Introduction to Data Science

done by Olesia Ved
Paris Dauphine University, INSA Lyon School of Engineering

FINANCE
FiNANCE
and
42
20
70
200
100
350
400

What is the DATA?

In the digital world:
Data is electronic information stored, processed, and transmitted in various formats.

Examples include website content, social media posts, photos, videos etc

DATA...
DATA...EVERYWHERE...
makeameme.org

Data Size in the world

How much data the average person stores ?

1 film: 2GB

600 photos: 1 GB

Data Size in the world

The average person stores : 500GB

Data Size in the world

The average person stores : 500GB

Seems not so much

Game of Go

But **Alpha Go (AI that plays Go)** has been trained only **with 44GB**

Data Size in the world

All US academic research libraries a **2 petabytes**

It is like **4000 persons’ personal storage**

Data Size in the world

Google (Google Doc, Google Search, Youtube etc) is around (very approximately) **10,000 Petabytes**

It is like **5000 US academic research libraries** in the world

Data Size in the world

it is estimated that the digital universe was approximately **44 zettabytes** in 2020

Or **4400 of Googles**

11 zeros!!!

Or **10^11** of **personal cloud storage’s**

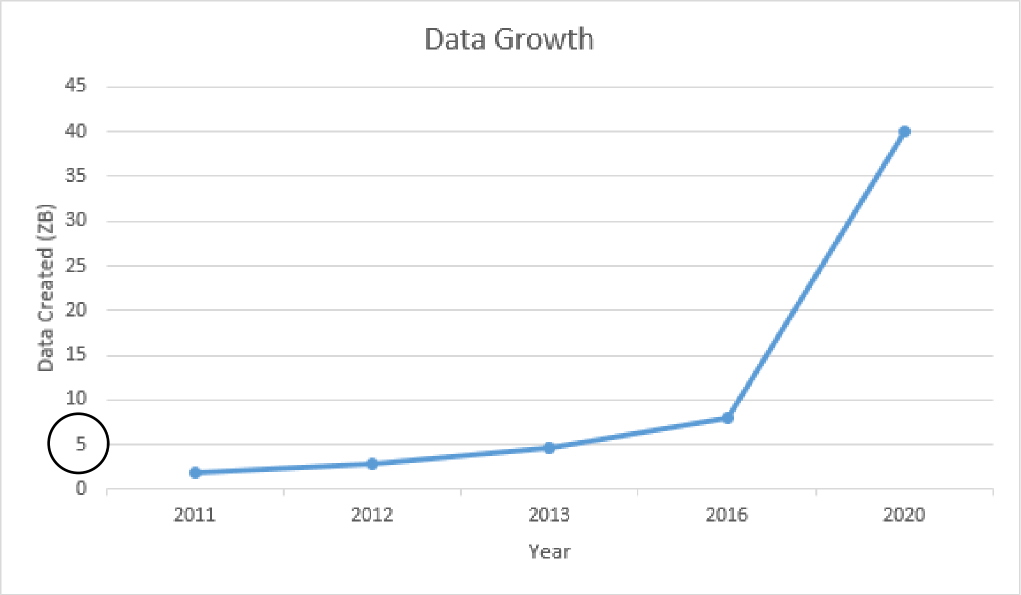
Question for you:

If in 2020 we had 44 zettabytes of data

How much data was back in 2013?

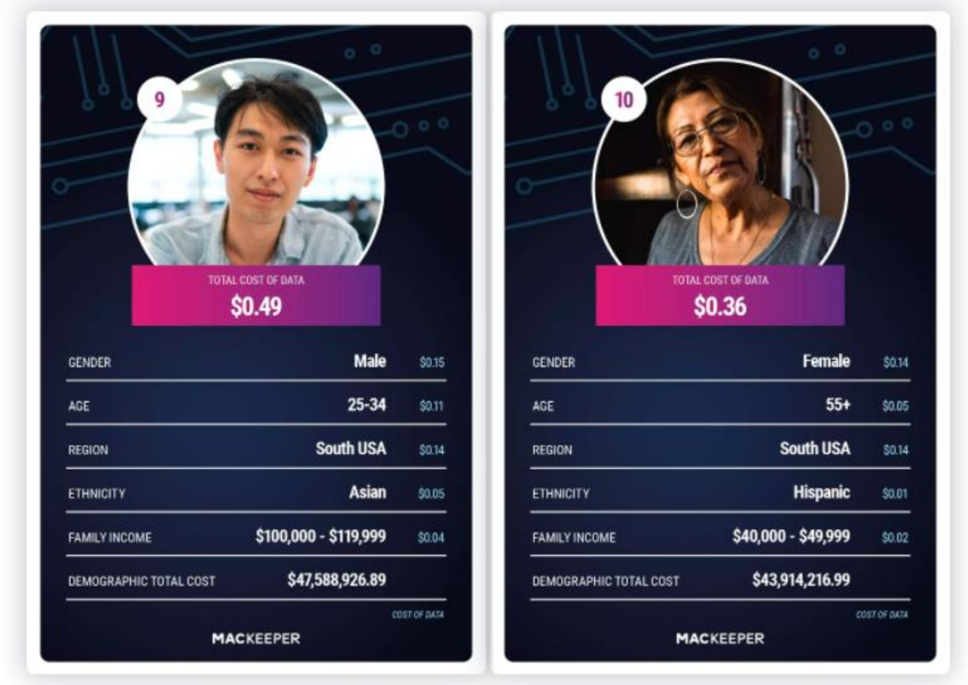
Rapid Data Growth

5 zettabytes in 2013!!!!!



Your personal Data is a big active

- Ads online
- Stories placement in Instagram

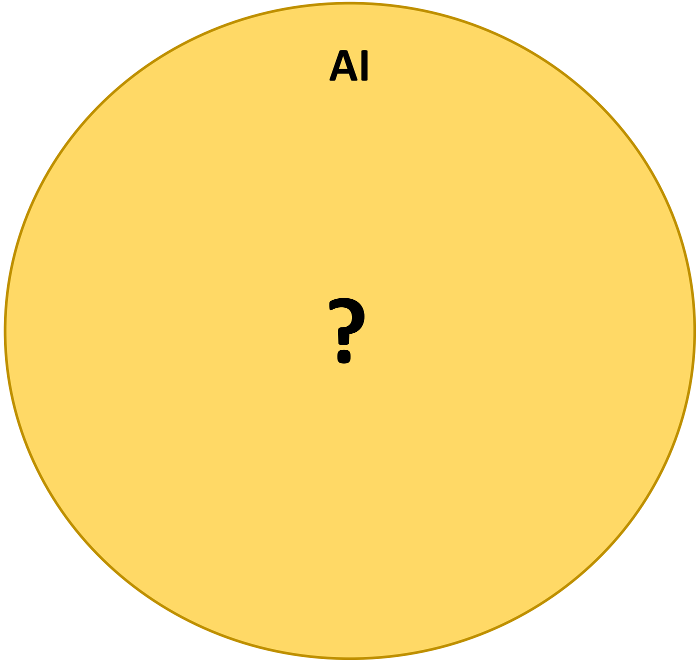


GENDER	Male	$0.15
AGE	25-34	$0.11
REGION	South USA	$0.14
ETHNICITY	Asian	$0.05
FAMILY INCOME	$100,000 - $119,999	$0.04
DEMOGRAPHIC TOTAL COST	$47,588,926.89	

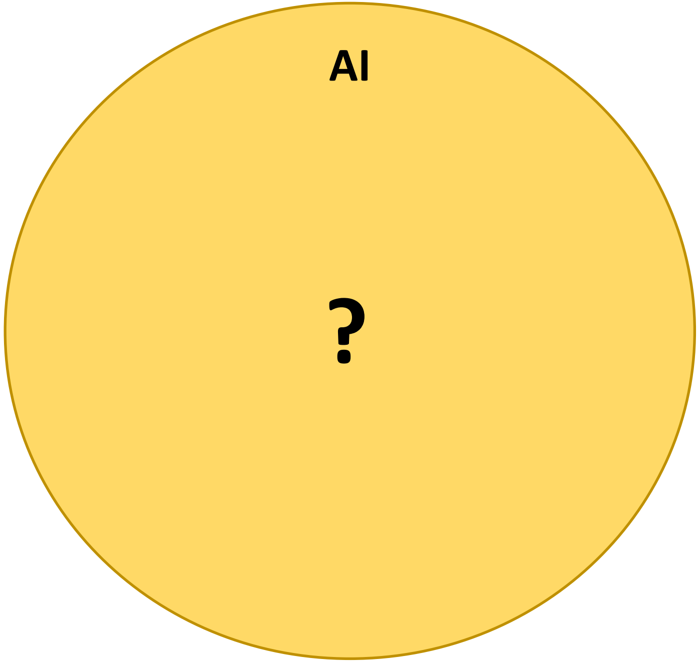
GENDER	Female	$0.14
AGE	55+	$0.05
REGION	South USA	$0.14
ETHNICITY	Hispanic	$0.01
FAMILY INCOME	$40,000 - $49,999	$0.02
DEMOGRAPHIC TOTAL COST	$43,914,216.99	

Multidisciplinary field?

Artificial Intelligence

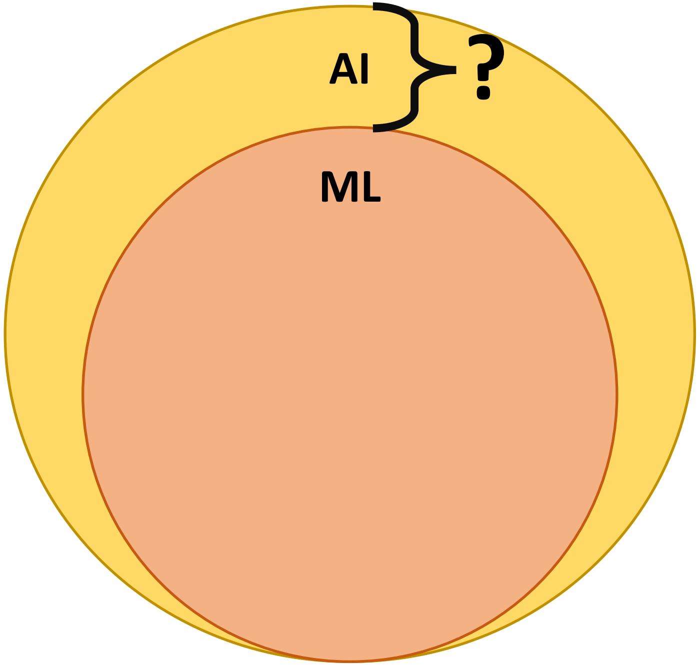


Artificial Intelligence

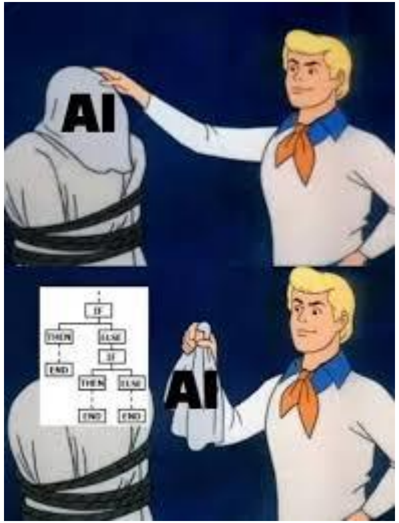


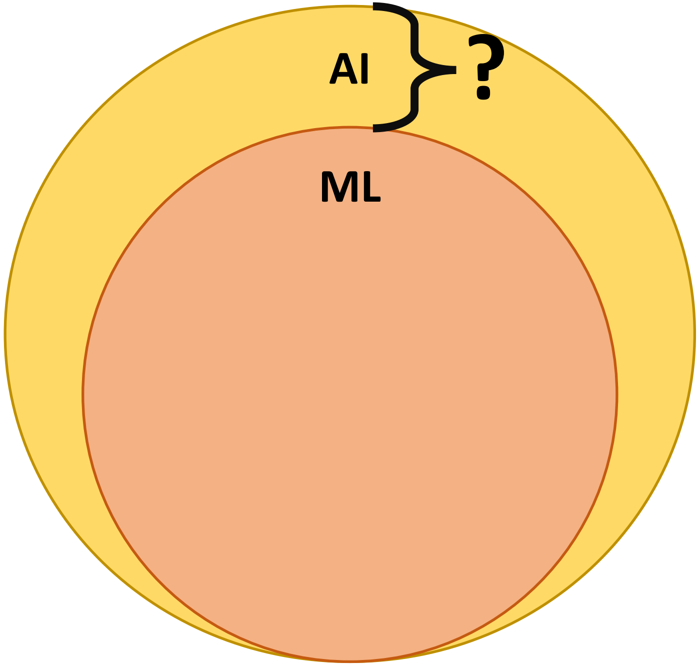
Algorithms performing
human-like decision making

AI & Machine Learning



AI & Machine Learning

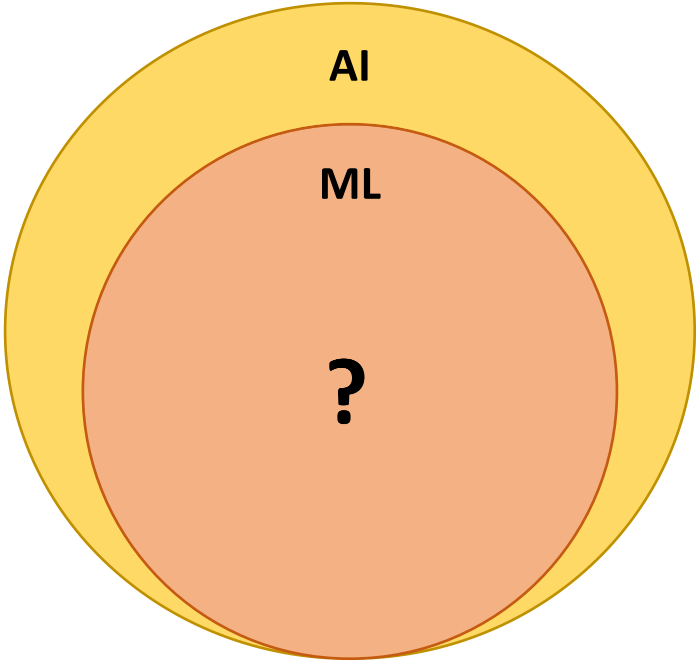




- Rule-based chatbots
- Visa type requirements on governmental web-site

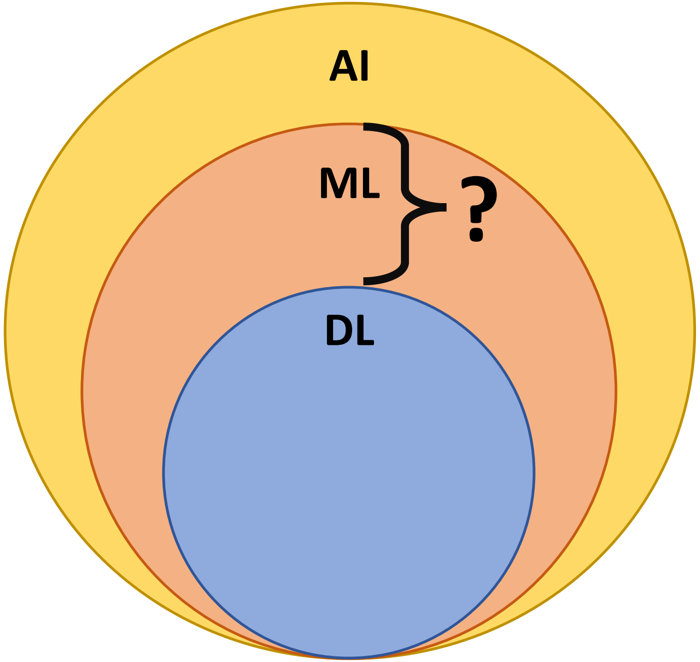
etc

Machine Learning



- Algorithms able to learn and adapt without following explicit instructions

ML & Deep Learning



Algorithms based on statistical knowledge:

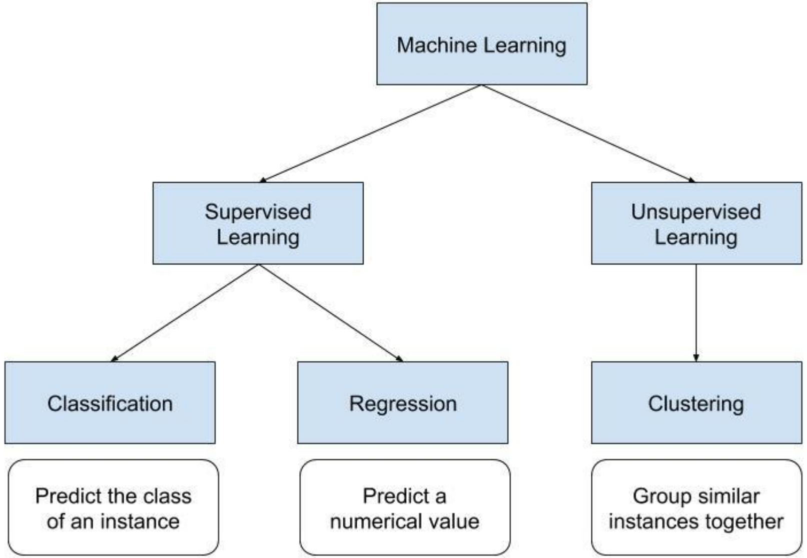
- Clustering
- Regressions

Machine Learning

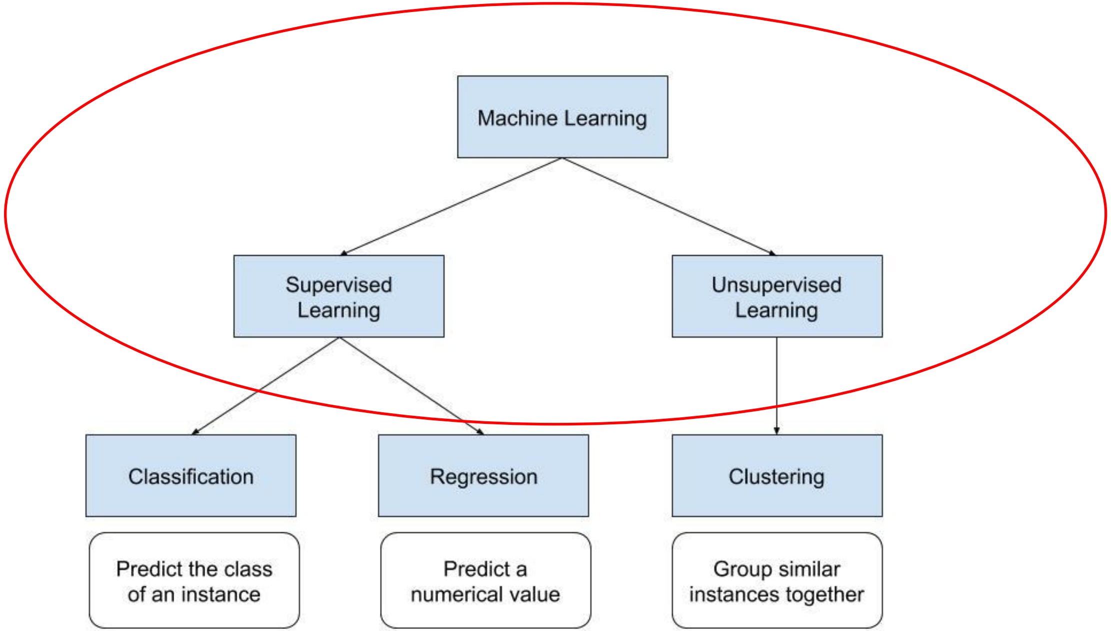
Further reading:

Here we are missing notions **of 2 types** that are less present in the market but very important:

- Semi-supervised
- Reinforcement Learning (e.g. used to teach robots to walk)



Supervised vs Unsupervised



Machine Learning

Supervised Learning

Unsupervised Learning

Classification

Regression

Clustering

Predict the class of an instance

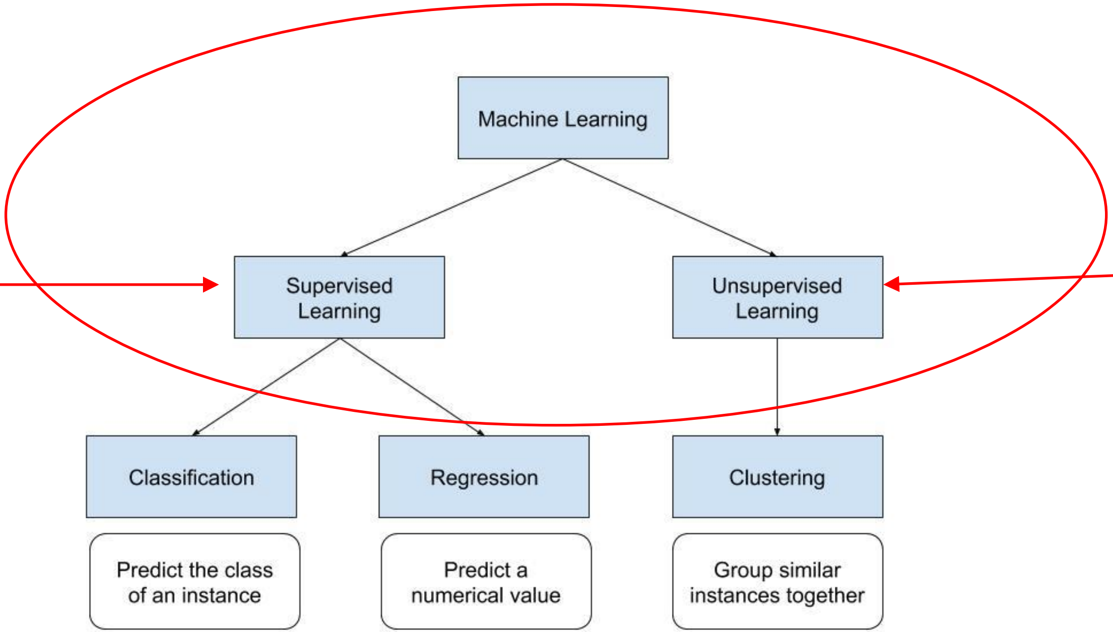
Predict a numerical value

Group similar instances together

Supervised vs Unsupervised

All datapoints are labeled

(Continuous variable for regression, Class for Classification)



Machine Learning

Supervised Learning

Unsupervised Learning

Classification

Regression

Clustering

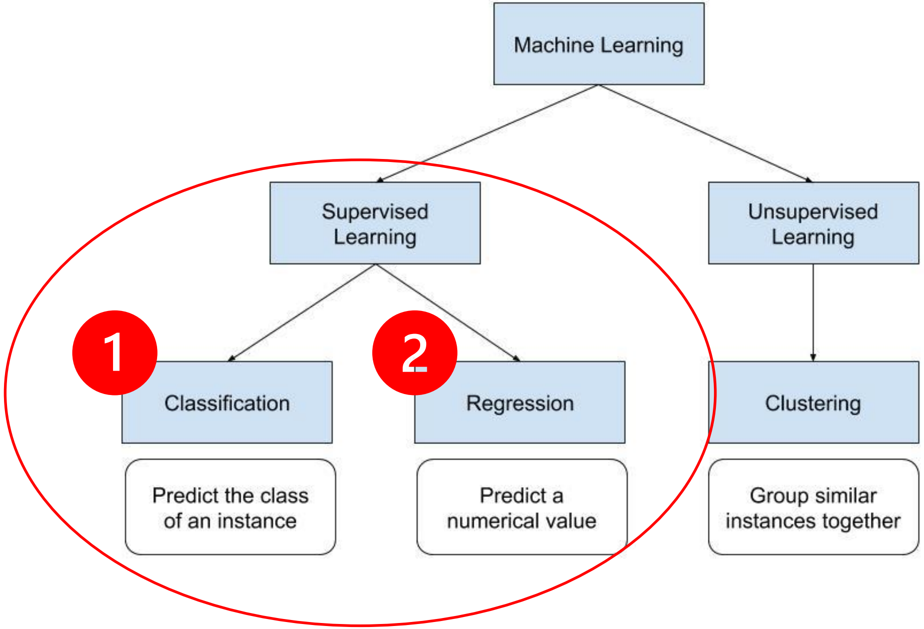
Predict the class of an instance

Predict a numerical value

Group similar instances together

None of datapoints is labeled

Supervised Learning



The categories = classes or labels
The training examples = pairs of input data and corresponding labels.

Classification

The goal -> to predict **which category** a new input data point belongs to, based on labeled examples from a training set

Classification

The categories = classes or labels
The training examples = pairs of input data and corresponding labels.

The goal -> to predict **which category** a new input data point belongs to, based on labeled examples from a training set

Step1: have a limited number of know categories. Example: dog breed

Chihuahua

Poodle

Dalmatian

The categories = classes or labels
The training examples = pairs of input data and corresponding labels.

Classification

The goal -> to predict **which category** a new input data point belongs to, based on labeled examples from a training set

Step2: have an information about characteristics on which the category depends

1) Dog size

2) Fur length (Animal’s hair)

Classification

The categories = classes or labels
The training examples = pairs of input data and corresponding labels.

The goal -> to predict **which category** a new input data point belongs to, based on labeled examples from a training set

Step3: Training Dataset

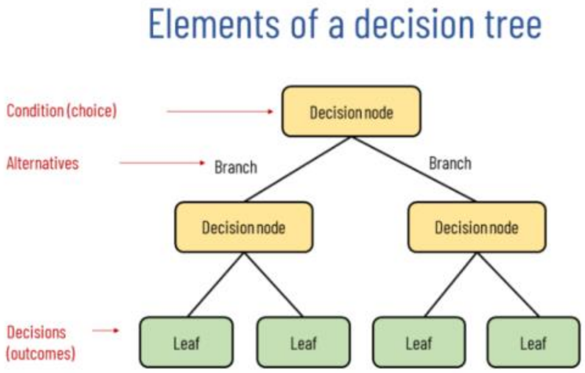
size	fur	class
medium	long	Poodle
big	long	Poodle
small	short	Chihuahua
big	short	Dalmatian

Now you have a classification task!

2 most common algorithms

- Decision Tree
- K nearest neighbours

Decision Tree



It splits data based on features to create decision rules and reaches conclusions at leaf nodes.

Decision node:

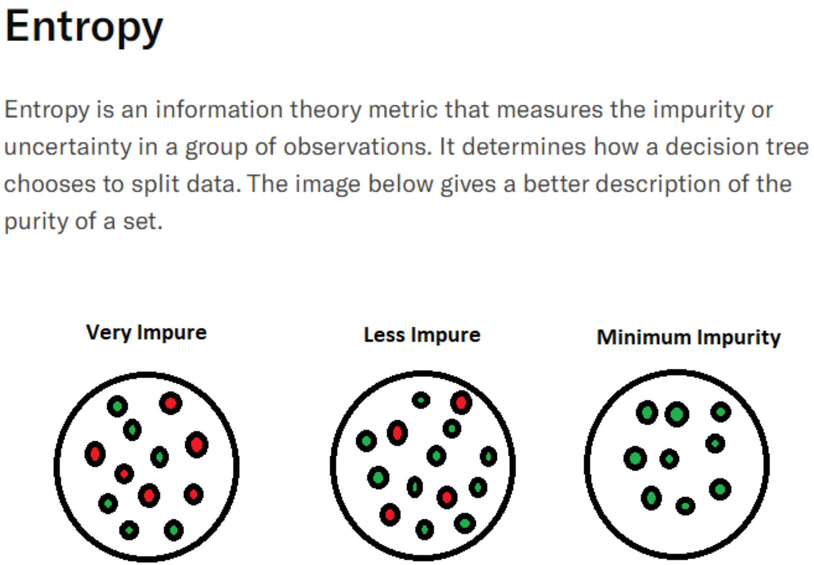
“Is dog big?” “Does it have short fur?”

It's like a diagram where each branch represents a decision based on a feature, leading to a final outcome.

Leaf=final outcome=one of dog breeds

Decision Tree: The criterion to split data

The difference of entropies:



Entropy

Entropy is an information theory metric that measures the impurity or uncertainty in a group of observations. It determines how a decision tree chooses to split data. The image below gives a better description of the purity of a set.

Basically, the more the majority class is present the less is entropy

$$E = -\sum_{i=1}^{N} p_i log_2 p_i$$

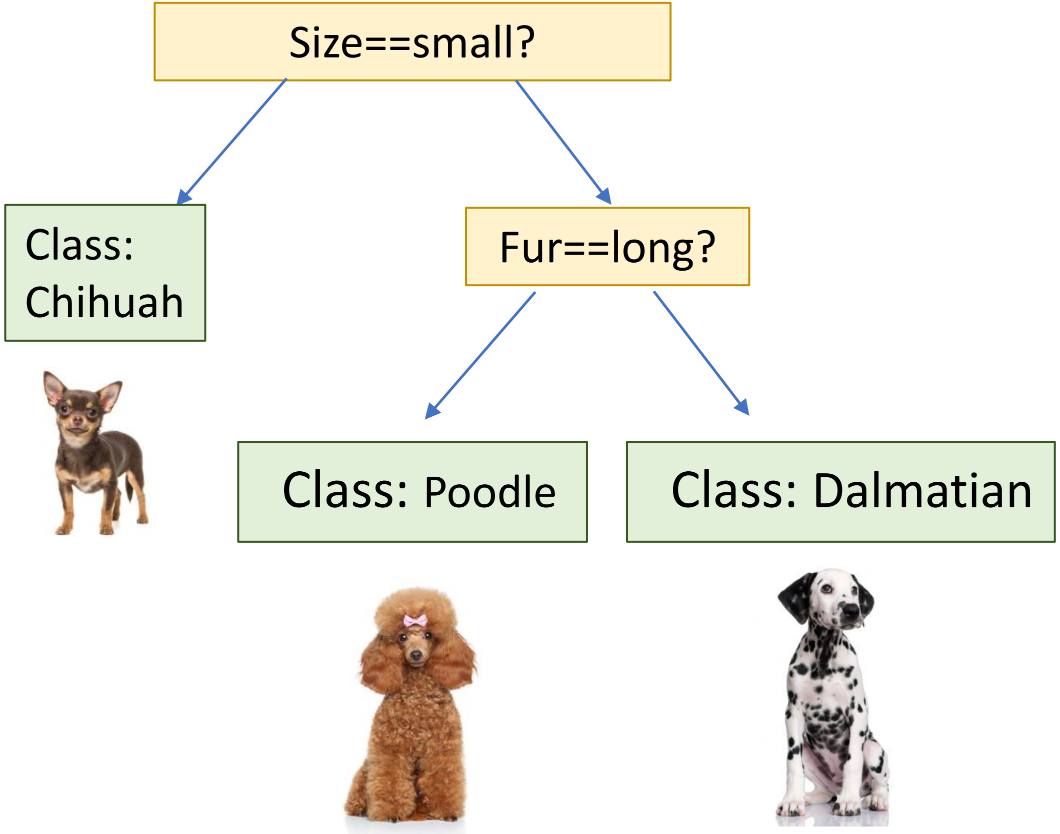
Where pi is the probability of randomly selecting an example in class i

Decision Tree: Dog example

What would be the first split that reduces the entropy the most?

size	fur	class
medium	long	Poodle
big	long	Poodle
small	short	Chihuahua
big	short	Dalmatian

Decision Tree: Dog example



Size==small?

Class: Chihuah

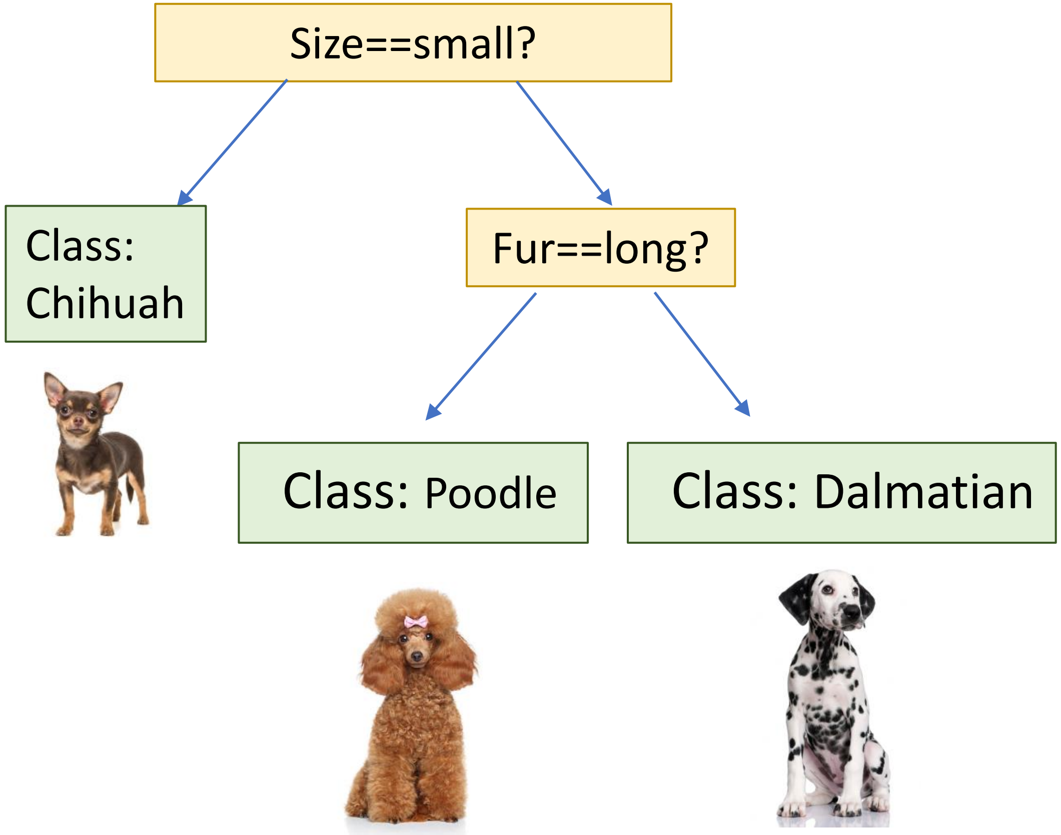
Fur==long?

Class: Poodle

Class: Dalmatian

size	fur	class
medium	long	Poodle
big	long	Poodle
small	short	Chihuahua
big	short	Dalmatian

Decision Tree: Dog example



Size==small?

Class: Chihuah

Fur==long?

Class: Poodle

Class: Dalmatian

Your tree is ready!
It has been deployed and now your algorithm needs to classify a new data point:

Classify this point:
size: medium
fur: medium

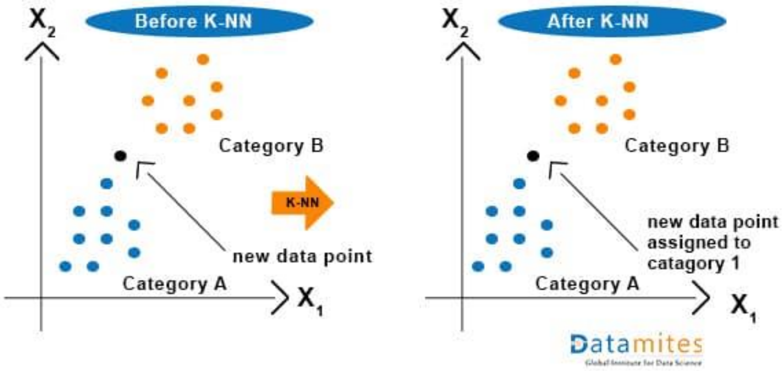
What would be the result?

K nearest neighbours

KNN is a simple algorithm that relies on the **"wisdom of the crowd"** principle

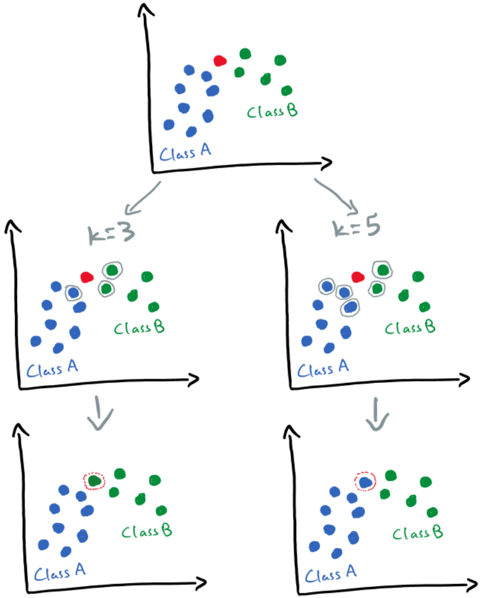
The class of a new data point is determined by **the majority** of its nearest neighbors.

The number of neighbors is determined by the parameter K



K nearest neighbours

One thing to keep in mind with KNN is that **the choice of K** can affect the algorithm's performance!



Why k should always be odd (2n+1)?

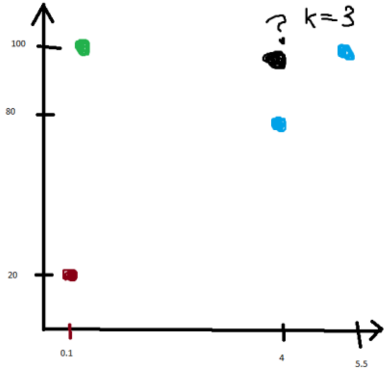
K nearest neighbours

What is the problem with our dataset that stops us from using KNN?

size	fur	class
medium	long	Poodle
big	long	Poodle
small	short	Chihuahua
big	short	Dalmatian

K nearest neighbours

Discreet values -> Continuous values



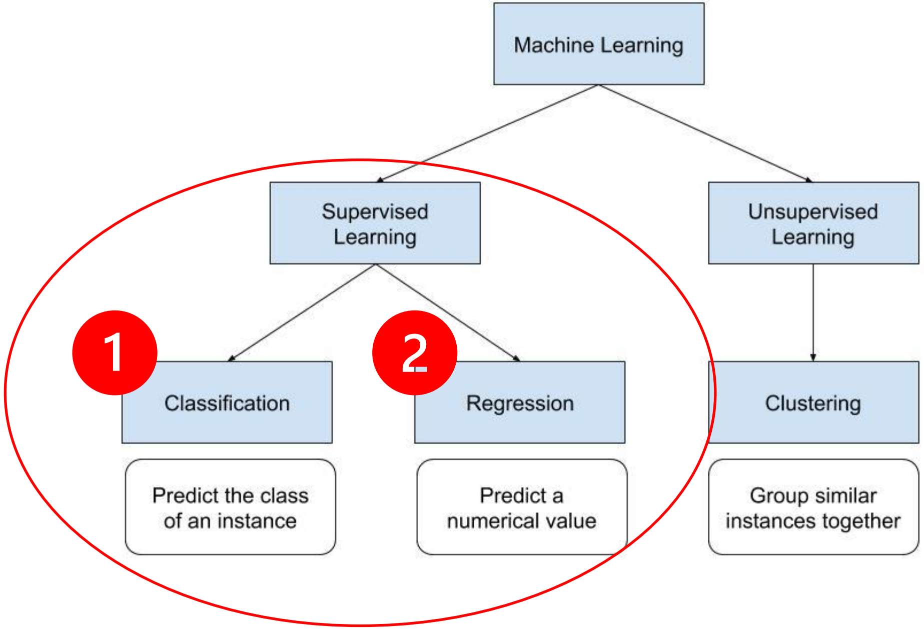
Size cm	Fur cm	class
100	5,5	Poodle
80	4	Poodle
20	0,09	Chihuahua
105	0,1	Dalmatian

What is the class of point K if red= Chihuahua, green=Dalmatian and blue = Poodle ?

Performance measure: Classification Task

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}}$$

Supervised Learning



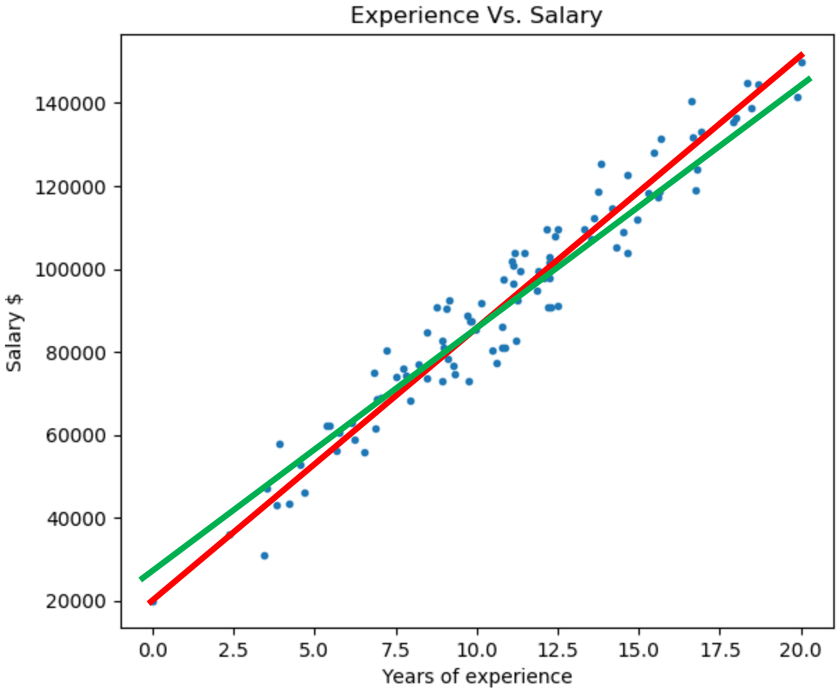
Regression

Regression is a type of machine-learning algorithm used for predicting continuous numerical values

Example:
Prediction of salary based on years of experience in the job

The most common type: Linear Regression

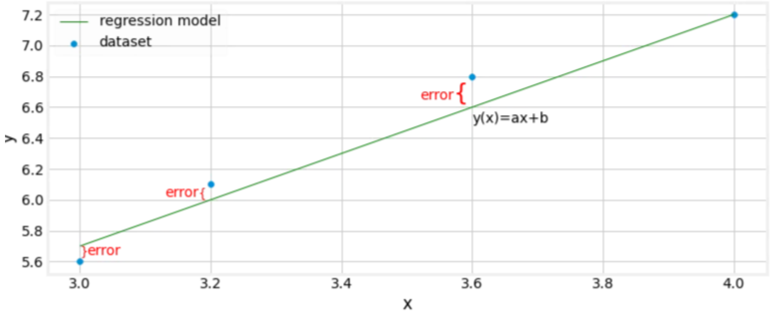
Linear regression =
We are trying to find the best line for a dataset



How to choose the right one ?

Linear Regression

1) Initialization:
   Computing a random line y(x)=a*x+b



2) Now we have an equation the error=> we can minimize it

The sum of squared errors is defined as:

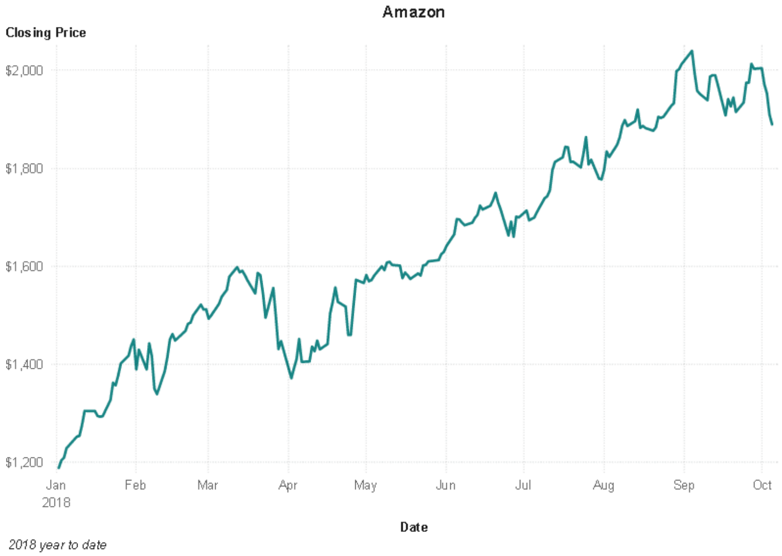
3) Compute **a** and **b** to minimize the error

$$S(a,b) = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - ax_i - b)^2$$

Example of Linear Regression

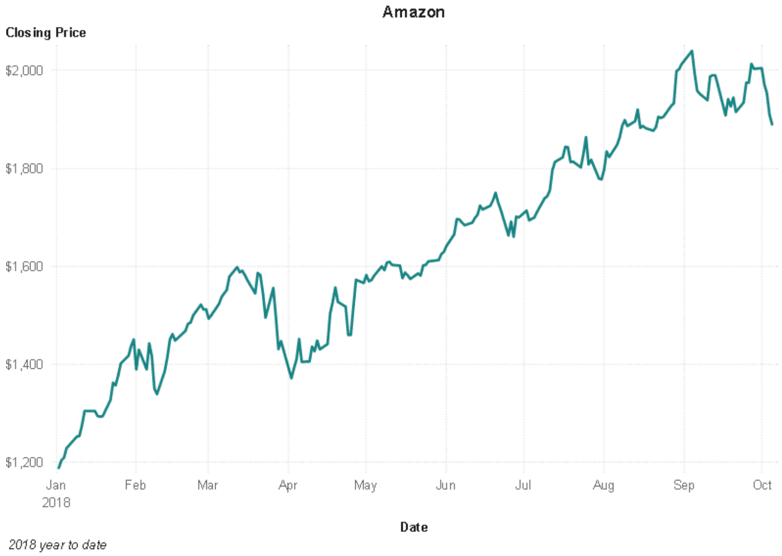
Fund Returns Predictions

Stock Market prices

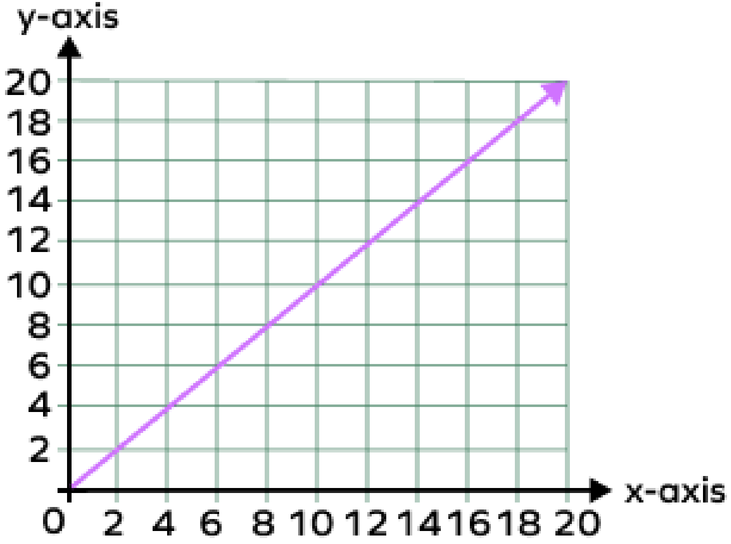


Nothing linear here !!!

Stock Market prices



Linear function

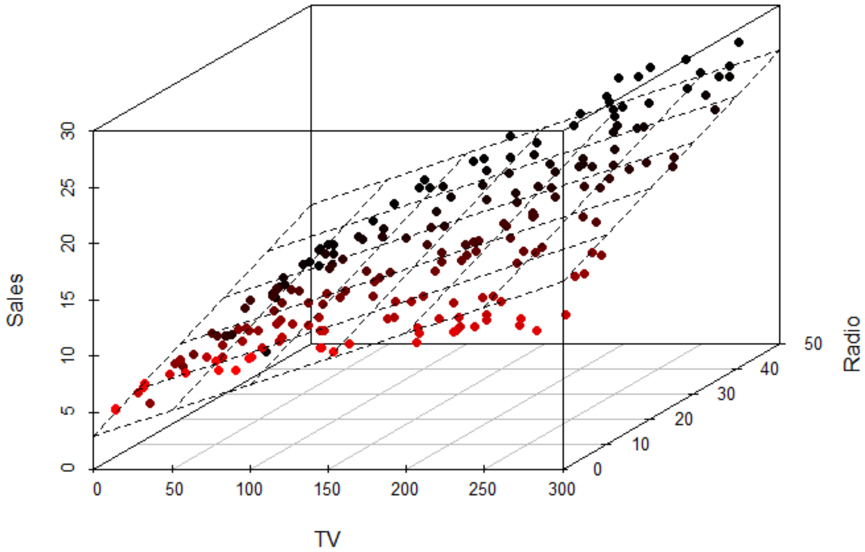


Well, actually

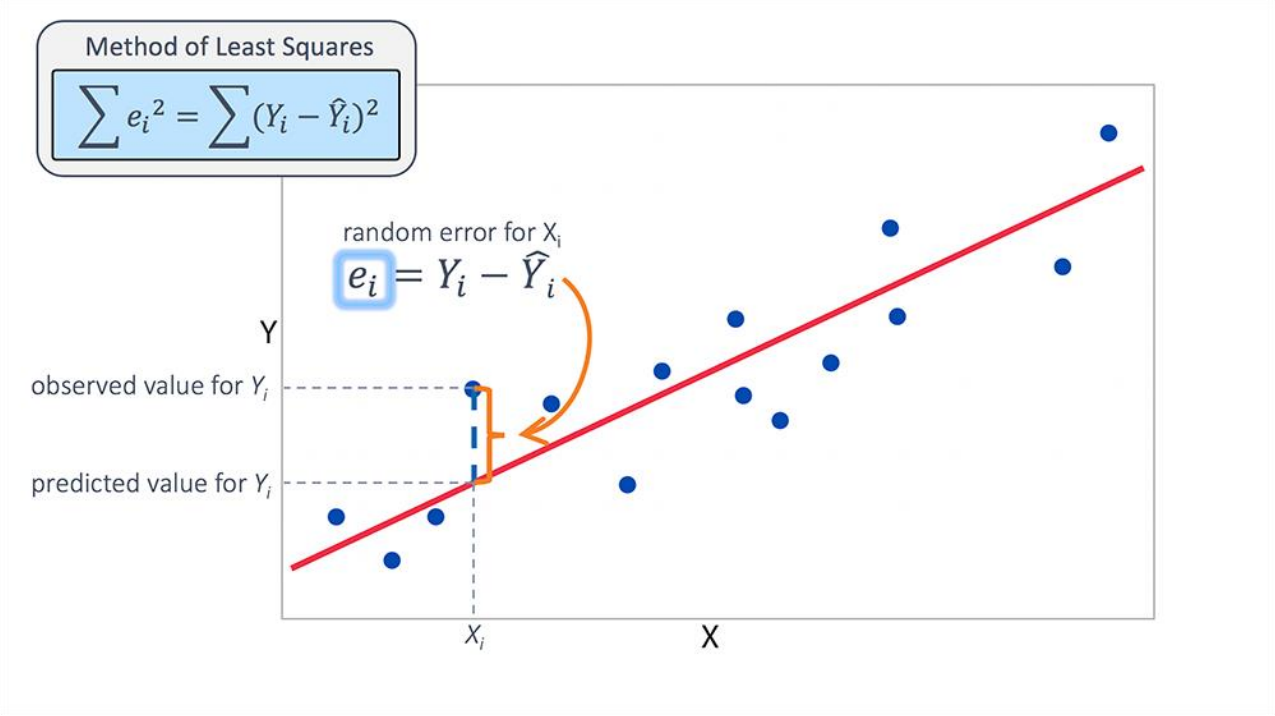
- Time is only one of the possible variables (and most of the time the least important)

- By increasing the number of variables (dimensions) we can achieve Linearity

- In case of Stock price, we can choose variables like: Company’s profit, Company’s dividends, the economical situation in the country etc



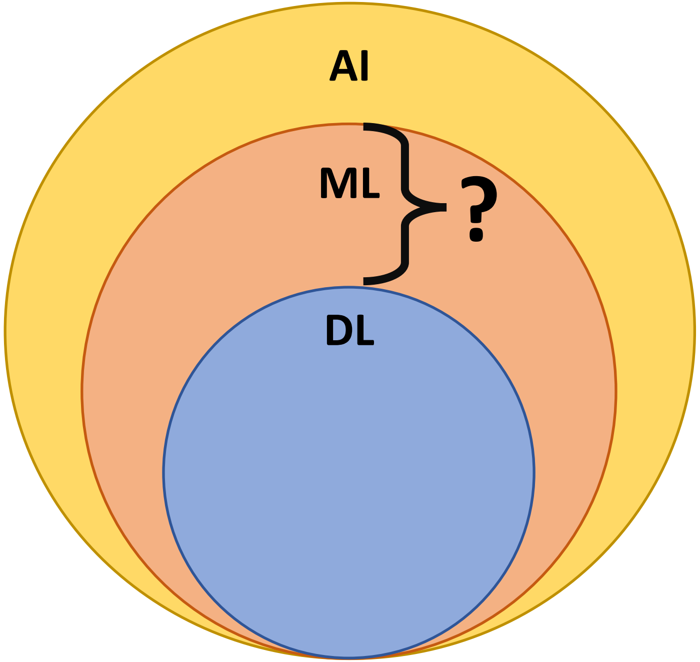
Performance measure : Linear regression



Linear Regression

The most important assumption:
all features (coordiantes) should be linearly correlated to the outcome!!!

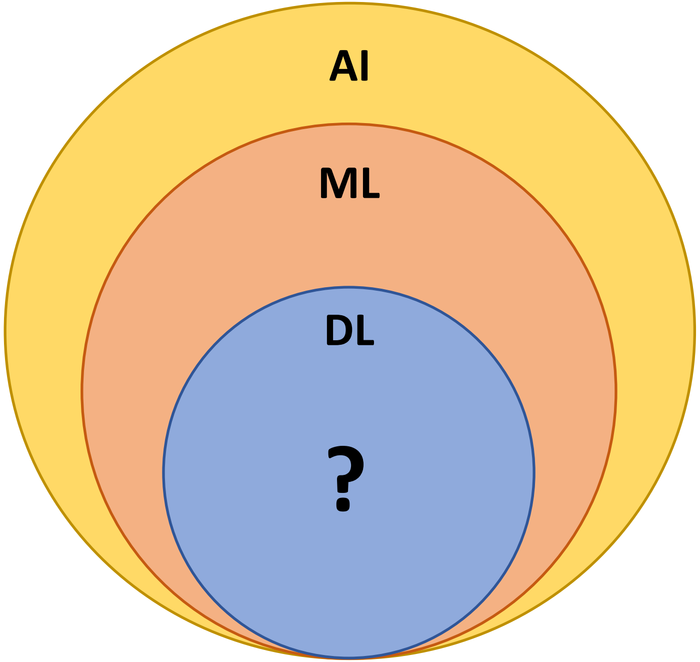
ML & Deep Learning



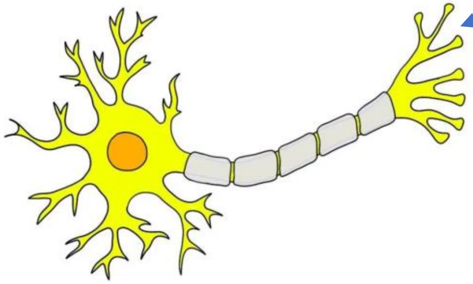
Algorithms based on statistical knowledge:

- Clustering
- Regressions

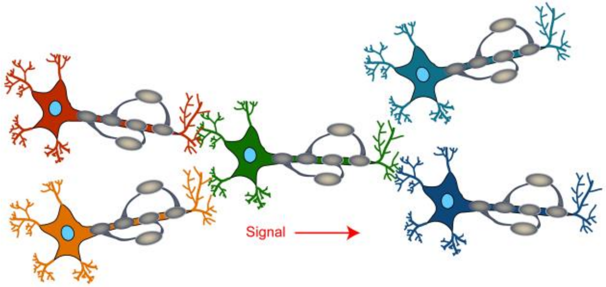
Deep Learning



- Technique that teaches computers to do what comes naturally to humans: learn by example.

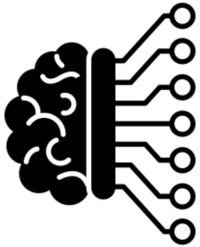


This is a neuron or a human brain cell



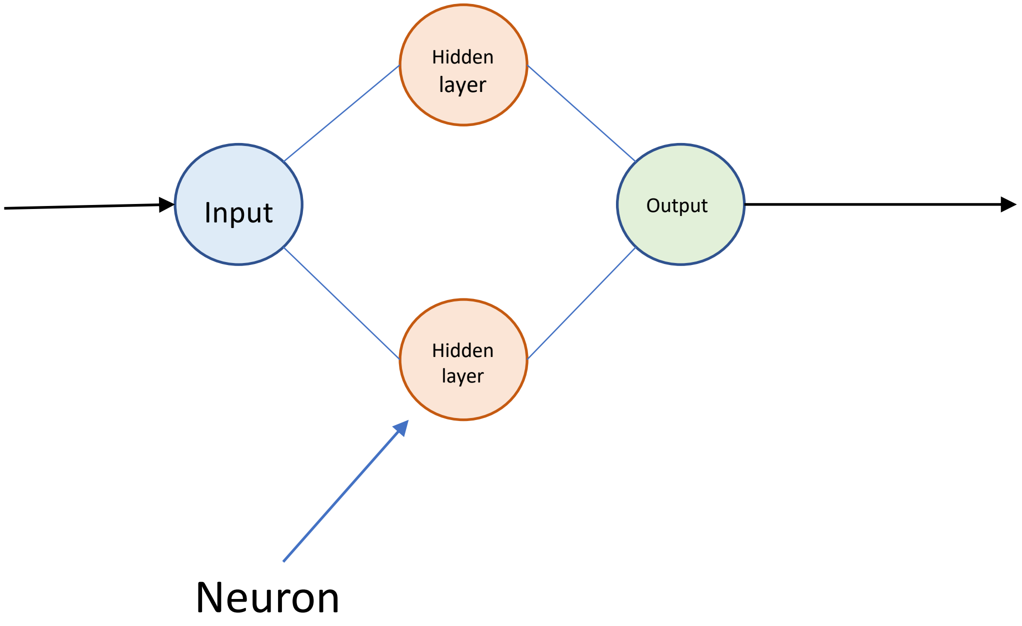
The Data Scientists have been inspired by neurons and created Neural networks

They function like neurons assembled together



And when you assemble A LOT of neurons, it is called **Deep Learning**

Deep Learning

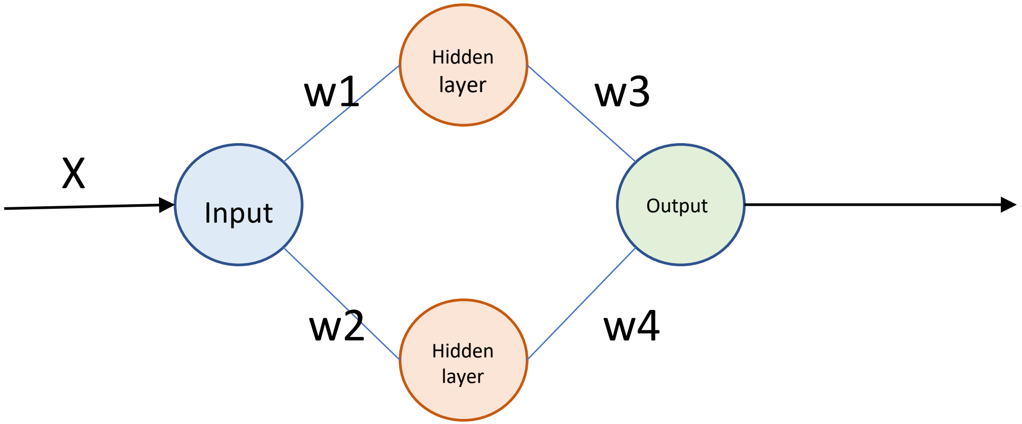


Imagine you have a dataset

X	Y
a	b
c	d
etc	

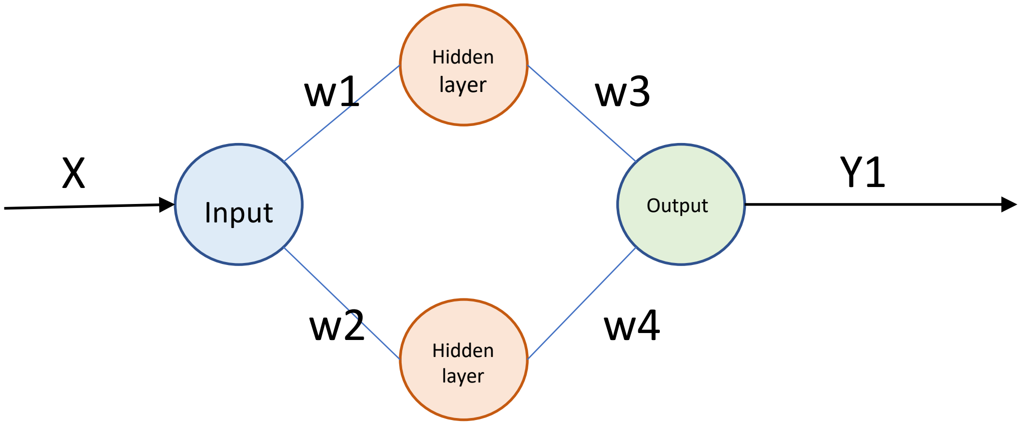
You want to predict Y given X

Deep Learning



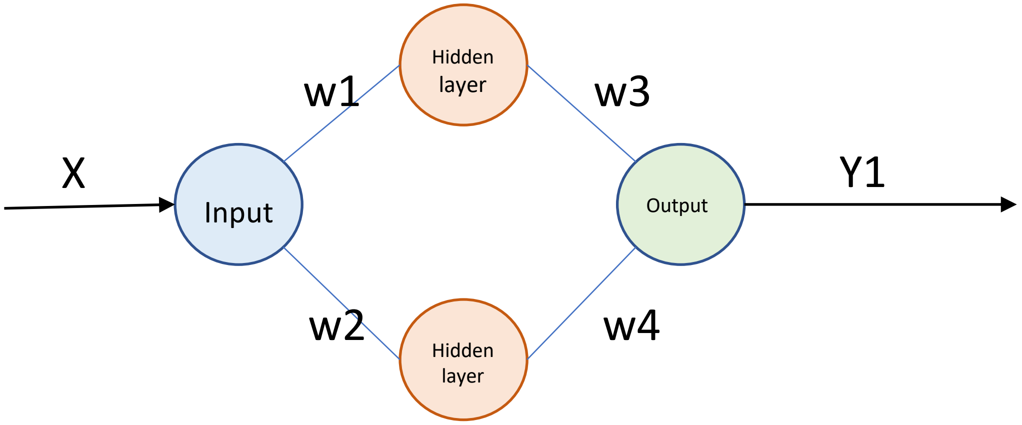
1. Initiation: Randomized weights

Deep Learning



1. Initiation: Randomized weights
2. You pass all links between neurons multiplying X by weights and summing up in the end. You get Y1!

Deep Learning



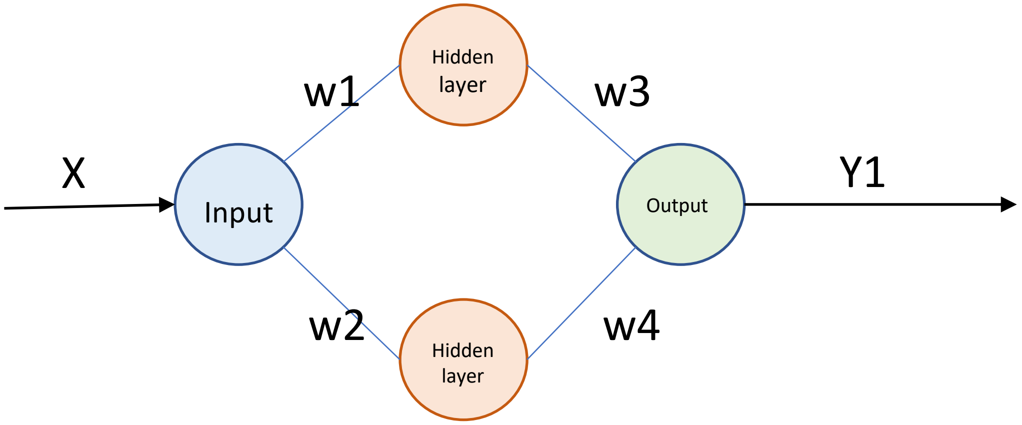
1. Initiation: Randomized weights

2. You pass through all links between neurons multiplying X by weights and summing up in the end. You get Y1!

3. Compute error Y-Y1 and compute the influence of each weight on the error

4. Adjust each weight a little bit

Deep Learning



Repeat a loooooot of times

And

Done !

Congrats! Now you are almost a Data Scientist ☺

Thank you for your attention