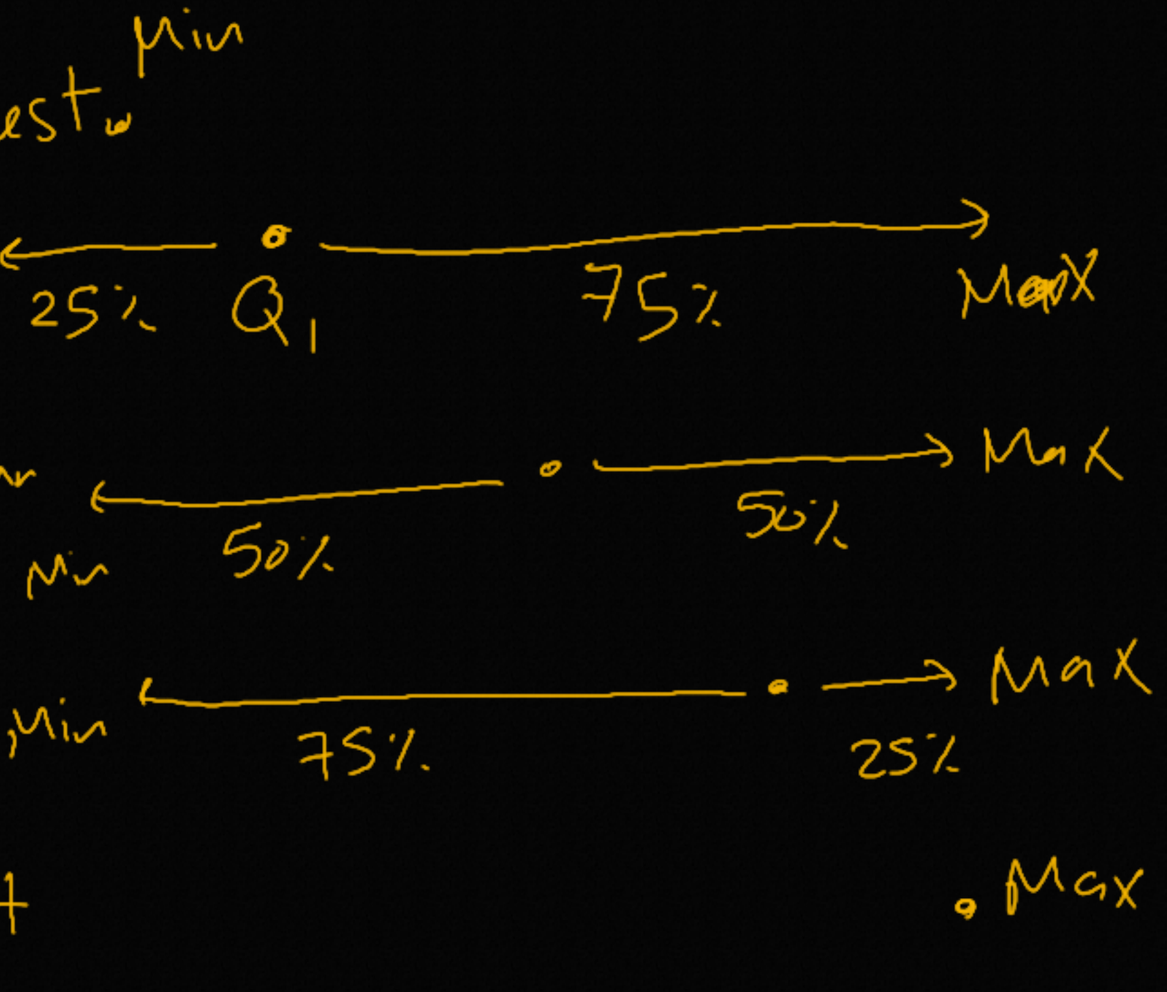
① Min. → Smallest

② Q1 → 25% of the data

③ Median (Q2) → 50% of the data

④ Q3 → 75% of the data

⑤ Max. → Highest



Avg

Disp (Range)

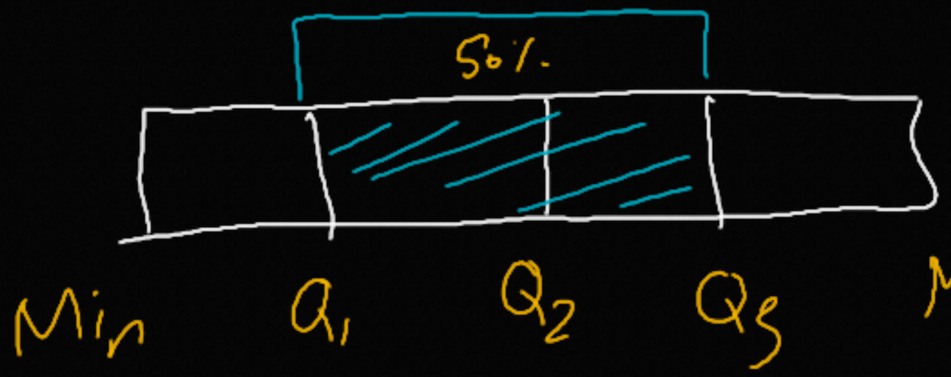
Col 9000 → 99000

1000 → 100 000

Col 9050 → 9000

6000 → 15000

Same ?



Dispersion (Spread) → 2 No

① Range : Max − Min

② IQR : $Q_3 - Q_1$ → 2 No.