Day 1

Detailed Introduction
into AI / Data Science and Big Data

Who is Ahmed?

Oregon State University

PhazeVentures

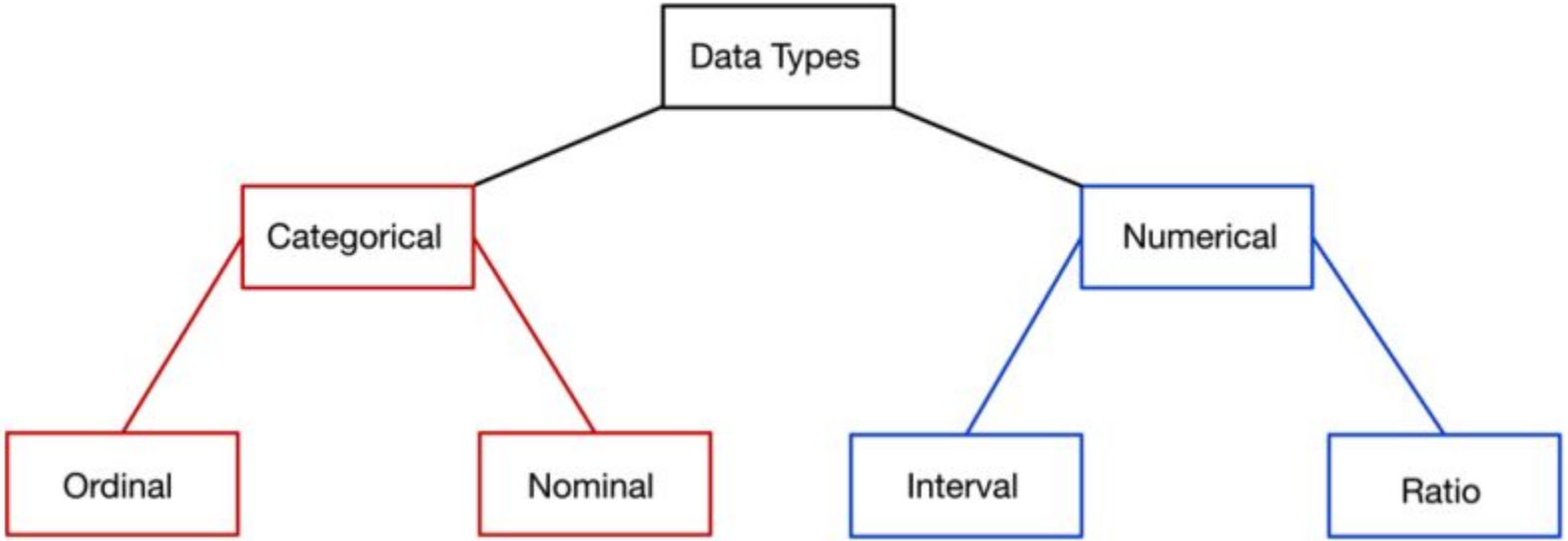
PhazeRo

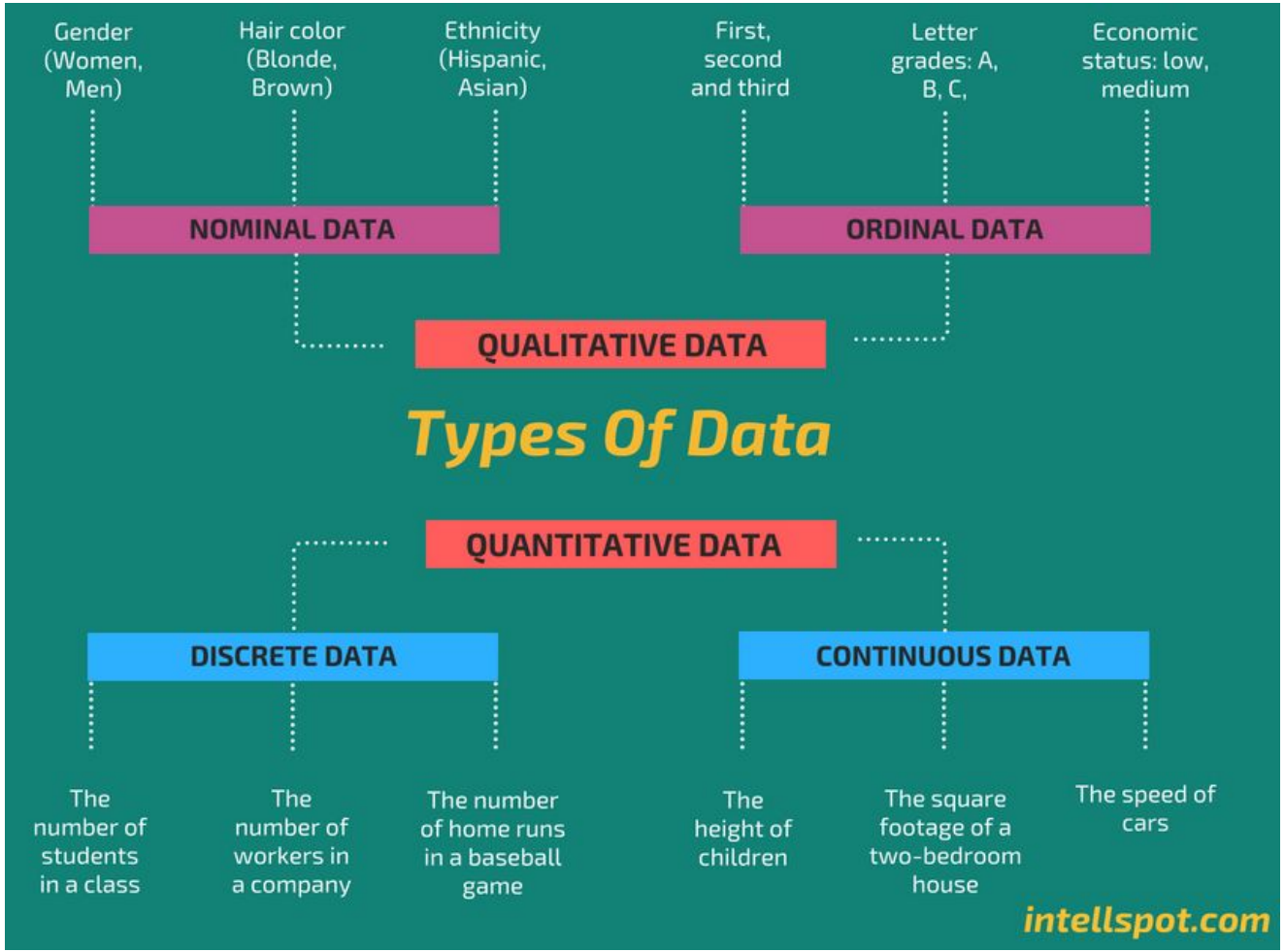
HARBOUR.SPACE UNIVERSITY

CENTRAL BANK OF OMAN

Pyra

What is Data?


Data Types
Categorical
Numerical
Ordinal
Nominal
Interval
Ratio


Gender (Women, Men)
Hair color (Blonde, Brown)
Ethnicity (Hispanic, Asian)
First, second and third
Letter grades: A, B, C,
Economic status: low, medium
NOMINAL DATA
ORDINAL DATA
QUALITATIVE DATA
Types Of Data
QUANTITATIVE DATA
DISCRETE DATA
CONTINUOUS DATA
The number of students in a class
The number of workers in a company
The number of home runs in a baseball game
The height of children
The square footage of a two-bedroom house
The speed of cars
intellspot.com

Statistics

Definition:

“Statistics concerns the collection, organization, analysis, interpretation, and presentation of data.”

Types of statistics:

- Descriptive statistics
- Inferential statistics

Descriptive Statistics

Summarizes (describes) a collection of data

Reveals useful or at least interesting things about the data, in particular “what is going in there”.

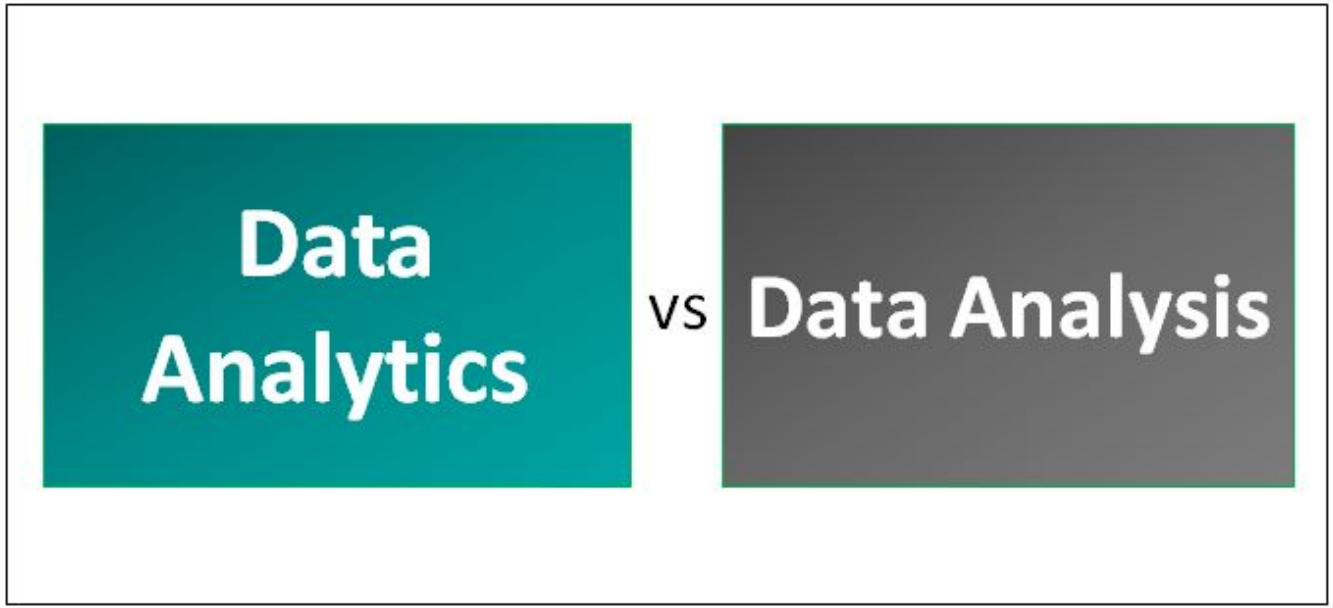
Inferential Statistics

- Helps us draw conclusions (inferences) about things that we do not fully grasp
- Helps us understand more about populations, phenomena, and systems from sample data (induction)
- Enables professionals to predict what is likely to happen if we make a specific decision or act on the entity in this or that manner

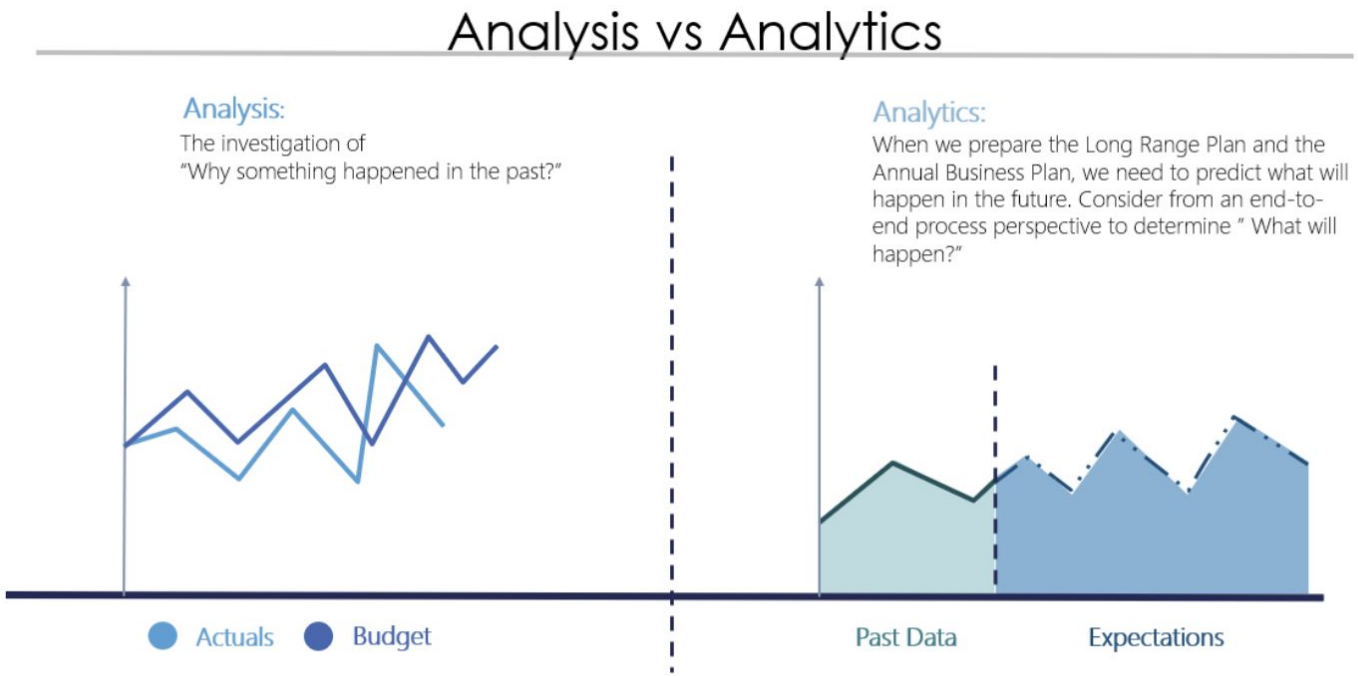
Business Intelligence (BI)

Definition:
“Business intelligence (BI) comprises the strategies and technologies used by enterprises for the data analysis of business information.”

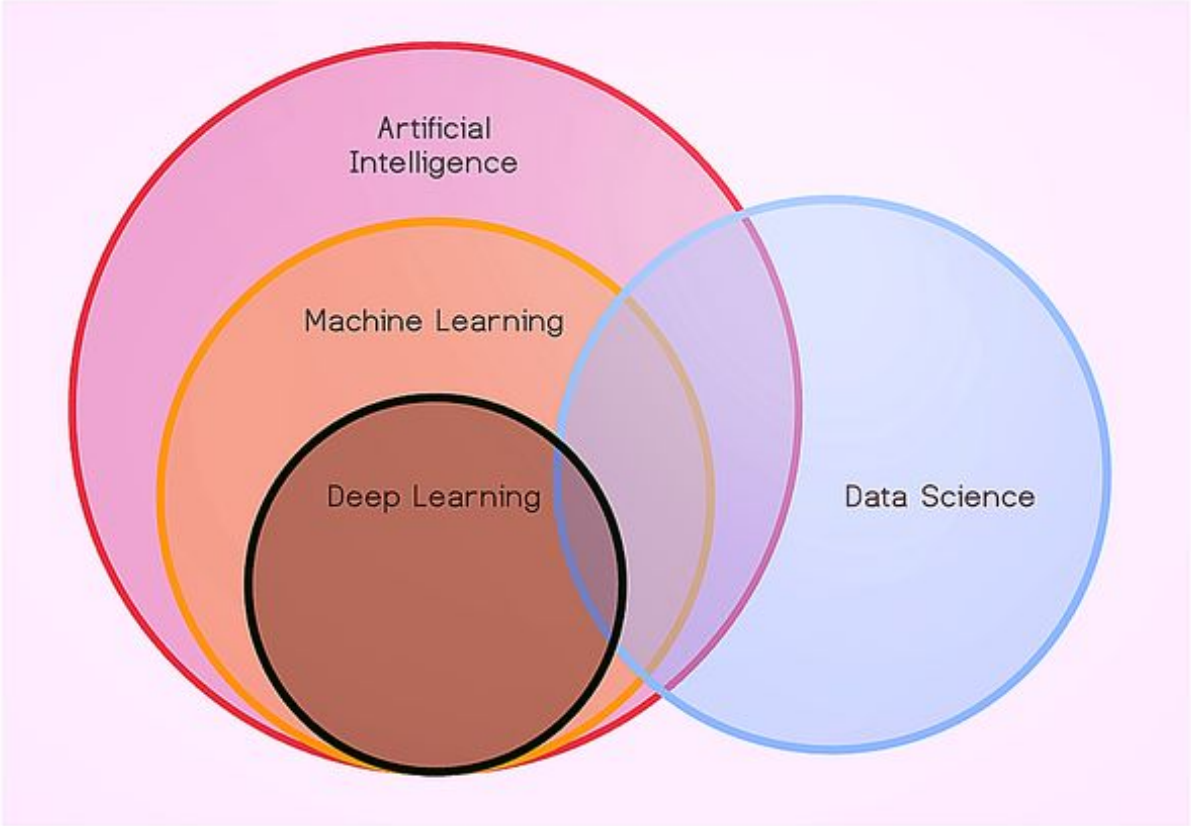
Data Science



Data Science



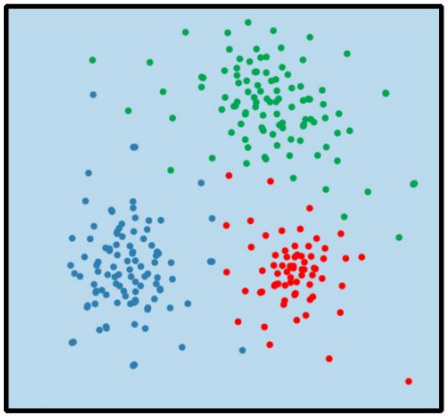
The Buzz Words



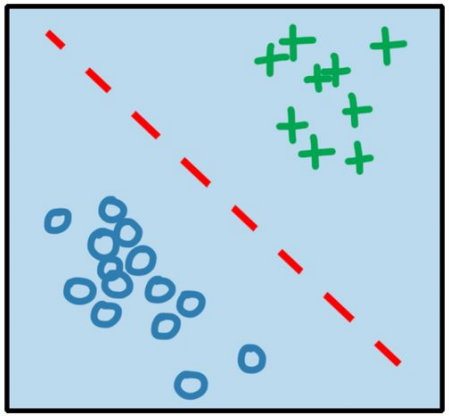
Types of AI

machine learning

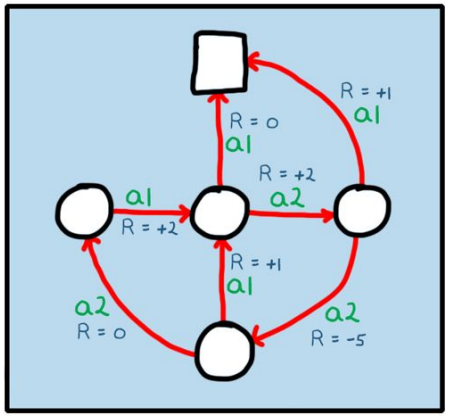
unsupervised learning



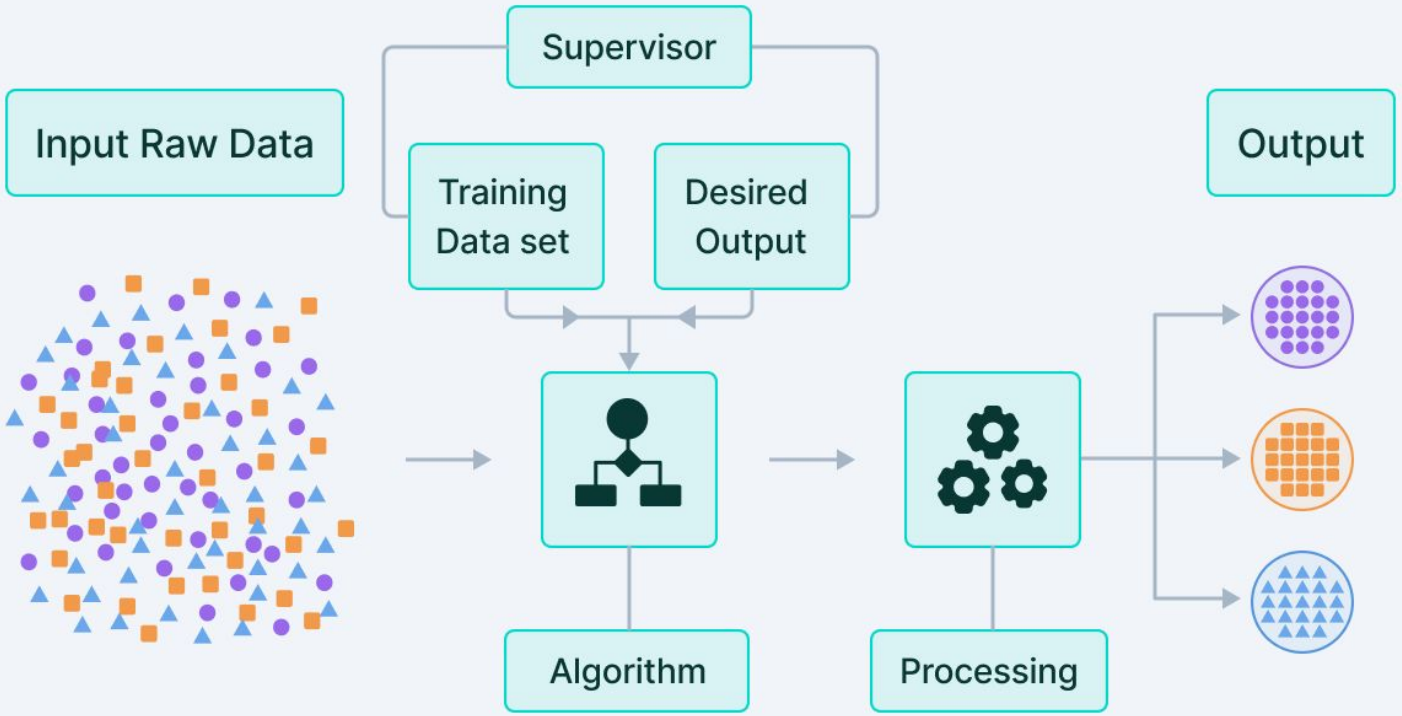
supervised learning



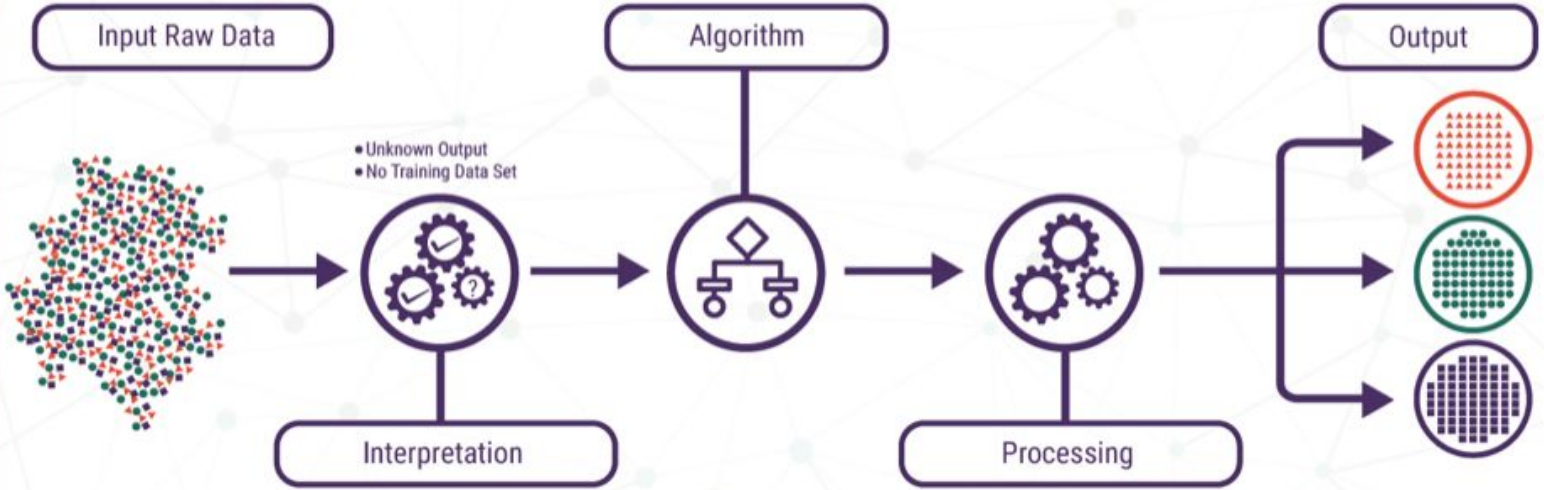
reinforcement learning



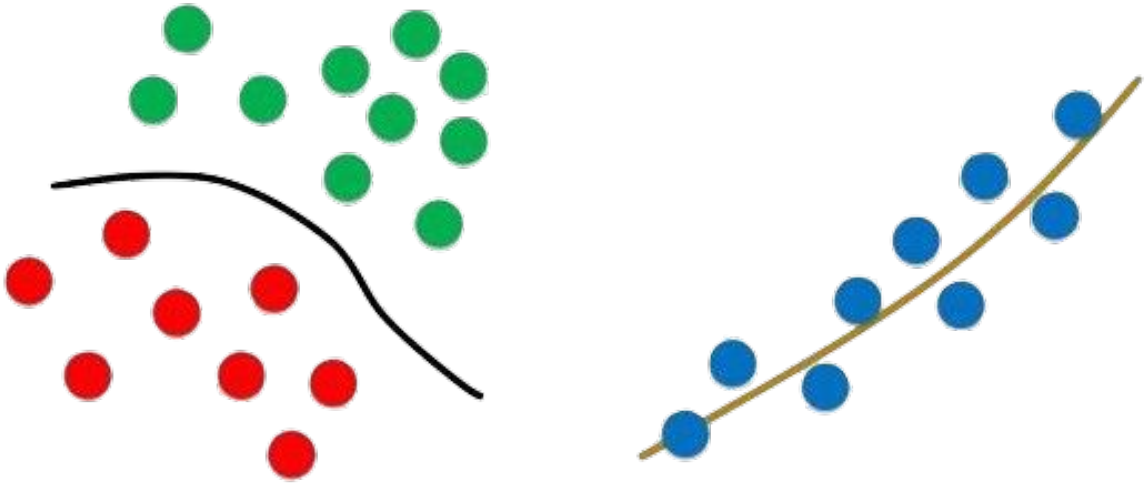
Supervised Learning



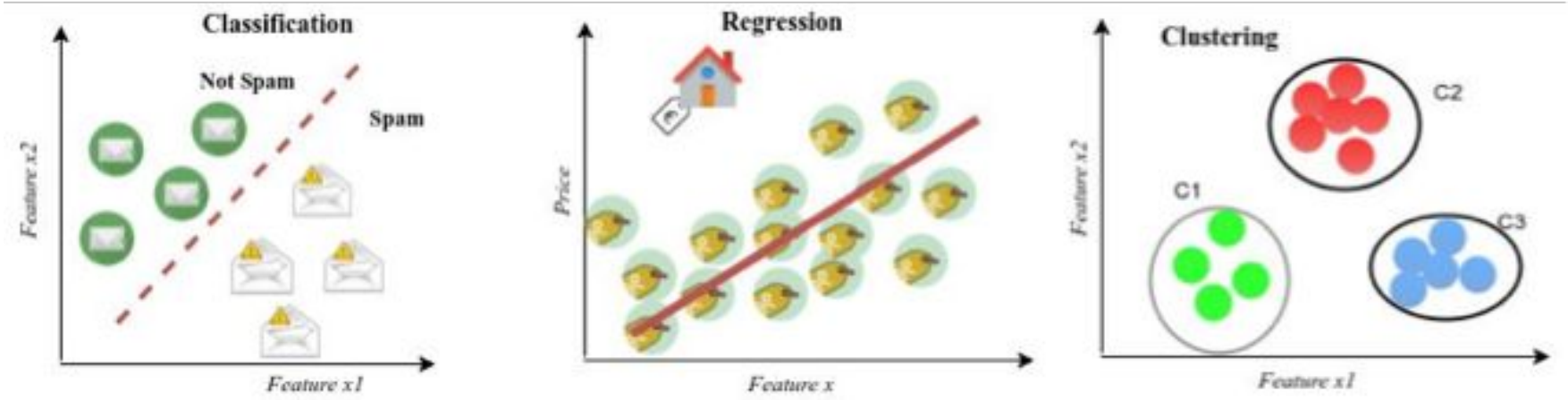
UNSUPERVISED LEARNING



Classification vs Regression

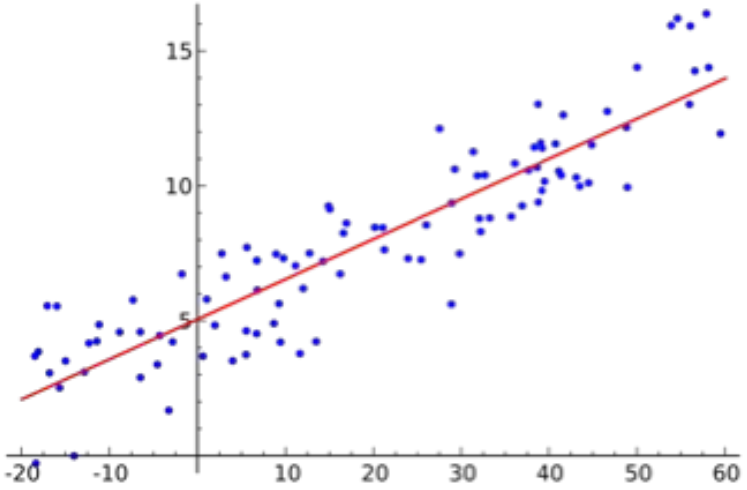


Classification vs Regression vs Clustering



ML / AI Algorithms

Linear Regression



Estimated (or predicted) Y value for observation i

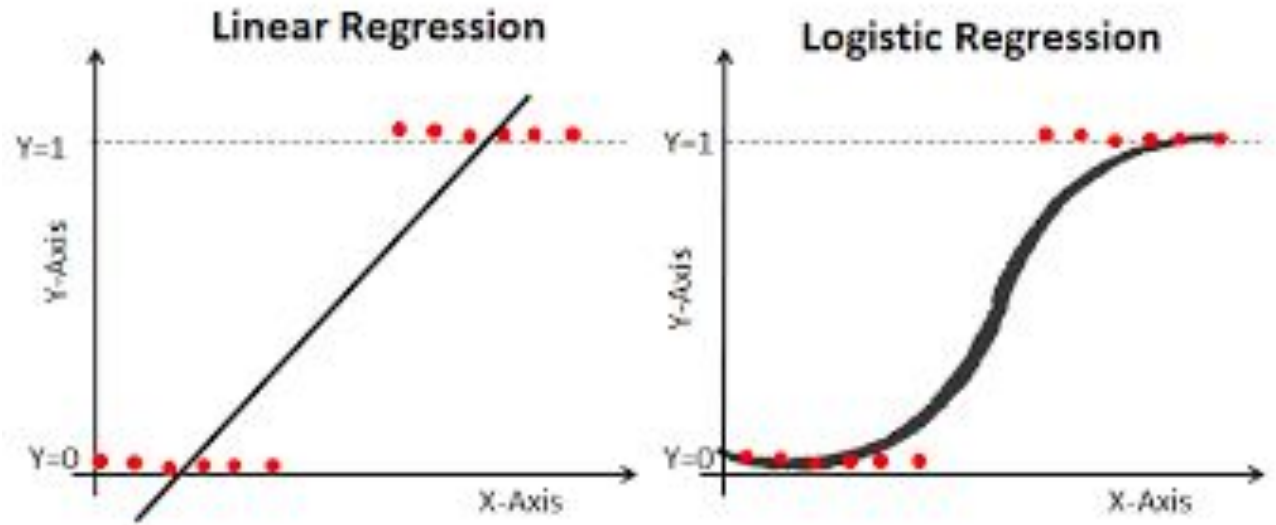
Estimate of the regression intercept

Estimate of the regression slope

Value of X for observation i

$$\hat{Y}_i = b_0 + b_1 X_i$$

Logistic Regression

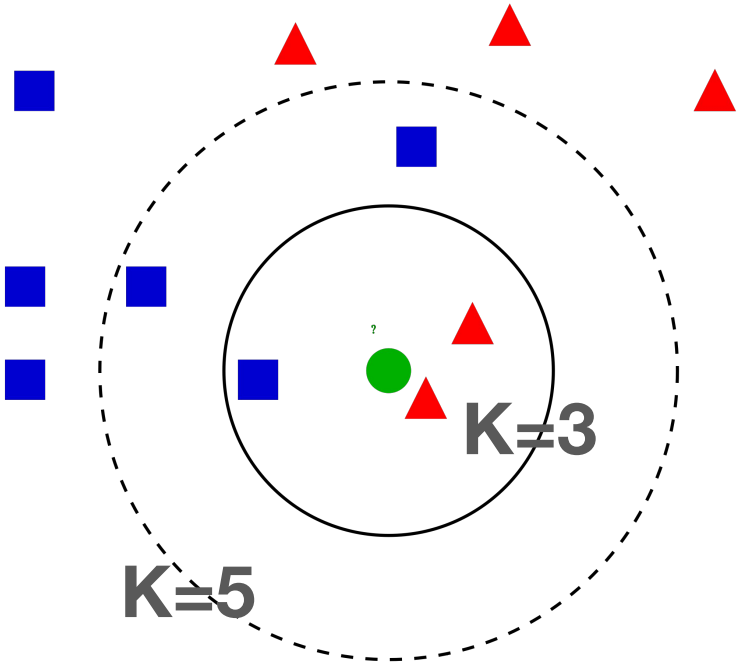


$$S(x) = \frac{1}{1 + e^{-x}}$$

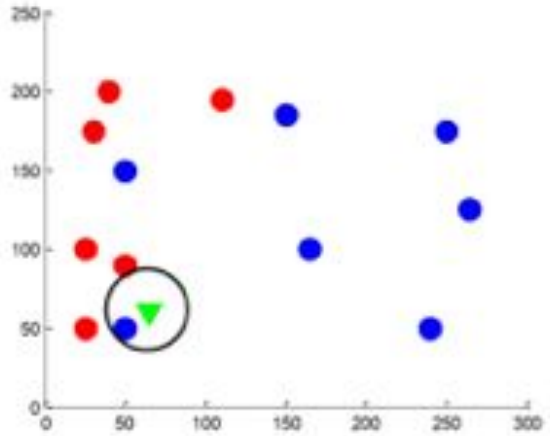
Naïve Bayes Classifier: Bayes Theorem

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

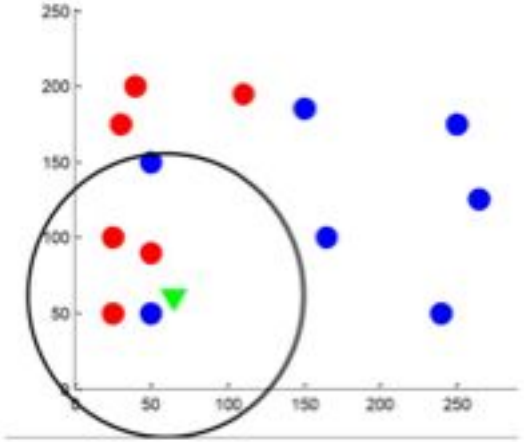
K-Nearest Neighbor



K-Nearest Neighbor – hyperparameter optimization



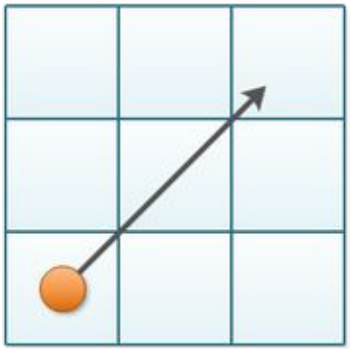
k = 1



k = 5

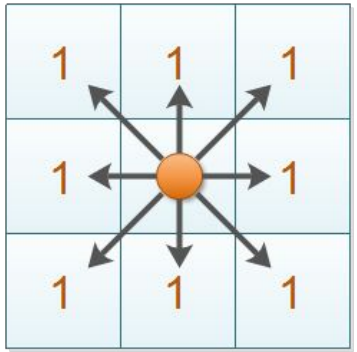
K-Nearest Neighbor – distance

Euclidean Distance



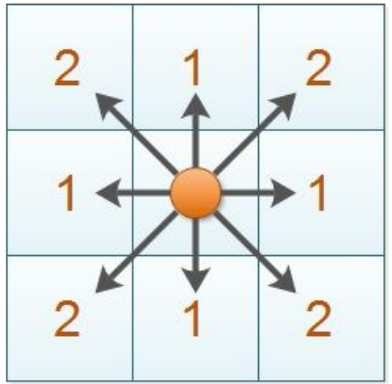
$$\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

Chebyshev Distance



$$\max(|x_1 - x_2|, |y_1 - y_2|)$$

Manhattan Distance

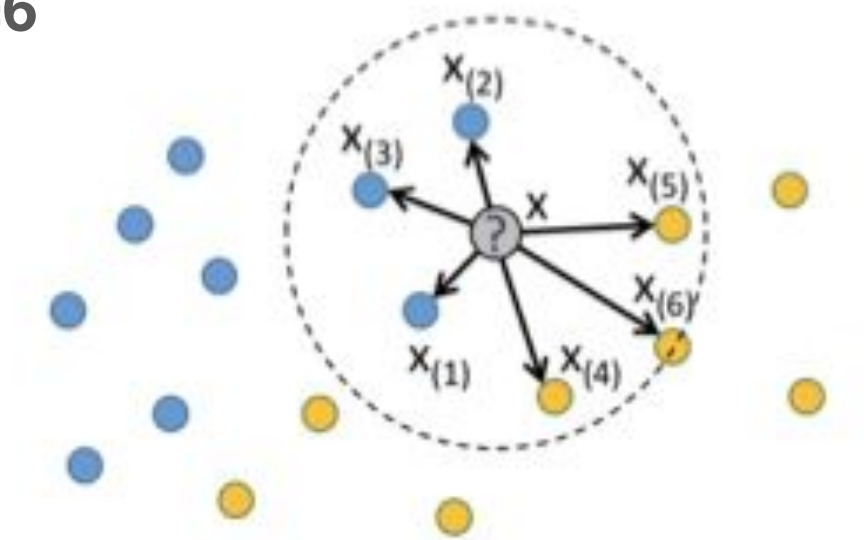


$$|x_1 - x_2| + |y_1 - y_2|$$

https://medium.com/@srishtisawla/k-nearest-neighbors-f77f6ee6b7f5

K-Nearest Neighbor – Weighted Neighbors

K=6



K-Nearest Neighbor – Weighted Neighbors

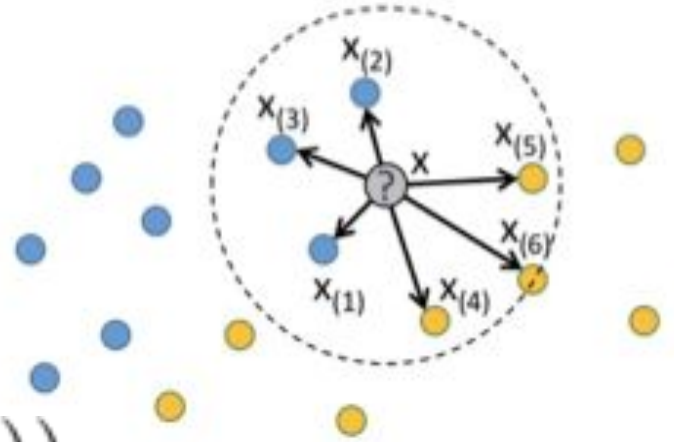
Weights Adjustments:

- Neighbors Orders

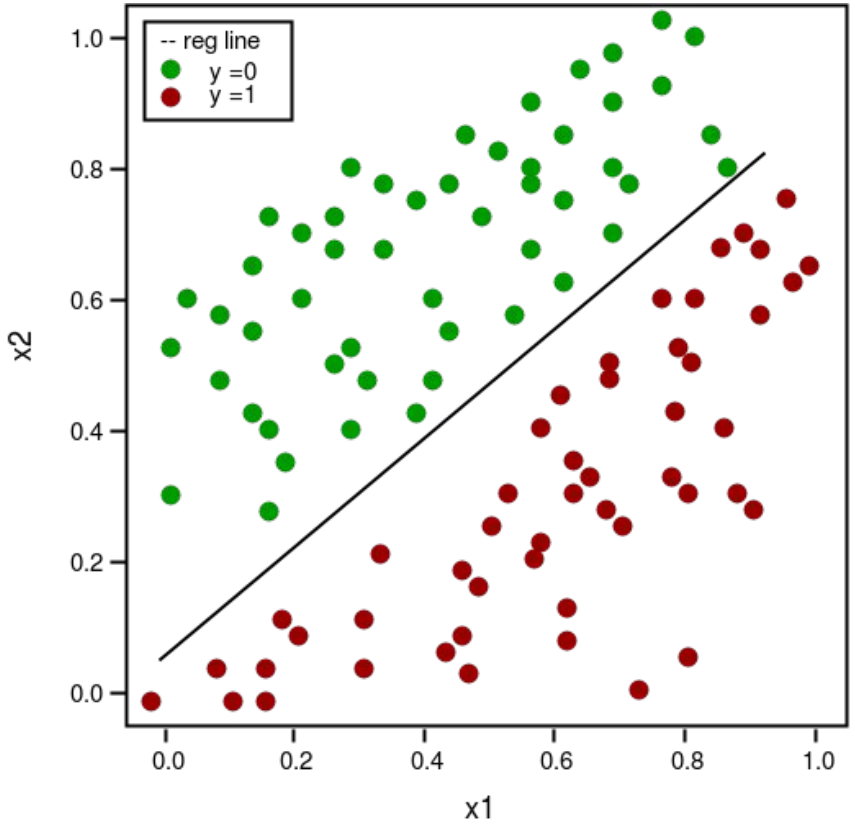
$$w(\mathbf{x}_{(i)}) = w_i$$

- Distance

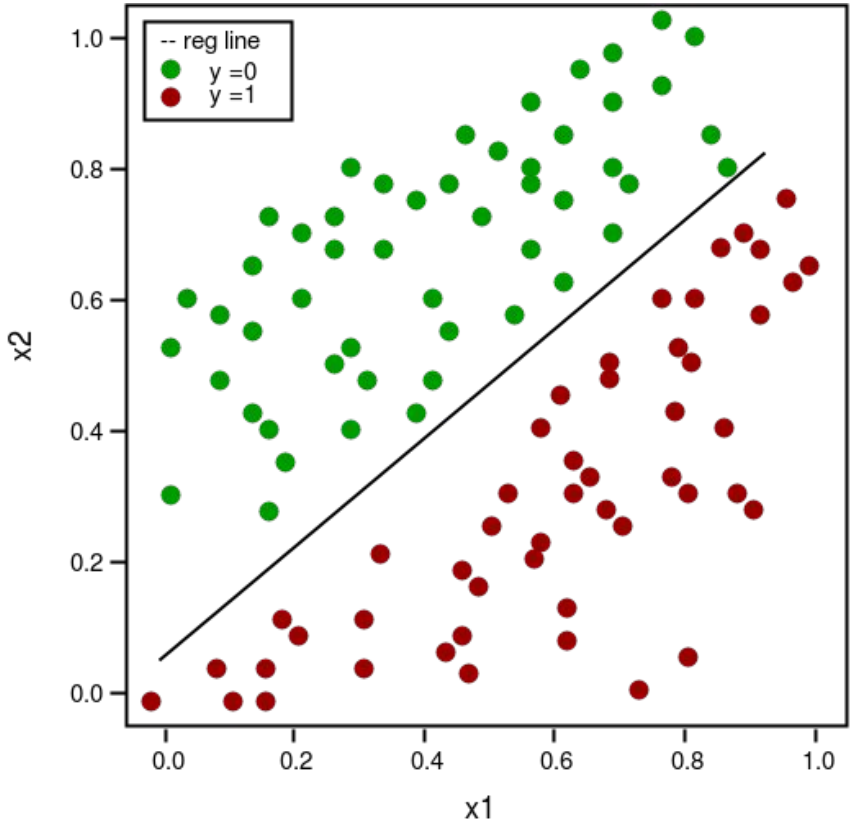
$$w(\mathbf{x}_{(i)}) = w(d(\mathbf{x}, \mathbf{x}_{(i)}))$$



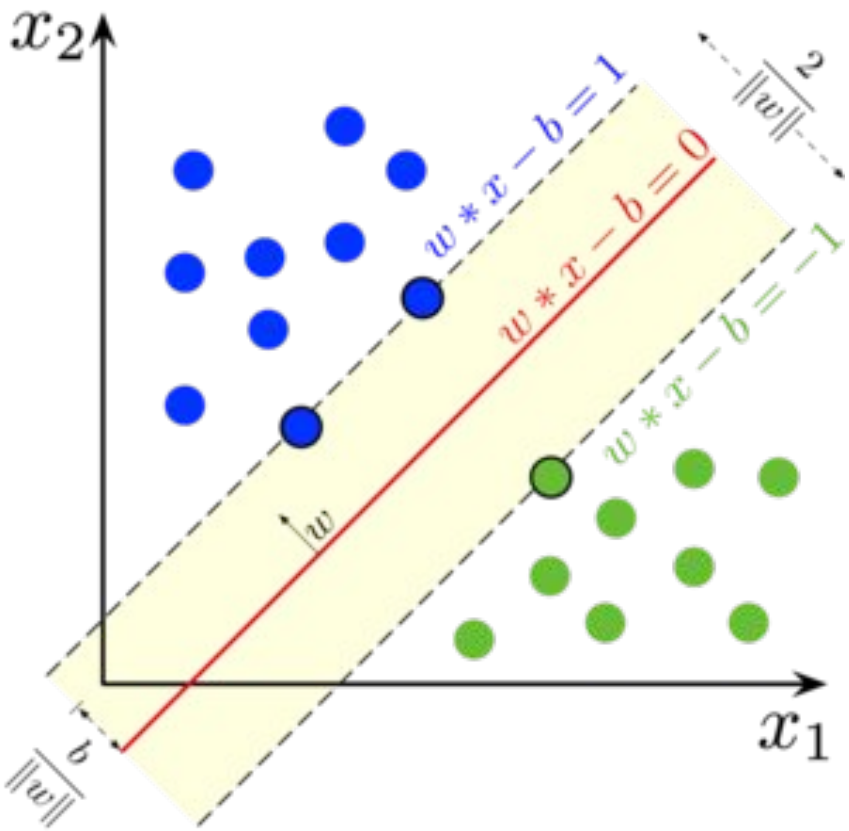
Linear Classification

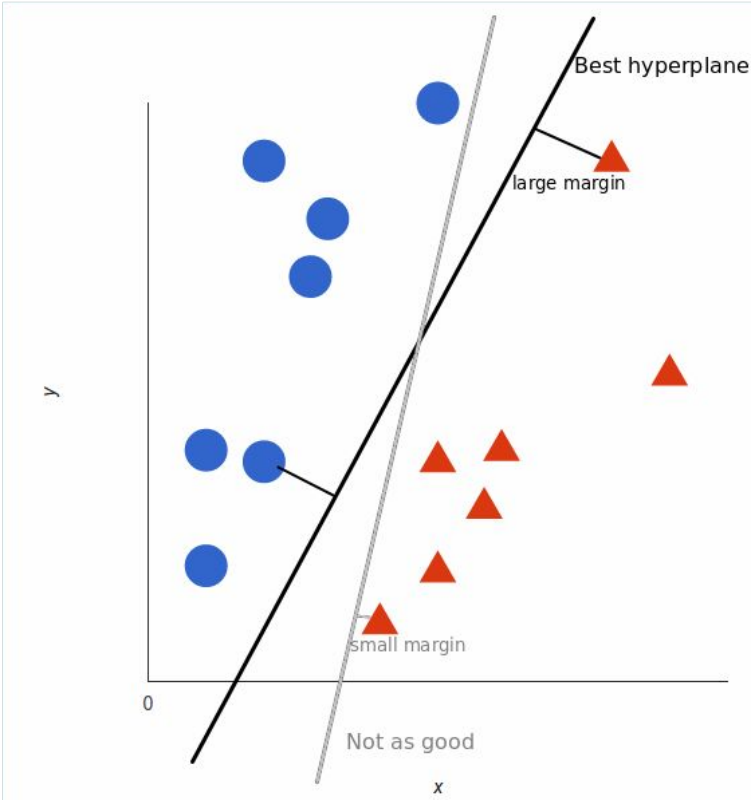


Support Vector Machines - SVMs

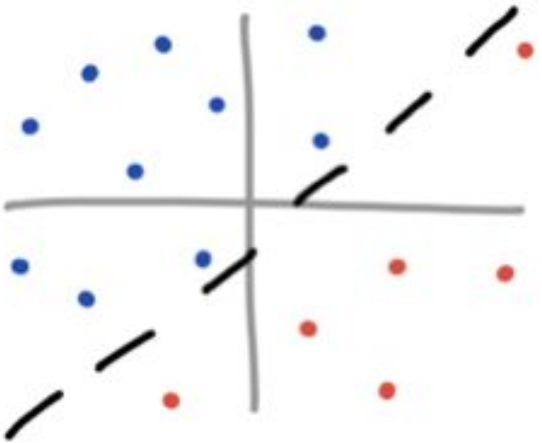


SVM


x_2
$w * x - b = 1$
$w * x - b = 0$
$w * x - b = -1$
$\frac{2}{\|w\|}$
w
$\frac{b}{\|w\|}$
x_1

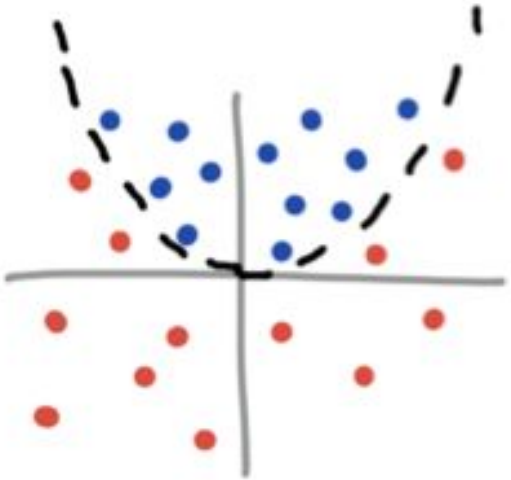

Best hyperplane
large margin
small margin
Not as good
y
x
0

SVM – Kernel Tricks



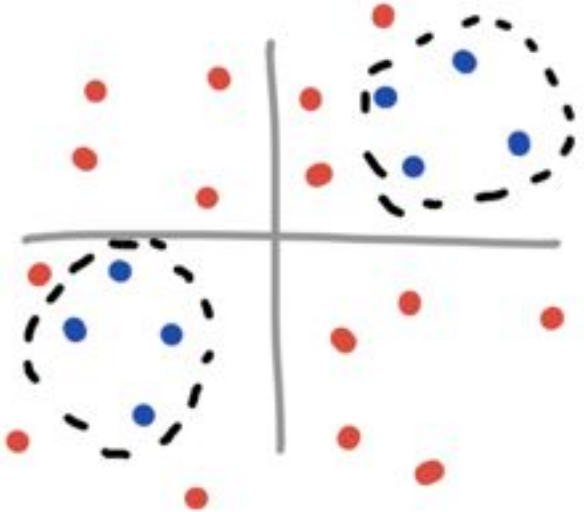
$K(w,x) =< w,x >$

Linear



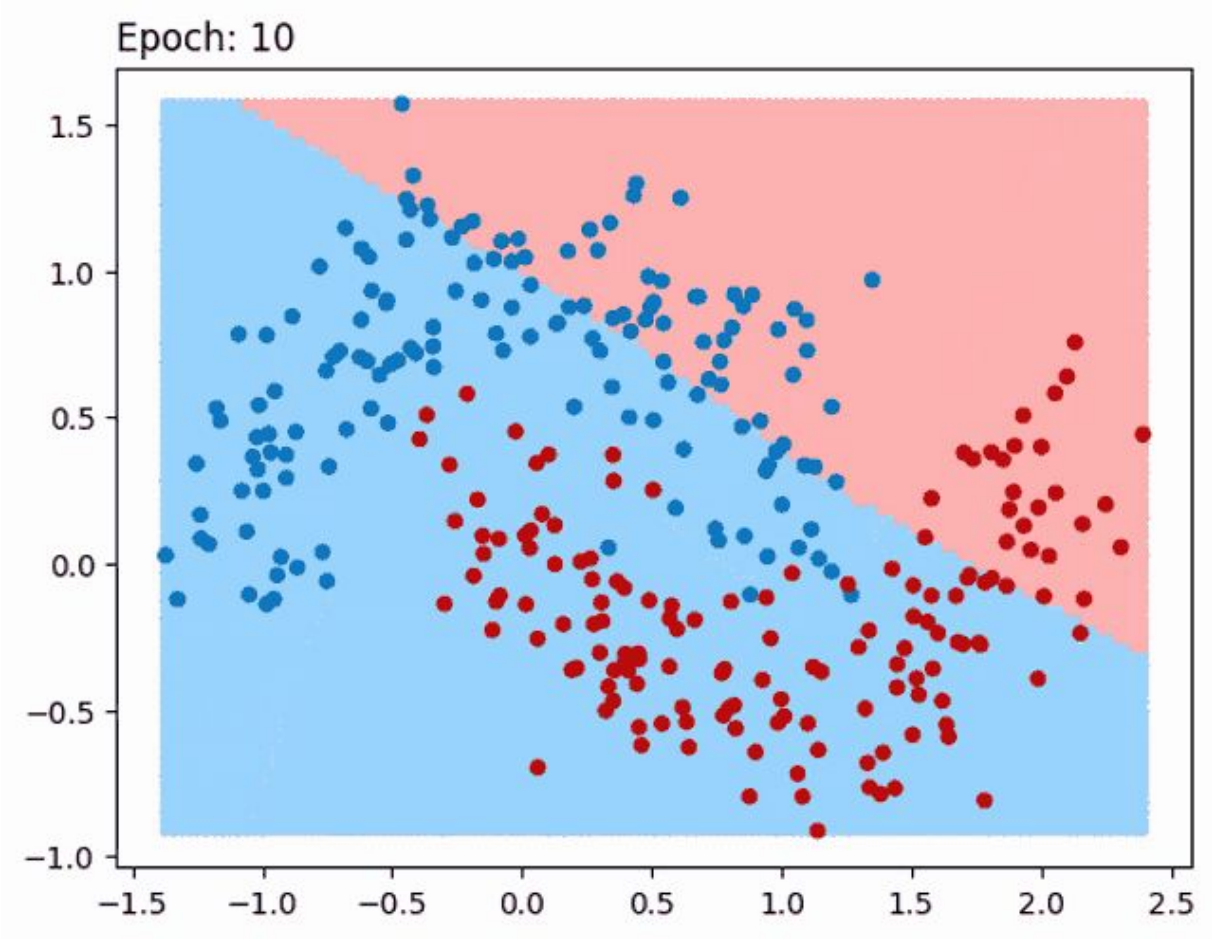
$K(w,x) = (\gamma < w,x > +r)^d$

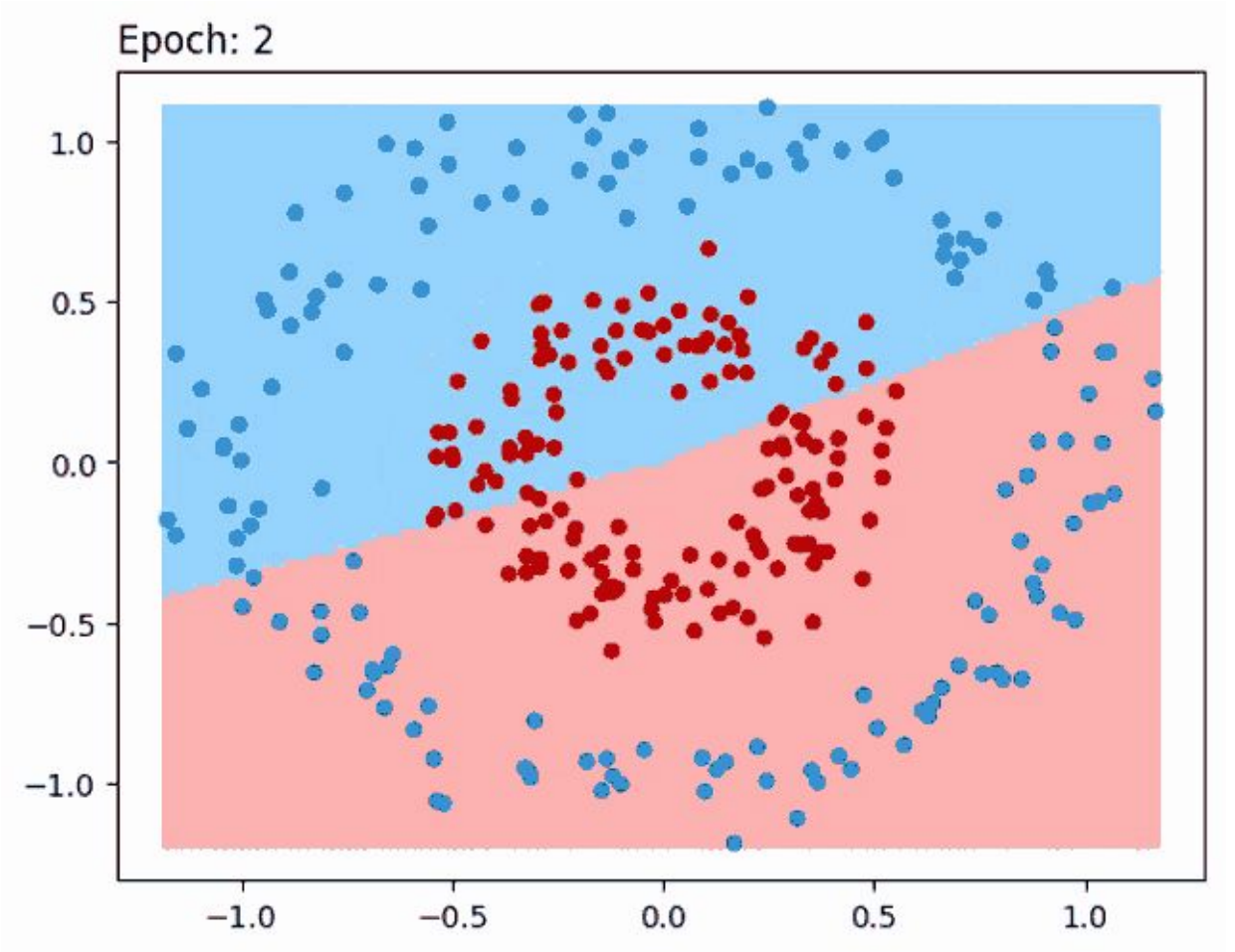
Polynomial



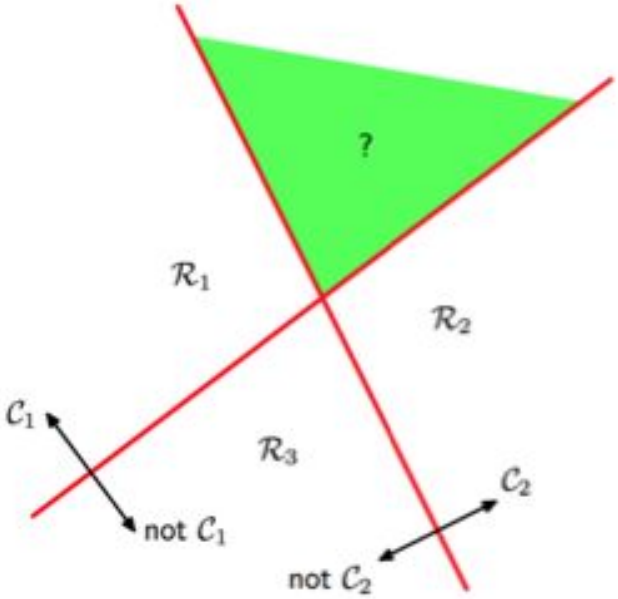
$K(w,x) = e^{-\gamma\|w-x\|^2}$

RBF

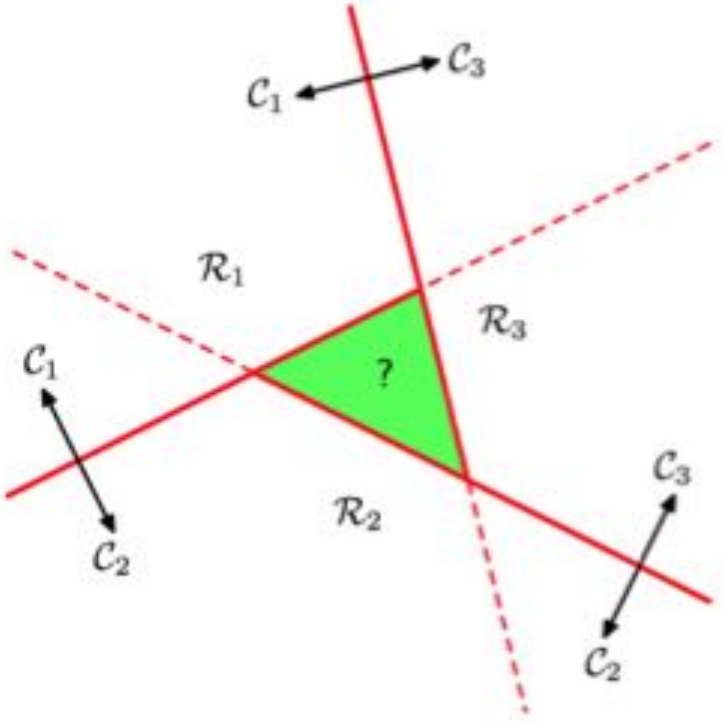

Epoch: 10
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
2.5


Epoch: 2
1.0
0.5
0.0
−0.5
−1.0
−1.0
−0.5
0.0
0.5
1.0

SVM – Multiclass Classification



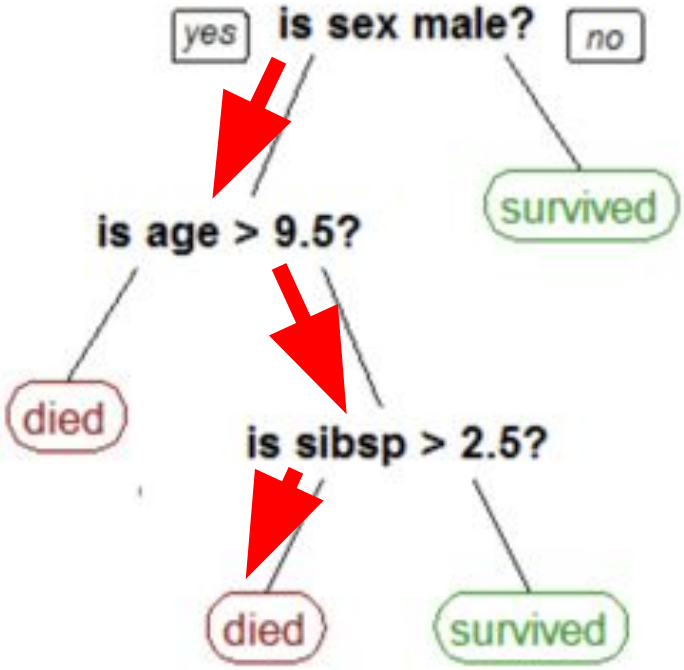
One vs Rest



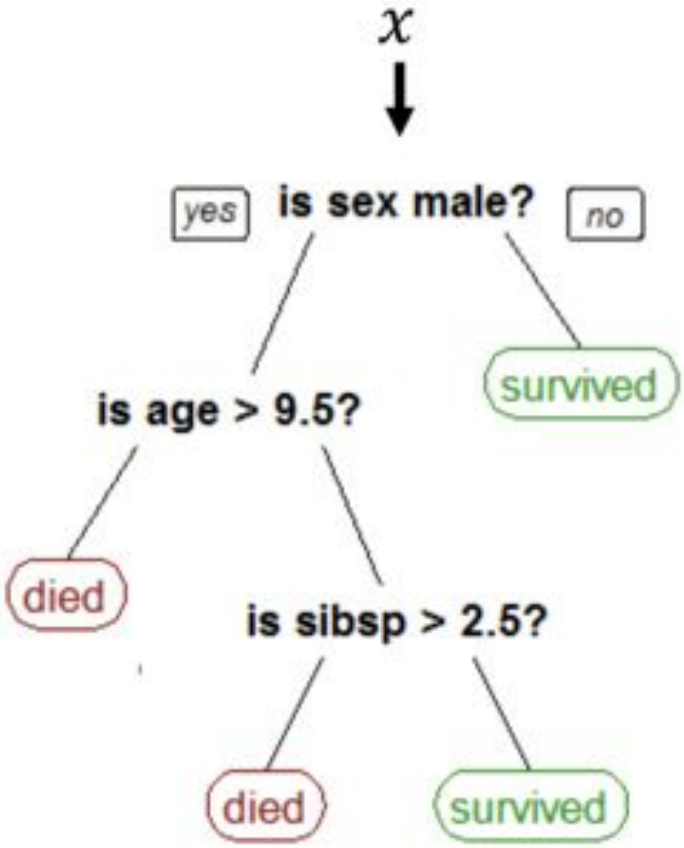
One vs One

Decision Trees

x = { age , *sex* , Sibsp } = { 9 , *male* , 3 }



Decision Trees



Decision Trees – Constructing Trees

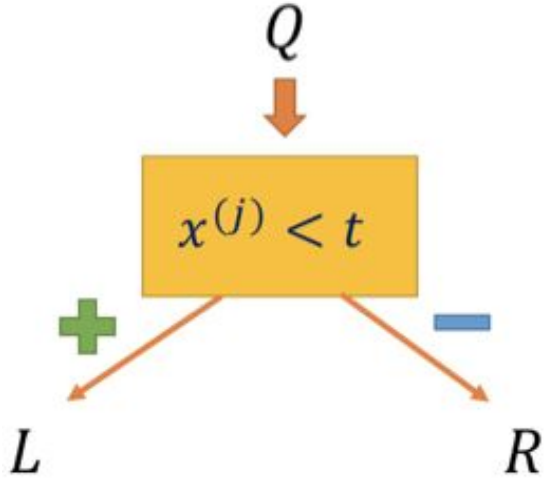
$$x^{(j)} < t$$

1- Make Split

2- Repeat

3- Repeat

Decision Trees – How to Split Properly?



$$G(j,t) = \frac{|L|}{|Q|} H(L) + \frac{|R|}{|Q|} H(R) \to \min_{j,t}$$

Decision Trees – Information criteria

Entropy Criteria :

$$H(R) = -\sum_k p_k \log_2 p_k$$

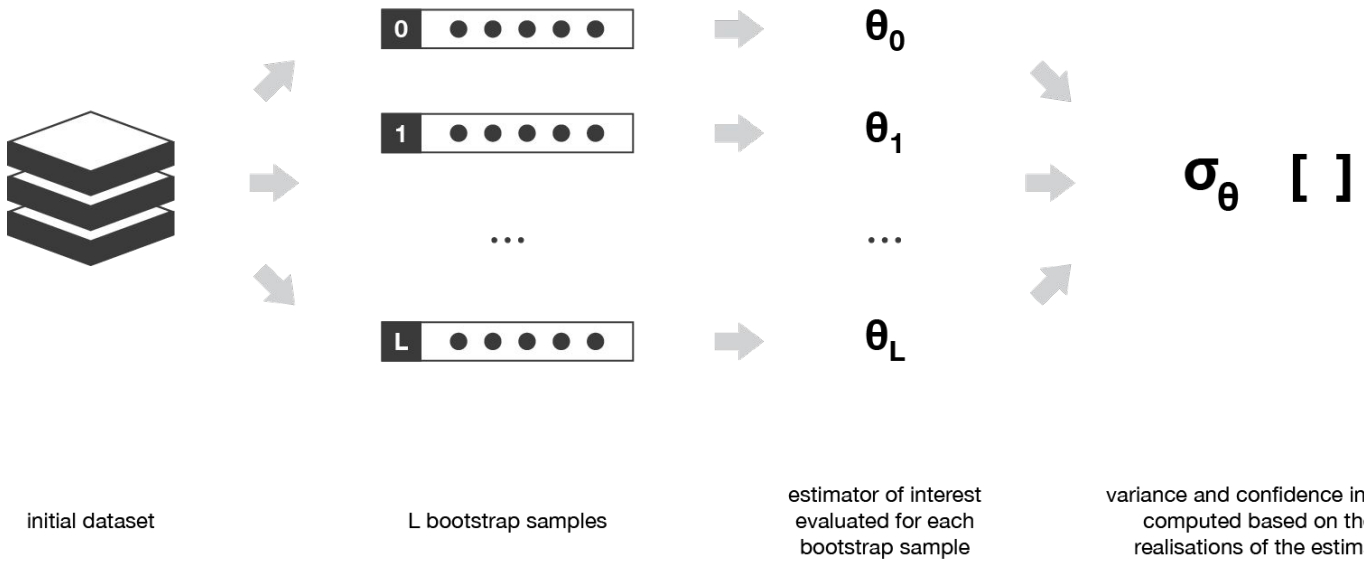
Gini impurity :

$$H(R) = 1 - \sum_k (p_k)^2$$

Ensembles: Bootstrap

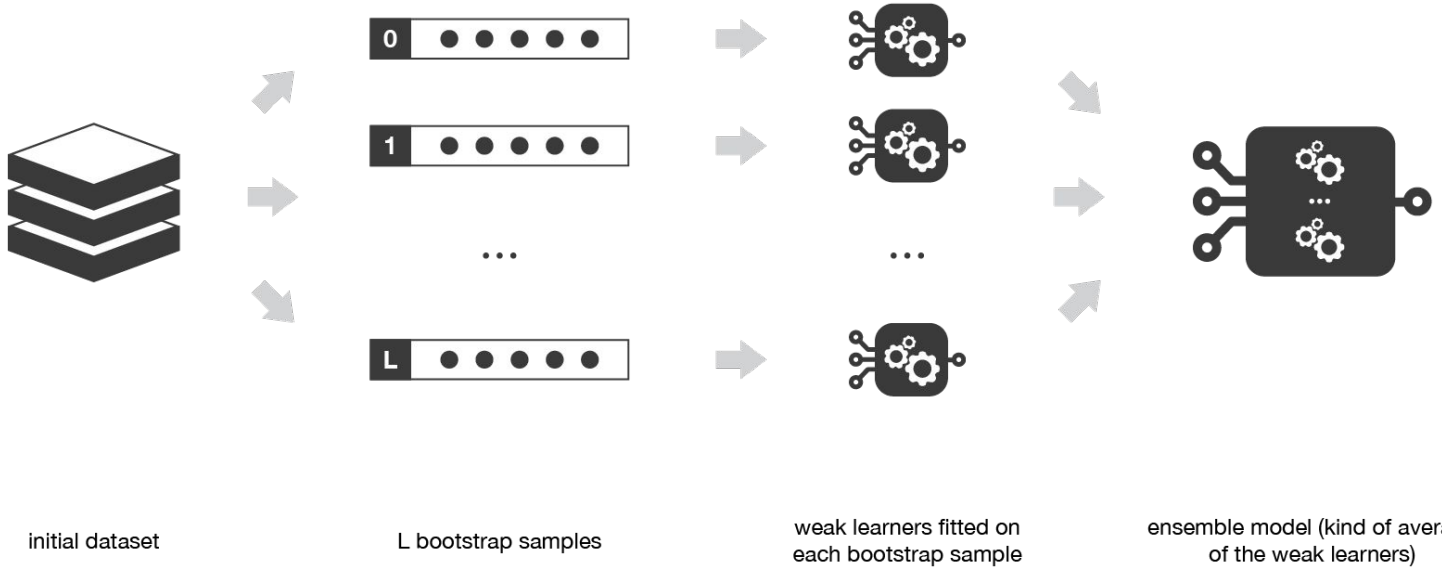
https://towardsdatascience.com/ensemble-methods-bagging-boosting-and-stacking-c9214a10a205

Ensembles: Bootstrap



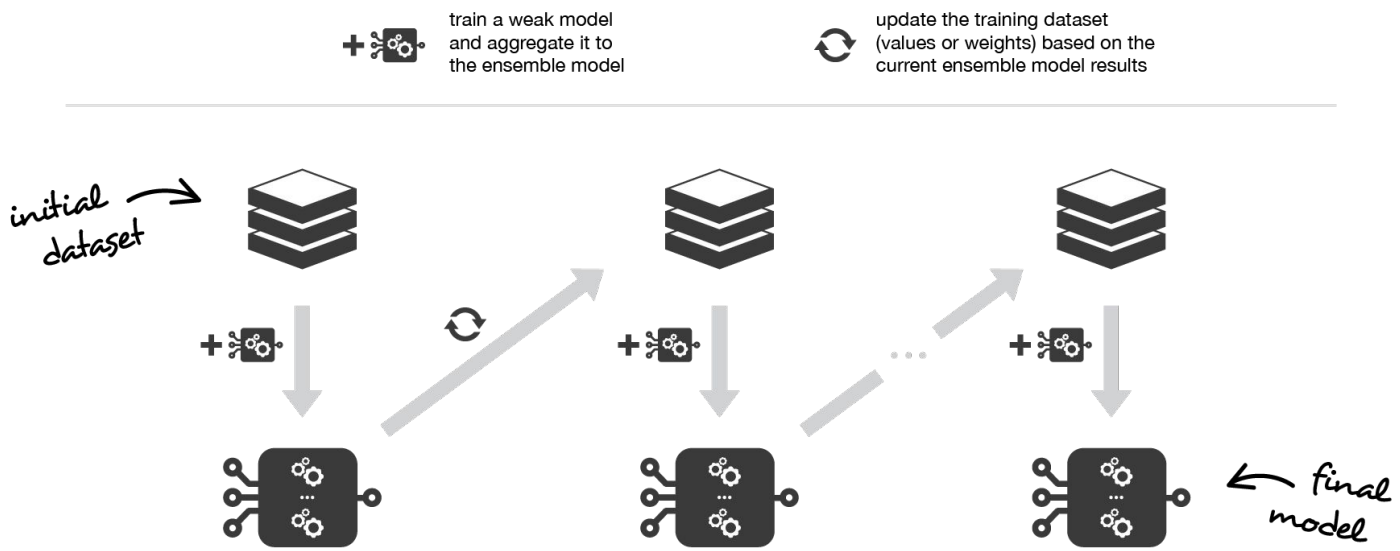
https://towardsdatascience.com/ensemble-methods-bagging-boosting-and-stacking-c9214a10a205

Ensembles: Bagging



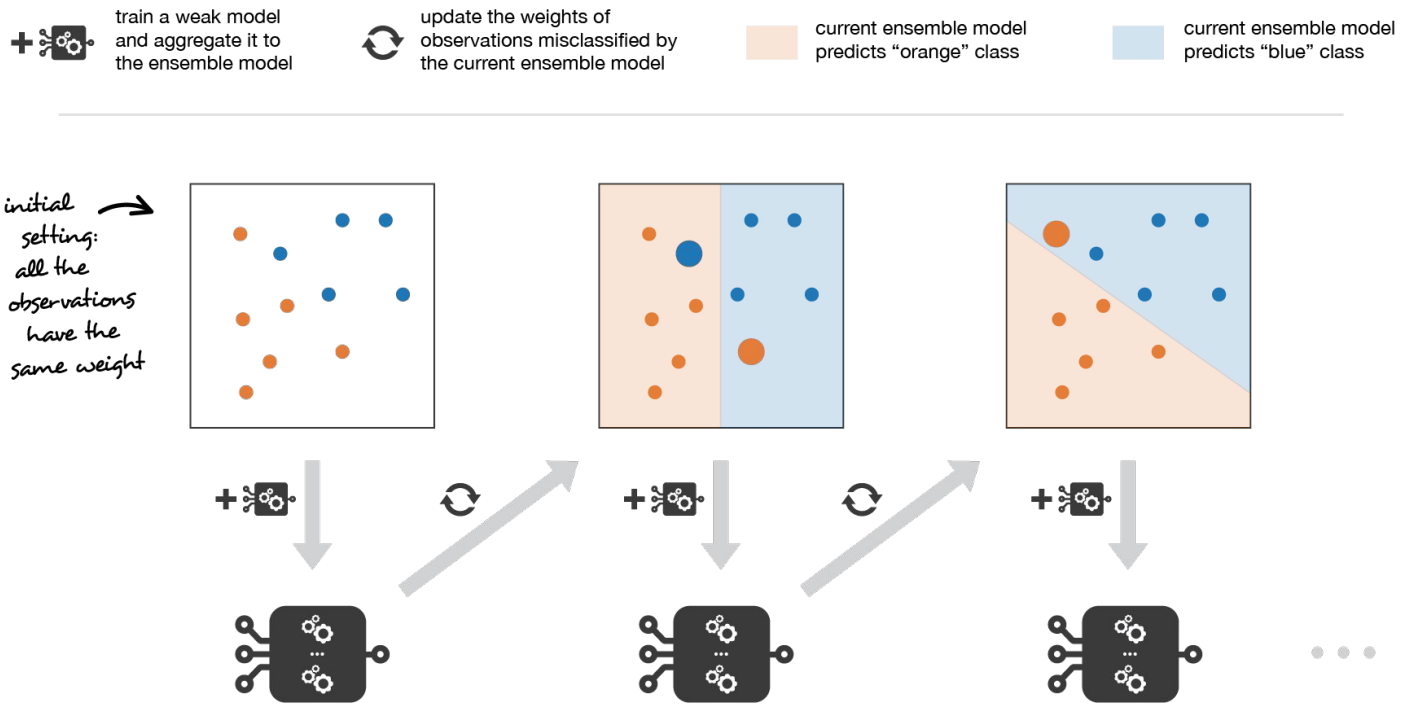
https://towardsdatascience.com/ensemble-methods-bagging-boosting-and-stacking-c9214a10a205

Ensembles : Boosting



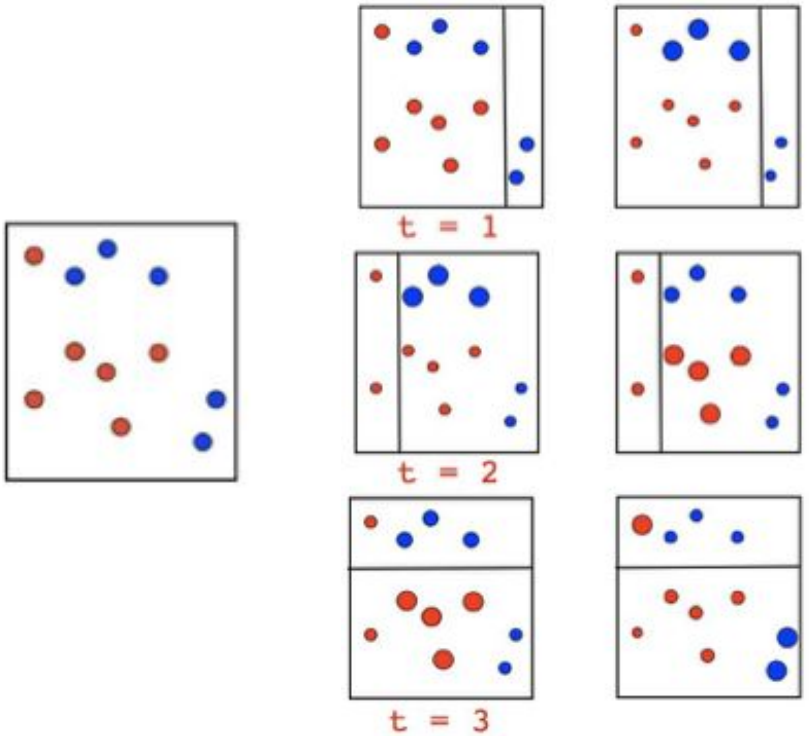
https://towardsdatascience.com/ensemble-methods-bagging-boosting-and-stacking-c9214a10a205

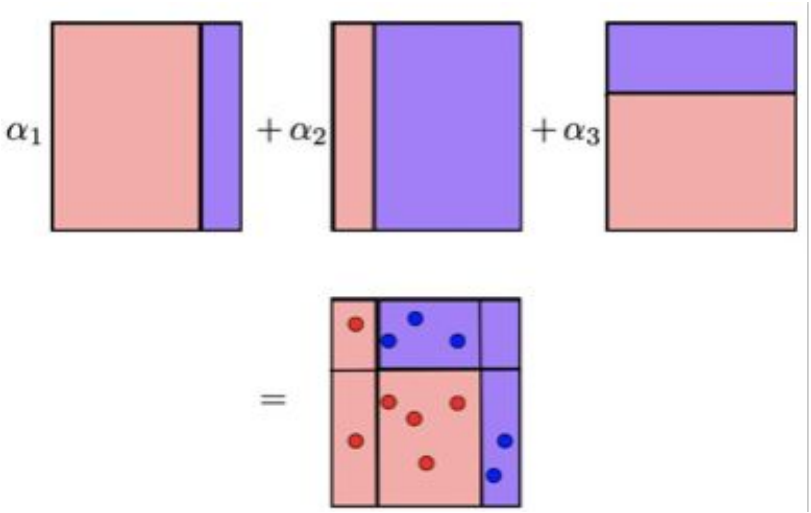
Ensembles : Boosting



https://towardsdatascience.com/ensemble-methods-bagging-boosting-and-stacking-c9214a10a205

Ensembles : Boosting



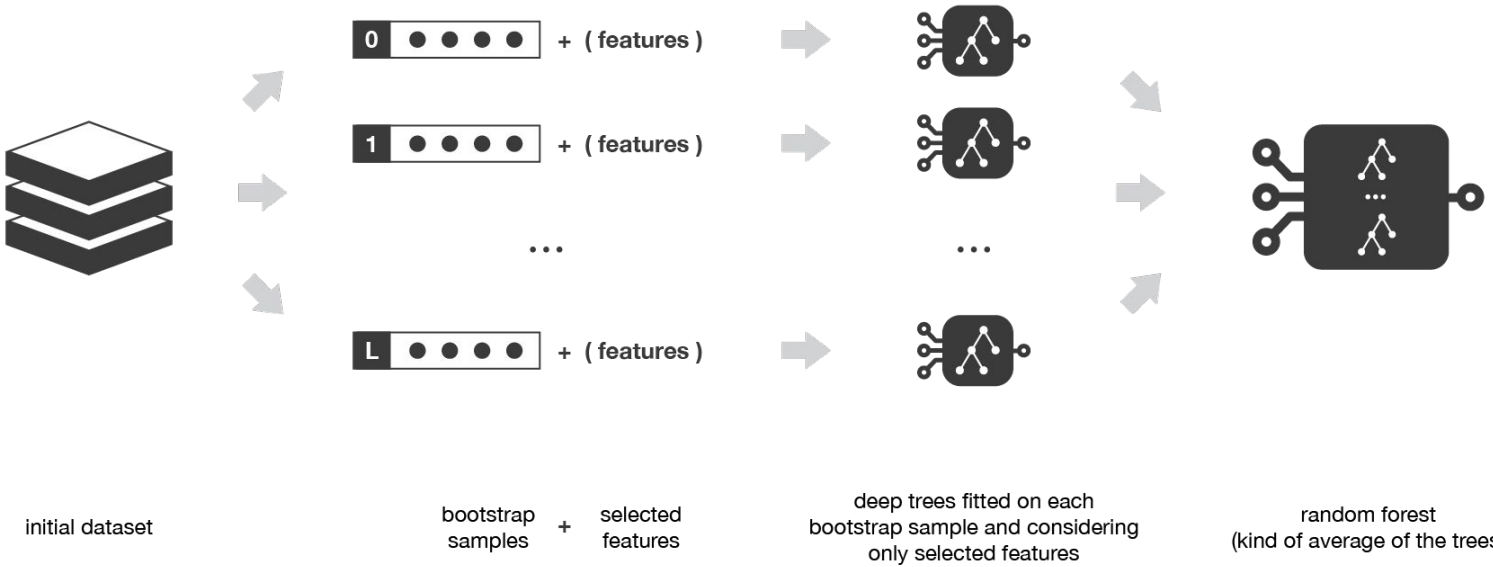


https://towardsdatascience.com/ensemble-methods-bagging-boosting-and-stacking-c9214a10a205

Ensembles: Stacking

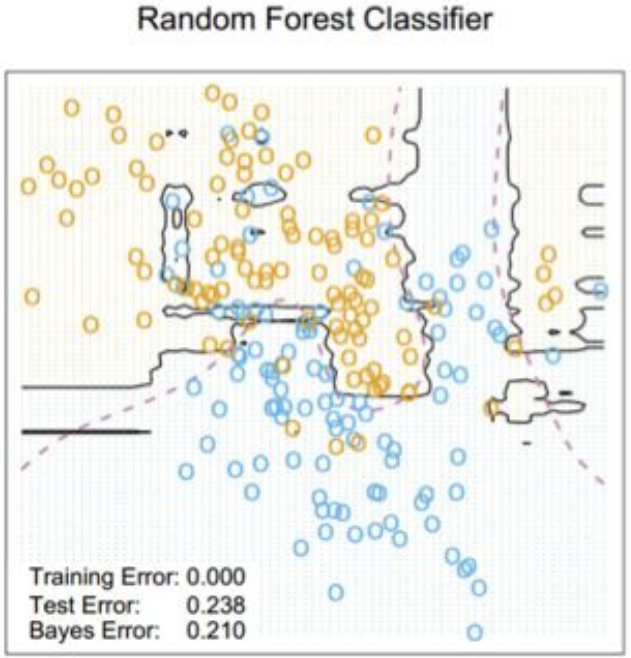
https://towardsdatascience.com/ensemble-methods-bagging-boosting-and-stacking-c9214a10a205

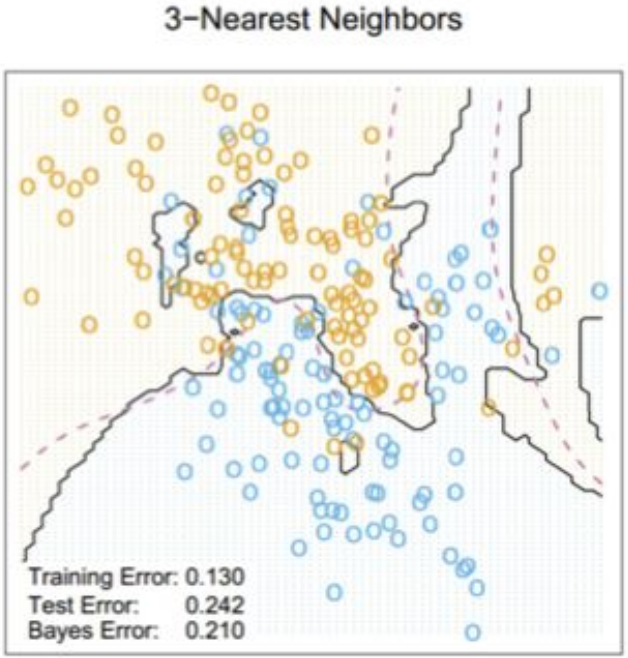
Ensembles Example: Random Forest



https://towardsdatascience.com/ensemble-methods-bagging-boosting-and-stacking-c9214a10a205

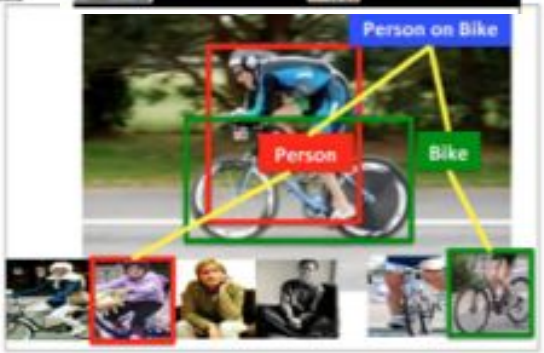
Ensembles Example: Random Forest



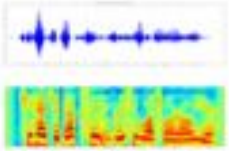


https://towardsdatascience.com/ensemble-methods-bagging-boosting-and-stacking-c9214a10a205

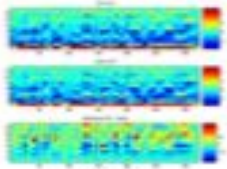
Deep Learning Applications



Audio Features



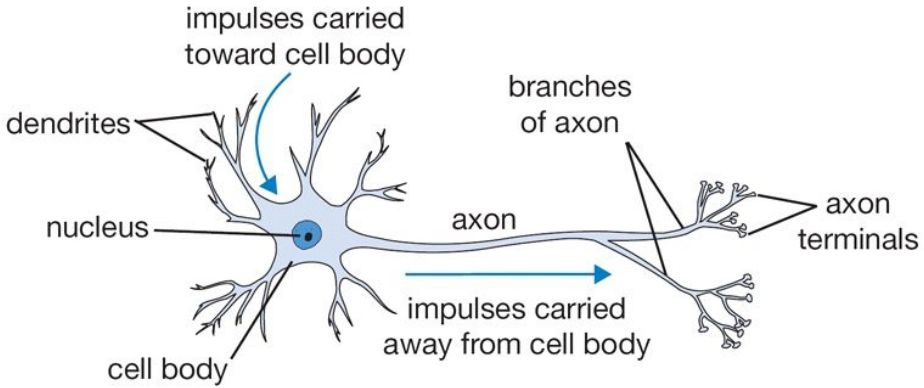
Spectrogram

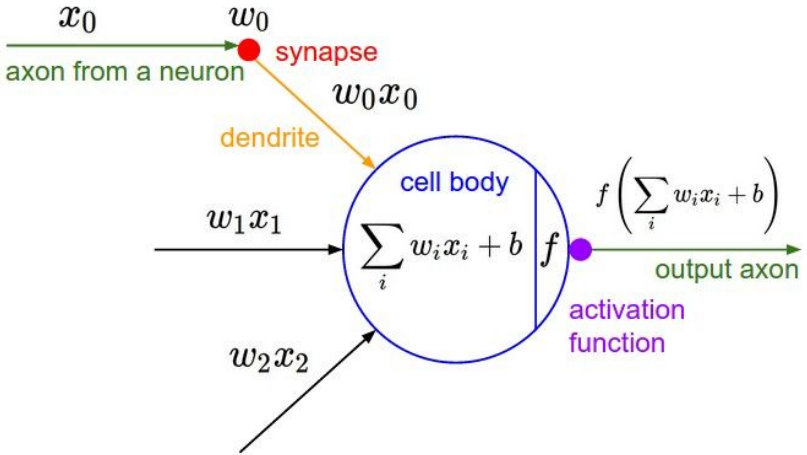


MFCC

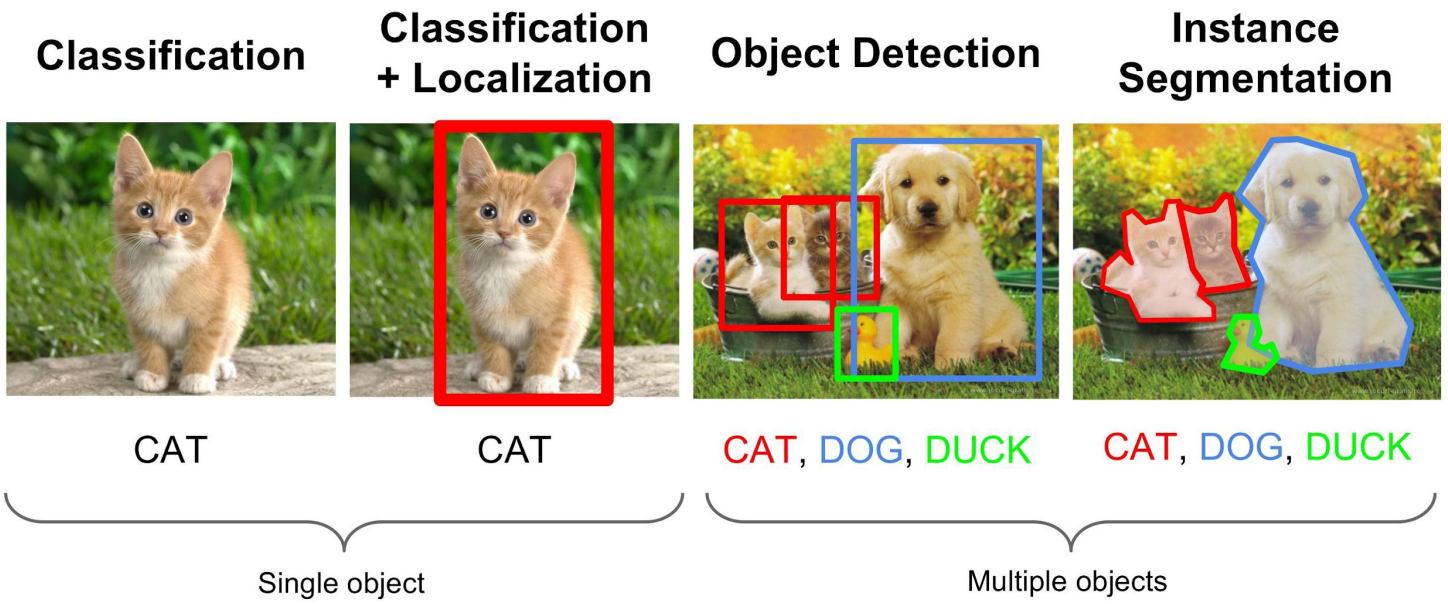
"man in black shirt is playing guitar."

Biological Motivation

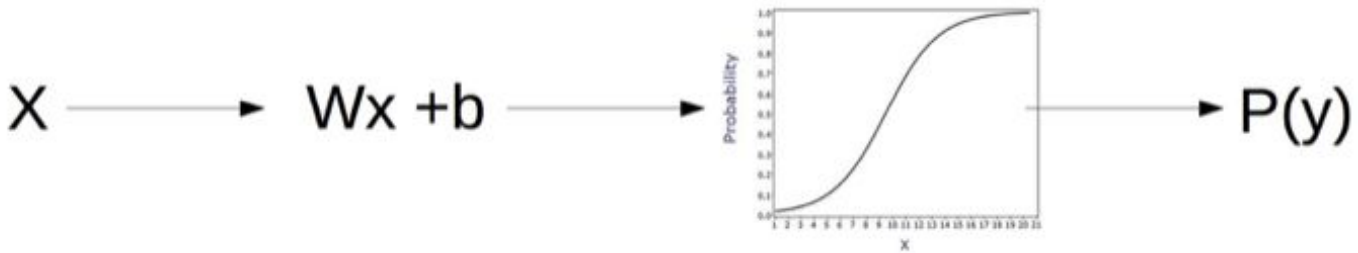




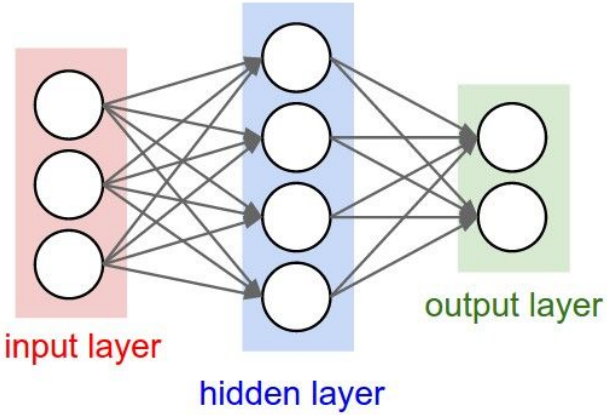
Deep Learning Applications

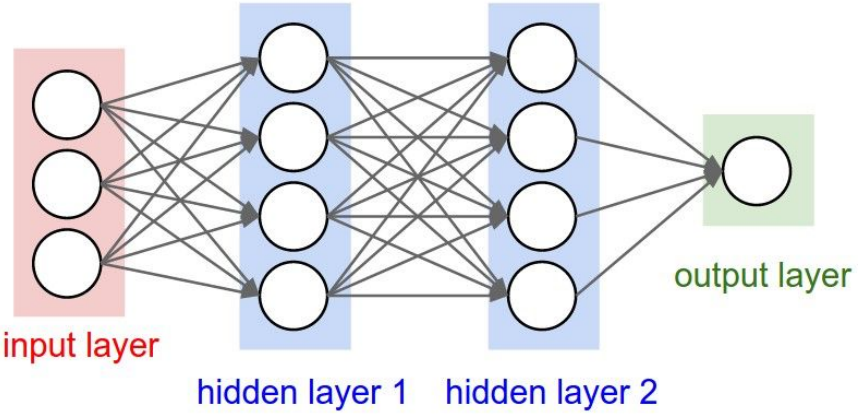


Logistic Regression

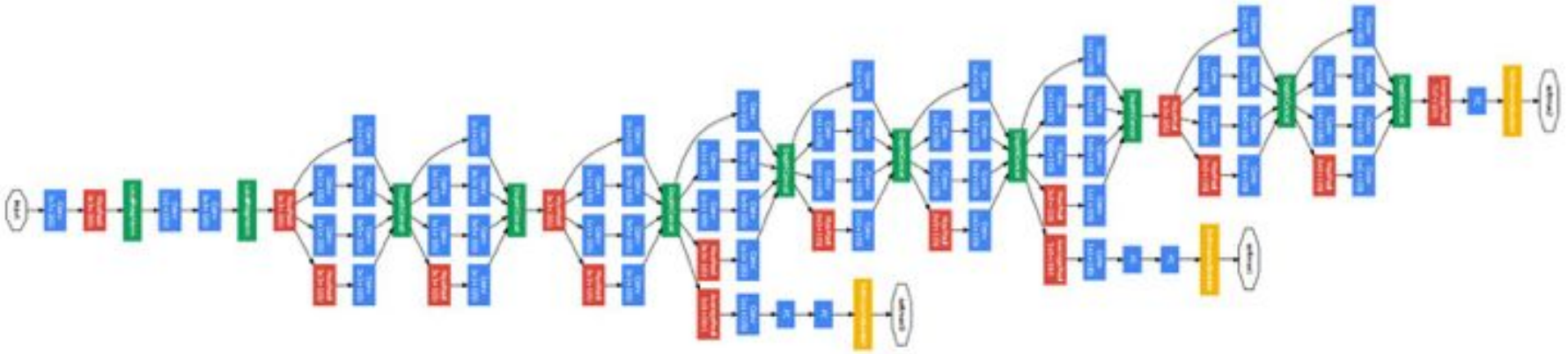


Neural Network Architectures

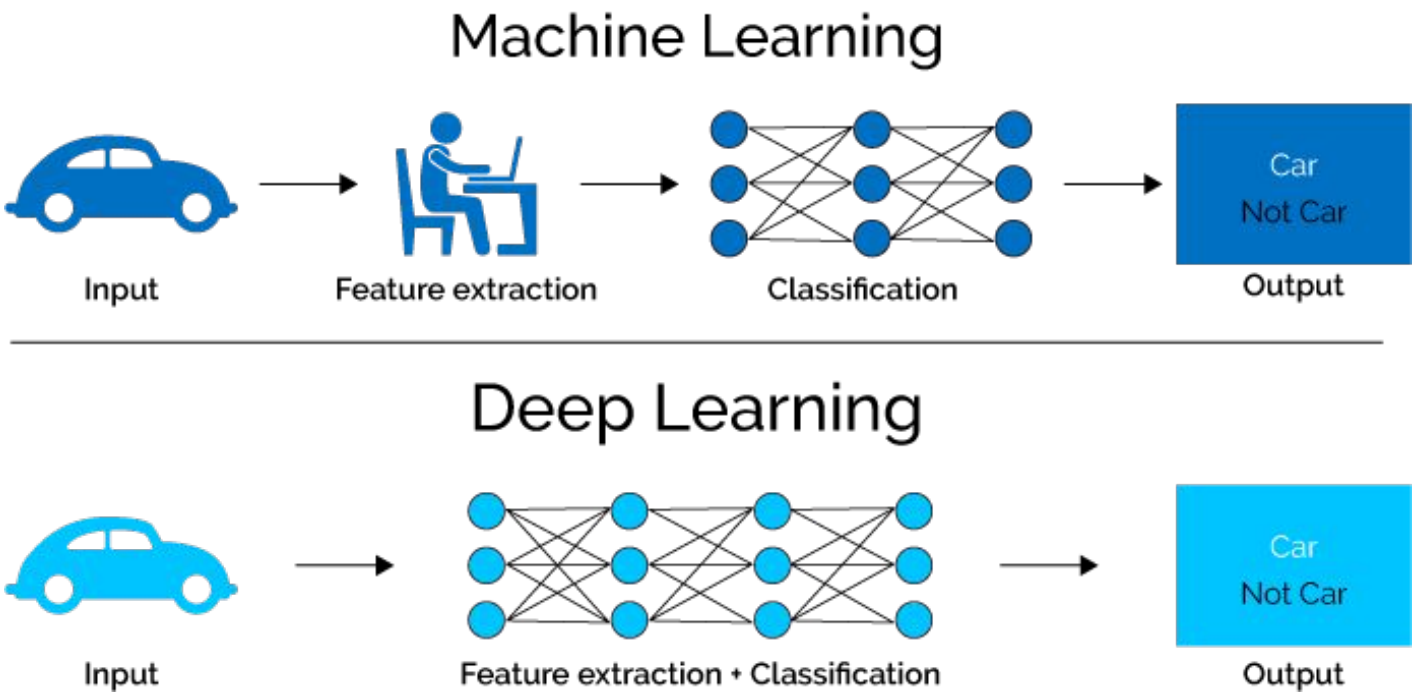




Neural Network Architectures



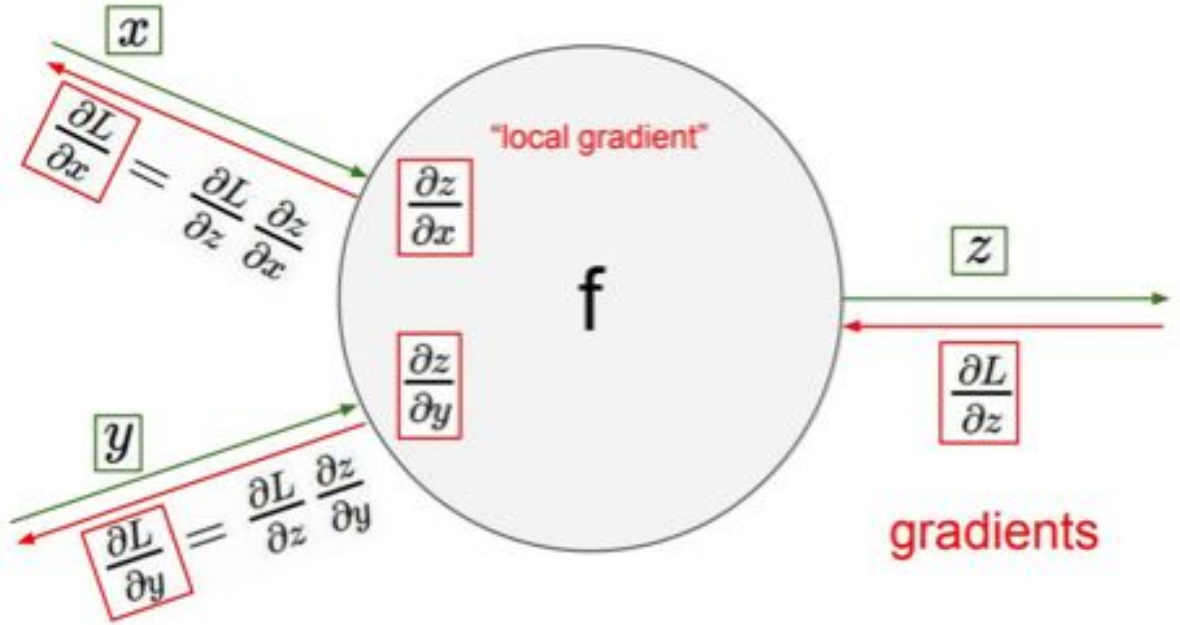
ML vs DL

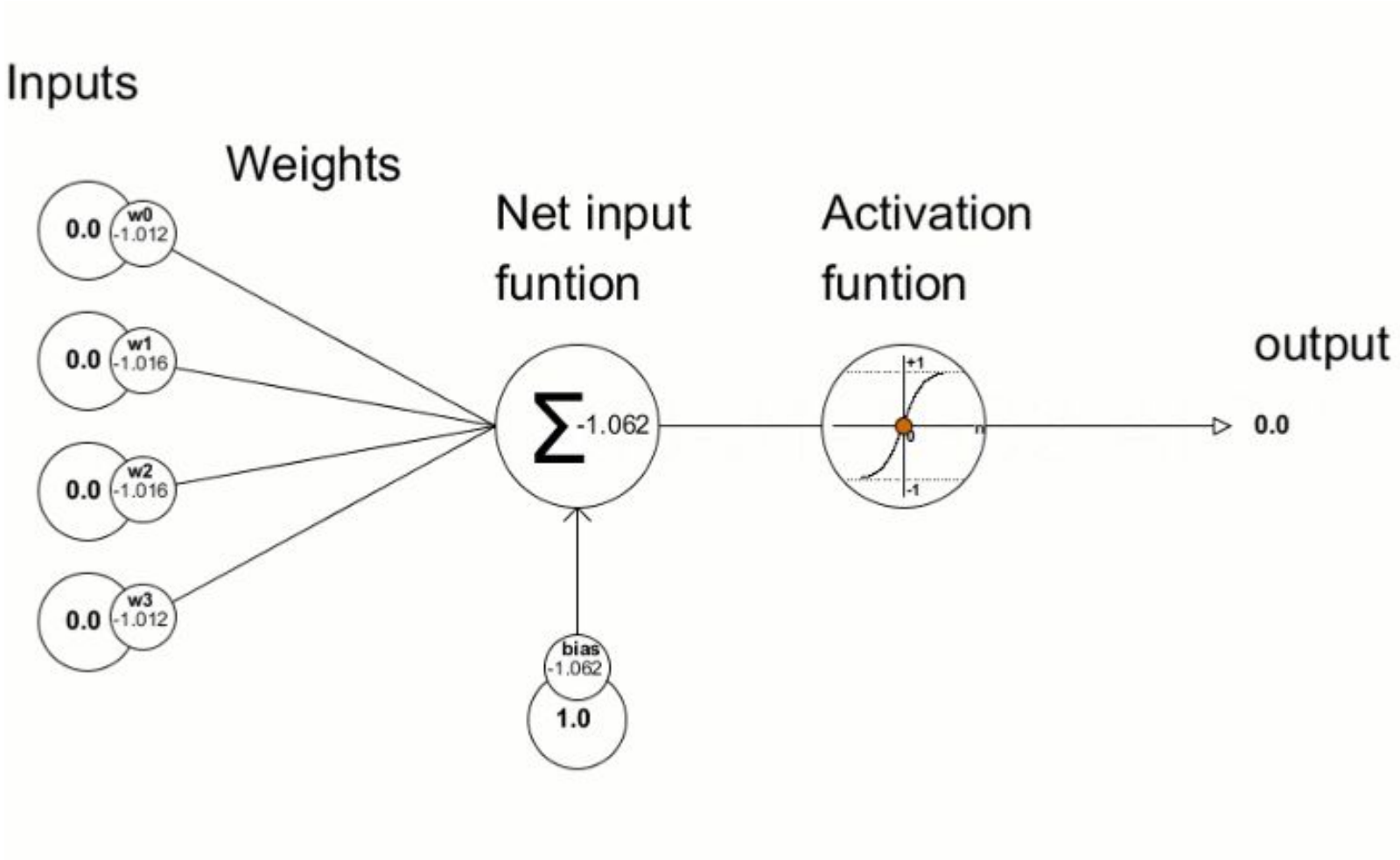

Machine Learning
Input
Feature extraction
Classification
Car
Not Car
Output
Deep Learning
Input
Feature extraction + Classification
Car
Not Car
Output

Backpropagation

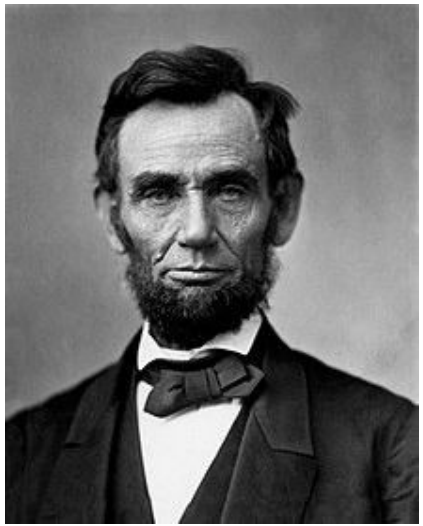
$$\frac{\partial L}{\partial x} = \frac{\partial L}{\partial z}\frac{\partial z}{\partial x}$$

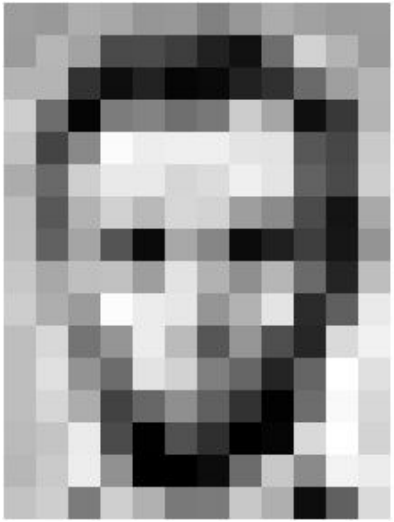
Backpropagation




Inputs
Weights
Net input funtion
Activation funtion
output
0.0
w0 -1.012
0.0
w1 -1.016
0.0
w2 -1.016
0.0
w3 -1.012
Σ -1.062
+1
0
n
-1
0.0
bias -1.062
1.0

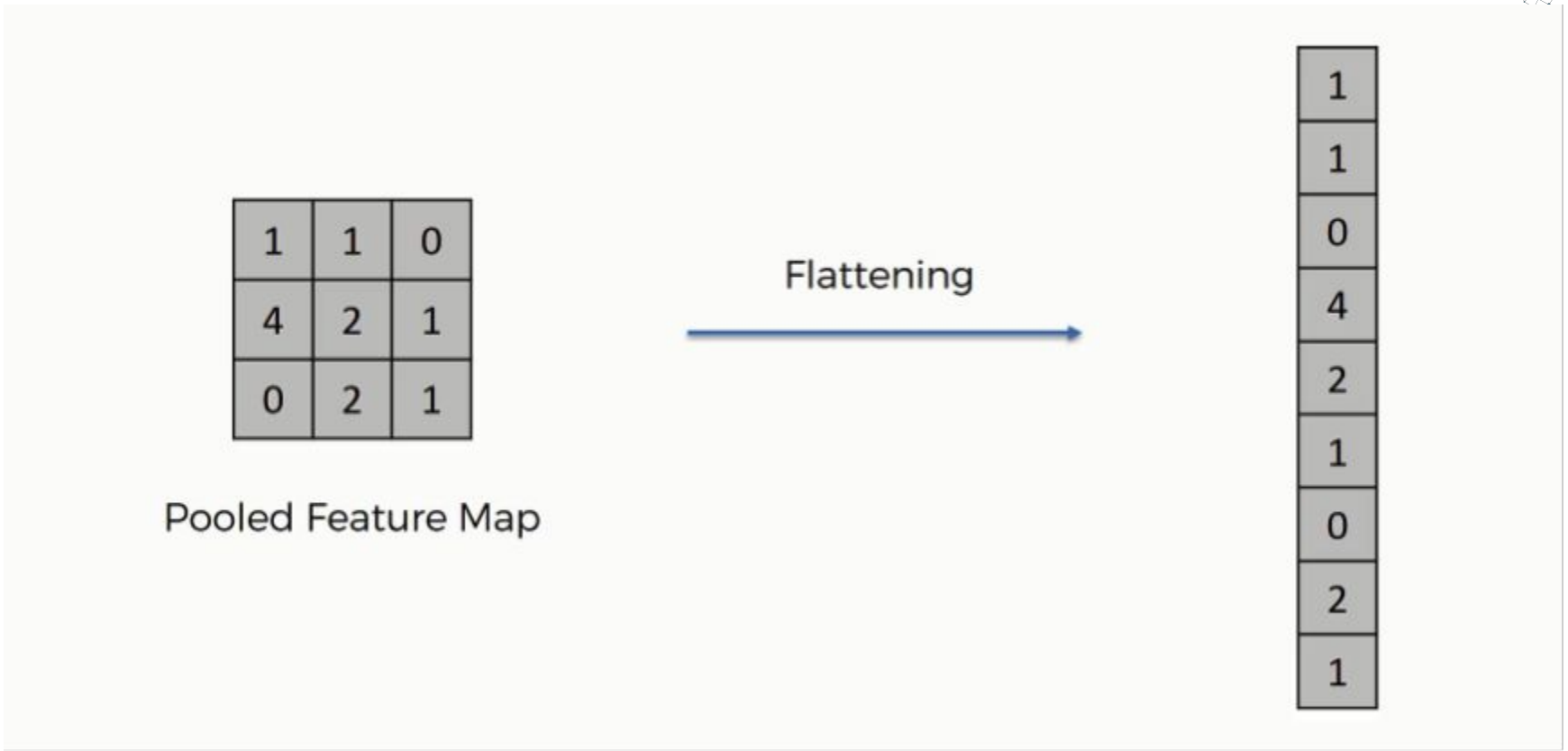
Neural Networks with Images

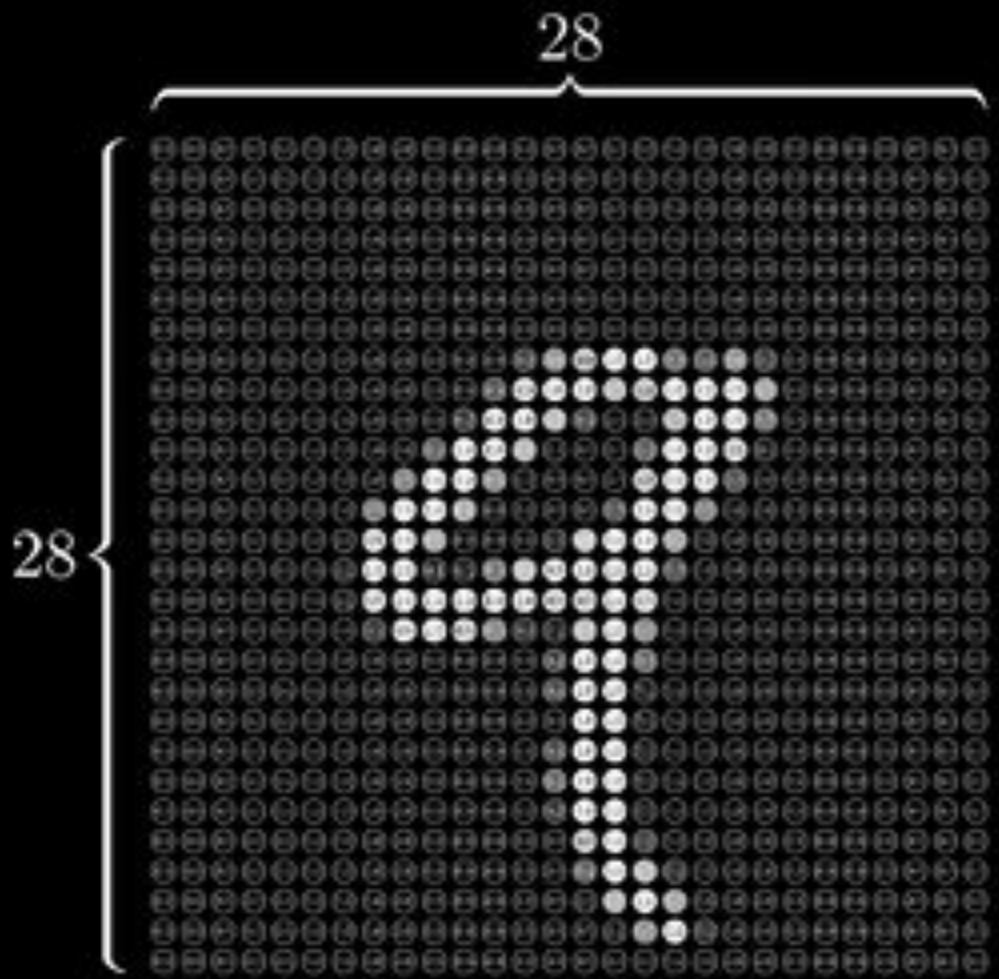




157	153	174	168	150	152	129	151	172	161	155	156
155	182	163	74	75	62	33	17	110	210	180	154
180	180	50	14	34	6	10	33	48	106	159	181
206	109	5	124	131	111	120	204	166	15	56	180
194	68	137	251	237	239	239	228	227	87	71	201
172	105	207	233	233	214	220	239	228	98	74	206
188	88	179	209	185	215	211	158	139	75	20	169
189	97	165	84	10	168	134	11	31	62	22	148
199	168	191	193	158	227	178	143	182	106	36	190
205	174	155	252	236	231	149	178	228	43	95	234
190	216	116	149	236	187	86	150	79	38	218	241
190	224	147	108	227	210	127	102	36	101	255	224
190	214	173	66	103	143	96	50	2	109	249	215
187	196	235	75	1	81	47	0	6	217	255	211
183	202	237	145	0	0	12	108	200	138	243	236
195	206	123	207	177	121	123	200	175	13	96	218

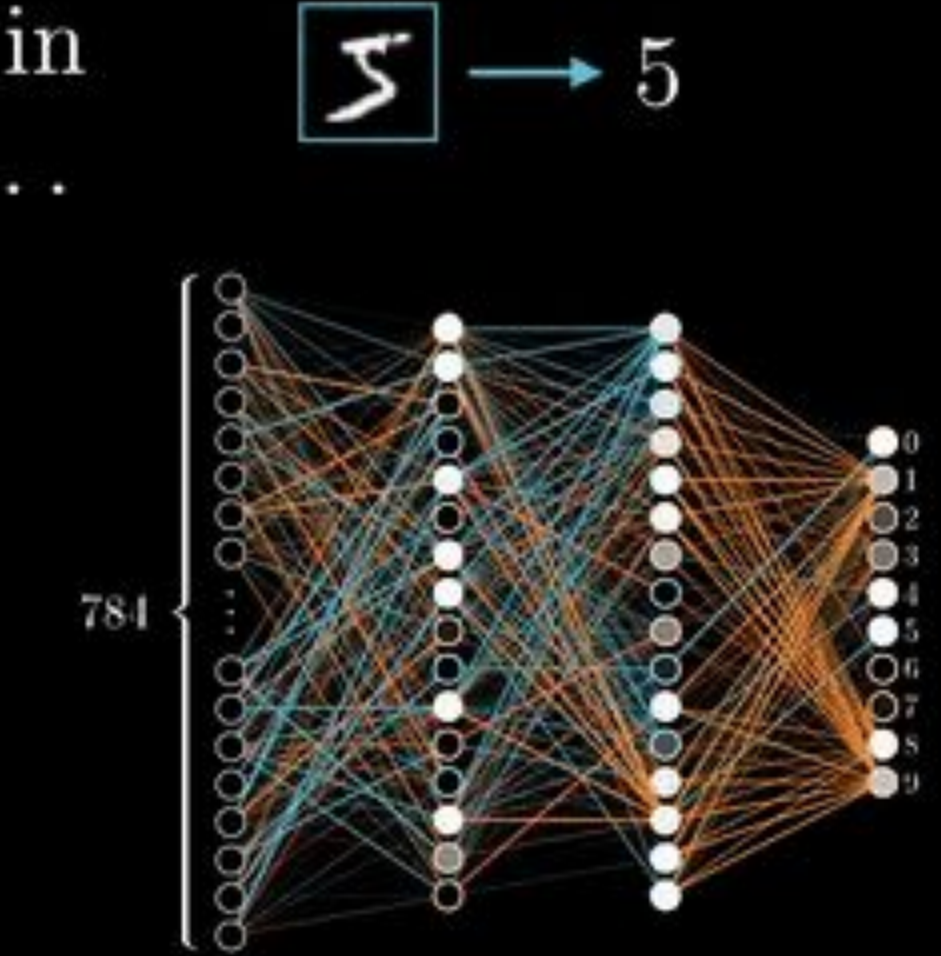
157	153	174	168	150	152	129	151	172	161	155	156
155	182	163	74	75	62	33	17	110	210	180	154
180	180	50	14	34	6	10	33	48	106	159	181
206	109	5	124	131	111	120	204	166	15	56	180
194	68	137	251	237	239	239	228	227	87	71	201
172	105	207	233	233	214	220	239	228	98	74	206
188	88	179	209	185	215	211	158	139	75	20	169
189	97	165	84	10	168	134	11	31	62	22	148
199	168	191	193	158	227	178	143	182	106	36	190
205	174	155	252	236	231	149	178	228	43	95	234
190	216	116	149	236	187	86	150	79	38	218	241
190	224	147	108	227	210	127	102	36	101	255	224
190	214	173	66	103	143	96	50	2	109	249	215
187	196	235	75	1	81	47	0	6	217	255	211
183	202	237	145	0	0	12	108	200	138	243	236
195	206	123	207	177	121	123	200	175	13	96	218

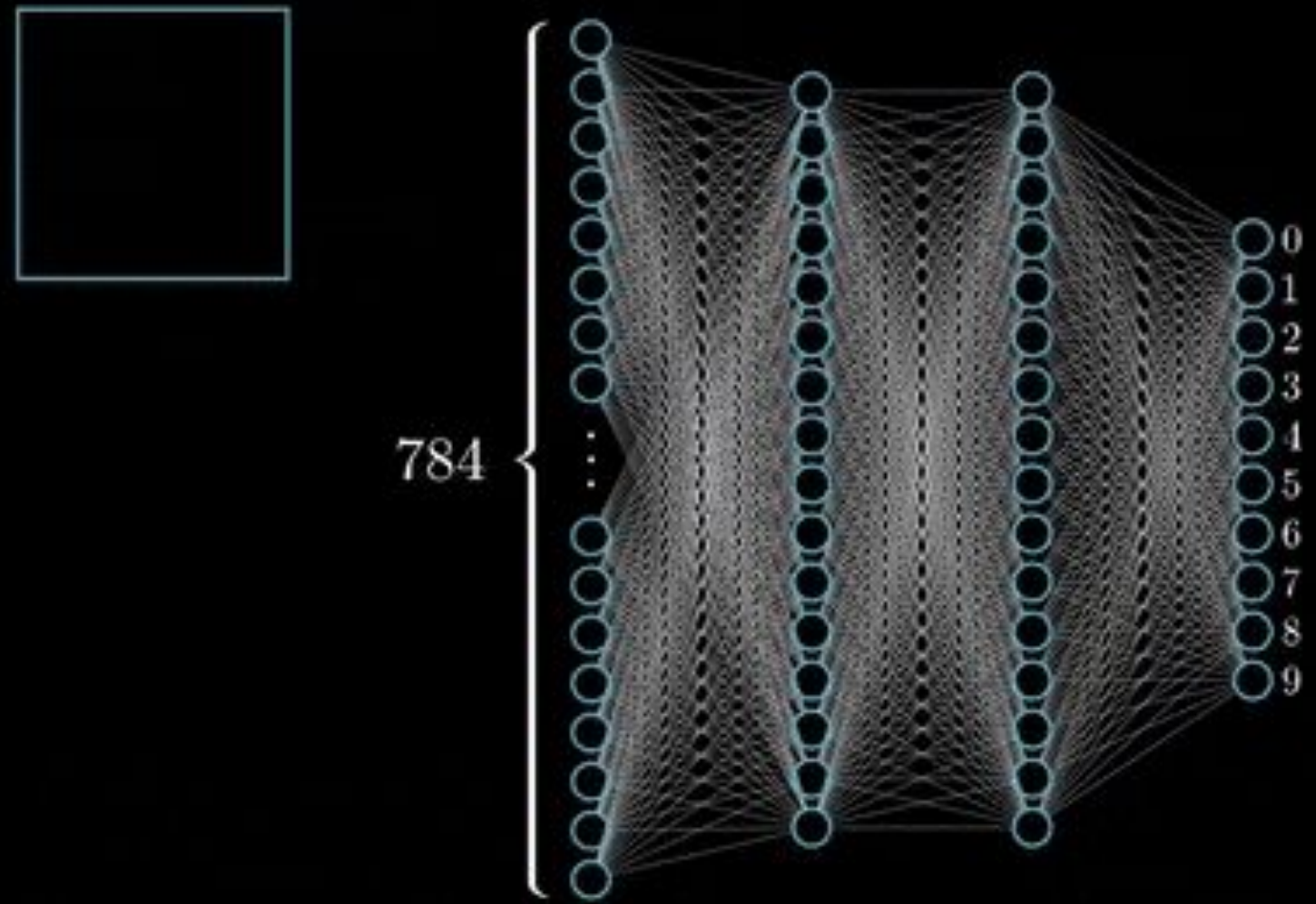

1 1 0
4 2 1
0 2 1
Pooled Feature Map
Flattening
1
1
0
4
2
1
0
2
1



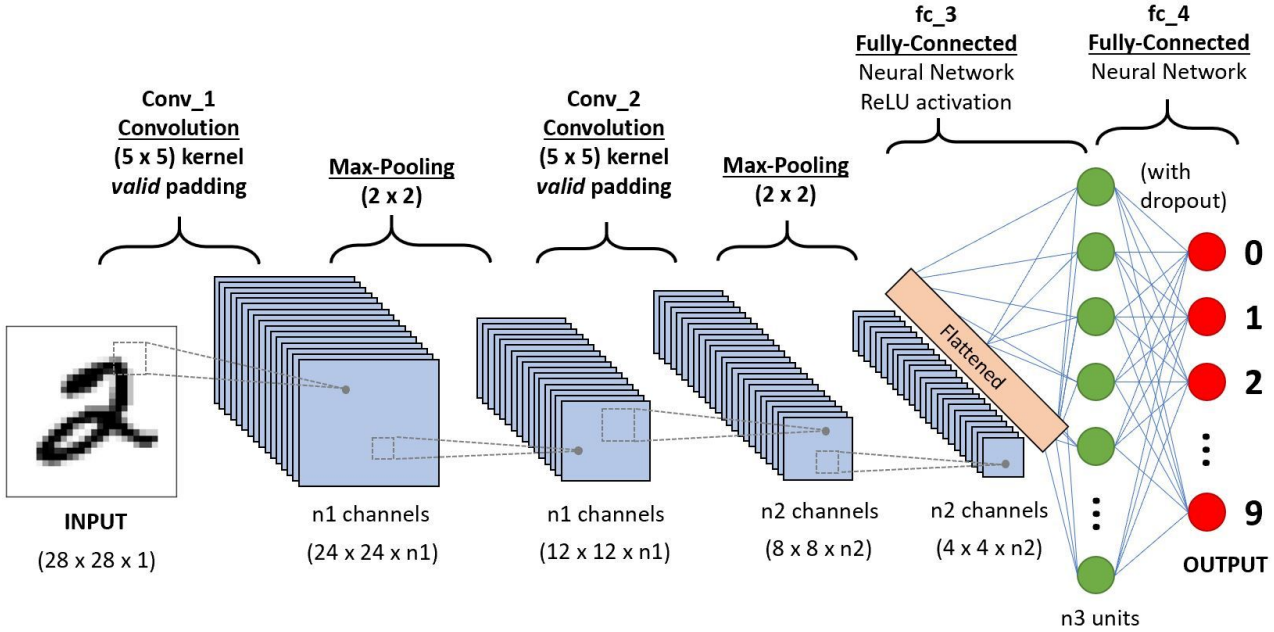
$$28 \times 28 = 784$$

Training in progress...




784
0
1
2
3
4
5
6
7
8
9

Convolutional Neural Networks - CNN



Interactive Exercise

Break

What is Big Data?

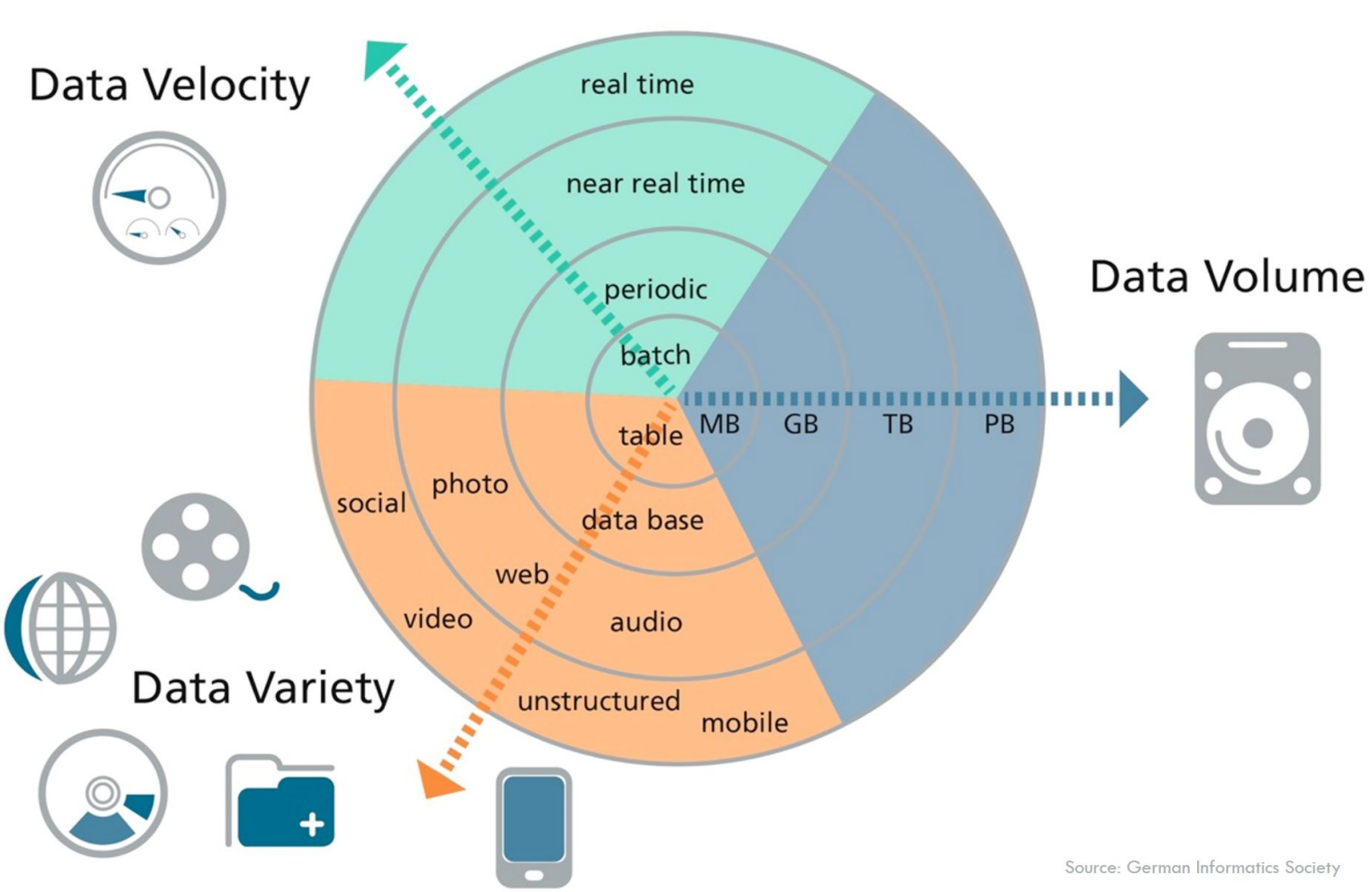
The six Vs of big data

Big data is a collection of data from various sources, often characterized by what's become known as the 3Vs: *volume, variety and velocity*. Over time, other Vs have been added to descriptions of big data:

VOLUME	VARIETY	VELOCITY	VERACITY	VALUE	VARIABILITY
The amount of data from myriad sources.	The types of data: structured, semi-structured, unstructured.	The speed at which big data is generated.	The degree to which big data can be trusted.	The business value of the data collected.	The ways in which the big data can be used and formatted.





Source: German Informatics Society

Big Data in Aviation

1. Encourage loyalty: United Airlines

Tailor-made offers will always appeal to the customer, thus encouraging loyalty. Airlines are in the fortunate position of being able to learn an enormous amount about their client base from data. Even a single booking contains data which can teach an airline a huge amount about its customers. **For instance, United Airlines use their “collect, detect, act” protocol to analyze over 150 variables in each customer profile.** These analyses measure everything from previous purchases to customer preferences in order to generate a tailor-made offer. The collect, detect, act initiative has increased United’s year-to-year revenue by over 15%.

3. Pricing Strategy: EasyJet

Many airlines go a step further than basic data collection. **With new technology, it is possible for companies to analyze big data accumulated from purchase activity to demand patterns.** For instance, if an airline sees the demand for a certain route increasing, they can adjust prices accordingly. From this information, the airline can also identify which customer segments are price sensitive, and determine a segment's price range for a given route.

A related big data case study comes from EasyJet. **EasyJet invested in an artificially intelligent algorithm that determines seat pricing automatically, depending on demand.** Furthermore, the system can also analyze historical data to predict demand patterns up to a year in advance. These analytics can also impact future decision-making about new routes, schedules, and codeshare alliances.

Examples of AI in Aviation

Face Recognition

- Biometric Face Recognition
- AirAsia
  a. https://www.youtube.com/watch?v=7dWdXZ7o_Bc
- Heathrow - British Airways
  a. https://www.youtube.com/watch?v=V78OOIbDe2c

Baggages

- **Bag Tag OCR**

  Bag Tag OCR AI reads text from images of luggage labels to enhance read rates during sortation and act as a failsafe when barcode reading systems fail. This is currently deployed in a major European hub airport to enhance remote processing of baggage.

- **Bag Classification**

  Through AI-based computer vision, classification and detection we have deployed systems that are able to automate luggage handling through classification as suitable for robotic loading to enable routing of bags that can be handled in this way to manual lines. Similar assessment to determine if a bag has straps and should be placed in a tub has also been assessed

Predictive Analytics

- Predictive analytics to enable the Digital Twin of the airport to enhance operations management team performance.
    a. Car park occupancy in 4 hours' time.
    b. The number of people present in an area; the gender and ethnicity ratios in a given population; or the remaining resource given the current situation.