

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

Three Types of in ML

Supervised Learning
Unsupervised Learning
Reinforcement Learning

Two Types of Supervised Learning

Classification, Regression

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

What is data?

Numbers

Statistics

Text

Records

Figures

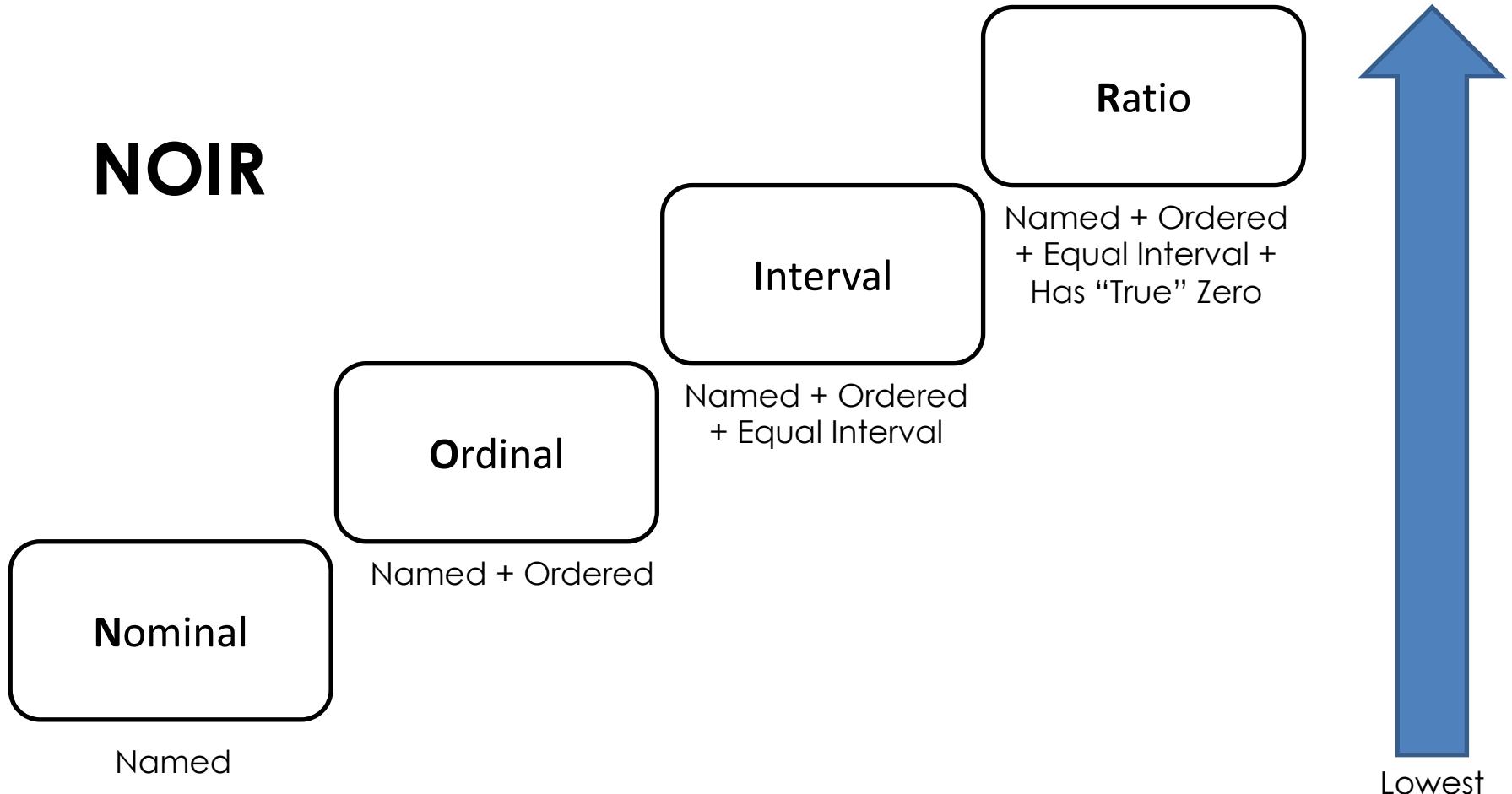
Facts

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

NOIR



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

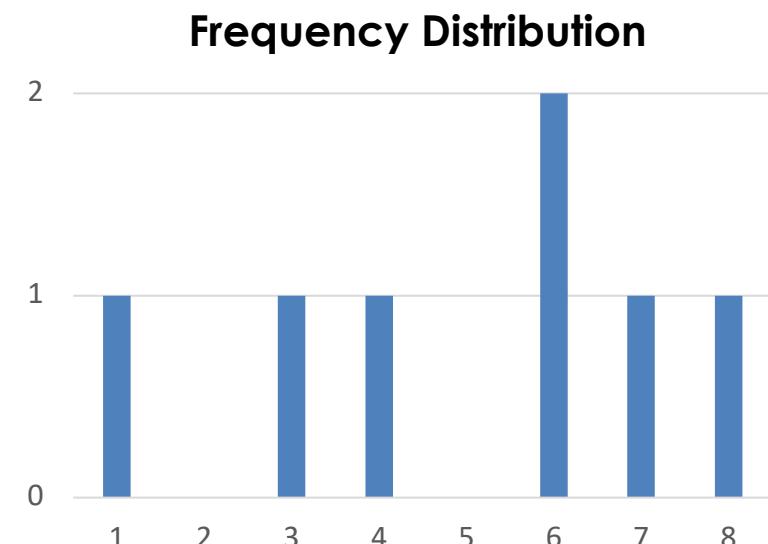
1, 3, 4, 6, 7, 8

$$(4+6)/2=5$$

Mode: number with highest frequency

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

6



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:

Gender



man



woman

Occupation



Doctor



Police



Teacher

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: Difficult, 5: Very Difficult

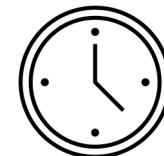
Interval Data

- 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
- 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



Ratio Data

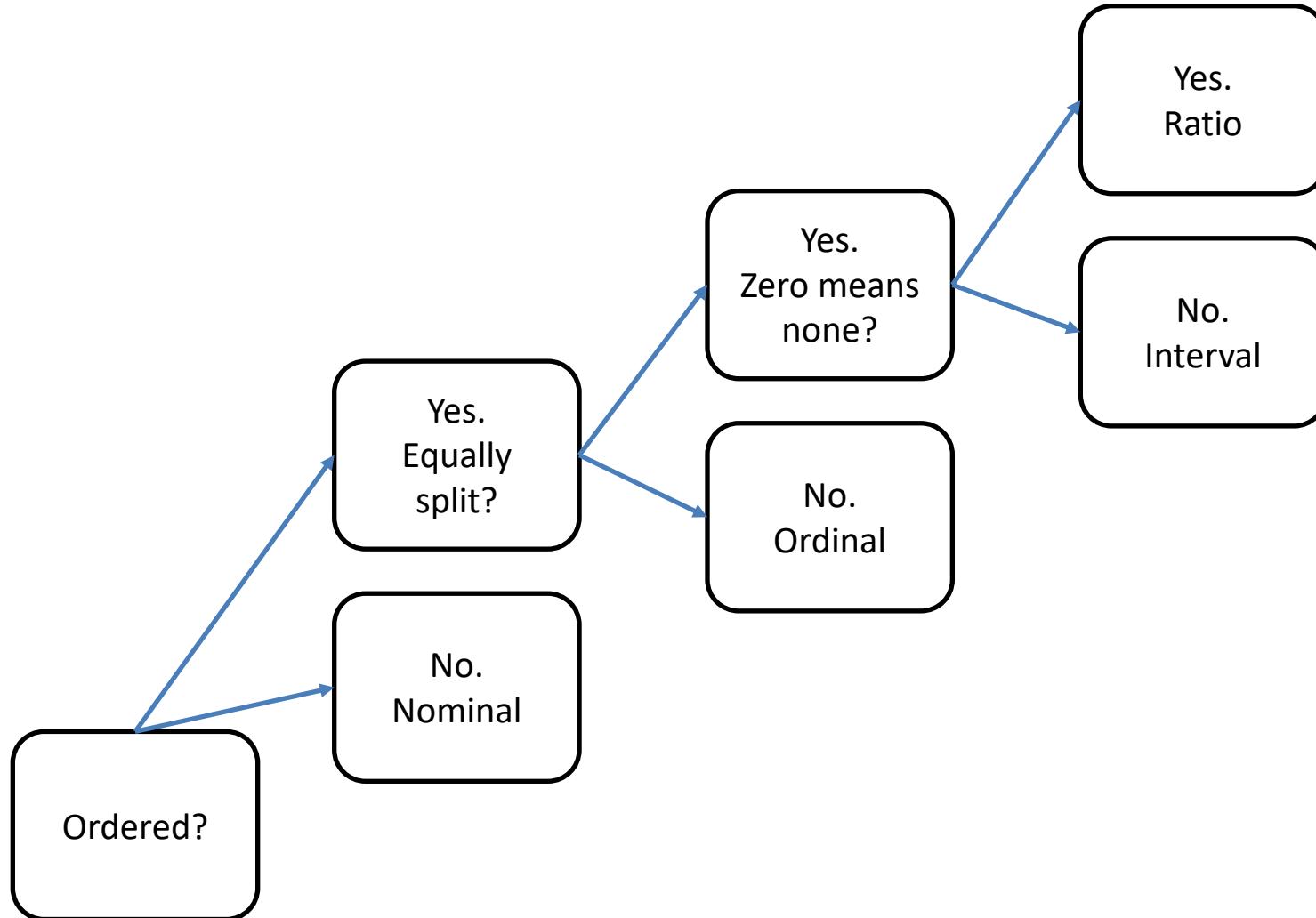
- Most precise and **highest** level of measurement
- Ordered
- Equal Intervals
- **Natural Zeros:**
 - If the variable equals zero, it means there is **none** of that variable
 - Not arbitrary
- Possible Measure: mode, frequency distribution + median + mean, standard deviation, addition/subtraction + multiplication and division (ratio)
- Example:
 - Weights
 - 10 KG, 20 KG, 30 KG
 - Time
 - 10 Seconds, 1 Hour, 1 Day



NOIR

We can estimate	Nominal	Ordinal	Interval	Ratio
Frequency 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

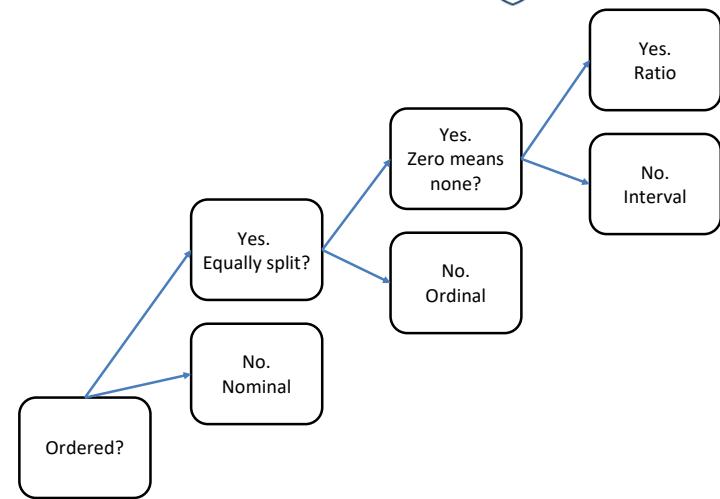
NOIR



Quick Quiz

- Which level of measurement?
Nominal, Ordinal, Interval, Ratio

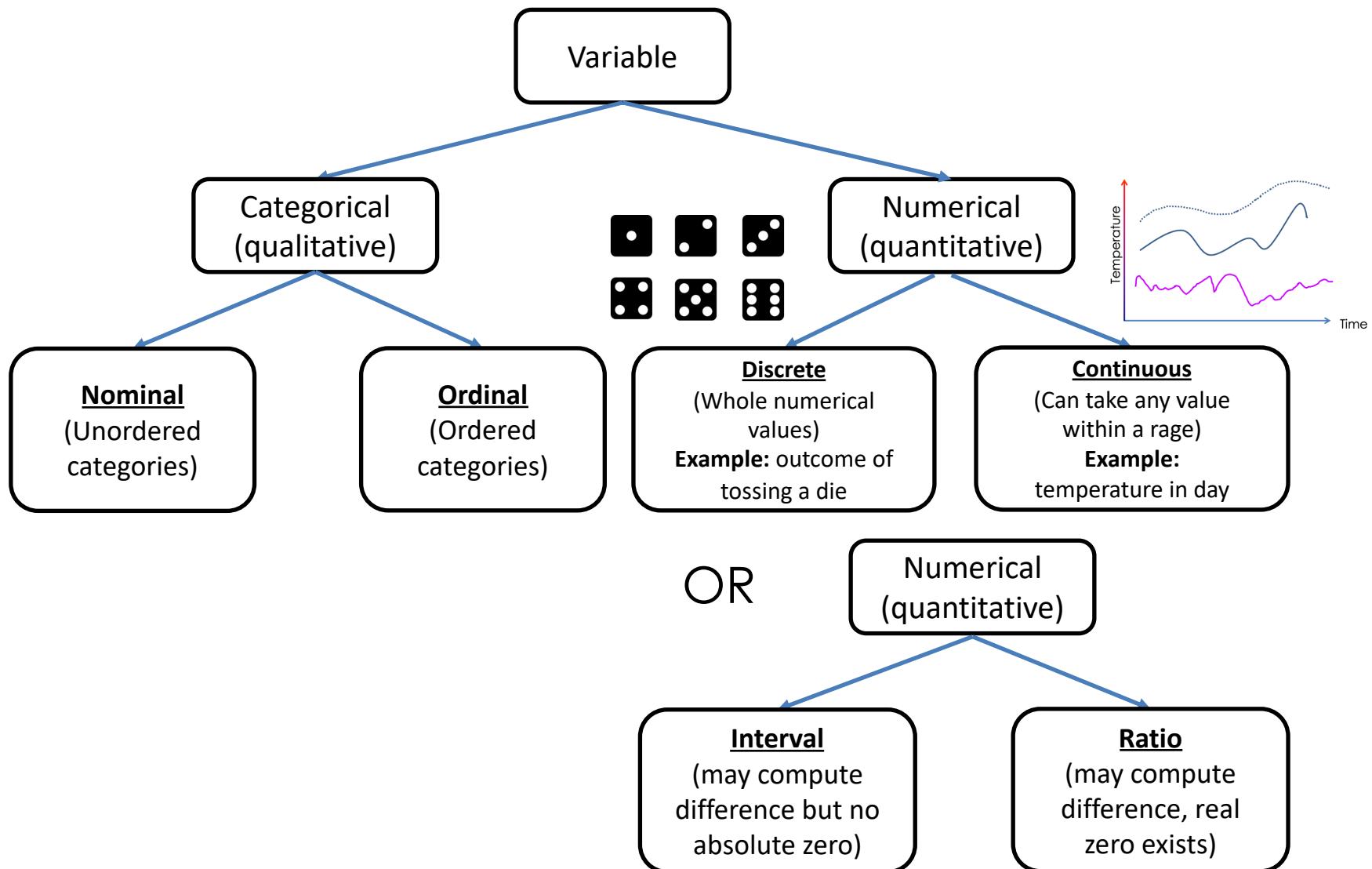
1. **Favorite Restaurant**
 - Mcdonald's, Burger King, Subway, KFC, ...
2. **Weight of luggage measured in KG**
3. **SAT Scores: note that, SAT ranges is [400, 1600]**
4. **Size of Packed Eggs in supermarkets**
 - Small, Medium, Large, Extra Large, ...
5. **Military rank**
 - General, Major, Captain, ...
6. **Number of people in a household**
 - 1, 2, 3, 4, 5, ...
7. **Credit Score in United States: the range is [300, 850]**



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



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

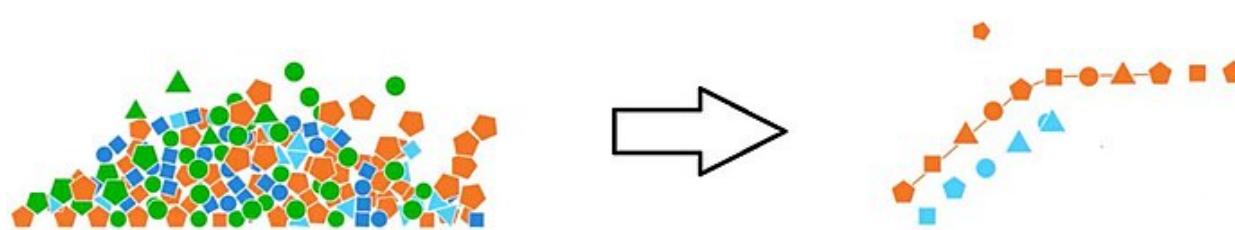
Types of data

Data
wrangling and
cleaning

Data integrity
and
visualization

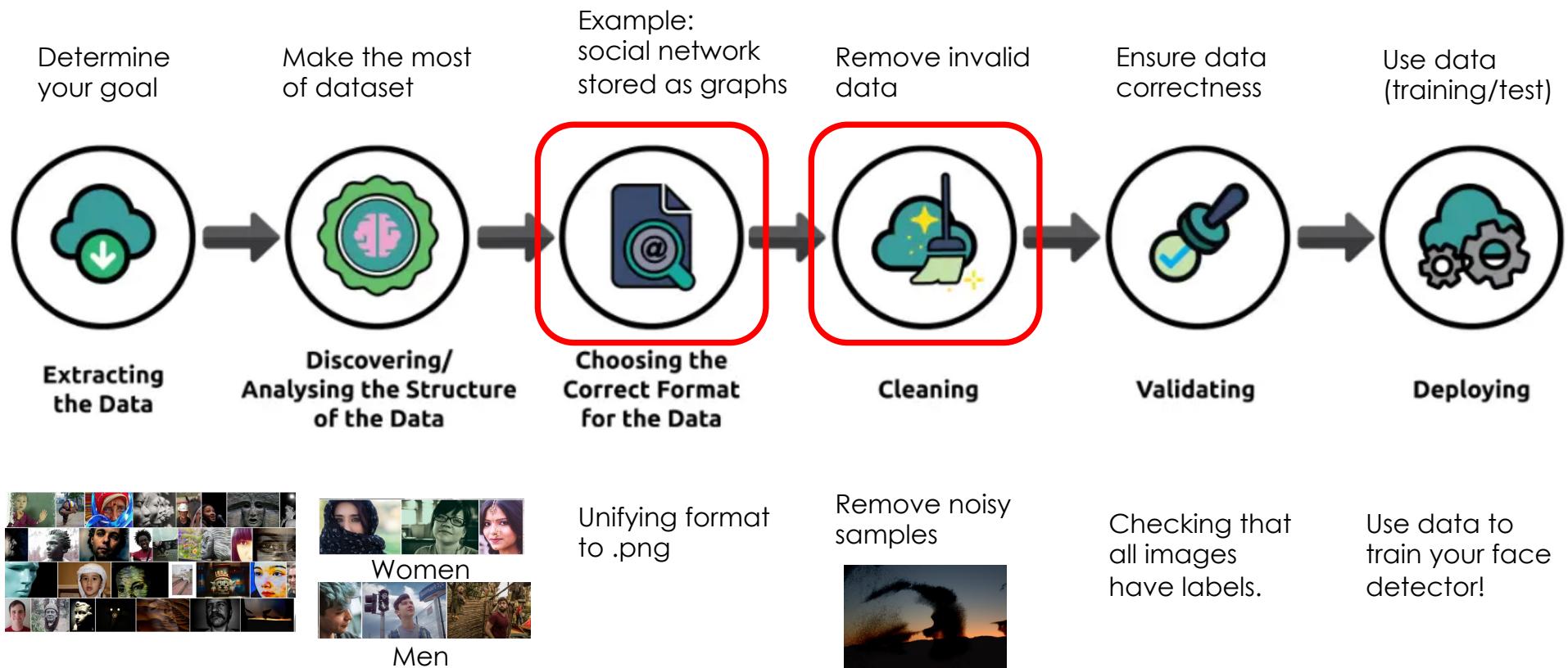
Data Wrangling

- 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



Collect Human Face Images for Face Detector

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

Formatting Data

- **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!

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

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

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

• Normalization

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

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

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

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

Data Cleaning

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

- Example:
 - Clipping outliers



- Handling missing features

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

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

- Method 2. Highlight the missing value
 - 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

We can only select one of these

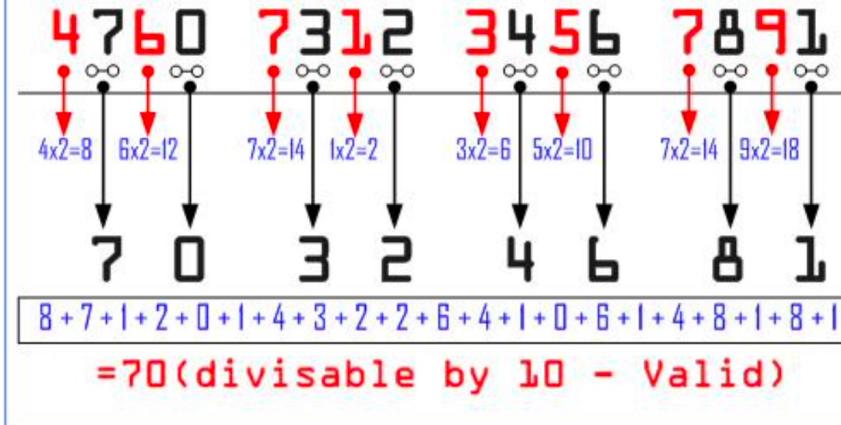
Organization	User Type	Is Emergency	External Profile Entered	Subject Areas		Bid	Relevance	Candidate Suggestion Rank	Tpms Rank	Quota	Number Of Assignments
				Primary	Secondary						
National University of Singapore	Student, >3 times as reviewer for CVPR, ICCV, or ECCV	<input type="checkbox"/> Yes <input type="checkbox"/> No		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	Not Entered	0.16	7				2

Data Integrity



4760 7312 3456 7891

Issuer Number Bank Number Account Number Check Digits

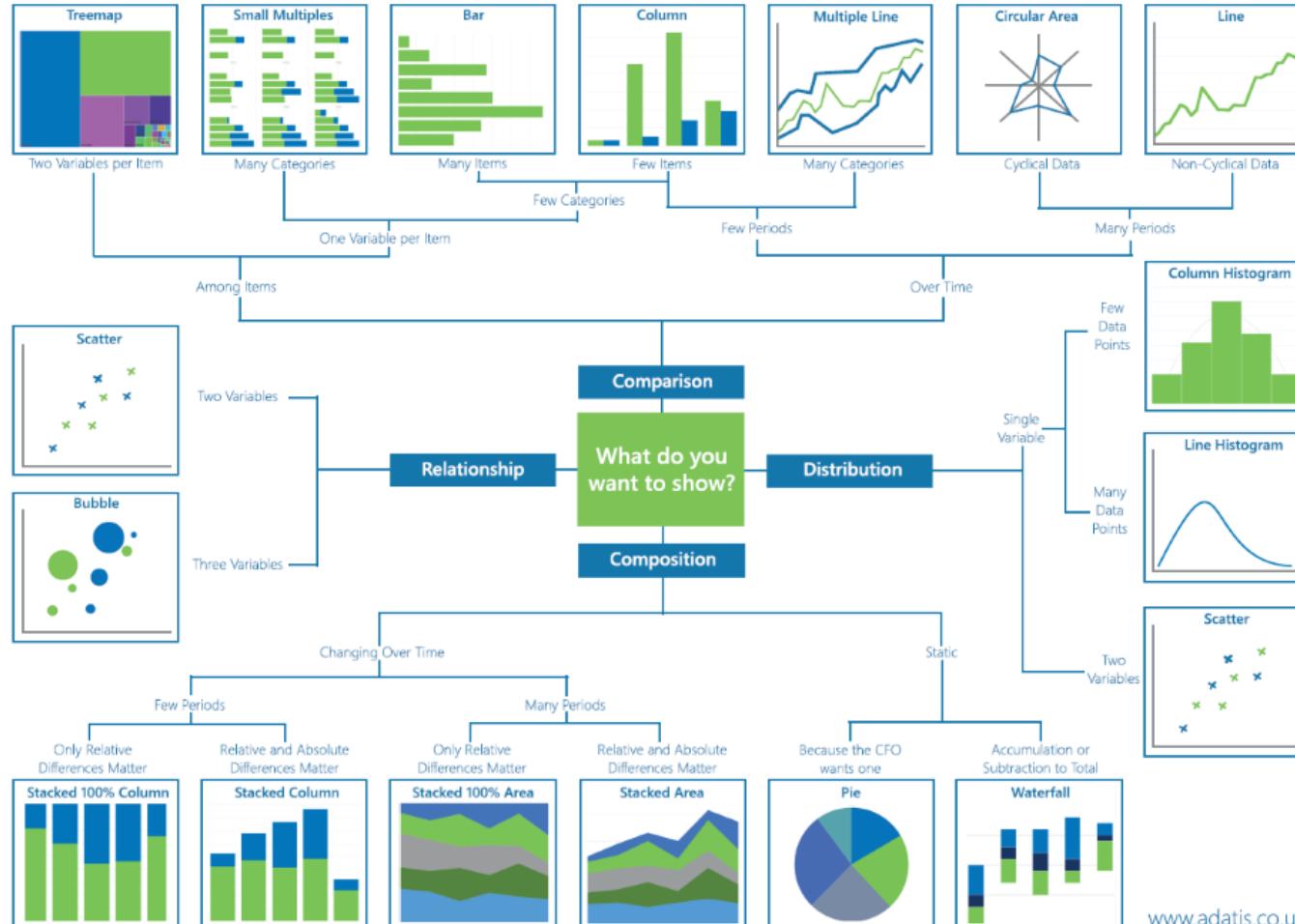


Data Visualization



Chart Types

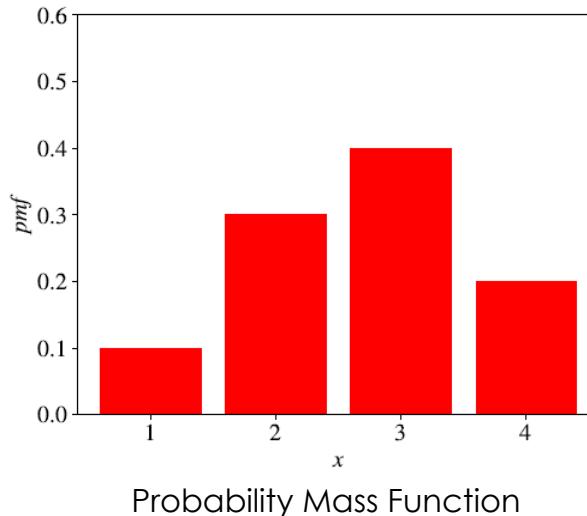
Microsoft Partner
Gold Data Analytics



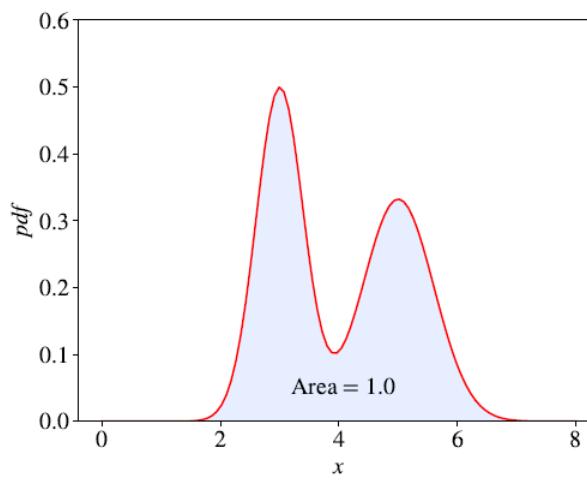
www.adatis.co.uk

Graphical Representation of data!

Visualization: Distribution



Probability Mass Function



Probability Density Function

Visualization: Bars

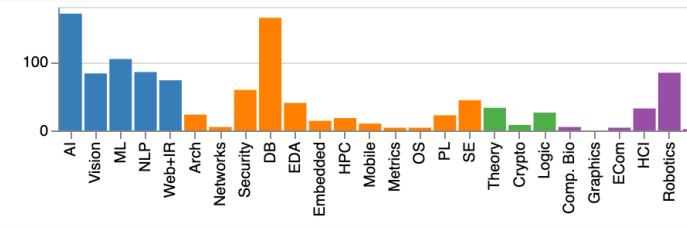
CSRankings: Computer Science Rankings

CSRankings is a metrics-based ranking of top computer science institutions around the world. Click on a triangle (\blacktriangleright) to expand areas or institutions. Click on a name to go to a faculty member's home page. Click on a chart icon (the  after a name or institution) to see the distribution of their publication areas as a bar chart. Click on a Google Scholar icon () to see publications, and click on the DBLP logo () to go to a DBLP entry.

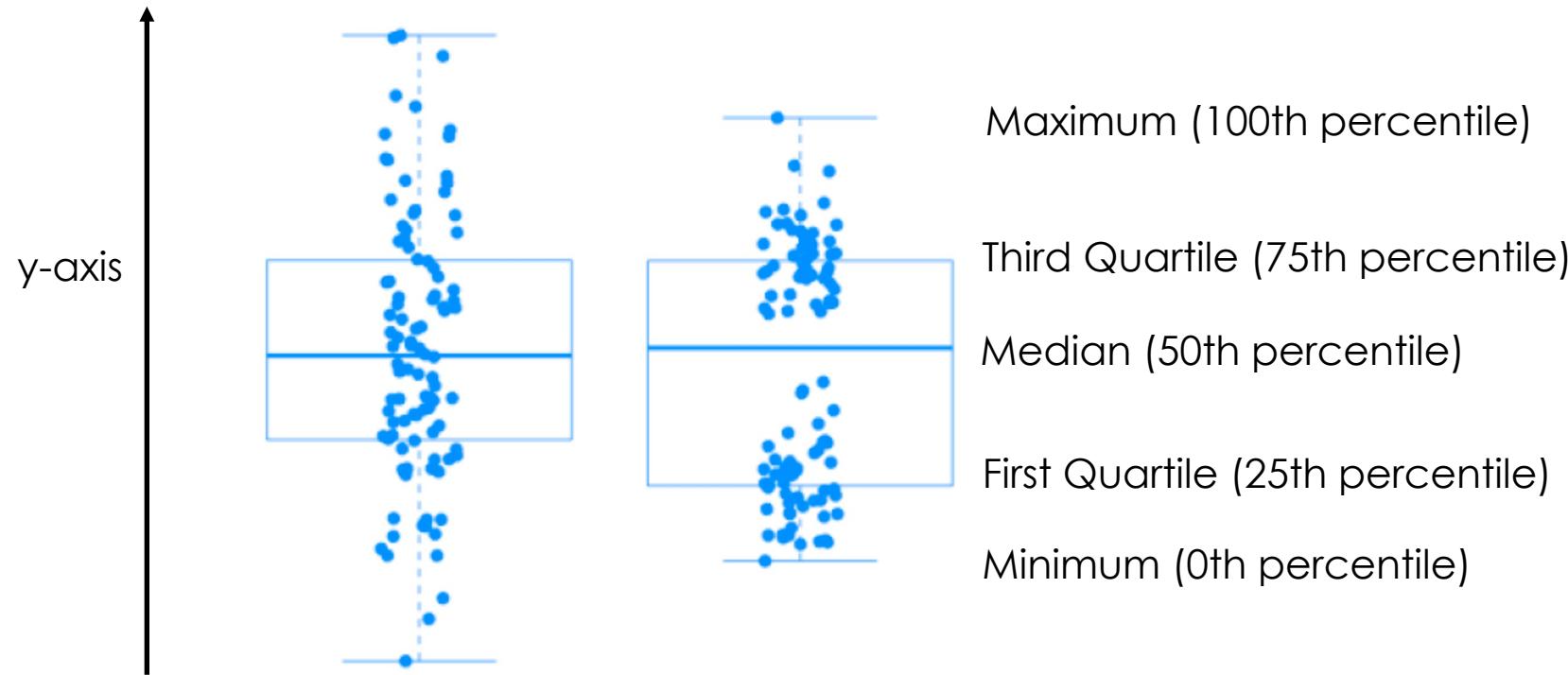
Applying to grad school? Read this first.

Rank institutions in by publications from to

12	 Georgia Institute of Technology  	9.1	94
13	 University of Maryland - College Park  	8.2	83
14	 University of Wisconsin - Madison  	7.6	65
15	 Columbia University  	7.4	55
15	 National University of Singapore  	7.4	66

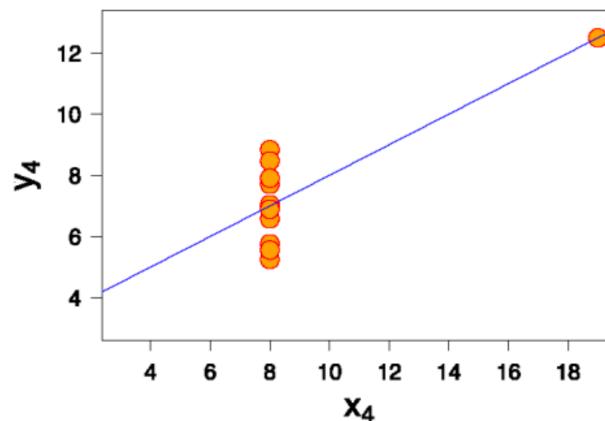
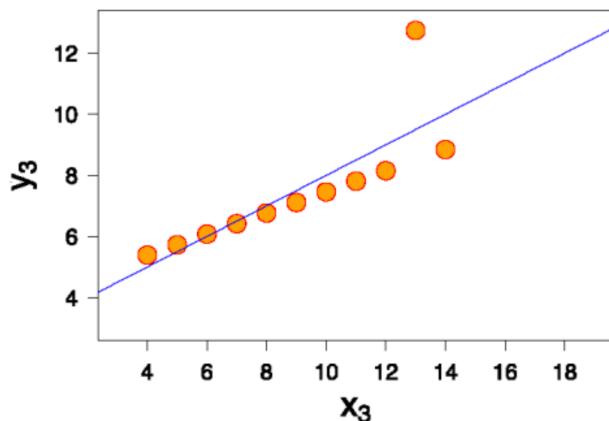
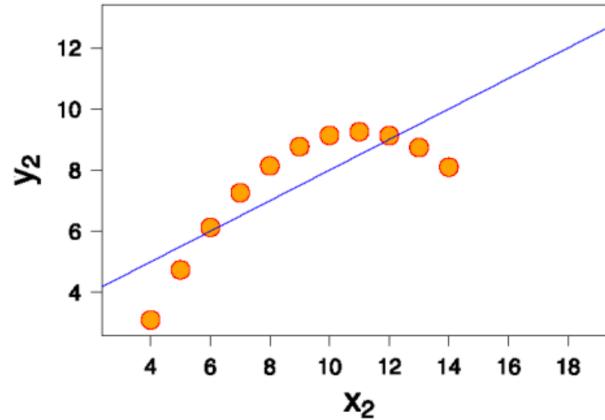
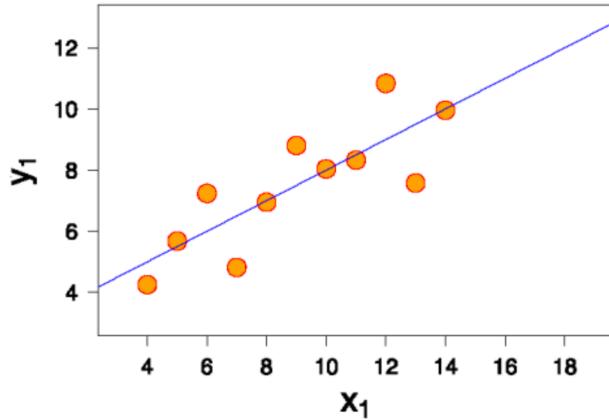


Visualization: Boxplots



- 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 (Q_3) is the middle number between the median and the highest value (i.e., Maximum) of the data set.

Why Visualization is Necessary

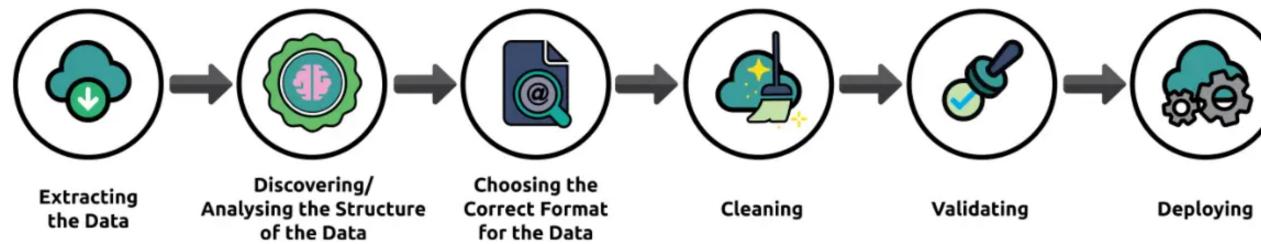


Four datasets with
identical means,
 variances and
 regression lines!

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)

- O
- ▲
- ◆
-
-
-
-
- ◆

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?

