Unsupervised Learning

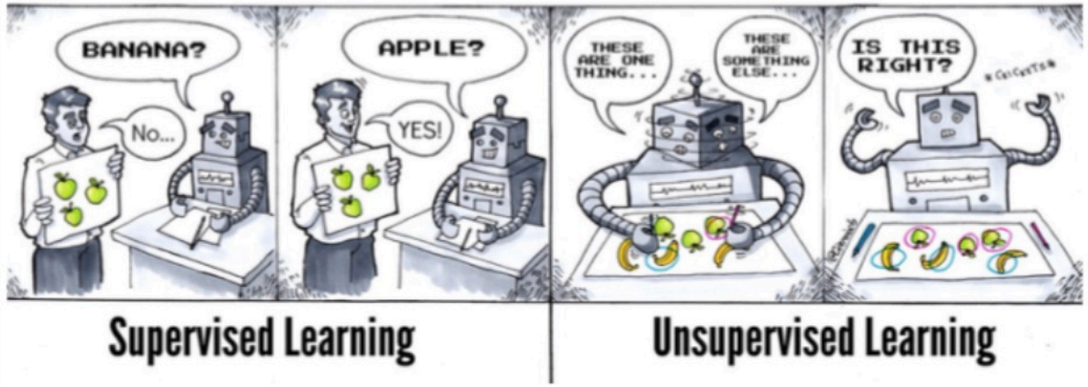
Session-17

- Unsupervised Recap
- K Means Clustering Theory
- Clustering Evaluation
- K Means Clustering with Python

Unsupervised Learning (Recap)

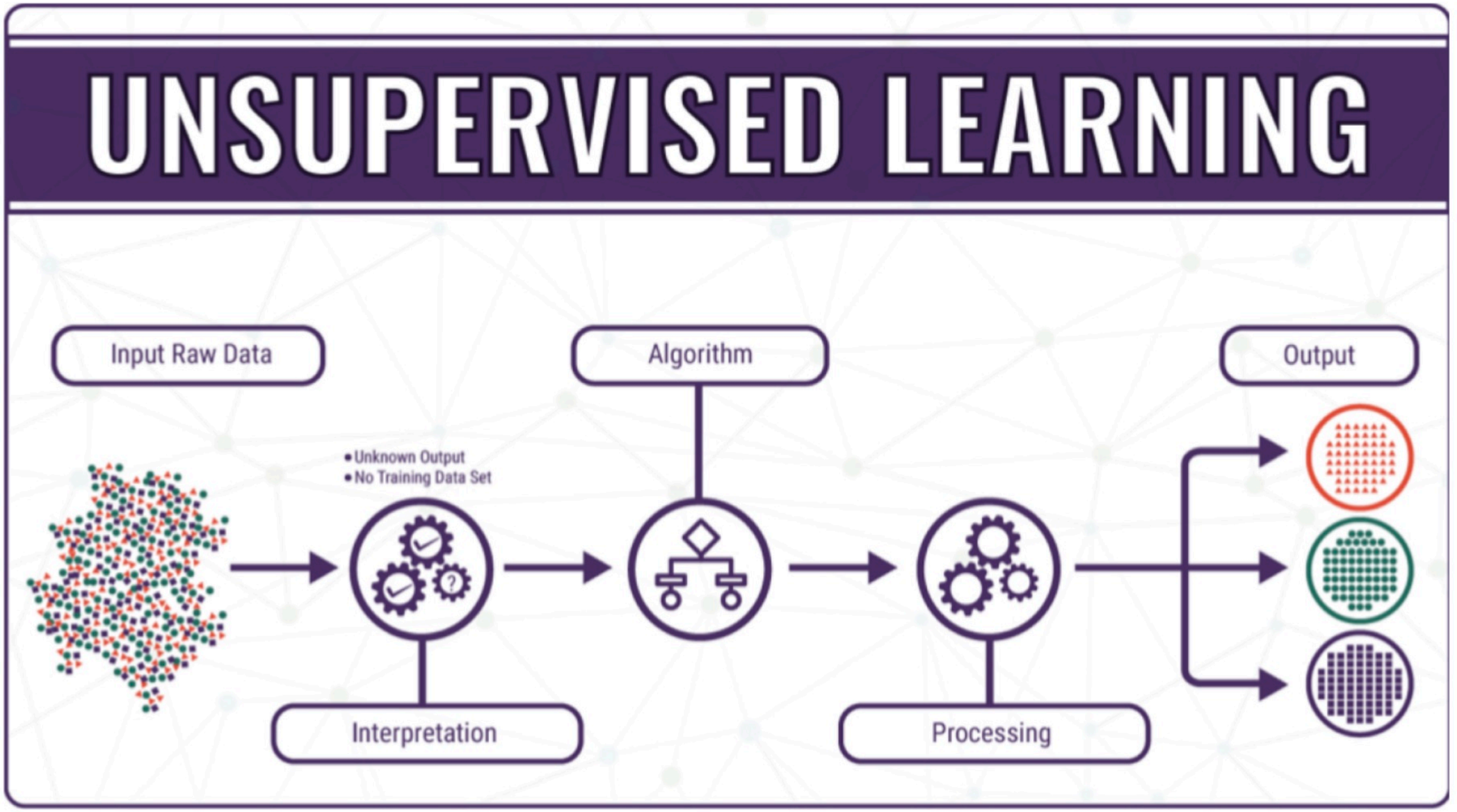
	Age	Cabin	Fare	Name	Parch	PassengerId	Pclass	Sex	SibSp	Ticket	Embarked	Survived
0	22.0	NaN	7.2500	Braund, Mr. Owen Harris	0	1	3	male	1	A/5 21171	S	0
1	38.0	C85	71.2833	Cumings, Mrs. John Bradley (Florence Briggs Th...	0	2	1	female	1	PC 17599	C	1
2	26.0	NaN	7.9250	Heikkinen, Miss. Laina	0	3	3	female	0	STON/O2. 3101282	S	1
3	35.0	C123	53.1000	Futrelle, Mrs. Jacques Heath (Lily May Peel)	0	4	1	female	1	113803	S	1
4	35.0	NaN	8.0500	Allen, Mr. William Henry	0	5	3	male	0	373450	S	0
5	NaN	NaN	8.4583	Moran, Mr. James	0	6	3	male	0	330877	Q	0
6	54.0	E46	51.8625	McCarthy, Mr. Timothy J	0	7	1	male	0	17463	S	0
7	2.0	NaN	21.0750	Palsson, Master. Gosta Leonard	1	8	3	male	3	349909	S	0
8	27.0	NaN	11.1333	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	2	9	3	female	0	347742	S	1
9	14.0	NaN	30.0708	Nasser, Mrs. Nicholas (Adele Achem)	0	10	2	female	1	237736	C	1

No Dependant Feature

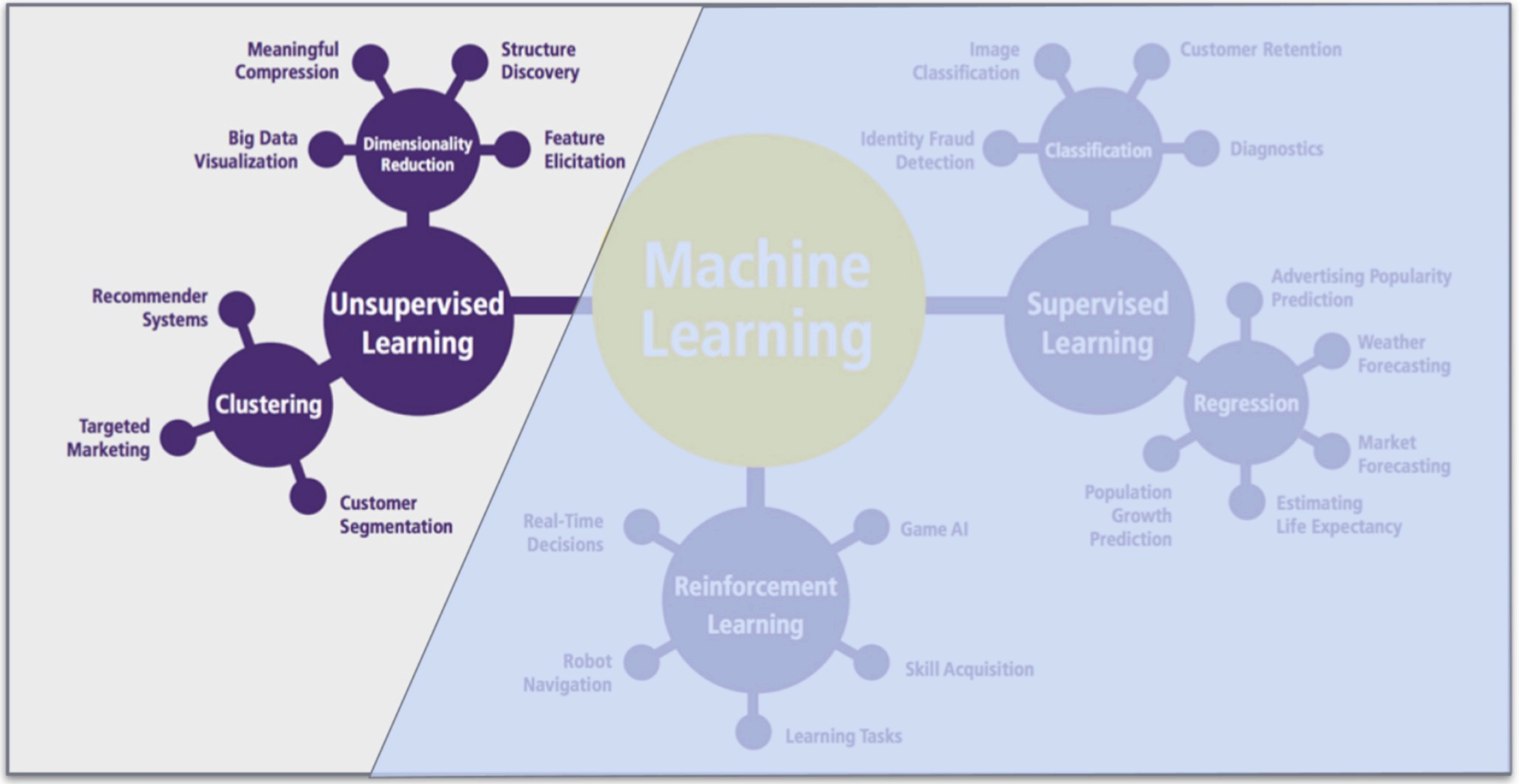


Learning problems where there is **no dependent or target variable** are called **unsupervised learning**. There is a cloud of variables in our dataset. Two of the most common unsupervised learning problems are **clustering** and **dimensionality reduction.**

Unsupervised Learning (Recap)



Unsupervised Learning (Recap)



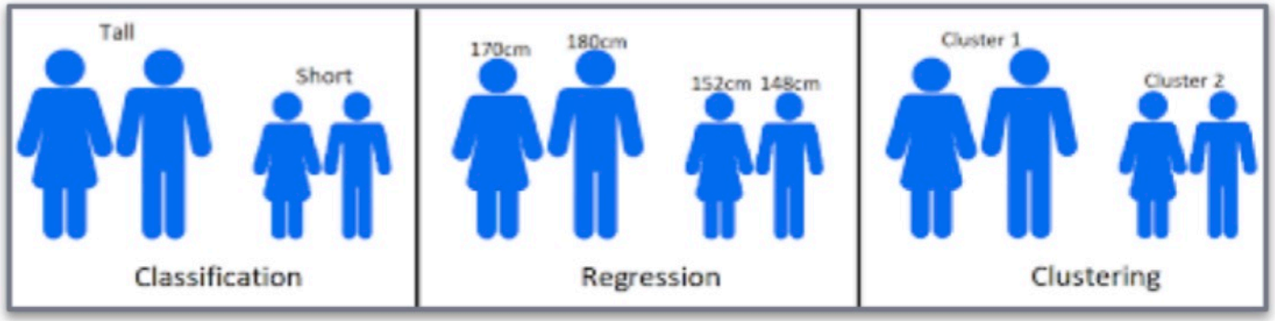
Clustering

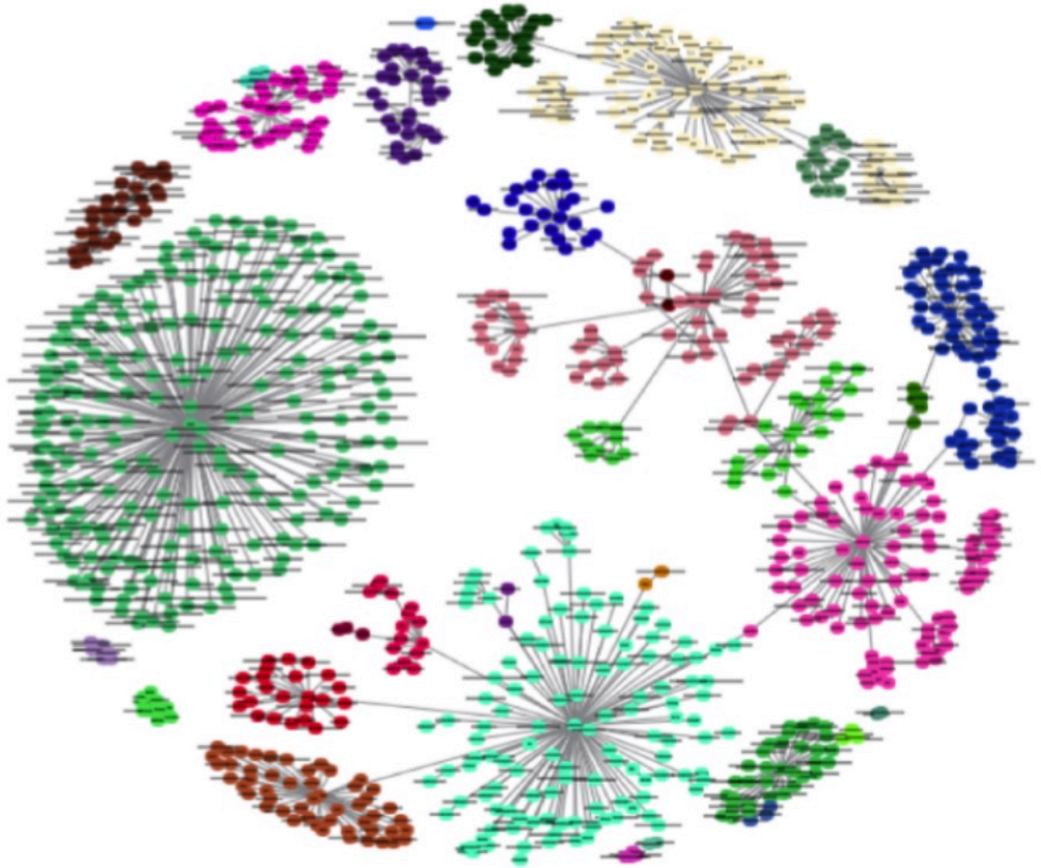
In clustering, we **group similar observations** with an algorithm. These are called **clustering algorithms**.

The clustering algorithm examines the set of variables and says, ***"these are interrelated observations."***

If the **number of clusters is too large**, the clusters will **start to differ insignificantly.**

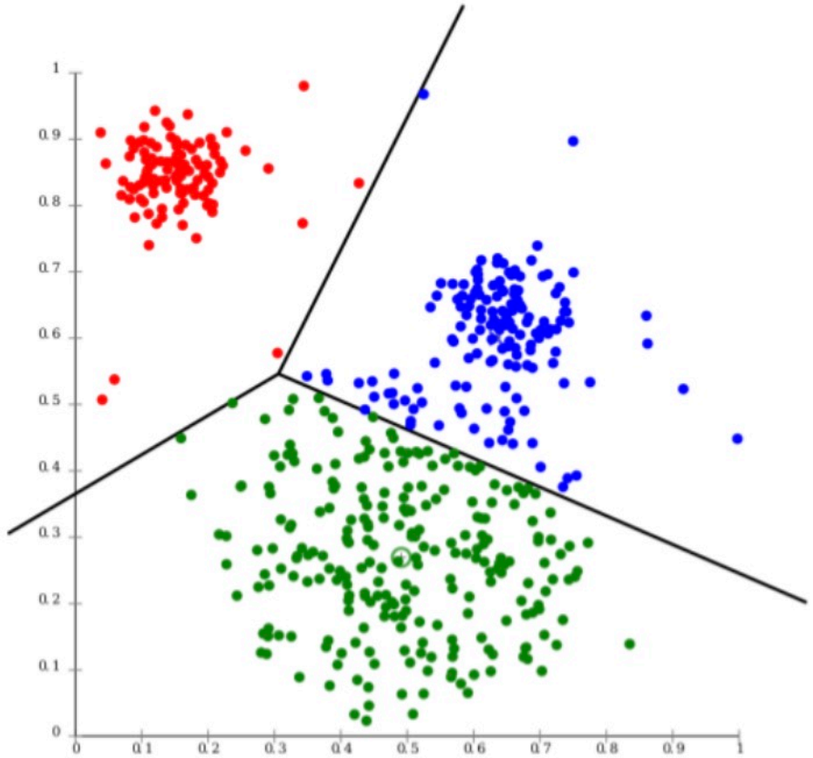
If there are too **few clusters**, we **won't be able to get too much information** from them and the real differences will be hidden.





pic.credit:https://towardsdatascience.com/machine-learning-birch-clustering-algorithm-clearly-explained-fb9838cbeed9

K-means Clustering Theory

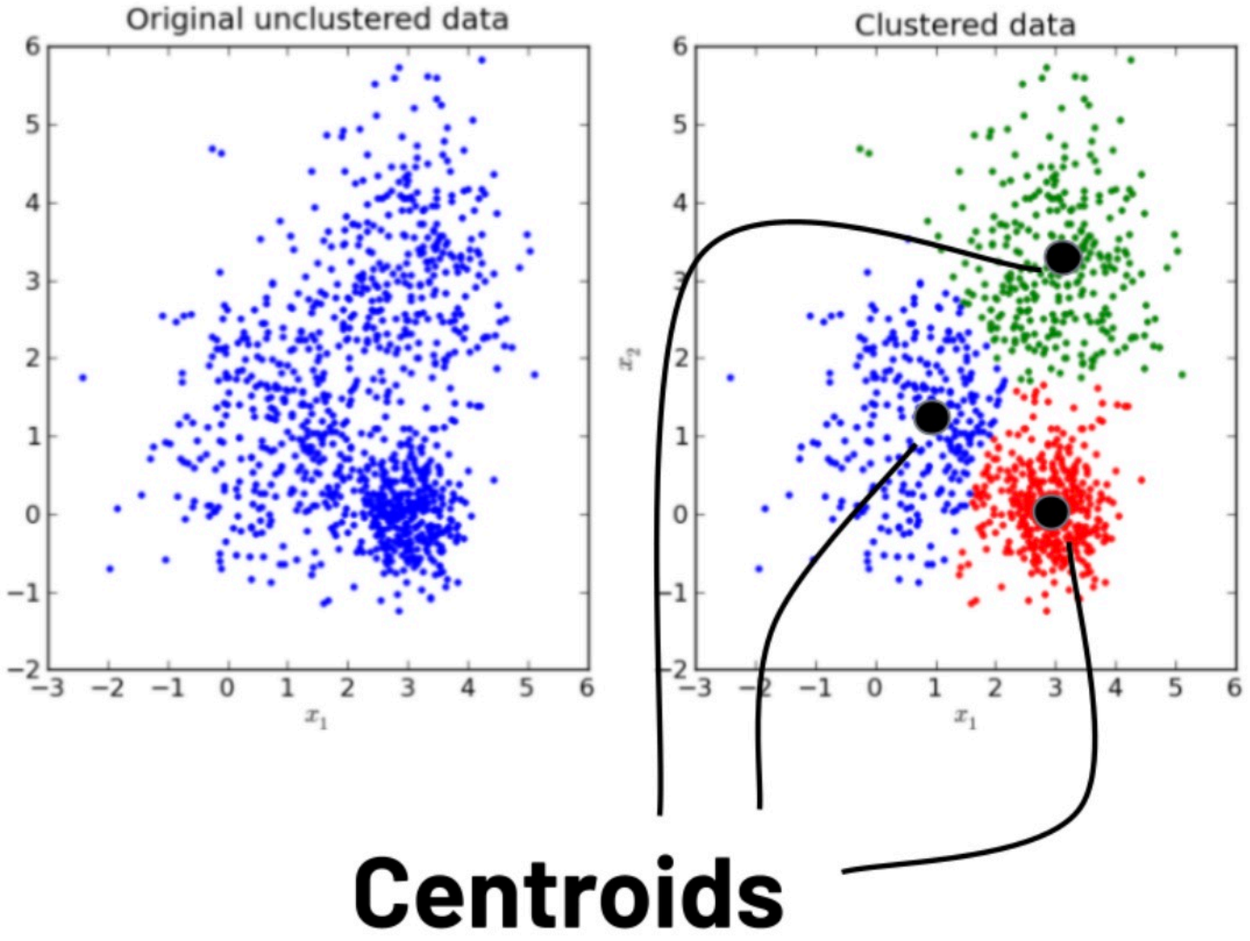


K-means Clustering is widely used **Unsupervised Learning** algorithm.

Some characteristics of K-means:

- The data will **self-cluster** itself
- **"K"** represents the **number of clusters**
- Each cluster consists of similar data within itself, but the clusters are not alike each other.

K-means Clustering Theory



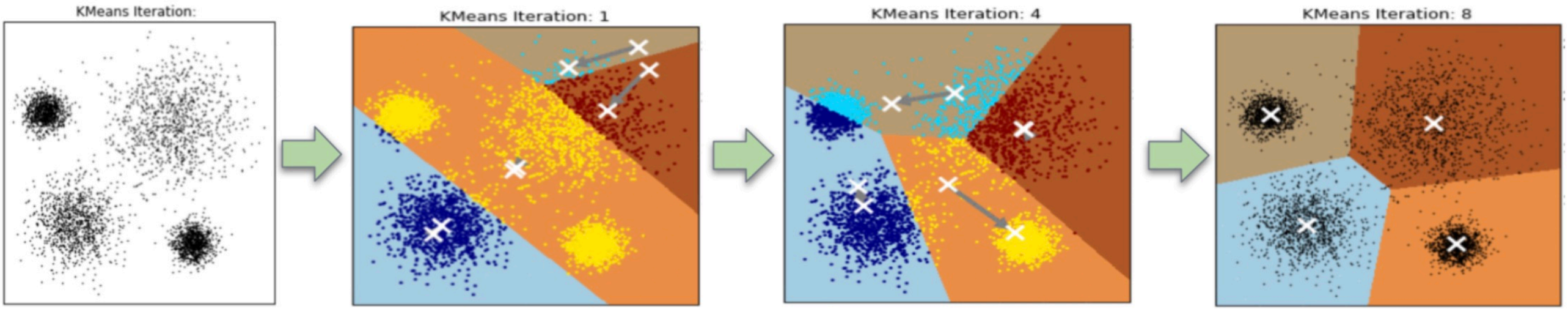
K-mean is an iterative algorithm that ultimately approaches a solution.

The basic idea in k-mean is to find the best points k that form the centers of k sets.

These points are called **centroids**.

Visualizing K-Means Clustering (naftaliharris.com)

K-means Clustering Theory



K-means Algorithm Working Cycle:

Step-1: Detect the number of clusters **(k)**

Step-2: Select random observations and make these observations **first centroids**.

Step-3: Distances to **first centroids** are **calculated for each observation**.

Step-4: Assign each data point to the nearest centroid.

Step-5: Centroid calculations are made again for the clusters formed after the assignment.

Step-6: After the iterations, the cluster structure of the observations in the case where the sum of intra-cluster error squares is minimum is selected as the final cluster.

K-means Clustering Theory

Distance Function: Assign objects to their closest cluster center according to the Euclidean distance function.

$$d(\mathbf{p}, \mathbf{q}) = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}$$

p, q = two points in Euclidean n-space

q_i, p_i = Euclidean vectors, starting from the origin of the space (initial point)

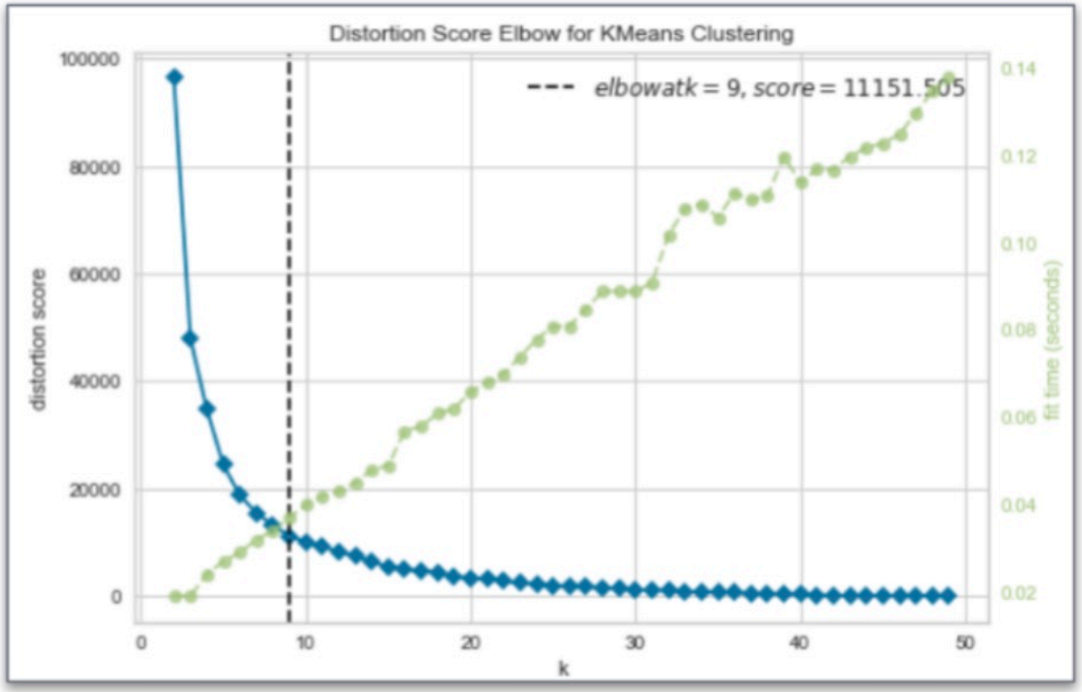
n = n-space

Optimization Criteria: when do we stop the algorithm?

$$\sum_{i=0}^{n} \min_{\mu_j \in C} (||x_i - \mu_j||^2)$$

K Means Clustering Theory

Detect the Number of Clusters (“k”) :

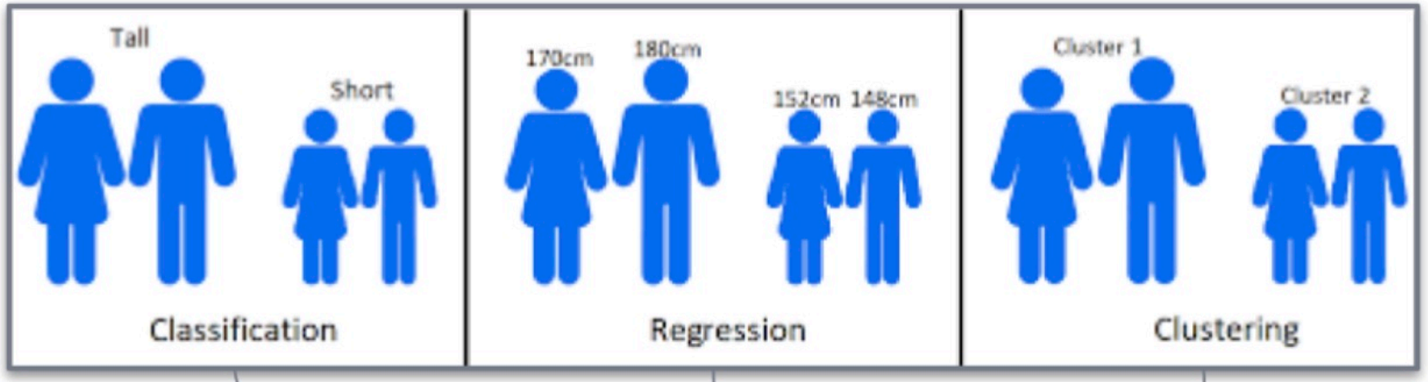


Domain Knowledge

Data Driven Approach

(Elbow Method)

Clustering Evaluation



Accuracy,
Recall,
Precision,
F1_Score

MAE,
MSE,
RMSE,
R^2

No Labels for Evaluation

Evaluation with Labels

Clustering Evaluation

Some Approaches for Clustering Evaluation

1 → Clustering Tendency

2 → Optimal Number of Clusters

3 → Clustering Quality

Clustering Evaluation

Clustering Tendency

Hopkins Test

If the data **does not contain clustering tendency**, then clusters identified by any clustering algorithms may be **irrelevant**.

Non-uniform distribution of points in data set becomes important in clustering.

Source: towardsdatascience.com/**Manimaran**

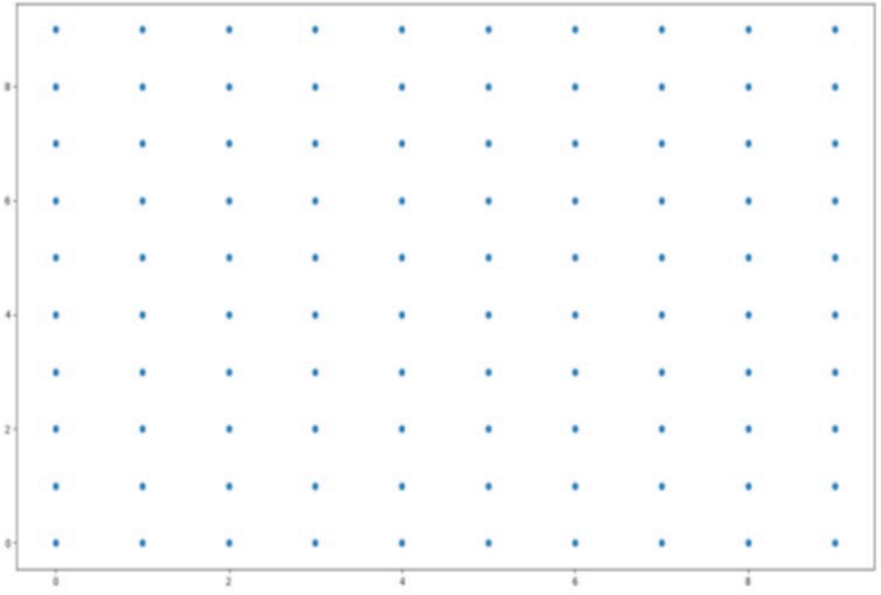
Clustering Evaluation

Clustering Tendency

Hopkins test, a statistical test for randomness of a variable.

Null Hypothesis (Ho) : Data points are generated by non-random, uniform distribution **(implying no meaningful clusters**)

Alternate Hypothesis (Ha): Data points are generated by random data points **(presence of clusters)**



```
>>> from sklearn import datasets
>>> from pyclustertend import hopkins
>>> X = datasets.load_iris().data
>>> hopkins(X,150)
0.16
```

Less suitable for clustering

0.5

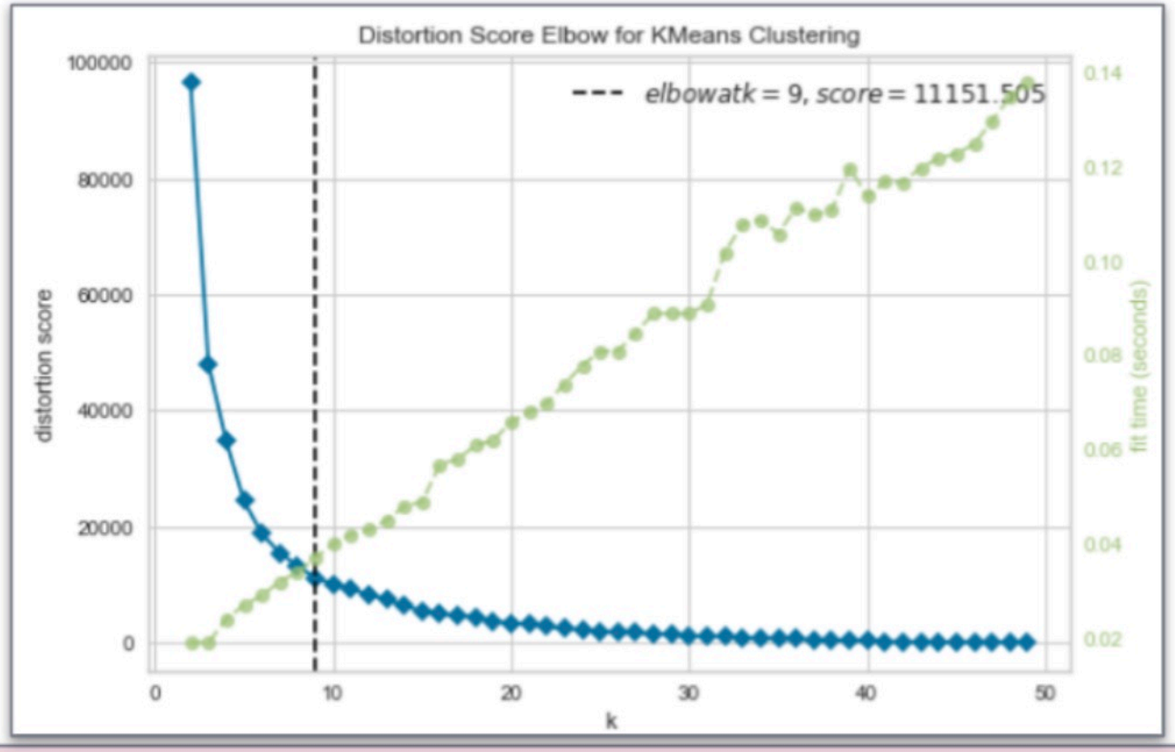
More suitable for clustering

Source: towardsdatascience.com/Manimaran

Clustering Evaluation

Optimal Number of Clusters

Domain Knowledge



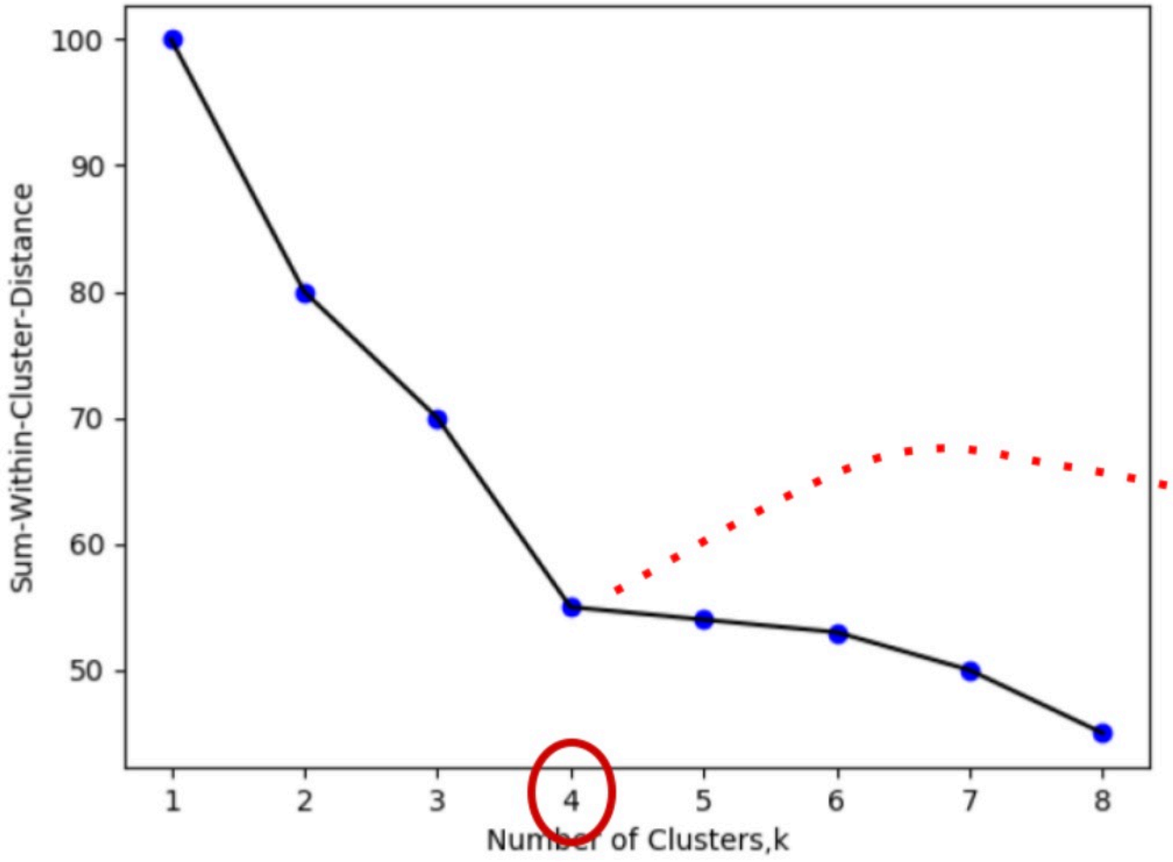
Data Driven Approach

(Elbow Method)

Clustering Evaluation

Optimal Number of Clusters

Elbow Method



Within-cluster variance is a measure of compactness of the cluster.

Inertia: It is calculated by measuring the distance between each data point and its centroid, squaring this distance, and summing these squares across one cluster. It measures how well a dataset was clustered by K-Means.

Plotting the k values and their corresponding sum of within-cluster variance helps in finding the number of clusters.

Clustering Evaluation

Clustering Quality

Ideal clustering is characterized by

minimal intra cluster distance and ***maximal inter cluster distance***.

External Metrics:

(Domain Knowledge-Need some labels)

- ★ ***Adjusted Rand index,***
- ★ Fowlkes-Mallows index,
- ★ Jaccard index/coefficient,

etc.

Internal Metrics:

(No Domain Knowledge)

- ★ ***Silhouette Coefficient,***
- ★ Davies-Bouldin Index,
- ★ Dunn index,

etc.

Clustering Evaluation

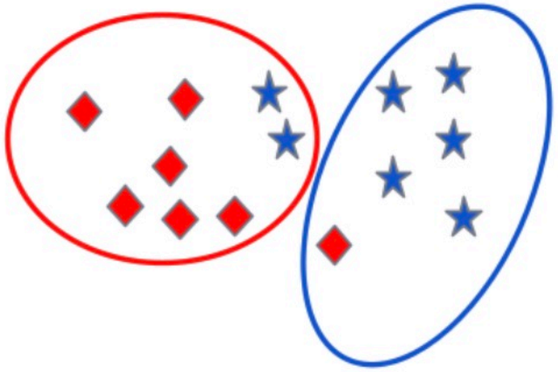
Adjusted Rand Index

sklearn.metrics.adjusted_rand_score

```
sklearn.metrics.adjusted_rand_score(labels_true, labels_pred)  [source]
```

*The Adjusted Rand Index **computes a similarity** measure between two clustering by considering all pairs of samples and counting pairs that are assigned in the same or different clusters in the predicted and true clustering.*

Note: Samples in the clusters are labeled by supervised way



*The value of ARI indicates **no good clustering if it is close to zero or negative**, and **a good cluster if it is close to 1.***

Clustering Evaluation

Silhouette Coefficient

sklearn.metrics.silhouette_score

```
sklearn.metrics.silhouette_score(X, labels, *, metric='euclidean', sample_size=None, random_state=None, **kwds)  [source]
```

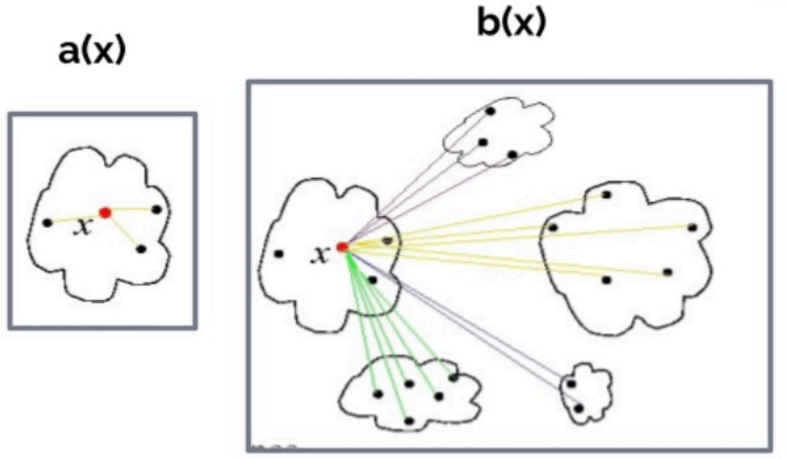
*If the **ground truth labels are not known**, evaluation must be performed using the model itself.* *(One of the evaluation method is Silhouette Coefficient)*

***A higher** Silhouette Coefficient score relates to a model with **better** defined clusters.*

- **a**: The mean distance between a sample and all other points in the same class.
- **b**: The mean distance between a sample and all other points in the *next nearest cluster*.

The Silhouette Coefficient *s* for a single sample is then given as:

$$s = \frac{b - a}{max(a, b)}$$



highly dense clustering

+1

S

-1

incorrect clustering

pic.credit :https://present5.com/clustering-methods-part-3-cluster-validation-pasi-franti/

K Means

Be ready for

KMeans Python

Session