Clustering in a nutshell

Hemza NEDJARI BENHADJ ALI

Open Week 3.0

01

General approach of clustering

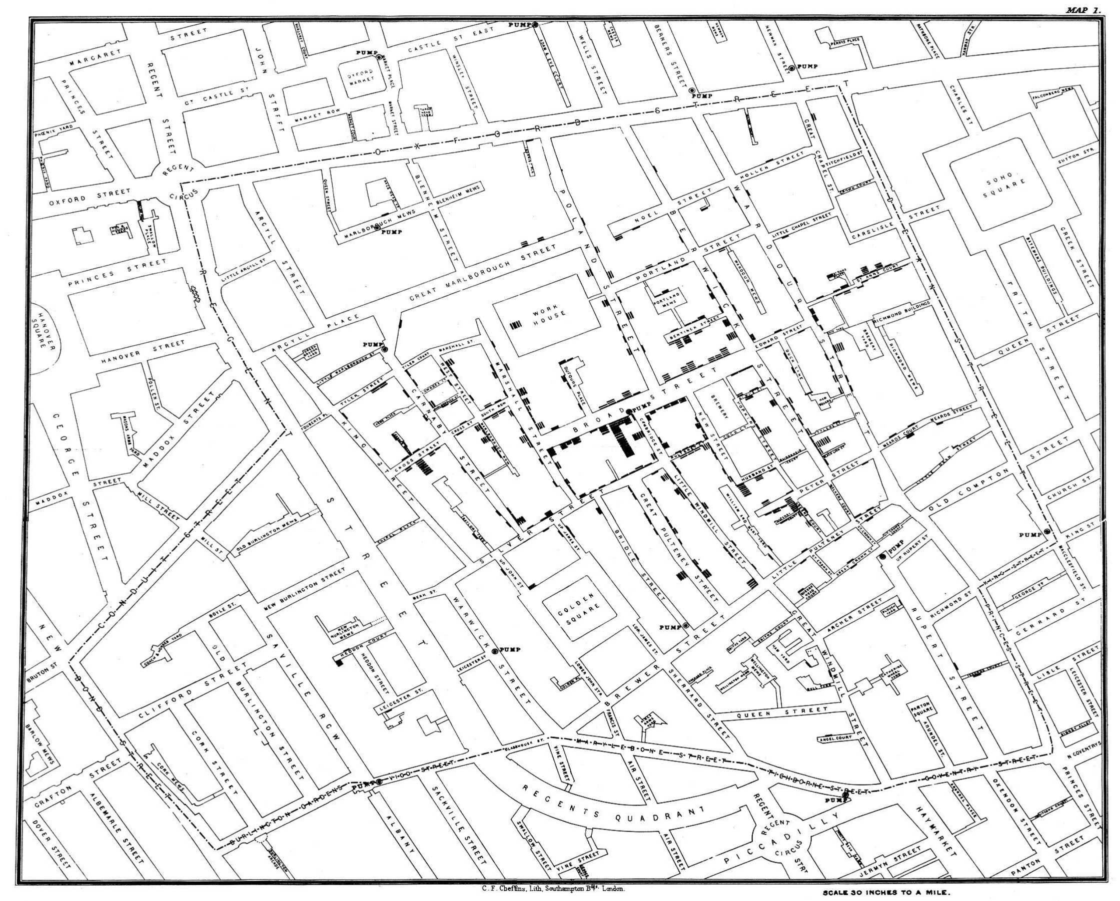
02

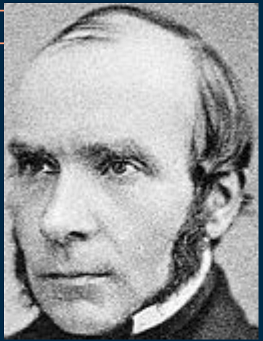
K-means Algorithm

03

Final thoughts

Once upon a time ...

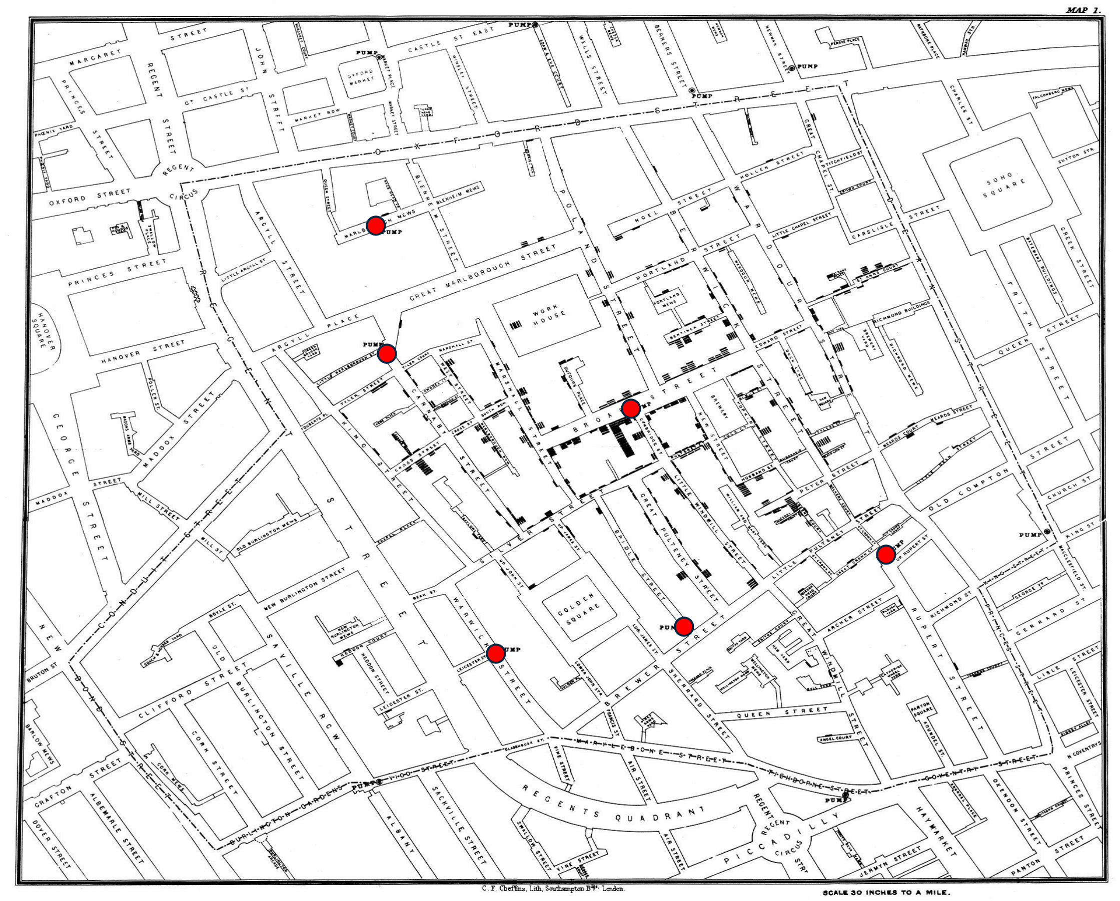


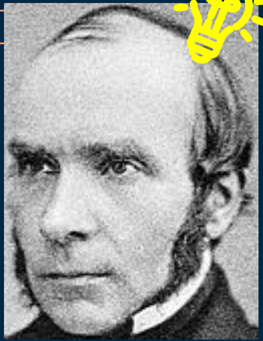


John SNOW

British physician
(1813 – 1858)

Once upon a time ...



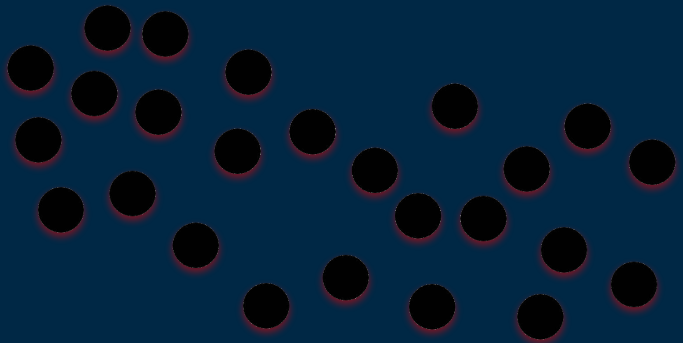


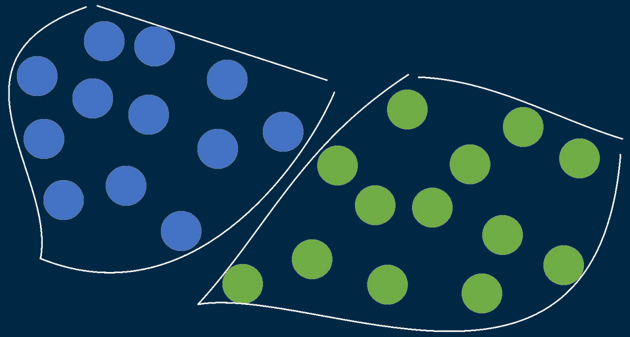
John SNOW

British physician
(1813 – 1858)

Clustering

- The organization of **unlabeled data** into **similarity** groups called **clusters**.
- A cluster is a collection of data items which are “similar” between them, and “dissimilar” to data items in other clusters





How can we do Clustering ?

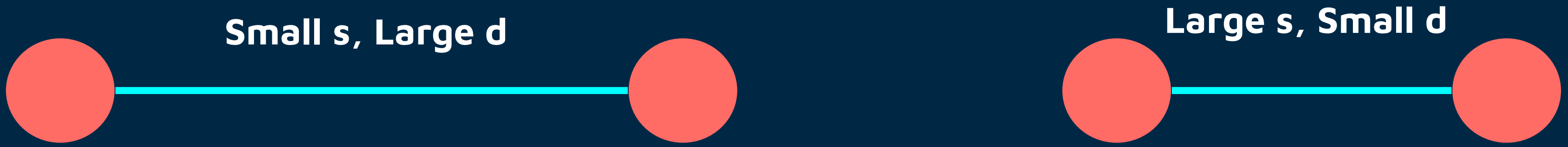
- Proximity measure
- Criterion function
- Algorithm

How can we do Clustering ?

1 . Proximity measure

- Similarity s(x,y) : Large if x and y are similar.

- Dissimilarity (distance) d(x,y) : small if x and y are similar.



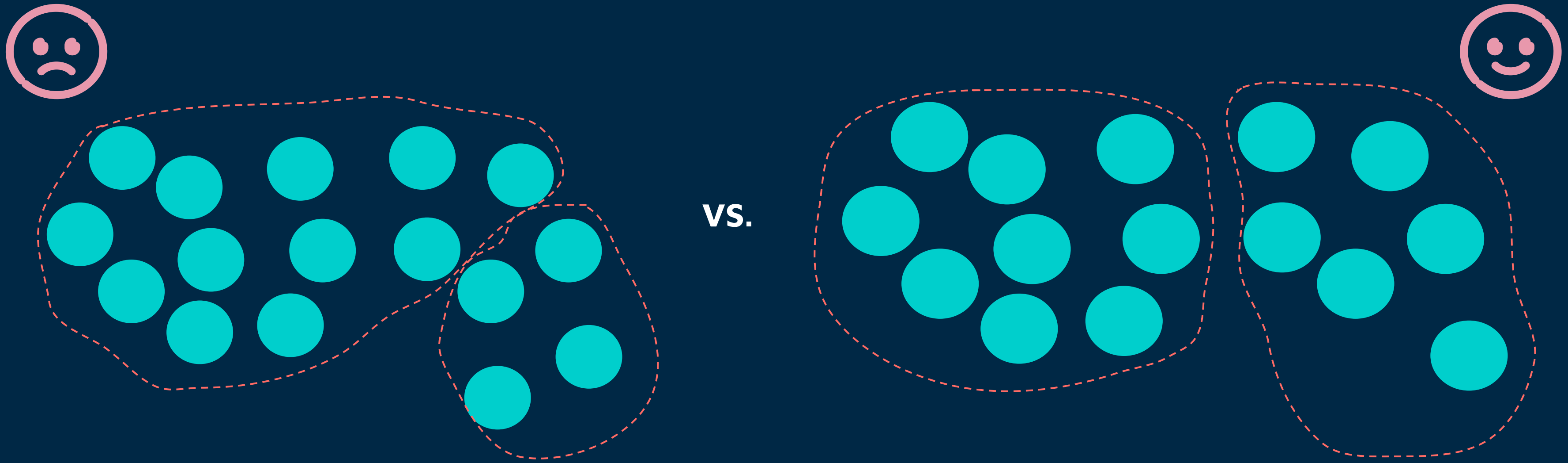
Small s, Large d

Large s, Small d

How can we do Clustering ?

2 . Criterion function

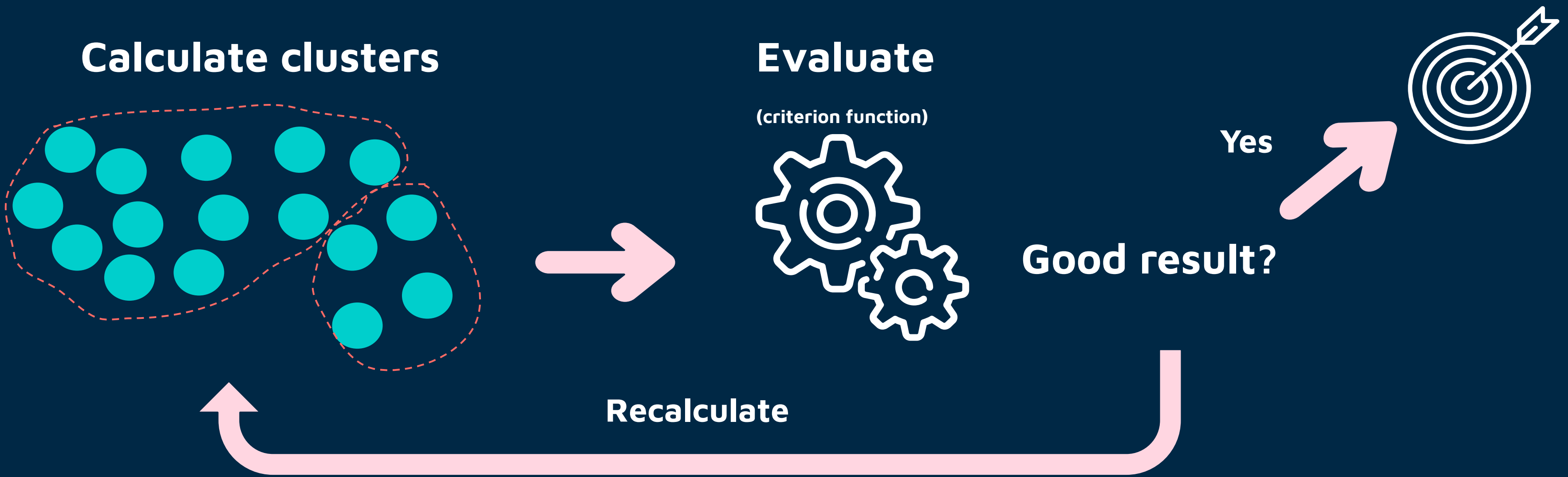
A formula to evaluate the quality of the clustering



How can we do Clustering ?

3 . Algorithm to compute the clustering

Iterate over data to optimize the criterion function



Calculate clusters

Evaluate

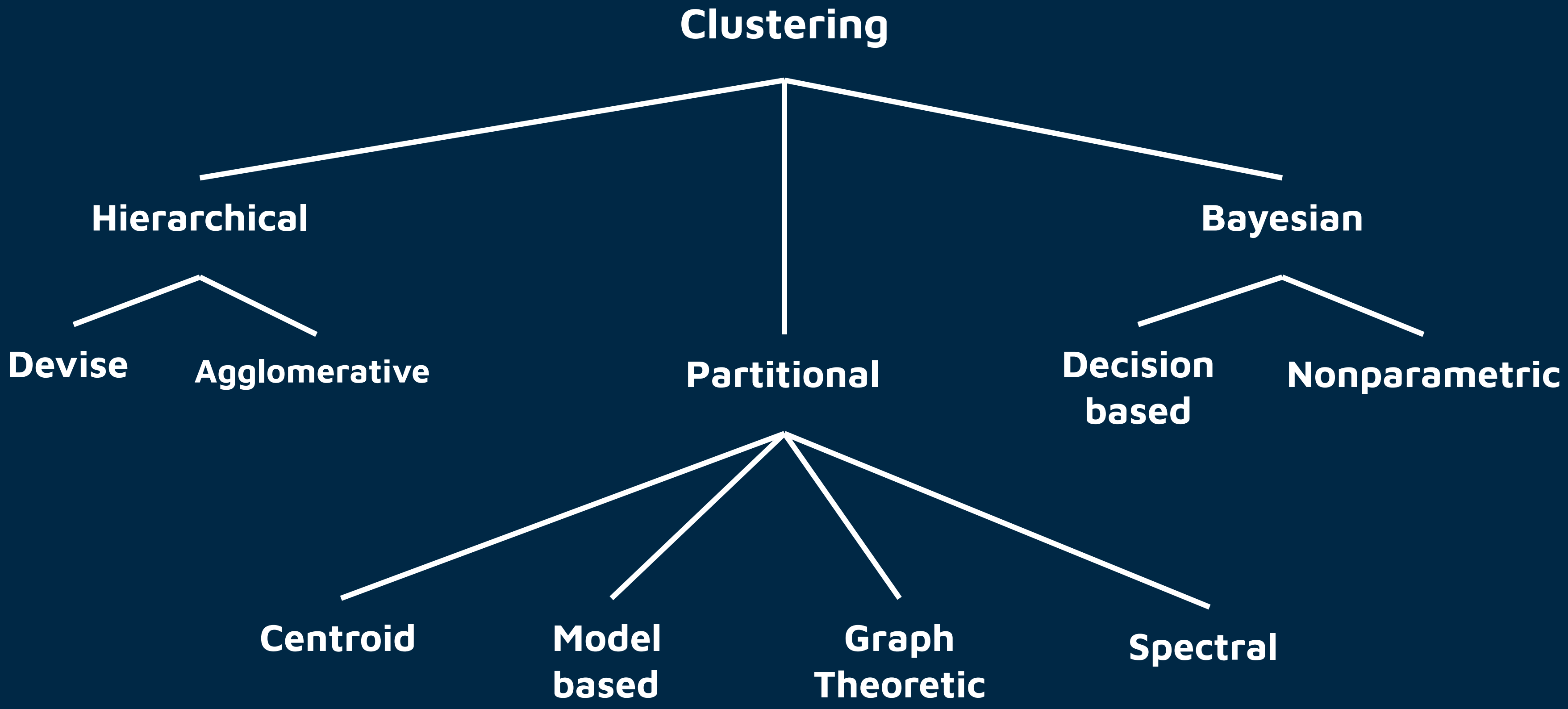
(criterion function)

Good result?

Yes

Recalculate

Clustering techniques



Clustering

Hierarchical

Partitional

Bayesian

Devise

Agglomerative

Decision based

Nonparametric

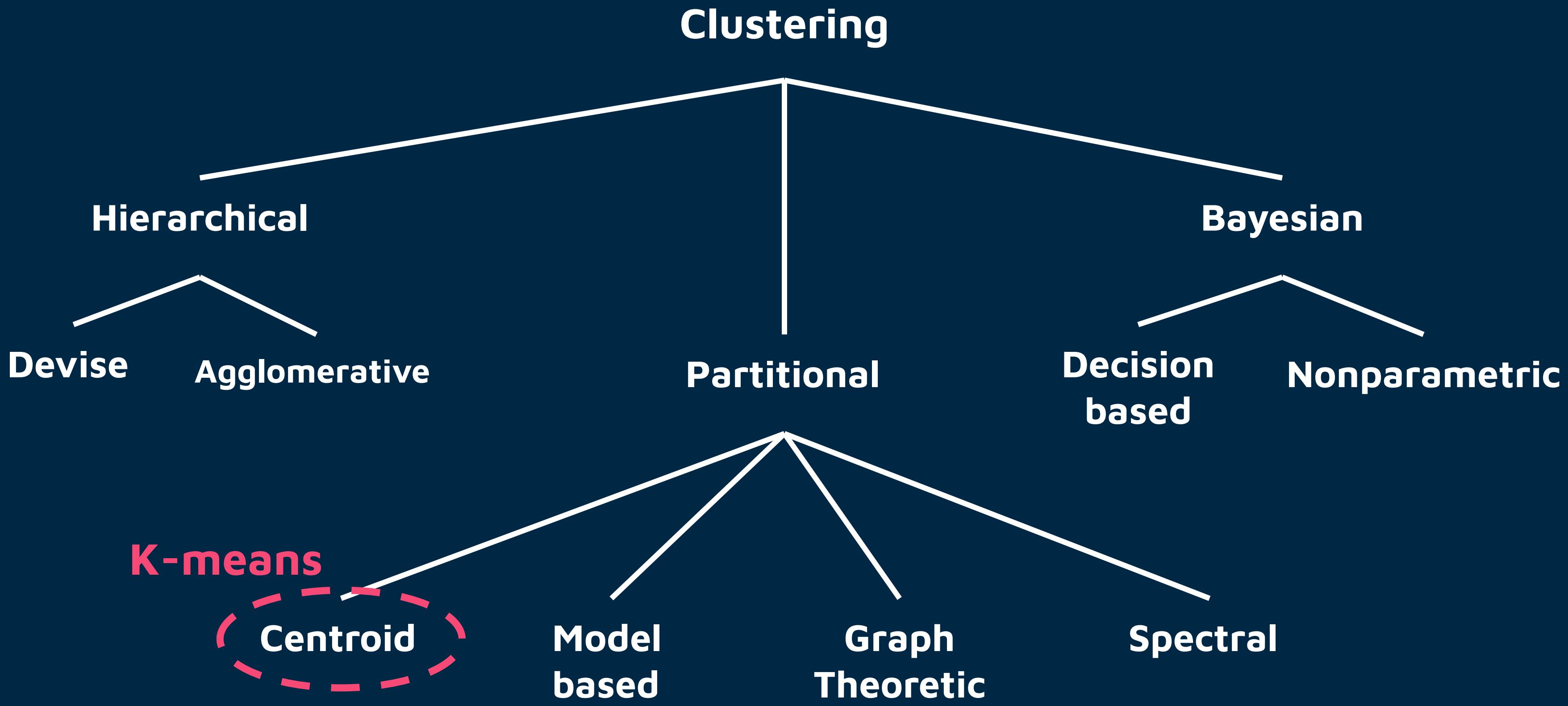
Centroid

Model based

Graph Theoretic

Spectral

Clustering techniques



Clustering

Hierarchical

Devise

Agglomerative

Partitional

K-means

Centroid

Model based

Graph Theoretic

Spectral

Bayesian

Decision based

Nonparametric

K-means

Partitioning the given data into **k clusters** :

- Each **cluster** has a cluster center, called **centroid**.
- Centroid is the mean point of the cluster.
- **K** is the **clusters number** specified by the user.

K-means

Partitioning the given data into **k clusters :**

- Each **cluster** has a cluster center, called **centroid**.
- **Centroid** is the **mean point** of a cluster.
- **K** is the **clusters number** specified by the user.

How can we calculate the centroid of a cluster ?

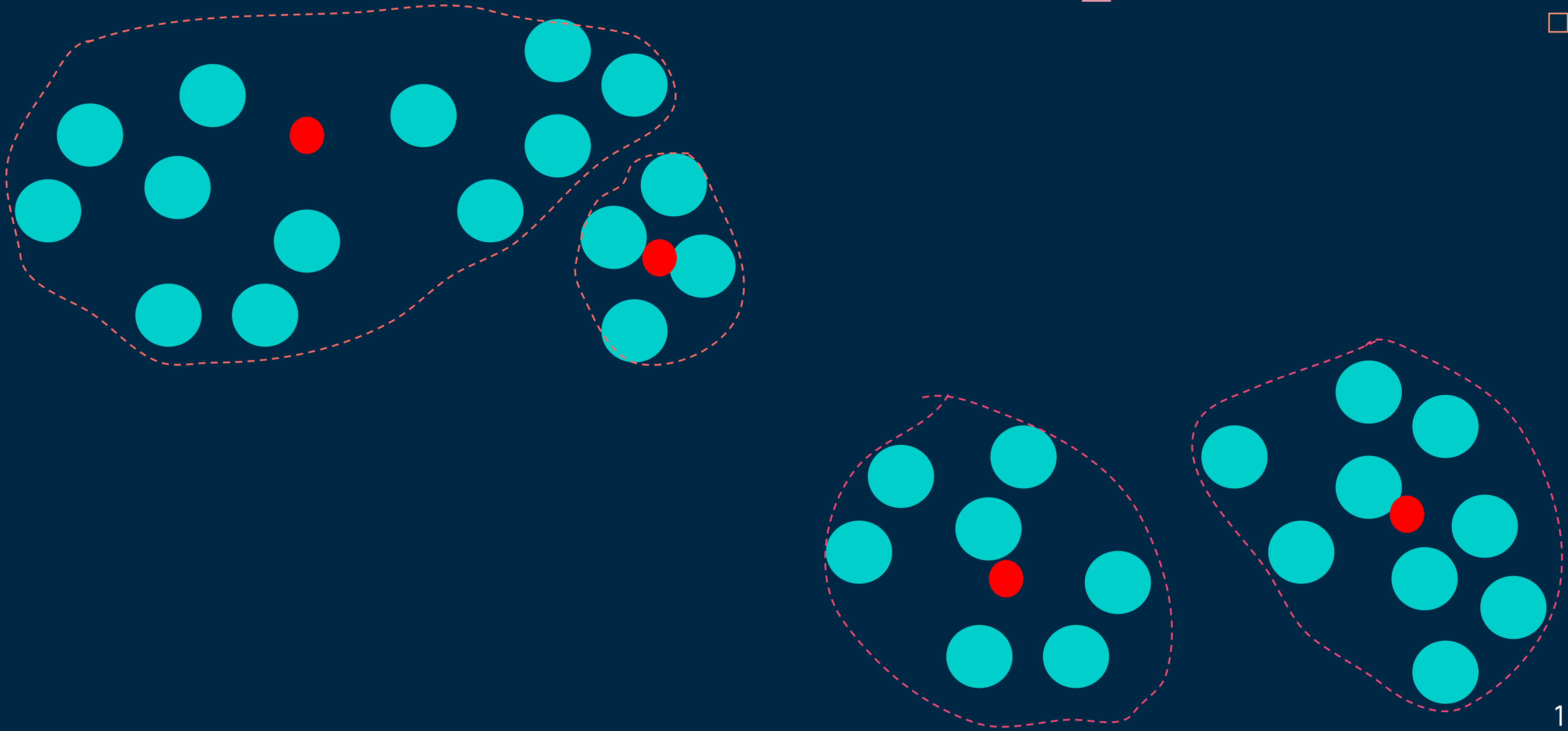
K-means

Sum of Squared Error **(SSE)**

$$SSE = \sum_{j=1}^{k} \sum\nolimits_{\mathbf{x} \in C_j} d(\mathbf{x}, \mathbf{m}_j)^2$$

1. **Cj** is the cluster number j.
2. **m$_j$** the **centroid** of cluster Cj.
3. **d(x,y)** is the **distance** between x and y.

K-means



K-means

Partitioning the given data into **k clusters :**

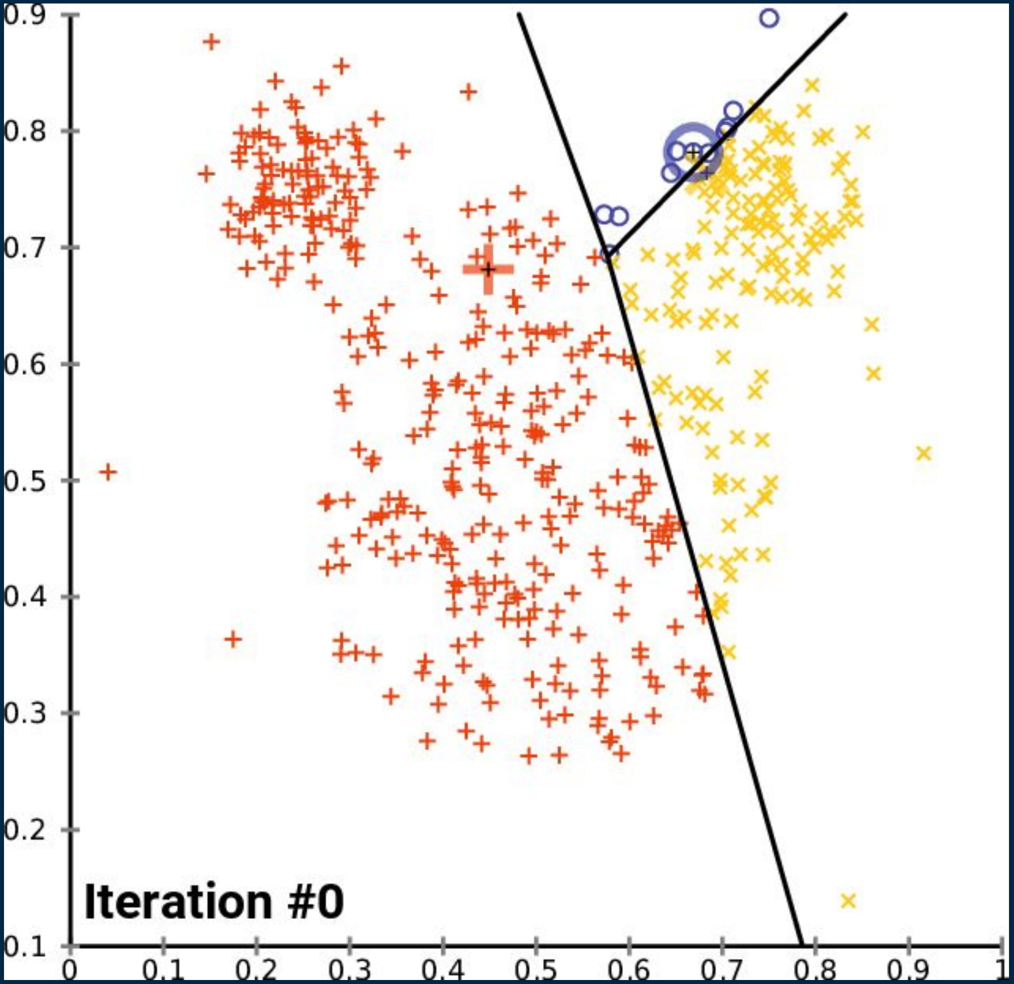
1. Choose **k (random)** data **points** as the **initial centroids**.
2. **Assign** each data point to the **closest centroid**.
3. **Re-compute** the **centroids** using the current cluster memberships.
4. Repeat 2 & 3 untill **SSE** is **minimized** or **stabillized**.

K-means

A cluster is a :

Centroid : C_i

Sum of Distances $d(C_i,X)$

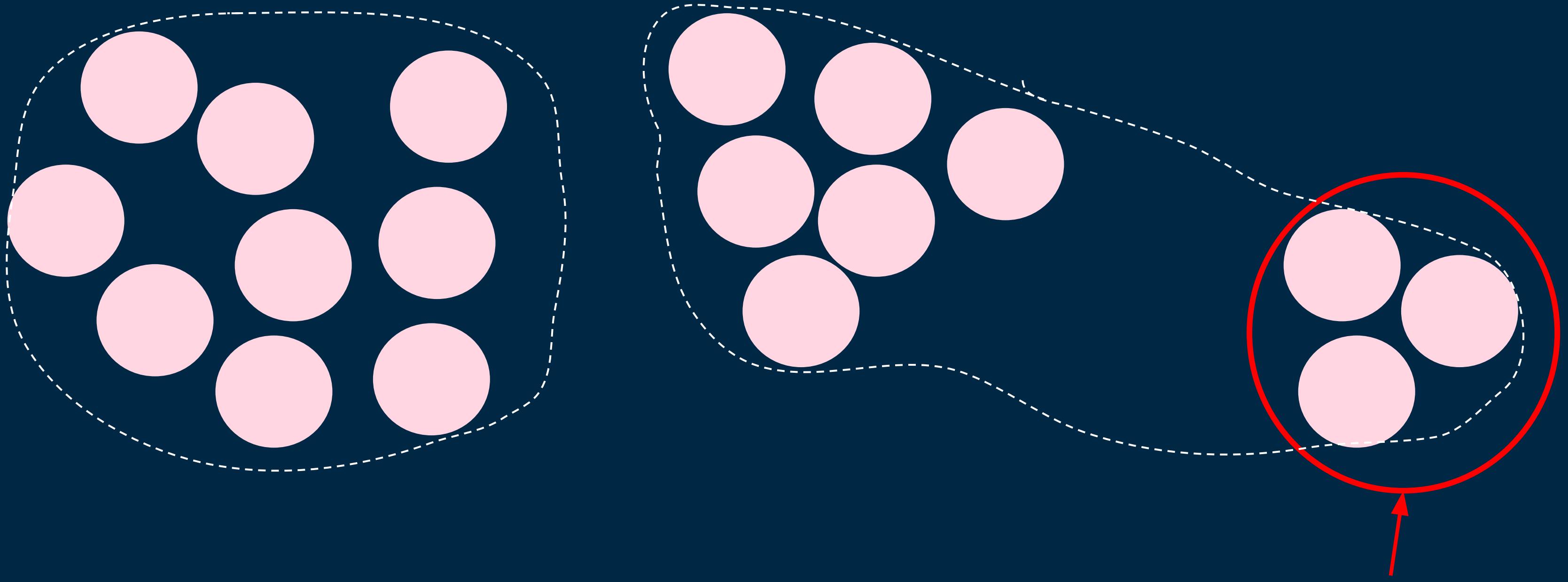


Why K-means ?

- **Simple**: easy to understand and to implement

- **Efficient** : time complexity :O(tkn) for the Euclidean distance
  - n : number of points
  - k : number of clusters
  - t : number of iterations

Yes! ... BUT



K-means

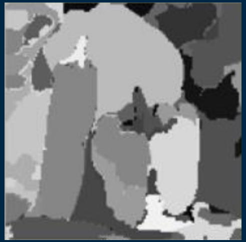
Code Demo

Check my github Repo :

https://github.com/Hamzandj/Open-Week-3.0

In a Nutshell

- **Clustering :** unsupervised machine learning.
- The clustering output can serve to **train a supervised model**.
- A good understanding of the data is mandatory.
- Results **interpretation** is a **key** to have a **good model that helps in decision making**.
- Clustering can be used to **solve complex problems**



THANKS

✉ nedjari.ba.hemza@gmail.com

in nedjari.ba.hemza

CREDITS: This presentation template was created by Slidesgo,
including icons by Flaticon, and infographics & images by Freepik