EE2211 Introduction to Machine Learning

Lecture 2

Wang Xinchao
xinchao@nus.edu.sg

Course Contents

- Introduction and Preliminaries (Xinchao)
  - Introduction
  - Data Engineering
  - Introduction to Probability and Statistics
- Fundamental Machine Learning Algorithms I (Yueming)
  - Systems of linear equations
  - Least squares, Linear regression
  - Ridge regression, Polynomial regression
- Fundamental Machine Learning Algorithms II (Yueming)
  - Over-fitting, bias/variance trade-off
  - Optimization, Gradient descent
  - Decision Trees, Random Forest
- Performance and More Algorithms (Xinchao)
  - Performance Issues
  - K-means Clustering
  - Neural Networks

Summary of Lec 1

Three Components in ML Definition

Task T, Performance P, Experience E

Two Types of Supervised Learning

Classification, Regression

Three Types of in ML

Supervised Learning
Unsupervised Learning
Reinforcement Learning

One Type of Unsupervised Learning

Clustering

Inductive and Deductive

Inductive: Probable
Deductive: Rule-based

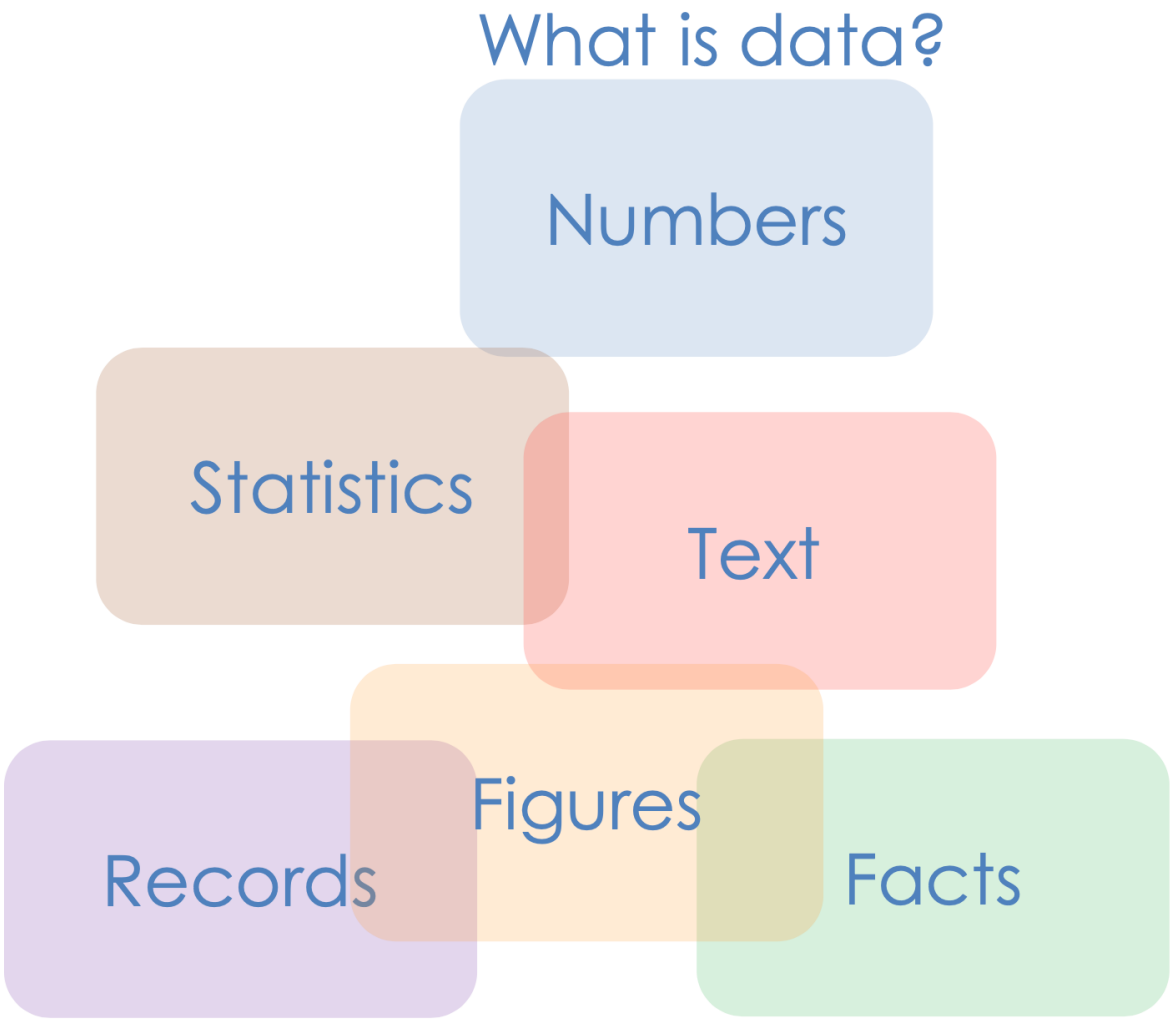
Example of a Classifier Model

Nearest Neighbor Classifier

Outline

- Types of data
- Data wrangling and cleaning
- Data integrity and visualization

Types of Data



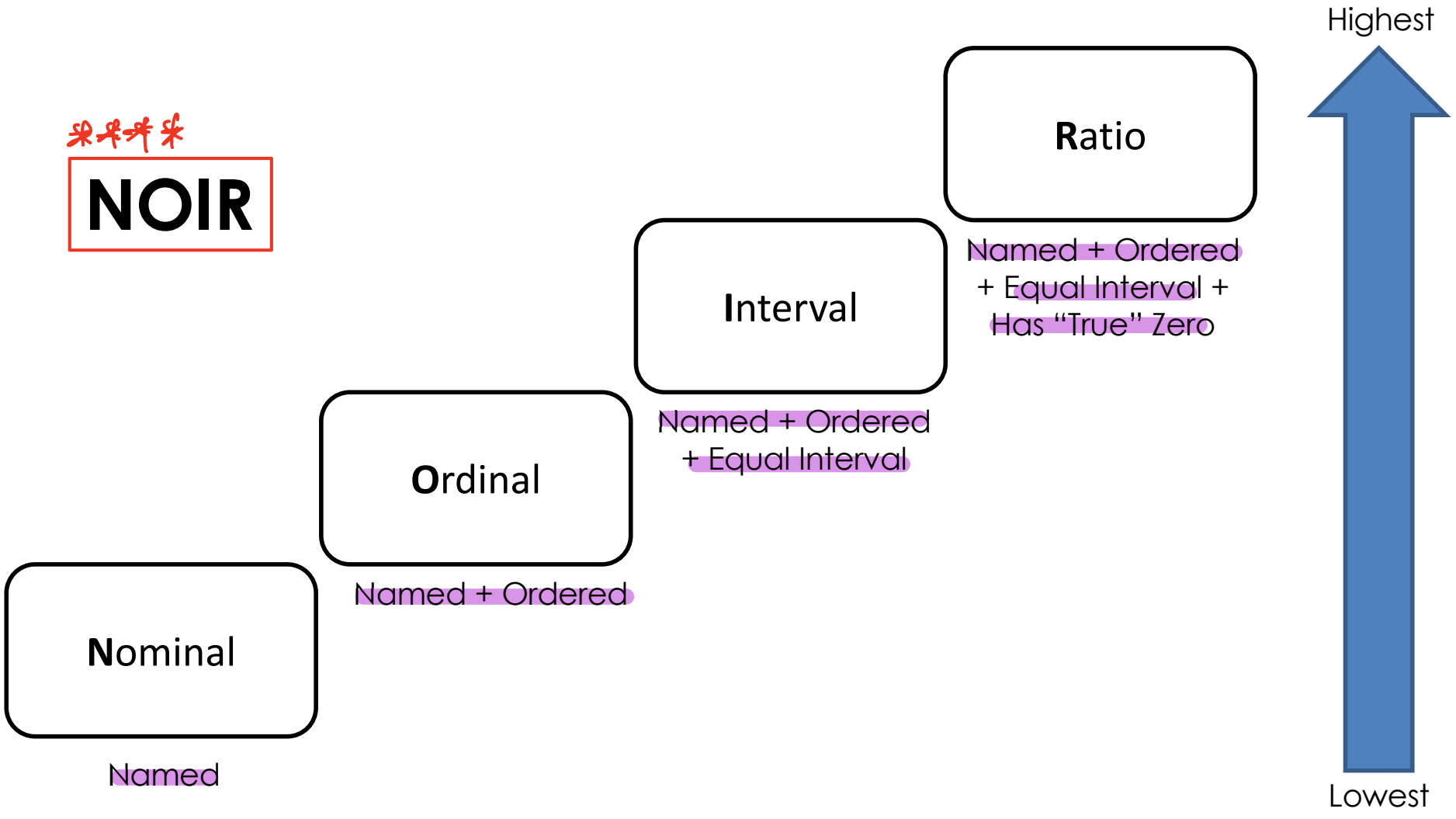
Ways of Viewing Data

- Based on Levels/Scales of Measurement
  - Nominal Data
  - Ordinal Data
  - Interval Data
  - Ratio Data

- Based on Numerical/Categorical
  - Numerical, also known as Quantitative
  - Categorical, also known as Qualitative

- Other aspects
  - Available or Missing Data

Levels/Scales of Measurement



A Quick Recap: Mean, Median, Mode

- If we are given a sequence of numbers:

1, 3, 4, 6, 6, 7, 8

Mean: computing the average

(1+3+4+6+6+7+8)/7 = 5

Median: number in the middle (after sorting)

1, 3, 4, 6, 6, 7, 8

*In case of even number of elements

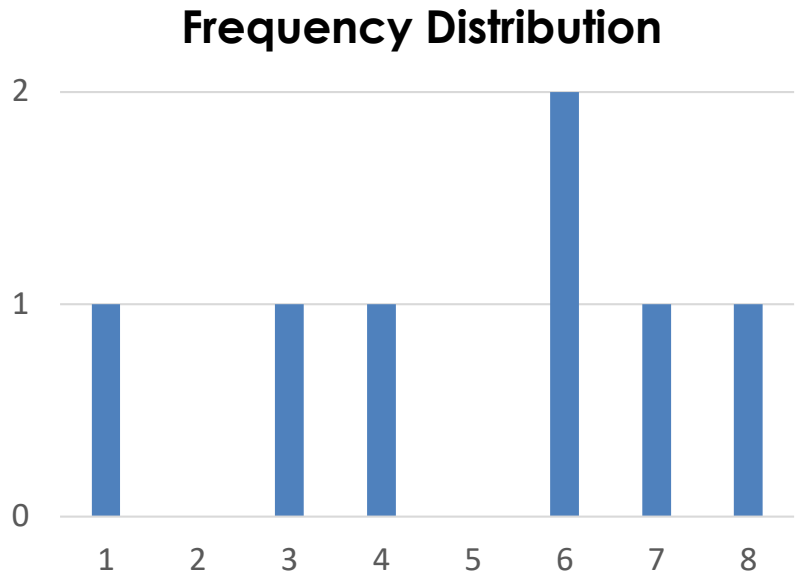
1, 3, 4, 6, 7, 8

(4+6)/2=5

Mode: number with highest frequency

1, 3, 4, 6, 6, 7, 8

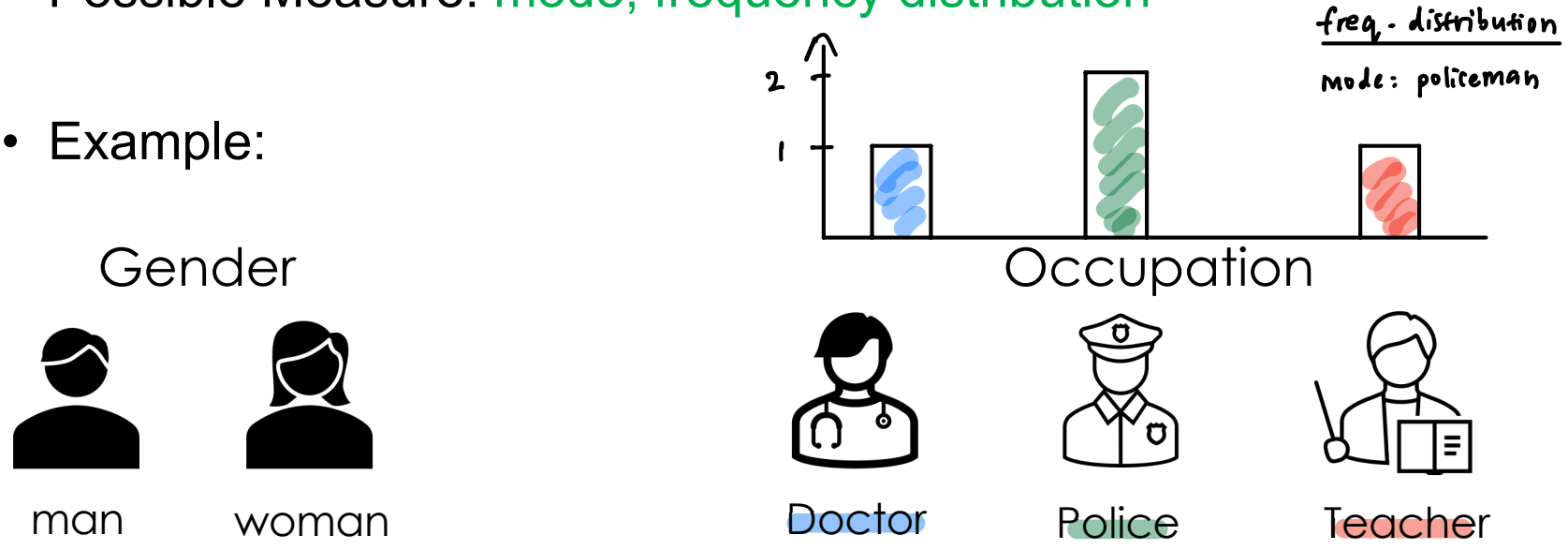
6



[NOIR]

Nominal Data

- Lowest Level of Measurement
- Discrete Categories
- **NO** natural order
- Estimating a mean, median, or standard deviation, would be meaningless.
- Possible Measure: mode, frequency distribution
- Example:



Ordinal Data

- **Ordered** Categories
- Relative Ranking
- Unknown "distance" between categories: orders matter but not the difference between values
- Possible Measure: mode, frequency distribution + median

- Example:
  - Evaluate the difficulty level of an exam
    - 1: Very Easy, 2: Easy, 3: About Right, 4: Difficulty, 5: Very Difficulty

Difference may be different, cannot compute, not well defined

Interval Data

- Ordered Categories
- Well-defined "**unit**" measurement:
  - Distances between points on the scale are measurable and well-defined
  - Can measure differences! ****
- **Equal Interval** (between two consecutive unit points)
- **Zero is arbitrary** (not absolute), in many cases human-defined
  - If the variable equals zero, it does not mean there is none of that variable
- **Ratio is meaningless** $\longrightarrow$ $\frac{20^\circ C}{10^\circ C} \rightarrow 2$ C no meaning, cannot do like that)
- Possible Measure: mode, frequency distribution + median + mean, standard deviation, addition/subtraction
- Example:
  - Temperature measured in Celsius
    - For instance: 10 degrees C, 28 degrees C
  - Year of someone's birth
    - For instance: 1990, 2005, 2010, 2022

e.g. if temp is 0°C, does not mean that there is no temperature

avg (10°C, 20°C) = 15°C

mean can be done hence, s.d. can be done

Ratio Data

- Most precise and **highest** level of measurement
- Ordered
- Equal Intervals
- **Natural Zeros**:
  - If the variable equals zero, it means there is none of that variable
  - Not arbitrary
- Possible Measure: mode, frequency distribution + median + mean, standard deviation, addition/subtraction + multiplication and division (ratio)

e.g. weight

$\frac{20kg}{10kg} = 2$ (have meaning, ie 2 times as heavy) **

$\frac{60\,sec}{10\,sec} = 6$ (period of 1min is 6 times of 10 sec)

- Example:
  - Weights
    - 10 KG, 20 KG, 30 KG

0kg = zero

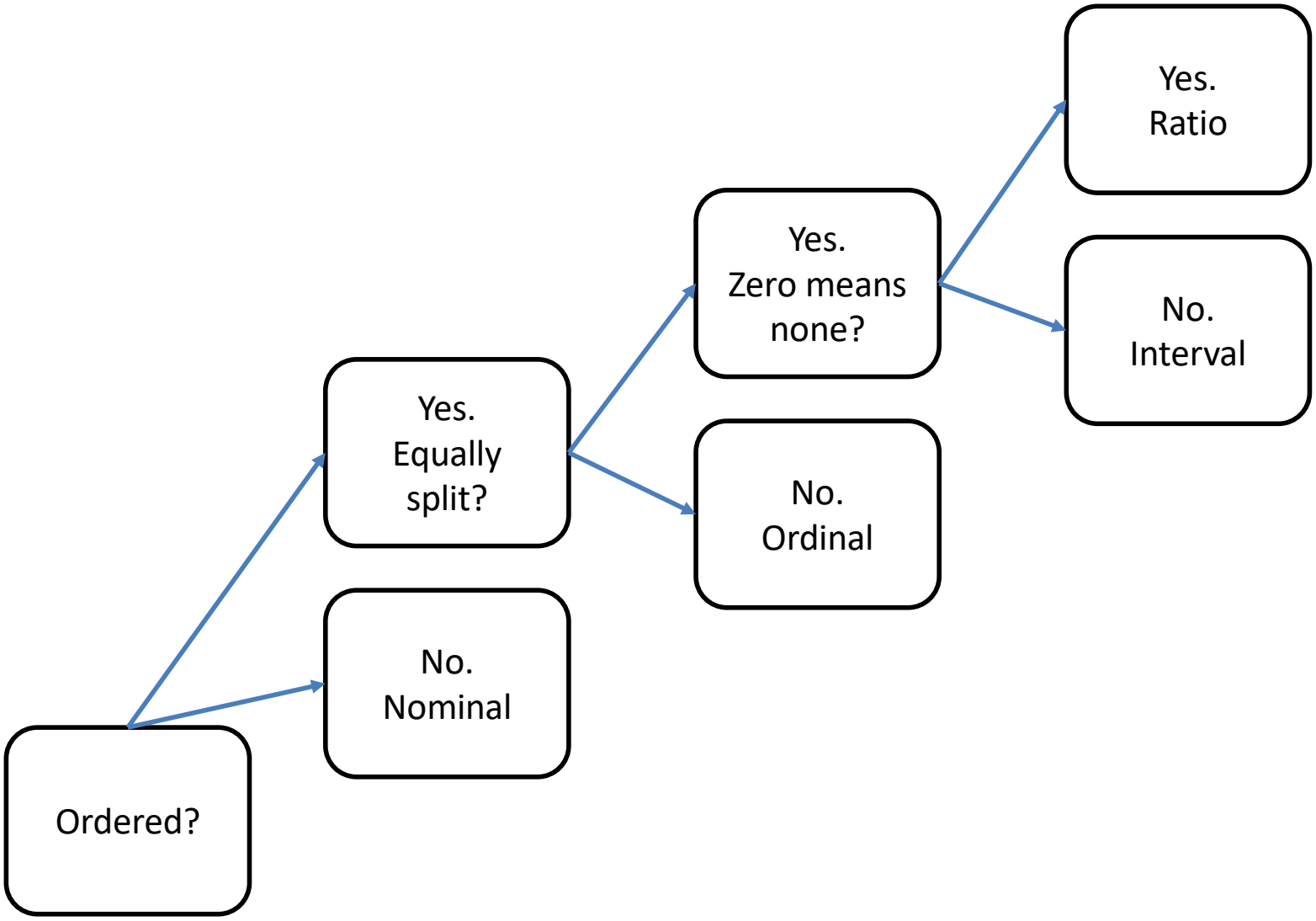
  - Time
    - 10 Seconds, 1 Hour, 1 Day

0 seconds = zero

NOIR

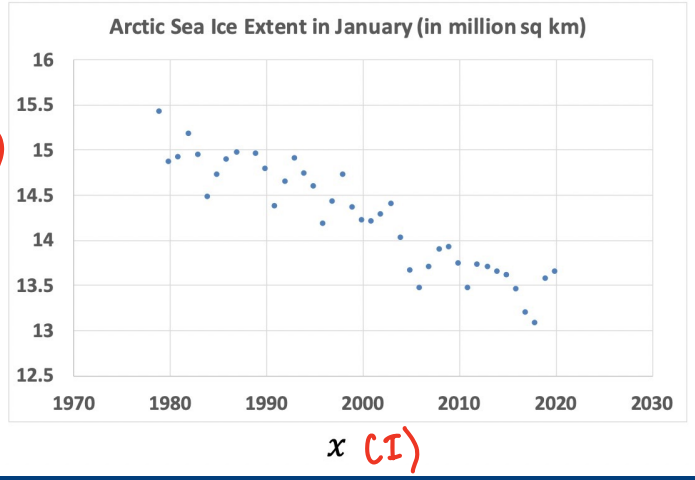
We can estimate	Nominal	Ordinal	Interval	Ratio
Frequency (mode) Distribution	Yes	Yes	Yes	Yes
Median	No	Yes	Yes	Yes
Add or subtract	No	No	Yes	Yes
Mean, standard deviation	No	No	Yes	Yes
Ratios	No	No	No	Yes

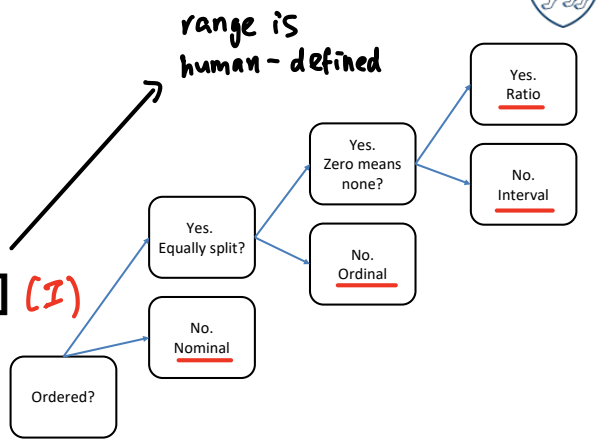
NOIR



- Which level of measurement?

Nominal, Ordinal, Interval, Ratio

1. **Favorite Restaurant**
   - Mcdonald's, Burger King, Subway, KFC, … (N)
2. **Weight of luggage measured in KG** (R)
3. **SAT Scores: note that, SAT ranges is [400, 1600]** (I)
4. **Size of Packed Eggs in supermarkets**
   - Small, Medium, Large, Extra Large, … (O)
5. **Military rank**
   - General, Major, Captain, … (O)
6. **Number of people in a household** → if 0 people in household, means zero,
   $\frac{3 \text{ people}}{1 \text{ people}} = 3$ (household is 3 times more people)
   - 1, 2, 3, 4, 5, … (R)
7. **Credit Score in United States: the range is [300, 850]** (I)
8. 



From this graph,

(Input) f(x) = y (output)

I R

different levels of measurement (no need to be same)

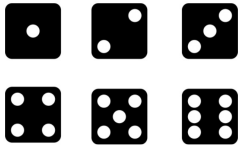
Ways of Viewing Data

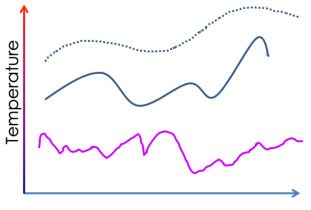
- Based on Levels/Scales of Measurement:
  - Nominal Data
  - Ordinal Data
  - Interval Data
  - Ratio Data

- Based on Numerical/Categorical
  - Numerical, also known as Quantitative
  - Categorical, also known as Qualitative

- Other aspects
  - Available or Missing Data

Numerical or Categorical





- Variable
  - Categorical (qualitative)
    - **Nominal** (Unordered categories)
    - **Ordinal** (Ordered categories)
  - Numerical (quantitative)
    - **Discrete** (Whole numerical values) **Example:** outcome of tossing a die
    - **Continuous** (Can take any value within a rage) **Example:** temperature in day

1st way

OR

- Numerical (quantitative)
  - **Interval** (may compute difference but no absolute zero)
  - **Ratio** (may compute difference, real zero exists)

2nd way

Ways of Viewing Data

- Based on Levels/Scales of Measurement:
  - Nominal Data
  - Ordinal Data
  - Interval Data
  - Ratio Data

- Based on Numerical/Categorical
  - Numerical, also known as Quantitative
  - Categorical, also known as Qualitative

- Other aspects
  - Available or Missing Data

Missing Data

- Missing data: data that is missing and you do not know the mechanism.
  - You should use a single common code for all missing values (for example, “NA”), rather than leaving any entries blank.

NUS student	Age	Country of birth
Olivia Tan	20	Singapore
Hendra Setiawan	19	Indonesia
John Smith	19	NA

Outline

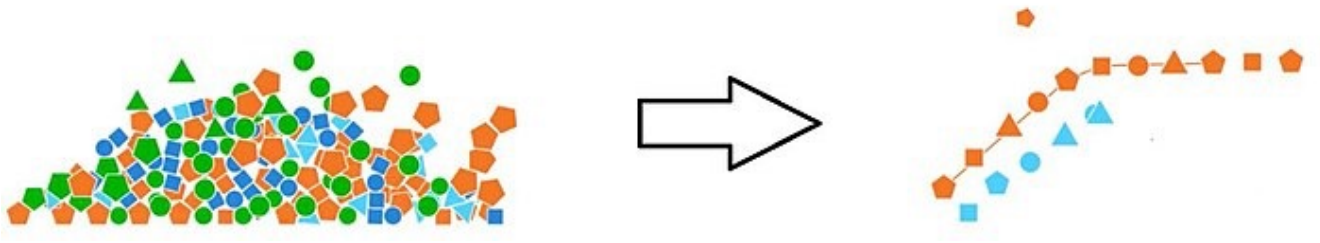
Types of data

Data wrangling and cleaning

Data integrity and visualization

Data Wrangling

- Data wrangling
  - The process of transforming and mapping data from one "raw" data form into another format with the intent of making it more appropriate and valuable for a variety of downstream purposes such as analytics.
  - In short, transforms data to gain insight
  - It is a general process!



Credit:https://en.wikipedia.org/wiki/Data_wrangling

Data Wrangling → happens before ML

[Start]

[End]

Determine your goal

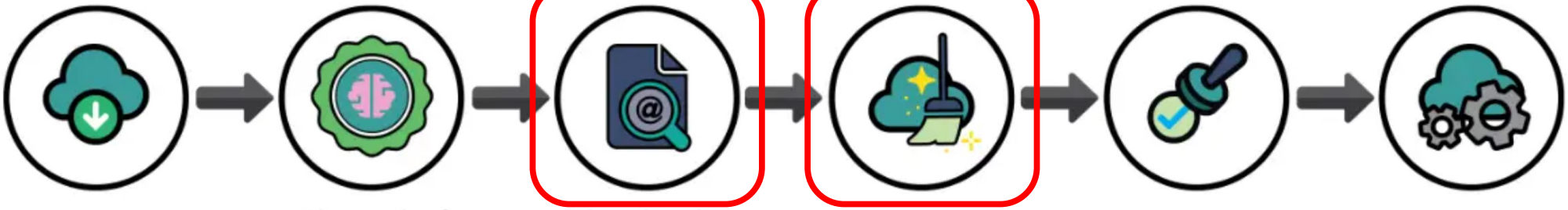
Make the most of dataset

Example: social network stored as graphs

Remove invalid data

Ensure data correctness

Use data (feature extraction/ training/test)



Extracting the Data

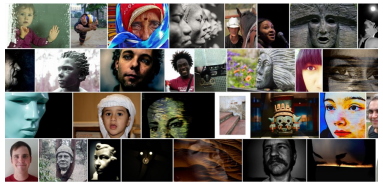
Discovering/ Analysing the Structure of the Data

Choosing the Correct Format for the Data

Cleaning

Validating

Deploying



Women

Men

Unifying format to .png

Remove noisy samples

Checking that all images have labels.

Use data for feature extraction and training your face detector!

Collect Human Face Images for Face Detector

Credit:https://understandingdata.com/what-is-data-wrangling/

Formatting Data

Multiple categories can use multiple one-hot encoding

(classes)

- **Binary Coding** to convert categories into binary form
  - One-hot encoding: unify several entities within one vector
    - Example: the color of a pixel can be red, yellow, or green
    - Very common in classification tasks!

R Y G

$$\text{red} = [1, 0, 0]$$
$$\text{yellow} = [0, 1, 0]$$
$$\text{green} = [0, 0, 1]$$

- **Normalization**
  - Linear Scaling:
    scale each variable to [0 1]

[1, 3, 5, 9, 11, 17 25]

$$x_i = \frac{x_i^{raw} - x^{min}}{x^{max} - x^{min}}, \quad i = 1, 2, ..., M$$

scaled between [0 1]

  - Z-score standardization:
    each independent dimension of data is normally distributed

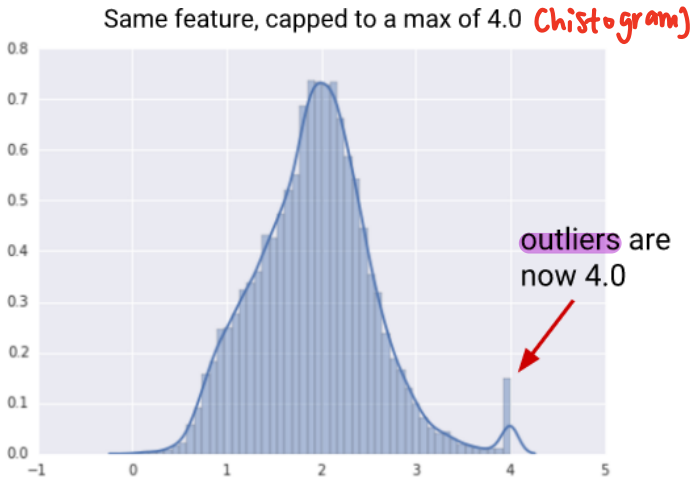
$$x_i = \frac{x_i^{raw} - E[X]}{\sigma(X)}, \quad i = 1, 2, ..., M.$$

→ mean

→ s.d.

Data Cleaning

- The process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database.
- Example:
  - Clipping outliers



*** Take note

  - Handling missing features

↓

how to handle the 'NA'

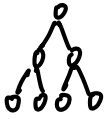
Students	Year of Birth	Gender	Height	GPA
Tan Ah Kow	1995	M	1.72	4.2
Ahmad Abdul	X NA	M	1.65	4.1
John Smith	1995	M	1.75	X NA
Chen Lulu	1995	F	X NA	4.0
Raj Kumar	1995	M	1.73	4.5
Li Xiuxiu	1994	F	1.70	3.8

Data Cleaning: Handling missing features

1. Removing the examples with missing features from the dataset
   - Can be done if the dataset is big enough so we can sacrifice some training examples

↓ if small dataset, should not remove

2. Using a learning algorithm that can deal with missing feature values
   - Example: random forest → Decision Trees



↳ random forest is NOT filling up missing data entry

3. Using a data imputation technique

Data Cleaning: Handling missing features: **Imputation**

- Method 1. Replace the missing value of a feature by an average value of this feature in the dataset:

$$\hat{x}^{(j)} \leftarrow \frac{1}{N}\sum_{i=1}^{N} x_i^{(j)}$$

↳ take avg value of data in dataset

- Method 2. Highlight the missing value (commonly used)
  - Replace the missing value with a value outside the normal range of values.
  - For example, if the normal range is [0, 1], then you can set the missing value to −1.
  - Enforce the learning algorithm to learn what is best to do when the feature has a value significantly different from regular values.

Outline

Types of data

Data wrangling and cleaning

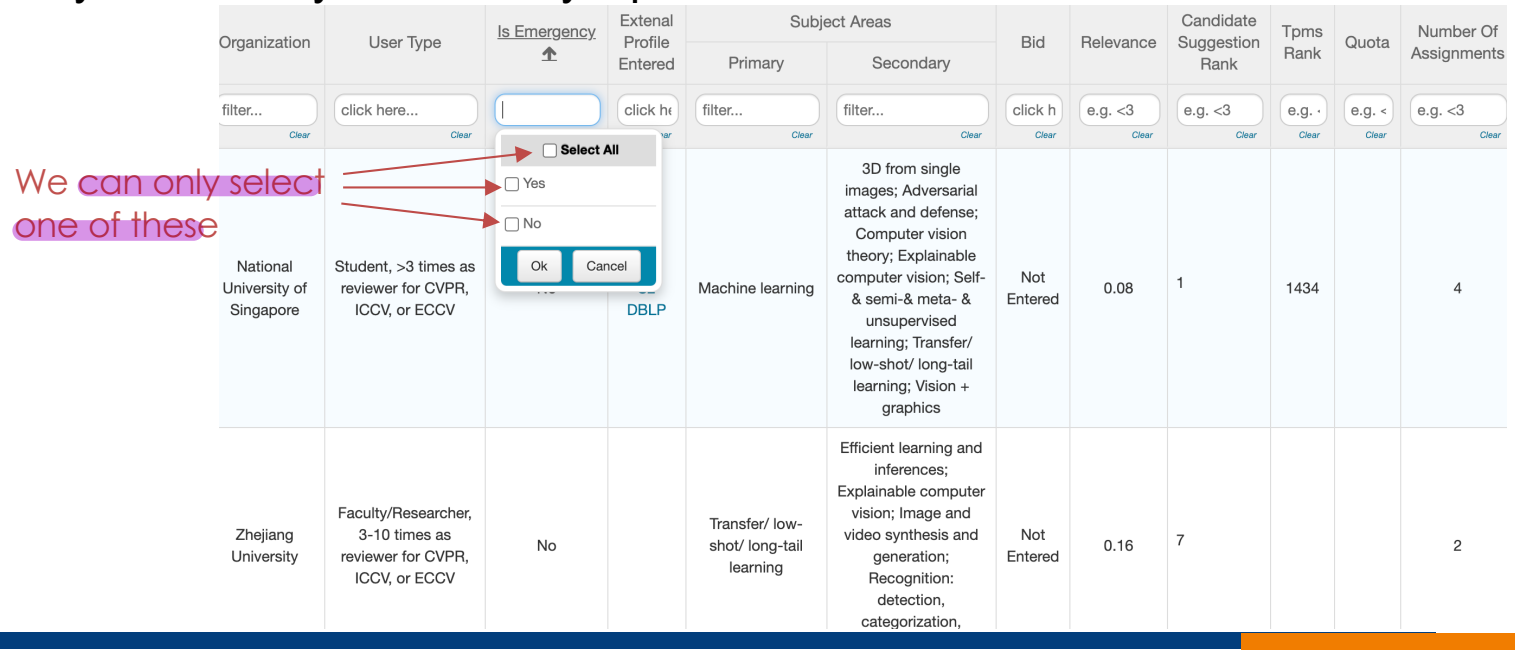
Data integrity and visualization

Data Integrity

- Data integrity is the maintenance and the assurance of data accuracy and consistency;
  - A critical aspect to the design, implementation, and usage of any system that stores, processes, or retrieves data.
  - Very broad concept!
- Example:
  - In a dataset, numeric columns/cells should not accept alphabetic data.
  - A binary entry should only allow binary inputs

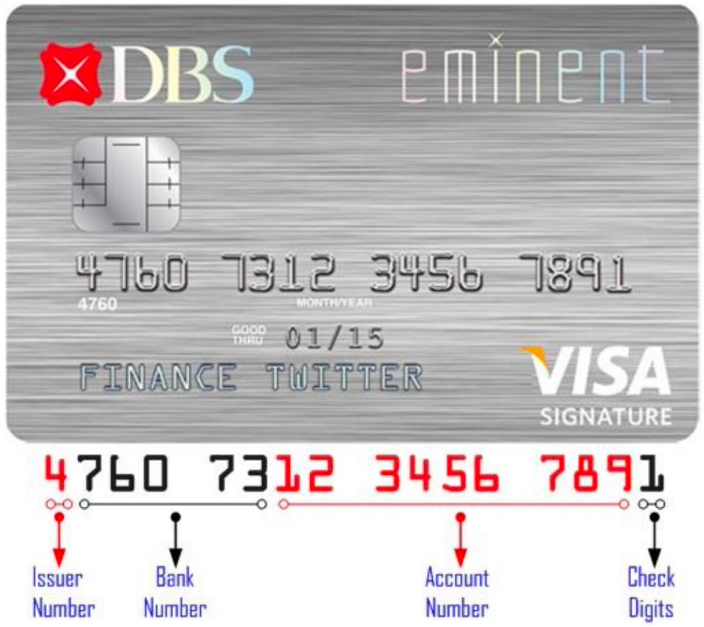
CMT

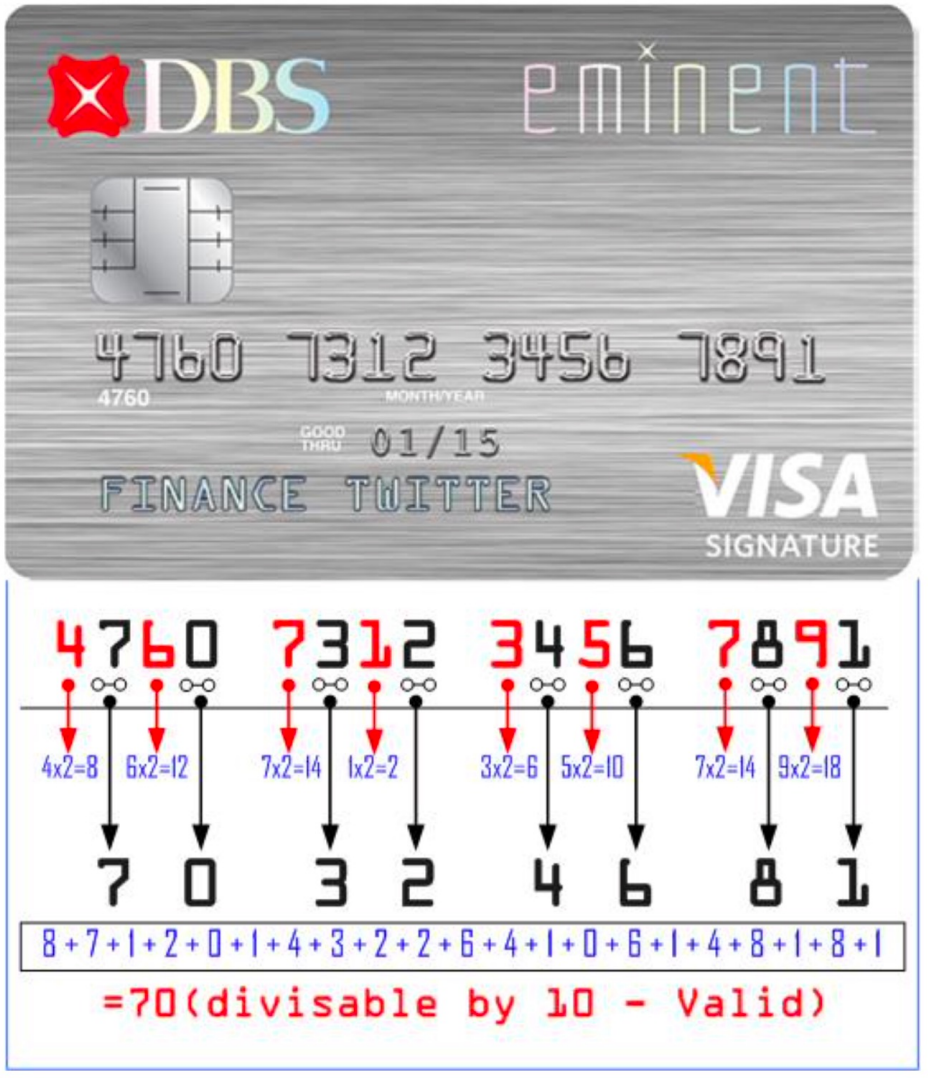
We can only select one of these



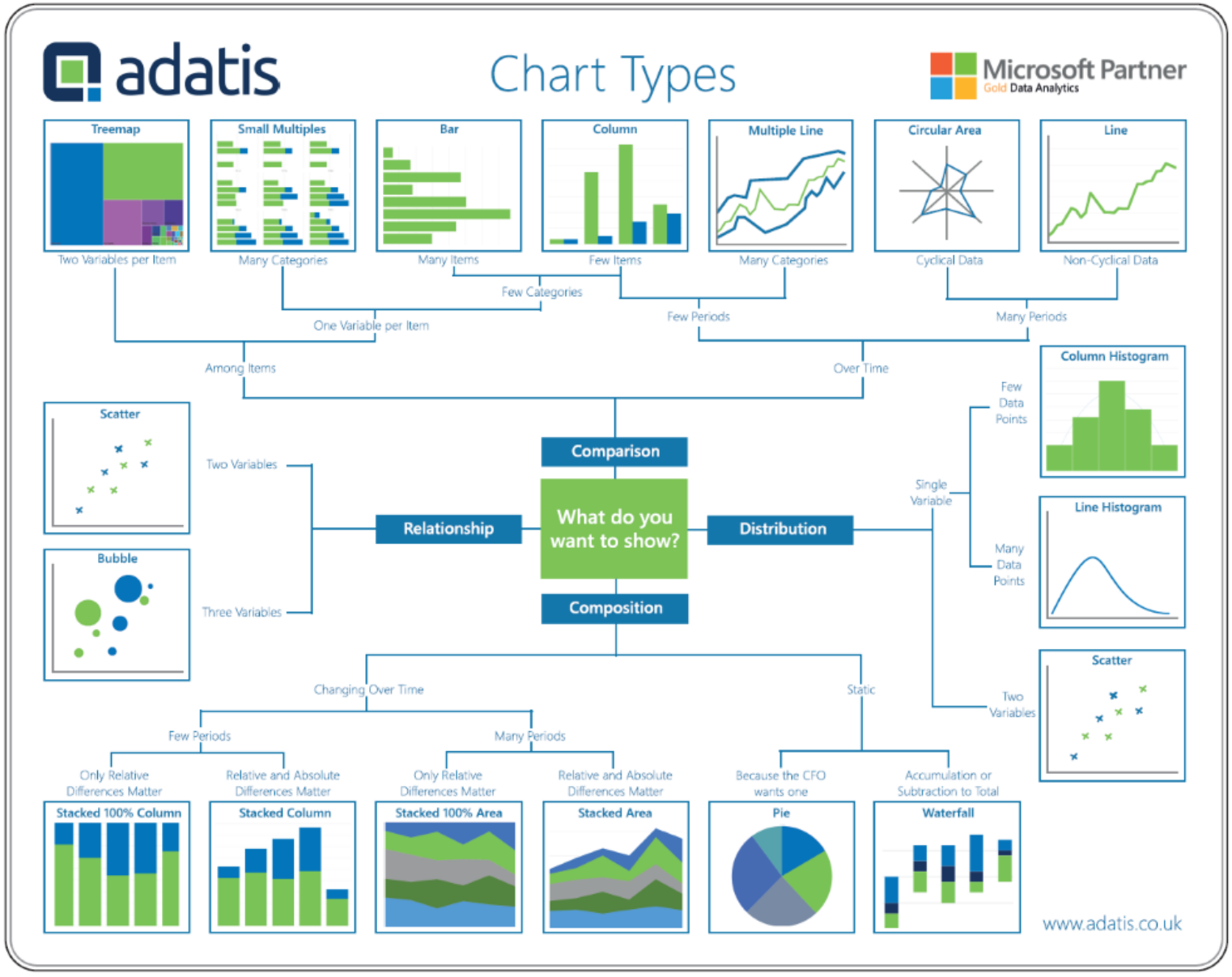
Organization	User Type	Is Emergency	Extenal Profile Entered	Subject Areas		Bid	Relevance	Candidate Suggestion Rank	Tpms Rank	Quota	Number Of Assignments
				Primary	Secondary						
filter...	click here...		click he	filter...	filter...	click h	e.g. <3	e.g. <3	e.g. ·	e.g. <	e.g. <3
National University of Singapore	Student, >3 times as reviewer for CVPR, ICCV, or ECCV		DBLP	Machine learning	3D from single images; Adversarial attack and defense; Computer vision theory; Explainable computer vision; Self- & semi-& meta- & unsupervised learning; Transfer/ low-shot/ long-tail learning; Vision + graphics	Not Entered	0.08	1	1434		4
Zhejiang University	Faculty/Researcher, 3-10 times as reviewer for CVPR, ICCV, or ECCV	No		Transfer/ low-shot/ long-tail learning	Efficient learning and inferences; Explainable computer vision; Image and video synthesis and generation; Recognition: detection, categorization,	Not Entered	0.16	7			2

Data Integrity



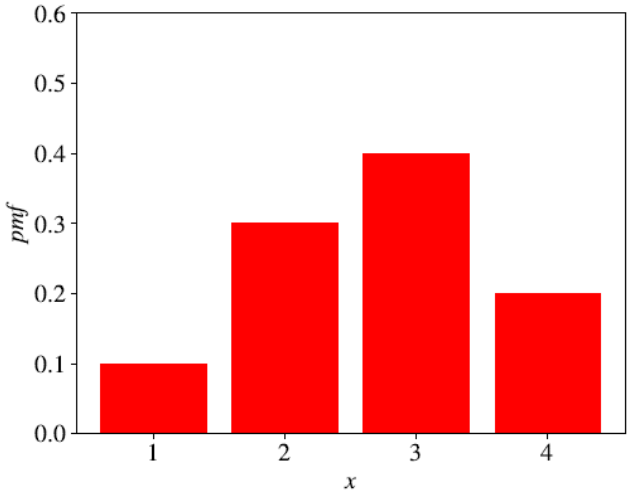


Data Visualization

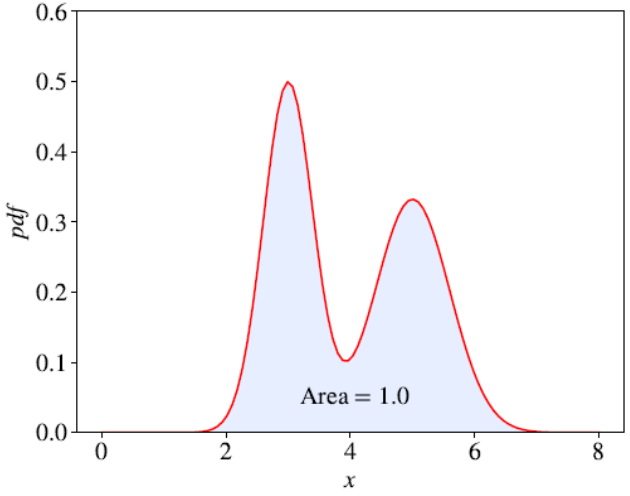


Graphical Representation of data!

Visualization: Distribution

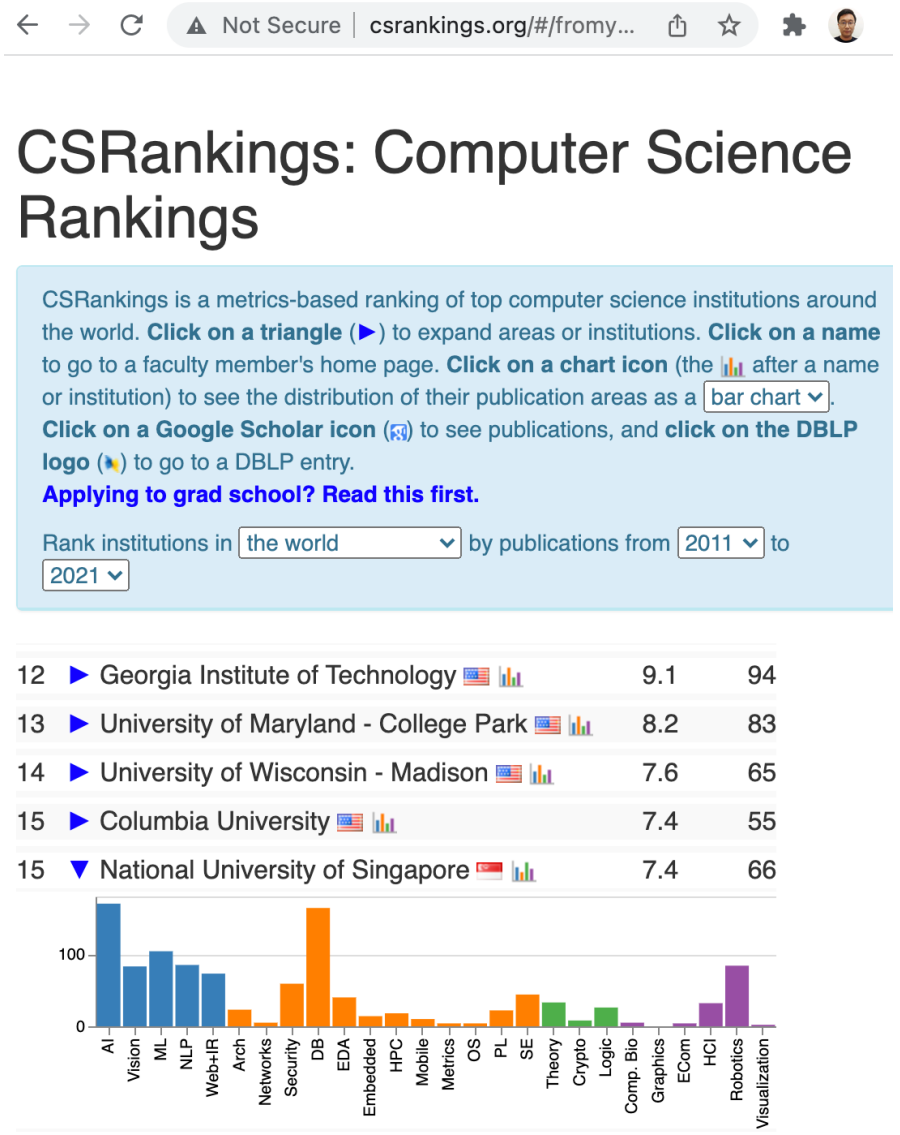


Probability Mass Function (Discrete random)

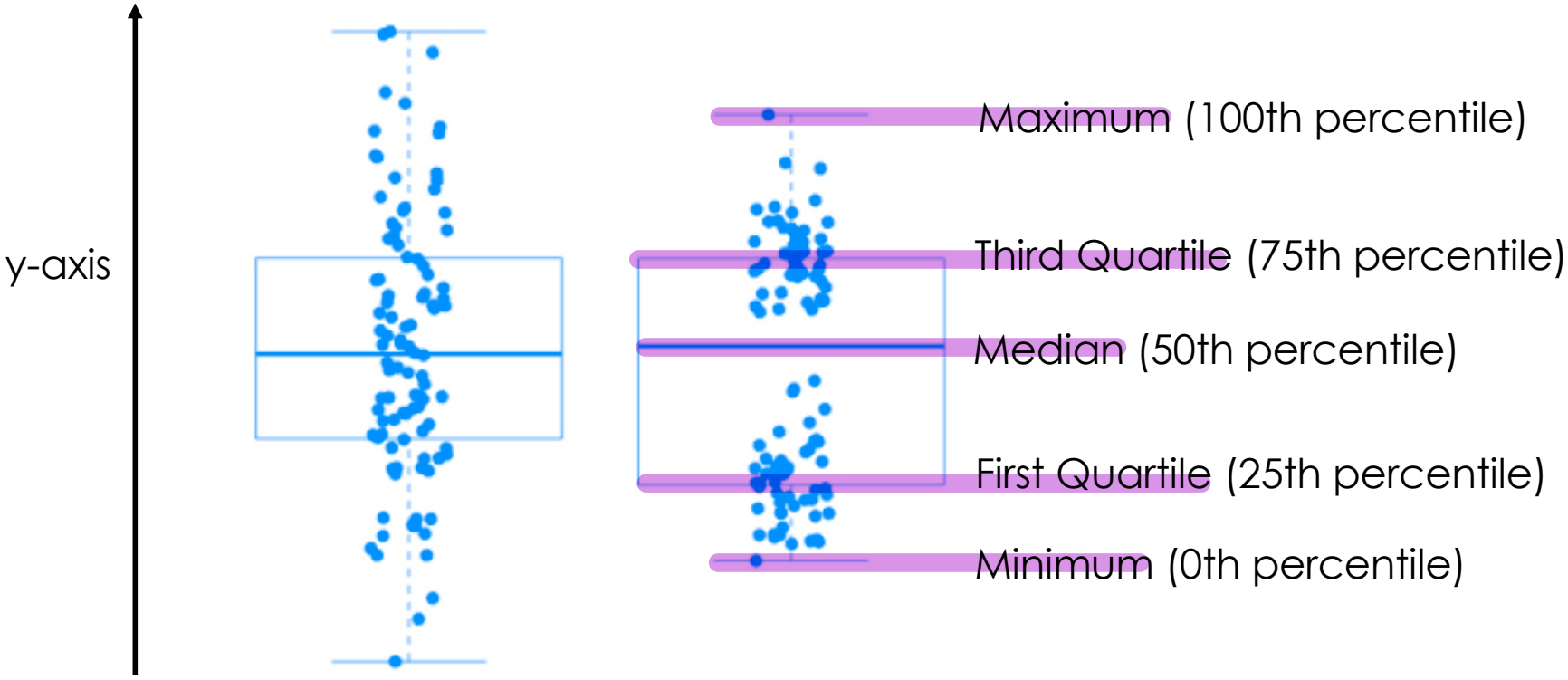


Probability Density Function (Continuous variables)

Visualization: Bars



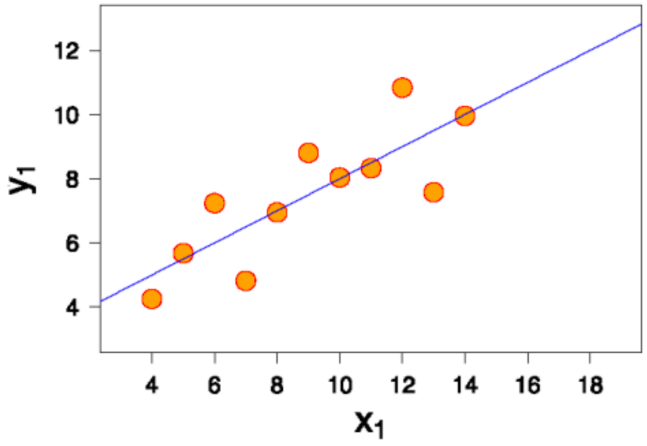
Visualization: Boxplots

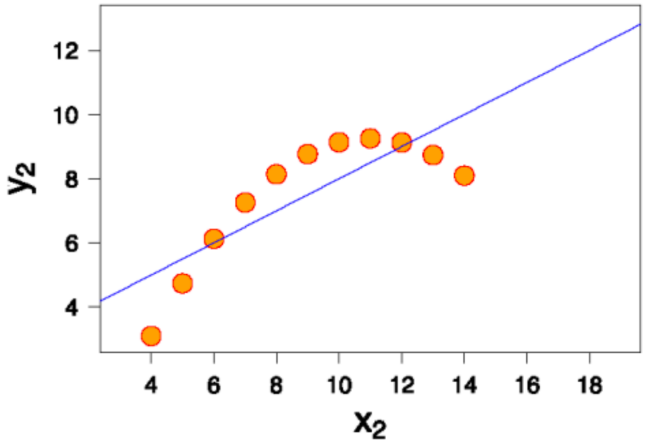


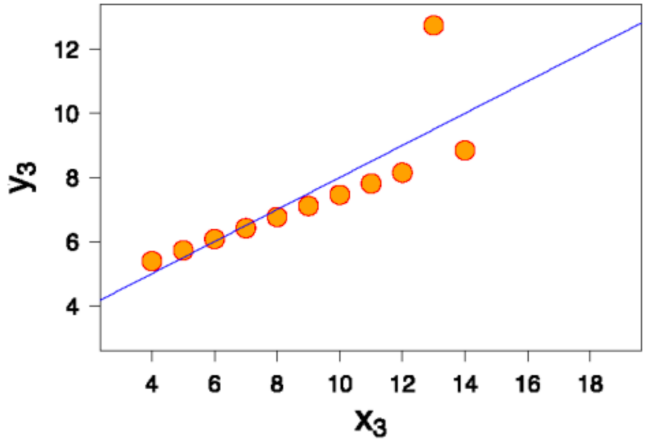
• The first quartile (Q_1) is defined as the middle number between the smallest number (i.e., Minimum) and the median of the data set.

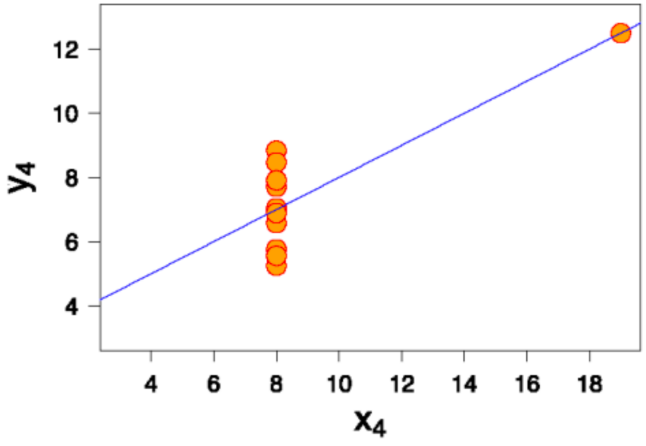
• The third quartile (Q3) is the middle number between the median and the highest value (i.e., Maximum) of the data set.

Why Visualization is Necessary









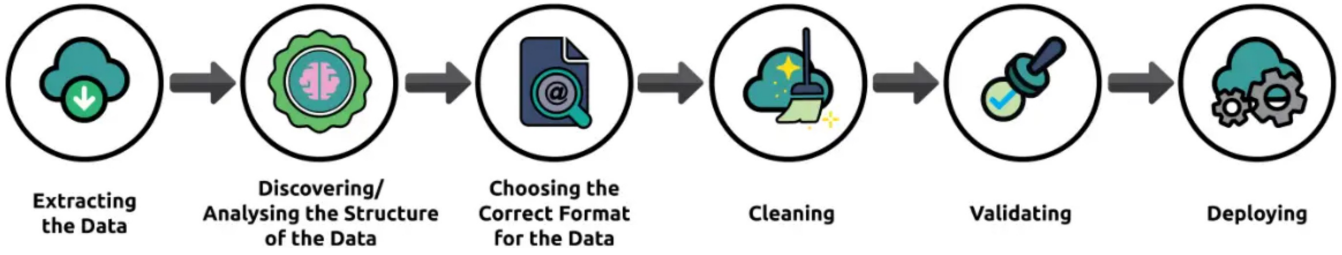
Four datasets with **identical** means, variances and regression lines!

****

Hence, we need visualization to show their difference!

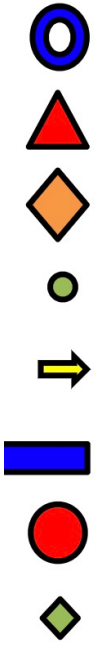
Summary

- Types of data
  - NOIR
- Data wrangling and cleaning



- Data integrity and visualization
  - Integrity: Design
  - Visualization: Graphical Representation

Practice Question
(Type of Question to Expect in Exams)



Color	Size	Shape
Blue	Large	Ring
Red	Large	Triangle
Orange	Large	Diamond
Green	Small	Circle
Yellow	Small	Arrow
Blue	Large	Rectangle
Red	Large	Circle
Green	Small	Diamond

What are the NOIR data types of *color*, *size*, and *shape* in the table?

Color : nominal (red, green, blue) → no order

Size : ordinal (XS, S, M, L) → have order but difference unknown

Shape : nominal (triangle, ring, diamond) → no order

What about their label, *yes*/*no*?

Label : Nominal (no order between yes/no)