

EE2211 Introduction to Machine Learning

Lecture 1

Introduction Step 1 of
workflow: problem
definition

Wang Xinchao
xinchao@nus.edu.sg

**Office Hour: Tuesday 9:30 – 10:30 AM
(Week 2-4, Week 10-12)**

Course Contents

- Introduction and Preliminaries (Xinchao)
 - Introduction
 - Data Engineering
 - Introduction to Probability and Statistics
- Fundamental Machine Learning Algorithms I (Helen)
 - Systems of linear equations
 - Least squares, Linear regression
 - Ridge regression, Polynomial regression
- Fundamental Machine Learning Algorithms II (Helen)
 - 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

Outline

- What is machine learning?
 - Three Definition(s)
- When do we need machine learning?
 - Sometimes we need, sometimes we don't
- Applications of machine learning
- Types of machine learning
 - Supervised, Unsupervised, Reinforcement Learning
- Walking through a toy example on classification
- Inductive vs. Deductive Reasoning

What is machine learning?

Learning is any process by which a system improves performance from experience.

- Herbert Simon

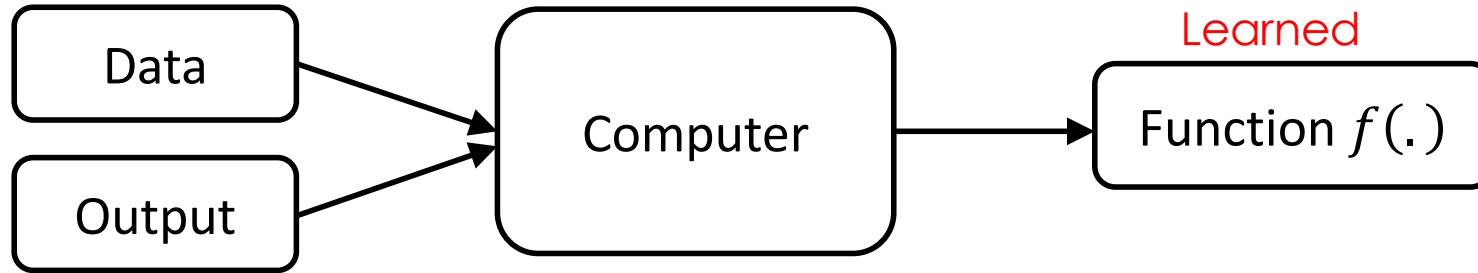
A computer program is said to learn

- from **experience E**
- with respect to some class of **tasks T**
- and **performance measure P**,

if its performance at tasks in T, as measured by P, improves with experience E.

- Tom Mitchell

Machine Learning (Supervised Learning)



Data Output



Cat

:



Dog

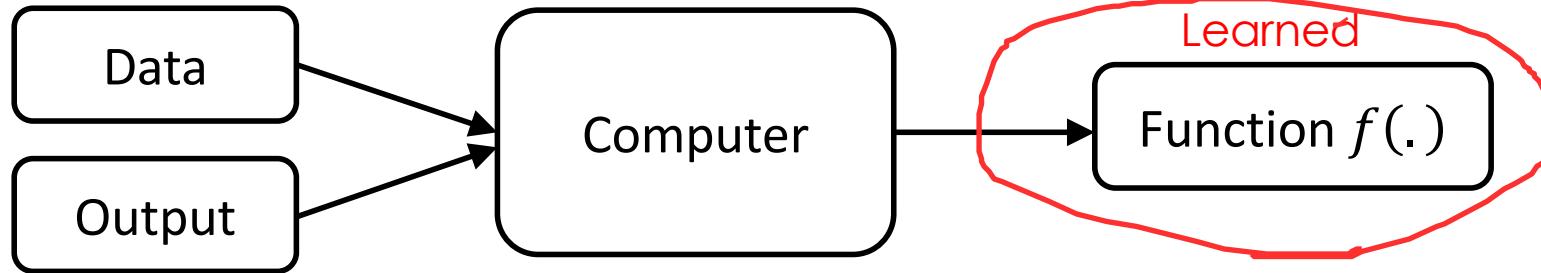
$\longrightarrow f(\cdot)$ such that

$f(\text{cat}) = \text{'cat'}$

$f(\text{dog}) = \text{'dog'}$

train the function to predict given a test data.

Machine Learning (Supervised Learning)



Data Output



Cat

:



Dog

$$\longrightarrow f(\cdot)$$

When applied

$$f(\text{New image}) \rightarrow \text{Cat !}$$

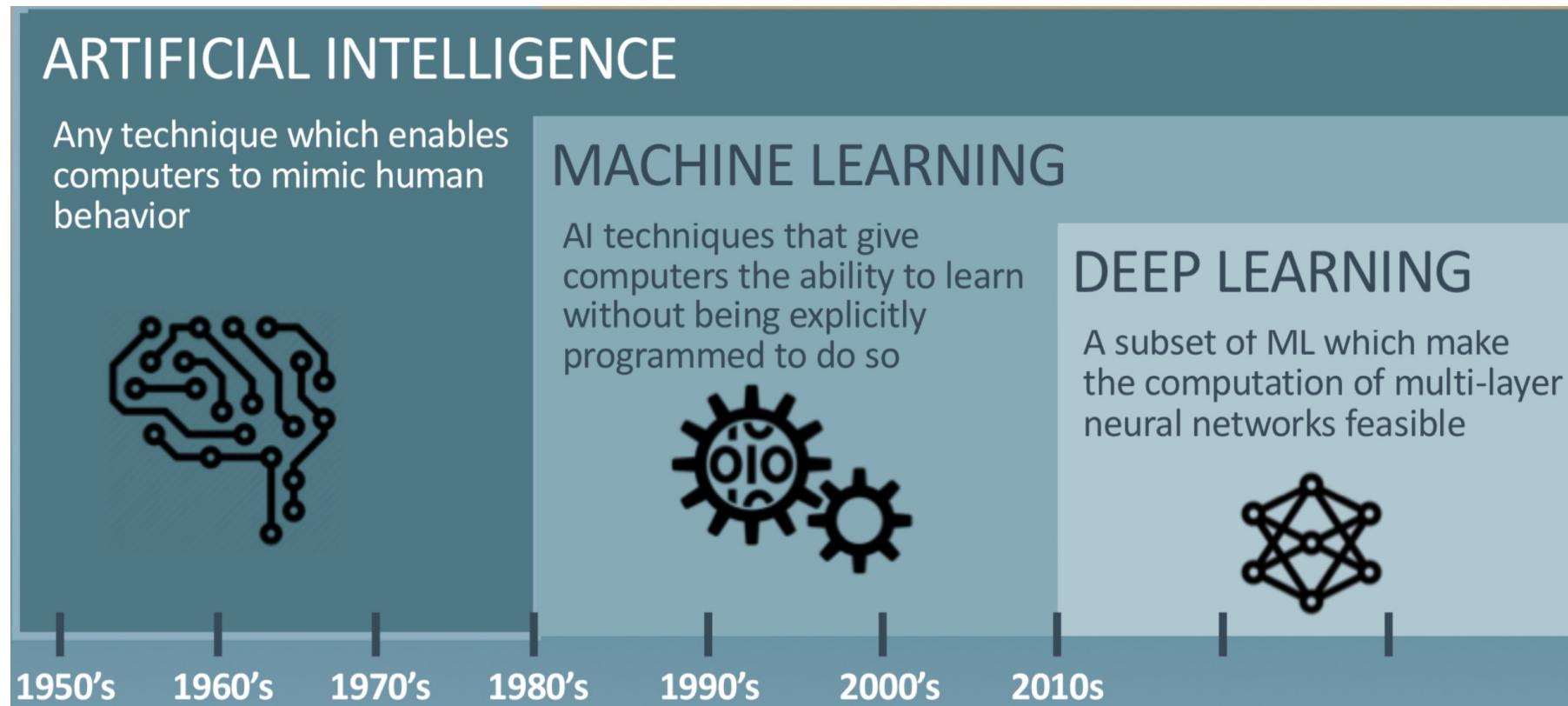


Machine Learning: field of study that gives computers the ability to learn without being explicitly programmed

- Arthur Samuel

AI, Machine Learning, and Deep Learning

Venn
Diagram



Example of AI but not ML: Deductive Reasoning

NUS is in Singapore, Singapore is in Asia \rightarrow NUS is in Asia

When do we need machine learning?

Lack of human expertise
(Navigating on Mars)



Involves huge amount of data
(Genomics)



Learning is not always useful:

No need to “learn” to calculate payroll!

My Salary = Days_of_work * Daily Salary + Bonus

Application of Machine Learning

Task T, Performance P, Experience E

T: Digit Recognition

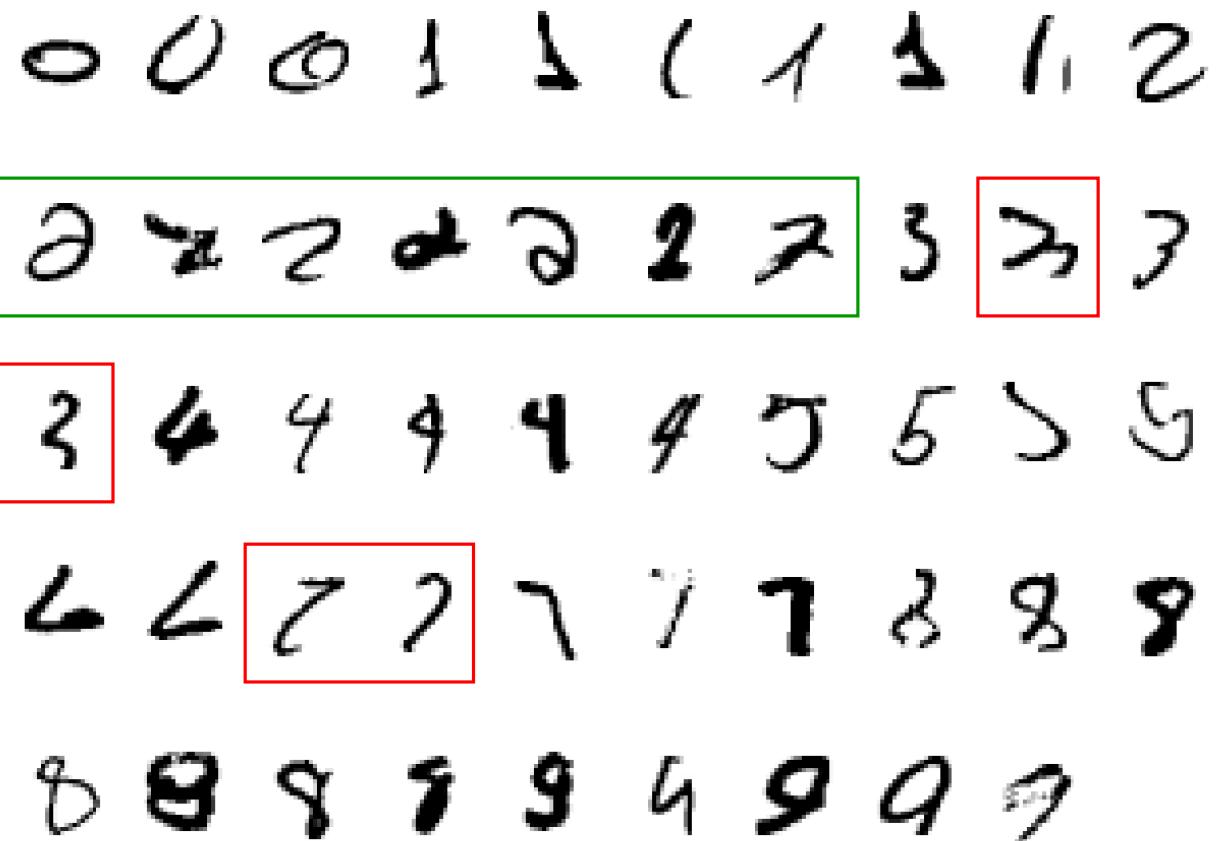
P: Classification Accuracy

E: Labelled Images

4 “four”

3 “three”

Labels -> Supervision!



Application of Machine Learning

Task T, Performance P, Experience E

T: Email Categorization

P: Classification Accuracy

E: Email Data, Some Labelled



Application of Machine Learning

Task T, Performance P, Experience E

T: Playing Go Game

P: Chances of Winning

E: Records of Past Games



Application of Machine Learning

Task T, Performance P, Experience E

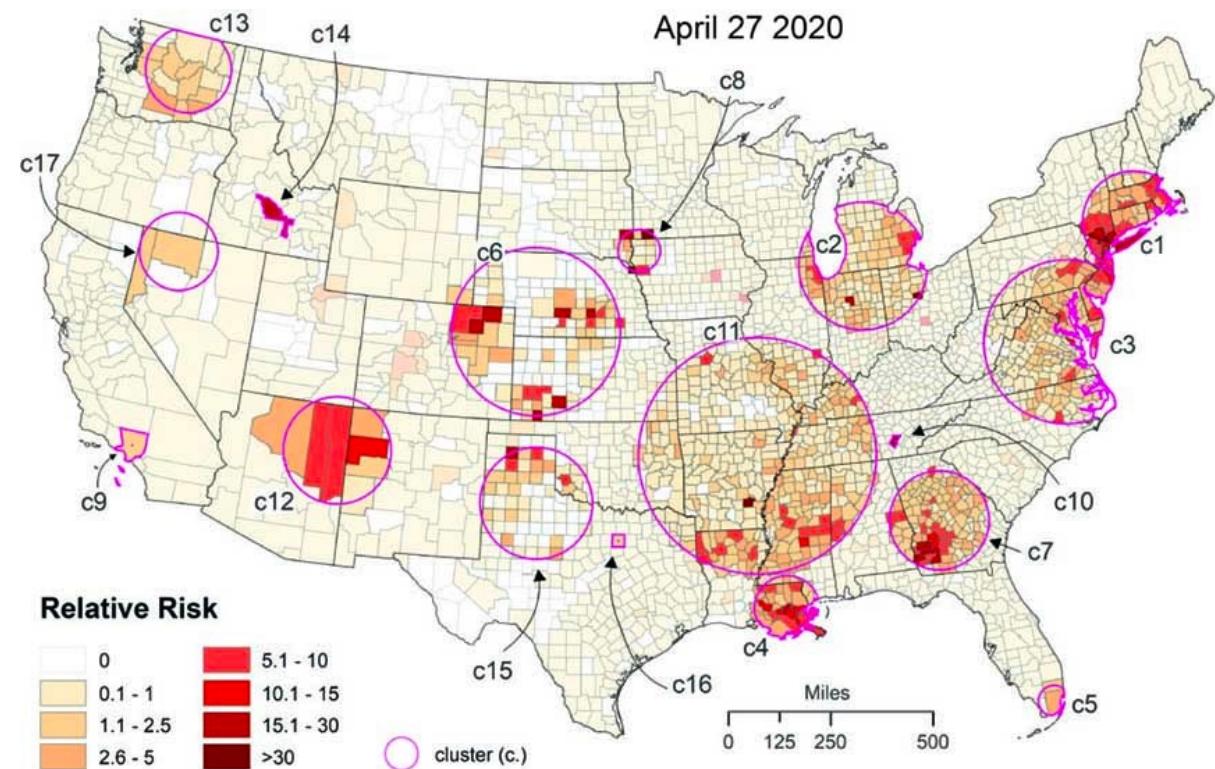
T: Identifying Covid-19 Clusters

P: Small Internal Distances

Larger External Distances

E: Records of Patients

Not supervised learning as the records of patients (samples) is given, not e.g. clusters of another map





Web Search Engine

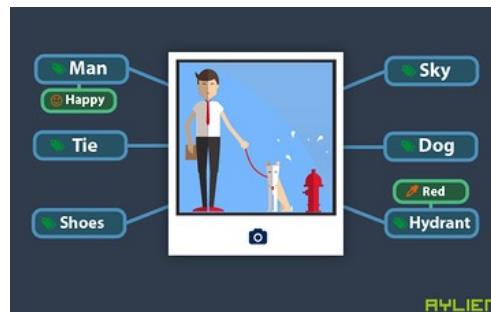
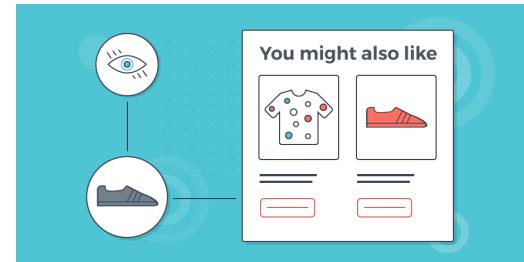


Photo Tagging



Product Recommendation



Virtual Personal Assistant



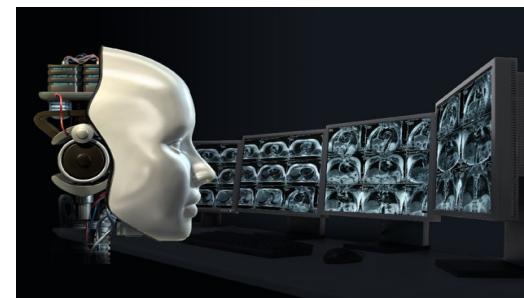
Language Translation



Portfolio Management



Traffic Prediction



Medical Diagnosis



Algorithmic Trading

Types of Machine Learning

Regression and
Classification Tasks

Clustering
Task

Supervised Learning

Input:

- 1) Training Samples,
- 2) Desired Output
(Teacher/Supervision)

Output:

A rule that maps input to output

Unsupervised Learning

Input:

Samples

Output:

Underlying patterns in data
NO DESIRED OUTPUT

Reinforcement Learning

Input:

Sequence of States,
Actions, and
Delayed Rewards

Output:

Action Strategy: a rule
that maps the environment to action

Important: Workflow of ML

Problem Definition -> Data Preparation -> Modelling -> Deployment.

Problem Definition:

1. Clarify the objective: What problem are you trying to solve?
2. Determine if ML is appropriate: Can this problem be learned from data?
3. Specify inputs and outputs
 - What input features will the model use?
 - What is the model trying to predict?

- * 4. Choose the task type
Classification? Regression? Clustering?...
- 5. Decide on performance measure
Accuracy? Mean Squared Error?...

Types of Machine Learning

Supervised Learning

Input:

- 1) Training Samples,
- 2) Desired Output
(Teacher/Supervision)

Output:

A rule that maps input to output

Data Output



Cat

:



Dog

→ $f(\cdot)$ such that

Unsupervised Learning

Input:

Samples

Output:

Underlying patterns in data

Reinforcement Learning

Input:

Sequence of States,
Actions, and
Delayed Rewards

Output:

Action Strategy: a rule
that maps the
environment to action

All these statements about classification and regression are false:

- Classification works with labels that are only continuous
- Classification works with labels that only belong to ordinal data
- Classification works with features that are only discrete
- Regression works with features that are only discrete

Supervised Learning

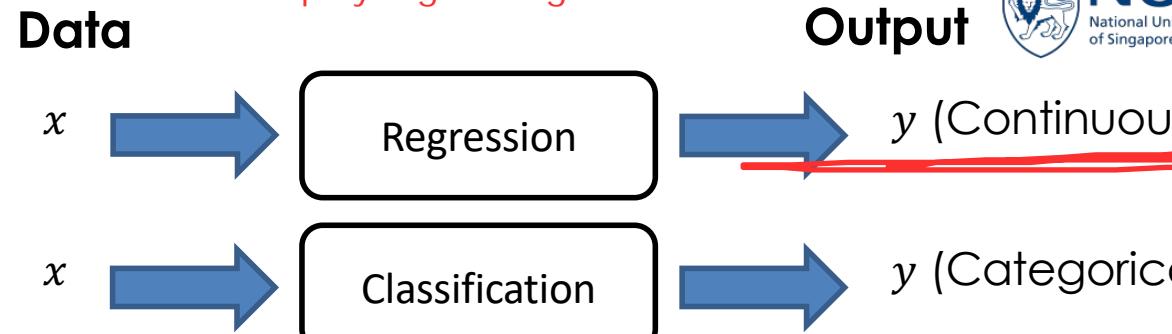
Techniques used includes linear regression, polynomial regression and ridge regression.
(k-NN also works as it is a supervised learning technique)

evaluation metrics for regression task: MSE, mean absolute error

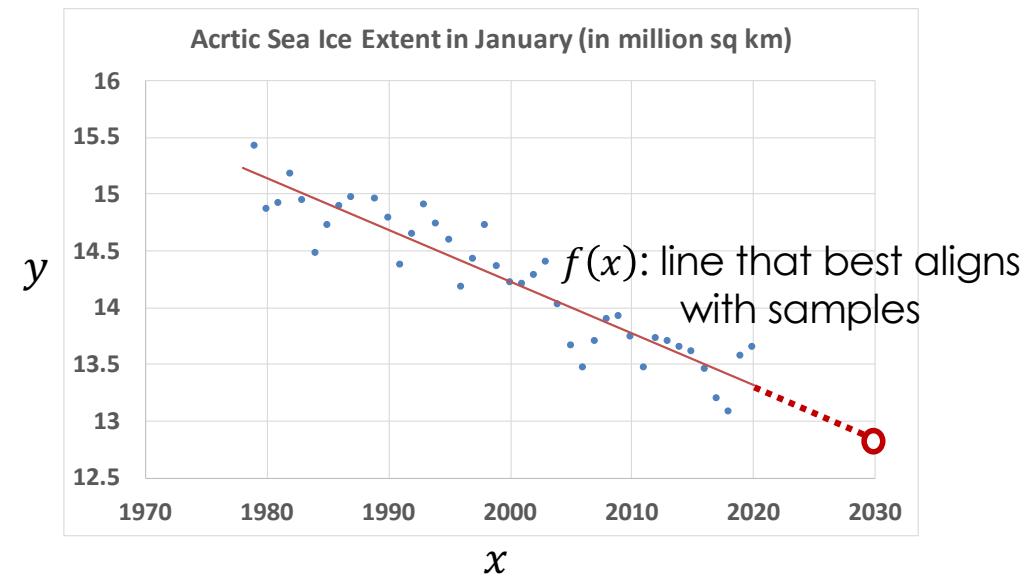
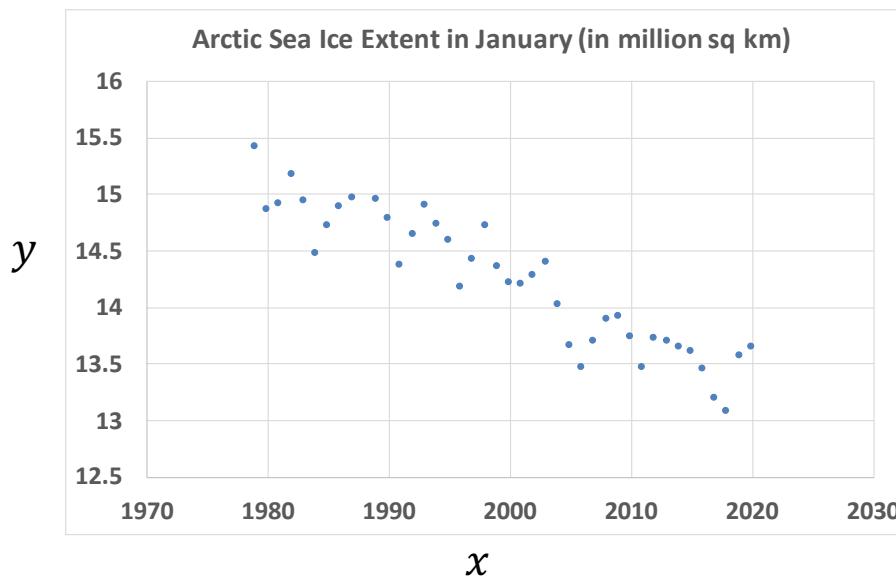
Regression

- Given $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_N, y_N)$
- Learn a function $f(\mathbf{x})$ to predict real-valued y given \mathbf{x}

Note that the "Regression" here refers to task, not techniques like linear/poly/logistic regression.



if y is continuous



Regression task is similar but not classification in the sense that y is no longer restricted to a finite set (classes), it now takes on uncountably many values. y here is not called label anymore, but the actual output value from trained function given a test data (x)

The technique of regression can be and are used for classification task:

- linear/poly: binary (sign function) and multi-category (one-hot encoding) classification
- logistic regression (which in fact is used for classification task and not regression task) - using soft threshold function

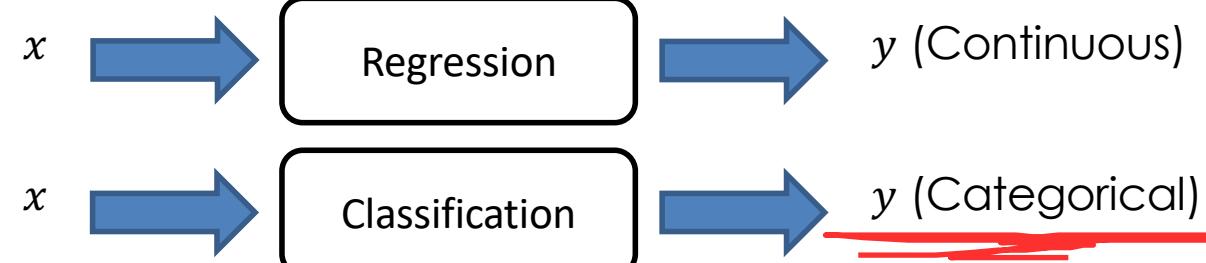


Supervised Learning

another technique that will be shown later is nearest neighbour classifier

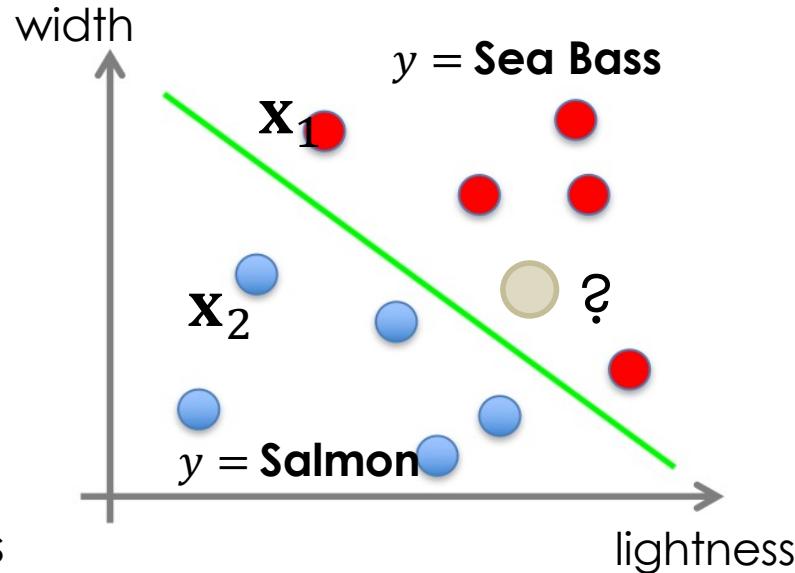
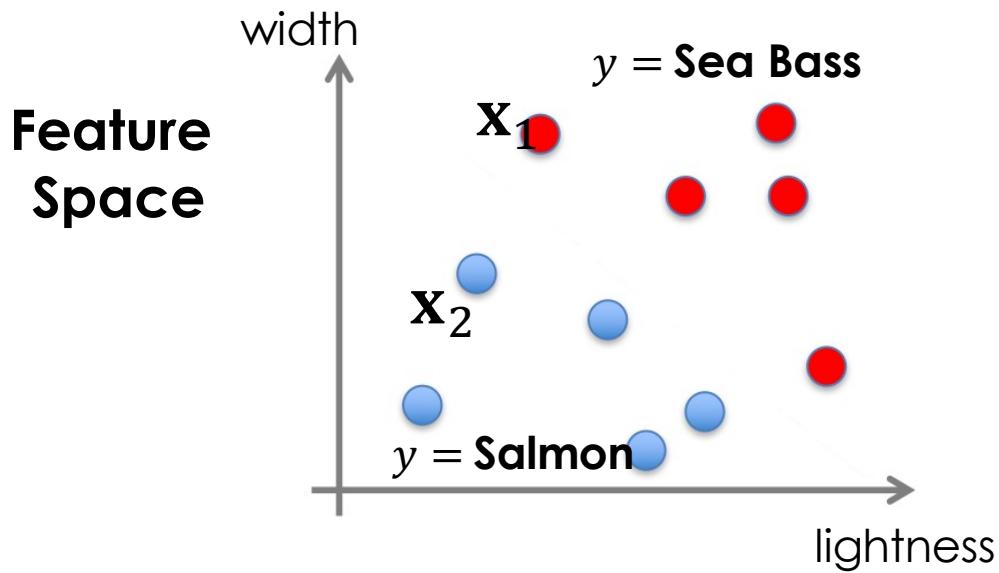
Data

Output



Classification

- Given $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_N, y_N)$
- Learn a function $f(\mathbf{x})$ to predict categorical y given \mathbf{x}



$f(x)$: line that separates two classes

Types of Machine Learning

Supervised Learning

Input:

- 1) Training Samples,
- 2) Desired Output
(Teacher/Supervision)

Output:

A rule that maps input to output

Unsupervised Learning

Input:

Samples

Output:

Underlying patterns in data

Reinforcement Learning

Input:

Sequence of States,
Actions, and
Delayed Rewards

Output:

Action Strategy: a rule
that maps the
environment to action

**Key different w.r.t. supervised learning:
No Label/Supervision is given!**

Techniques used for clustering
task includes:

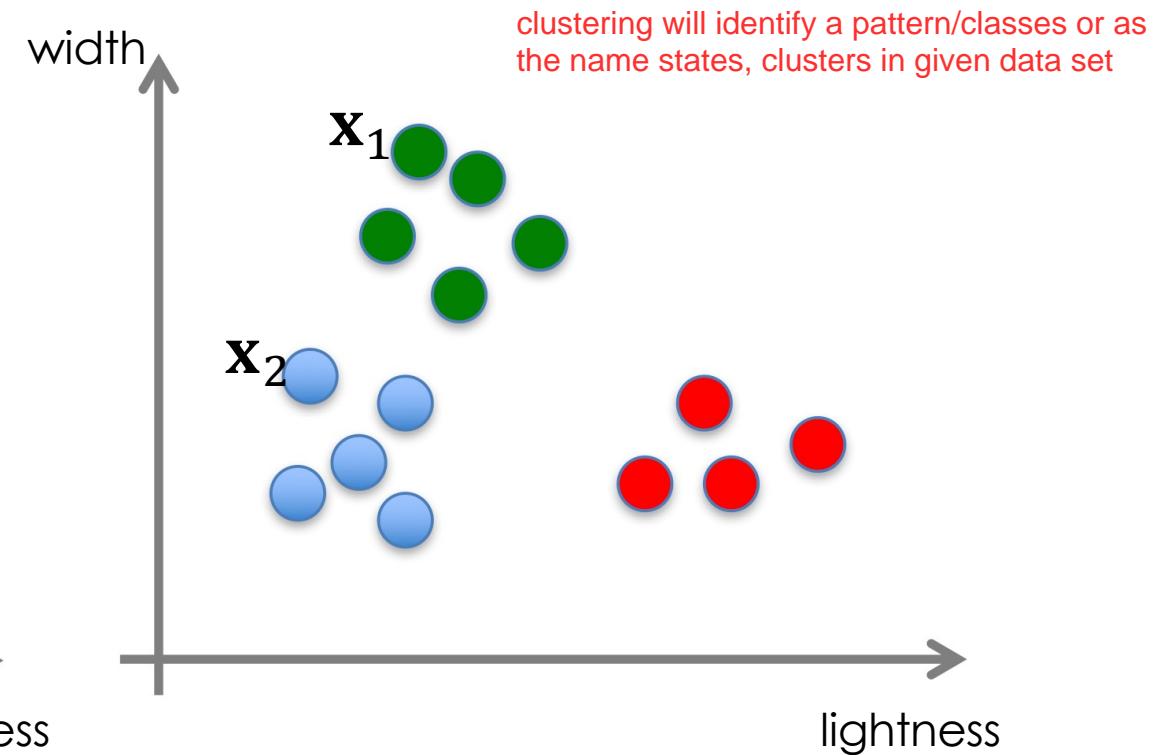
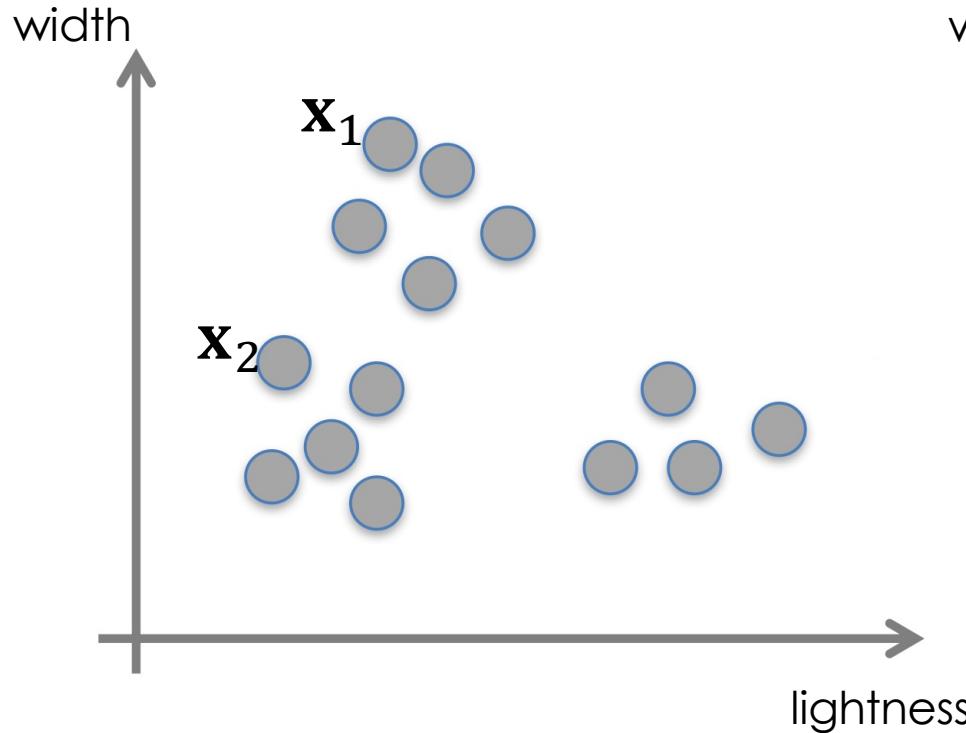
K-means
Fuzzy C-means

Unsupervised Learning

labels refer to DESIRED OUTCOMES: what we want the model to label the test data after prediction.
Here it is not specified. Only to cluster the data and output pattern/class

Clustering

- Given $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N$, without labels
- Output Hidden Structure Behind



No Label/Supervision is given!

Types of Machine Learning

Supervised Learning

Input:

- 1) Training Samples,
- 2) Desired Output
(Teacher/Supervision)

Output:

A rule that maps input to output

Unsupervised Learning

Input:

Samples

Output:

Underlying patterns in data

Reinforcement Learning

Input:

Sequence of States,
Actions, and
Delayed Rewards

Output:

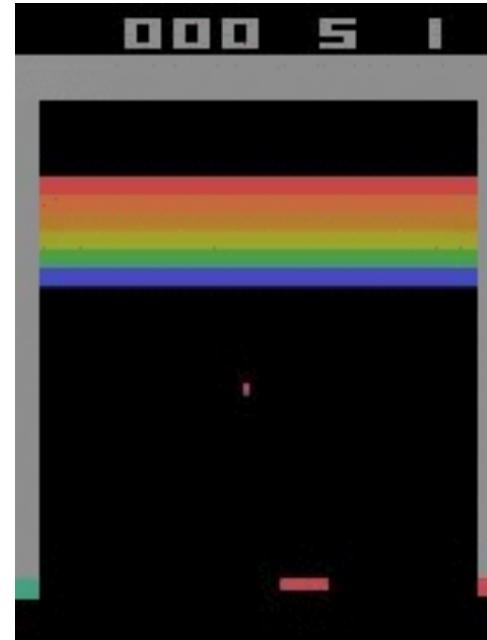
Action Strategy: a rule
that maps the environment to action

Reinforcement Learning

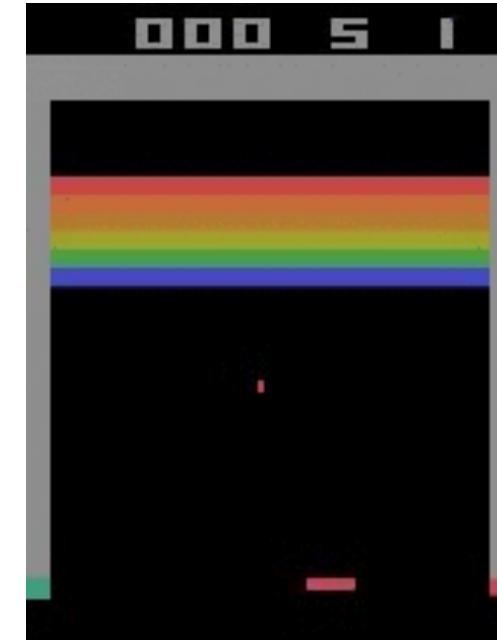
Breakout Game



Initial Performance



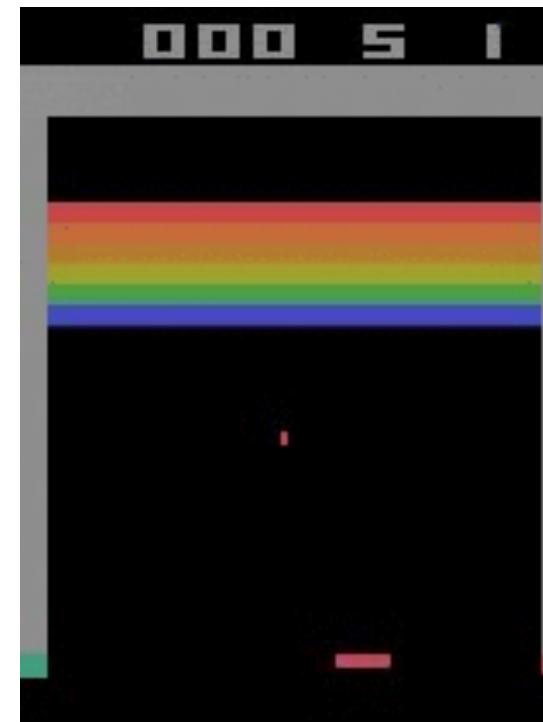
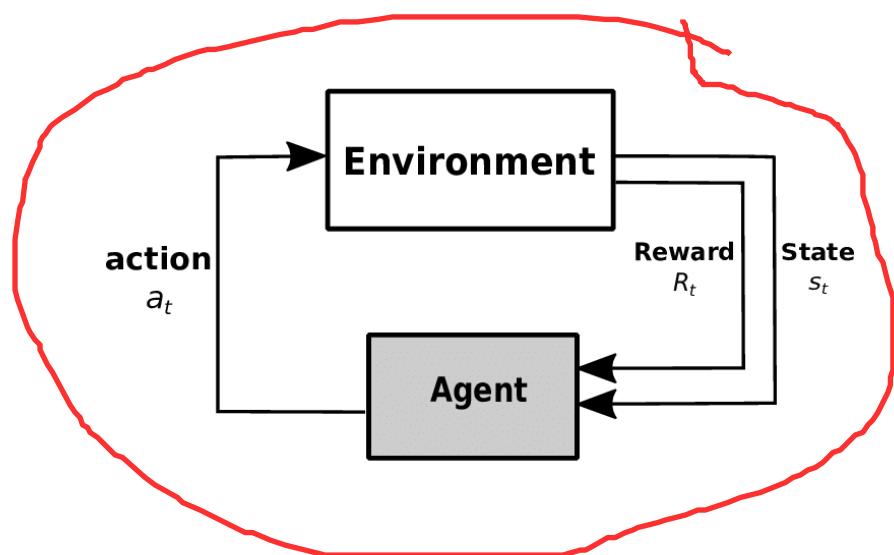
Training 15 minutes



Training 30 minutes

Reinforcement Learning

- Given sequence of states \underline{S} and actions \underline{A} with (delayed) rewards \underline{R}
- Output a policy $\pi(a, s)$, to guide us what action a to take in state s



\underline{S} : Ball Location,
Paddle Location, Bricks
 \underline{A} : left, right
 \underline{R} :
positive reward
Knocking a brick,
clearing all bricks
negative reward
Missing the ball
zero reward
Cases in between

Supervised
Unsupervised
Reinforcement

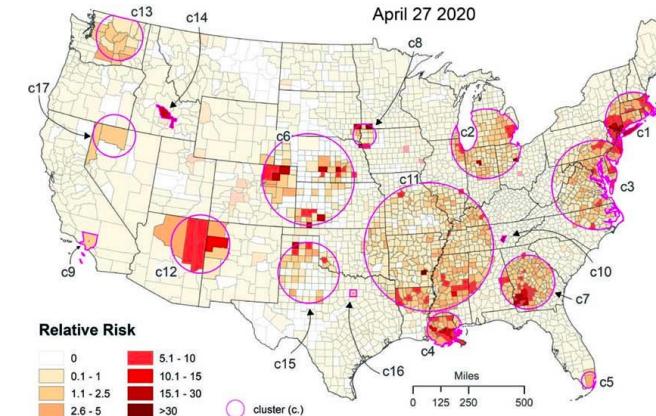
Quiz Time!

0 0 0 1 1 1 1 1 1 2
2 2 2 2 2 2 2 3 3 3
3 4 4 4 4 4 5 5 5 5
6 6 7 7 7 7 7 8 8 8
8 8 8 8 9 9 9 9 9 9

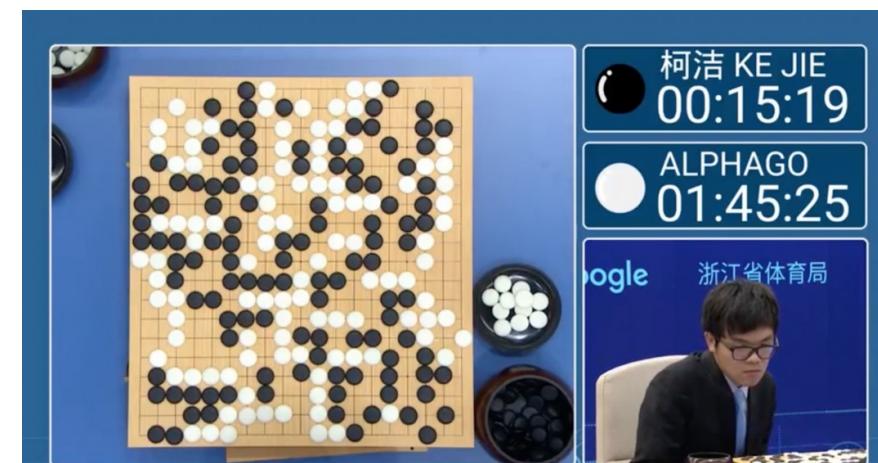
Supervised



Supervised



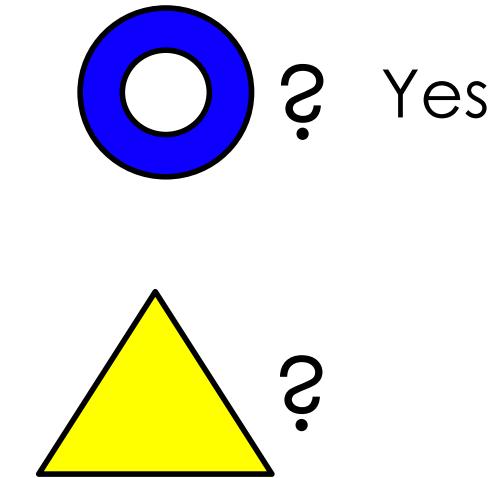
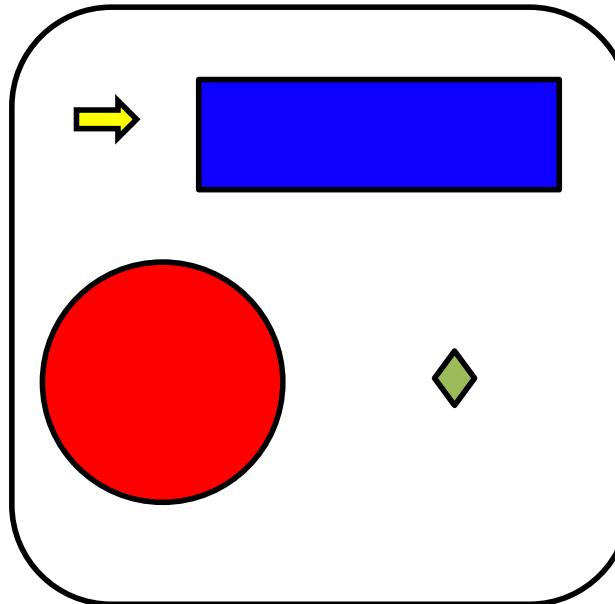
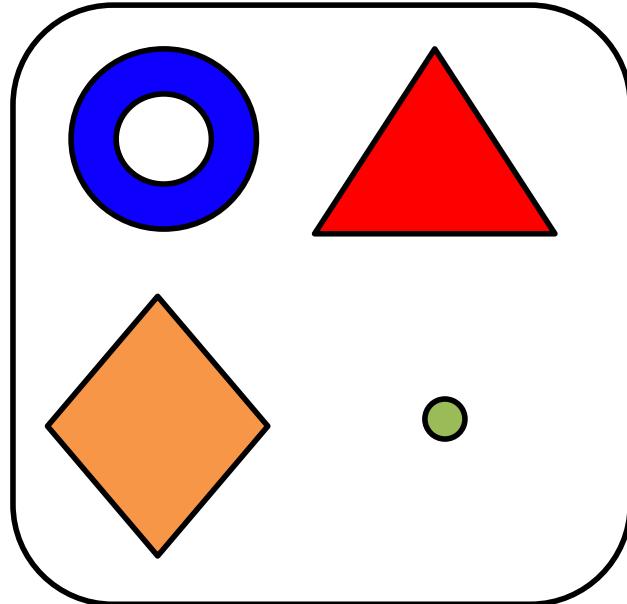
Unsupervised



Reinforcement

This is a classification task, hence need to use supervised learning techniques. The example uses k-nearest neighbours

Walking Through A Toy Example: Token Classification



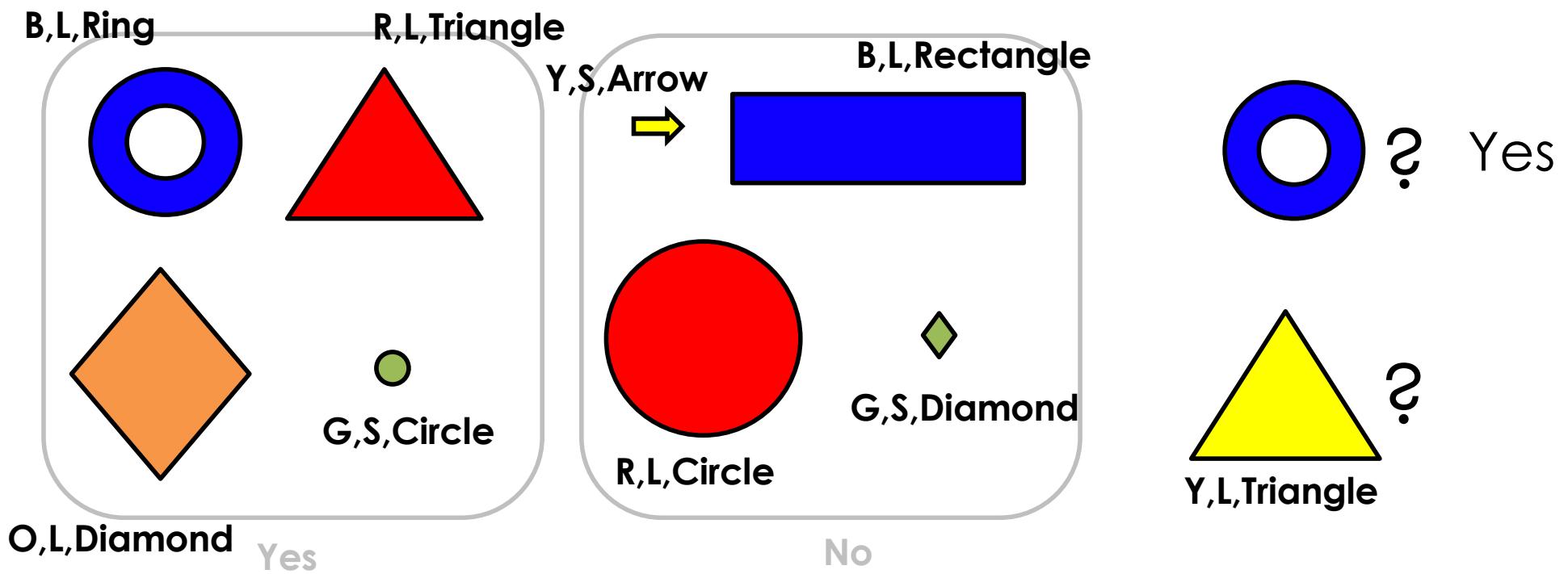
feature (attributes of sample) extraction - for the model to look out for (colour, shape)
decide labels to classify the data into - for the model to label each data (correct shape, wrong shape)

Step 1: Feature Extraction
Extract Attributes of Samples



Step 2: Sample Classification
Decide Label for a Sample

Walking Through A Toy Example: Token Classification



Step 1: Feature Extraction
Color, Size, Shape

Walking Through A Toy Example: Token Classification

Feature Extraction

then based on features,
decide a label

	<u>Color</u>	<u>Size</u>	<u>Shape</u>	Label
O	Blue	Large	Ring	Yes
▲	Red	Large	Triangle	Yes
◆	Orange	Large	Diamond	Yes
●	Green	Small	Circle	Yes
→	Yellow	Small	Arrow	No
■	Blue	Large	Rectangle	No
●	Red	Large	Circle	No
◆	Green	Small	Diamond	No
▲	Yellow	Large	Triangle	?

this forms the feature space of the
training data set

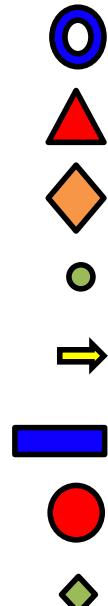
Walking Through A Toy Example: Token Classification

Feature Extraction

	Color	Size	Shape	Label
O	Blue	Large	Ring	Yes
▲	Red	Large	Triangle	Yes
◆	Orange	Large	Diamond	Yes
●	Green	Small	Circle	Yes
→	Yellow	Small	Arrow	No
■	Blue	Large	Rectangle	No
○	Red	Large	Circle	No
◆	Green	Small	Diamond	No

Walking Through A Toy Example: Token Classification

Feature Extraction



Color	Size	Shape	Label
Blue	Large	Ring	Yes
Red	Large	Triangle	Yes
Orange	Large	Diamond	Yes
Green	Small	Circle	Yes
Yellow	Small	Arrow	No
Blue	Large	Rectangle	No
Red	Large	Circle	No
Green	Small	Diamond	No

the metric used here is $\max(\text{number of similar features})$
or it will be introduced as nearest neighbour later
0 means different
1 means same

Similarity

	Color	Size	Shape	Total
	0	1	0	1
	0	1	1	2
	0	1	0	1
	0	0	0	0
	1	0	0	1
	0	1	0	1
	0	1	0	1
	0	0	0	0

Walking Through A Toy Example: Token Classification

k-NN classifier:

For classification: The algorithm finds the k closest data points to the new point and assigns the class that is most common among those k neighbours.

For regression: It averages (or otherwise aggregates) the values of the k nearest points to make a prediction.

Similarity

	Color	Size	Shape	Total
▲ (Yellow)	0	1	0	1
▲ (Yellow)	0	1	1	2
▲ (Yellow)	0	1	0	1
● (Green)	0	0	0	0
→ (Yellow)	1	0	0	1
■ (Blue)	0	1	0	1
○ (Red)	0	1	0	1
◆ (Green)	0	0	0	0

Step 2: Sample Classification

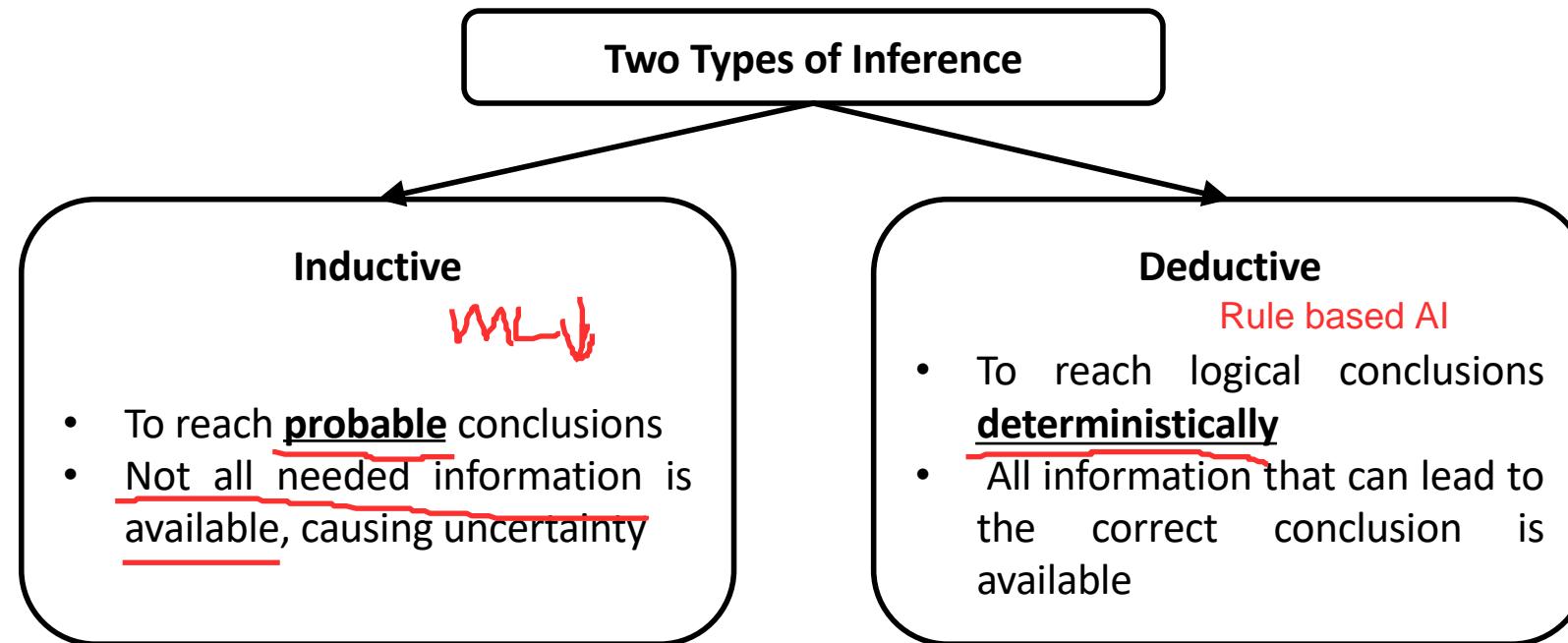
Nearest Neighbor Classifier:

- 1) Find the “nearest neighbor” of a sample in the feature space
Here based on number of similar features
- 2) Assign the label of the nearest neighbor to the sample

Nearest Neighbor Classifier is just one possible classifier. We may use other types of classifiers, given the extracted features.

Inductive vs. Deductive Reasoning

- Main Task of Machine Learning: to make inference



Probability and Statistics



Rule-based reasoning

NUS is in Singapore, Singapore is in Asia ->
NUS is in Asia

Inductive Reasoning

Note: humans use inductive reasoning all the time and not in a formal way like using probability/statistics.



Ref: Gardner, Martin (March 1979). "MATHEMATICAL GAMES: On the fabric of inductive logic, and some probability paradoxes" (PDF). *Scientific American*. 234

Summary by Quick Quiz

Three Components in ML Definition

Task T, Performance P, Experience E

Three Types of in ML

Supervised Learning
Unsupervised Learning
Reinforcement Learning

Inductive and Deductive

Inductive: Probable
Deductive: Rule-based

Two Types of Supervised Learning

Classification, Regression

One Type of Unsupervised Learning

Clustering

Example of a Classifier Model

Nearest Neighbor Classifier

Practice Question

(Type of Question to Expect in Exams)

Which of the following statement is true?

- A. Nearest Neighbor Classifier is an example of unsupervised learning
- B. Nearest Neighbor Classifier is an example of deductive learning
- C. Nearest Neighbor Classifier is an example of feature selection
- D. None of the above is correct.

EE2211 Introduction to Machine Learning

Lecture 2

Data Engineering
Step 2: Data Preparation

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 (Helen)
 - Systems of linear equations
 - Least squares, Linear regression
 - Ridge regression, Polynomial regression
- Fundamental Machine Learning Algorithms II (Helen)
 - 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

Outline

Types of data

Data
wrangling and
cleaning

Data integrity
and
visualization

Types of Data

Negative examples include opinions, speculations, guesses, ideas/theories.

Data must be a representation of information that can be stored, processed or analysed.

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

NoIR

- ✓ Based on Numerical/Categorical

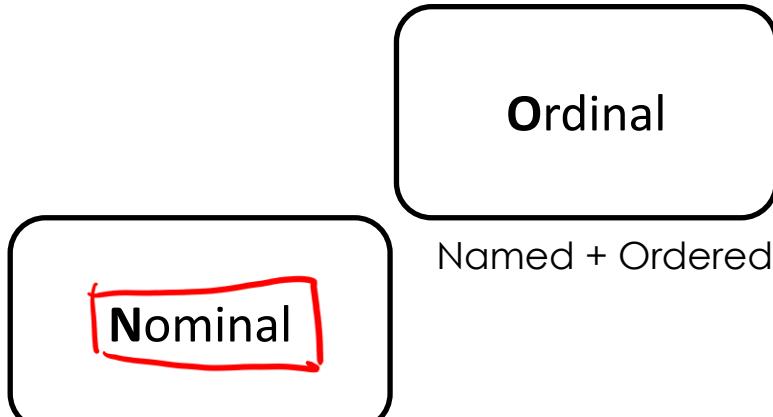
- Numerical, also known as Quantitative
- Categorical, also known as Qualitative

- ✓ Other aspects

- Available or Missing Data

Levels/Scales of Measurement

NOIR



Named + Ordered
+ Equal Interval

Named + Ordered
+ Equal Interval
+ Has “True” Zero
absolute

Ratio

Highest

Lowest

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

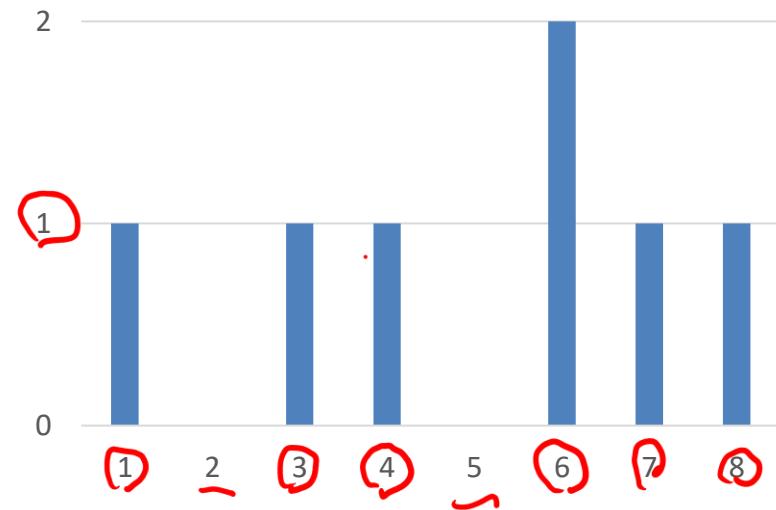
✓ 1, 3, 4, 6, 7, 8
 $(4+6)/2=5$

Mode: number with highest frequency

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

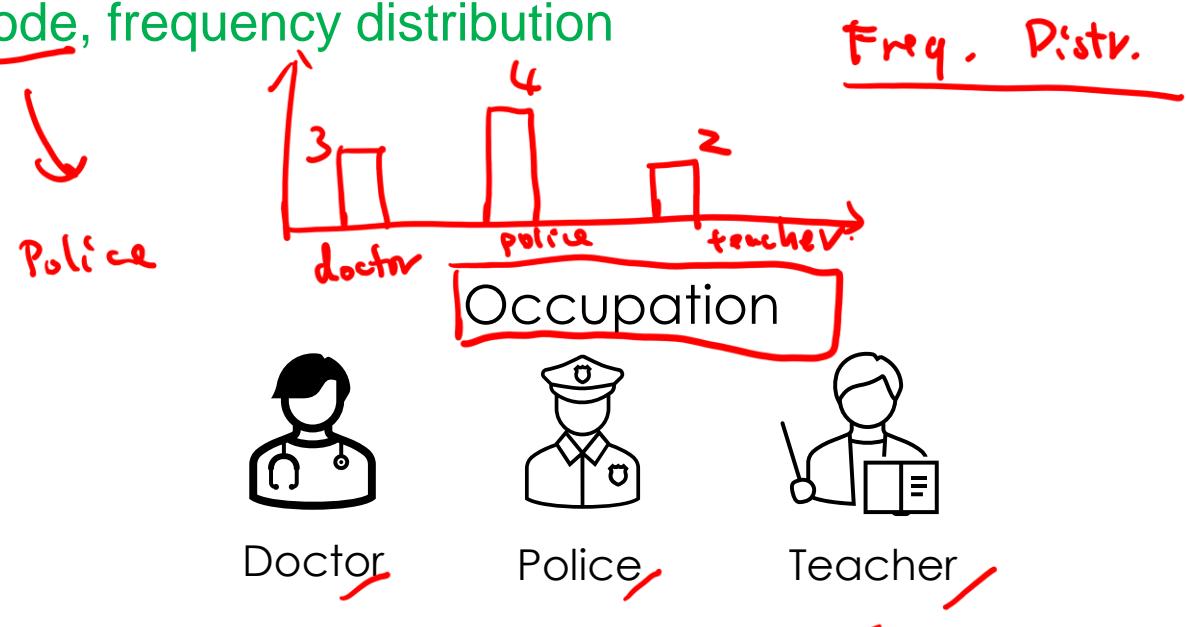
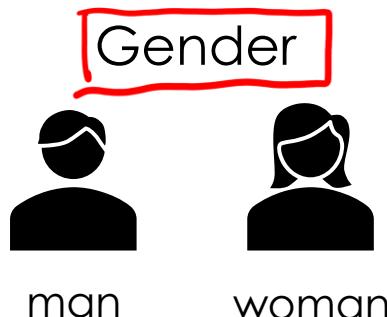
histogram

Frequency Distribution



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

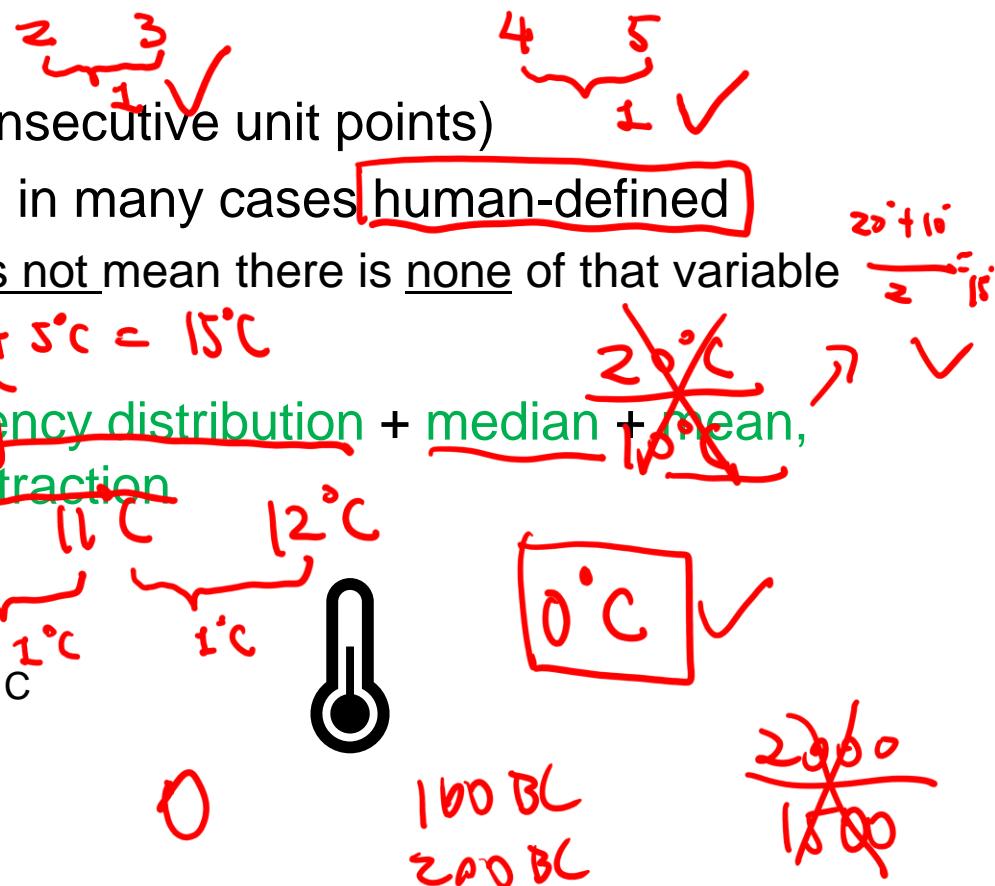
NoIR



- 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

N O I R

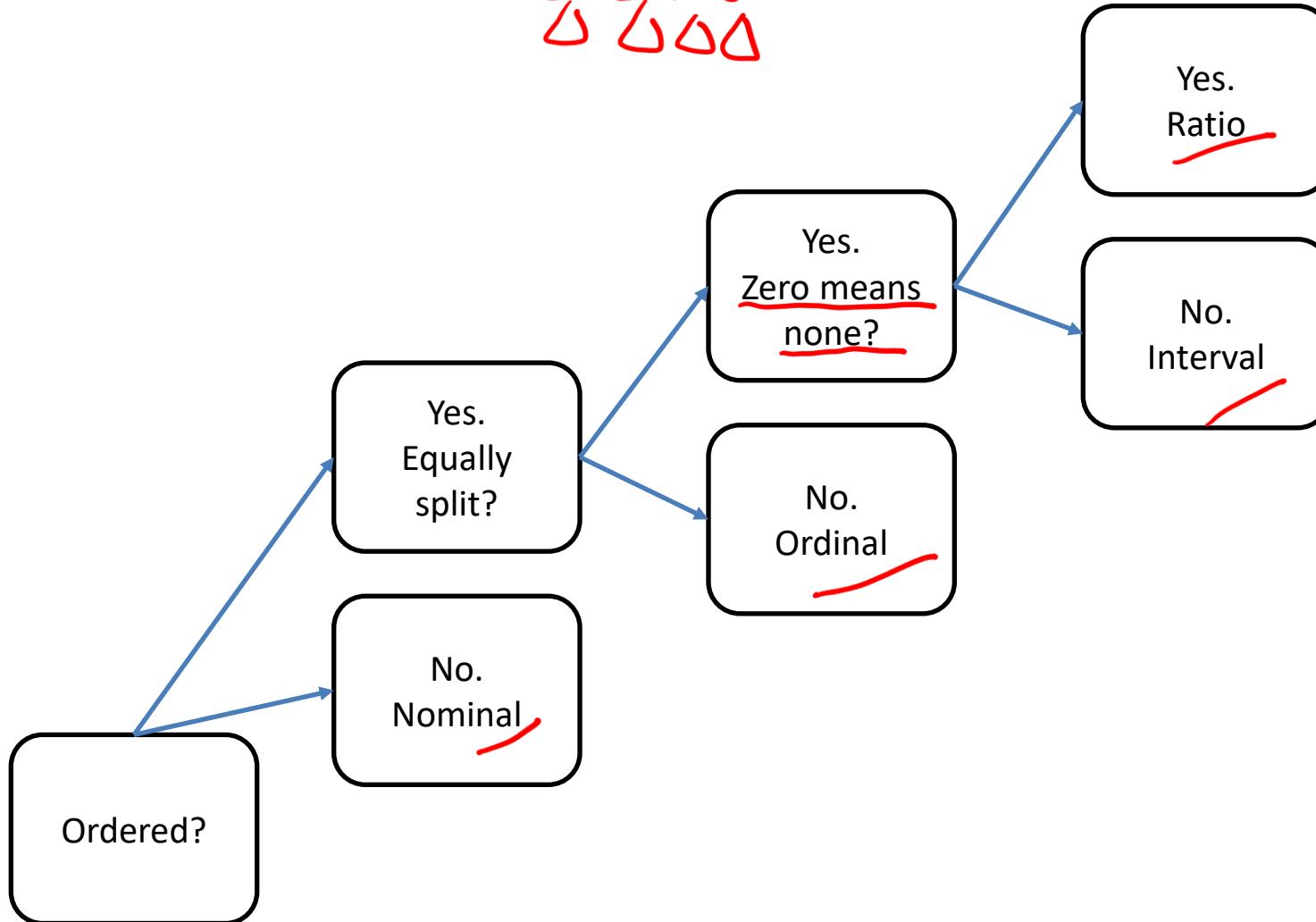
- Most precise and highest level of measurement
 - Ordered
 - Equal Intervals
 - Natural Zeros:
 - If the variable equals zero, it means there is none of that variable
 - Not arbitrary
 - Possible Measure: mode, frequency distribution + median + mean, standard deviation, addition/subtraction + multiplication and division (ratio)
 - Example:
 - Weights
 - 10 KG, 20 KG, 30 KG
 - Time
 - 10 Seconds, 1 Hour, 1 Day
- 
- 0 KG
- $$\frac{20 \text{ KG}}{10 \text{ KG}} = 2.$$
- 
- 0 sec
- $$\frac{60 \text{ sec}}{20 \text{ sec}} = 3.$$

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

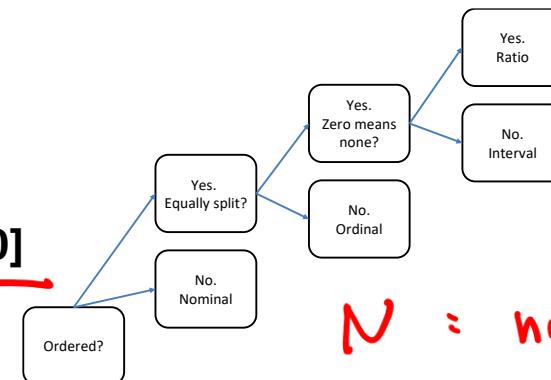
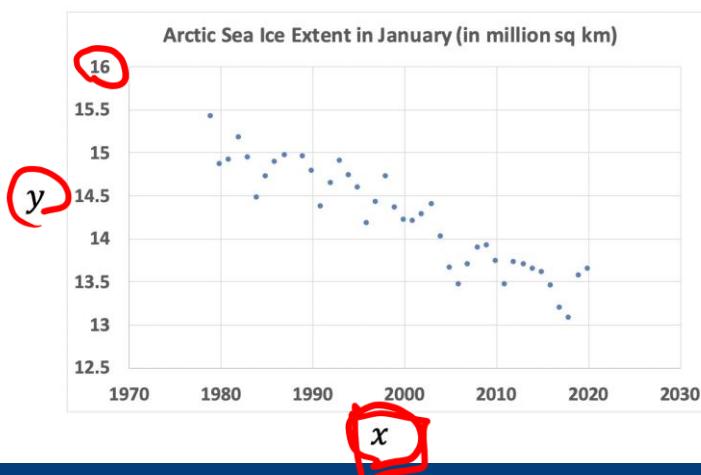
ဗုဒ္ဓဘာ



- 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]
- 8.



N : named
 O : ordered
 I : equally split
 R : ratio

2000
14.5

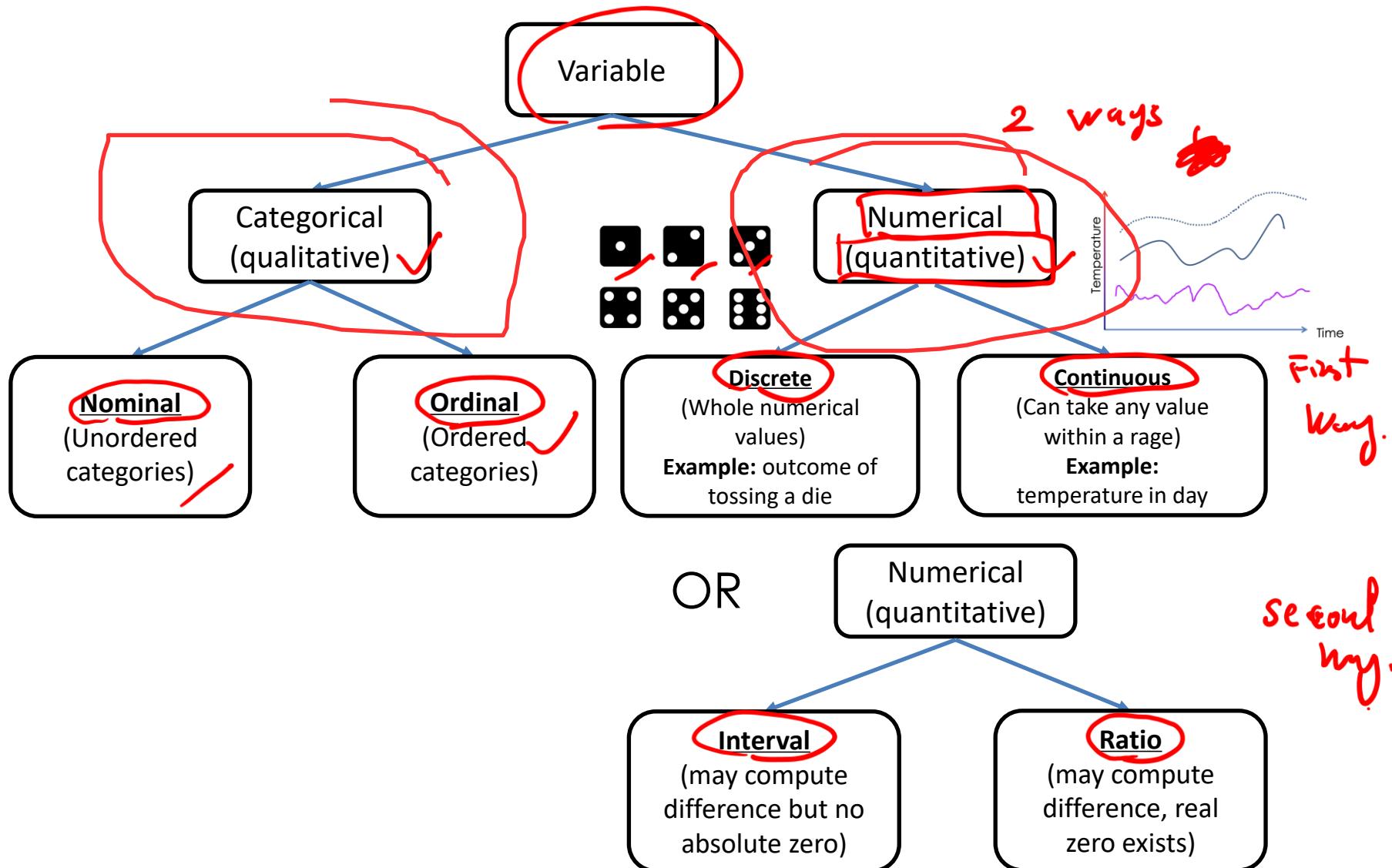
Ways of Viewing Data

- 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

unavailable.

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

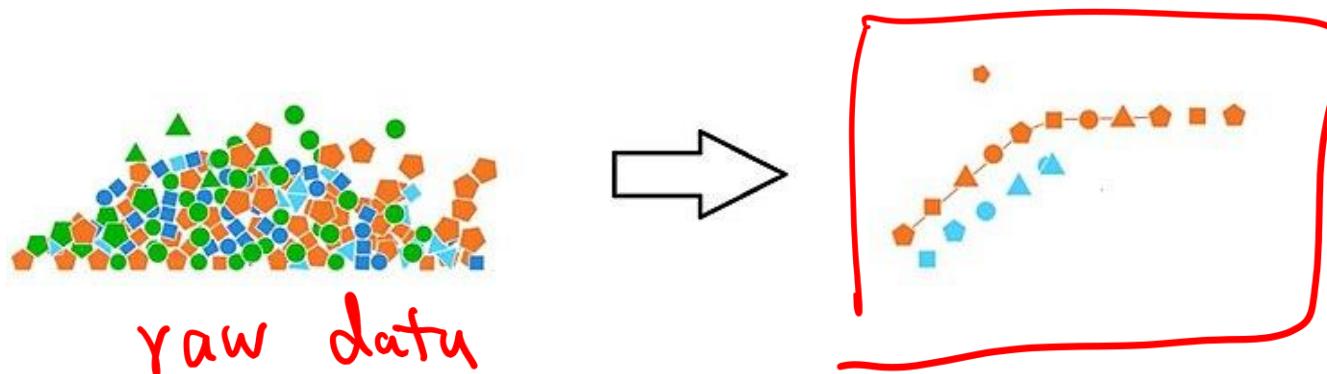
Data integrity
and
visualization

Data Wrangling

raw to a more useful format
before ML

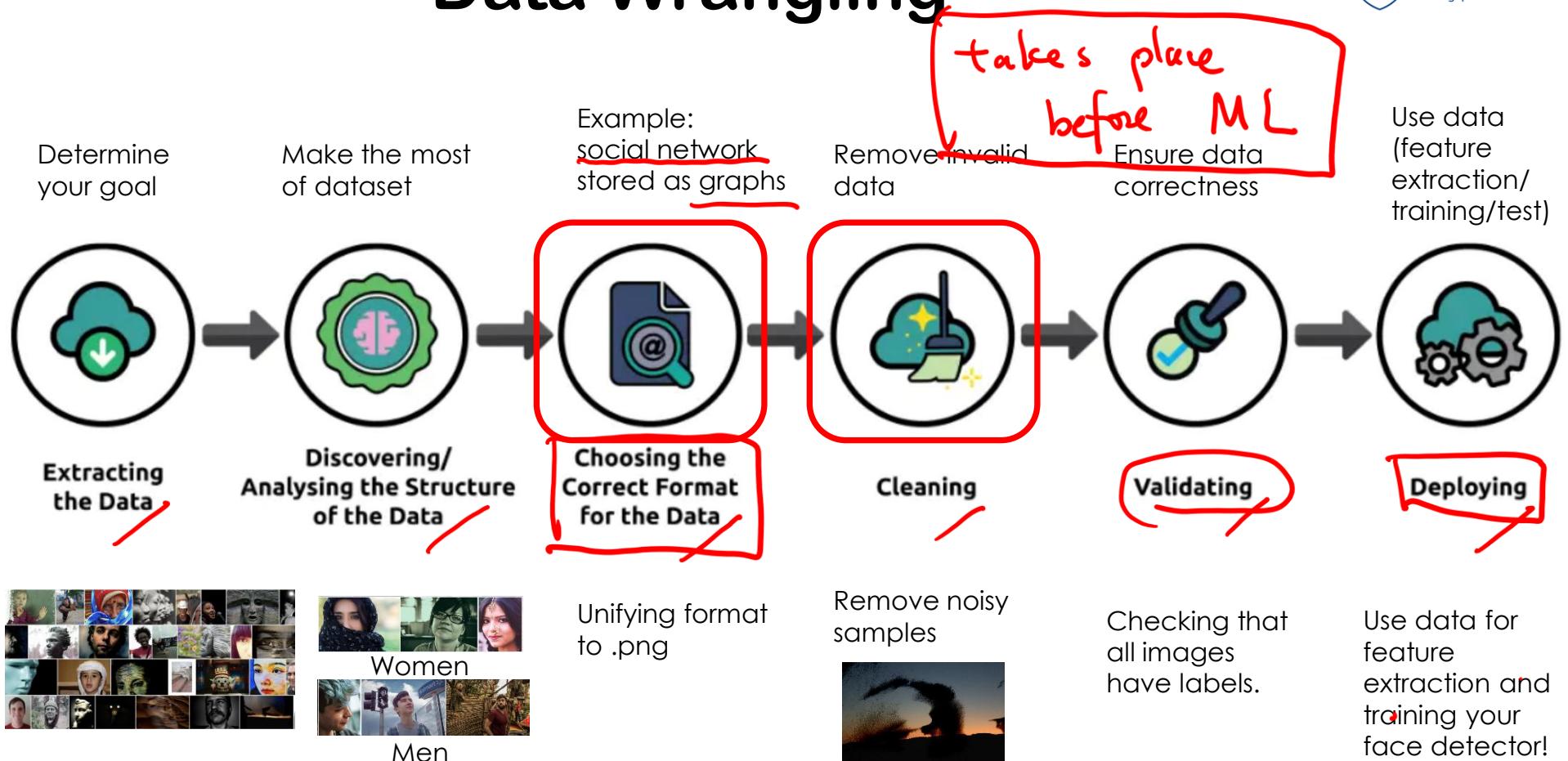
- Data wrangling

- 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!

training data may have values that have much larger numerical ranges/variance. This causes feature space to be highly skewed to those and have zigzags in space. We NORMALISE the data to ensure each feature contributes fairly across samples

and across features. Features $\begin{bmatrix} 2, 5, 17, 50, 99 \end{bmatrix}$ are comparable and no one data/feature dominate the feature space.

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

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

red =	$\begin{bmatrix} 1, 0, 0 \end{bmatrix}$
yellow =	$\begin{bmatrix} 0, 1, 0 \end{bmatrix}$
green =	$\begin{bmatrix} 0, 0, 1 \end{bmatrix}$

only 1 dimension is turned on.

Data Cleaning

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

- Example:

Clipping outliers

Robust Algorithm



histogram

throw away.

Handling missing features

Students	Year of Birth	Gender	Height	GPA
1 Tan Ah Kow	1995	M	1.72	4.2
2 Ahmad Abdul	X NA	M	1.65	4.1
3 John Smith	1995	M	1.75	X NA
4 Chen Lulu	1995	F	X NA	4.0
5 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

week 9.
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. **NA**.
 - 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

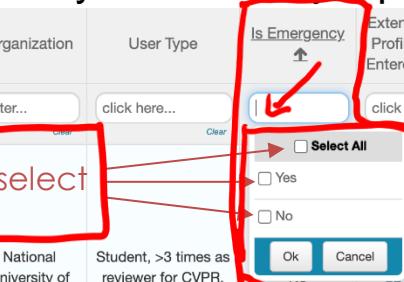
Data
wrangling and
cleaning

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 **C M T**

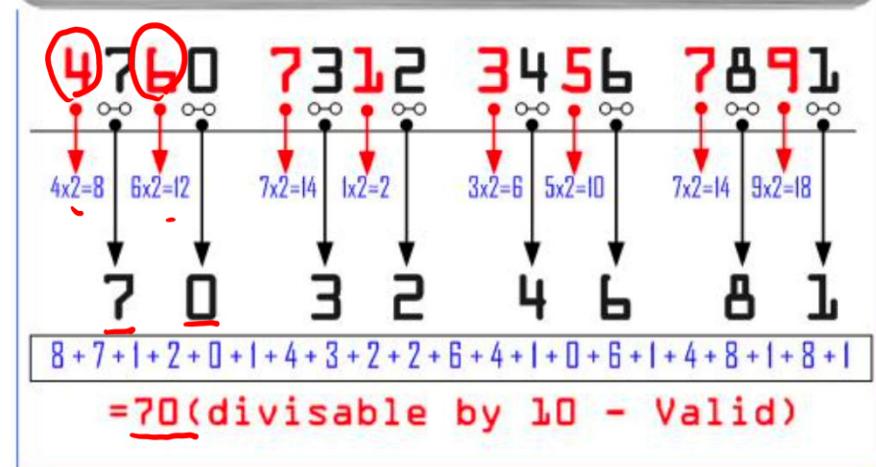
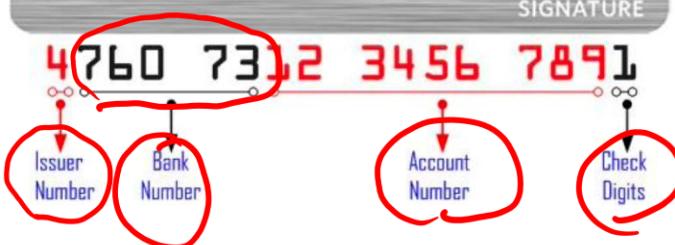
We can only select one of these



The table below shows data entries for different organizations. The 'Is Emergency' column is highlighted with a red box and a red arrow pointing to the 'Yes' radio button in the modal dialog. The 'C M T' text is also highlighted with a red box.

Organization	User Type	Is Emergency	External Profile Entered	Subject Areas	Bid	Relevance	Candidate Suggestion Rank	Tpms Rank	Quota	Number Of Assignments
National University of Singapore	Student, >3 times as reviewer for CVPR, ICCV, or ECCV	<input checked="" type="radio"/> Yes <input type="radio"/> No	click here...	Primary Secondary Machine learning	filter... Clear	e.g. <3 Clear	e.g. <3 Clear	e.g. . Clear	e.g. < Clear	1434 4
Zhejiang University	Faculty/Researcher, 3-10 times as reviewer for CVPR, ICCV, or ECCV	No	Transfer/ low-shot/ long-tail 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			

Data Integrity

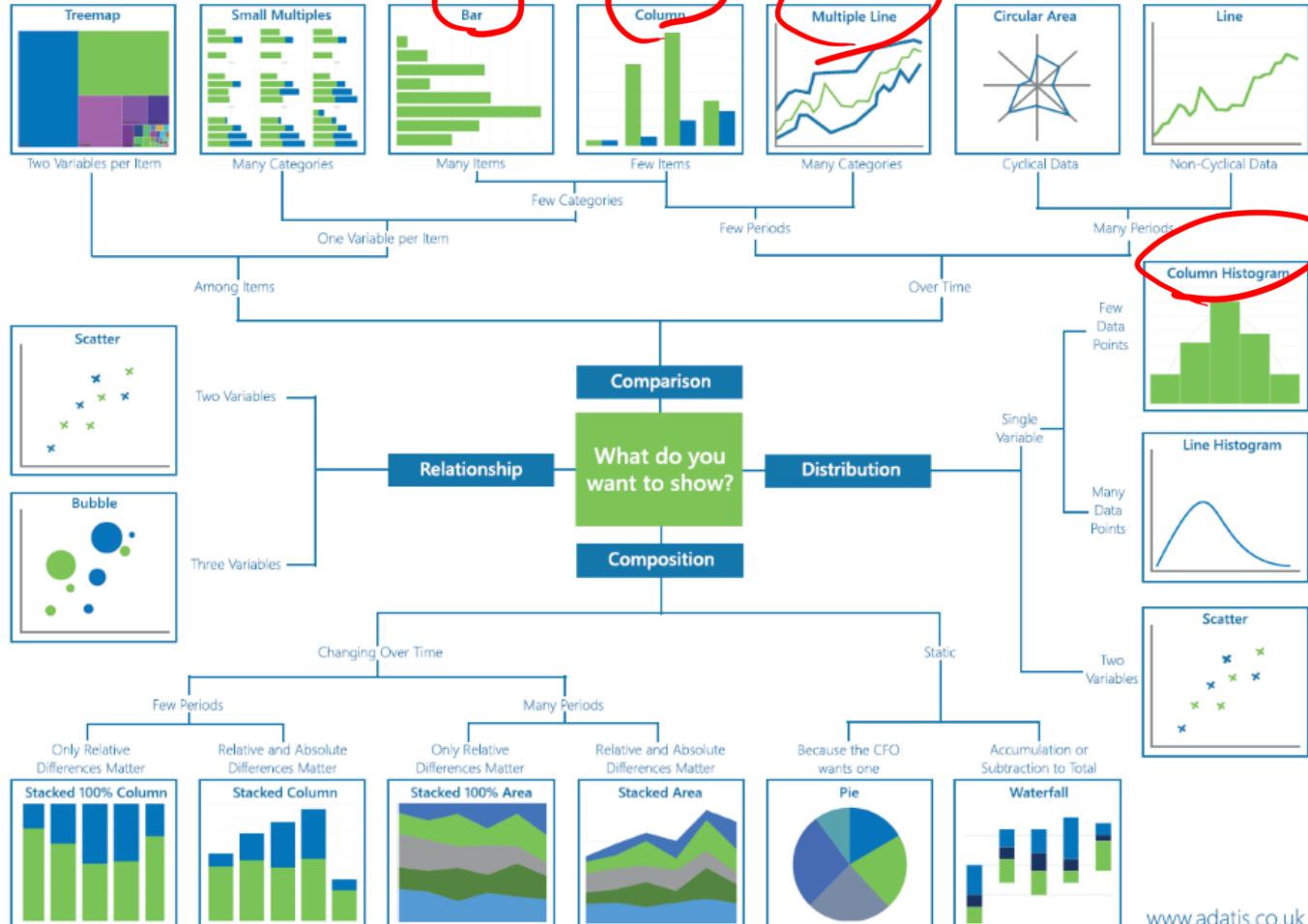


Data Visualization



Chart Types

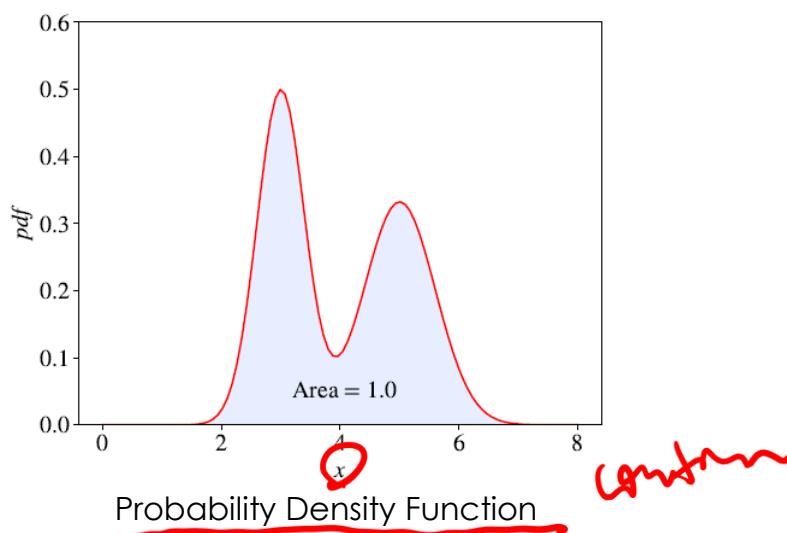
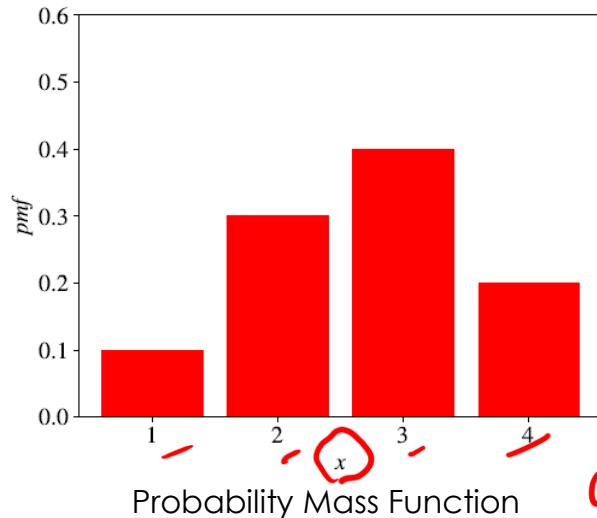
Microsoft Partner
Gold Data Analytics



www.adatis.co.uk

Graphical Representation of data!

Visualization: Distribution



Visualization: Bars

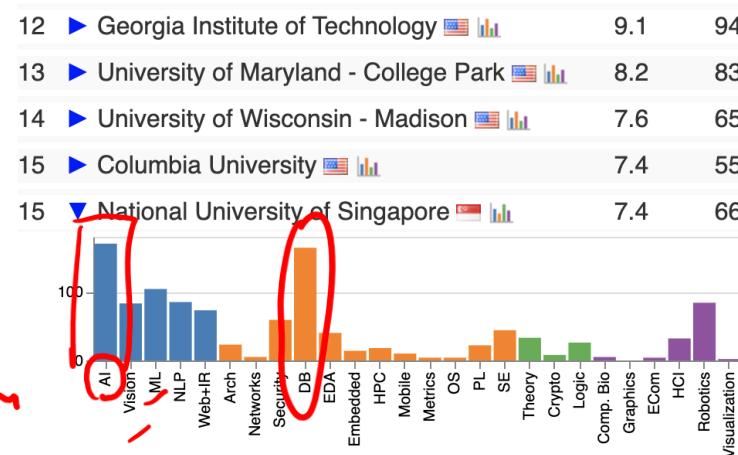
← → ⌂ Not Secure csrankings.org/#/from... ↑ ☆ ↗ ↙

CSRankings: Computer Science Rankings

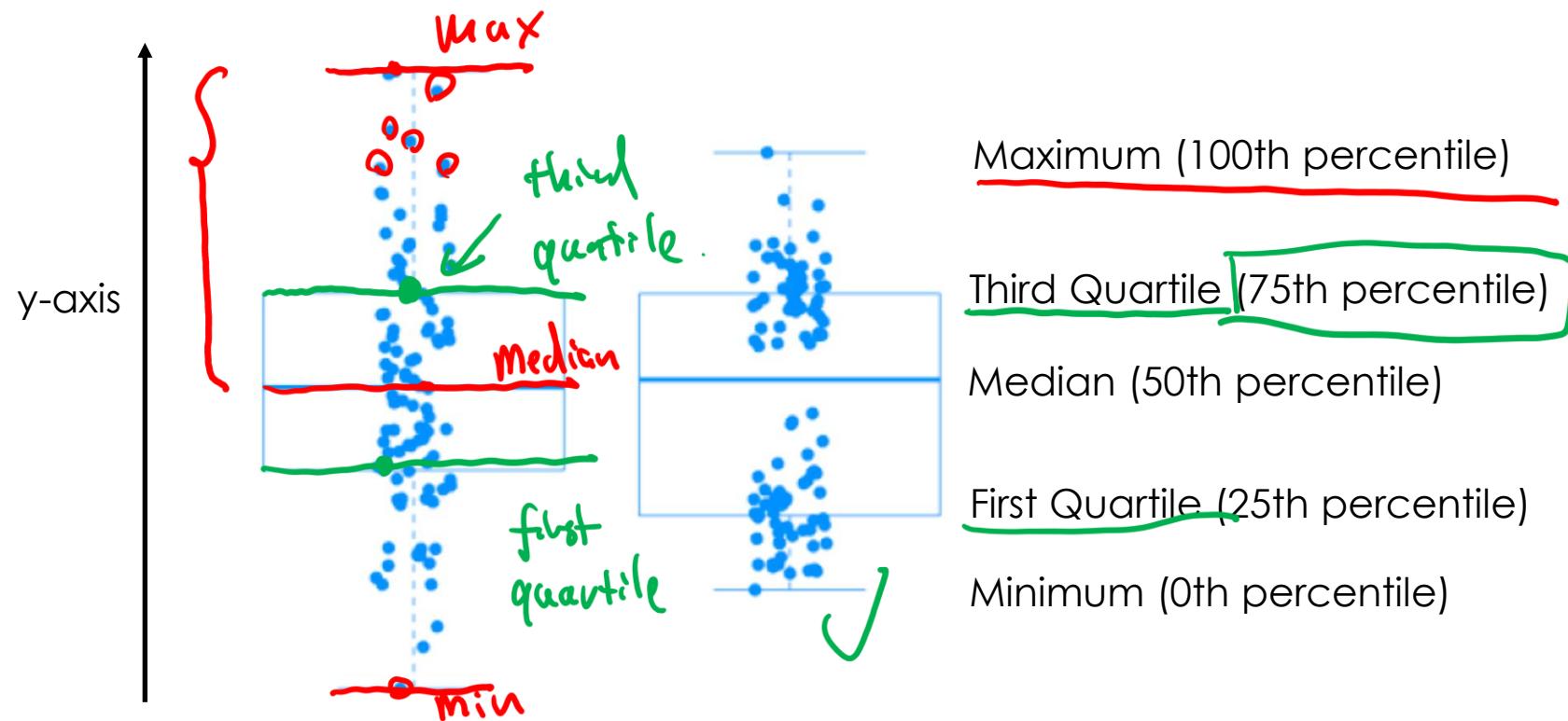
CSRankings is a metrics-based ranking of top computer science institutions around the world. Click on a triangle (\blacktriangleright) to expand areas or institutions. Click on a name to go to a faculty member's home page. Click on a chart icon (the  after a name or institution) to see the distribution 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 the world by publications from 2011 to 2021

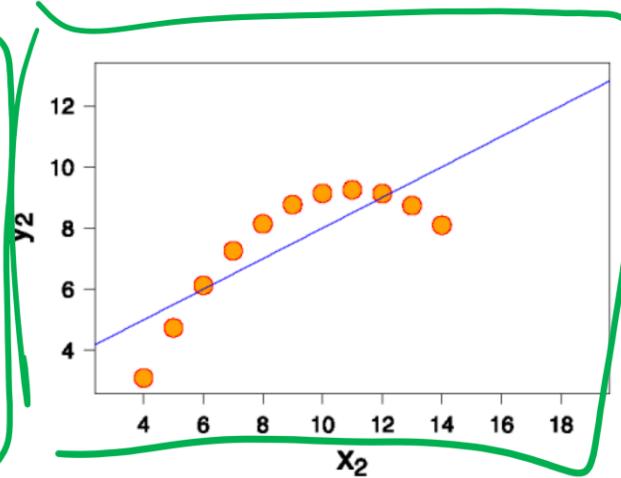
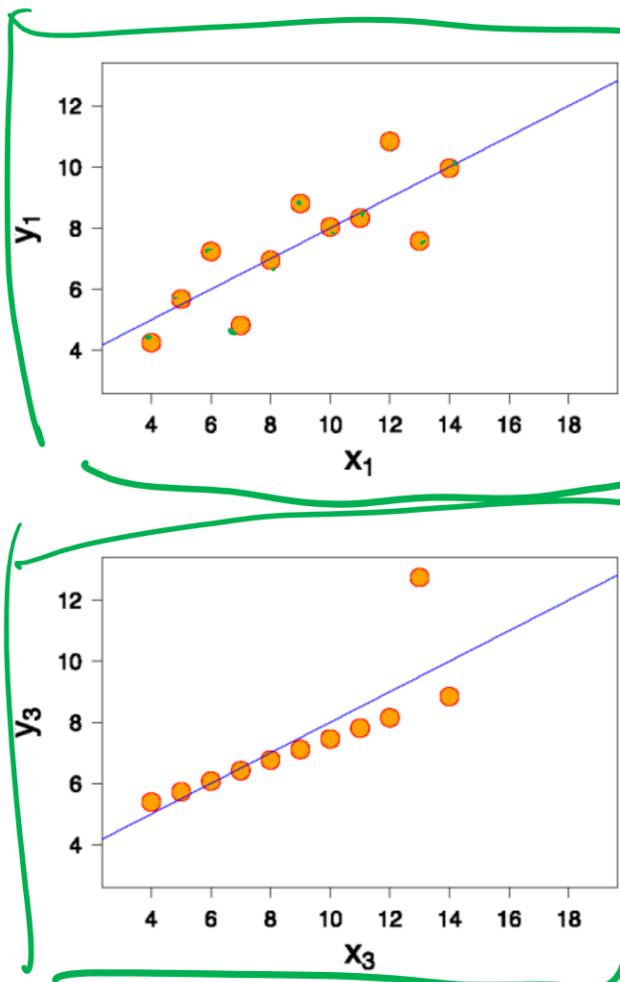


Visualization: Boxplots

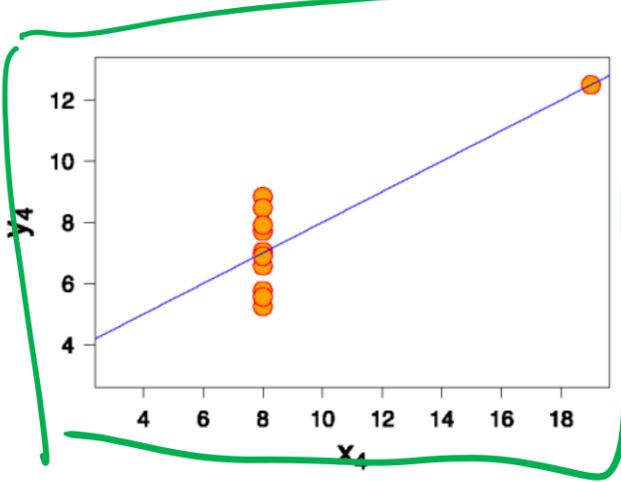
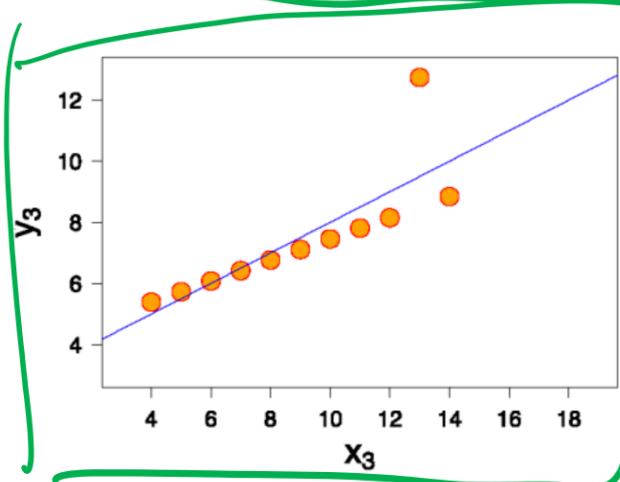


- The first quartile (Q_1) is defined as the middle number between the smallest number (i.e., Minimum) and the median of the data set.
- The third quartile (Q_3) is the middle number between the median and the highest value (i.e., Maximum) of the data set.

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
- ▲
- ◆
-
-
-
-
- ◆

Annotations above the table:

- N (green) points to the first column header "Color".
- O (green) points to the second column header "Size".
- N (green) points to the third column header "Shape".

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?

What about their label, yes/no?

Blue
Large
Ring

input

Yes [No.]

out put

N

EE2211 Introduction to Machine Learning

Lecture 3

lin alg, prob, stats

Wang Xinchao
xinchao@nus.edu.sg

Course Contents

- Introduction and Preliminaries (Xinchao)
 - Introduction
 - Data Engineering
 - **Introduction to Linear Algebra, Probability and Statistics**
- Fundamental Machine Learning Algorithms I (Helen)
 - Systems of linear equations
 - Least squares, Linear regression
 - Ridge regression, Polynomial regression
- Fundamental Machine Learning Algorithms II (Helen)
 - 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

Outline

- (Very Gentle) Introduction to Linear Algebra
 - Prof. Helen's part will follow up
- Causality and Simpson's paradox
 - Understanding at intuitive level is sufficient
- Random Variable, Bayes' Rule

(Very Gentle) Introduction to Linear Algebra

- A **scalar** is a simple numerical value, like 15 or -3.25
 - Focus on **real** numbers
- Variables or constants that take scalar values are denoted by an *italic* letter, like x or a

Notations, Vectors, Matrices

- A **vector** is an ordered list of scalar values
 - Denoted by a **bold character**, e.g. **x** or **a**
 - In many books, vectors are written column-wise:
- $$\mathbf{a} = \begin{bmatrix} 2 \\ 3 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} -2 \\ 5 \end{bmatrix}, \quad \mathbf{c} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \neq \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$
- 2D Vectors*
- The three vectors above are two-dimensional, or have two elements

Notations, Vectors, Matrices

- We denote an entry or attribute of a vector as an italic value with an index, e.g. $\underline{a}^{(j)}$ or $\underline{x}^{(j)}$.
 – The index j denotes a specific dimension of the vector, the position of an attribute in the list

$$\mathbf{a} = \begin{bmatrix} a^{(1)} \\ a^{(2)} \end{bmatrix} = \begin{bmatrix} 2 \\ 3 \end{bmatrix}$$

or more commonly

$$\mathbf{a} = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} 2 \\ 3 \end{bmatrix}$$

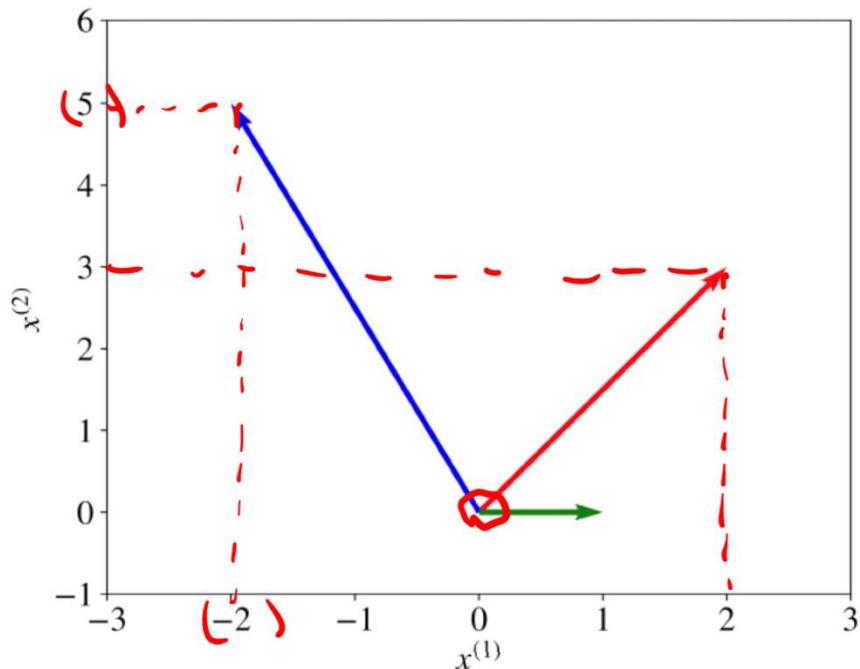
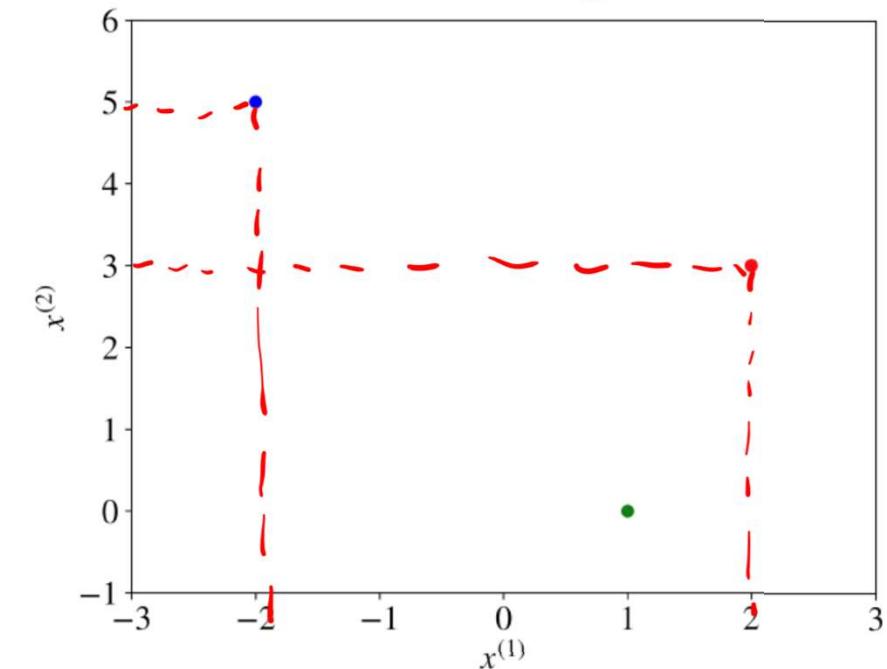
- Note:
 - $x^{(j)}$ is not to be confused with the power operation, e.g., x^2 (squared)
 - Square of an indexed attribute of a vector is denoted as $(x^{(j)})^2$.

Notations, Vectors, Matrices

- **Vectors** can be visualized as, in a multi-dimensional space,
 - arrows that point to some directions, or
 - points

Illustrations of three two-dimensional vectors, $\mathbf{a} = \begin{bmatrix} 2 \\ 3 \end{bmatrix}$, $\mathbf{b} = \begin{bmatrix} -2 \\ 5 \end{bmatrix}$, and $\mathbf{c} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$

$$\mathbf{a} = \begin{bmatrix} 2 \\ 3 \end{bmatrix}, \mathbf{b} = \begin{bmatrix} -2 \\ 5 \end{bmatrix}$$



Notations, Vectors, Matrices

- A **matrix** is a rectangular array of numbers arranged in rows and columns

– Denoted with bold capital letters, such as **X** or **W**

– An example of a matrix with two rows and three columns:

	<i>F₁</i>	<i>F₂</i>	<i>F₃</i>
<i>C₁</i>	2	4	3
<i>C₂</i>	21	-6	1
<i>C₃</i>	<i>x_{1,1}</i>	<i>x_{1,2}</i>	<i>x_{1,3}</i>
	<i>x_{2,1}</i>	<i>x_{2,2}</i>	<i>x_{2,3}</i>

X = [R₁ | R₂]

Sample 1 Sample 2

- A **set** is an unordered collection of unique elements

$$S = \{ 1, 3, 5 \}$$

– When an element x belongs to a set S, we write $x \in S$

– A special set denoted R includes all real numbers from minus infinity to plus infinity

- Note:

– For elements in matrix X, we shall use the indexing $x_{1,1}$ where the first and second indices indicate the row and the column position.

– Usually, for input data, rows represent samples and columns represent features

Notations, Vectors, Matrices

- **Capital Sigma:** the summation over a collection $\{x_1, x_2, x_3, x_4, \dots, x_m\}$ is denoted by:

$$\sum_{i=1}^m x_i = \underline{x_1} + \underline{x_2} + \dots + x_{m-1} + \underline{x_m}$$

starting index Index
Σ

- **Capital Pi:** the product over a collection $\{x_1, x_2, x_3, x_4, \dots, x_m\}$ is denoted by:

$$\prod_{i=1}^m x_i = \underline{x_1} \cdot \underline{x_2} \cdot \dots \cdot x_{m-1} \cdot \boxed{x_m}$$

Systems of Linear Equations

Linear dependence and independence

- A collection of d -vectors $\mathbf{x}_1, \dots, \mathbf{x}_m$ (with $m \geq 1$) is called **linearly dependent** if

$$\beta_1 \mathbf{x}_1 + \cdots + \beta_m \mathbf{x}_m = 0$$

holds for some β_1, \dots, β_m that are not all zero.

- A collection of d -vectors $\mathbf{x}_1, \dots, \mathbf{x}_m$ (with $m \geq 1$) is called **linearly independent** if it is not linearly dependent, which means that

$$\checkmark \beta_1 \mathbf{x}_1 + \cdots + \beta_m \mathbf{x}_m = 0$$

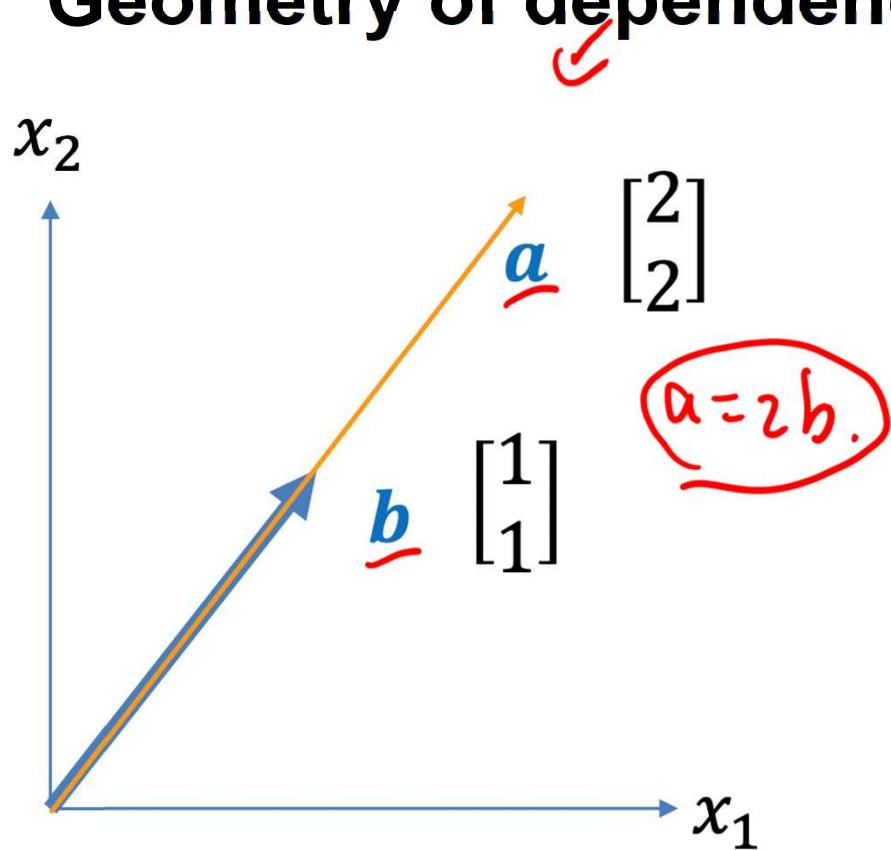
only holds for $\beta_1 = \cdots = \beta_m = 0$.

Note: If all rows or columns of a square matrix X are **linearly independent**, then X is **invertible**.

full rank

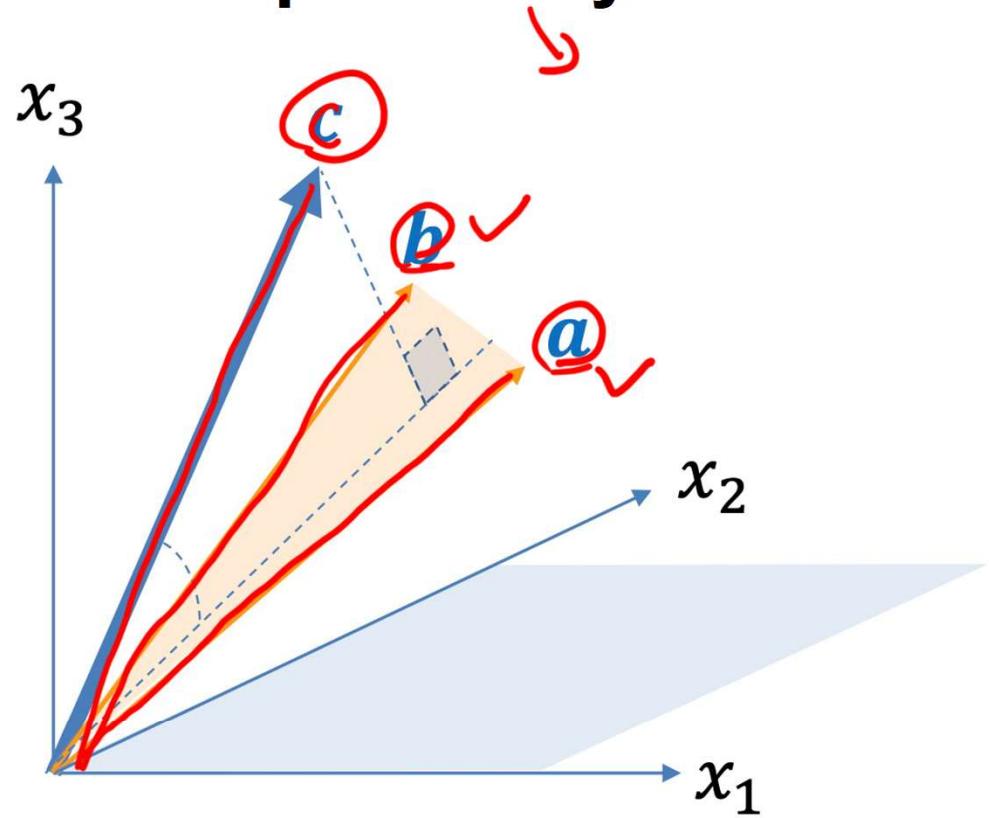
Systems of Linear Equations

Geometry of dependency and independency



$$\beta_1 \mathbf{a} + \beta_2 \mathbf{b} = 0$$

$$\beta_1 = 1, \quad \beta_2 = -2$$



$$\beta_1 \mathbf{a} + \beta_2 \mathbf{b} + \beta_3 \mathbf{c} = 0$$

$$\beta_1, \beta_2, \beta_3 \rightarrow = 0$$

Systems of Linear Equations

These equations can be written compactly in matrix-vector notation:

Where

$$\text{data matrix } \mathbf{X} = \begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m,1} & x_{m,2} & \dots & x_{m,d} \end{bmatrix}^d_m, \quad \mathbf{w} = \begin{bmatrix} w_1 \\ \vdots \\ w_d \end{bmatrix}^d, \quad \mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix}^m.$$

$\mathbf{Xw} = \mathbf{y}$

Given Given goal, unknown

Note:

- The data matrix $\mathbf{X} \in \mathbb{R}^{m \times d}$ and the target vector $\mathbf{y} \in \mathbb{R}^m$ are given
- The unknown vector of parameters $\mathbf{w} \in \mathbb{R}^d$ is to be learnt
- The rank(\mathbf{X}) corresponds to the maximal number of linearly independent columns/rows of \mathbf{X} .

Exercises

- The principled way for computing rank is to do Echelon Form
 - <https://stattrek.com/matrix-algebra/echelon-transform.aspx#MatrixA>
- For small-size matrices, however, the rank is in many cases easy to estimate

- What is the rank of

$$\begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix}$$

2

$$\begin{bmatrix} 1 & 1 \\ 100 & 100 \end{bmatrix}$$

1

$$\begin{bmatrix} 1 & -2 & 3 \\ 0 & -3 & 3 \\ 1 & 1 & 0 \end{bmatrix}$$

2

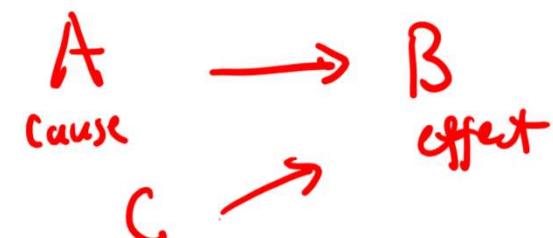
X (2,) X

Outline

- (Very Gentle) Introduction to Linear Algebra
- Causality and Simpson's paradox
- Random Variable, Bayes' Rule

Causality

- Causality, or causation is:
 - The influence by which one event or process (i.e., **cause**) contributes to another (i.e. **effect**),
 - The **cause** is partly responsible for the **effect**, and the **effect** is partly dependent on the **cause**
- Causality relates to an extremely very wide domain of subjects: philosophy, science, management, humanity.
- Causality research is extremely complex
 - Researcher can never be completely certain that there are **no other factors** influencing the causal relationship,
 - In most cases, we can only say “probably” causal.



Causality

- (Probable) causal relations or non-causal?
 - New web design implemented ? Web page traffic increased **C**
 - Your height and weight ? Gets A in EE2211 **N**
 - Uploaded new app store images ? Downloads increased by 2X **C**
 - One works hard and attends lectures/tutorials ? Gets A in EE2211 **C**
 - Your favorite color ? Your GPA in NUS **N**

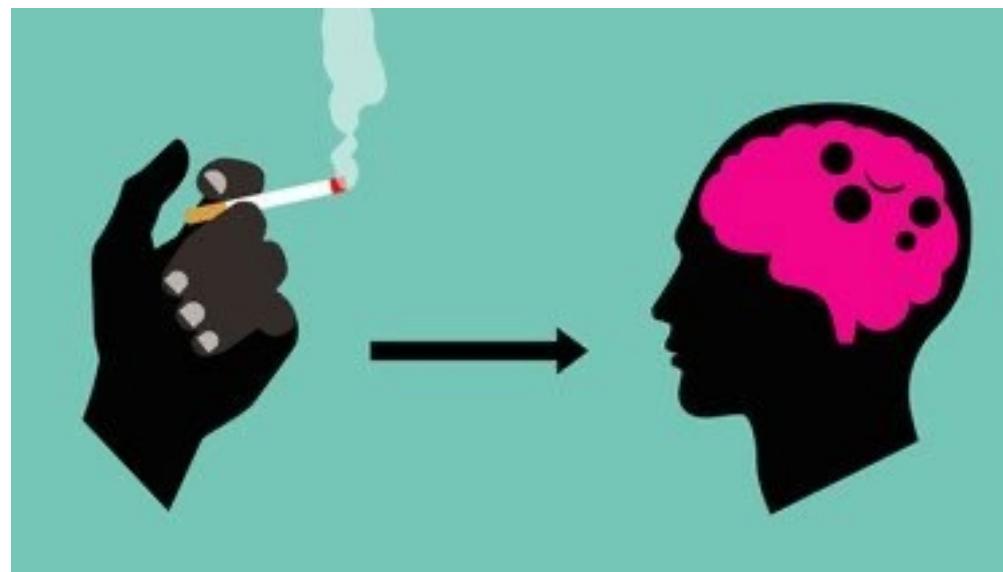
Causality

- One popular way to causal data analysis is Randomized Controlled Trial (RCT)
 - A study design that randomly assigns participants into an experimental group or a control group.
 - As the study is conducted, the only expected difference between two groups is the outcome variable being studied.
- Example:
 - To decide whether smoking and lung cancer has a causal relation, we put participants into experimental group (people who smoke) and control group (people who don't smoke), and check whether they develop lung cancer eventually.
- RCT is sometimes infeasible to conduct, and also has moral issues.

Causality is a statistical relationship

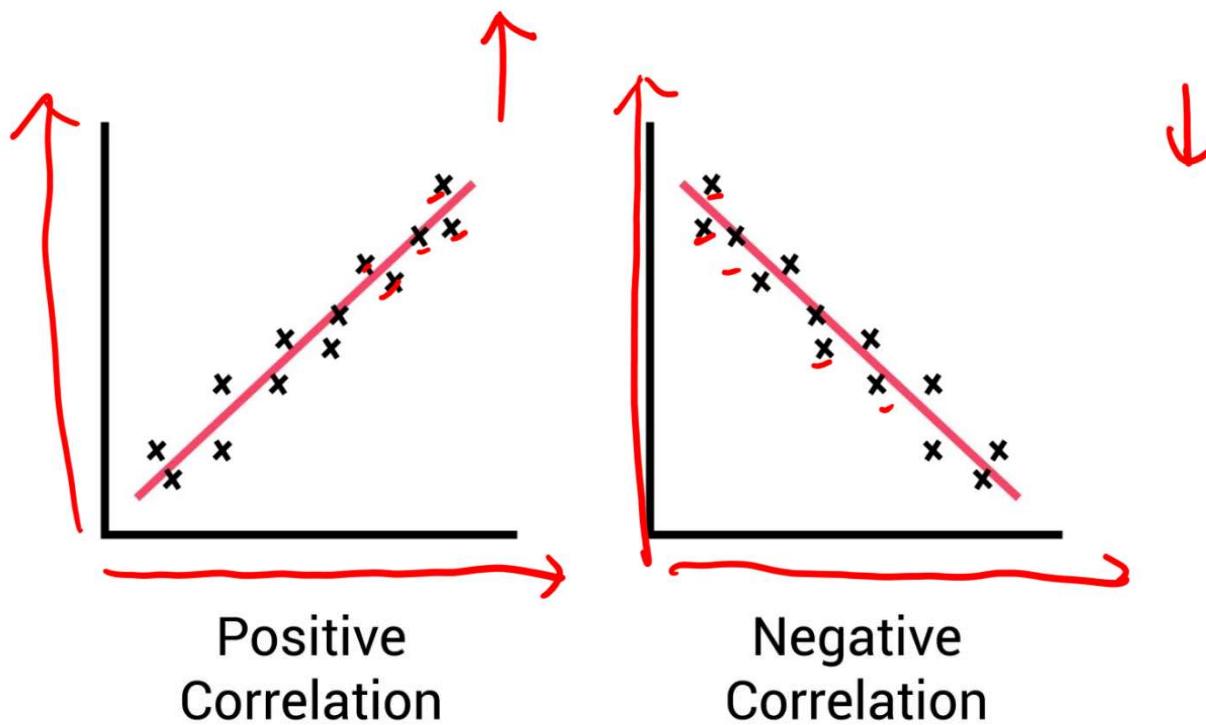


- Decades of data show a clear causal relationship between smoking and cancer.
- If one smokes, it is a sure thing that his/her risk of cancer will increase.
- But it is not a sure thing that one will get cancer.
- The relationship is not deterministic.



Correlation (vs Causality)

- In statistics, **correlation** is any **statistical relationship**, whether causal or not, between two random variables.
- Correlations are useful because they can indicate a **predictive relationship** that can be exploited in practice.

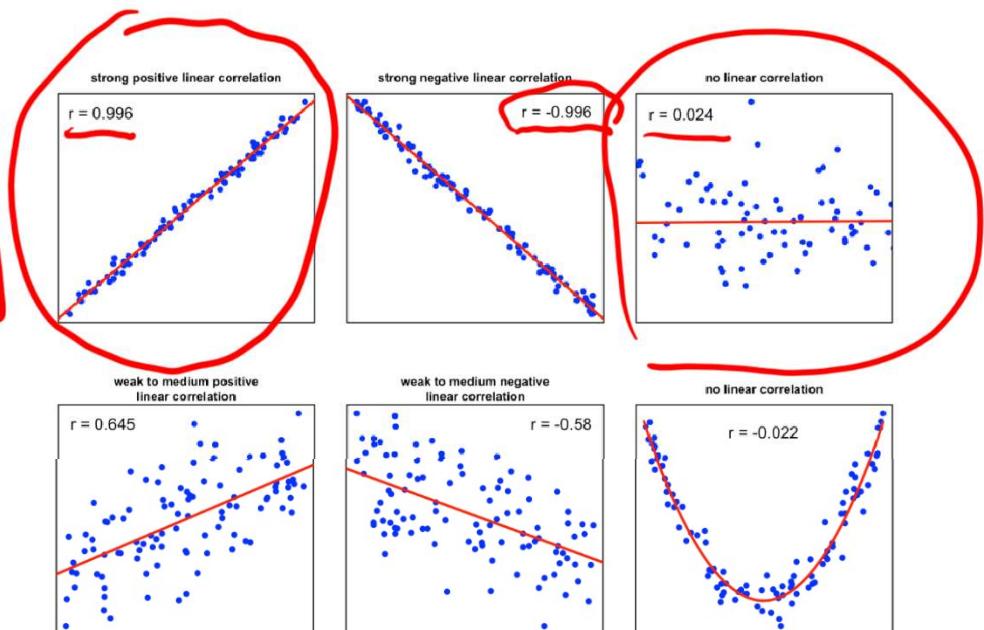


Correlation (vs Causality)

- Linear correlation coefficient, r , which is also known as the Pearson Coefficient.

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} = \frac{s_{xy}}{s_x s_y},$$

Strong linear relationship	$r > 0.9$
Medium linear relationship	$0.7 < r \leq 0.9$
Weak linear relationship	$0.5 < r \leq 0.7$
No or doubtful linear relationship	$0 < r \leq 0.5$

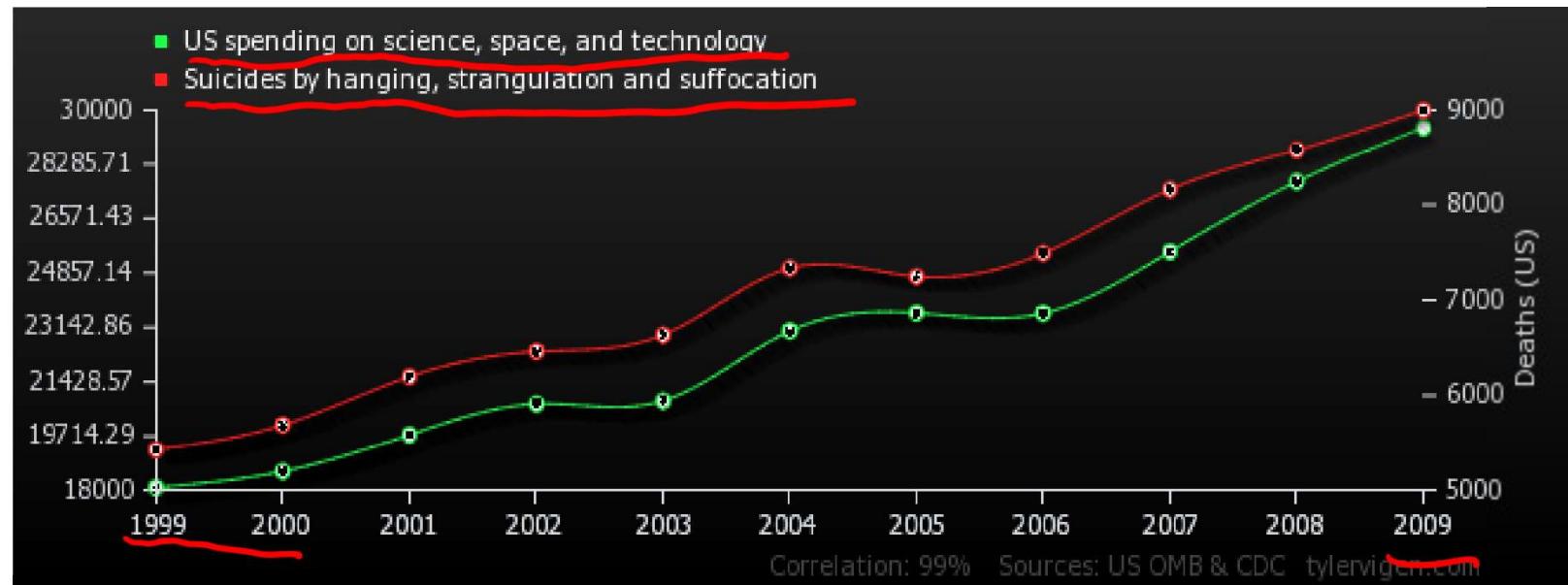


The same holds for negative values.

1.<https://www.geo.fu-berlin.de/en/v/soga/Basics-of-statistics/Descriptive-Statistics/Measures-of-Relation-Between-Variables/Correlation/index.html>

Correlation does not imply causation!

- Some great examples of correlations that can be calculated but are clearly not causally related appear at <http://tylervigen.com/>

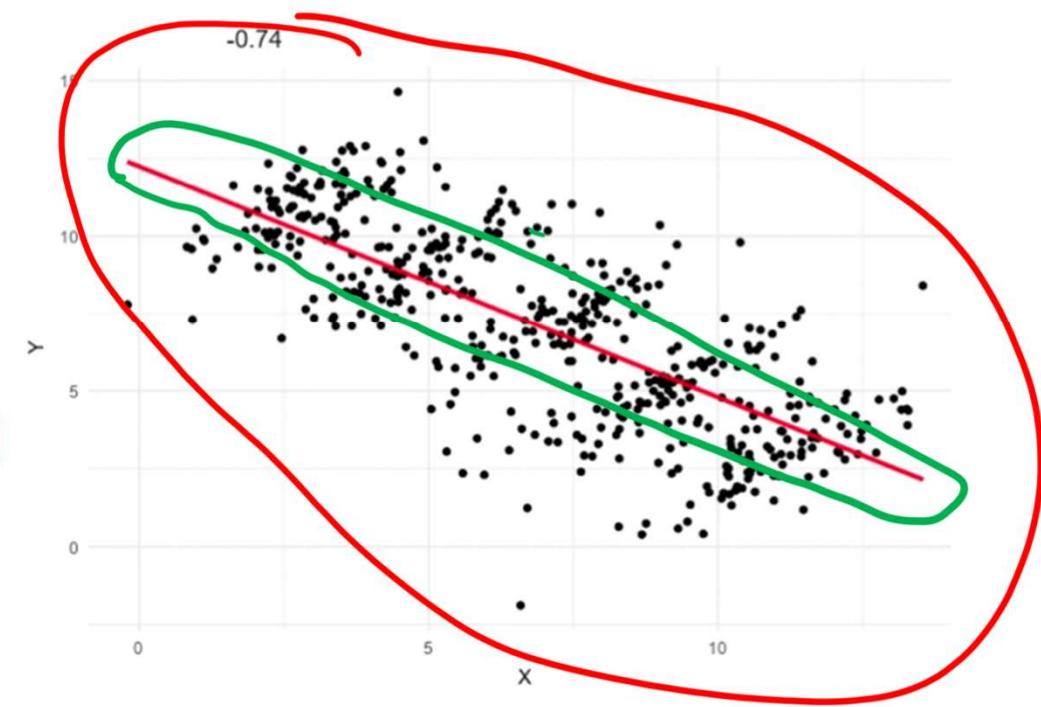
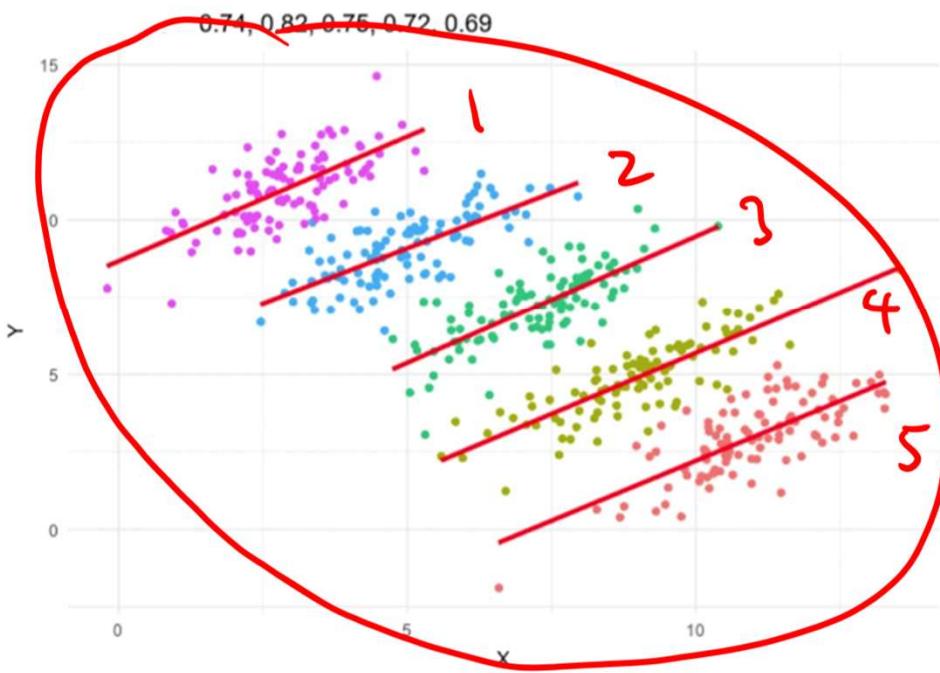


Simpson's paradox

fact

- Simpson's paradox is a phenomenon in probability and statistics, in which a trend appears in several different groups of data but disappears or reverses when these groups are combined.

small trends reverse or disappear in big



The same set of samples!

Example

- Batting Average in professional baseball game
- Two well-known players, Derek Jeter and David Justice

Batter \ Year	1995	1996	Combined
Derek Jeter	12/48	.250	195/630 .310
David Justice	104/411	.253	149/551 .270

#of wins #of games

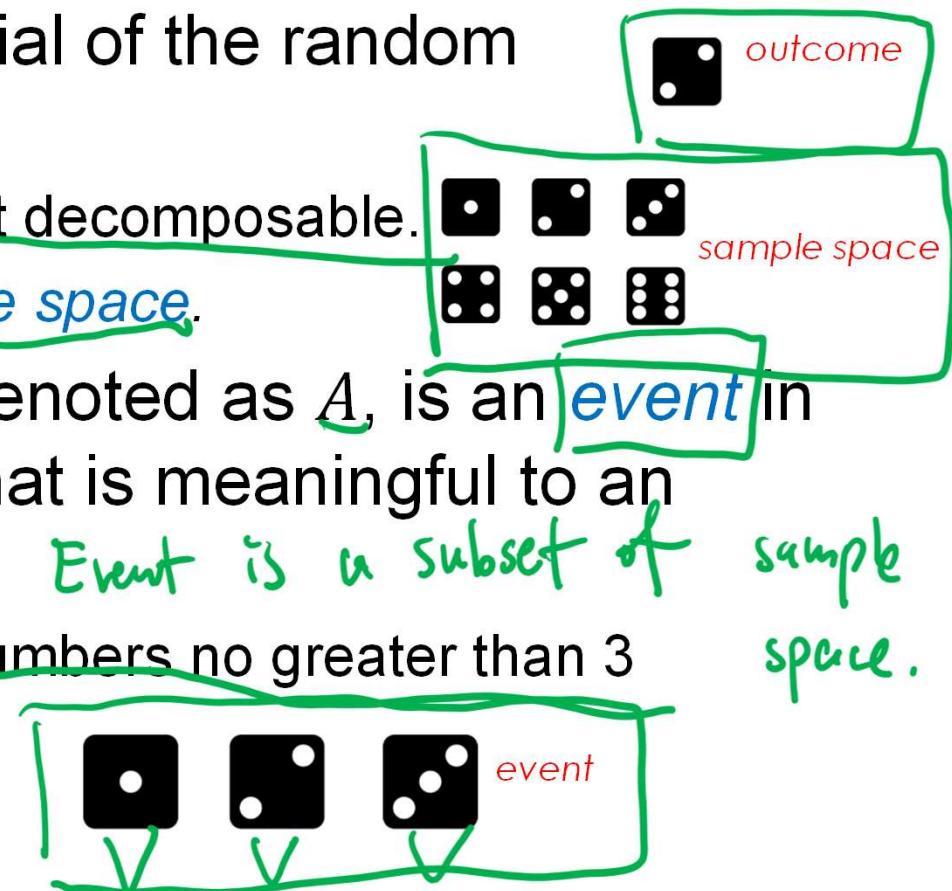
Outline

- (Very Gentle) Introduction to Linear Algebra
- Causality and Simpson's paradox
- Random Variable, Bayes' Rule

Probability

outcomes (not decomposable) vs events (specific group of possible outcomes, hence decomposable and subset of the sample space)

- We describe a *random experiment* by describing its procedure and observations of its outcomes.
- Outcomes* are mutual exclusive in the sense that only one outcome occurs in a specific trial of the random experiment.
 - This also means an outcome is not decomposable.
 - All unique outcomes form a sample space.
- A subset of sample space S , denoted as A , is an event in a random experiment $A \subset S$, that is meaningful to an application.
 - Example of an event: faces with numbers no greater than 3



Axioms of Probability

Assuming events $A \subseteq S$ and $B \subseteq S$, the probabilities of events related with and must satisfy,

$$1. \Pr(A) \geq 0$$

$$2. \Pr(S) = 1$$

$$3. \text{If } A \cap B = \emptyset, \text{ then } \Pr(A \cup B) = \Pr(A) + \Pr(B)$$

$$\text{*otherwise, } \Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$$

[https://en.wikipedia.org/wiki/Union_\(set_theory\)](https://en.wikipedia.org/wiki/Union_(set_theory))

[https://en.wikipedia.org/wiki/Intersection_\(set_theory\)](https://en.wikipedia.org/wiki/Intersection_(set_theory))

Random Variable RV.

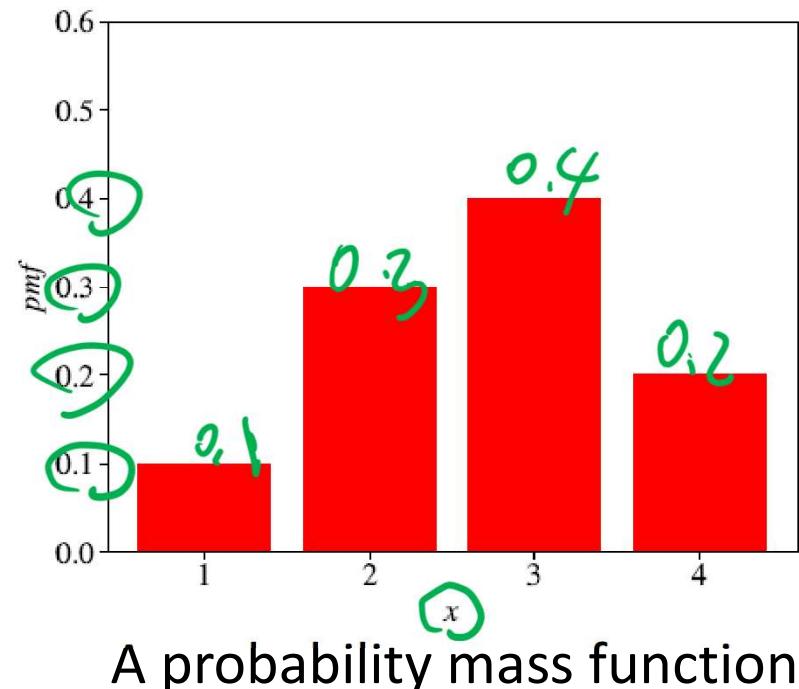
- A **random variable**, usually written as an *italic capital letter*, like X , is a variable whose possible values are numerical outcomes of a random event.
- There are two types of random variables: **discrete** and **continuous**.

Notations

- Some books used $P(\cdot)$ and $p(\cdot)$ to distinguish between the probability of discrete random variable and the probability of continuous random variables respectively.
- We shall use $\boxed{Pr(\cdot)}$ for both the above cases

Discrete random variable

- A discrete random variable (DRV) takes on only a countable number of distinct values such as red, orange, blue or 1, 2, 3.
- The probability distribution of a discrete random variable is described by a list of probabilities associated with each of its possible values.
- This list of probabilities is called a probability mass function (pmf).
 - Like a histogram, except that here the probabilities sum to 1



Discrete random variable

- Let a **discrete random variable X** have k possible values

$$\{x_i\}_{i=1}^k.$$

- The **expectation** of X denoted as $E(x)$ is given by,

$$\begin{aligned} E(x) &\stackrel{\text{def}}{=} \sum_{i=1}^k [x_i \cdot \Pr(X = x_i)] \\ &= \underline{x_1} \cdot \underline{\Pr(X = x_1)} + \underline{x_2} \cdot \underline{\Pr(X = x_2)} + \cdots + \underline{x_k} \cdot \underline{\Pr(X = x_k)} \end{aligned}$$

where $\Pr(X = x_i)$ is the probability that X has the value x_i according to the **pmf**.

- The **expectation** of a random variable is also called the **mean, average** or **expected value** and is frequently denoted with the letter μ .

Discrete random variable

- Another important statistic is the standard deviation, defined as,

$$\sigma \stackrel{\text{def}}{=} \sqrt{E[(X - \mu)^2]} .$$

- Variance, denoted as σ^2 or $var(X)$, is defined as,

$$\sigma^2 = E[(X - \mu)^2]$$

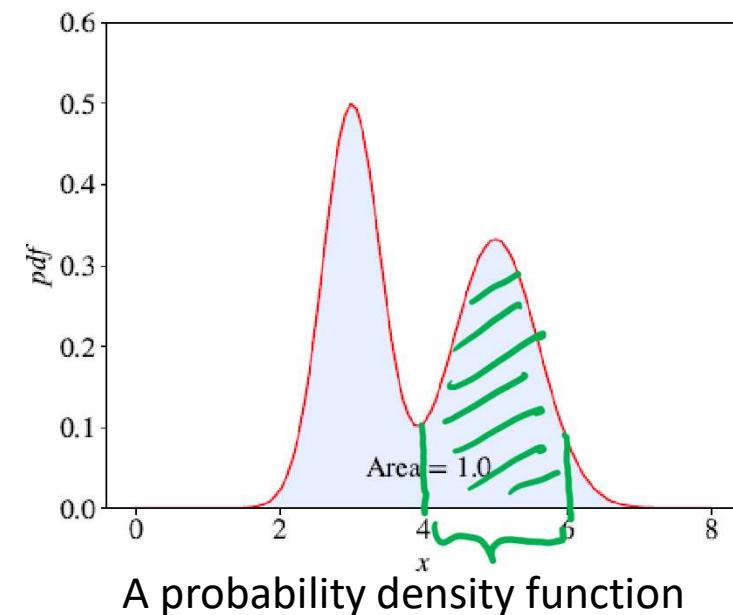
- For a discrete random variable, the standard deviation is given by

$$\sigma = \sqrt{\Pr(X = x_1)(x_1 - \mu)^2 + \cdots + \Pr(X = x_k)(x_k - \mu)^2}$$

where $\mu = E(X)$.

Continuous random variable

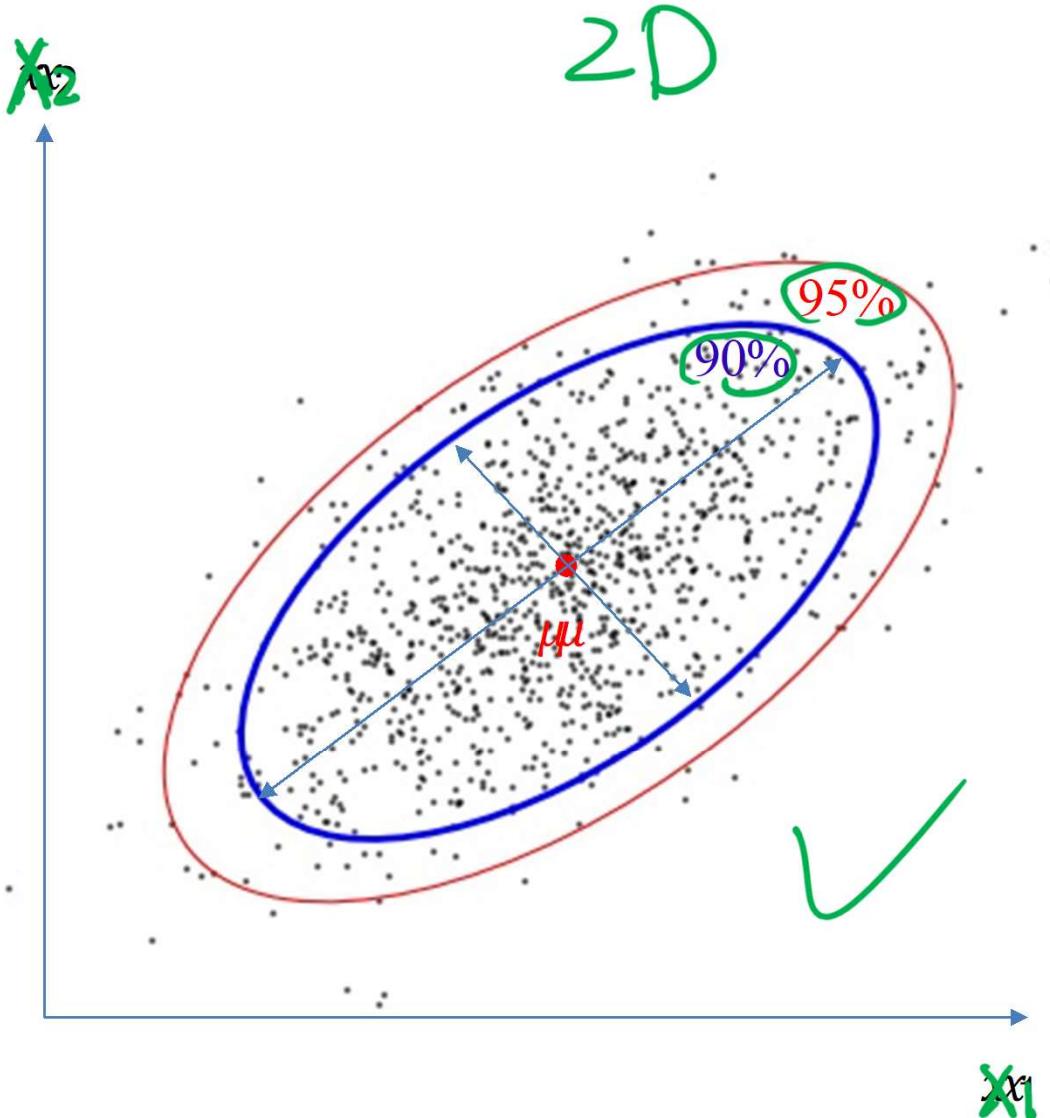
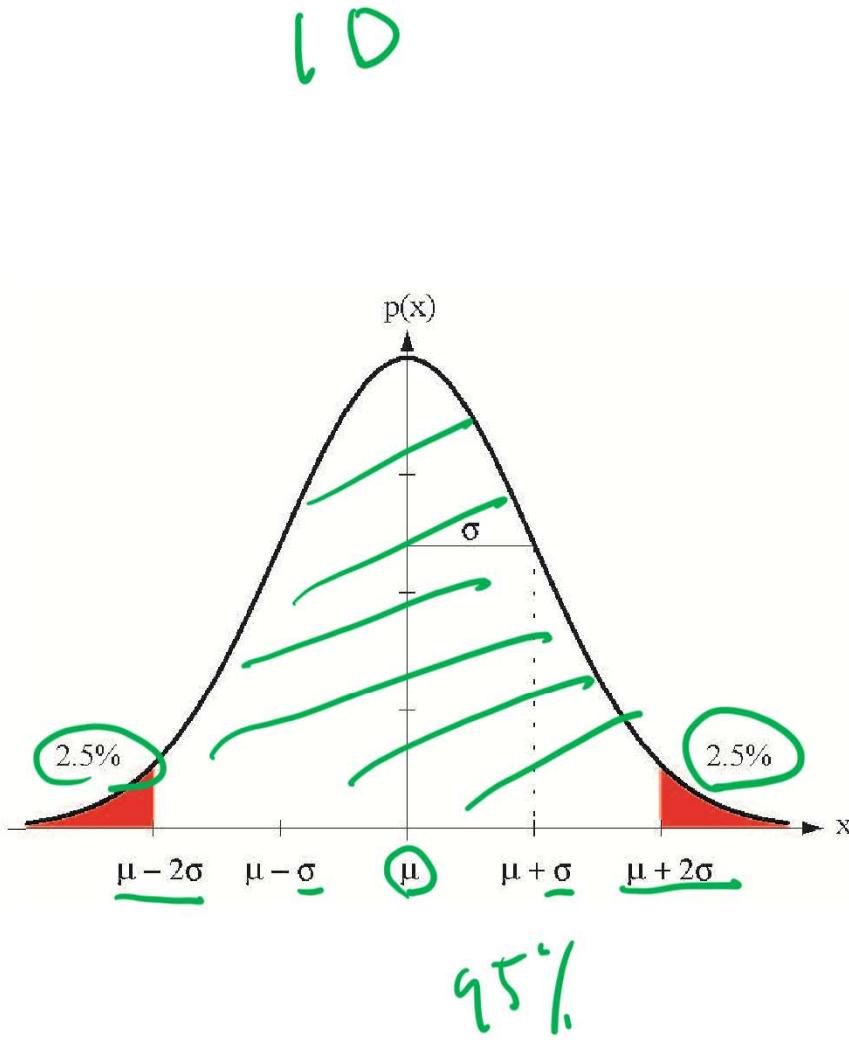
- A continuous random variable (CRV) takes an infinite number of possible values in some interval.
 - Examples include height, weight, and time. *1.7325 m*
 - The number of values of a continuous random variable X is infinite, the probability $\Pr(X = c)$ for any c is 0
 - Therefore, instead of the list of probabilities, the probability distribution of a CRV (a continuous probability distribution) is described by a probability density function (pdf).
 - The pdf is a function whose range is nonnegative and the area under the curve is equal to 1.



Continuous random variable

- The **expectation** of a continuous random variable X is given by $E[x] \stackrel{\text{def}}{=} \int_R x f_X(x) dx$ where f_X is the **pdf** of the variable X and \int_R is the integral of function $x f_X$. $\sum x f(x)$
- The **variance** of a continuous random variable X is given by $\sigma^2 \stackrel{\text{def}}{=} \int_R (X - \mu)^2 f_X(x) dx$ $\int x^2 f(x) - (\text{S} \text{u} \text{t}(x))^2$
- Integral is an equivalent of the summation over all values of the function when the function has a continuous domain.
- It equals the area under the curve of the function.
- The property of the pdf that the area under its curve is 1 mathematically means that $\int_R f_X(x) dx = 1$

Mean and Standard Deviation of a Gaussian Distribution



Example 1

- **Independent random variables**
- Consider tossing a fair coin twice, what is the probability of having (H,H)? Assuming a coin has two sides, H=head and T=Tail

$$-\Pr(x=H, y=H) = \Pr(x=H)\Pr(y=H) = (1/2)(1/2) = 1/4$$

Example 2

- **Dependent random variables**
- Given 2 balls with different colors (Red and Black), what is the probability of first drawing B and then R? Assuming we are drawing the balls without replacement.

don't put it back

- The space of outcomes of taking two balls sequentially without replacement:

B-R R-B

– Thus having B-R is $1/2$.

$$P(y=R \mid x=B)$$

something has happened.

- Mathematically:

$$\Pr(x=B, y=R) = \Pr(y=R \mid x=B) \Pr(x=B) = 1 \times (1/2) = 1/2$$

Conditional Probability

Example 3

- **Dependent random variables**
- Given 3 balls with different colors (R, G, B), and we draw 2 balls. What is the probability of first having B and then G, if we draw without replacement?
- The space of outcomes of taking two balls sequentially without replacement:

R-G | G-B | B-R

R-B | G-R | B-G

Thus, $\Pr(y=G, x=B) = \underline{1/6}$

- Mathematically:

$$\begin{aligned}
 \Pr(y=G, x=B) &= \Pr(y=G | x=B) \Pr(x=B) \\
 &= (1/2) \times (1/3) \\
 &= 1/6
 \end{aligned}$$

Two Basic Rules

1. • Sum Rule

$$\Pr(X = x) = \sum_Y \Pr(X = x, Y = y_i)$$

of variable within ().

2 • Product Rule

$$\underbrace{\Pr(X = x, Y = y)}_{\text{joint}} = \underbrace{\Pr(Y = y | X = x)}_{\text{conditional}} P(X = x)$$

[Independent Assumption]

Bayes' Rule

- The conditional probability $\Pr(Y = y|X = x)$ is the probability of the random variable Y to have a specific value y , given that another random variable X has a specific value of x .
- The **Bayes' Rule** (also known as the **Bayes' Theorem**):

$$\Pr(Y = y|X = x) = \frac{\Pr(X = x|Y = y) \Pr(Y = y)}{\Pr(X = x)}$$

Annotations on the equation:

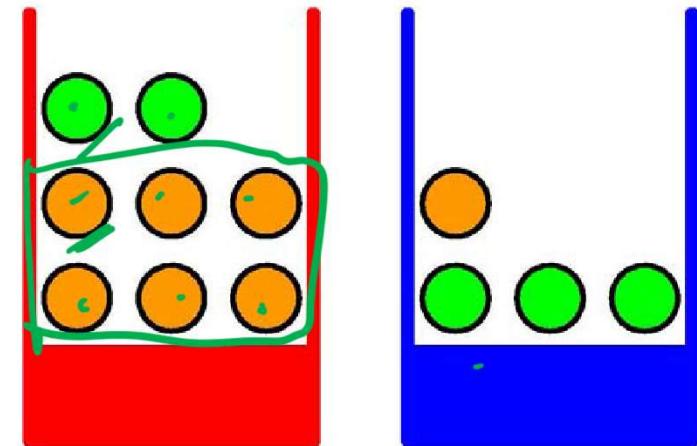
- posterior**: A green bracket under the term $\Pr(Y = y|X = x)$.
- likelihood**: A green bracket under the term $\Pr(X = x|Y = y)$.
- prior**: A green bracket under the term $\Pr(Y = y)$.
- evidence**: A red bracket under the term $\Pr(X = x)$.

Example

- Drawing a sample of fruit from a box
 - First pick a box, and then draw a sample of fruit from it
 - B : variable for Box, can be r (red) or b (blue)
 - F : variable for Fruit, can be o (orange) or a (apple)

- $\Pr(B=r) = 0.4$
- $\Pr(F=o | B=r) = 0.75$
- $\Pr(F=o) = 0.45$

picked red box
 prior
 likelihood
 evidence



✓

- $\Pr(B=r | F=o) = \Pr(F=o | B=r) * \Pr(B=r) / \Pr(F=o)$

$$= 0.75 * 0.4 / 0.45 = 2/3$$

posterior

$\boxed{\Pr(B=b | F=o)} = 1/3 ?$

$1 - \frac{2}{3} \approx \frac{1}{3}$

Summary

- (Very Gentle) Introduction to Linear Algebra
- Causality and Simpson's paradox
- Random Variable, Bayes' Rule

Practice Question

(Type of Question to Expect in Exams)

Suppose the random variable X has the following probability mass function (pmf) listed in the table below. k is unknown.

	1	2	3	4	5
$\Pr[X]$	0.1	0.05	0.05	0.6	$k = 0.2$

Handwritten annotations: A green arrow points from the value 0.8 at the top right towards the row for Pr[X]. The cell for Pr[X=1] is circled in green. The cell for Pr[X=3] is circled in green. The cell for Pr[X=5] is circled in green. The cell for Pr[X=k] is circled in green, with the value 0.2 written next to it. There are also green checkmarks and U-shaped arrows under the Pr[X] column for the values 0.1, 0.05, and 0.6.

What is the probability that X takes a value of odd numbers?

1) $0.1 + 0.05 + 0.2 = 0.35$.

2) $1 - (0.05 + 0.6) = 0.35$