Basics of Machine Learning and Kmeans Clustering

Agenda

Through this presentation we will try to understand:

- Background of Machine Learning
- Types of Machine Learning Algorithms
- Introduction to Unsupervised and Supervised Learning Algorithms
- Evaluation of Machine Learning Algorithms

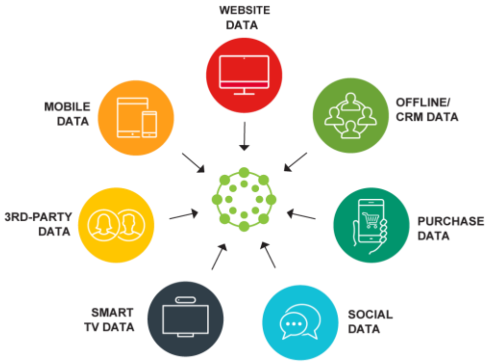
How much data do we create every single day?

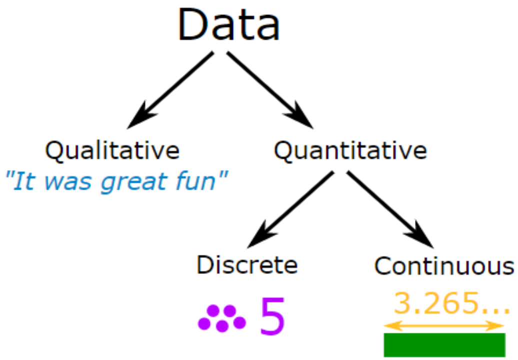
2,500,000,000,000,000,000

(Two and half quintillion)

What is data?

Data is a collection of facts, such as numbers, words, measurements, observations or even just descriptions of things.





A DAY IN DATA

The exponential growth of data is undisputed, but the numbers behind this explosion – fuelled by internet of things and the use of connected devcies – are hard to comprehend, particularly when looked at in the context of one day

500m
tweets are sent every day
Twitter

4PB
of data created by Facebook, including
350m photos
100m hours of video watch time
Facebook Research

294bn
billion emails are sent
Radicati Group

320bn
emails to be sent each day by 2021

306bn
emails to be sent each day by 2020

3.9bn
people use emails

4TB
of data produced by a connected car
Intel

DEMYSTIFIYING DATA UNITS

From the more familiar 'bit' or 'megabyte', larger units of measurement are more frequently being used to explain the masses of data

Unit		Value	Size
b	bit	0 or 1	1/8 of a byte
B	byte	8 bits	1 byte
KB	kilobyte	1,000 bytes	1,000 bytes
MB	megabyte	$1{,}000^2$ bytes	1,000,000 bytes
GB	gigabyte	$1{,}000^3$ bytes	1,000,000,000 bytes
TB	terabyte	$1{,}000^4$ bytes	1,000,000,000,000 bytes
PB	petabyte	$1{,}000^5$ bytes	1,000,000,000,000,000 bytes
EB	exabyte	$1{,}000^6$ bytes	1,000,000,000,000,000,000 bytes
ZB	zettabyte	$1{,}000^7$ bytes	1,000,000,000,000,000,000,000 bytes
YB	yottabyte	$1{,}000^8$ bytes	1,000,000,000,000,000,000,000,000 bytes

*A lowercase "b" is used as an abbreviation for bits, while an uppercase "B" represents bytes.

65bn
messages sent over WhatsApp and two billion minutes of voice and video calls made
Facebook

Searches made a day	5bn
Searches made a day from Google	3.5bn

Smart insights

463EB
of data will be created every day by 2025
IDC

95m
photos and videos are shared on Instagram
Instagram Business

28PB
to be generated from wearable devices by 2020
Statista

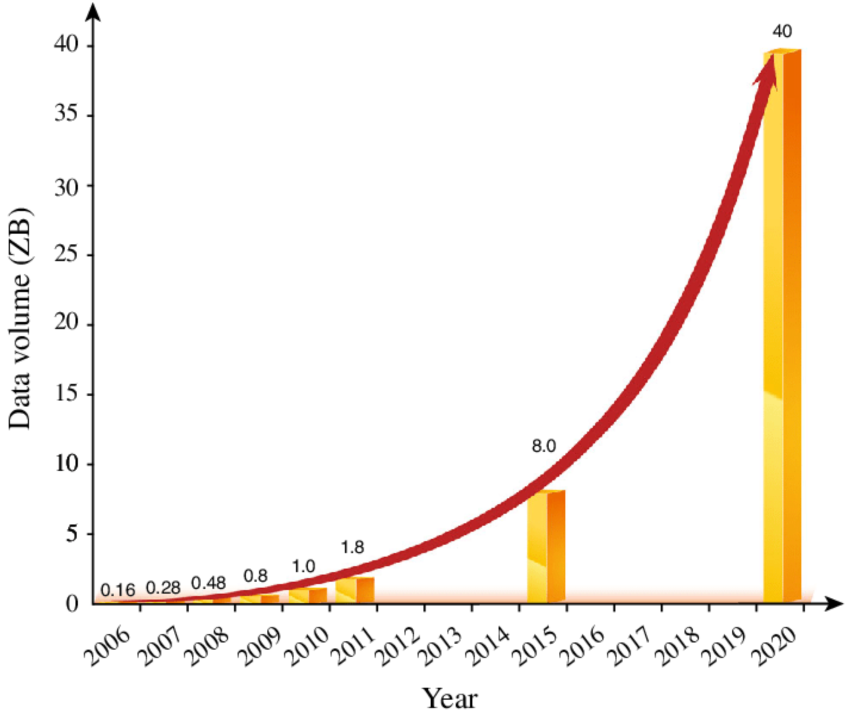
ACCUMULATED DIGITAL UNIVERSE OF DATA

4.4ZB
44ZB

Source: Visual Capitalist

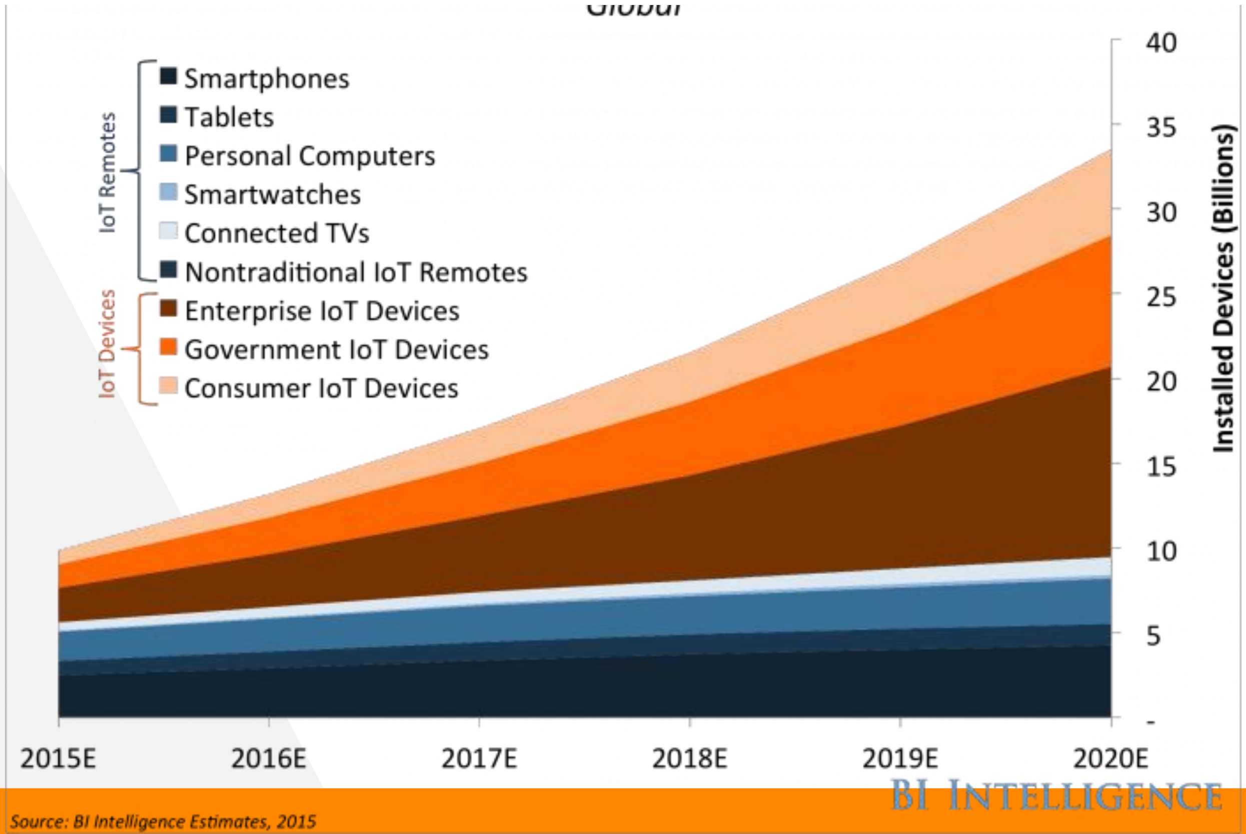
Significant growth in Data Science and AI

Explosion of data volume

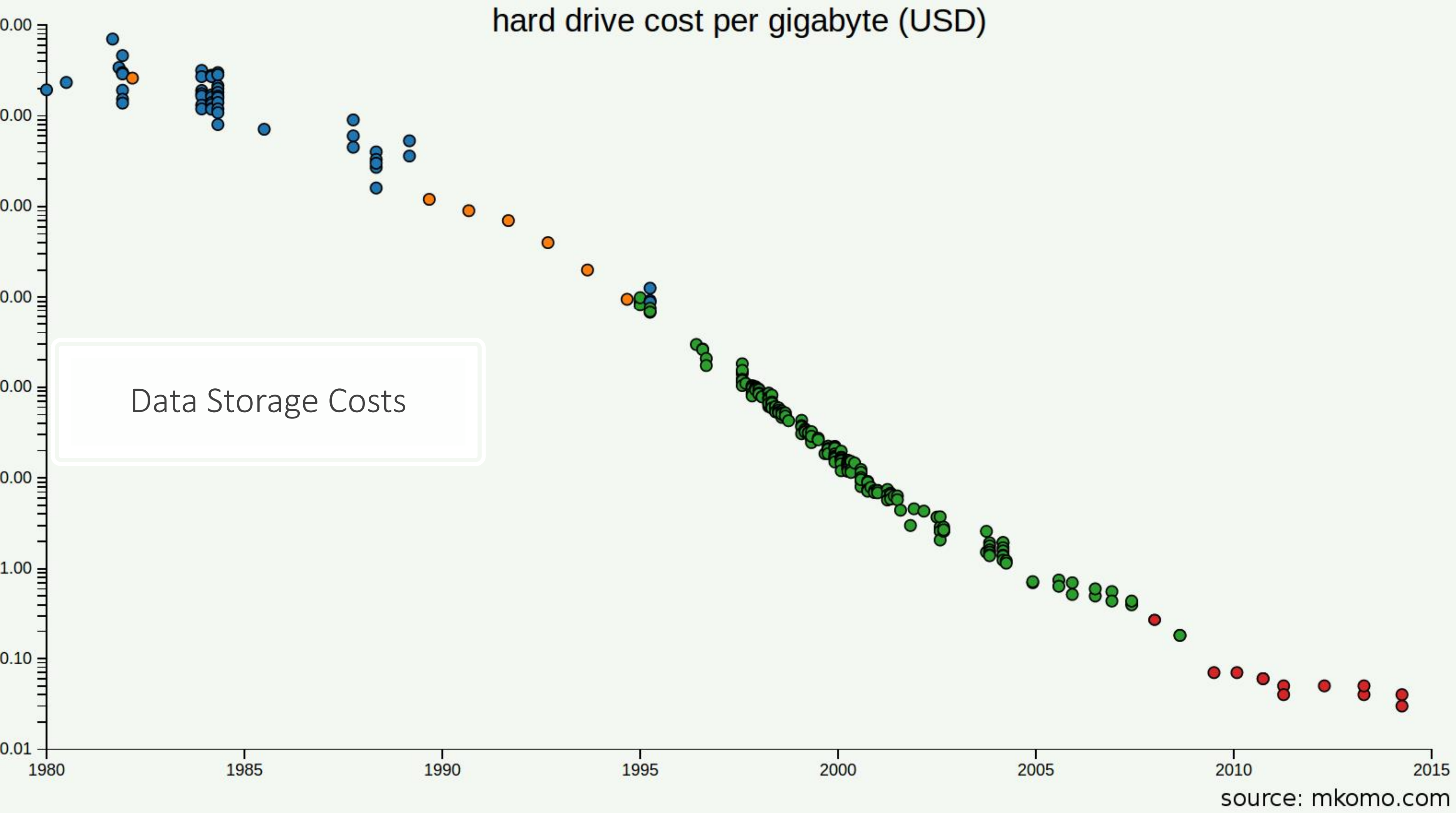


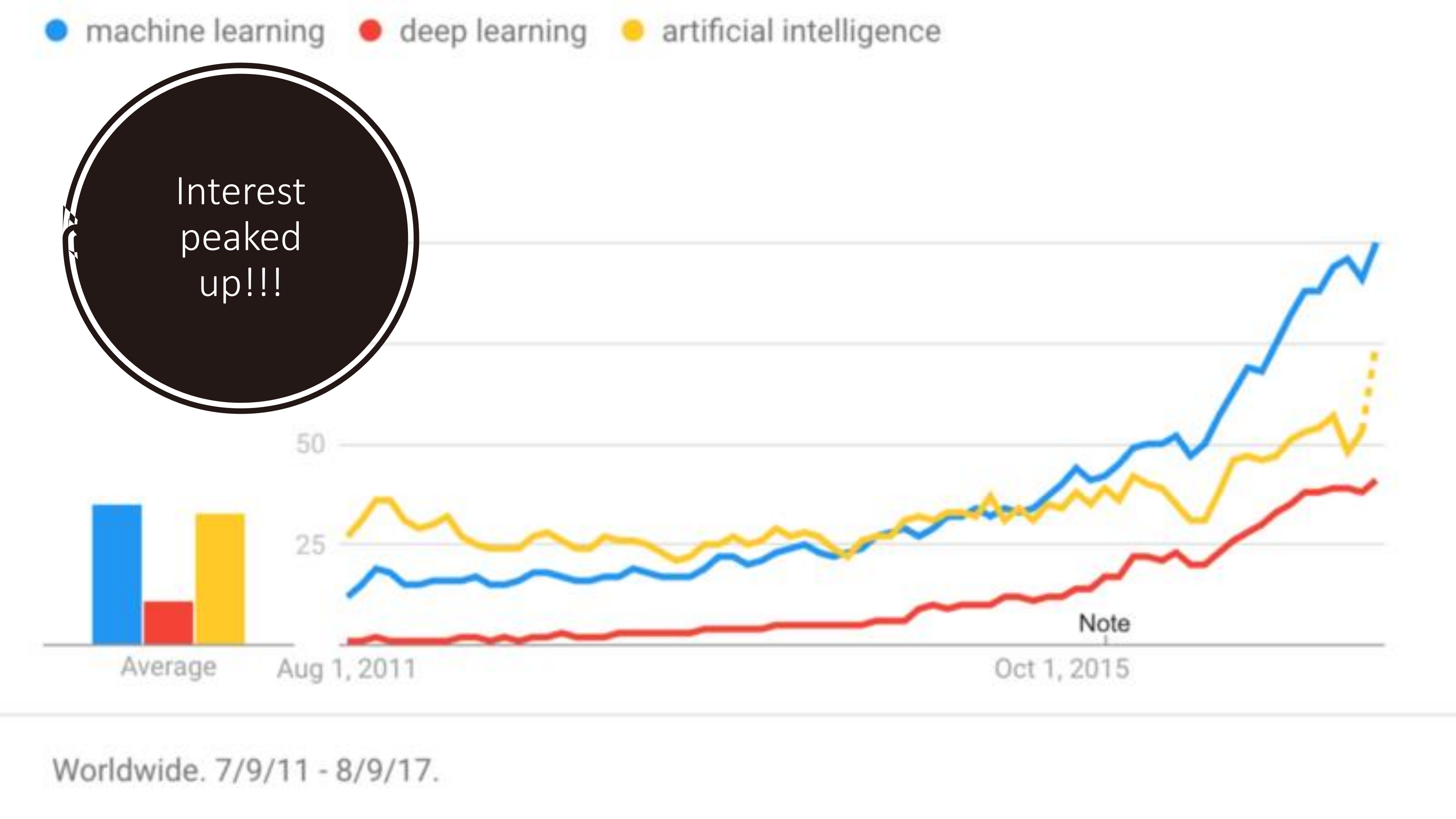
Source: ResearchGate

Devices connected to the internet



Data Storage Costs




machine learning
deep learning
artificial intelligence
Interest peaked up!!!
50
25
Average
Aug 1, 2011
Oct 1, 2015
Note
Worldwide. 7/9/11 - 8/9/17.

"DATA IS THE NEW OIL."
Coined in 2006 by Clive Humby, a British data commercialization entrepreneur this now famous phrase was embraced by the World Economic Forum in a 2011 report, which considered data to be an economic asset, like oil.
From the beginning of recorded time until 2003, we created 5 exabytes (5 billion gigabytes) of data.
In 2011 the same amount was created every two days.
By 2013, it's expected that the time will shrink to 10 minutes.
There are nearly as many bits of information in the digital universe as there are stars in our actual universe.
There are 133 million BLOGS on the web.
As of August 2012, there were just over
Just as a study of activity on Twitter gave residents, family members, and journalists advance warning of details about the devastating earthquake and tsunami in Japan, high-frequency traders, with the help of computer algorithms, use Big Data to follow trends and to act quickly
The Economist @TheEconomist · 2h
The world's most valuable resource is no longer oil, but data
amazon
UBER
Microsoft
Google
TESLA
from the current 65 milliseconds it takes for trading instructions to travel between New York City and London.
they save 5 milliseconds
The new cable takes a shallower, therefore shorter route.
UK
Every hour, we create enough Internet traffic to fill 7 billion DVDs.
Side by side, that's that's seven times the height of Everest.
millions of users
50% of 5-year-old kids in the U.S. are given access to a smartphone.
Information Graphics by Nigel Holmes

Let’s discuss

What is **machine learning**?

What is Learning?

1

"Learning denotes changes in a system that ... enable a system to do the same task ... more efficiently the next time." - Herbert Simon

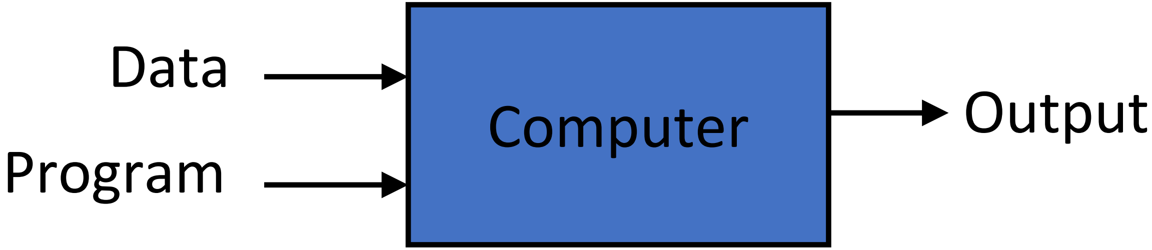
2

"Learning is constructing or modifying representations of what is being experienced." - Ryszard Michalski

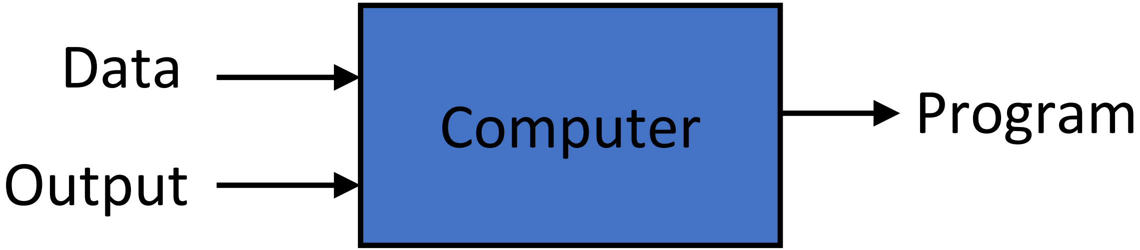
3

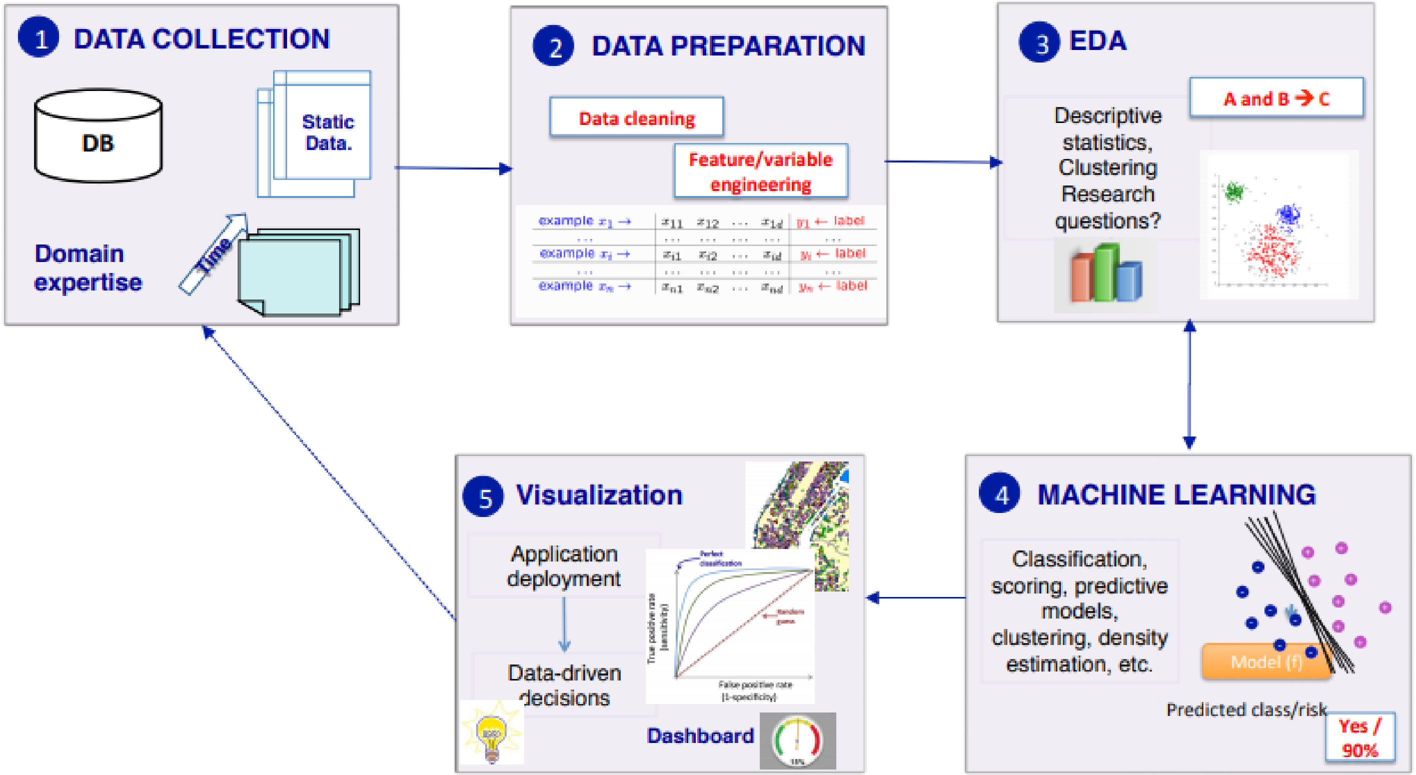
"Machine learning refers to a system capable of the autonomous acquisition and integration of knowledge."

Traditional Programming

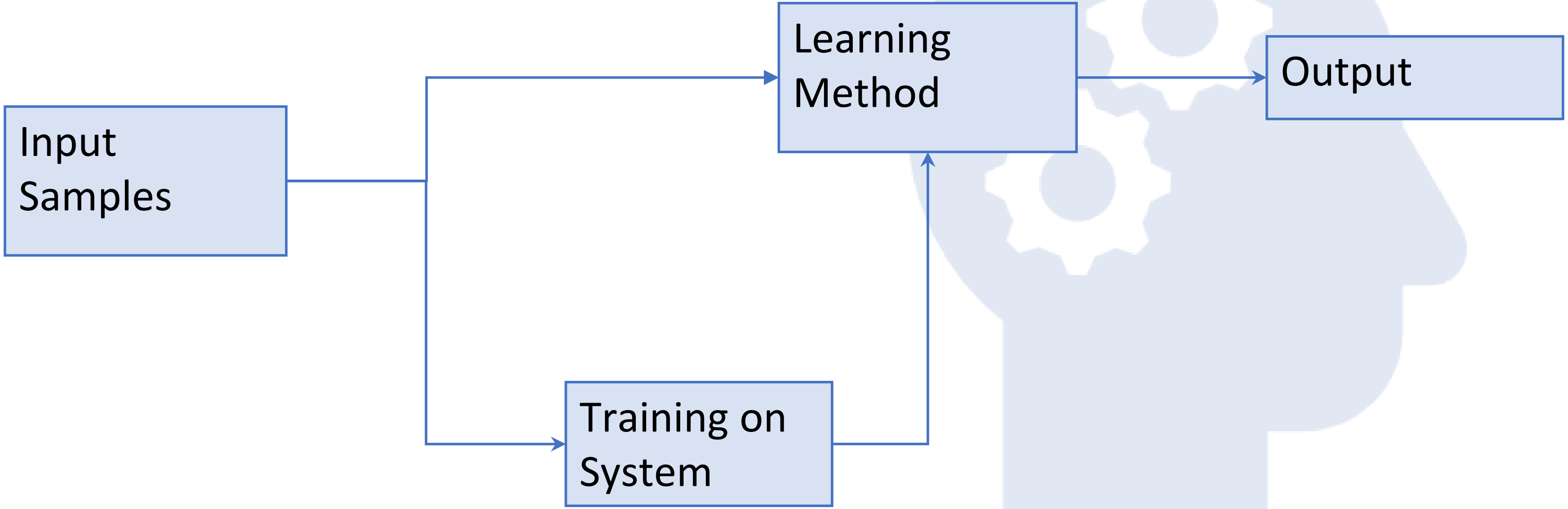


Machine Learning




1 DATA COLLECTION
DB
Static Data.
Domain expertise
Time
2 DATA PREPARATION
Data cleaning
Feature/variable engineering
example x_1 → x_{11} x_{12} ... x_{1d} y_1 ← label
example x_i → x_{i1} x_{i2} ... x_{id} y_i ← label
example x_n → x_{n1} x_{n2} ... x_{nd} y_n ← label
3 EDA
A and B → C
Descriptive statistics, Clustering Research questions?
4 MACHINE LEARNING
Classification, scoring, predictive models, clustering, density estimation, etc.
Model (f)
Predicted class/risk
Yes / 90%
5 Visualization
Application deployment
Data-driven decisions
Dashboard

Learning System Model



Why is machine learning required?

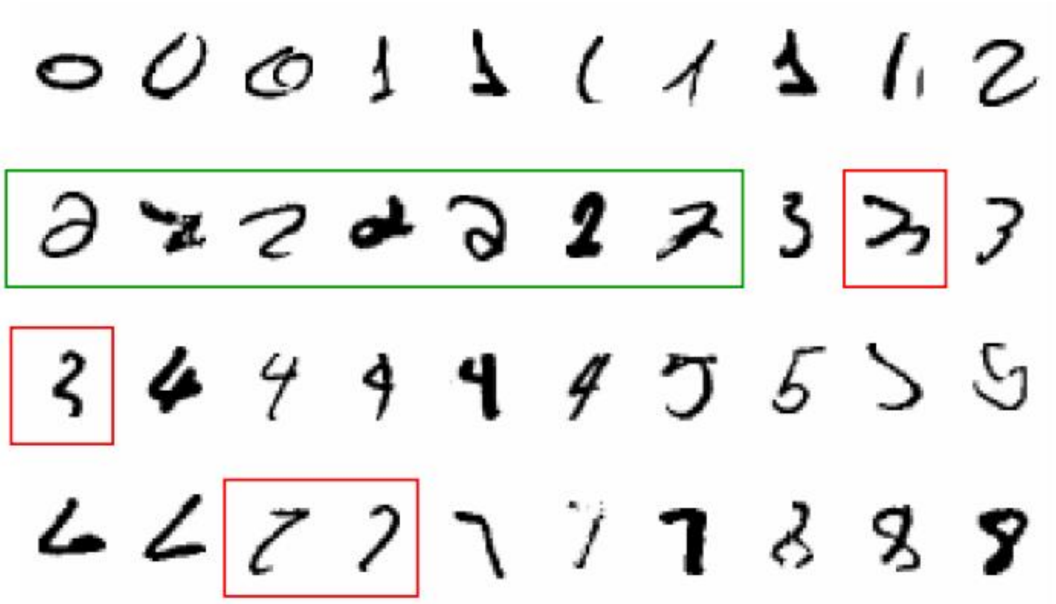
Lack of human experts

Black-box human expertise

Rapidly changing phenomenon

Need for customization and personalization

A classic example of a task that requires machine learning: It is very hard to say what makes a 2



Some examples that machine learning solves

Recognizing patterns:

- Facial identities or facial expressions
- Handwritten or spoken words
- Medical images

Generating patterns:

- Generating images or motion sequences

Recognizing anomalies:

- Unusual credit card transactions
- Unusual patterns of sensor readings in a nuclear power plant

Prediction:

- Future stock prices or currency exchange rates

3 vital things to define

Task: Recognizing hand-written words

Performance Metric: Percentage of words correctly classified

Experience: Database of human-labeled images of handwritten words

Types of Learning

Supervised (inductive) learning –

- Given: training data + desired outputs (labels)

Unsupervised learning –

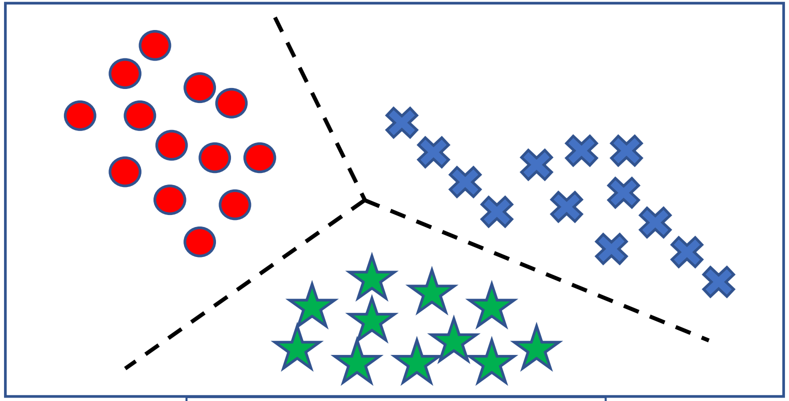
- Given: training data (without desired outputs)

Semi-supervised learning –

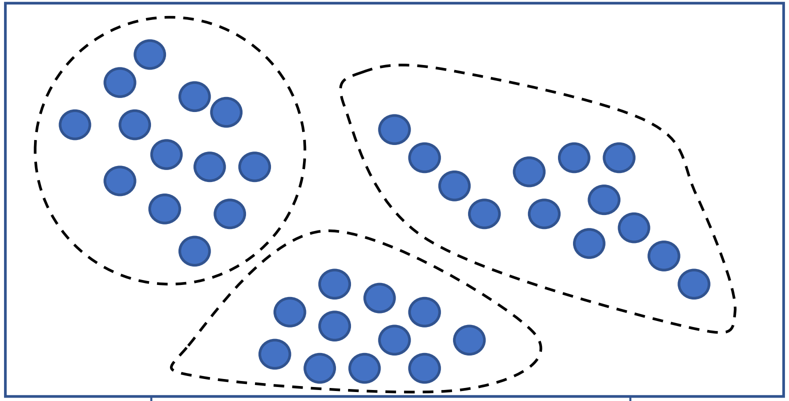
- Given: training data + a few desired outputs

Reinforcement learning –

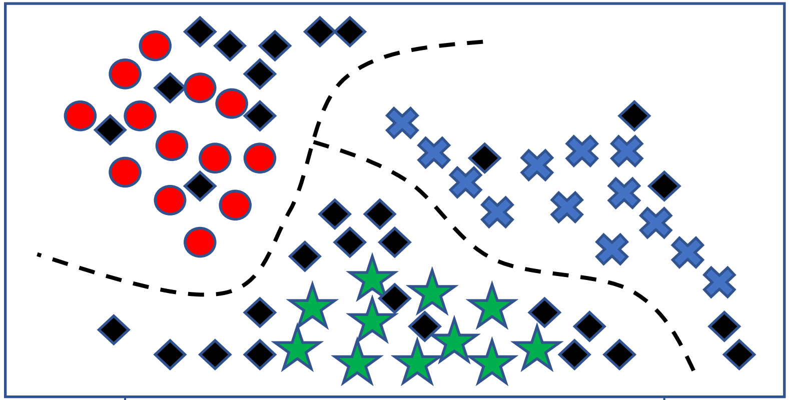
- Rewards from sequence of actions



Supervised learning

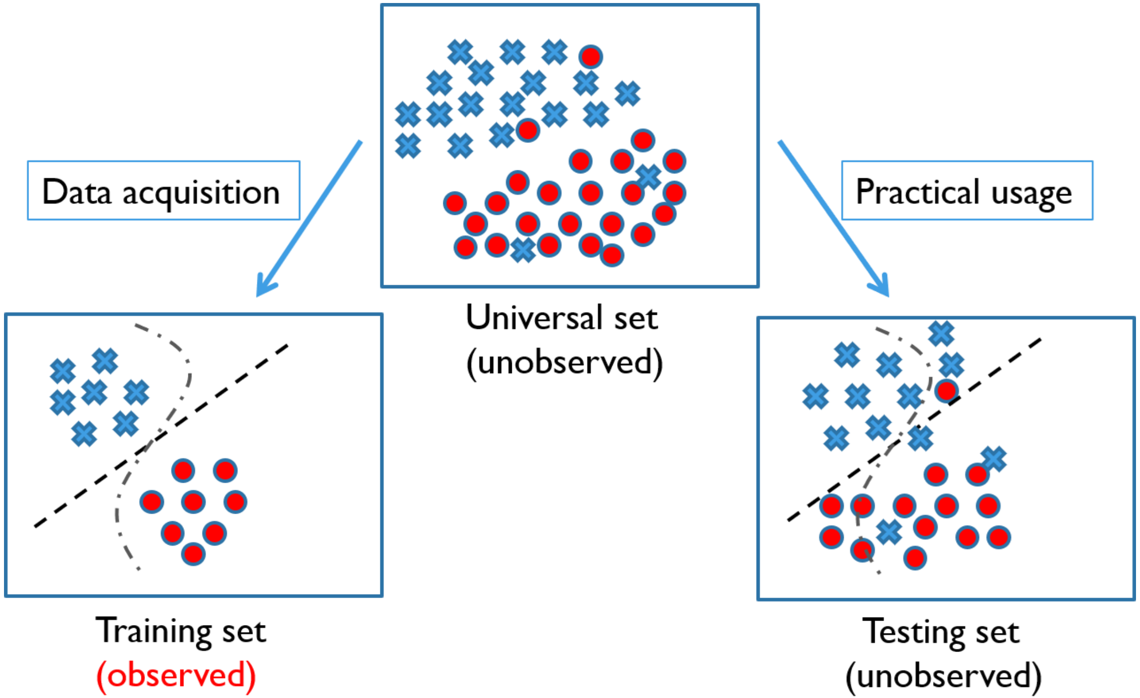


Unsupervised learning



Semi-supervised learning

Training and Test Sets



Unsupervised Learning

Data has only input values and without target results

Unsupervised Learning

- The data has no target attribute.
- We want to explore the data to find some intrinsic structures in them.

What is Clustering?

Clustering

Clustering is a technique for finding similarity groups in data, called **clusters**. I.e.,

- It groups data instances that are similar to (near) each other in one cluster and data instances that are very different (far away) from each other into different clusters.

What's a cluster?

Intuitive definition:

Grouping of data points that are close to each other

To make this computer friendly, need a mathematical definition of "close."

Closeness (most common definitions):

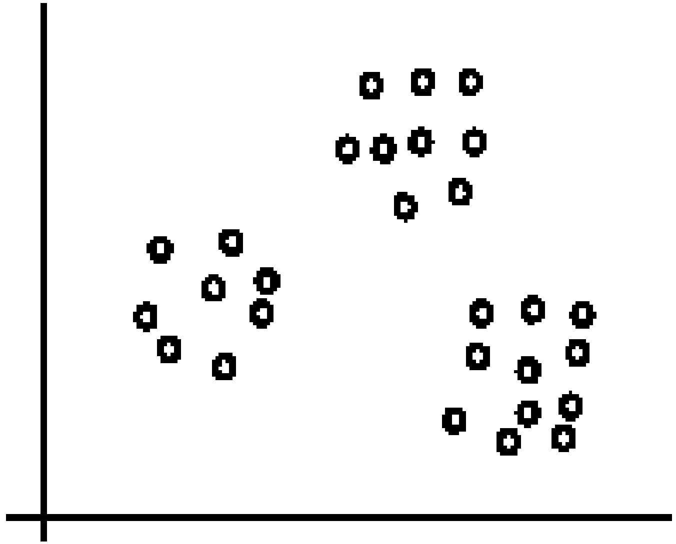
based on distance or density

Clustering as unsupervised learning

Unlabeled data $\xrightarrow{\text{algorithm}}$ Structured data

New data (unlabeled) $\xrightarrow{\text{assignment}}$ New data included in structure

Think of it like this – In layman figures



A Clustering Technique

K-Means Algorithm

K-means is a partitional clustering algorithm

The *k*-means algorithm partitions the given data into *k* clusters.

Each cluster has a cluster **center**, called **centroid**.

k is specified by the user

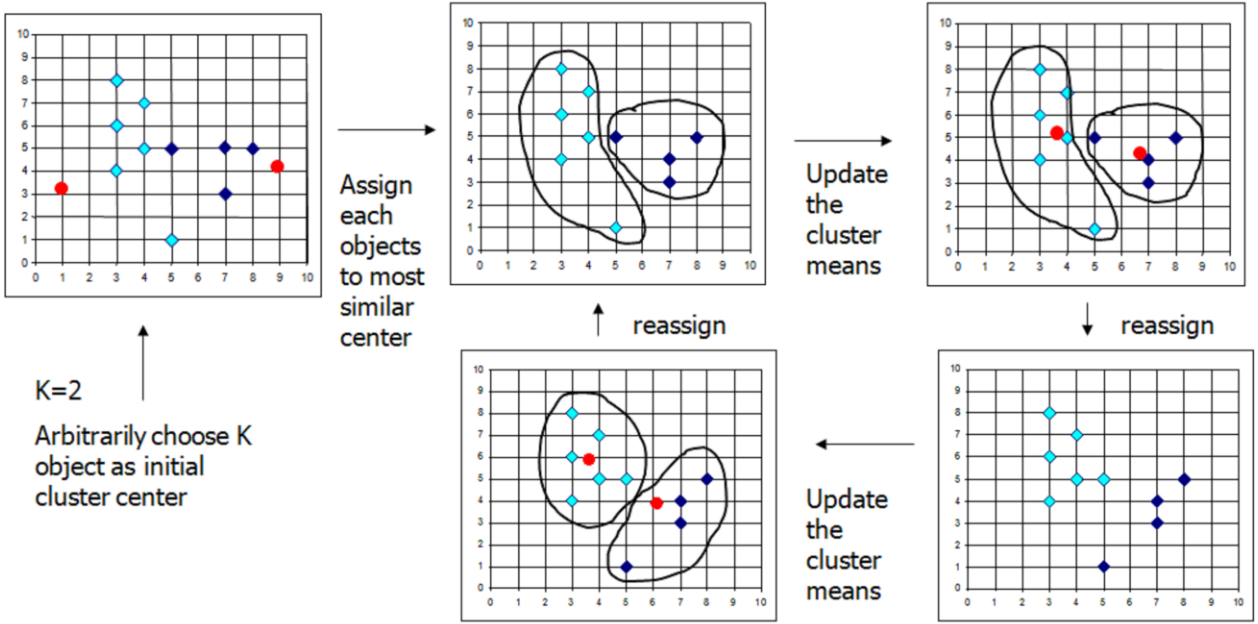
k-means clustering: the **algorithm**

- Choose *k* centroids
- Assign points to cluster based on nearest centroid
- Recompute centroids
- Repeat steps (2) and (3) until there is no more change to the centroids

k-means: simple example

k-means: simple example

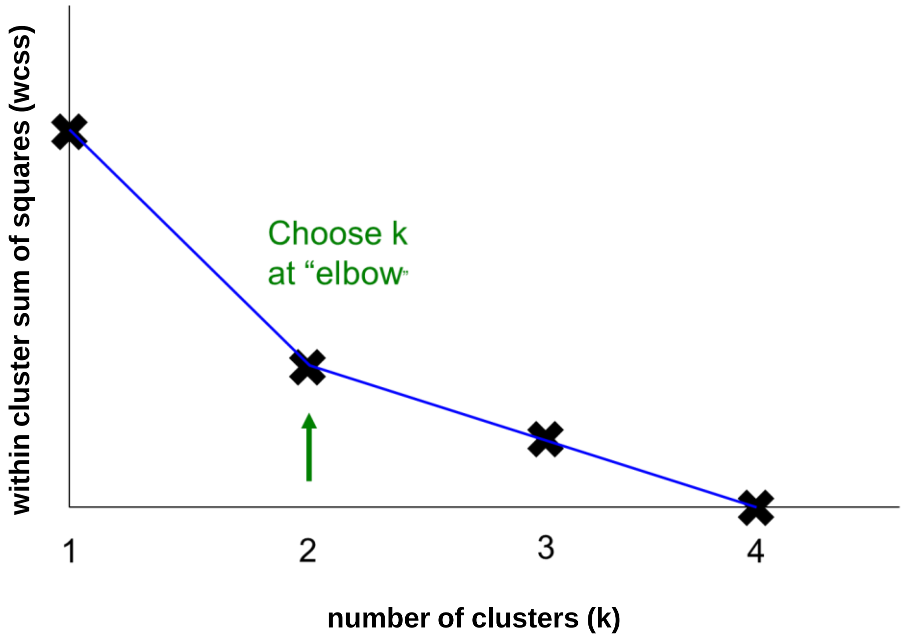
k-means: simple example


K=2
Arbitrarily choose K object as initial cluster center
Assign each objects to most similar center
Update the cluster means
reassign
reassign
Update the cluster means

k-means performance

good clustering → points close to cluster centroids

k-means performance



k-means: adding new data

Add	Add new data to nearest cluster
Treat	Treat clusters as labeled data
Use	Use this data to train a classifier
Apply	Apply classifier to new data

k-means: strengths and weaknesses

Strengths:

- Simple—one parameter (*k* clusters)
- Typically fast
- Easy to implement

Weaknesses:

- Optimal *k* is often not obvious
- Sensitive to outliers
- Scaling affects results

Clustering - Real life Examples

Example 1: groups people of similar sizes together to make “small”, “medium” and “large” T-Shirts.

Tailor-made for each person: too expensive	One-size-fits-all: does not fit all.

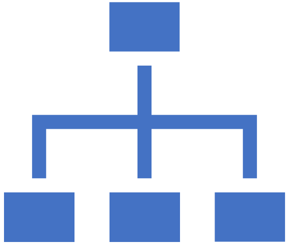
Example 2: In marketing, segment customers according to their similarities

To do targeted marketing.

Let’s dive straight to the Hands-on using Jupyter notebooks

Other clustering algorithms

Self Organizing Maps (SOM)

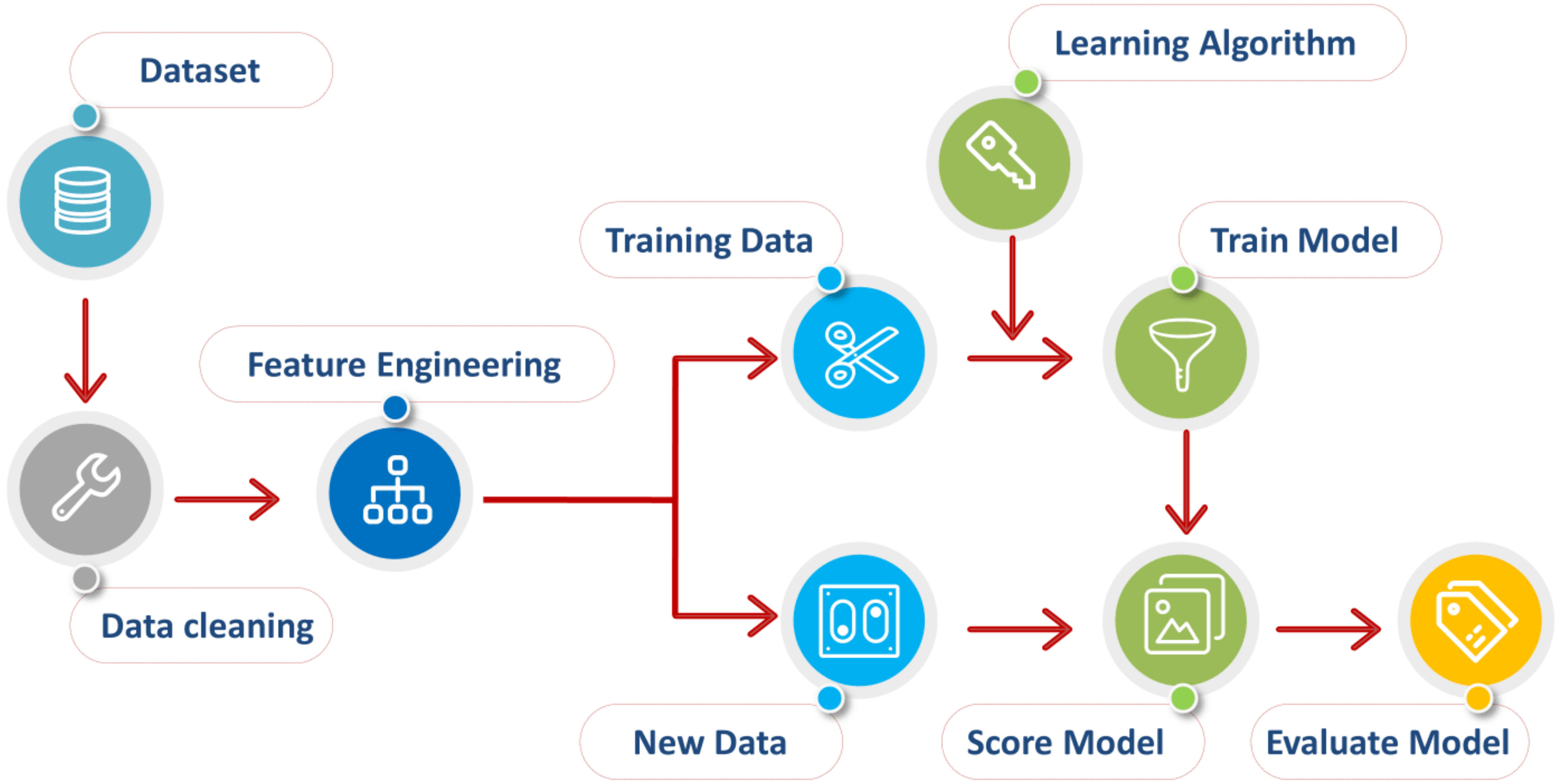


Agglomerative Hierarchical Clustering

Supervised Learning

Data includes both the input and the desired results.

General Machine Learning Process


Dataset
Data cleaning
Feature Engineering
Training Data
New Data
Learning Algorithm
Train Model
Score Model
Evaluate Model

Think of the following examples.

- An emergency room in a hospital measures 17 variables (e.g., blood pressure, age, etc) of newly admitted patients.
- **A decision is needed: whether to put a new patient in an intensive-care unit.**
- Due to the high cost of ICU, those patients who may survive less than a month are given higher priority.
- **Problem**: to predict high-risk patients and discriminate them from low-risk patients.

Another example..

- A credit card company receives lots of applications for new cards. Each application contains information about the applicant for the card,
  - age
  - Marital status
  - annual salary
  - location
  - outstanding debts
  - credit rating
  - Family information etc
- **Problem**: to decide whether an application should be approved or not approved.

Jargons to be aware of!

Model Inputs: Features, Attributes, Predictors, Inputs, Independent Variables, Dimensions, probably more.

Model Outputs (what we're trying to predict): Target, Response, Output, Dependent Variable, Labels

Row of Data (Inputs + Outputs): Observation, Datapoint, Record, Row

Labels: The values on the target variables in Supervised Learning

Target Data Types vs Algorithm Types

Supervised Learning

Continuous
- Regression

Categorical
- Classification

Agenda Part 2 – Supervised Machine Learning

1

Regression Type

- Properties
- Algorithms
- Performance Metrics

2

Classification Type

- Properties
- Algorithms
- Performance Metrics

Regression Properties

- Prediction of a *continuous* (numerical) output variable
- Can have real or discrete input variables
- Multiple input variables – Multivariate Regression problem
- Input variables ordered by time – Time Series Forecasting problem

Regression Algorithms

- Linear Regression
- K-Nearest Neighbors (KNN)
- Support Vector Machines (SVM)
- Decision Tree
- Random Forest
- Artificial Neural Network (ANN)

Regression Performance Metrics

- We can evaluate a regression algorithm performance by:
  - Mean Squared Error (MSE)
  - Root Mean Squared Error (RMSE)
  - Mean Absolute Error (MAE)
  - R-squared / Adjusted R-squared

Classification Properties

- Prediction of a *discreet* (categorical) output variable
- Can have real or discrete input variables
- Output with 2 classes – Binary Classification problem
- Output with more than 2 classes – Multi-class Classification problem
- When there is an unequal distribution of classes - Imbalanced Classification problem
  - Many real-world classification problems have imbalanced class distribution: Fraud detection, Spam detection, Churn prediction

Classification Algorithms

- Logistic Regression
- K-Nearest Neighbors (KNN)
- Naïve Bayes
- Support Vector Machines (SVM)
- Decision Tree
- Random Forest
- Artificial Neural Network (ANN)

Classification Performance Metrics

- We can evaluate a classification algorithm performance by
  - Accuracy
  - Confusion Matrix
  - Precision
  - Recall
  - F1-Score
  - Area under ROC curve (ROC AUC)

Linear Regression

Getting our line straight!

Introduction to Regression Analysis

- Regression analysis is used to:
  - Predict the value of a dependent variable based on the value of at least one independent variable
  - Explain the impact of changes in an independent variable on the dependent variable
- Dependent variable:

The variable we wish to predict or explain

- Independent variable:

The variable used to explain the dependent variable

Simple Linear Regression Model

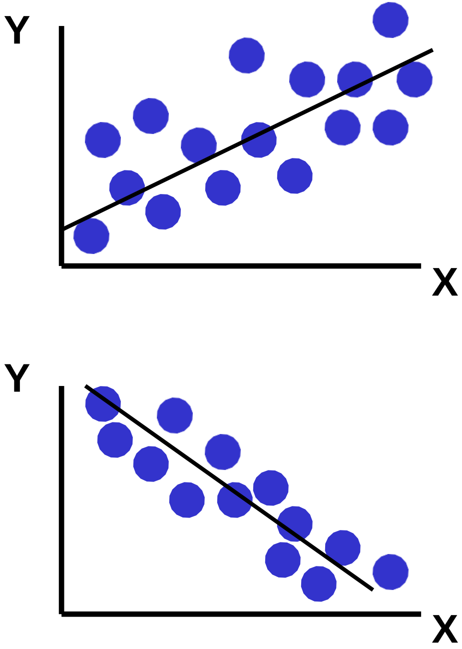
Only **one** independent variable, X

Relationship between X and Y is described by a linear function.

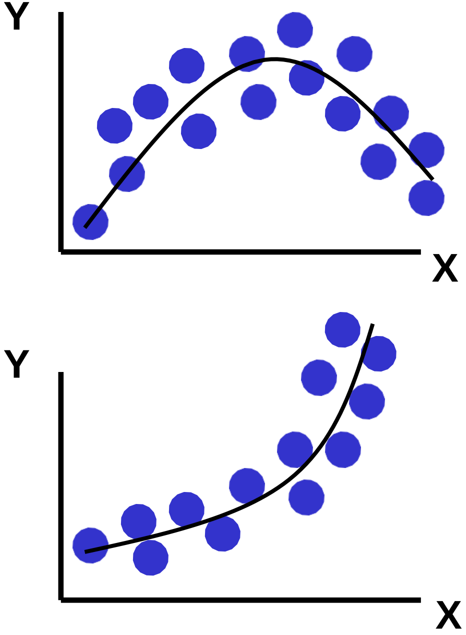
Changes in Y are assumed to be caused by changes in X

Types of Relationships

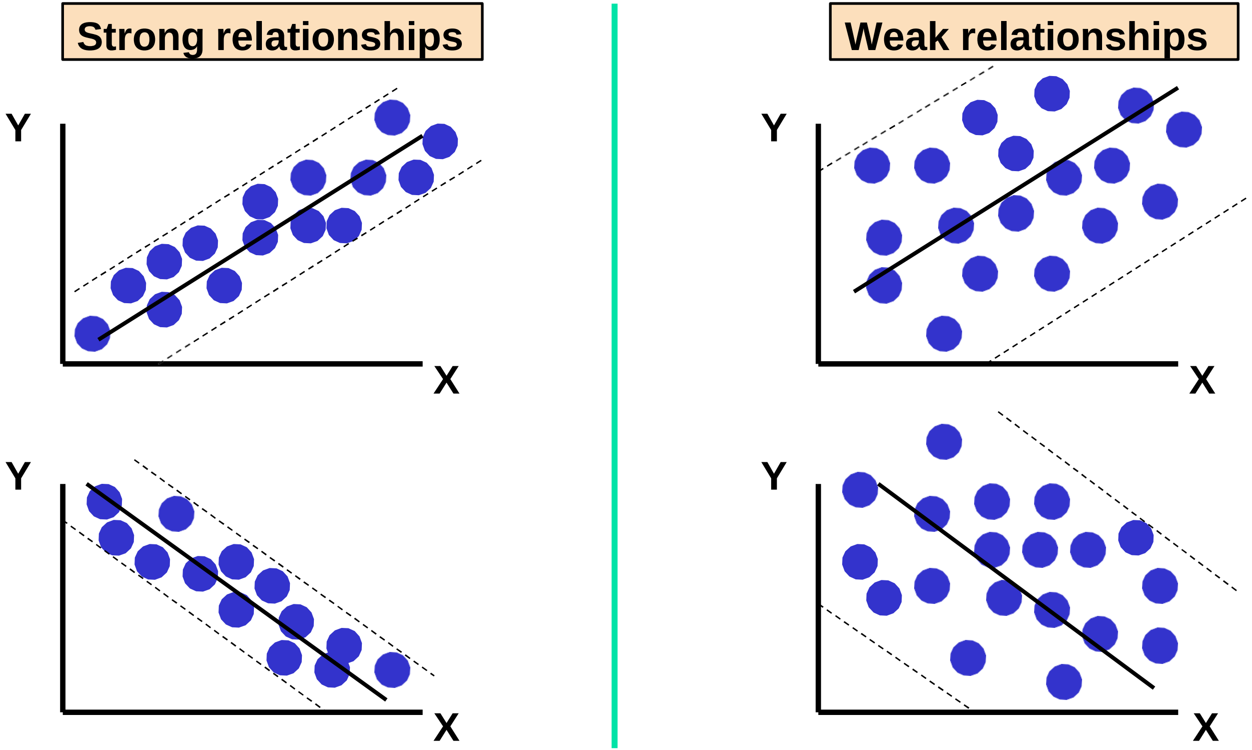
Linear relationships



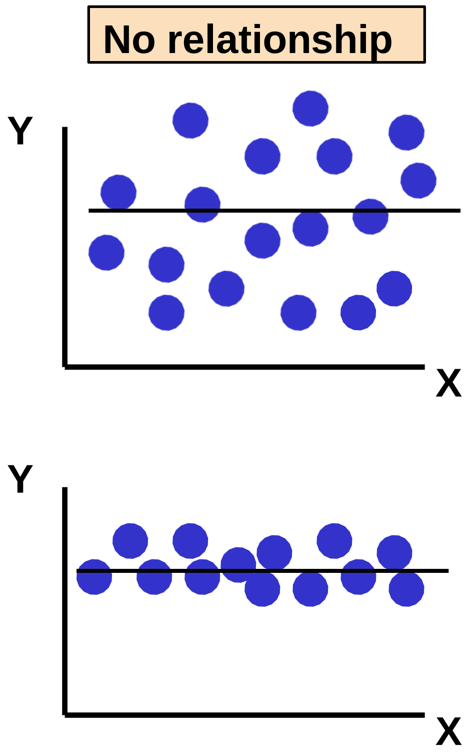
Curvilinear relationships



Types of Relationships



Types of Relationships



Simple Linear Regression Model

Dependent Variable

Population Y intercept

Population Slope Coefficient

Independent Variable

Random Error term

$$Y_i = \underbrace{b + MX_i}_{\text{Linear component}} + \underbrace{\varepsilon_i}_{\text{Random Error component}}$$

How do we determine if our Regression model is doing well or not?

Performance Metrics (Regression)

Mean Absolute Error - Sum of the absolute differences between predictions and actual values.

Mean Squared Error - Measures the average of the squares of the errors—that is, the average squared difference between the estimated values and what is estimated.

Logistic Regression

What is it and what is the algorithm?

What is the difference between Linear Regression & Logistic Regression?

Recap: What is linear regression?

- ***Linear regression*** quantifies the relationship between one or more *predictor variables* and one *outcome variable.*

- For example, linear regression can be used to quantify the relative impacts of age, gender, and diet (the predictor variables) on height (the outcome variable).

Recap: Example

Year	Sales (Million Euro)	Advertising (Million Euro)
1	651	23
2	762	26
3	856	30
4	1,063	34
5	1,190	43
6	1,298	48
7	1,421	52
8	1,440	57
9	1,518	58

Sales = 168 + 23 Advertising

What is logistic regression?

- Logistic regression is the appropriate regression analysis to conduct when the dependent variable is **binary**.
- Like all regression analyses, the logistic regression is a predictive analysis.
- Logistic regression is used to describe data and to explain the relationship between one dependent **binary variable** and **one or more nominal, ordinal, interval or ratio-level independent variables.**

Good to know!

Nominal

- Nominal scales are used for labeling variables, without any quantitative value. “Nominal” scales could simply be called “labels.”
  - E.g Male/Female, Red/Green/Yellow

Ordinal

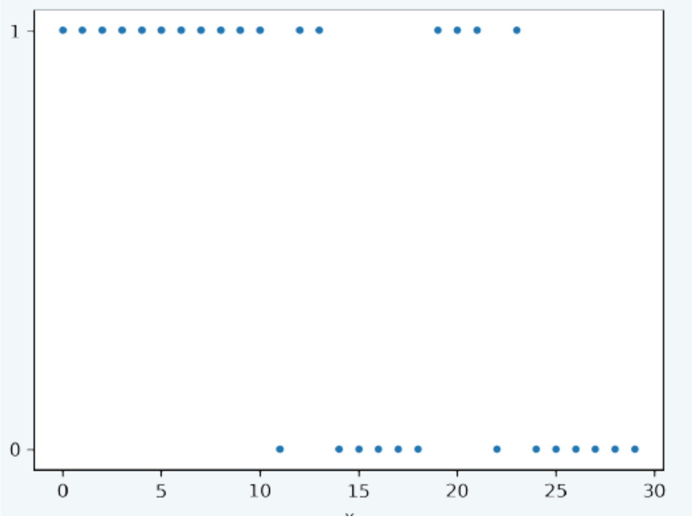
- With ordinal scales, the order of the values is what’s important and significant, but the differences between each one is not really known.
  - E.g Good, Very good, Excellent, Fantastic – 1#, 2#, 3#, 4#

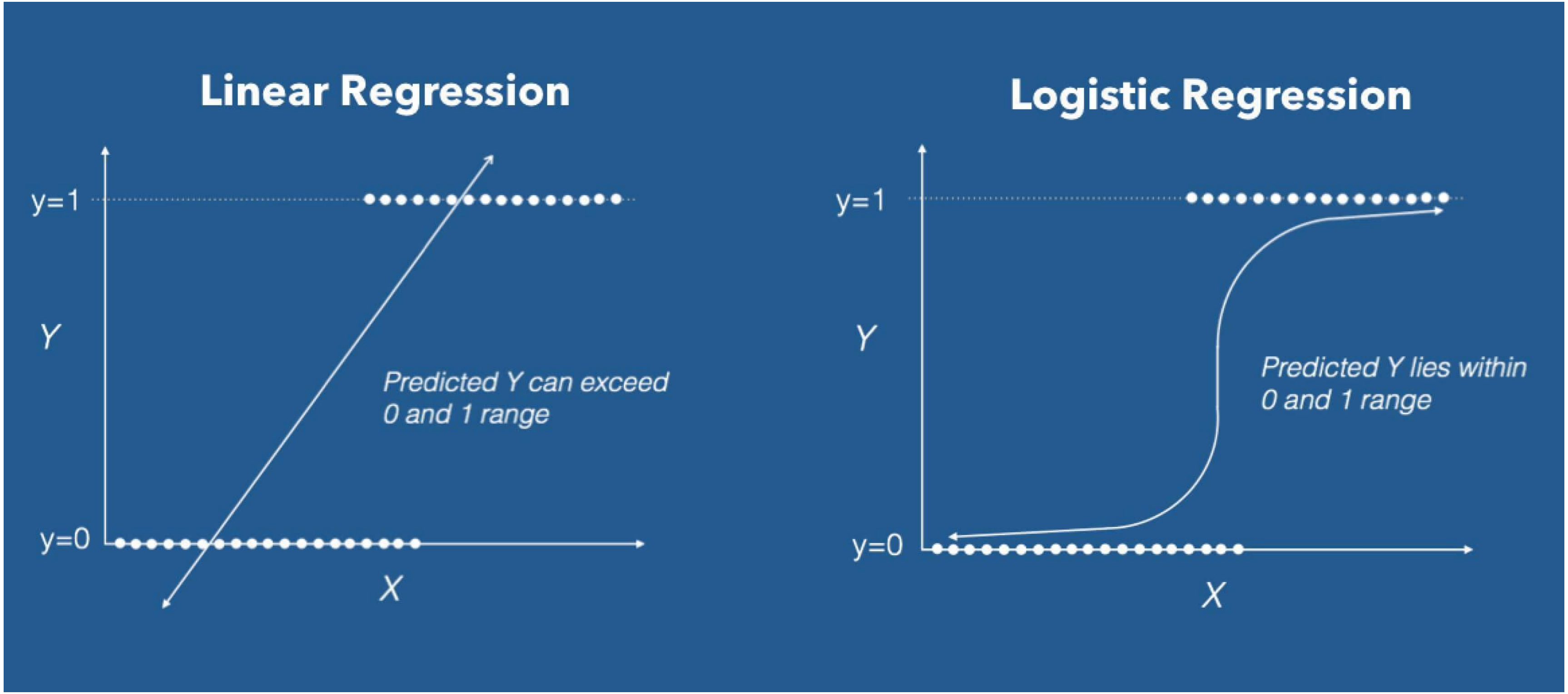
Interval

- Interval scales are numeric scales in which we know both the order and the exact differences between the values.
  - E.g Temp Celsius - because the difference between each value is the same.

Example – Log Reg – Scoring Goals!

- If we are kicking our soccer ball from a variety of distances.
- The results are going to be only Goal or no Goal.
- Our Standard Linear Regression will not work in this scenario!





Linear Regression
y=1
Y
y=0
X
Predicted Y can exceed
0 and 1 range
Logistic Regression
y=1
Y
y=0
X
Predicted Y lies within
0 and 1 range

The Sigmoid function

- We apply sigmoid function on the linear regression equation.
- By doing so, we will push our straight line to be a S shape or Sigmoid Curve.

$$y = \frac{1}{1 + e^{-x}}.$$

Model Evaluation

Model Evaluation is an integral part of the model development process.

It helps to find the best model that represents our data and how well the chosen model will work in the future.

Performance Metrics (Classification)

Confusion Matrix

Accuracy

Precision and Recall

How do you evaluate classifiers?

Accuracy!

$$Accuracy = \frac{\text{Number of correct classifications}}{\text{Total number of test cases}}$$

Confusion Matrix

It is a performance measurement for machine learning classification problem where output can be two or more classes.

It is a table with 4 different combinations of predicted and actual values.

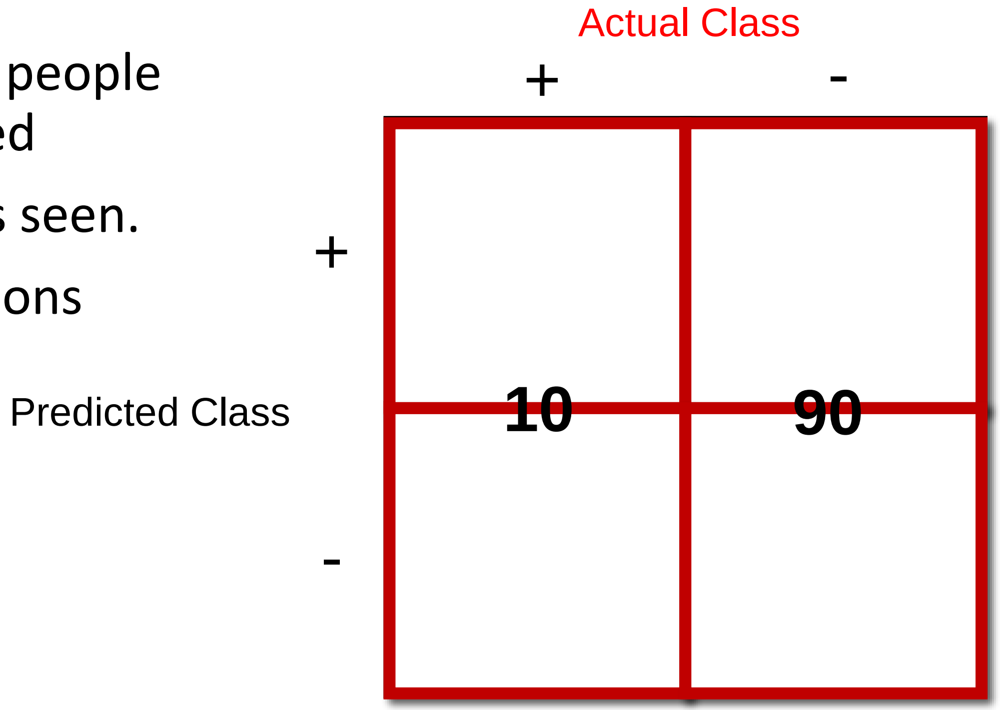
Let’s take an example of Confusion matrix

- Assuming there are 100 people which are to be predicted

		Actual Class	
		+	-
Predicted Class	+		
	-		

Let’s take an example of Confusion matrix

- Assuming there are 100 people which are to be predicted
- The actual classes are as seen.
- Now we get our predictions from our model.



Let’s take an example of Confusion matrix

- Assuming there are 100 people which are to be predicted
- Now we get our predictions from our model.

		Actual Class	
		+	-
Predicted Class	+	10	7
	-		83

Let’s take an example of Confusion matrix

- Assuming there are 100 people which are to be predicted
- Now we get our predictions from our model.

		Actual Class	
		+	-
Predicted Class	+	8	7
	-	2	83

Let's take an example of Confusion matrix

- Assuming there are 100 people which are to be predicted
- Now we get our predictions from our model.

		Actual Class	
		+	-
Predicted Class	+	**True +**	**False +**
	-	**False -**	**True -**

So how can we use the metrics?

Say we have 2 confusion matrix from 2 models

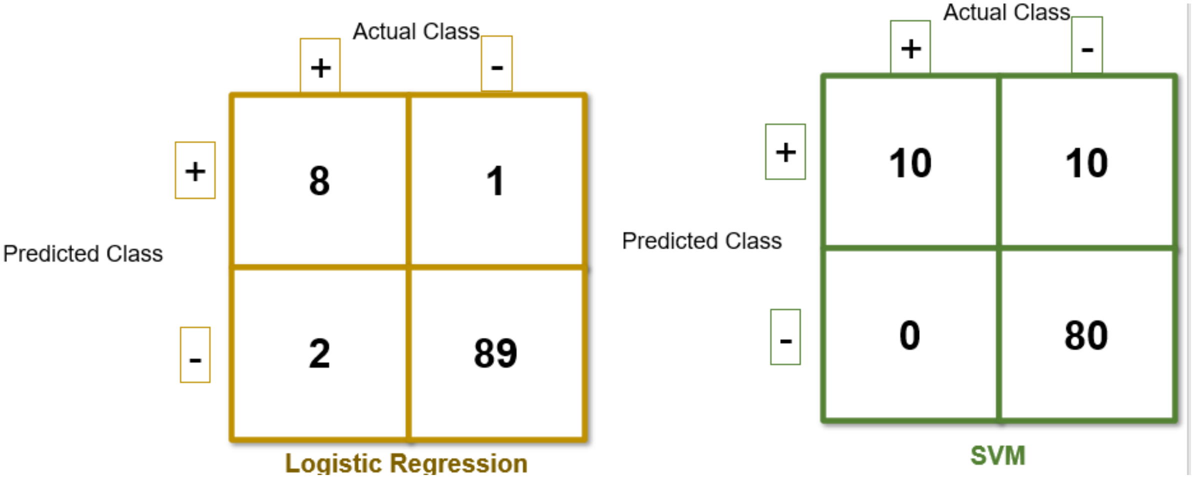
Predicted Class \ Actual Class	+	-
+	**8**	**1**
-	**2**	**89**

Logistic Regression

Predicted Class \ Actual Class	+	-
+	**10**	**10**
-	**0**	**80**

SVM

We can compare them!



Accuracy: (TP+TN)/(TP+TN+FP+FN)	97%	90%
Precision: TP/(TP+FP)	89%	50%
Recall: TP/(TP+FN)	80%	100%

	Predicted class POSITIVE (spam)	Predicted class NEGATIVE (normal)	
Actual class POSITIVE (spam)	TRUE POSITIVE (TP) **320**	FALSE NEGATIVE (FN) **43**	*Recall* $= \frac{TP}{TP+FN} = \frac{320}{320+43} = 0.882$
Actual class NEGATIVE (normal)	FALSE POSITIVE (FP) **20**	TRUE NEGATIVE (TN) **538**	
	Precision $= \frac{TP}{TP+FP} = \frac{320}{320+20} = 0.941$		

Precision and Recall

Precision attempts to answer the following question:

What proportion of positive identifications was correct?

Recall attempts to answer the following question:

What proportion of actual positives was identified correctly?

Let’s try to answer the following questions!

What of the following is not a type of machine learning process?

- Unsupervised Learning
- Semi-supervised Learning
- Supervised Learning
- Pro-supervised Learning

A Self Organizing Map (SOM) is an example of which type of learning algorithm?

- Unsupervised Learning
- Supervised Learning

Imagine, you are solving a classification problems with highly imbalanced class.

The majority class is observed 99% of times in the training data. Which of the following is a suitable metric to look at?

- Accuracy
- Precision
- Mean Absolute Error
- None of the above

A feature F can take certain value: A, B, C, D, E, & F and represents grade of students from a college.

Which of the following statement is true?

- Feature F is an example of nominal variable
- Feature F is an example of ordinal variable
- Both the above
- None of the above

THANK YOU!

Any questions?