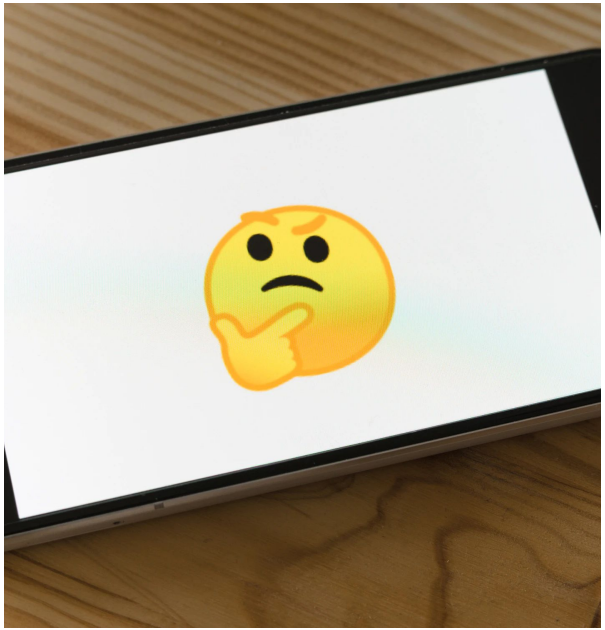
Intro to Machine Learning

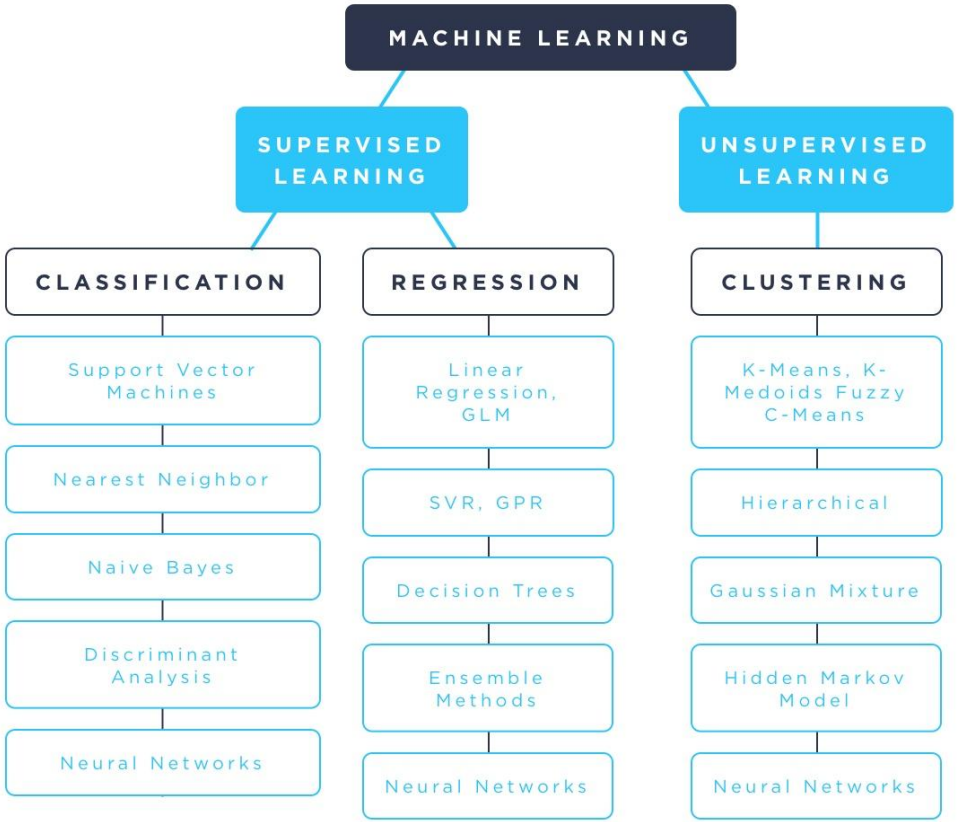
YOUR COURSE NAME | IRONHACK



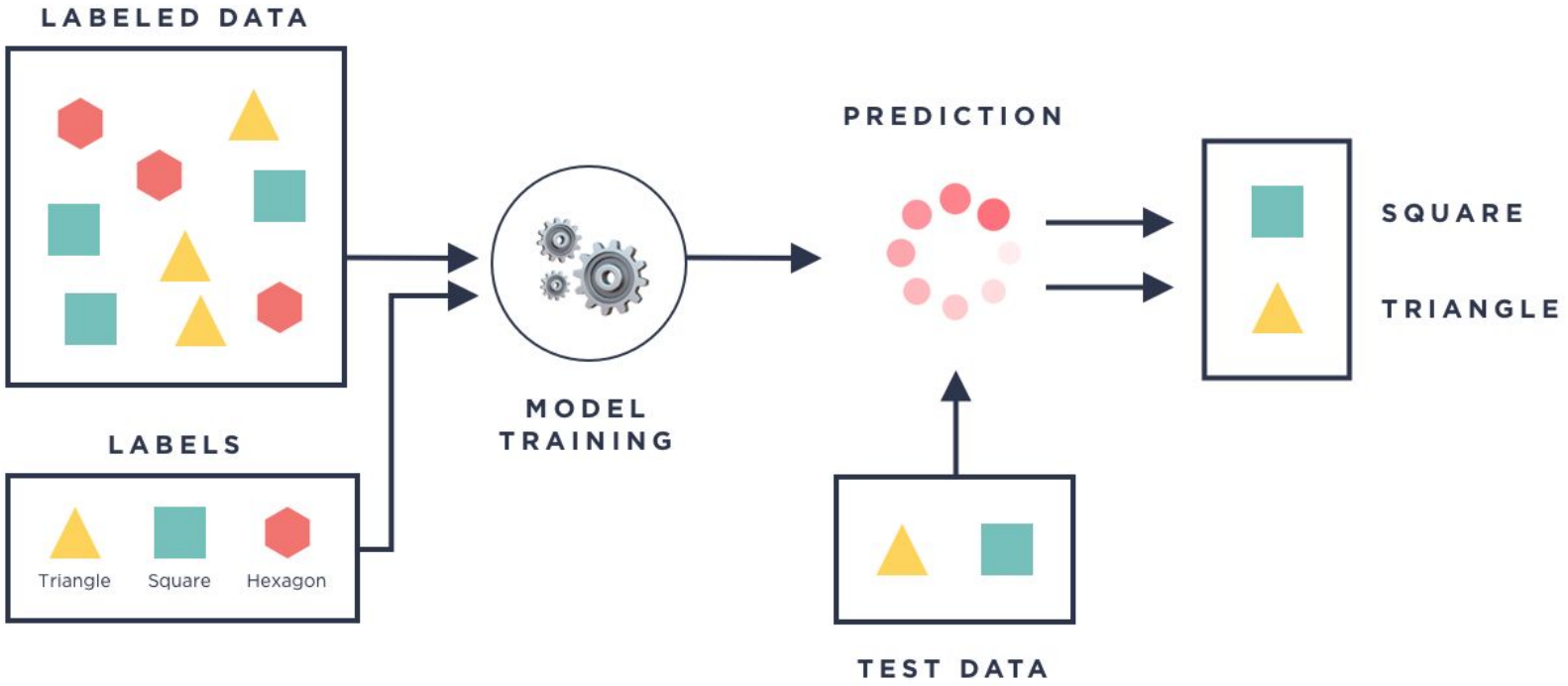
Credit: Unsplash

Agenda

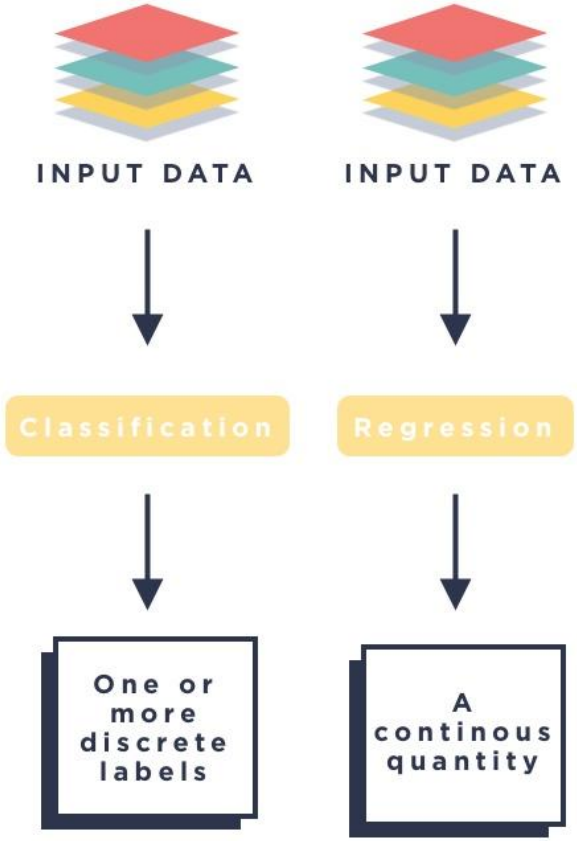
1. Machine Learning
2. Supervised Learning
3. Unsupervised Learning


MACHINE LEARNING
SUPERVISED LEARNING
UNSUPERVISED LEARNING
CLASSIFICATION
REGRESSION
CLUSTERING
Support Vector Machines
Nearest Neighbor
Naive Bayes
Discriminant Analysis
Neural Networks
Linear Regression, GLM
SVR, GPR
Decision Trees
Ensemble Methods
Neural Networks
K-Means, K-Medoids Fuzzy C-Means
Hierarchical
Gaussian Mixture
Hidden Markov Model
Neural Networks

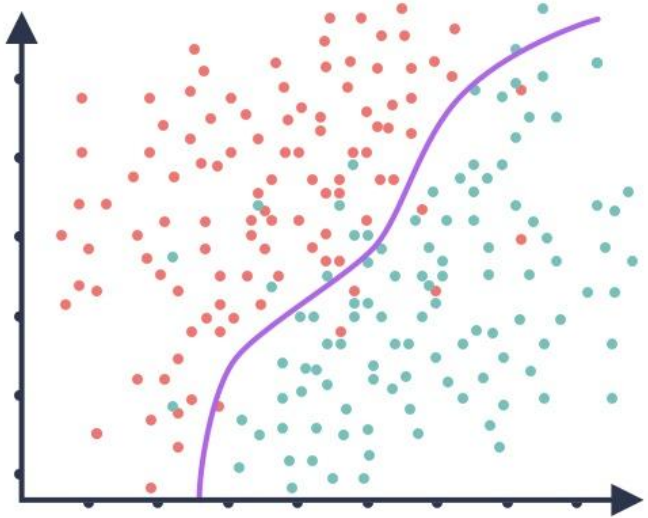
Supervised Learning.


LABELED DATA
LABELS
Triangle
Square
Hexagon
MODEL TRAINING
PREDICTION
TEST DATA
SQUARE
TRIANGLE

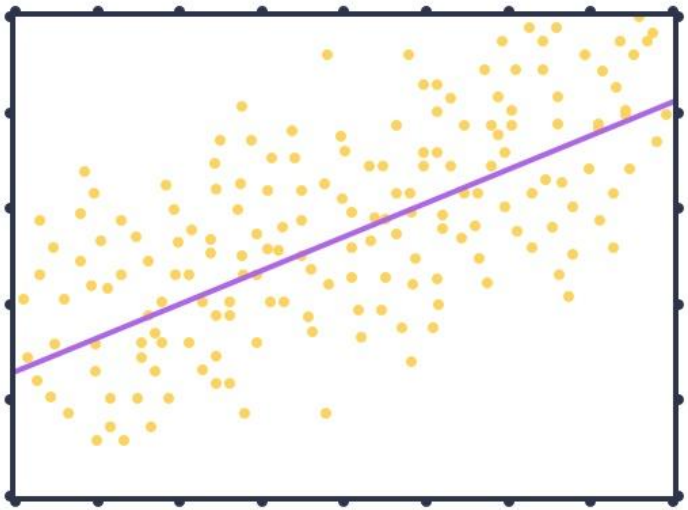
HV1	IC1	IC2	IC3	IC4	IC5	AVGGIFT	TARGET_D
2346	420	446	468	503	14552	15.5	21
497	350	364	357	384	11696	3.08	3
1229	469	502	507	544	17313	7.5	20
325	148	181	171	209	6334	6.7	5
768	174	201	220	249	7802	8.78571429	10
557	211	188	221	205	5550	13	16
2145	474	492	522	554	18340	11.5714286	15
2184	351	376	394	419	16480	12.5	20
1442	369	394	445	488	26462	7.84615385	10
1708	437	586	551	684	29098	9.76923077	20
1054	584	644	652	726	26074	13.5384615	20
1062	486	550	555	584	17908	15.3333333	20
849	457	508	470	519	16386	12.8	25
213	222	273	283	329	12227	5.125	5
574	289	318	315	363	11250	3.55555556	4
2506	449	455	501	517	16302	8.875	50
622	347	378	401	416	15808	15	25
764	272	361	346	424	16257	7.91304348	15
681	335	398	356	419	14011	30.75	51


INPUT DATA
INPUT DATA
Classification
Regression
One or more discrete labels
A continous quantity

What is the difference between…



CLASSIFICATION



REGRESSION

Unsupervised Learning.

WHAT IS UNSUPERVISED LEARNING?

- learning in which we don't provide the computer sets of labelled data. **We don't have a target variable**.
- We just give the computer a set of "rules" to find out the solution to the problem and let it work on it.
- we do not have a clear way to tell whether the algorithm is doing well or not - in SL, predictions are compared with true values (labels) - here there are no performance metrics with which to supervise the task.

WHAT?? No target?

Then what are we doing?!

Instead of telling the machine “Predict Y for our data X,” we’re asking
“What can you tell me about X?”

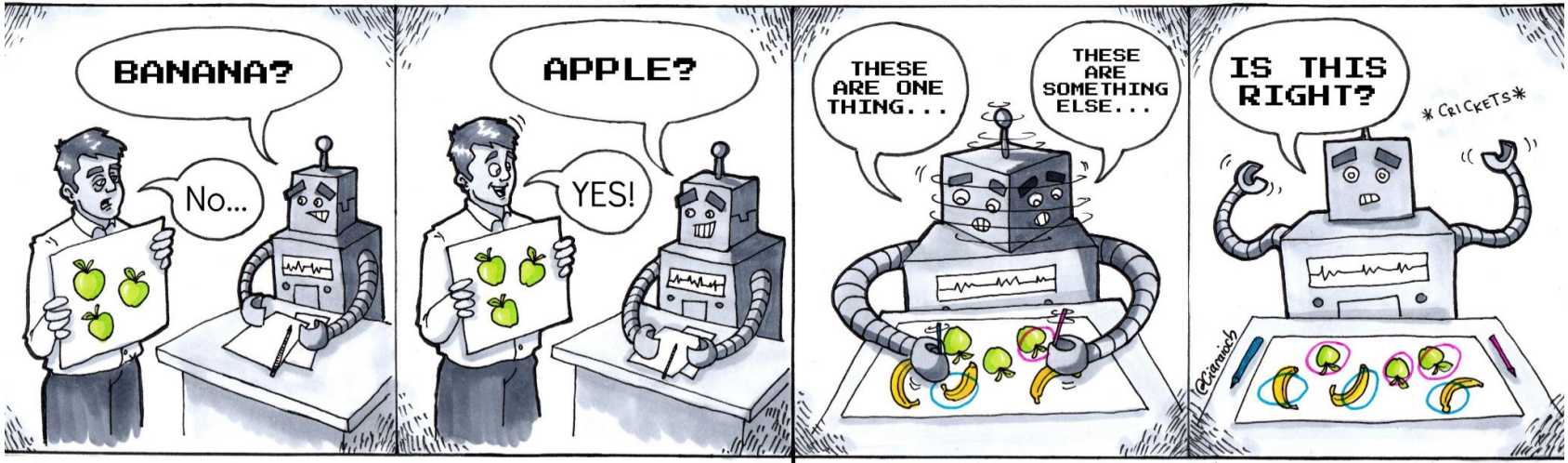
Things we ask the machine to tell us about X may be “What are the six best groups we can make out of X?”
Or
“What three features occur together most frequently in X?”

EXAMPLES OF UNSUPERVISED LEARNING

- **Clustering**: We ask the computer to make groups of observations based on the features.
- **Principal Component Analysis**: We look for another set of decorrelated features.
- **Anomaly detection**: We look for observations which have features clearly different from the others.
- **Generative modelling:** We let the computer learn patterns in order to generate fake ones.

The most popular task in UL is **Clustering**

- companies usually have big unlabeled datasets of customers.
- segment those customers to understand behaviour, target promotions or tailor new products
- Clustering algorithms find observations with characteristic similarities and group them accordingly

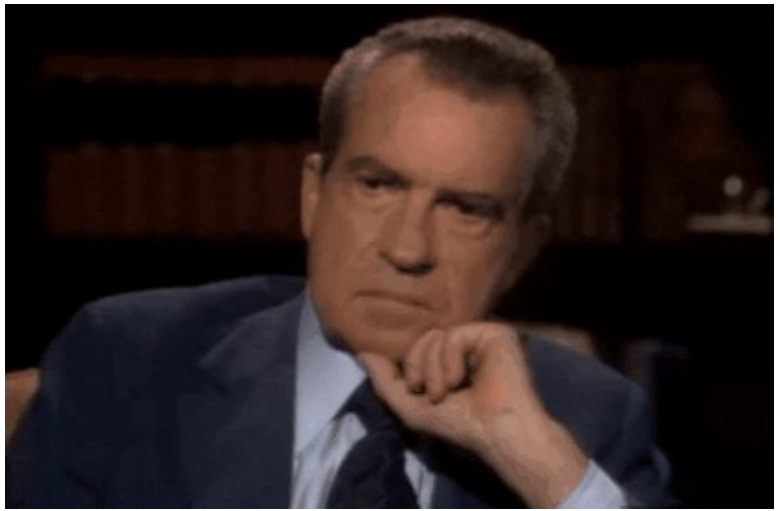


Supervised Learning

Unsupervised Learning

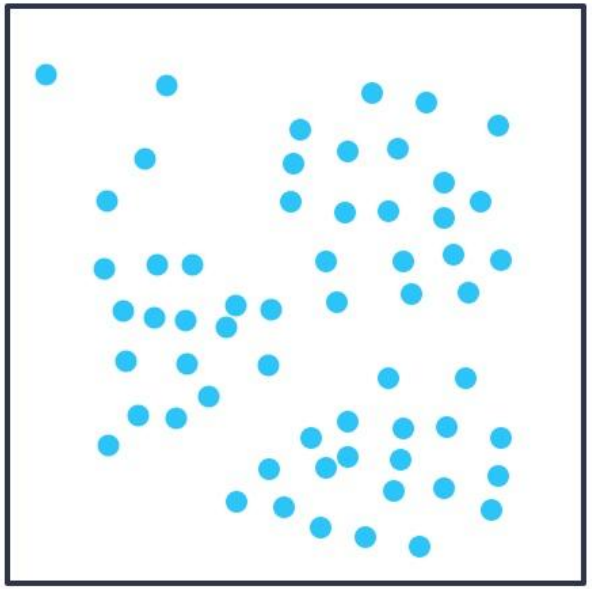
Hmm… unsupervised. What could POSSIBLY go wrong?

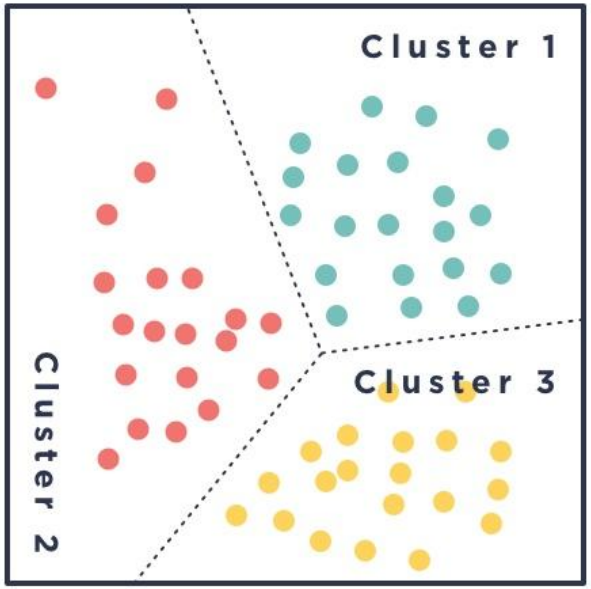
- Computational complexity due to a high volume of data
- Longer **training** times.
- Higher risk of inaccurate results.
- Human intervention to validate output variables.
- Lack of transparency into the basis on which data was clustered.



Class method

CLUSTERING





Unsupervised clustering methods

- Affinity Propagation
- Agglomerative Clustering
- BIRCH
- DBSCAN
- K-Means
- Mini-Batch K-Means
- Mean Shift
- OPTICS
- Spectral Clustering
- Mixture of Gaussians

K means clustering - unsupervised

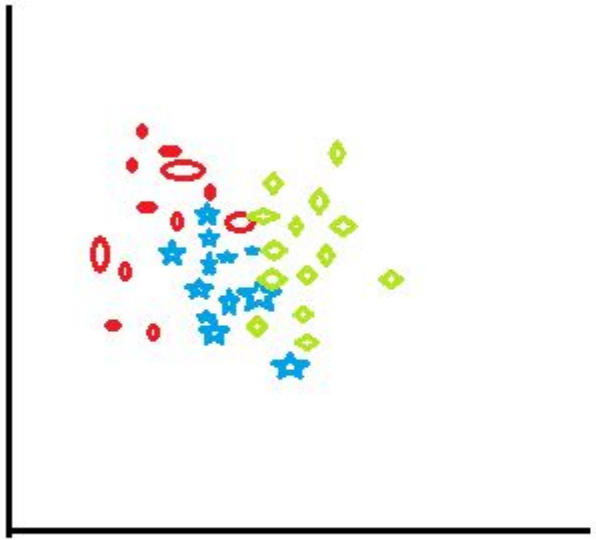


fig 1: before applying k-means clustering

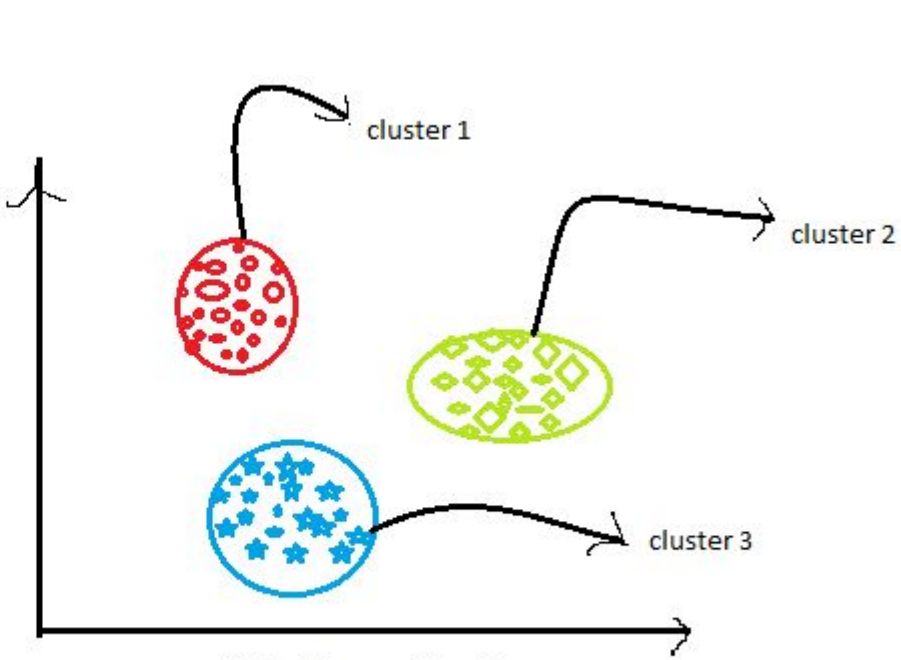


fig 2: After applying K-means clustering

K means clustering - unsupervised

PROS : Easy to implement

Computationally efficient

CONS :

Requires no of clusters in advance

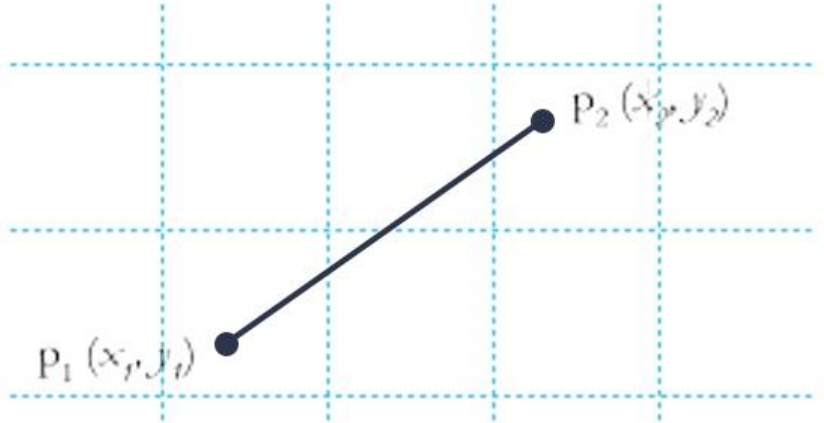
Results vary in presence of outliers

Not great for convex shaped clusters

For numerical values only

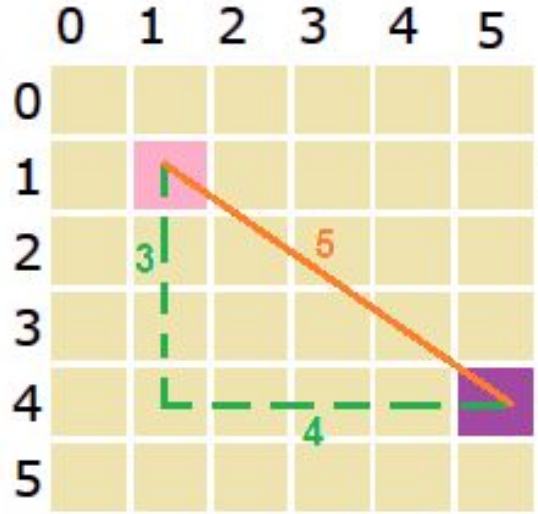
- ➔ First, the k centroids are randomly assigned to a point.
- ➔ Next, each point in the dataset is assigned to a cluster. The assignment is done by finding the closest centroid and assigning the point to that cluster.
- ➔ After this step, the centroids are all updated by taking the mean value of all the points in that cluster.

EUCLIDEAN DISTANCE



Euclidean Distance (d)= $\sqrt{(x_2-x_1)^2 + (y_2-y_1)^2}$

MANHATTAN DISTANCE



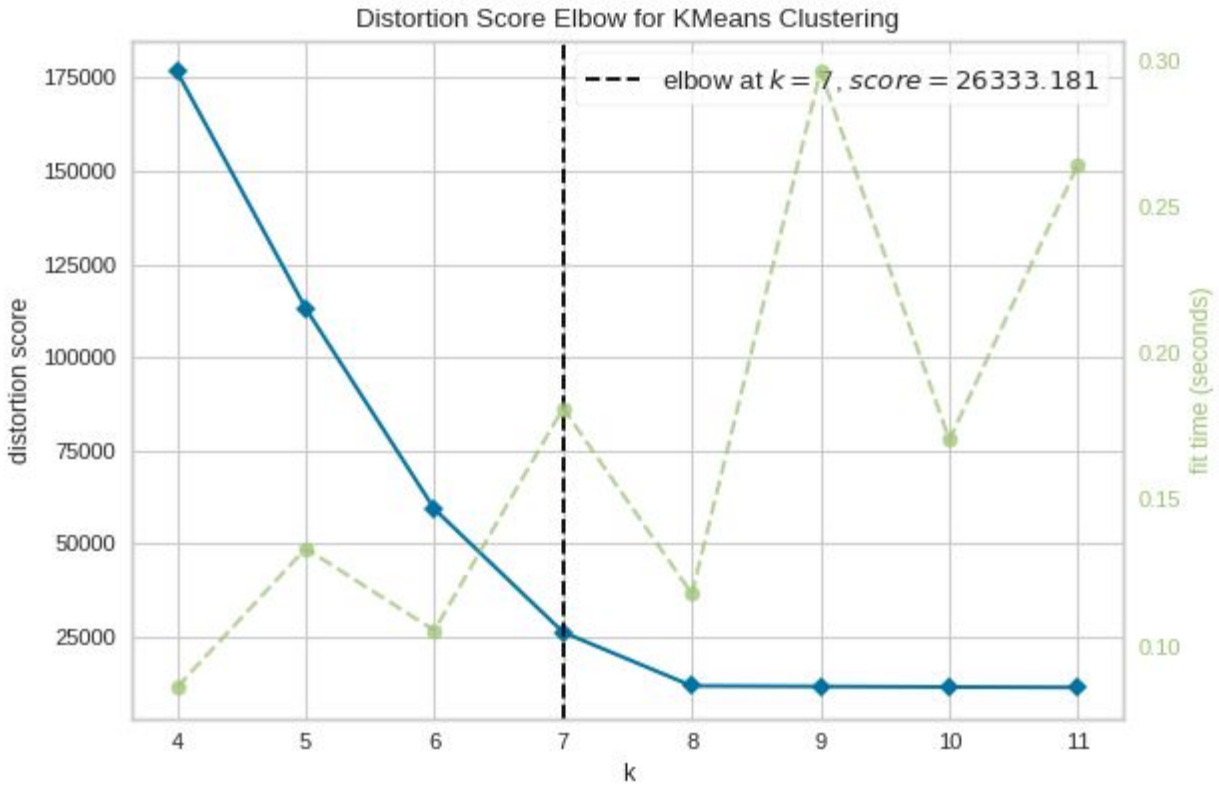
Euclidean distance = $\sqrt{(5-1)^2 + (4-1)^2} = 5$

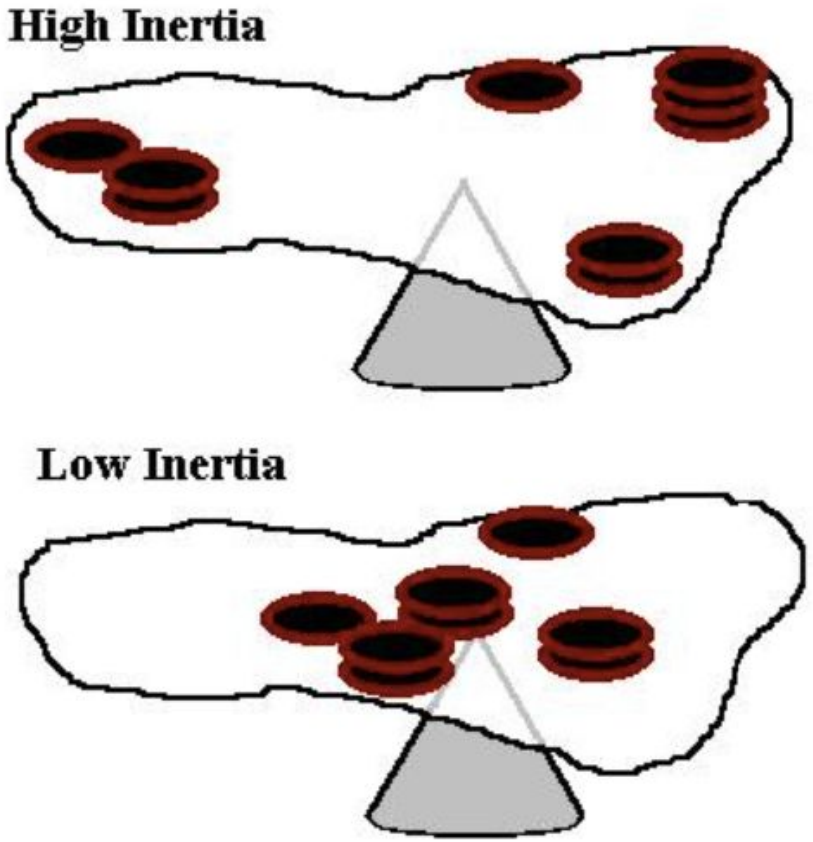
Manhattan distance = $|5-1| + |4-1| = 7$

Picking K - elbow method

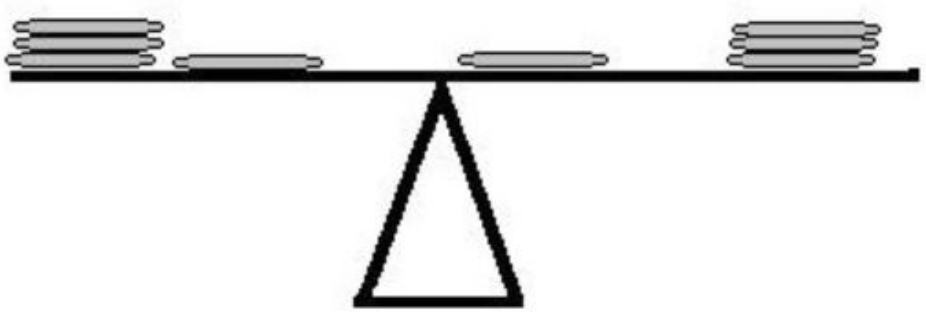
Distortion:average of the squared distances from the cluster centers of the respective clusters.

Inertia: It is the sum of squared distances of samples to their closest cluster center.

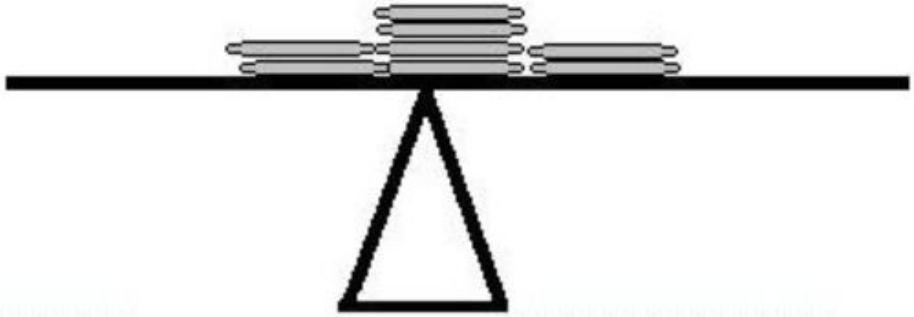



High Inertia
Low Inertia

Inertia on a scale

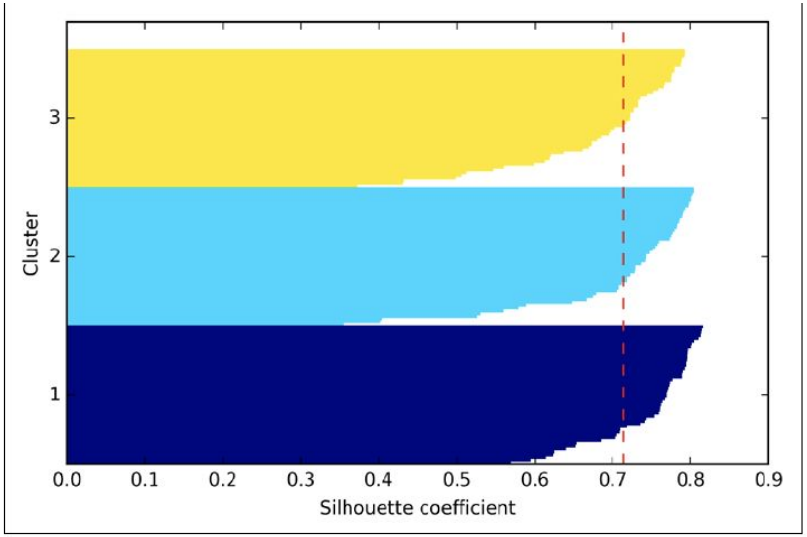


High variance



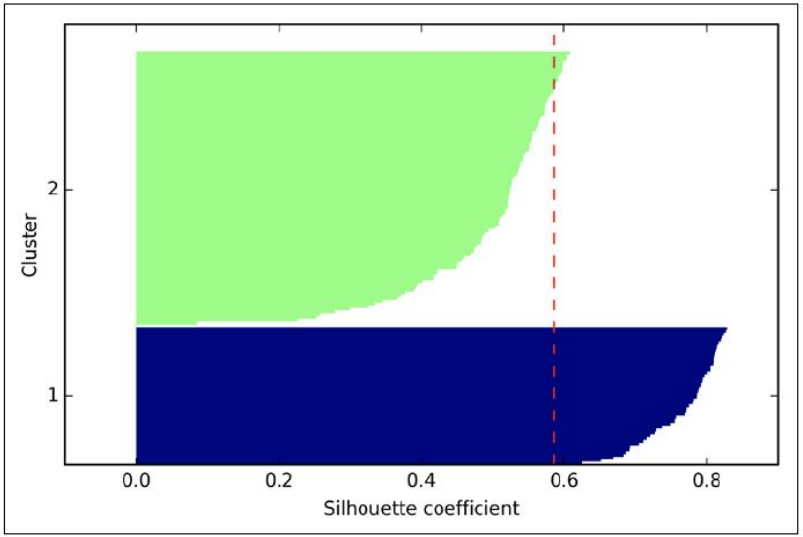
Low variance

IRONHACI



Picking K - silhouette coefficient

Reveals the difference between cluster cohesion and separation.




UNSUPERVISED:
ME TOO. BUT AT LEAST THEY TOLD YOU THE ANSWERS
SUPERVISED:
THEY GAVE ME SO MUCH TO READ, AND TESTS!
REINFORCEMENT:
AT LEAST Y'ALL DON'T MAKE YOUR OWN BOOK!