APPRENTISSAGE NON SUPERVISÉ

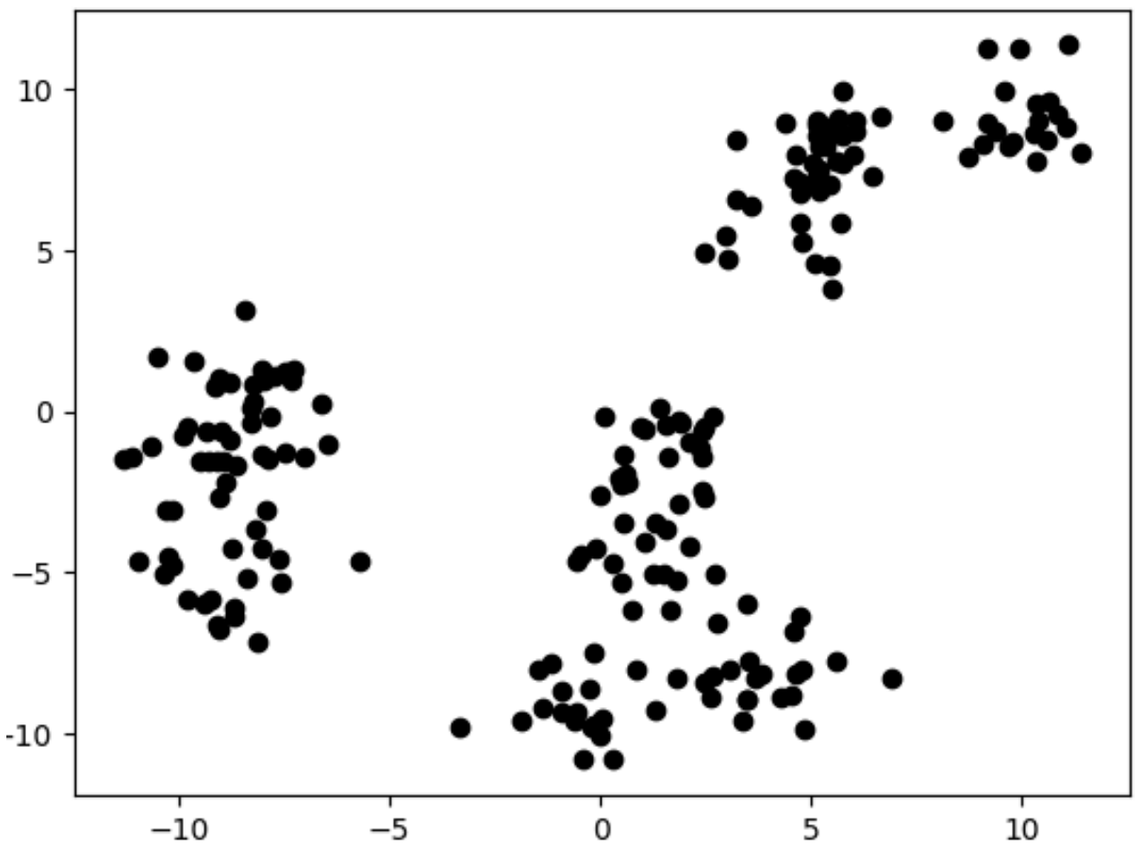
Cours 4

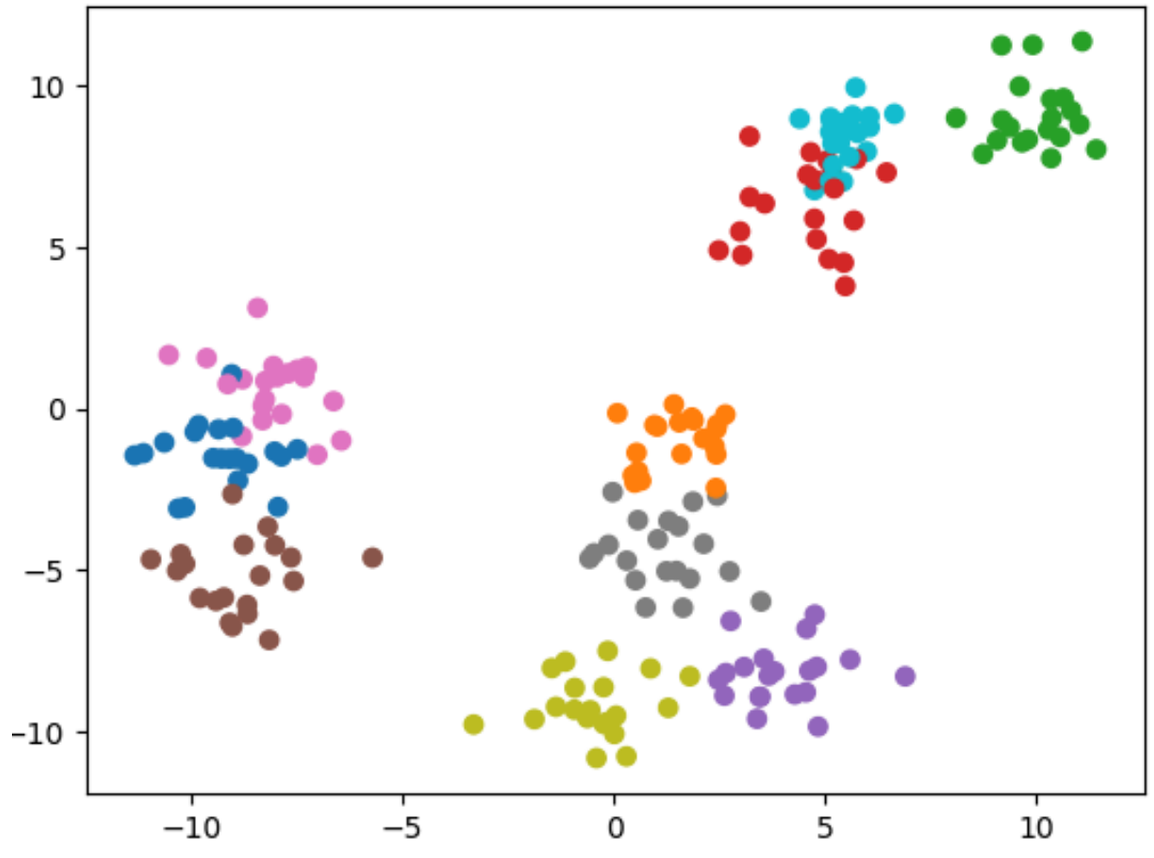
Caio Corro

APPRENTISSAGE NON SUPERVISÉ

Clustering ou partitionnement des données

- Recherche de la structure sous-jacente des données
- Regrouper les points similaires dans les mêmes clusters / groupes / classes





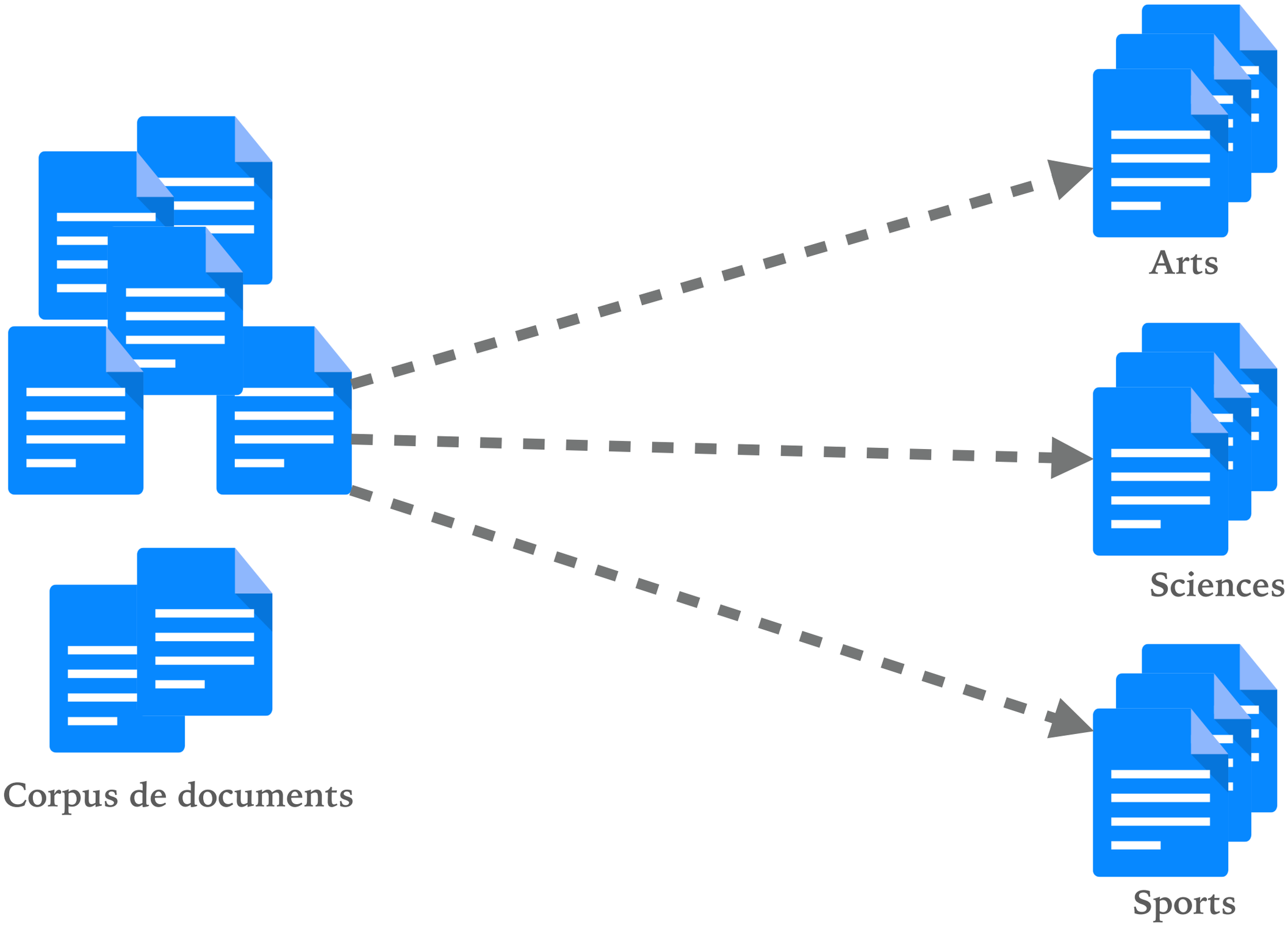
EXEMPLE : CLUSTERING D'ANIMAUX

Oiseaux

Mammifères

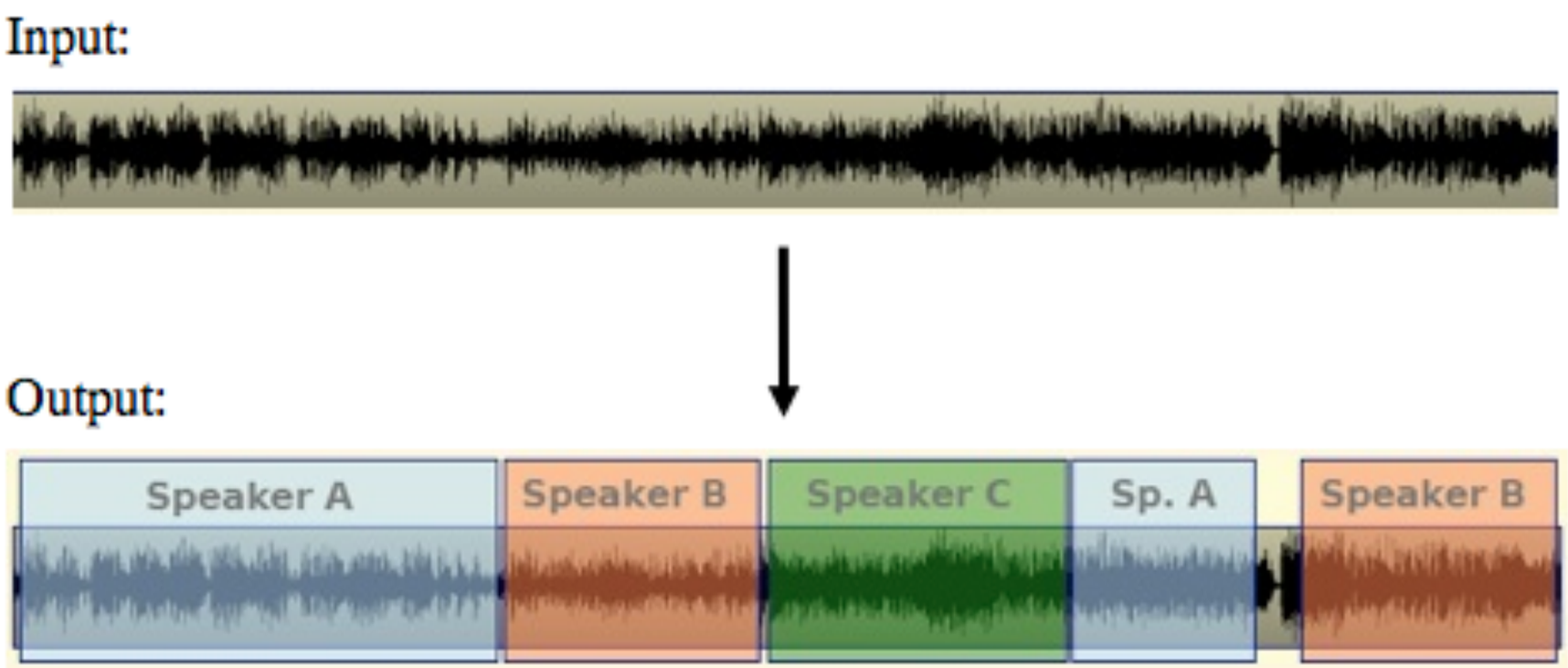
Reptiles

EXEMPLE : CLUSTERING DE DOCUMENTS

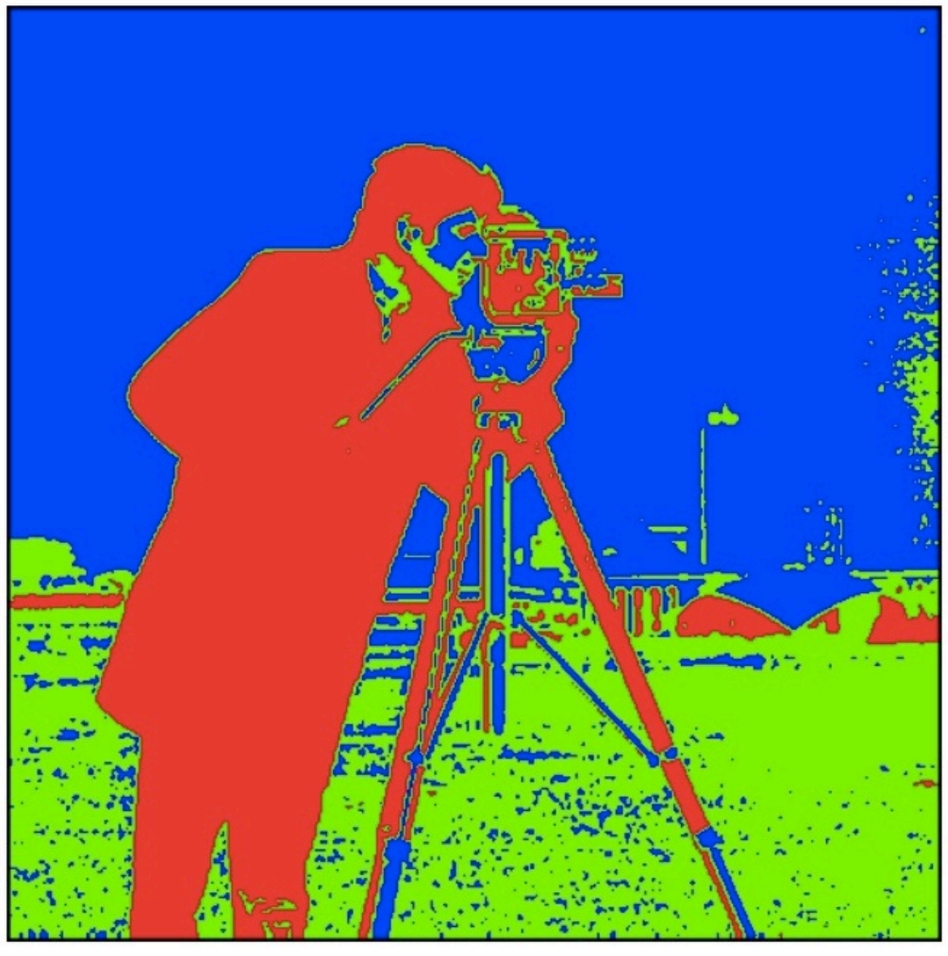


EXEMPLE : SPEAKER DIARISATION

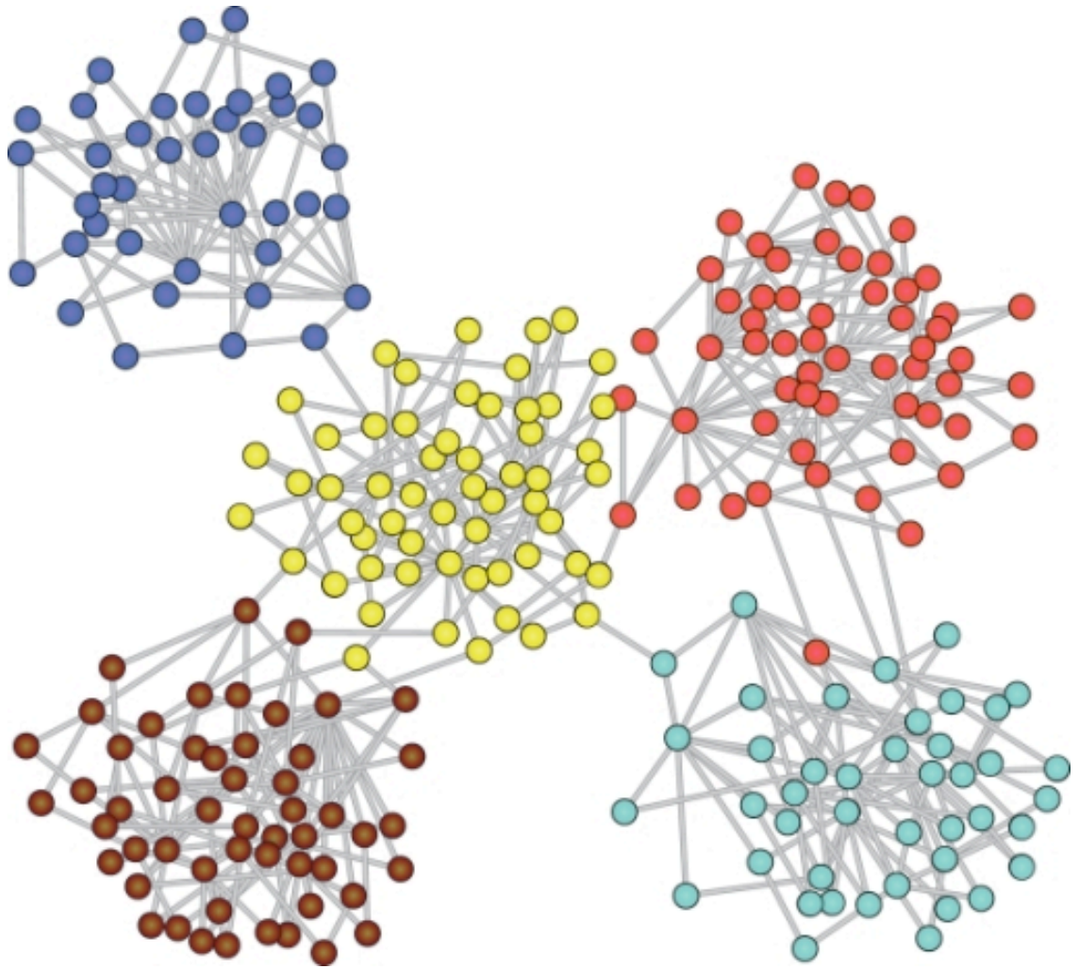
Découper un signal sonore en locuteurs



EXEMPLE : SEGMENTATION D'IMAGES



EXEMPLE : DÉTECTION DE COMMUNAUTÉS

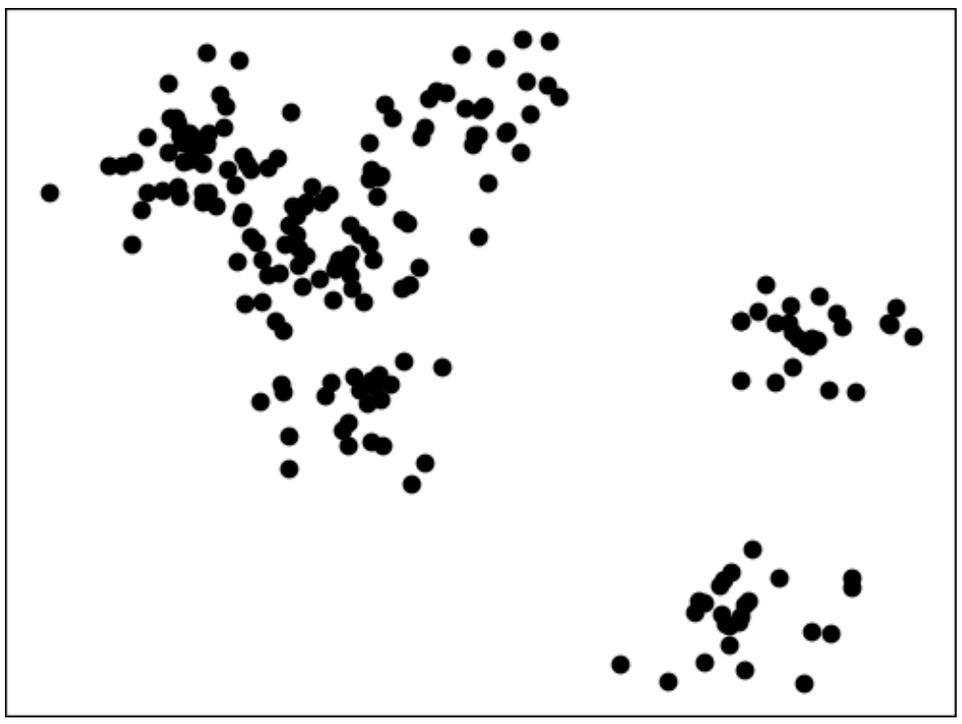


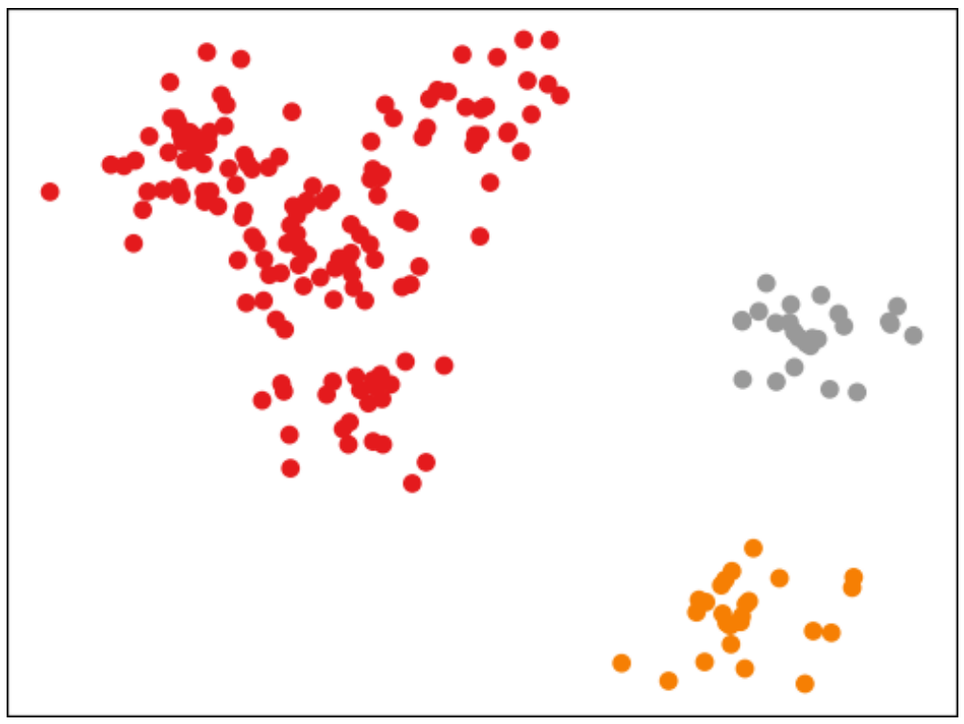
Détection de communautés dans les réseaux sociaux

CLUSTERING

Problème de clustering

- On a des points de données dans un espace
- On veut les regrouper en groupes (clusters) homogènes tel que :
  - les points "similaires" doivent être dans le même cluster
  - les clusters doivent être différents les uns des autres





K-MEANS

Hier : clustering hiérarchique ascendant

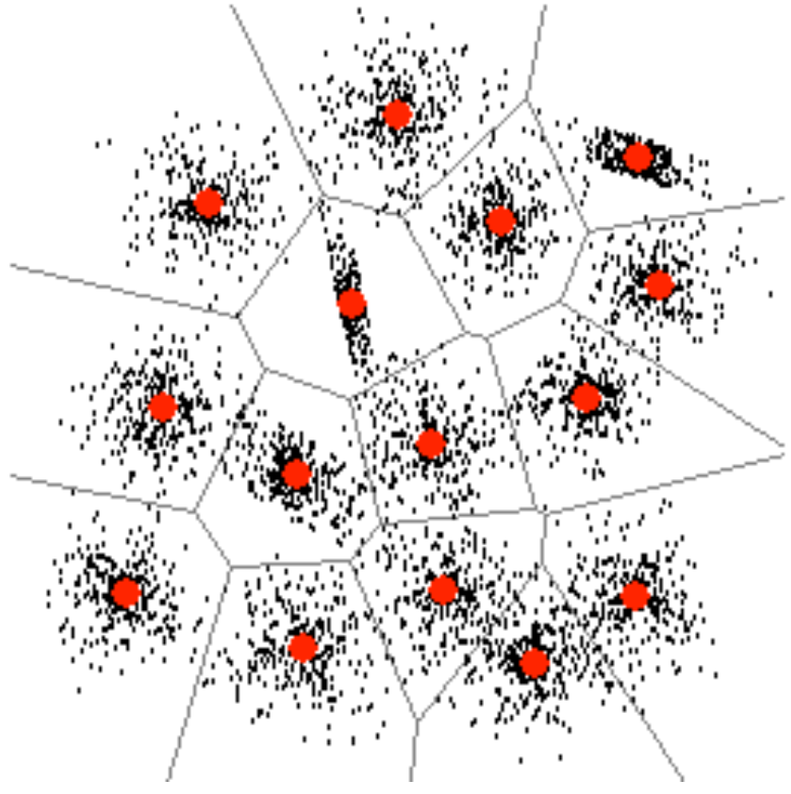
- On regroupe les points de façon hierarchique
- Cela produit une séquence de partitions des données

Aujourd'hui : k-means

- On préfixe le nombre de cluster à trouver
  (le k dans k-means fait référence au nombre de clusters)
- On va chercher itérativement la partition des données
  en k clusters qui minimise un critère de qualité

Sur la figure

- Points noirs : données
- Points rouges : représentants des clusters
- Chaque point est affectés au cluster dont le représentant
  est le plus proche



K-MEANS : CRITÈRE

Notations

- Ensemble des données : $X = \{x^{(1)}, x^{(2)}, \ldots, x^{(n)}\}$
- Partition des données : $\pi = \{C^{(1)}, \ldots C^{(k)}\}$

Représentants

On dénote $m^{(i)}$ le centroïde du cluster $C^{(i)}$: $m^{(i)} = \frac{1}{|C^{(i)}|} \sum_{x \in C^{(i)}} x$

Dispersion intra-clusters

Dispersion des éléments qui composent chaque cluster autour de son centroïde :

$$\sum_{i=1}^{k} \sum_{x \in C^{(i)}} \|x - m^{(i)}\|_2^2$$

Comment trouver la partition des données qui minimise la dispersion intra-clusters ?!

K-MEANS : CRITÈRE

Notations

- Ensemble des données : $X = \{x^{(1)}, x^{(2)}, \ldots, x^{(n)}\}$
- Partition des données : $\pi = \{C^{(1)}, \ldots C^{(k)}\}$
- Représentant de chaque cluster dans la partition : $\bar{c}^{(1)}, \ldots, \bar{c}^{(k)}$
- Ensemble des partitions de k parties de X : $\mathscr{P}_k(X)$

Attention à la différence entre C majuscule et c minuscule

Minimisons la dispersion intra-clusters

C'est désormais une variable !

Minimisation sur la partition

$$\min_{\substack{\pi \in \mathscr{P}_k(X), \\ \bar{c}^{(1)}, \ldots, \bar{c}^{(k)}}} \sum_{i=1}^{k} \sum_{x \in C^{(i)}} \|x - \bar{c}^{(i)}\|_2^2$$

Minimisation sur les représentants des clusters

Détail important

Ici, $m^{(i)}$ n'est pas forcément le centroïde de $C^{(i)}$, c'est son représentant.

Cependant, il s'avère que le meilleur représentant est le centroïde... (exercice à la fin)

K-MEANS

Minimisons la dispersion intra-clusters

C'est désormais une variable !

$$\min_{\substack{\pi \in \mathscr{P}_k(X), \\ \bar{c}^{(1)}, \ldots, \bar{c}^{(k)}}} \sum_{i=1}^{k} \sum_{x \in C^{(i)}} \|x - \bar{c}^{(i)}\|_2^2$$

Résolution itérative du problème de minimisation

On va résoudre de façon itérative le problème :

1. Minimiser sur π en gardant les représentants $c^{(i)}$ fixes
2. Minimiser sur $\bar{c}^{(i)}$ en gardant la partition π fixe

On fait ces deux étapes plusieurs fois (ou jusqu'à ce que plus rien ne change)

K-MEANS

Minimisons la dispersion intra-clusters

C'est désormais une variable !

$$\min_{\substack{\pi \in \mathscr{P}_k(X), \\ \bar{c}^{(1)}, \ldots, \bar{c}^{(k)}}} \sum_{i=1}^{k} \sum_{x \in C^{(i)}} \|x - \bar{c}^{(i)}\|_2^2$$

Minimisation sur la partition

Minimisation sur la partition

$$\min_{\pi \in \mathscr{P}_k(X)} \sum_{i=1}^{k} \sum_{x \in C^{(i)}} \|x - \bar{c}^{(i)}\|_2^2$$

Solution : on assigne chaque point au cluster dont le représentant est le plus proche

Pourquoi ? chaque point contribue exactement une fois à l'objectif dans le terme de distance, et donc assigner chaque point de tel sorte à ce que cette distance soit minimisée donne la partition qui minimise la distance intra-clusters.

K-MEANS

Minimisons la dispersion intra-clusters

$$\min_{\substack{\pi \in \mathscr{P}_k(X), \\ c^{(1)}, \ldots, c^{(k)}}} \sum_{i=1}^{k} \sum_{x \in C^{(i)}} \|x - c^{(i)}\|_2^2$$

Minimisation sur les représentants

$$\min_{\bar{c}^{(1)}, \ldots, \bar{c}^{(k)}} \sum_{i=1}^{k} \sum_{x \in C^{(i)}} \|x - \bar{c}^{(i)}\|_2^2$$

Minimisation sur les représentants des clusters

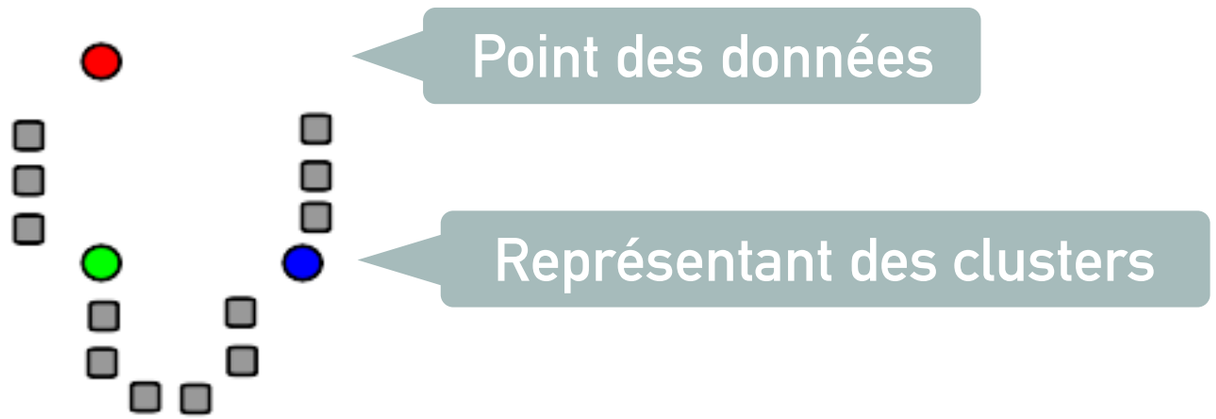
La solution de ce problème est triviale, on définie simplement : $\bar{c}^{(i)} = \frac{1}{|C^{(i)}|} \sum_{x \in C^{(i)}} x$

(pourquoi ? exercice à la fin !)

K-MEANS : EXEMPLE

(wikipedia)

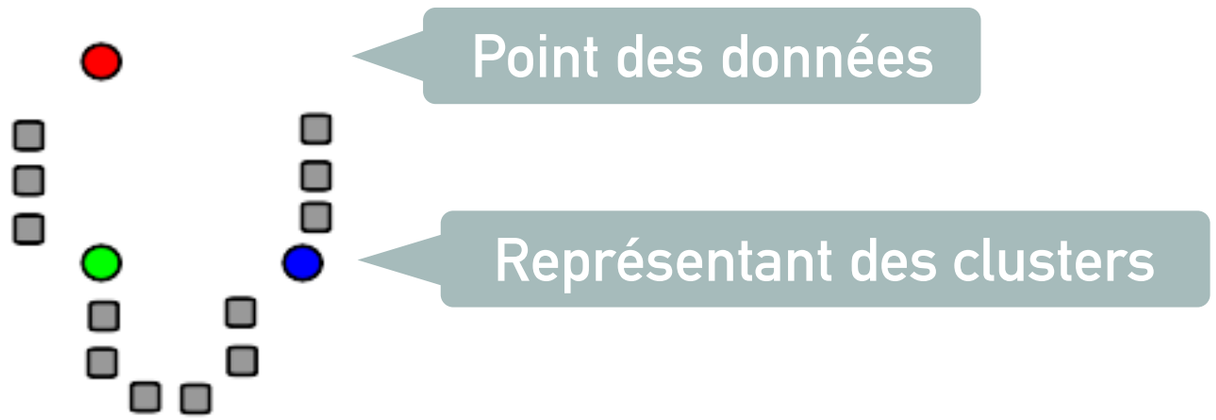
1 Initialisation : on choisi des représentants aléatoires



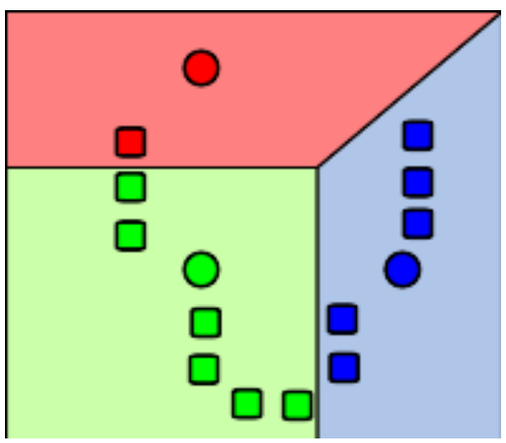
K-MEANS : EXEMPLE

(wikipedia)

1 Initialisation : on choisi des représentants aléatoires



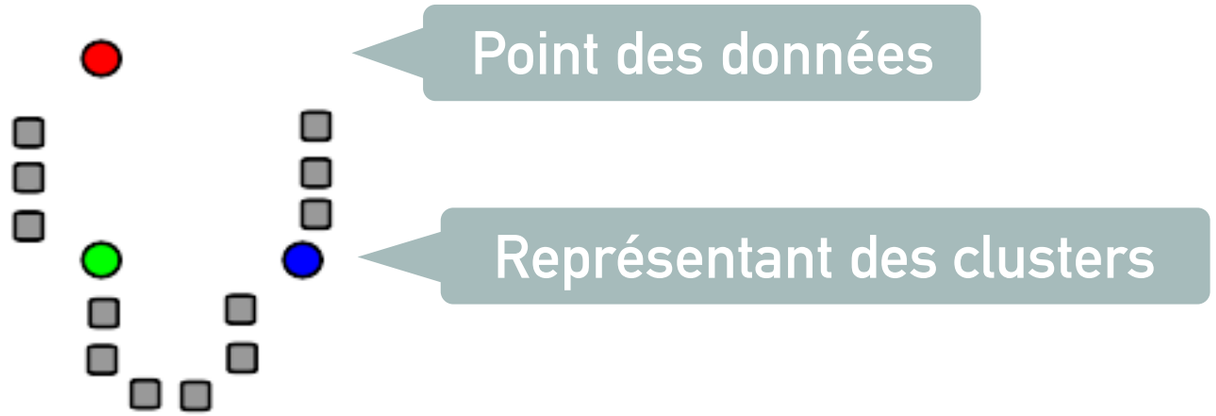
2 On assigne chaque point à son cluster le plus proche



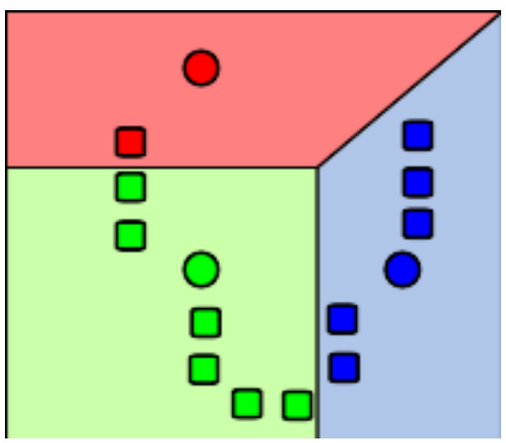
K-MEANS : EXEMPLE

(wikipedia)

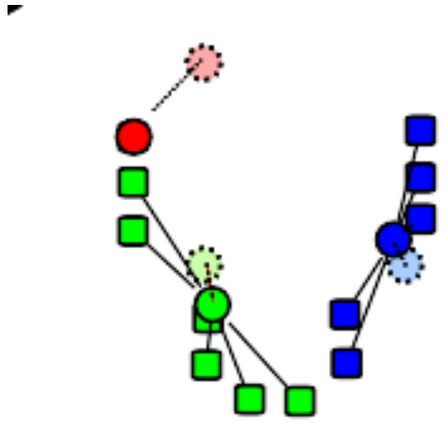
1 Initialisation : on choisi des représentants aléatoires



2 On assigne chaque point à son cluster le plus proche



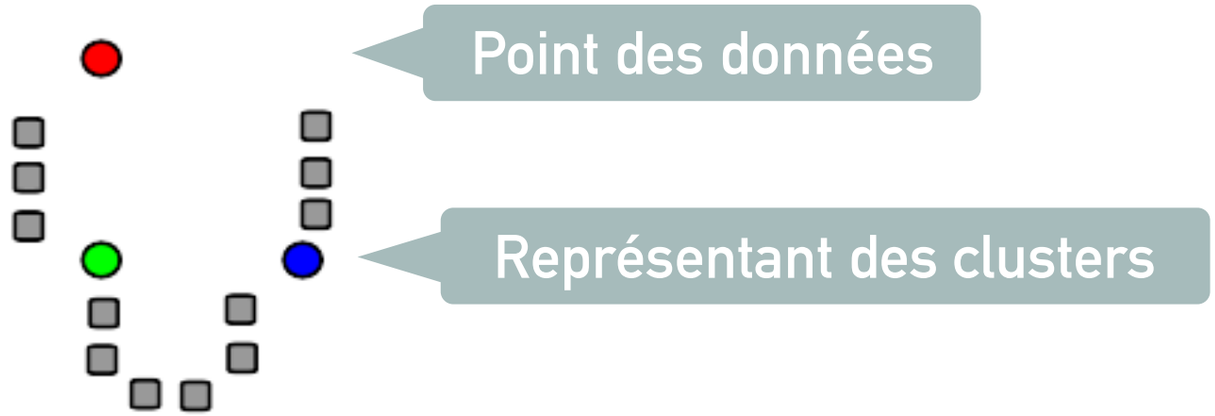
3 On recalcule les représentants des clusters



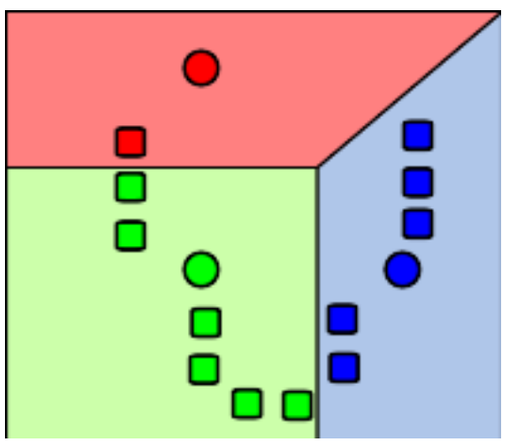
K-MEANS : EXEMPLE

(wikipedia)

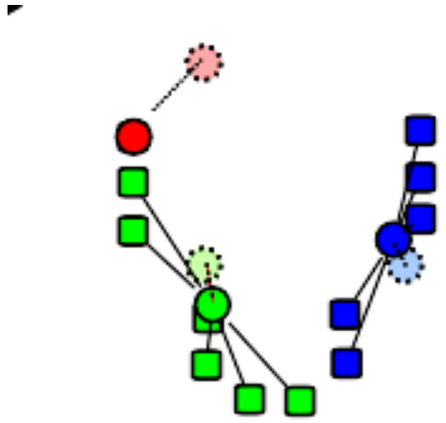
1 Initialisation : on choisi des représentants aléatoires



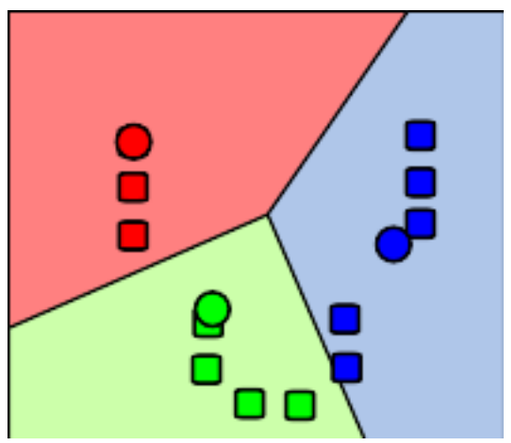
2 On assigne chaque point à son cluster le plus proche



3 On recalcule les représentants des clusters



4 On assigne chaque point à son cluster le plus proche



Etc etc

EXERCICE

Minimisation sur les représentants

$$\min_{\bar{c}^{(1)},\dots,\bar{c}^{(k)}} \sum_{i=1}^{k} \sum_{x \in C^{(i)}} \|x - \bar{c}^{(i)}\|_2^2$$

Minimisation sur les représentants des clusters

Exercice

Calculer la solution optimale à ce problème.