Faculté d'informatique
Departement IA & SD

DATA MINING

TP4: le clustering hiérarchique (AGENS & DIANA)

BENADJAL.N
Phd student in computer science
LAB: MOVEP
Email: nesrinebenadjel63@gmail.com

INGSOFT4

2025/2026

Introduction

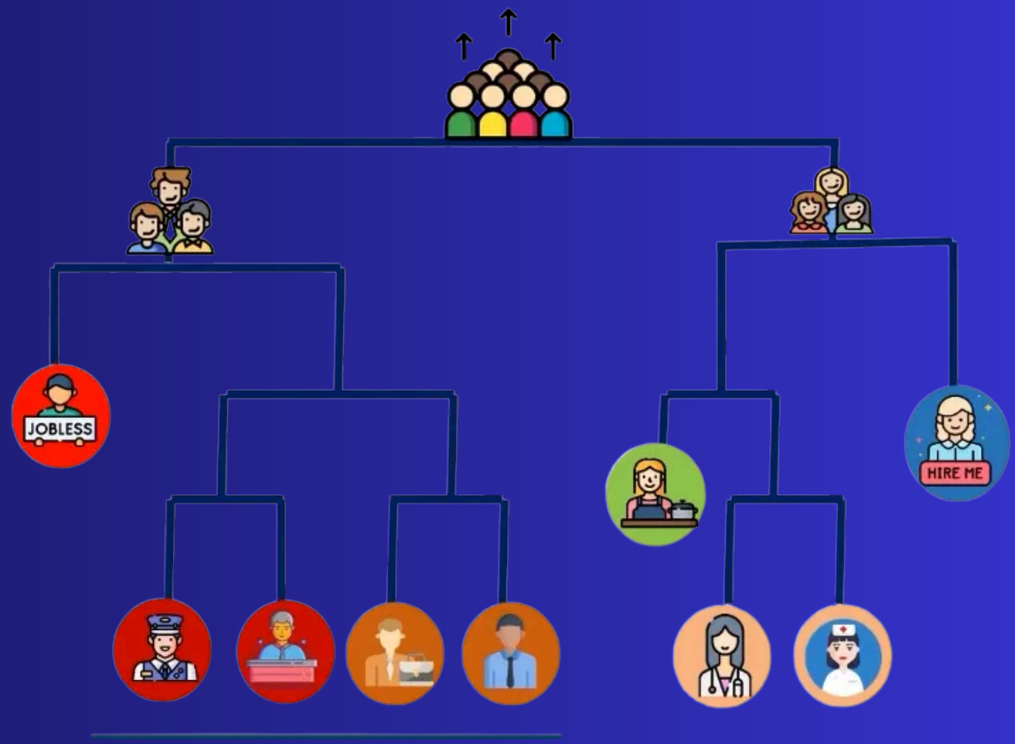
Dans les séances de TP passées, nous avons découvert les algorithmes d'apprentissage automatique basés sur le calcul des distances avec un centre ou médoïde. Dans ce TP , nous allons voir deux nouveaux algorithmes de clustering hiérarchiques :

- L’algorithme AGNES
- L’algorithme DIANA

Qu'est-ce que le clustering hiérarchique ?

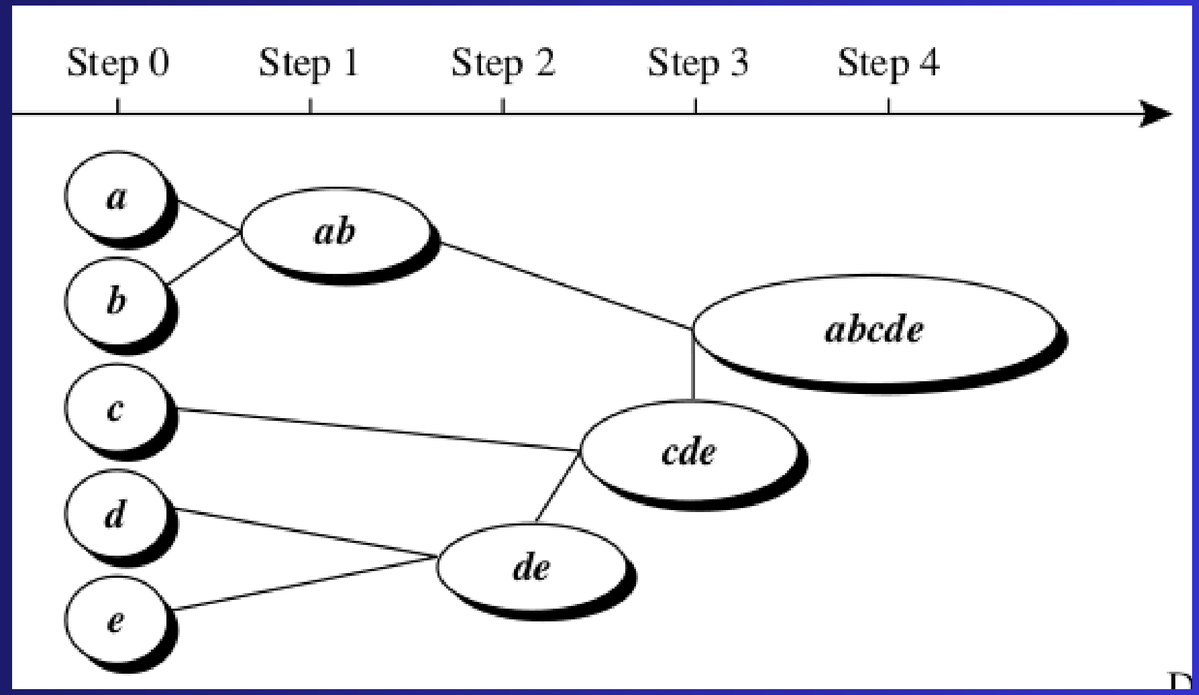
le clustering hiérarchique

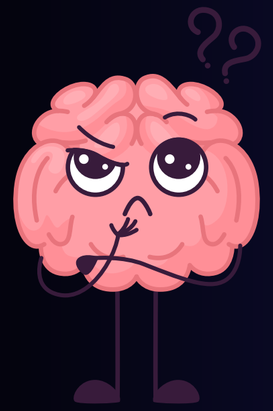
Le clustering ou regroupement hiérarchique consiste à créer une arborescence de cluster pour représenter les données.
Chaque noeud de l’arborescence contient un groupe de données similaires, et les noeuds sont regroupés en fonction de leurs similitudes.
Le nombre total de clusters n’est pas prédéterminé avant la création du graphique.



AGENS:

l'algorithme AGENS considère initialement chaque point de données comme un cluster unique, puis commence à combiner les paires de clusters les plus proches.
Le processus se répète jusqu'à ce que tous les clusters soient fusionnés en un seul cluster contenant tous les ensembles de données.





Il s'agit du calcul des distances entre les clusters. Comment cela se fait-il ?

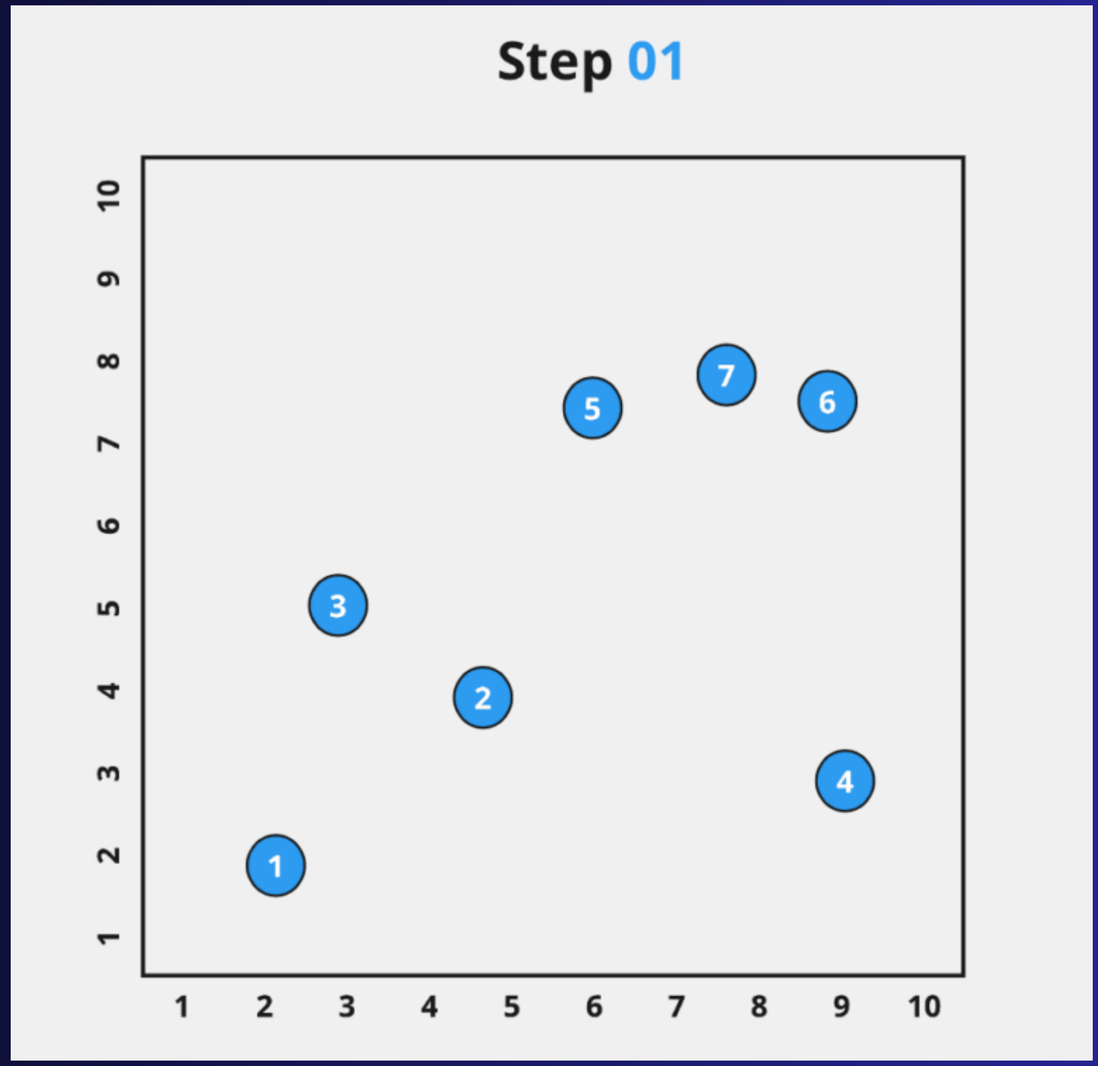
Utiliser les méthodes de liaison.

Critère de Liaison (Lien)

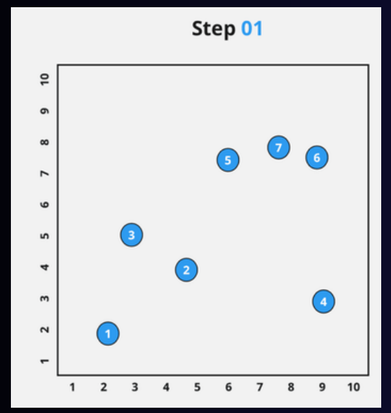
Il existe plusieurs façons de mesurer la distance entre les éléments afin de déterminer les règles de regroupement, et elles sont souvent appelées méthodes de liaison.
Voici quelques méthodes de liaison courantes :

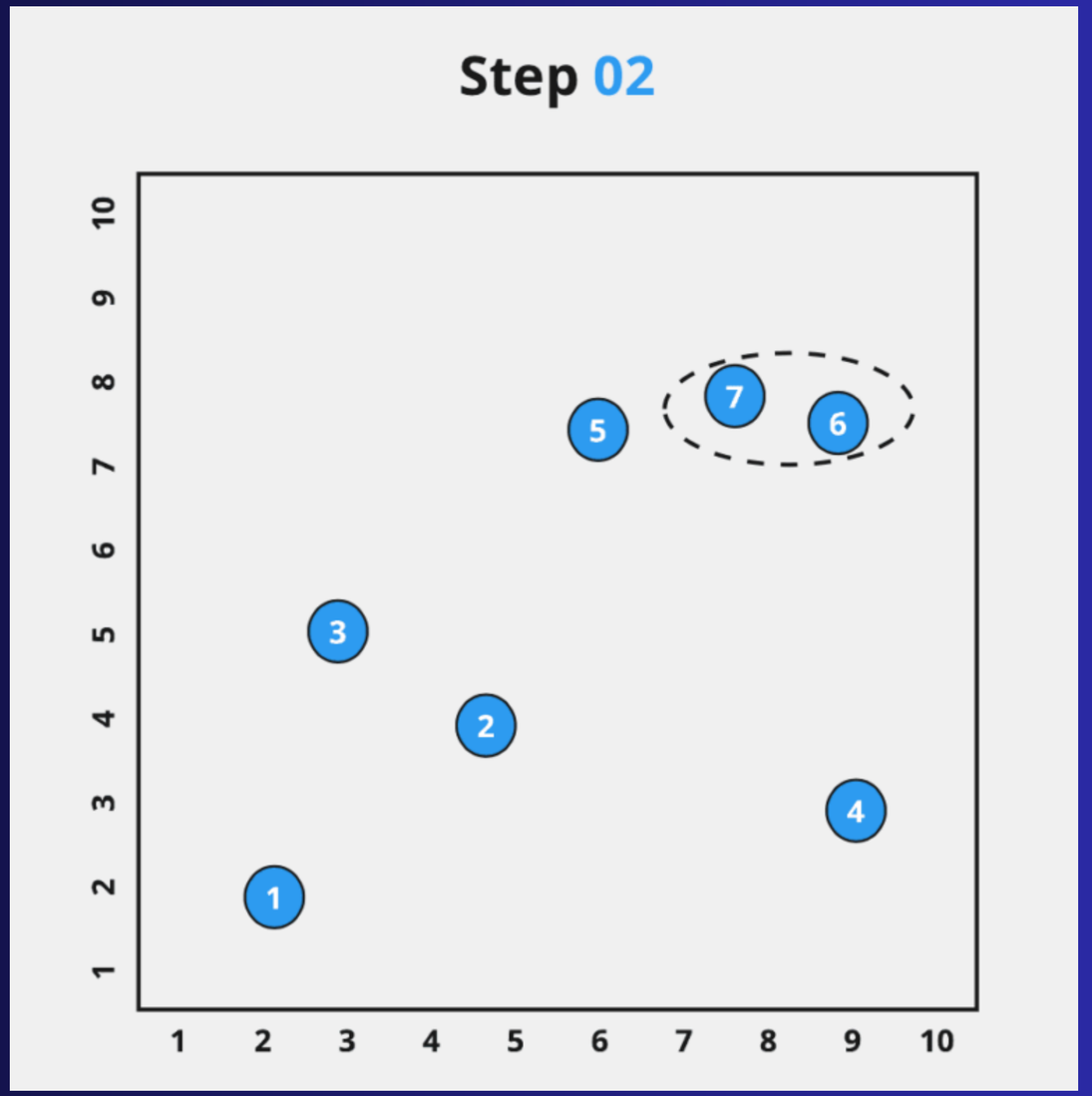
Liaison semple(min):
Distance minimale entre un point de C_i et un point de C_j
Liasion Complet(max):
Distance maximale enter un point de C_i et un point de C_j
Liasion moyenne :
Distance moyenne entre tous les points de C_i et tous les points de C_j

Déroulement de l'algorithme AGENS

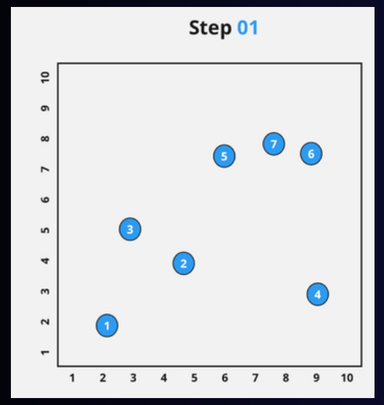


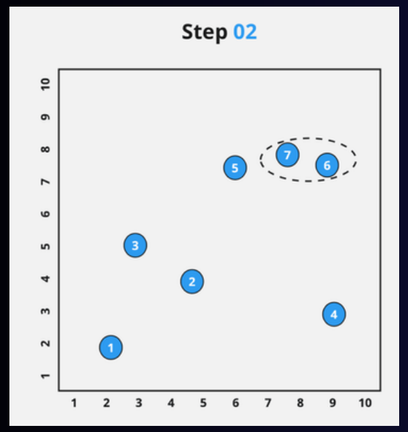
Déroulement de l'algorithme AGENS

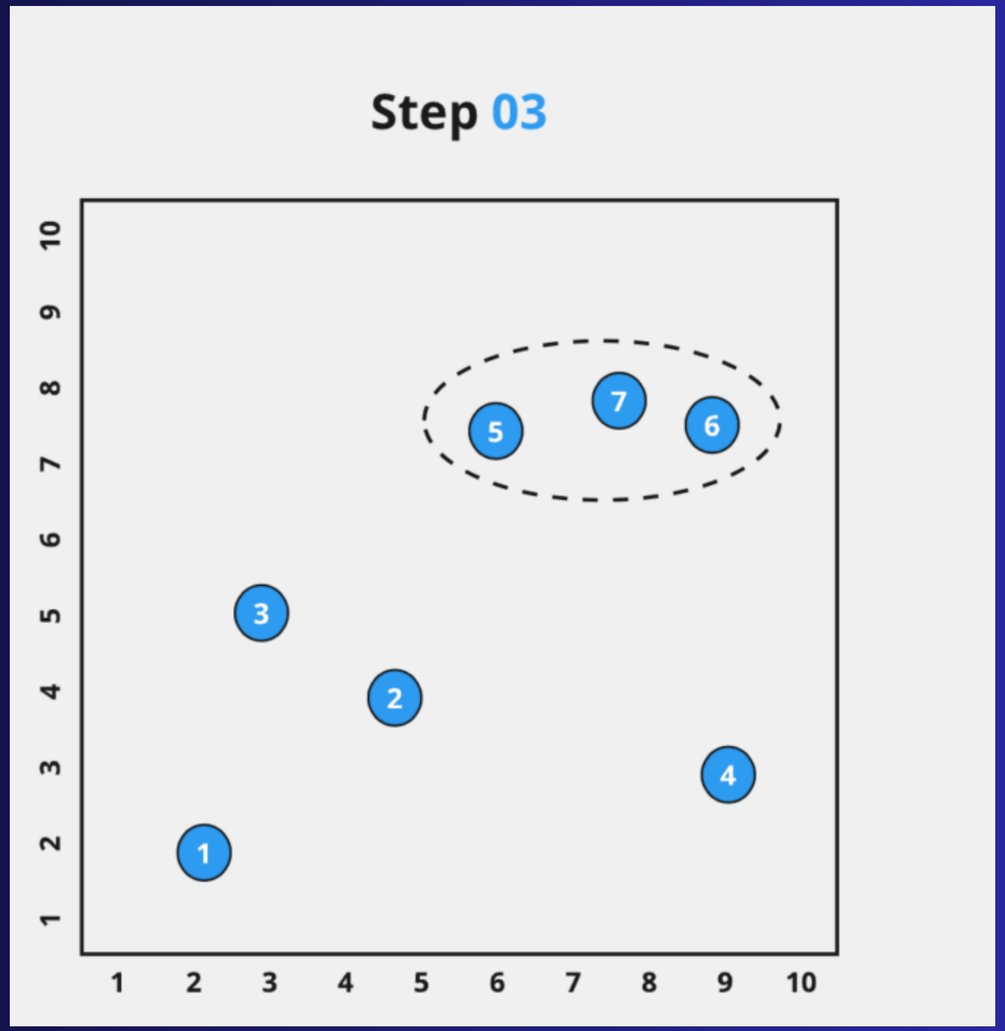




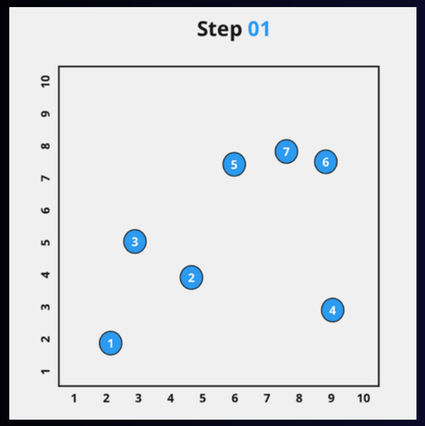
Déroulement de l'algorithme AGENS

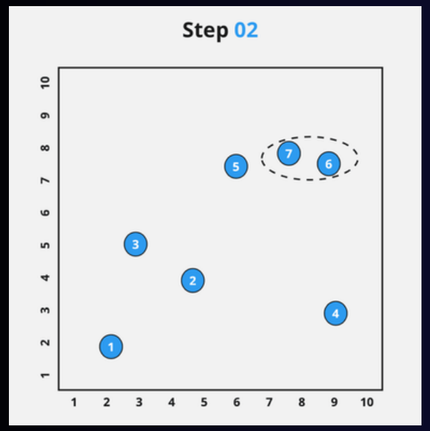


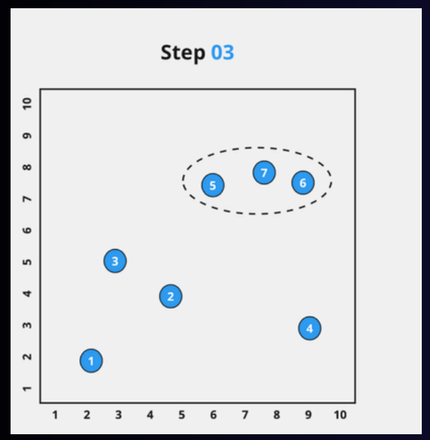


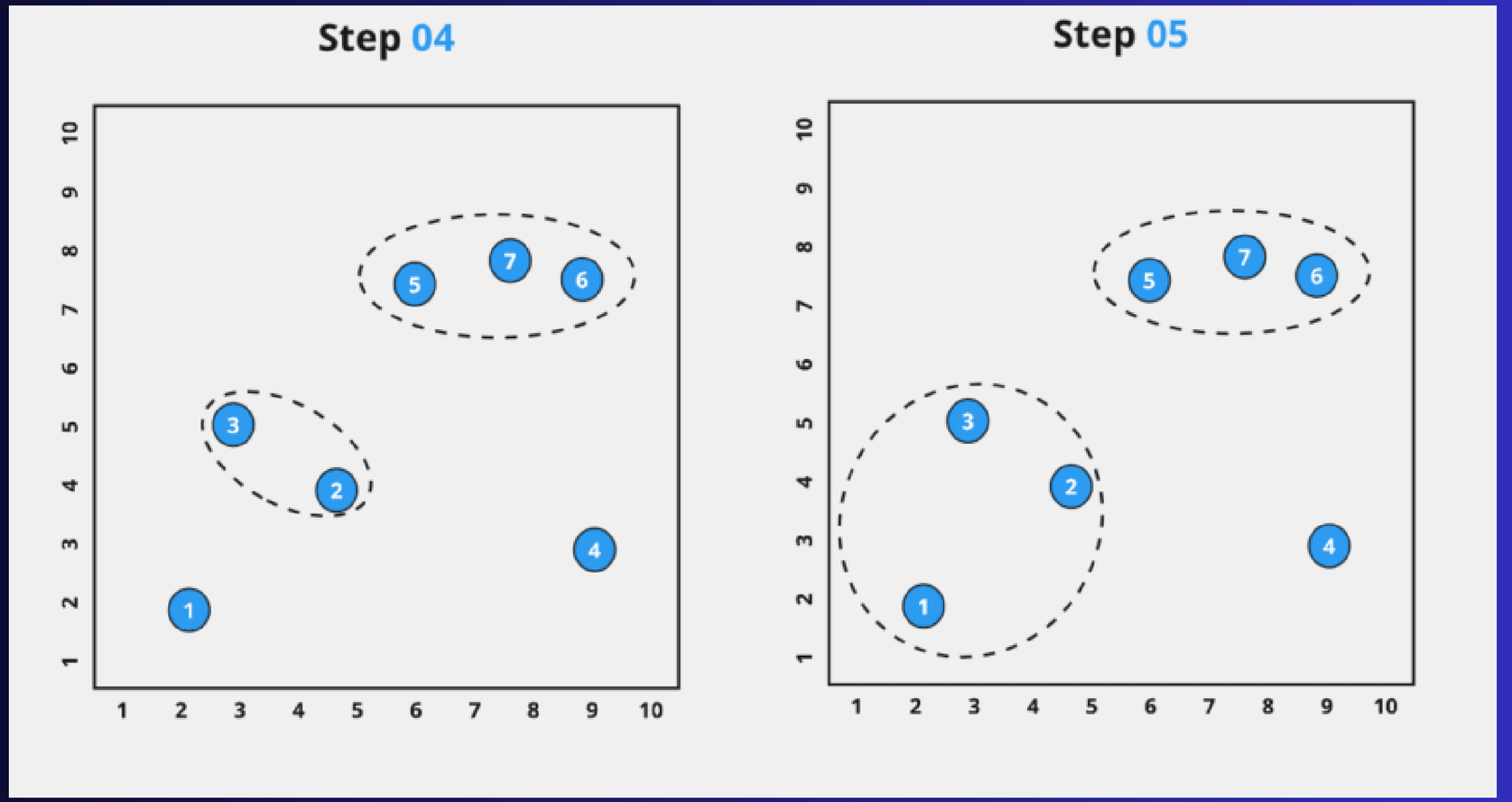


Déroulement de l'algorithme AGENS

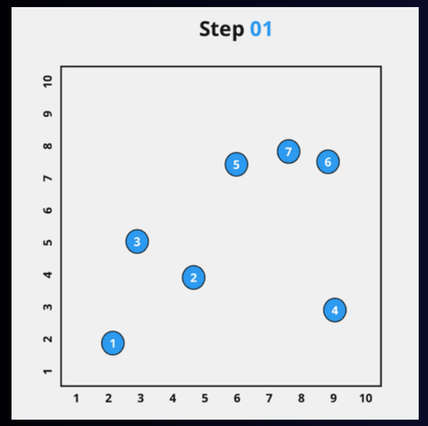


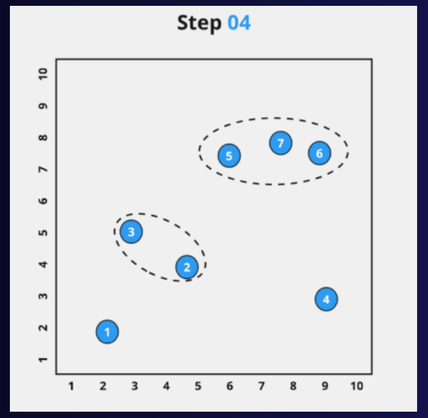


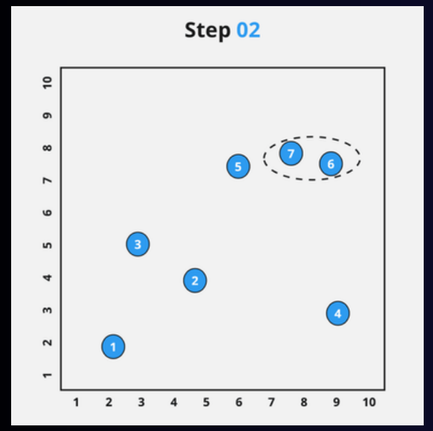


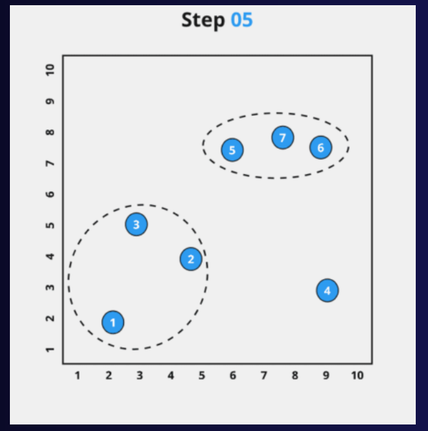


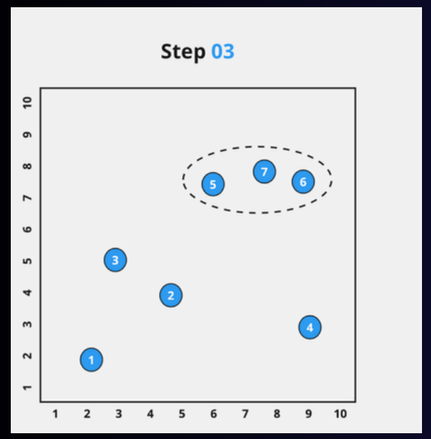
Déroulement de l'algorithme AGENS

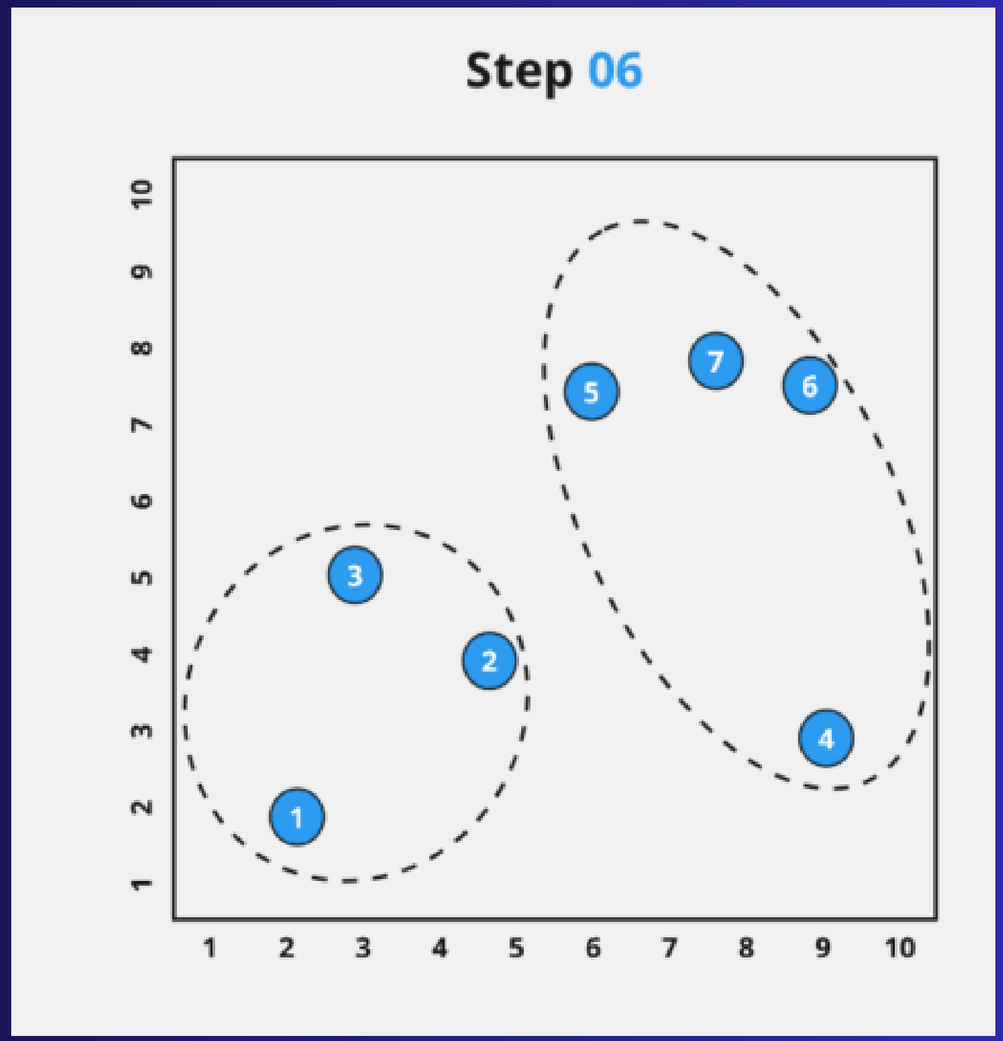




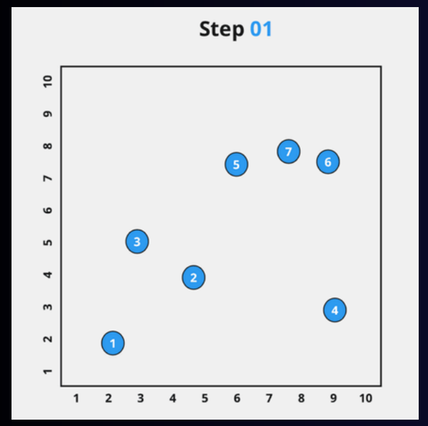


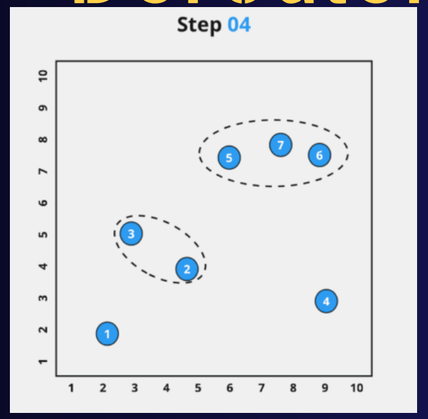


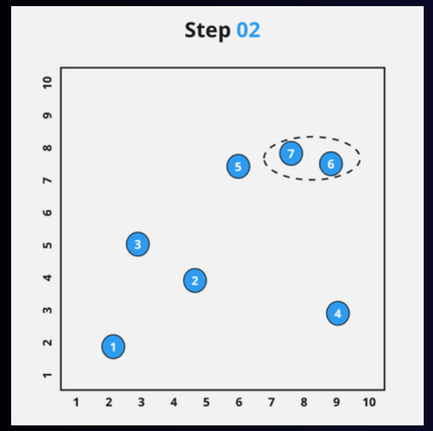


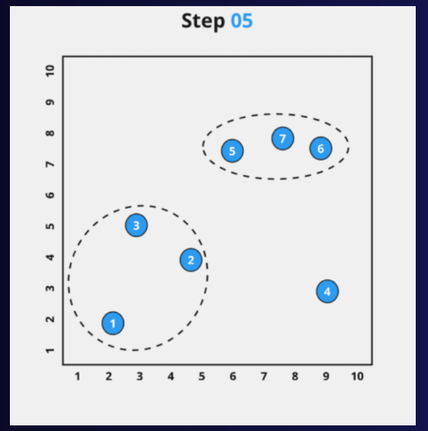


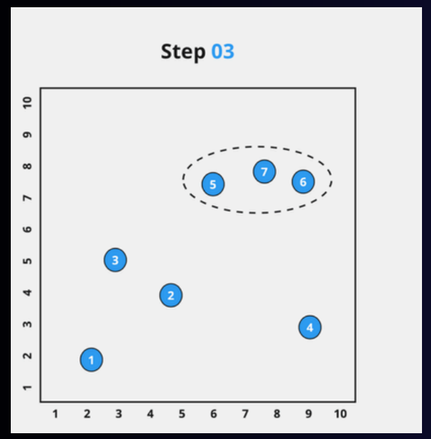
Déroulement de l'algorithme AGENS

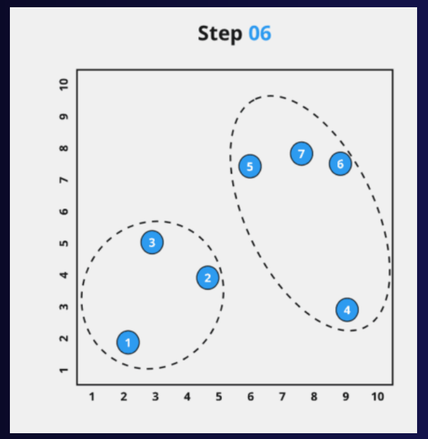


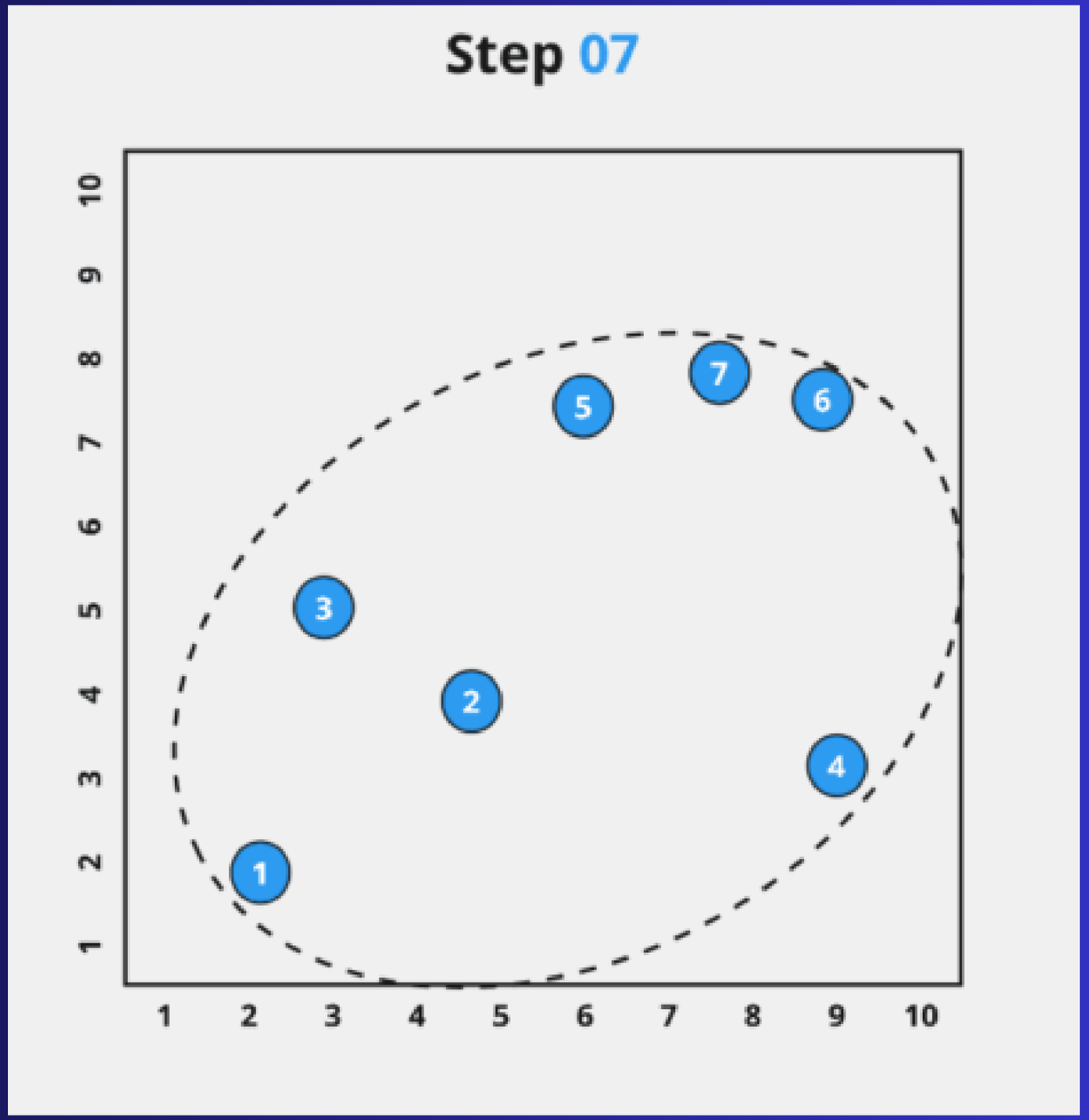








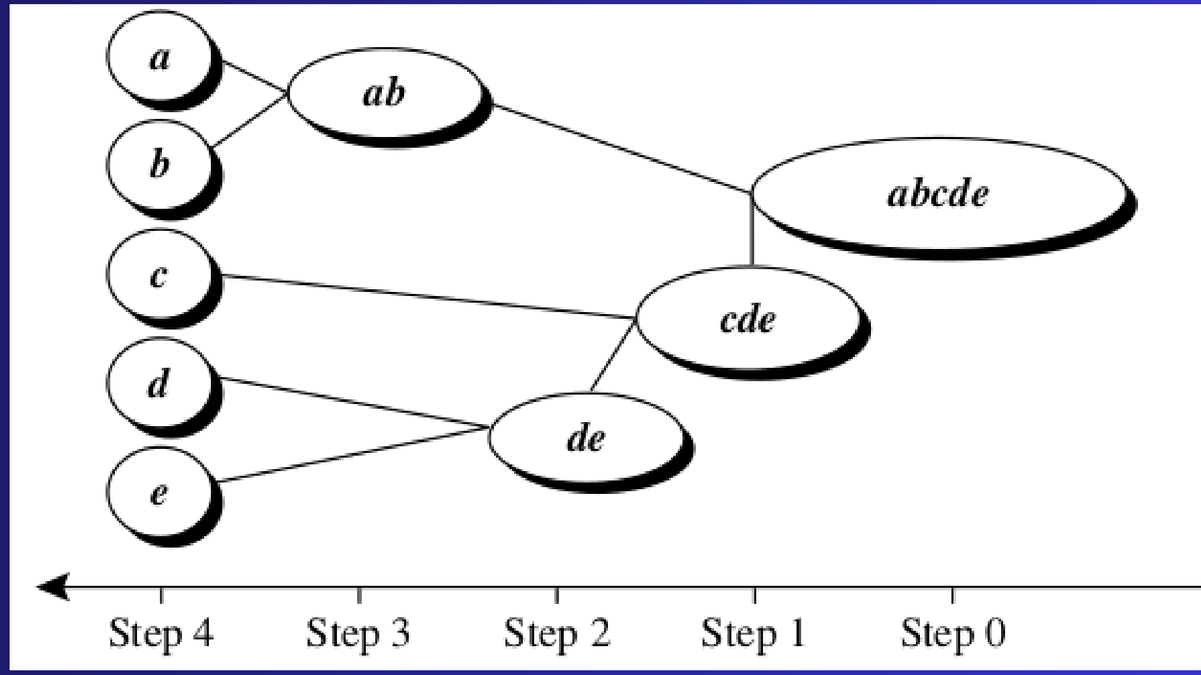




DIANA

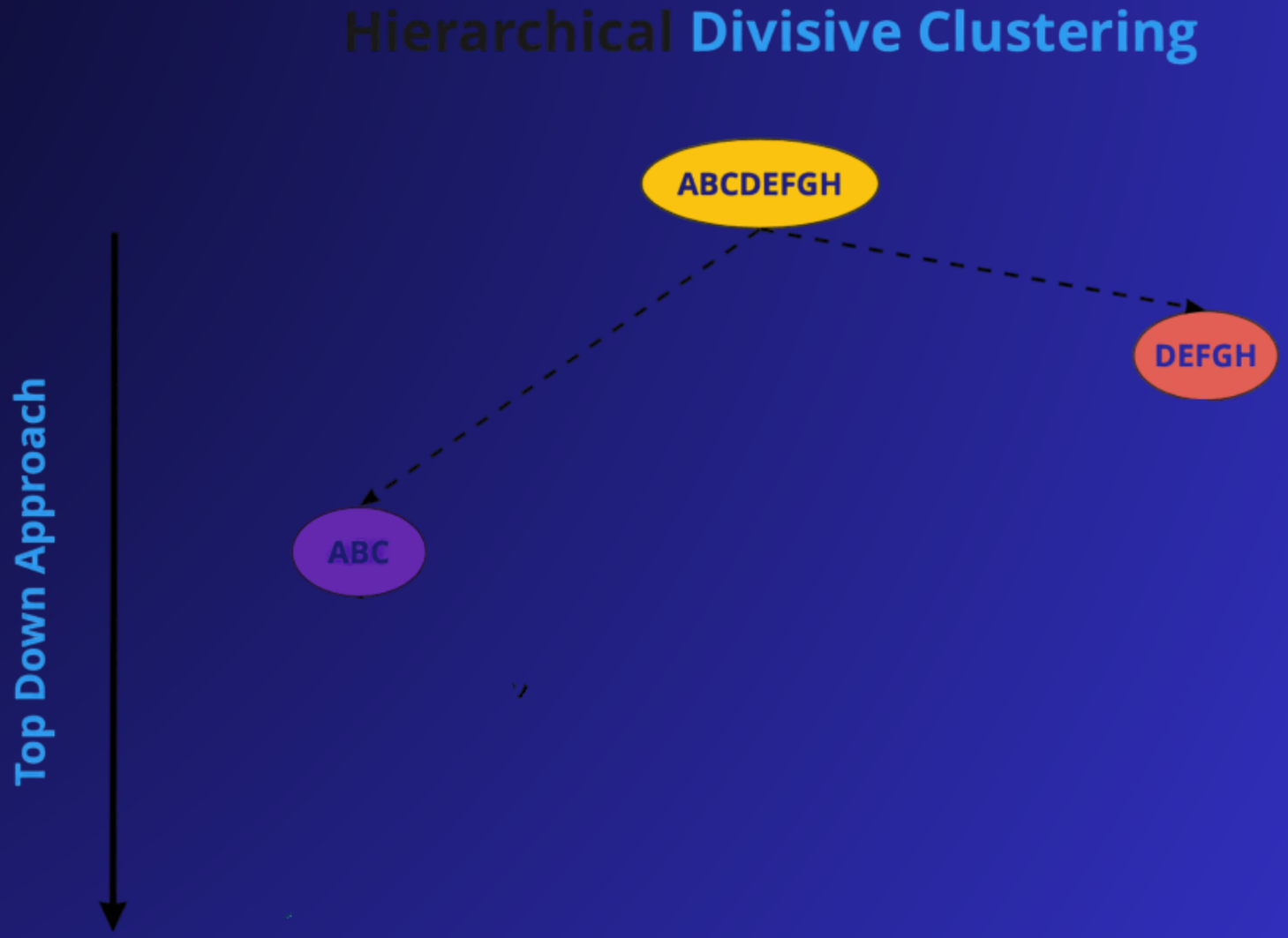
Il s'agit d'une approche de clustering descendante. Son fonctionnement est similaire à celui du clustering agglomératif, mais dans le sens inverse .

- Commencez par un seul cluster
- Calculez la mesure de dissimilarité entre toutes les paires de points et identifiez le point de données le plus disparate.
- Affectez les points identifiés à un autre cluster.
- Continuer le processus de division sur les clusters créés jusqu'à ce que chaque point forme un cluster individuel.



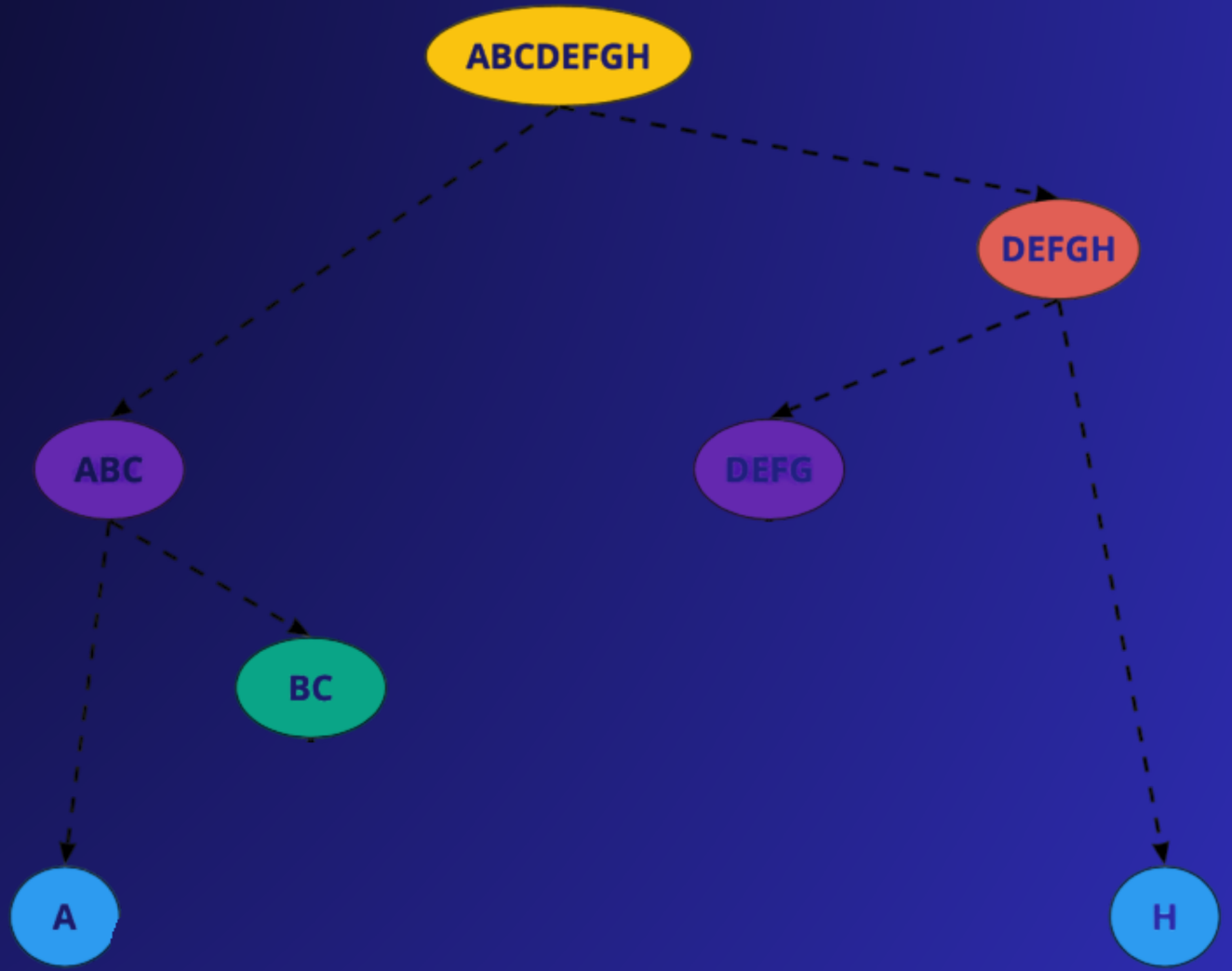
Déroulement de l'algorithme DIANA

1



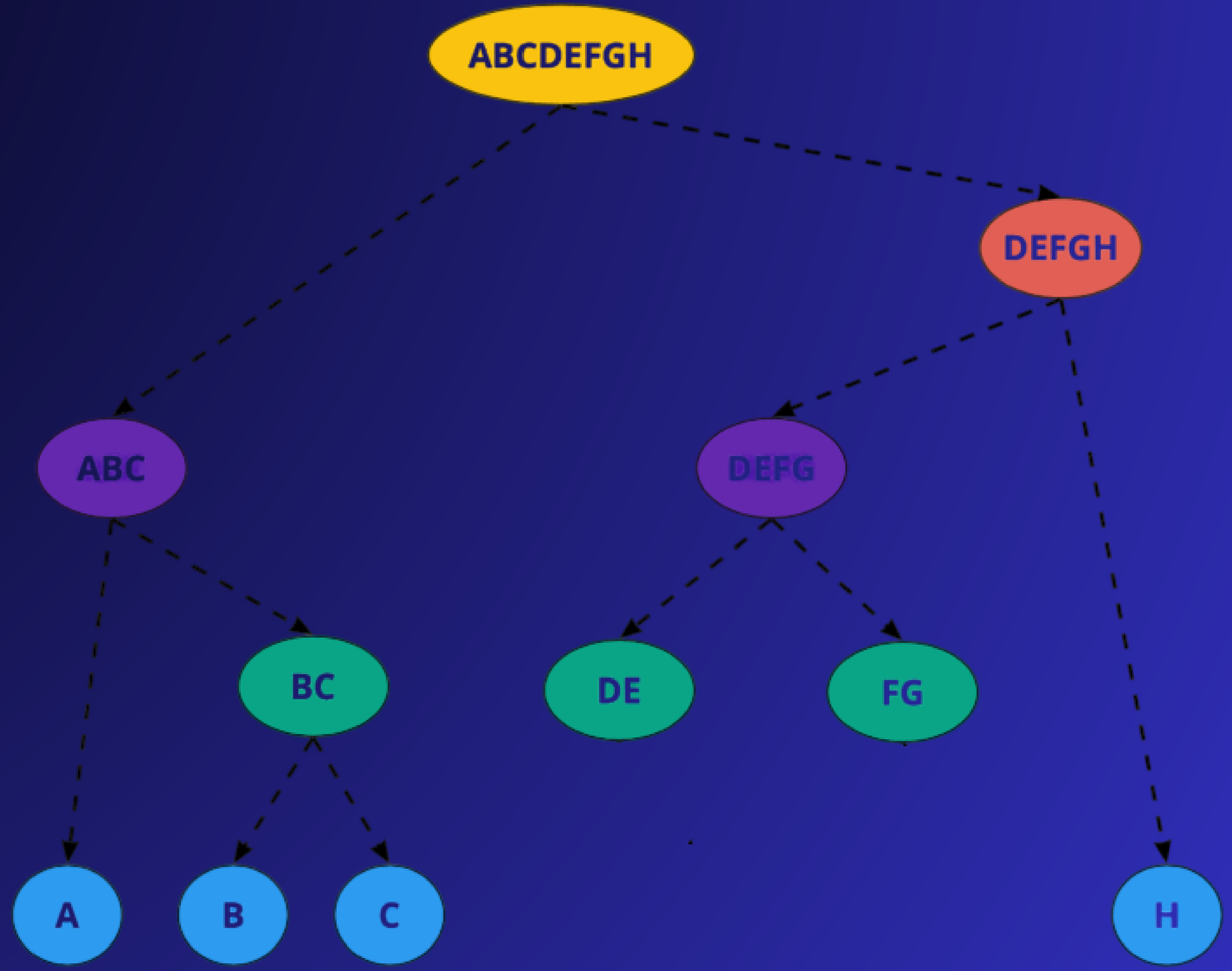
Déroulement de l'algorithme DIANA

2



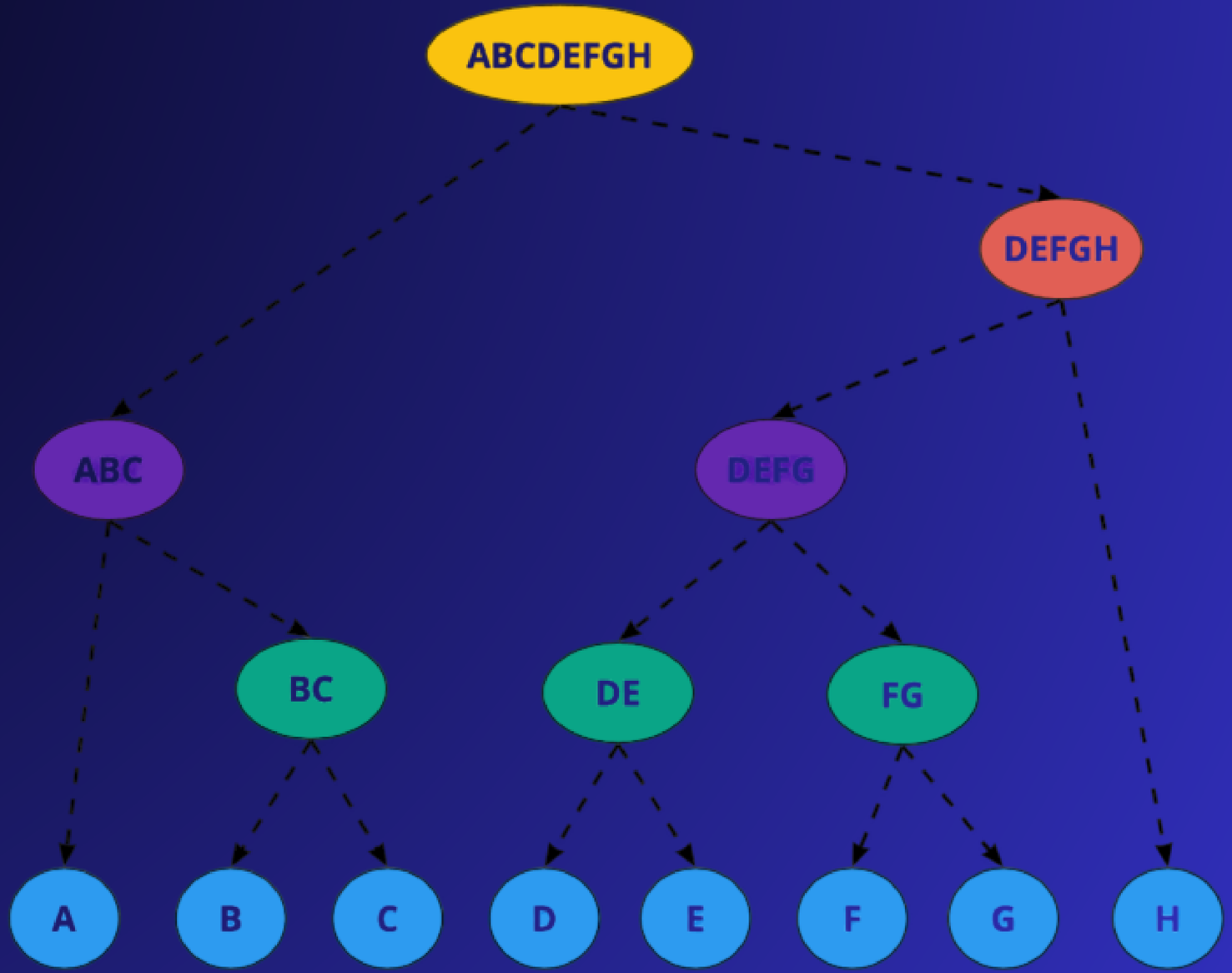
Déroulement de l'algorithme DIANA

3



Déroulement de l'algorithme DIANA

4

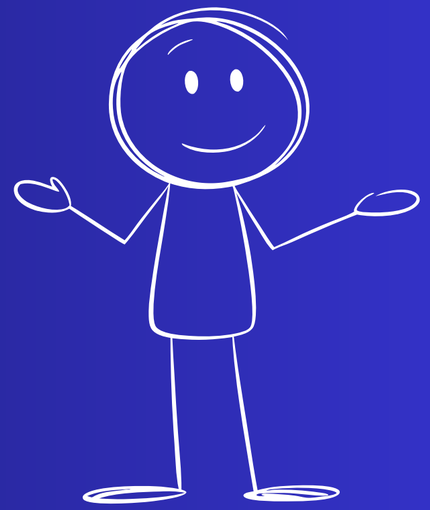


Résultat(AGENS, DIAINA)

le schéma final représente une arborescence claire et lisible des données. Ce regroupement hiérarchique offre un visuel instantané de l’ensemble de données. Cette arborescence de clusters porte le nom de dendrogramme.

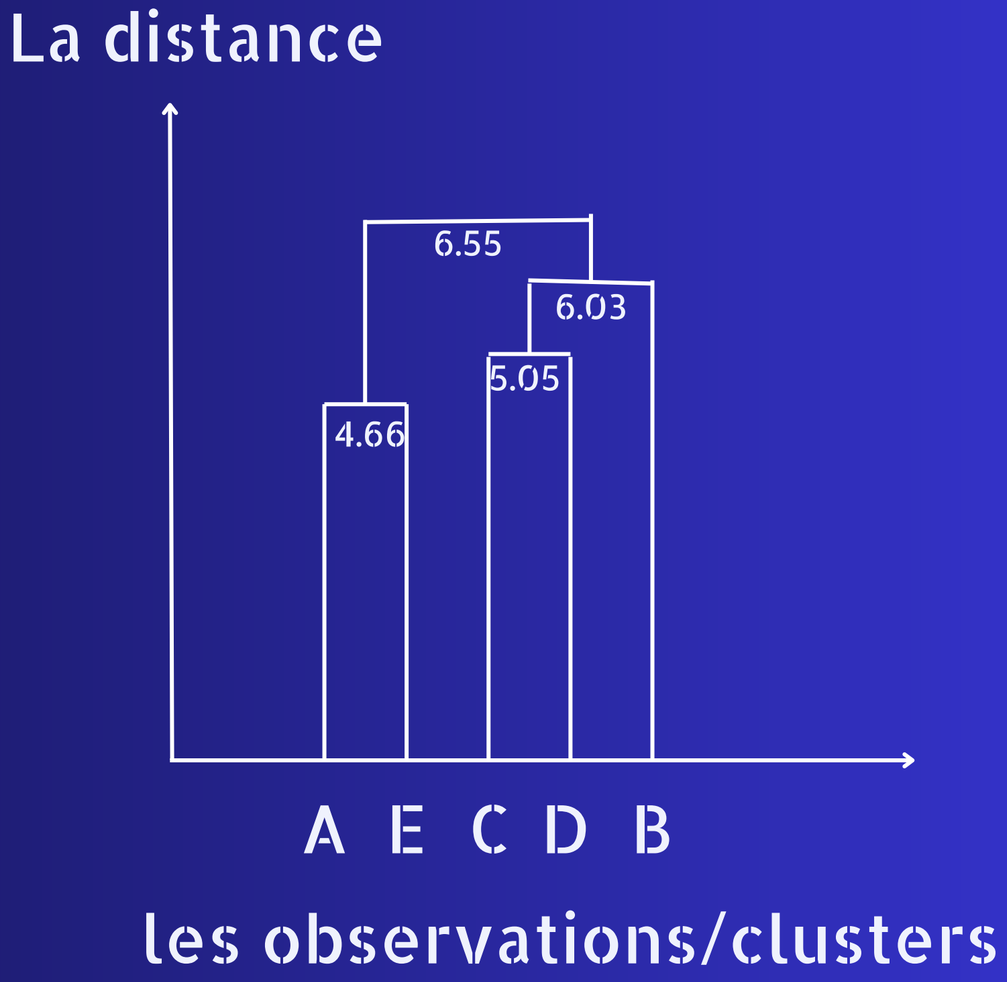
Qu'est-ce qu'un dendrogramme ?

Un Dendrogamme est un diagramme de regroupement hiérarchique, permettant d'organiser des données en arborescence en fonction de leurs similitudes.



Dendrogramme :

Le Dendrogramme est donc le type de diagramme en arborescence que l’on utilise pour présenter le clustering hiérarchique, à savoir les relations entre des ensembles de données similaires.



Faculté d'informatique
Departement IA & SD

DATA MINING

TP5: Clustering DBSCAN

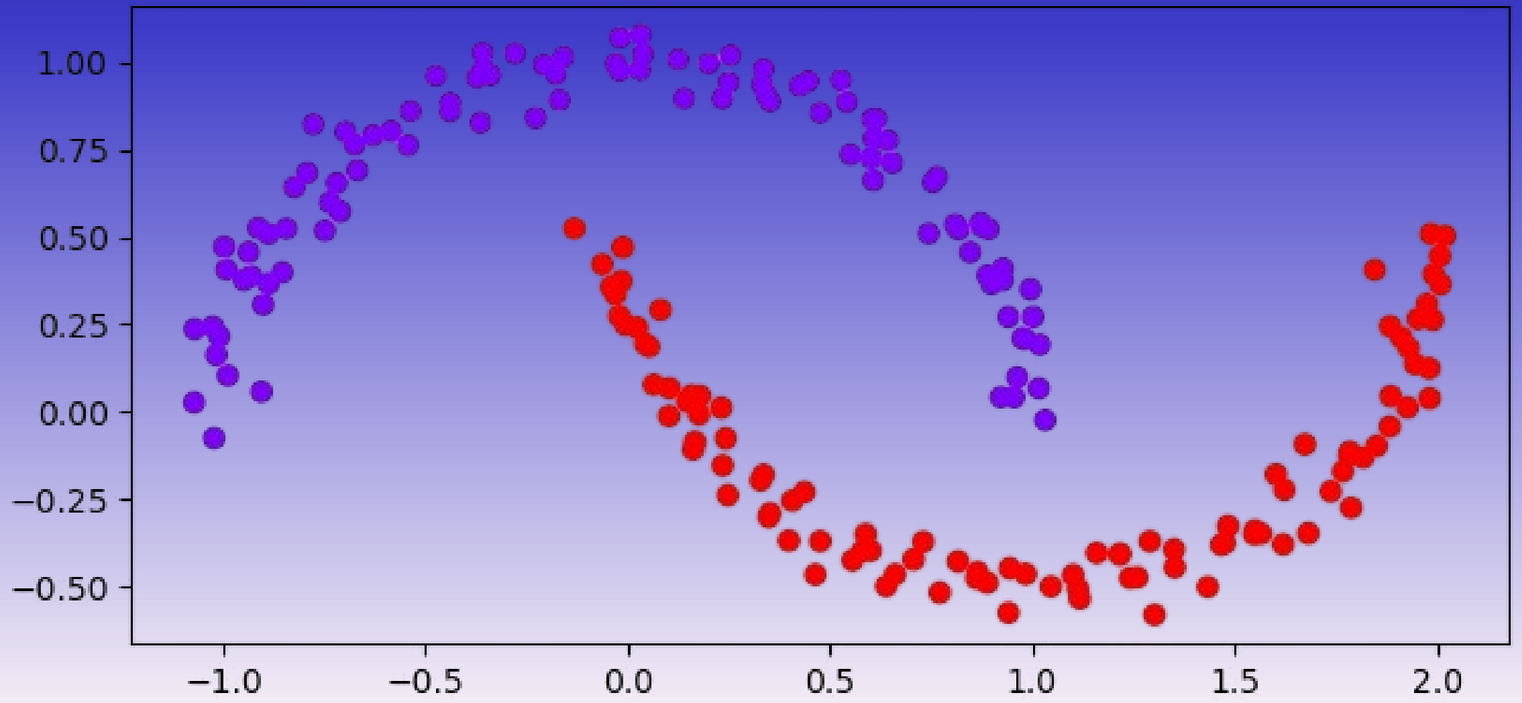
BENADJAL.N
Phd student in computer science
LAB: MOVEP
Email: nesrinebenadjel63@gmail.com

INGSOFT4

2025/2026

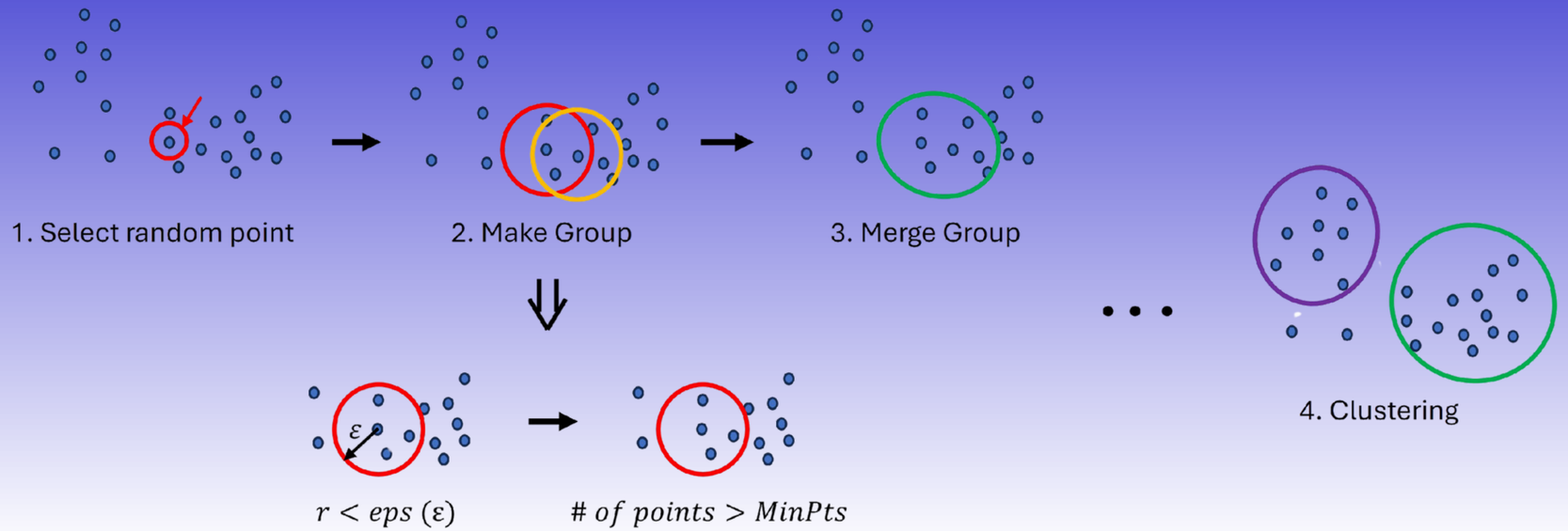
Introduction

Dans les séances de TP passées, nous avons découvert les algorithmes d'apprentissage automatique non supervisé hiérarchiques. Dans ce TP, nous allons voir un nouvel algorithme basé sur la densité : c’est l’algorithme de DBSCAN.



DBSCAN

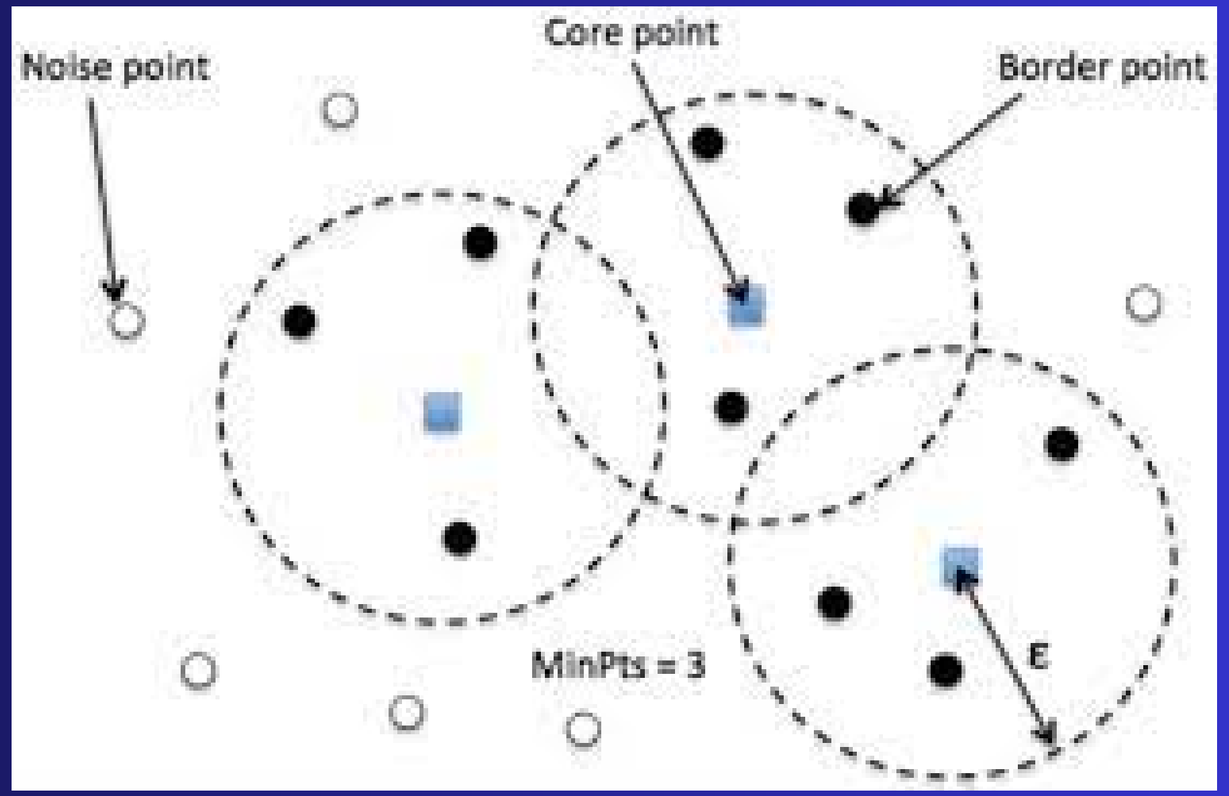
DBSCAN (Density-Based Spatial Clustering of Applications with Noise) est un algorithme de clustering qui regroupe les points de données selon leur densité, ce qui le rend idéal pour détecter des clusters de formes arbitraires. Contrairement aux algorithmes de clustering basés sur les centroïdes, tels que K-Means, DBSCAN ne nécessite pas de spécifier le nombre de clusters à l'avance. Il identifie également les valeurs aberrantes comme du bruit, ce qui le rend robuste pour les jeux de données présentant des anomalies.



Paramètres de l'algorithme DBSCAN

Epsilon (ε):
Epsilon définit le rayon maximal à l'intérieur duquel deux points voisins sont considérés comme appartenant au même groupe.

MinPts:
MinPts spécifie le nombre minimal de points requis pour former une région dense, qui définit le cœur d'un cluster.

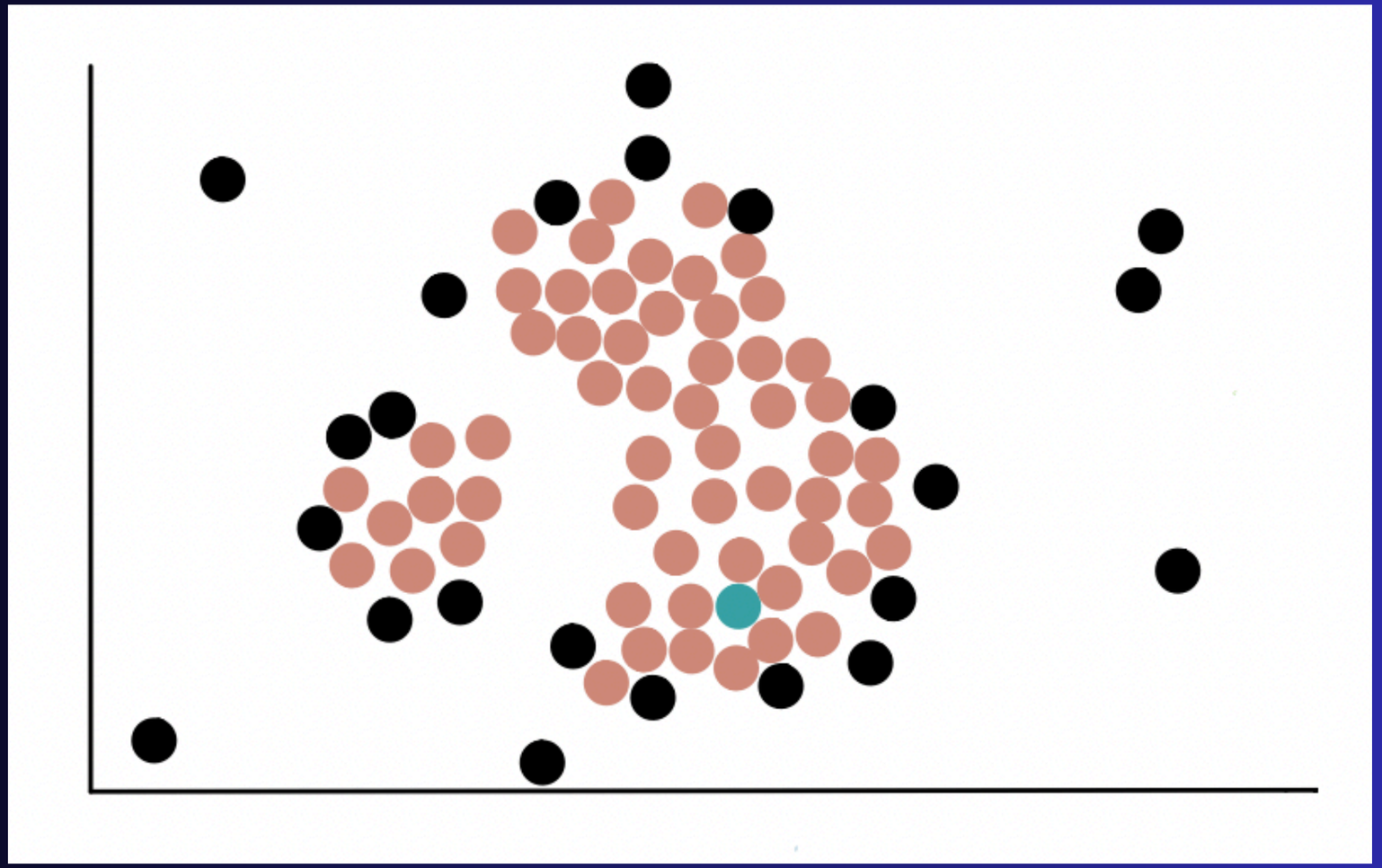


Pseudocode pour DBSCAN

```
for each unvisited point P:
    mark P as visited
    neighbors = find_neighbors(P, epsilon)
    if len(neighbors) < MinPts:
        label P as noise
    else:
        create new cluster
        expand_cluster(P, neighbors, epsilon, MinPts)
```

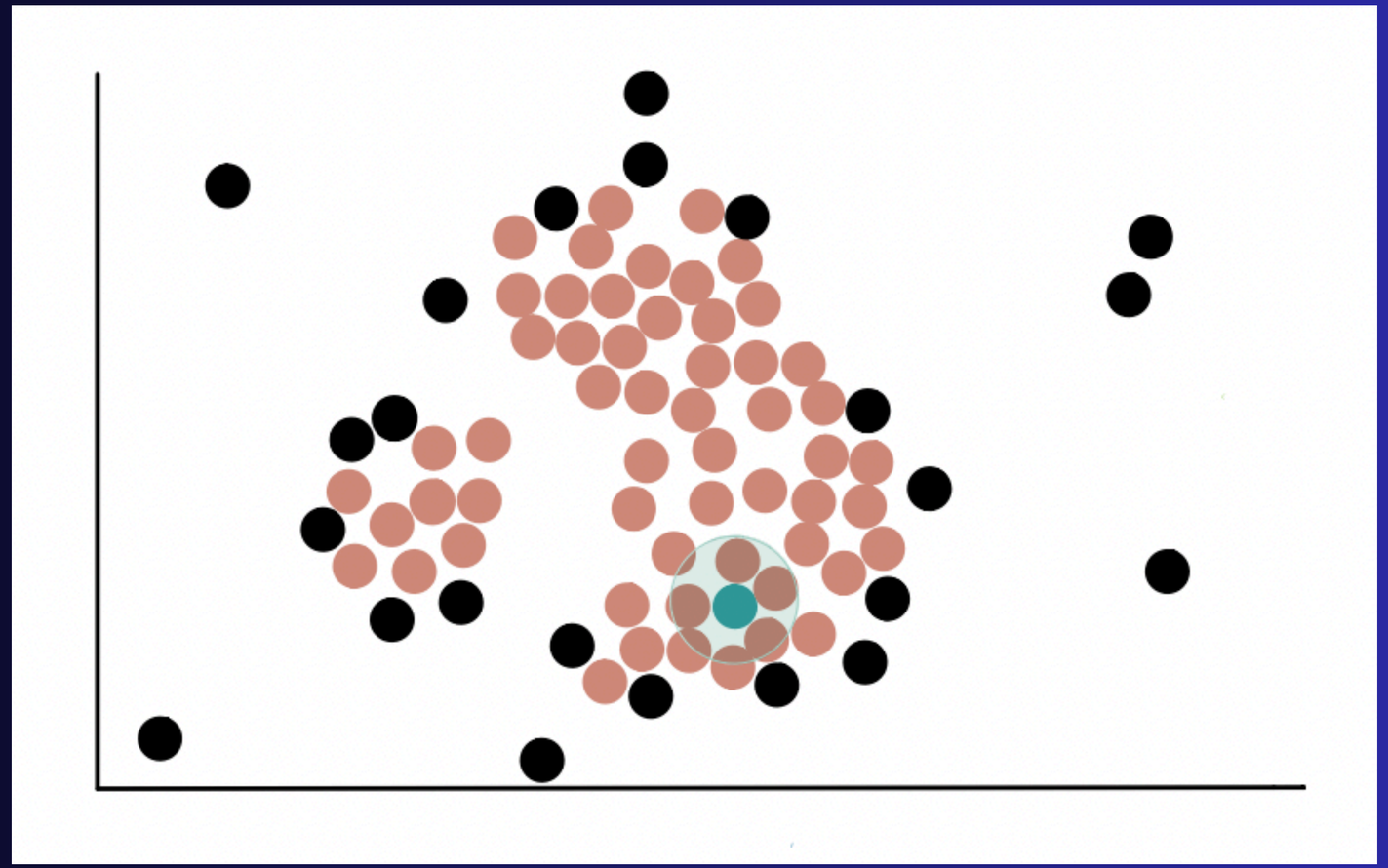
Déroulement de l'algorithme DBSCAN

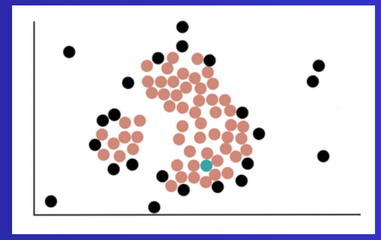
1



Déroulement de l'algorithme DBSCAN

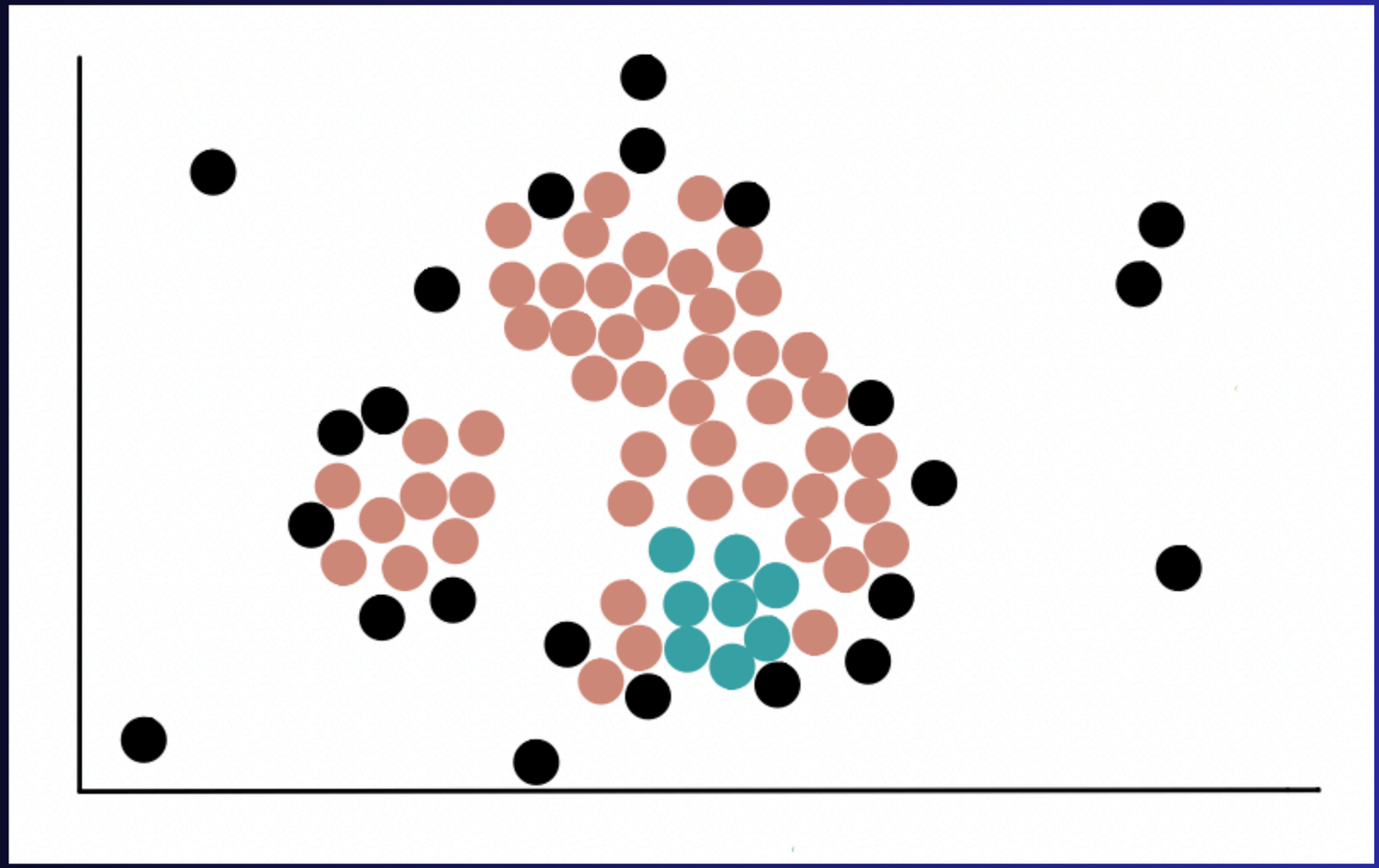
2

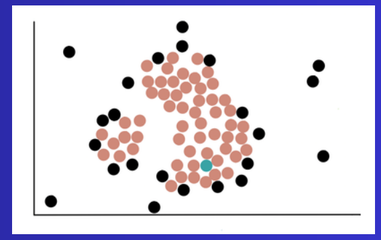


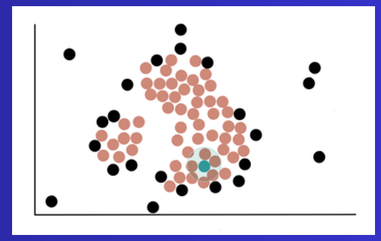


Déroulement de l'algorithme DBSCAN

3

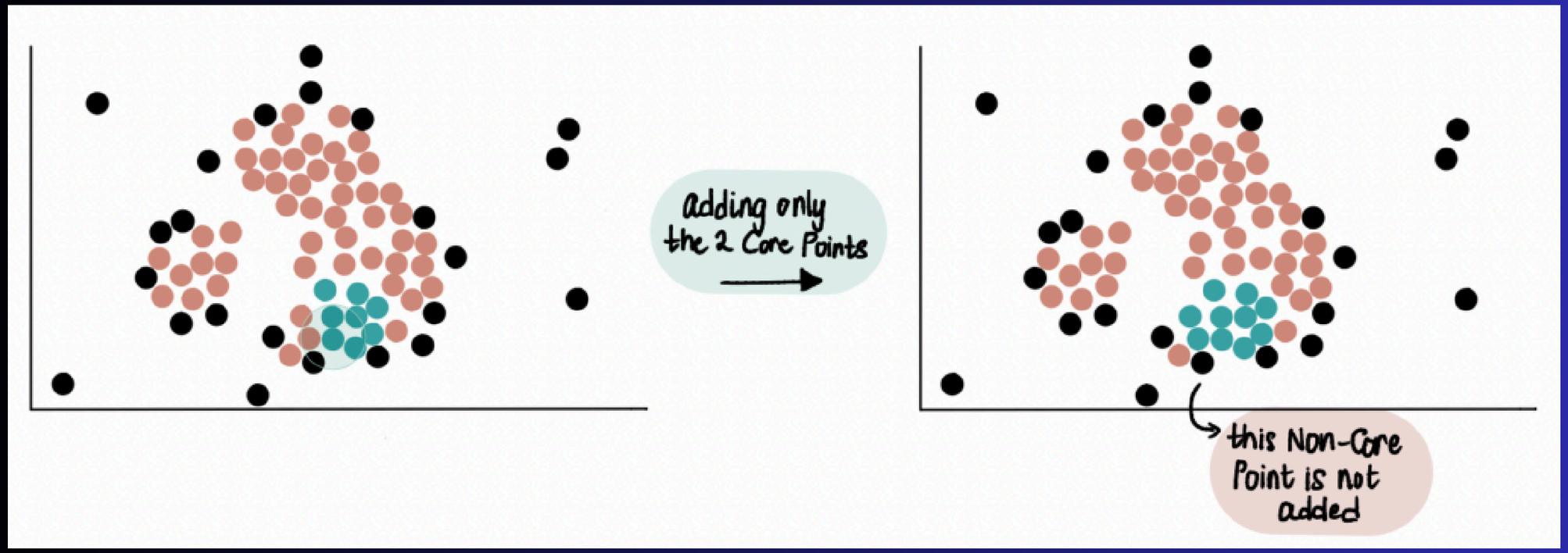


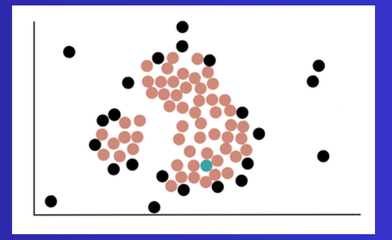


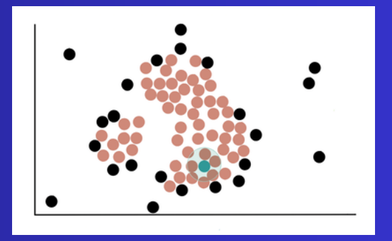


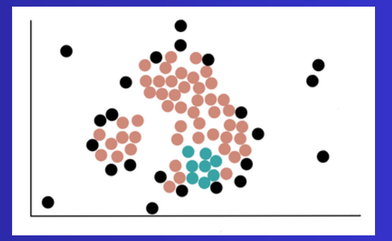
4

Déroulement de l'algorithme DBSCAN



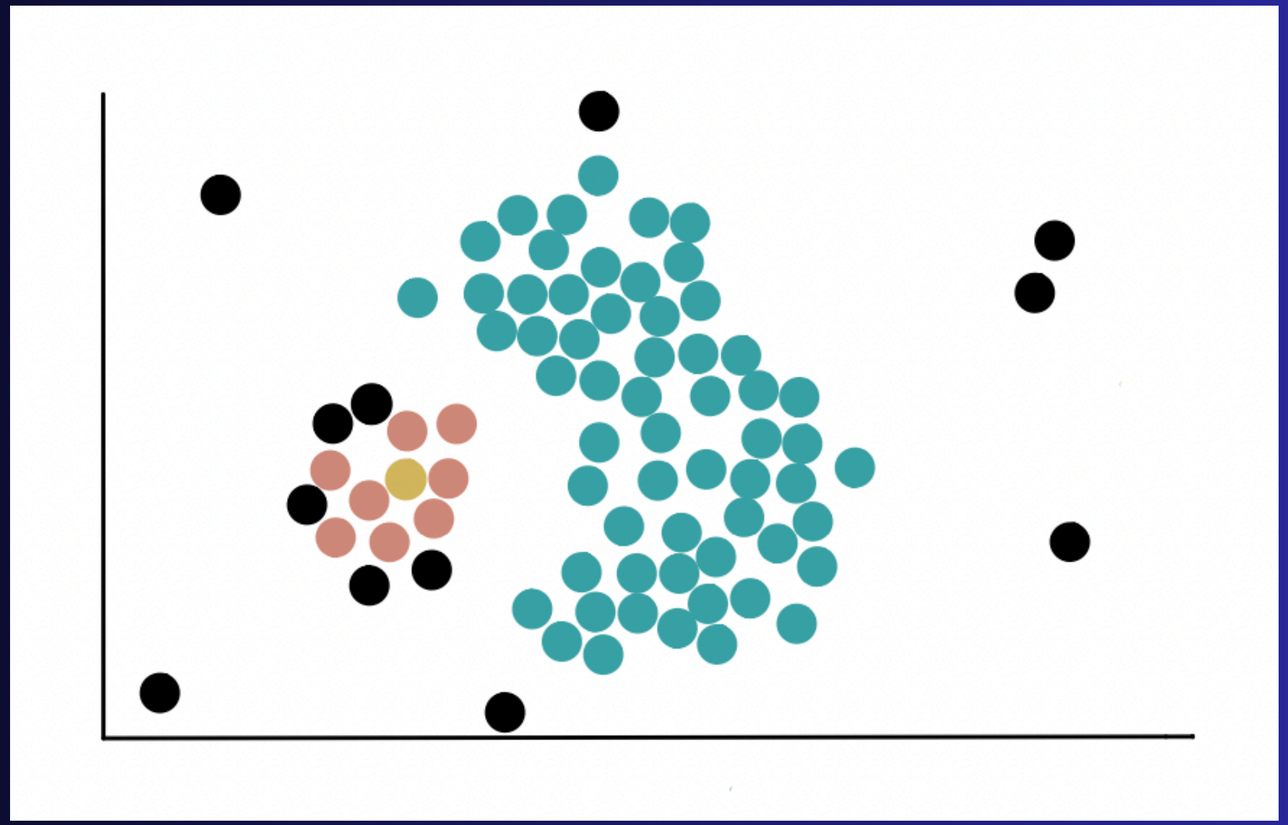


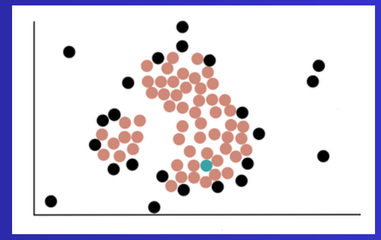


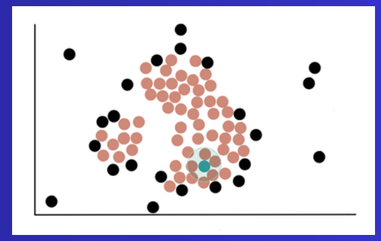


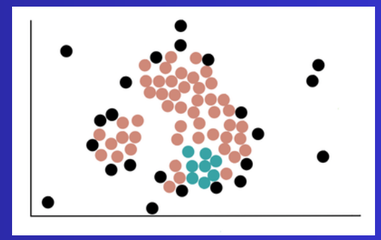
Déroulement de l'algorithme DBSCAN

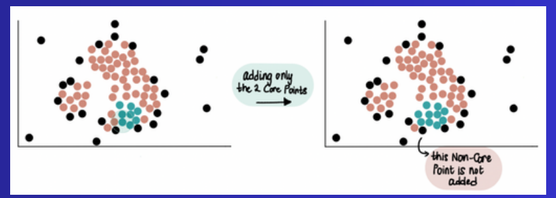
5











Déroulement de l'algorithme DBSCAN

6