Machine Learning

N. Hascoët – Prof. F. Chinesta

4 juillet 2019

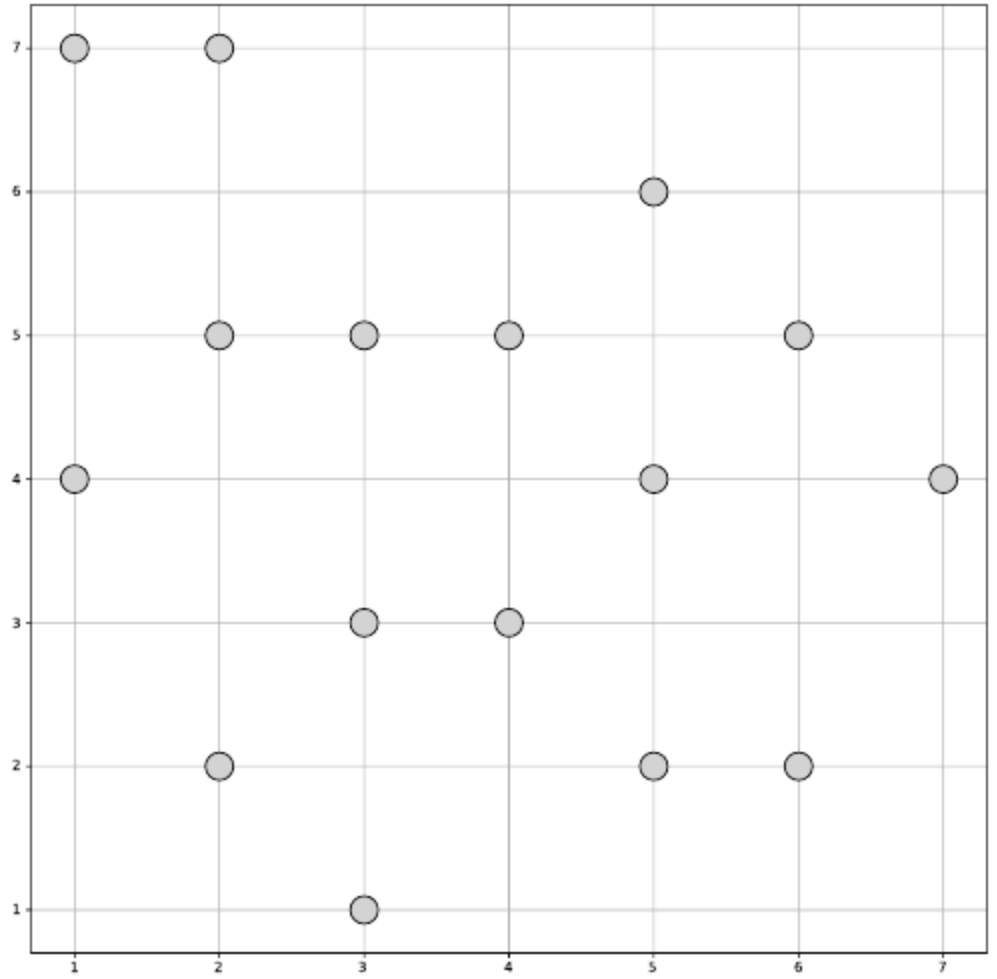
ARTS ET MÉTIERS ParisTech

Clustering

- Apprentissage NON-SUPERVISÉ
- Données de résultat INCONNU
- Nécessite une MÉTRIQUE ou un CRITÈRE
- Regrouper les données LES PLUS PROCHES

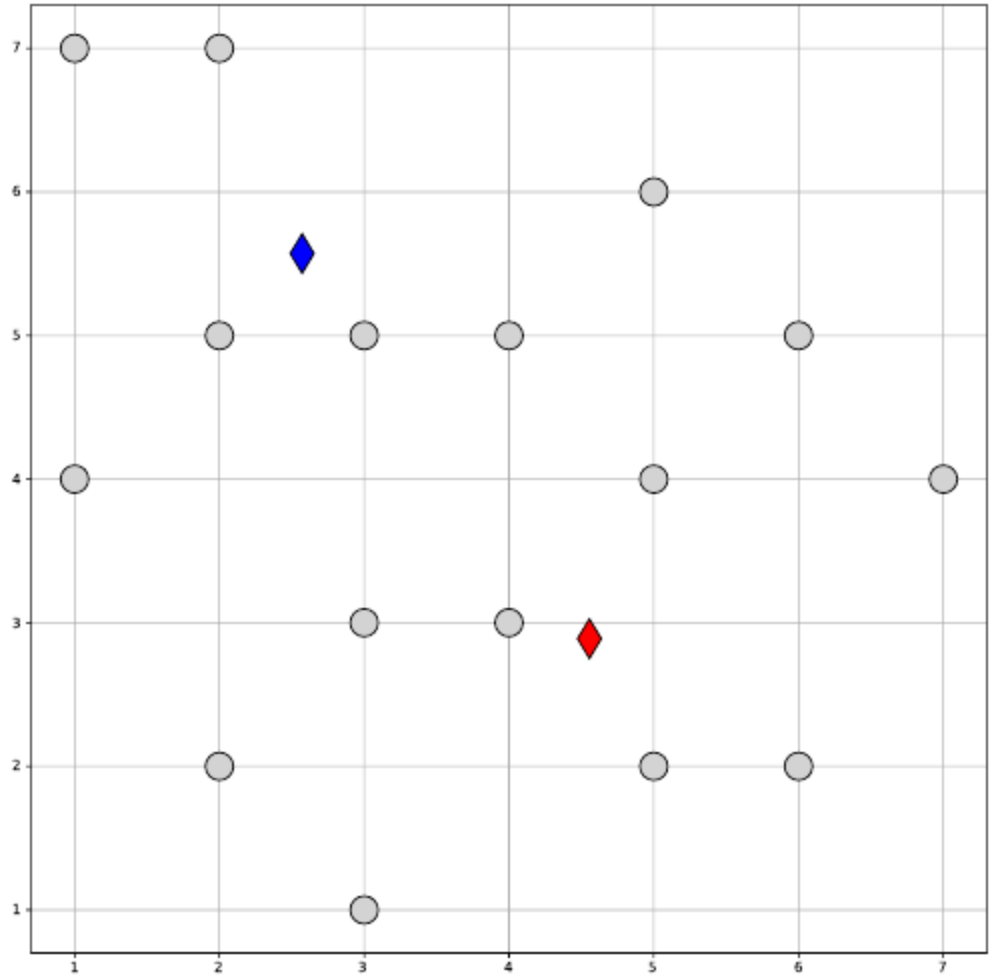
Clustering

- Jeu de données inconnues



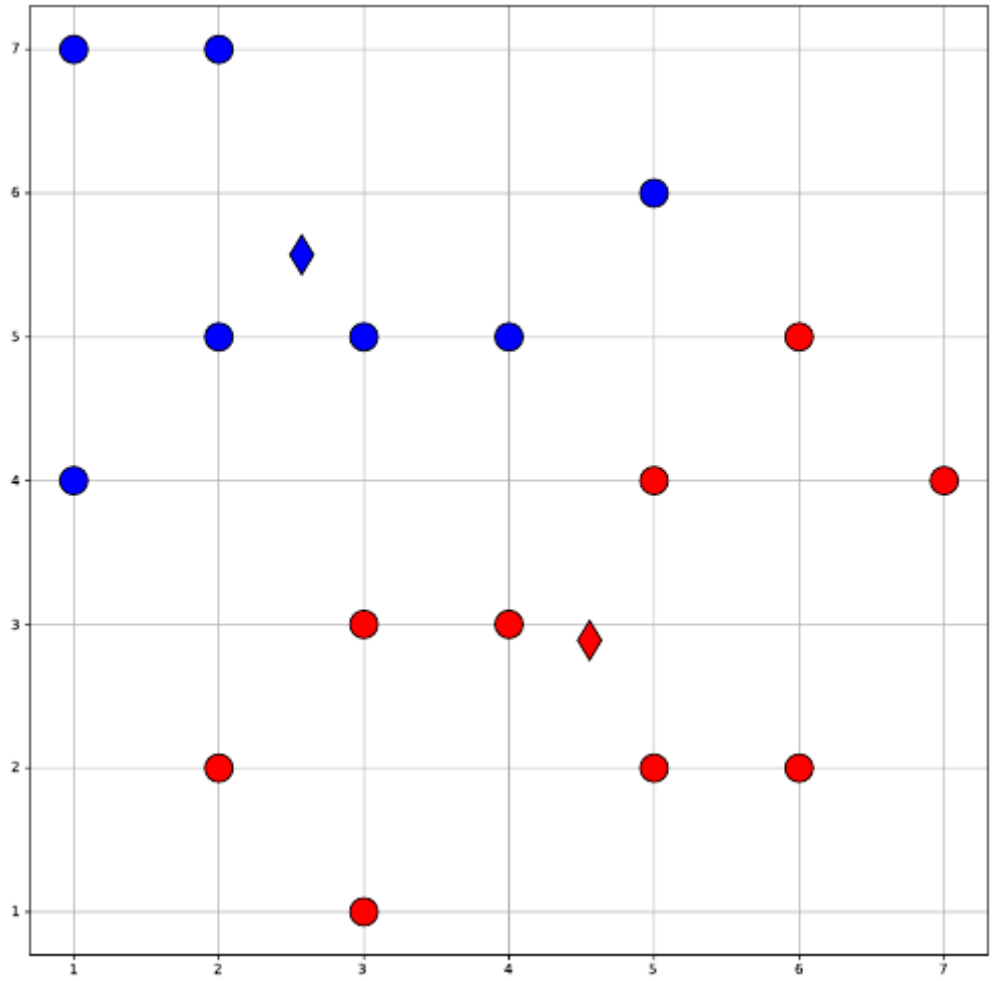
Clustering

- Jeu de données inconnues
- Calcul/sélection de CENTROÏDES



Clustering

- Jeu de données inconnues
- Calcul/sélection de CENTROÏDES
- Attribution d'un CLUSTER à chaque donnée (suivant une distance prédéfinie)



K-moyennes

- Regroupe la donnée en échantillons de K GROUPES (à choisir)
- VARIANCE équivalente dans chaque groupe

$$V = \sum_j \sum_{x_i \to \mu_j} d(x_i, \mu_j)^2$$

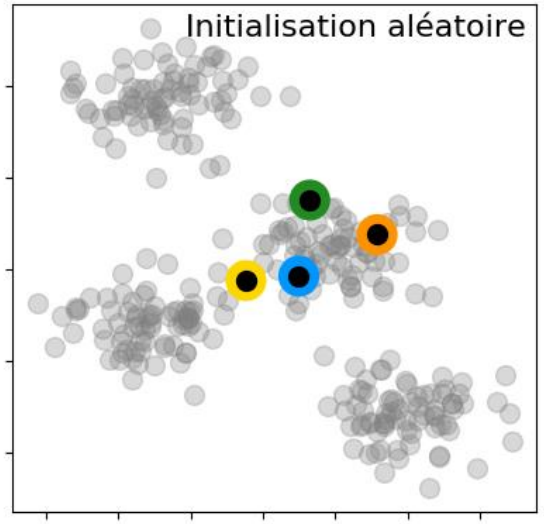
$\begin{cases} n & \text{nombre d'échantillons} \\ C & \text{ensemble des clusters} \\ \mu_j & \text{centroïde du cluster } j \end{cases}$

- Minimisation d'un critère d'INERTIE dans un cluster

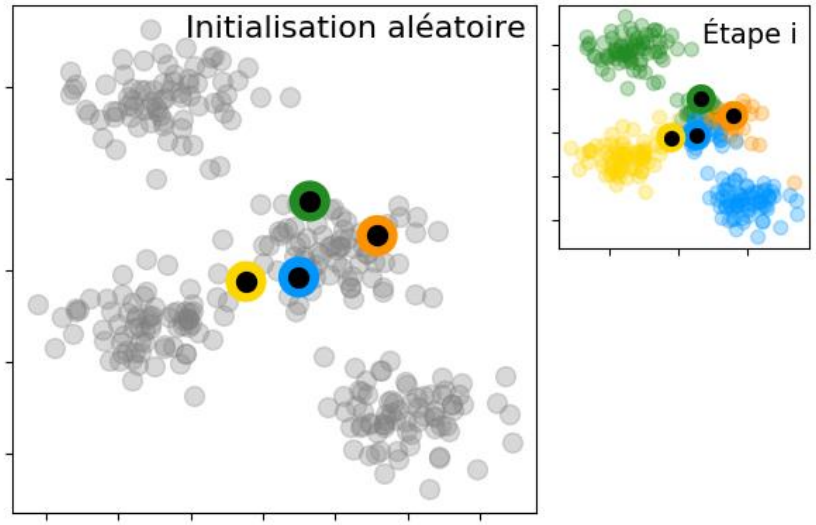
$$\sum_{i=1}^{n} \min_{\mu_j \in C} \left(\|x_i - \mu_j\|^2 \right)$$

- NORME euclidienne $d(a, b) = \sqrt{\sum_{i=1}^{n} (a_i - b_i)^2}$

K-moyennes

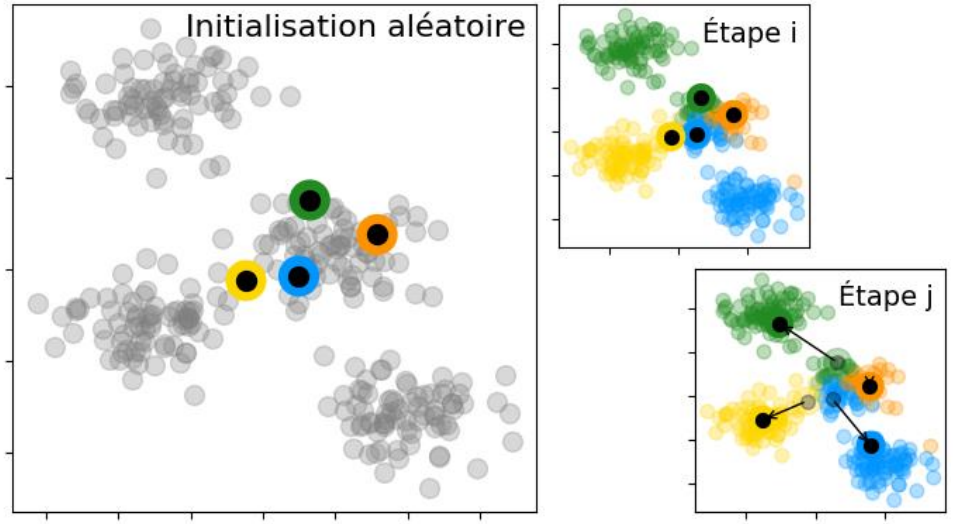


K-moyennes



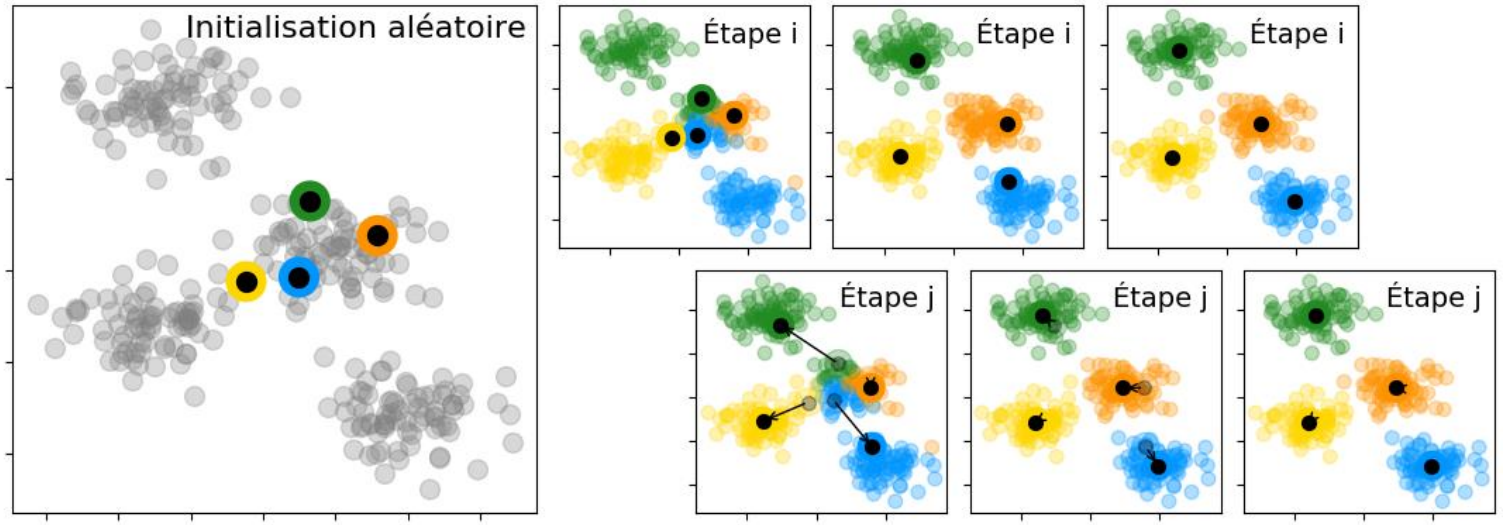
- Étape i : affecter chaque point au cluster dont le centroïde est le plus proche

K-moyennes



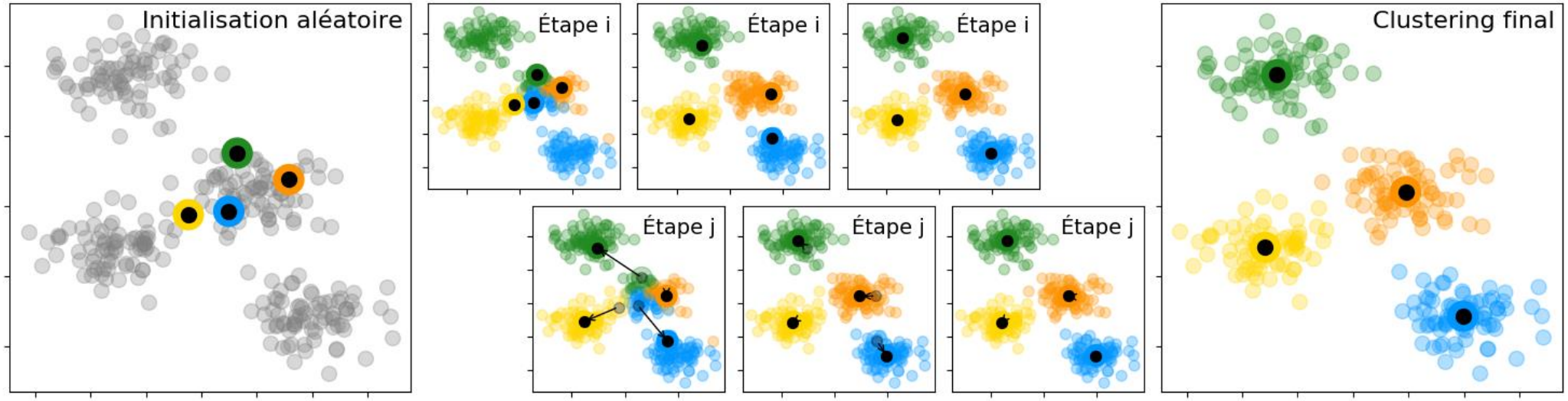
- Étape i : affecter chaque point au cluster dont le centroïde est le plus proche
- Étape j : recalculer la moyenne de chaque cluster et définir le centroïde

K-moyennes



- Étape i : affecter chaque point au cluster dont le centroïde est le plus proche
- Étape j : recalculer la moyenne de chaque cluster et définir le centroïde
- Répéter (i) et (j) jusqu'à convergence

K-moyennes



- Étape i : affecter chaque point au cluster dont le centroïde est le plus proche
- Étape j : recalculer la moyenne de chaque cluster et définir le centroïde
- Répéter (i) et (j) jusqu'à convergence

K-moyennes

- En paramètre : le NOMBRE DE CLUSTERS K
- SCALABLE pour un grand nombre d'échantillons
- Adapté pour un nombre moyen de clusters
- Cas d'utilisation générique pour des clusters de TAILLES SIMILAIRES

Clustering hiérarchique

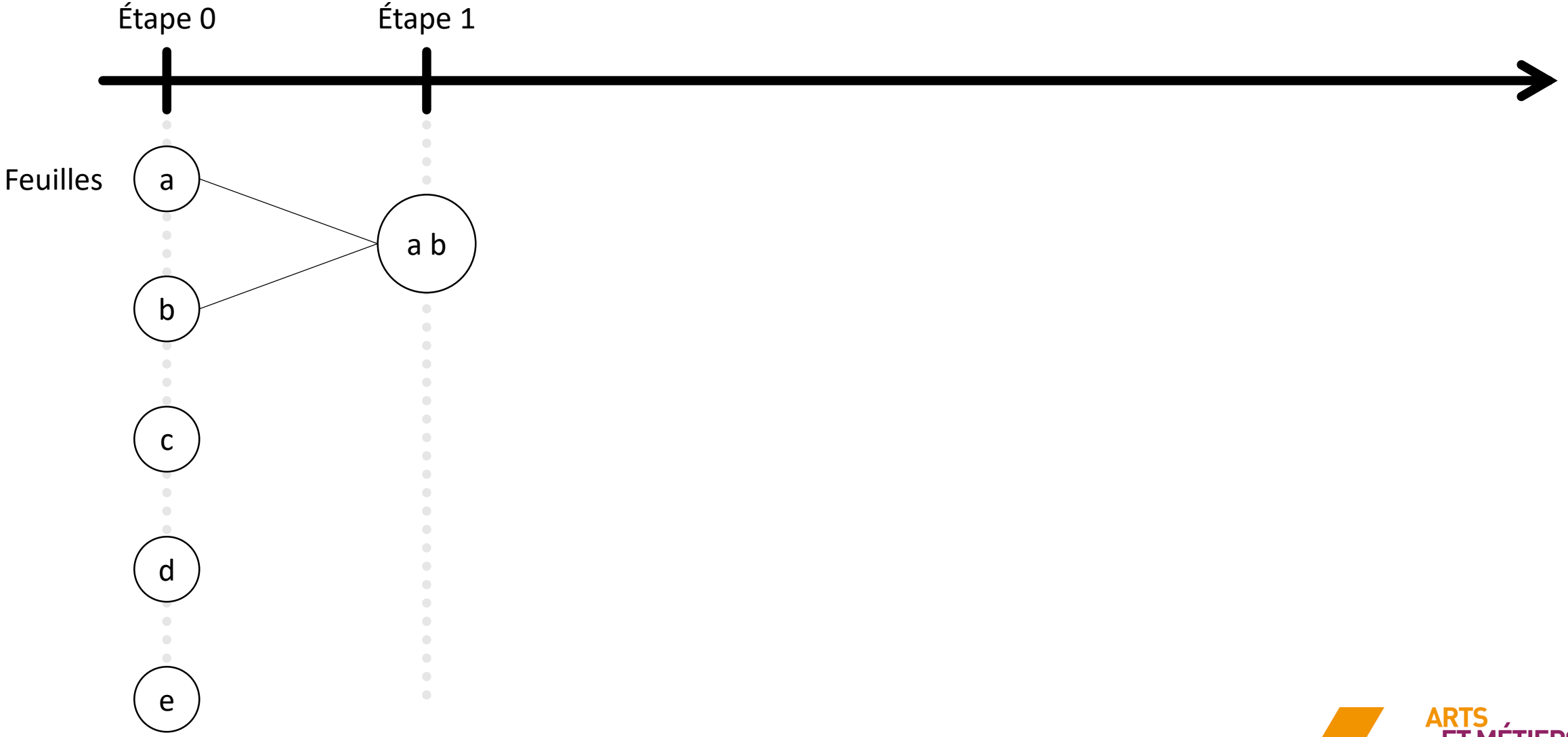
- Approche ASCENDANTE
- Initialement, chaque donnée x représente une classe C
- L'objectif est de diminuer le nombre de classes
- Définition d'une mesure de DISSIMILARITÉ par pair

$$C_1 = \{x_1\}, C_2 = \{x_2\}$$
$$dissim(C_1, C_2) = dissim(x_1, x_2)$$

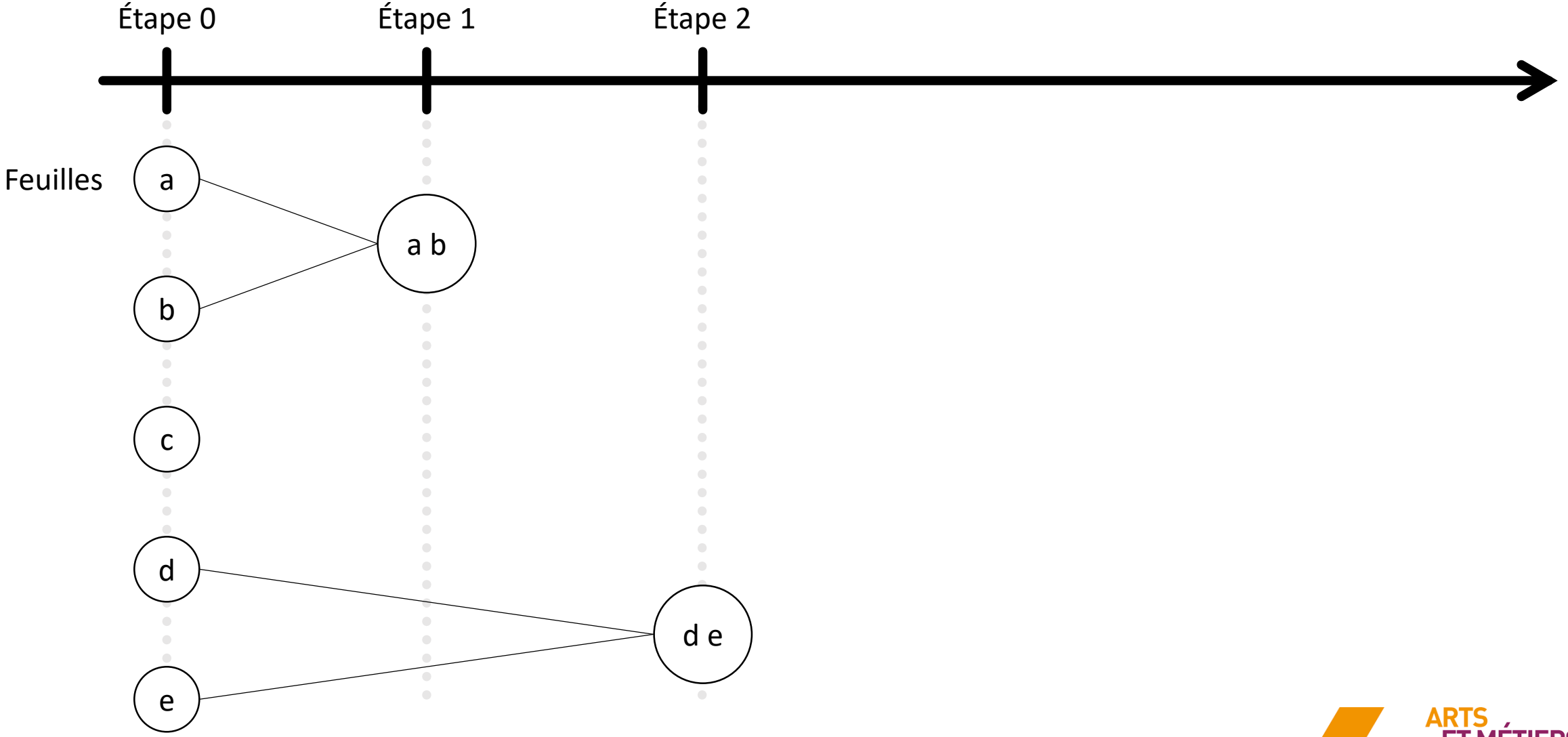
- Les classes dont la dissimilarité est MINIMALE sont regroupées

Clustering hiérarchique

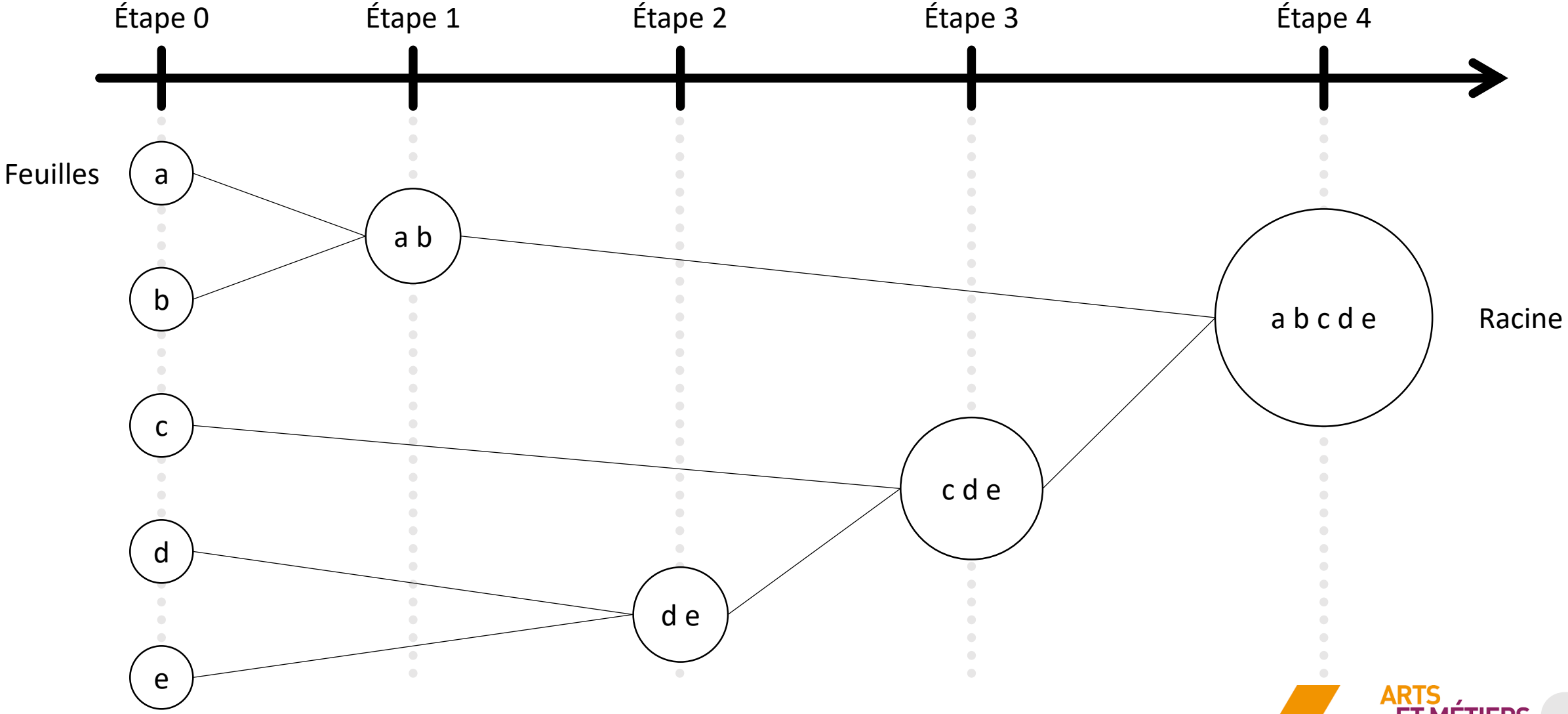
Clustering hiérarchique



Clustering hiérarchique



Clustering hiérarchique



Clustering hiérarchique

- En paramètre :
  - nombre de clusters ou distance seuil
  - MÉTRIQUE de regroupement
- SCALABLE sur un grand nombre de données
- Adapté pour jusqu'à un grand nombre de clusters
- Possibilité de rajouter des CONTRAINTES DE CONNECTIVITÉ
- Approche DESCENDANTE possible

Modèle de mélange gaussien

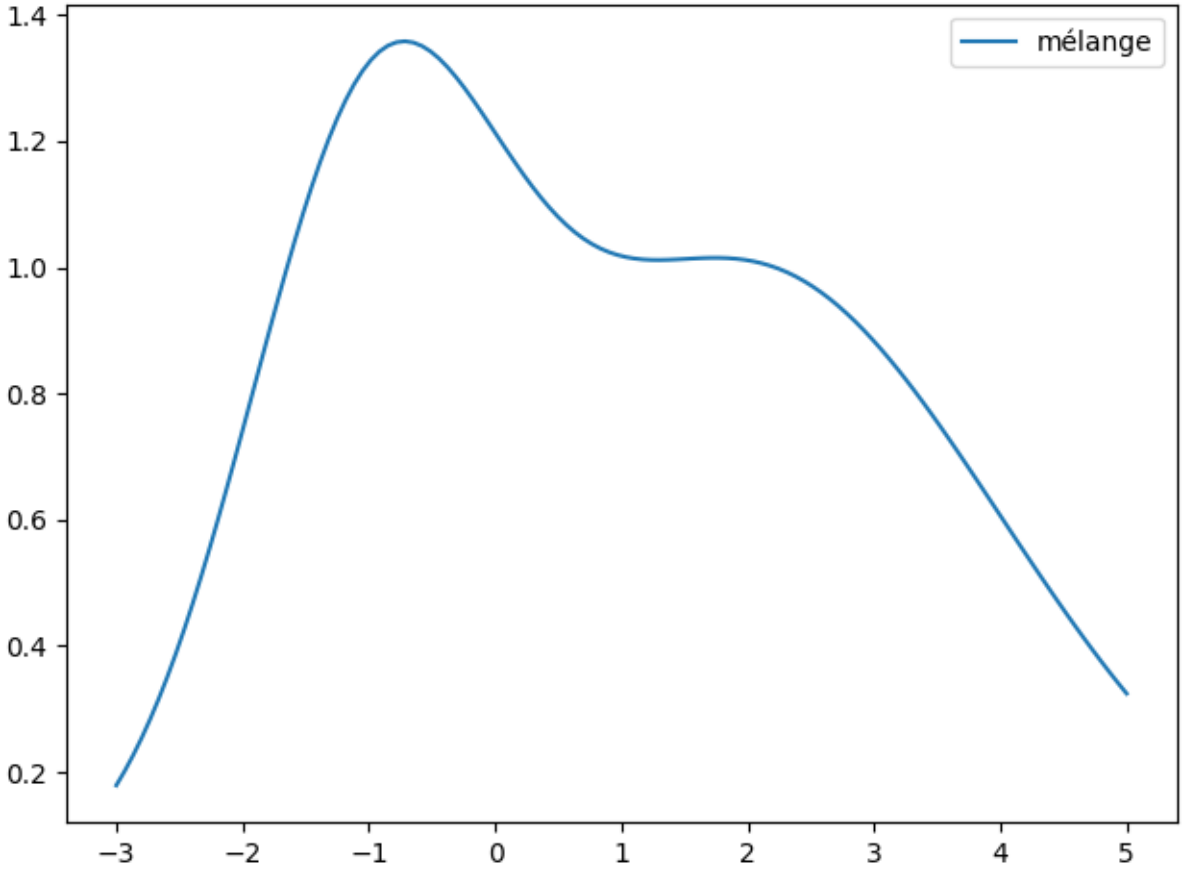
- DÉCOMPOSITION d'une représentation quelconque en plusieurs composantes GAUSSIENNES

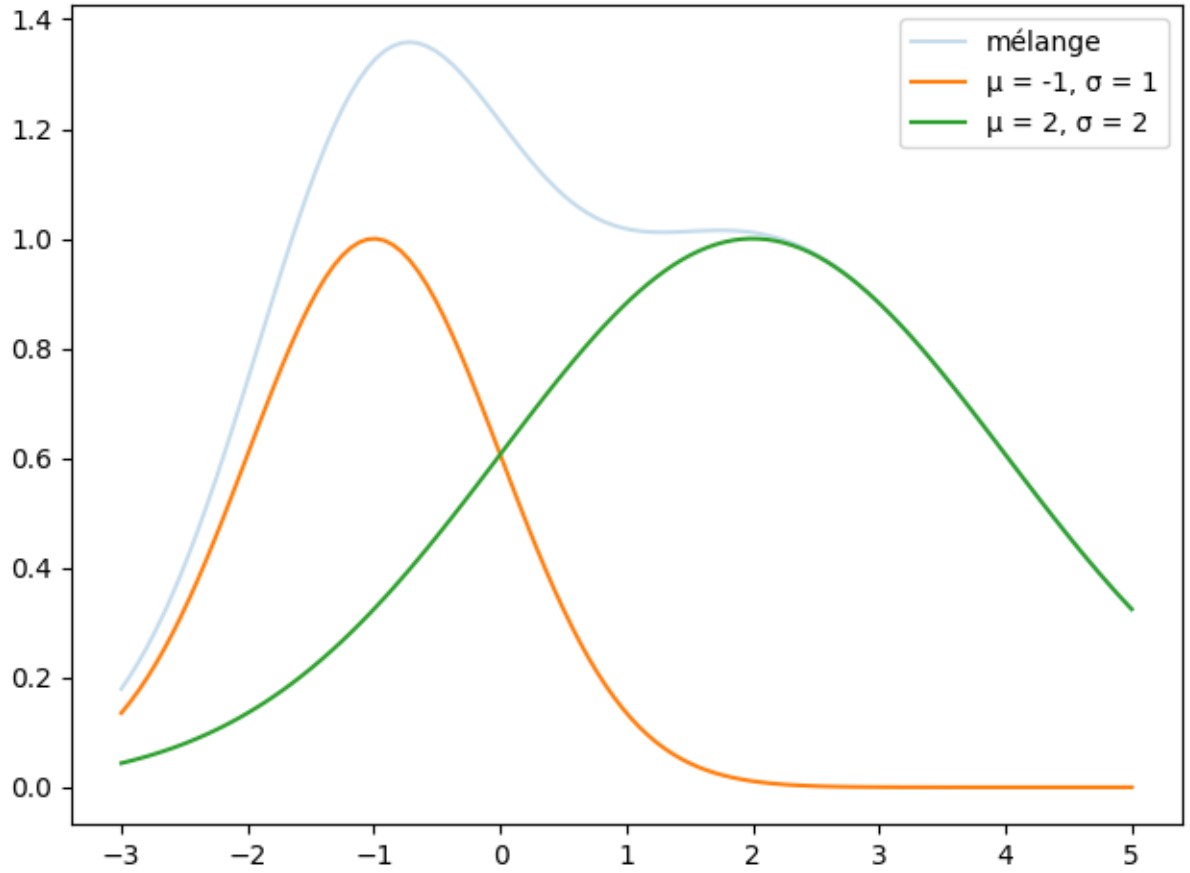
- Algorithme ESPÉRANCE-MAXIMISATION
  - Soit un échantillon $x = (x_1, \dots, x_n)$ suivant une loi $f(x_i, p)$
  - On cherche p maximisant la log-vraisemblance

$$L(x, p) = \sum_{i=1}^{n} \log f(x_i, p)$$

  - Résulte en un ensemble de paires (μ, σ)

Modèle de mélange gaussien





Modèle de mélange gaussien

- Adapté à des problèmes d'ESTIMATION de densité
- Très PEU SCALABLE à un grand nombre de données
- Adapté pour un PETIT NOMBRE de clusters