Fouille de données: Analyse de clusters

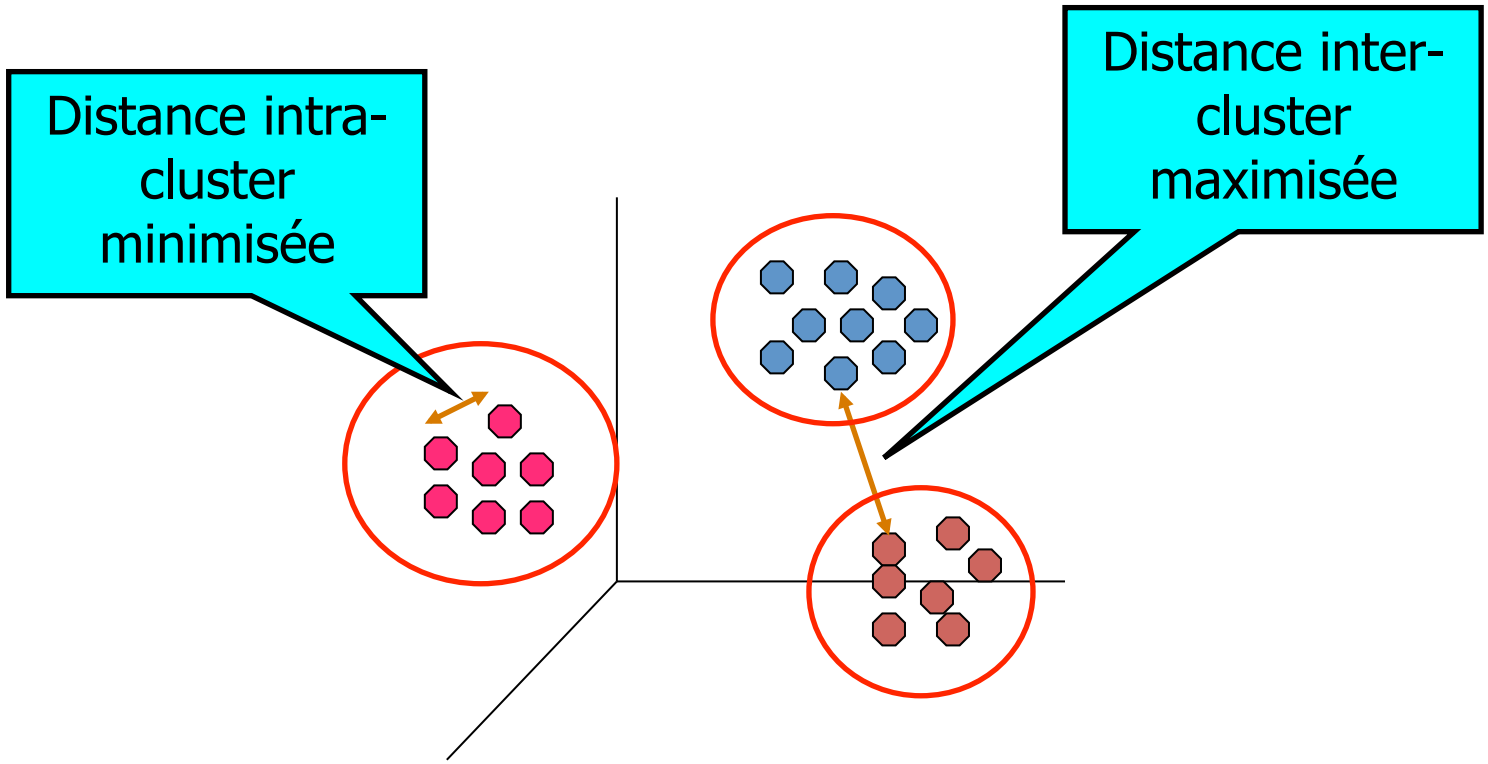
Translated from Tan, Steinbach, Kumar

Lecture Notes for Chapter 8

Introduction to Data Mining

Qu’est ce que l’analyse de clusters ?

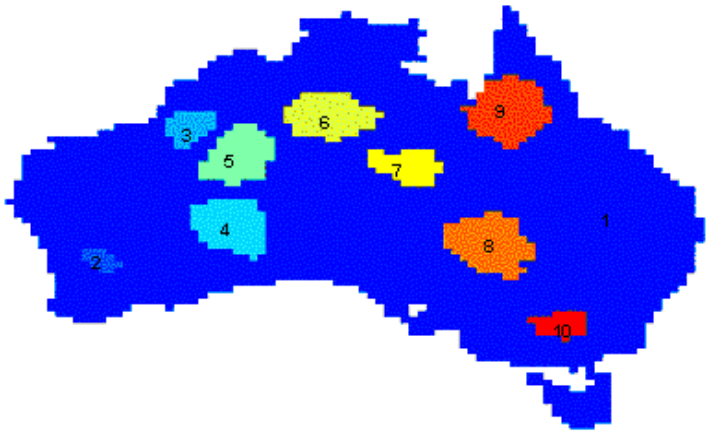
- Trouver des groupes d’objets tels que les objets d’un groupe sont similaires entre eux mais différents des objets des autres groupes



Applications de l'analyse de clusters

- Comprendre
  - Regrouper des documents similaires pour naviguer, grouper des gènes et protéines qui ont les mêmes fonctions, ou des actions dont le cours varie de façon similaire
- Résumer
  - Réduire la taille des jeux de données

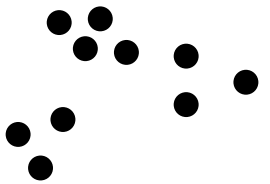
	Discovered Clusters	*Industry Group*
1	Applied-Matl-DOWN,Bay-Network-Down,3-COM-DOWN, Cabletron-Sys-DOWN,CISCO-DOWN,HP-DOWN, DSC-Comm-DOWN,INTEL-DOWN,LSI-Logic-DOWN, Micron-Tech-DOWN,Texas-Inst-Down,Tellabs-Inc-Down, Natl-Semiconduct-DOWN,Oracl-DOWN,SGI-DOWN, Sun-DOWN	Technology1-DOWN
2	Apple-Comp-DOWN,Autodesk-DOWN,DEC-DOWN, ADV-Micro-Device-DOWN,Andrew-Corp-DOWN, Computer-Assoc-DOWN,Circuit-City-DOWN, Compaq-DOWN, EMC-Corp-DOWN, Gen-Inst-DOWN, Motorola-DOWN,Microsoft-DOWN,Scientific-Atl-DOWN	Technology2-DOWN
3	Fannie-Mae-DOWN,Fed-Home-Loan-DOWN, MBNA-Corp-DOWN,Morgan-Stanley-DOWN	Financial-DOWN
4	Baker-Hughes-UP,Dresser-Inds-UP,Halliburton-HLD-UP, Louisiana-Land-UP,Phillips-Petro-UP,Unocal-UP, Schlumberger-UP	Oil-UP

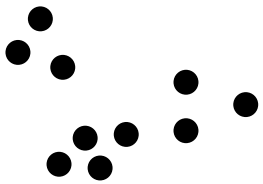


Ce que l'analyse de clusters n'est pas

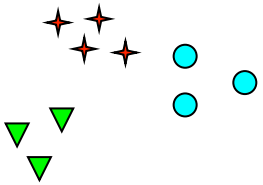
- Classification supervisée
  - Pas de label de classe
- Segmentation simple
  - Diviser les étudiants en groupes suivant l'ordre alphabétique
- Résultat d'une requête
  - Grouper suivant une spécification externe
- Partitionnement de graphe
  - Des similarités, mais les domaines sont différents

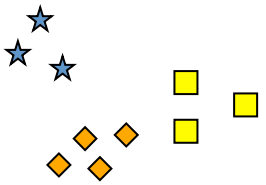
La notion de cluster peut être ambigue



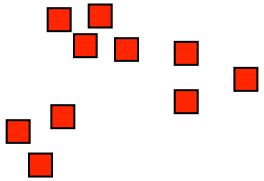


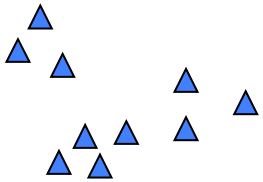
Combien de clusters ?



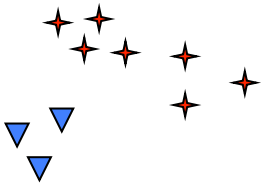


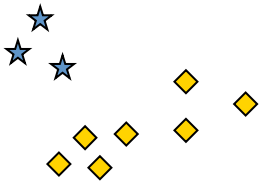
6 Clusters





2 Clusters



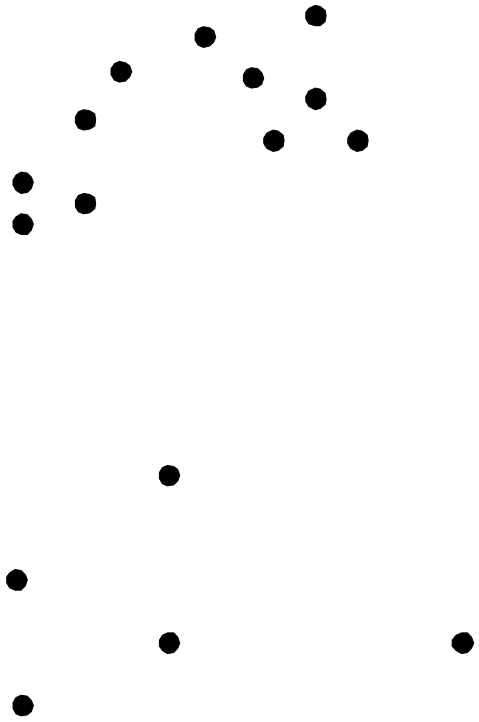


4 Clusters

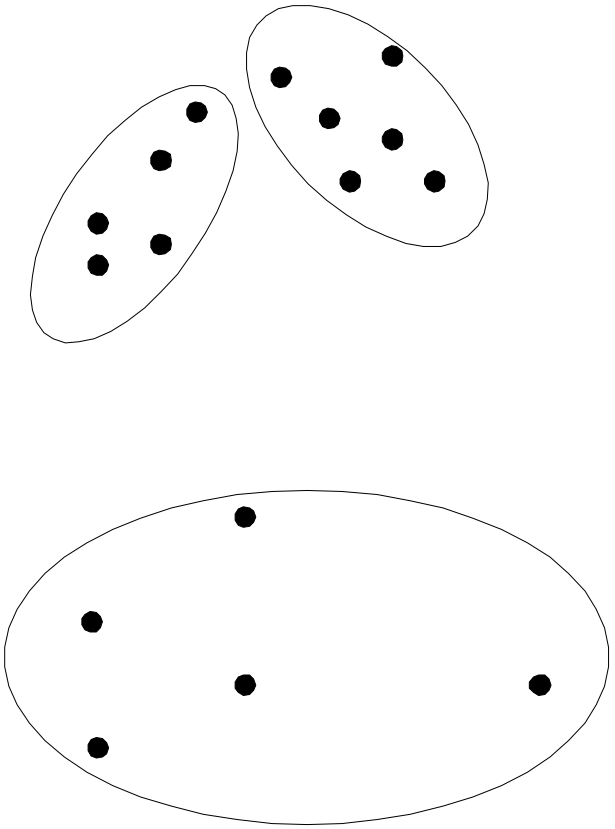
Types de clusering

- Un clustering est un ensemble de clusters
- Il y a une distinction entre le clustering hiérarchique et un partitionnement en clusters
- Clustering en partitionnement
  - Une division des données en sous-ensembles ne se chevauchant pas
- Clustering hiérarchique
  - Un ensemble de clusters imbriqués organisée en arbre hiérarchique

Clustering partitionnant

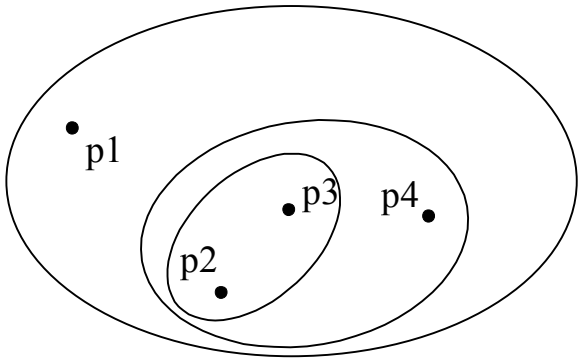


Points de données

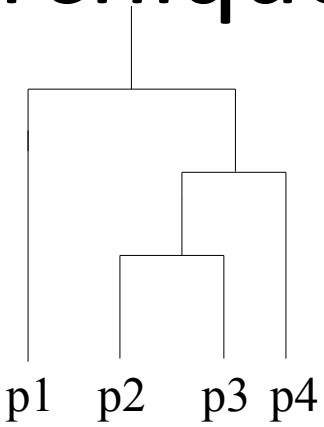


Un clustering partitionnant

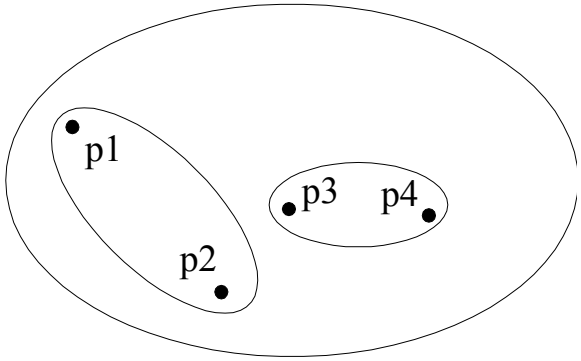
Clustering hiérarchique



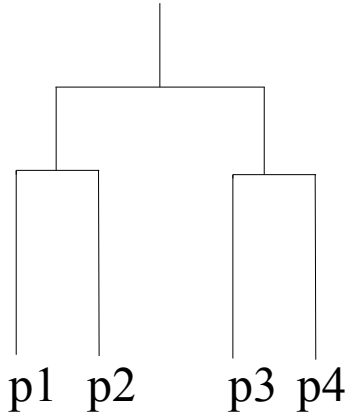
Clustering hiérarchique traditionnel



Dendrogram traditionnel



Clustering hiérarchique non-traditionnel



Autres distinctions entre des ensembles de clusters

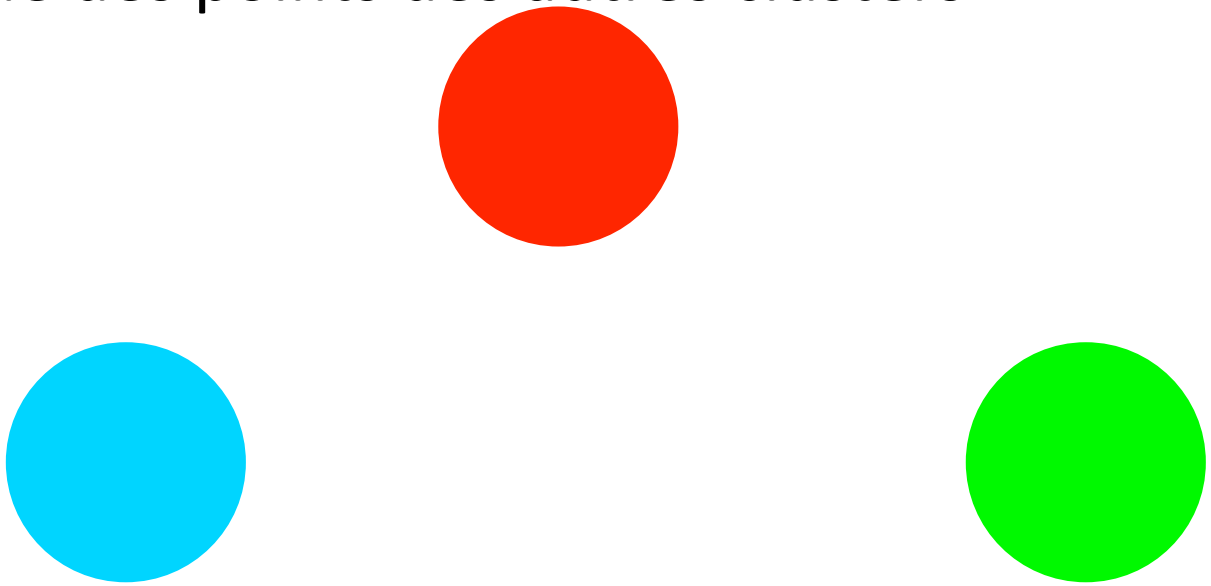
- Exclusif / non-exclusif
  - Dans un clustering non exclusif, un point peut appartenir à plusieurs clusters
  - Peut représenter des classes multiples, ou des points à la frontière
- Clustering flou (fuzzy)
  - Dans le clustering flou, un point appartient à chaque cluster avec un poids compris entre 0 et 1
  - La somme des poids d'un point est 1
  - Notion de probabilité
- Partiel / complet
  - Dans certain cas, nous ne voulons clusteriser qu'une partie des données
- Hétérogène / homogène
  - Les clusters peuvent varier en aille, forme, et densité

Types de clusters

- Clusters bien séparés
- Clusters basés sur un centre
- Clusters contigus
- Clusters basés sur la densité
- …

Types de clusters : Clusters bien séparés

- Clusters bien séparés:
  - Un cluster est un ensemble de points tel que chaque point d'un cluster est plus proche (ou plus similaire) de tous les autres points de son cluster que des points des autres clusters



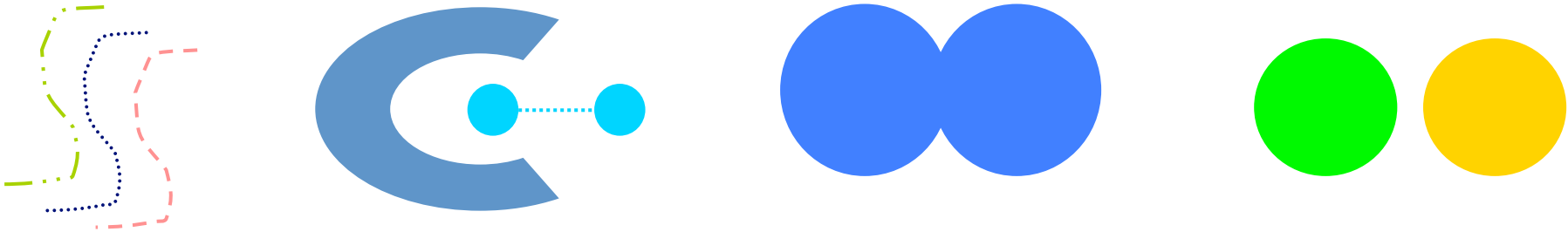
3 clusters bien séparés

Types de clusters: basés sur un centre

- Basés sur un centroïd
  - Un cluster est un ensemble d'objets tels qu'un objet d'un cluster est plus proche (ou similaire) du centre du cluster que du centre des autres clusters
  - Le centre du cluster est souvent un centroïde, la moyenne de tous les moints du cluster, ou un medoïd, le point le plus « représentatif » du cluster

Types de clusters: contigus

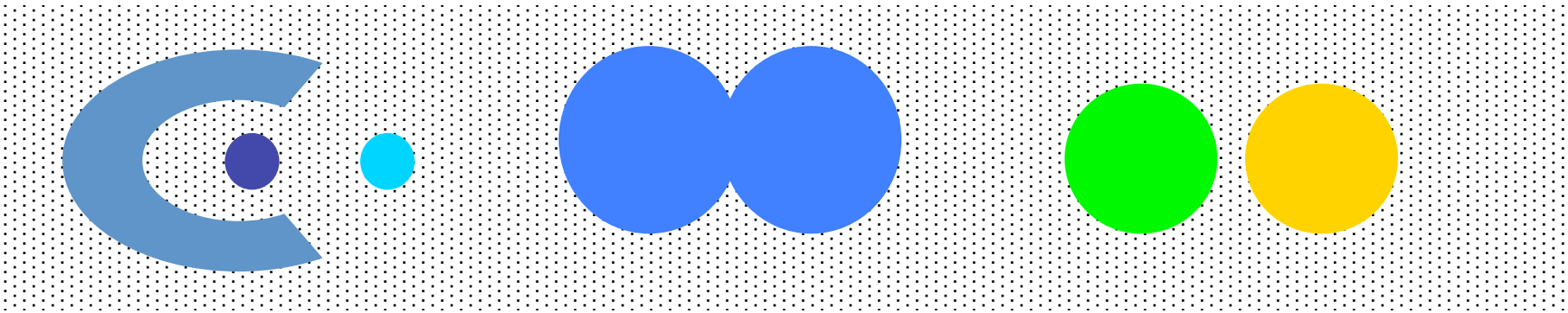
- Clusters contigus (transitifs par plus proche voisin)
  - Un cluster est un ensemble de points tels qu'un point d'un cluster est plus proche (ou similaire) d'un ou plus autres points du cluster que d'un point d'un autre cluster



8 clusters contigus

Types de clusters: basés sur la densité

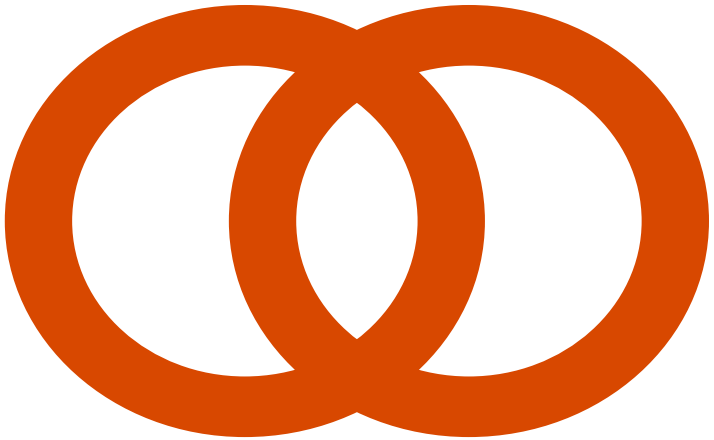
- Basés sur la densité
  - Un cluster est une région de points dense, qui est séparée des autres clusters par des régions non denses
  - Utilisé quand les clusters sont irréguliers ou mélangés, ou en présence de bruit



6 clusters basés sur la densité

Types de clusters: clusers conceptuels

- Propriété partagée, ou clusters conceptuels
  - Trouve des clusters qui partagent certaines propriétés ou représentent un concept particulier



2 cercles se recouvrant

Types de clusters: fonction objective

- Clusters définis par une fonction objective
  - Trouve des clusters qui minimisent ou maximisent une fonction objective
  - Énumère toutes les façons possibles de diviser les points en clusters et évalue la qualité de chaque ensemble de clusters potentiel en utilisant la fonction objective donnée (NP hard)
  - Peut avoir des objectifs globaux ou locaux
    - Les algorithmes de clustering hiérarchique ont en général des objectifs locaux
    - Les algorithmes de partitionnement ont en général des objectifs globaux

Types de clusters: fonction objective …

- Transférer le problème du clustering dans un domaine différent et résoudre ce problème dans ce nouveau domaine
  - La matrice de proximité définit un graphe pondéré, dans lequel les sommets sont les points, et les arcs représentent la proximité entre les points
  - Le clustering est équivalent à diviser le graphe en composante connectées, une par cluster
  - On veut minimiser le poids des arcs entre les clusters et maximiser le poids des arcs dans un cluster

Les caractéristiques des données d'entrée sont importantes

- Type de mesure de proximité ou densité
  - Impacte important sur le clustering
- Type des attributs
  - Joue sur la fonction de similarité
- Type de données
- Nombre de dimensions
- Bruit / valeurs aberrantes
- Distribution des données

Algorithmes de clustering

- K-means et ses variantes
- Clustering hiérarchique
- Clustering basé sur la densité

Algorithme de clustering K-means

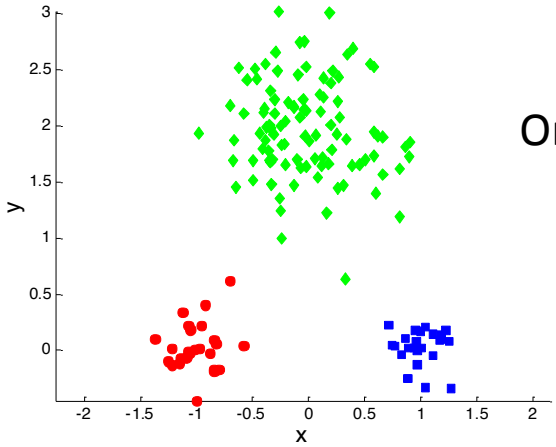
- Algorithme de clustering par partitionnement
- Chaque cluster a un centroid associé (centre)
- Chaque point est assigné au cluster dont le centroid est le plus proche
- Le nombre de clusters, K, doit être spécifié
- L'algorithme est très simple

1: Select K points as the initial centroids.
2: **repeat**
3: Form K clusters by assigning all points to the closest centroid.
4: Recompute the centroid of each cluster.
5: **until** The centroids don't change

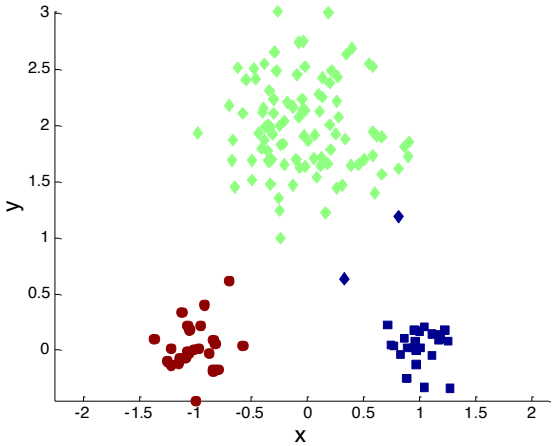
Algorithme K-means, détails

- Les centroïdes initiales sont choisies au hasard
  - Les clusters générés peuvent varier d'une exécution à l'autre
- Les centroïdes sont en général la moyenne des points du cluster
- La proximité est mesurée par distance Euclidienne, similarité cosinus, corrélation …
- K-means converge pour les mesures citées
- La majorité de la convergence a lieu dans les premières itérations
  - Souvent, la condition d'arrêt devient « jusqu'à ce que peu de changements aient lieu »
- La complexité est O(n*K*I*d)
  - n=nombre de points, K = nombre de clusters, I=nombre d'itérations, d = nombre d'attributs

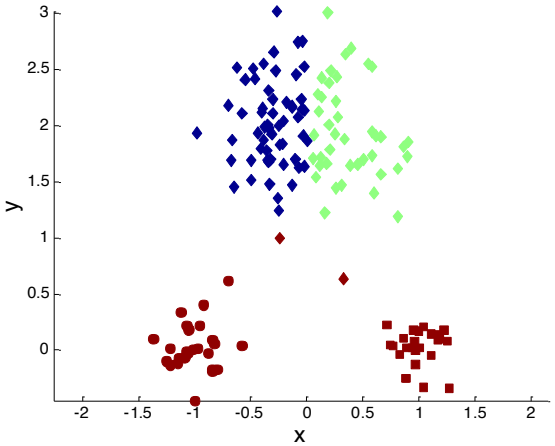
2 clusterings différents produits par K-means



Original Points

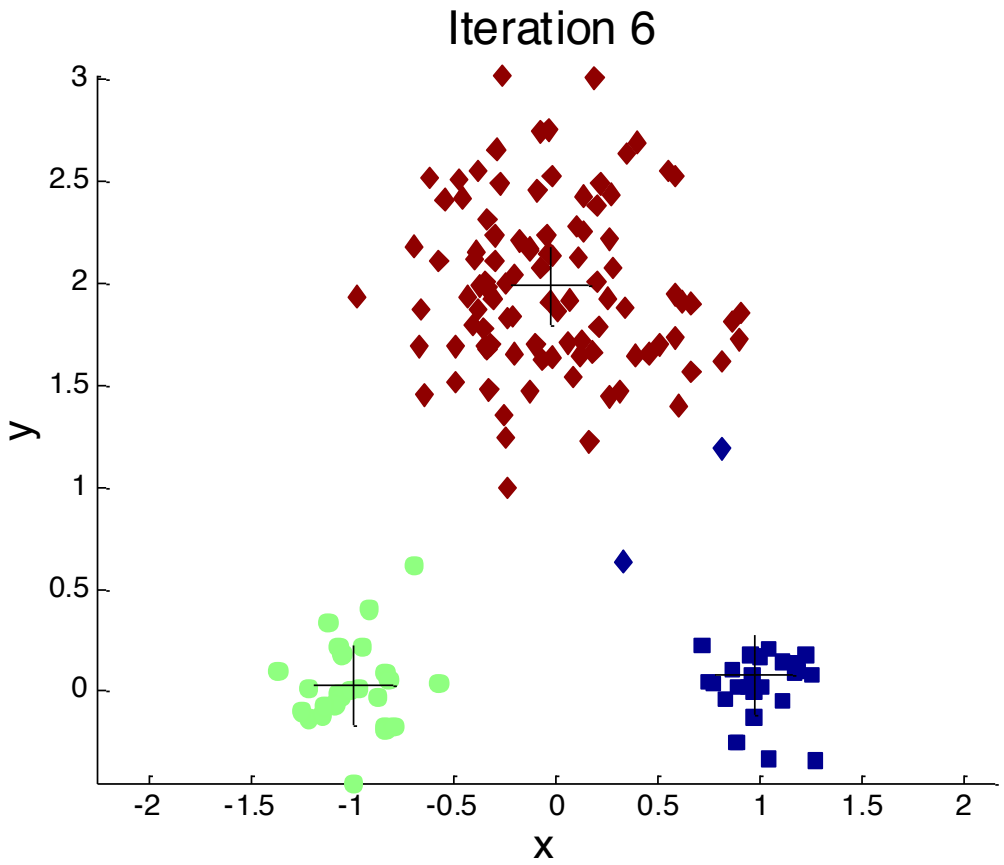


Optimal Clustering

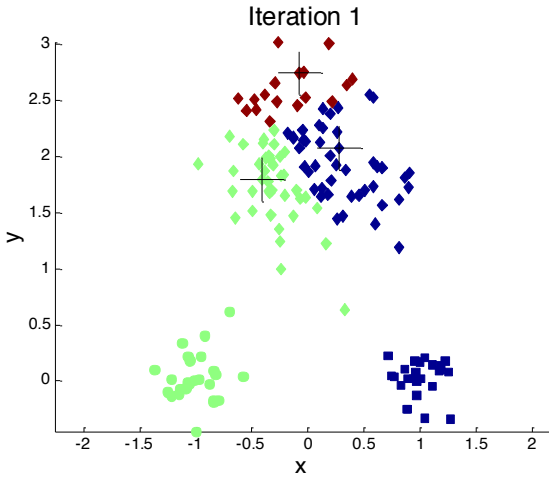


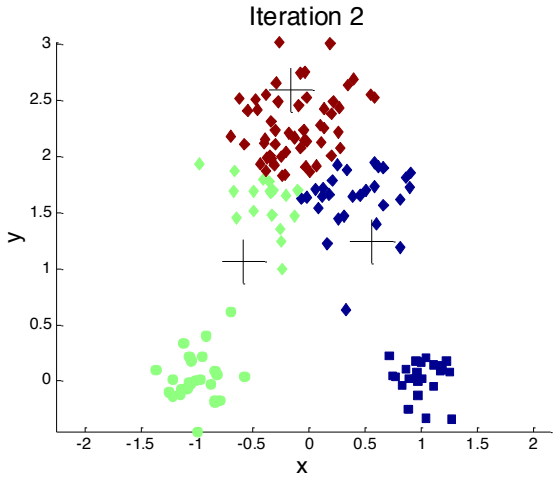
Sub-optimal Clustering

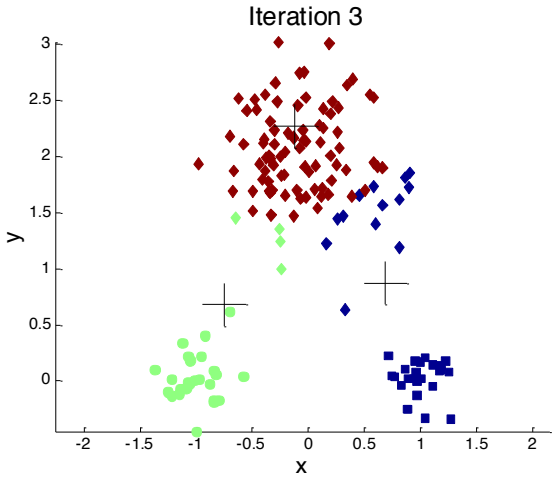
Importance du choix des centroïdes initiales

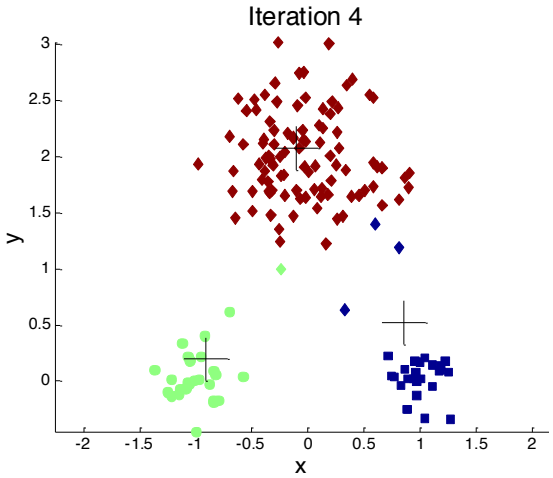


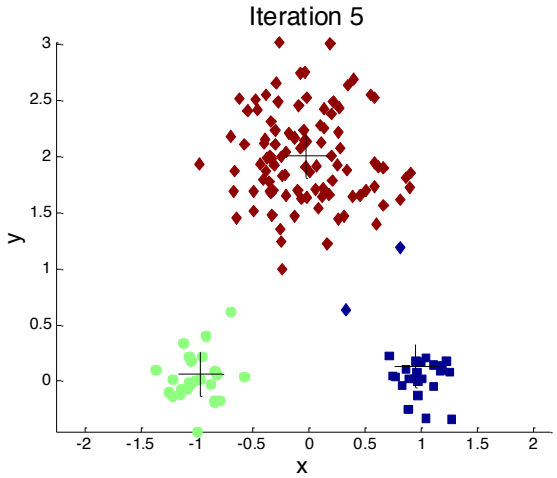
Importance du choix des centroïdes initiales

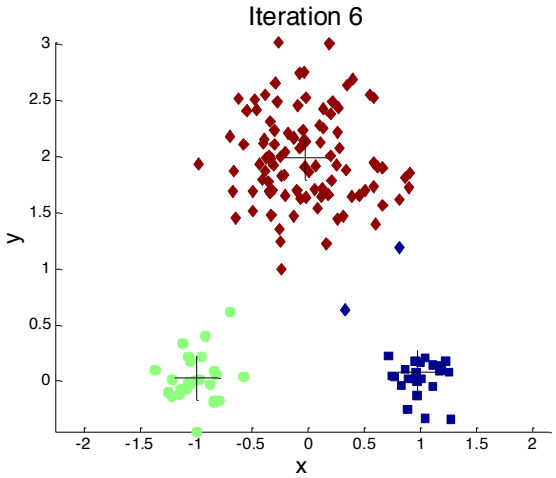












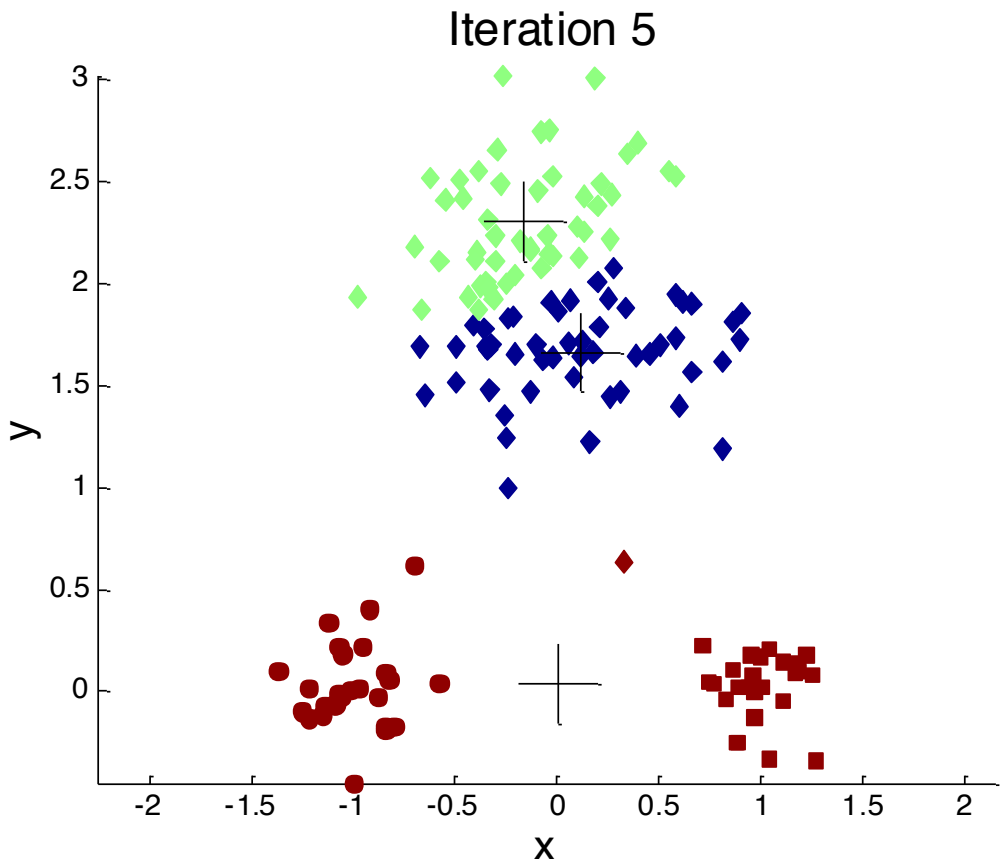
Évaluation des clusters conçus par K-means

- Mesure la plus commune = Sum of Squared Error (SSE)
  - Pour chaque point, l'erreur est la distance au cluster le plus proche
  - Pour obtenir la SSE, on fait la somme de ces erreurs au carré

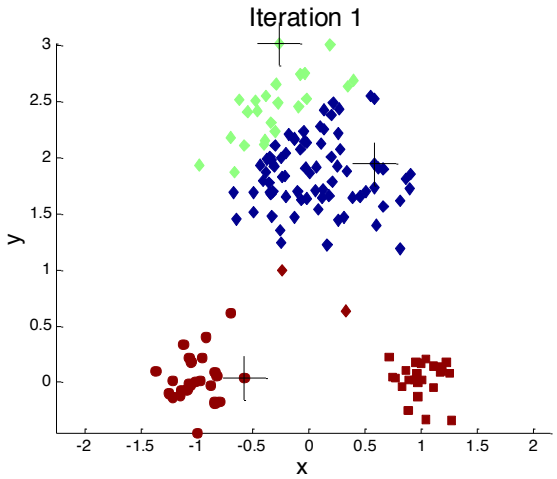
$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} dist^2(m_i, x)$$

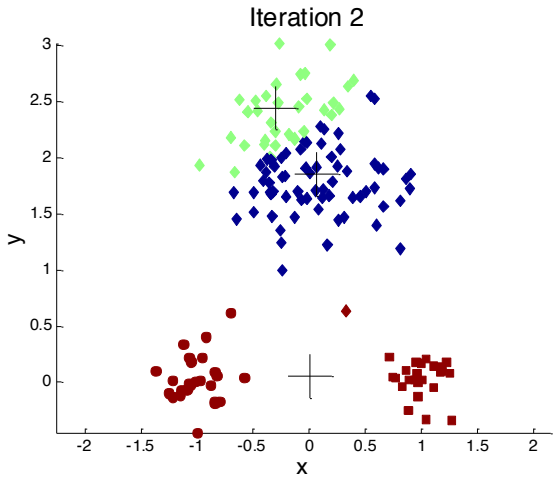
  - *x* est un point dans le cluster C_i et m_i est le représentatif du cluster C_i
    - On peut montrer que le *x* qui minimise l'erreur est le centre du cluster
  - Pour 2 clustering, on choisir celui qui produit la plus petite erreur
  - One façon simple de diminuer la SSE est d'augmenter K, le nombre de clusters
    - Un bon clustering avec un petit K peut avoir une SSE plus faible qu'un mauvais clustering avec un K élevé

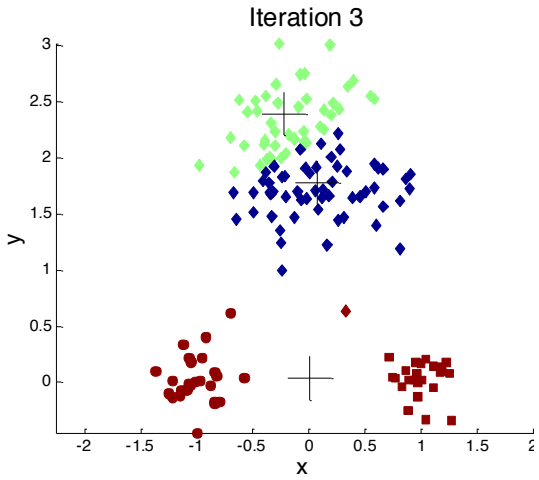
Importance du choix initial des centroïdes

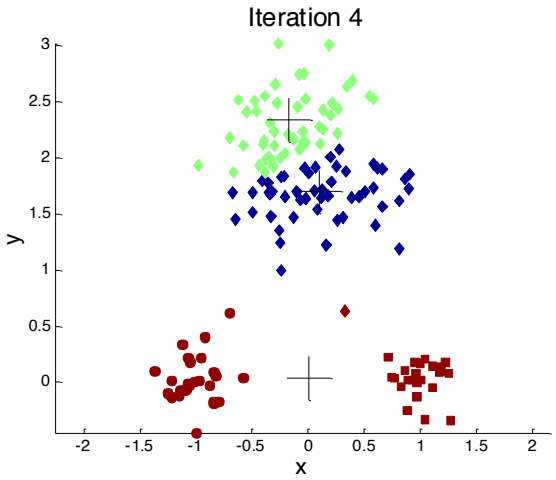


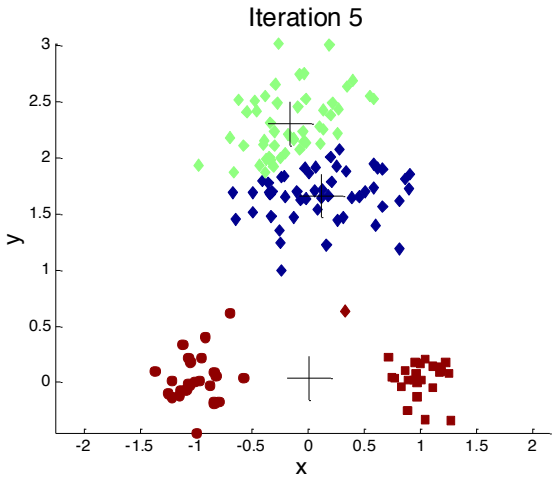
Importance du choix initial des centroïdes …











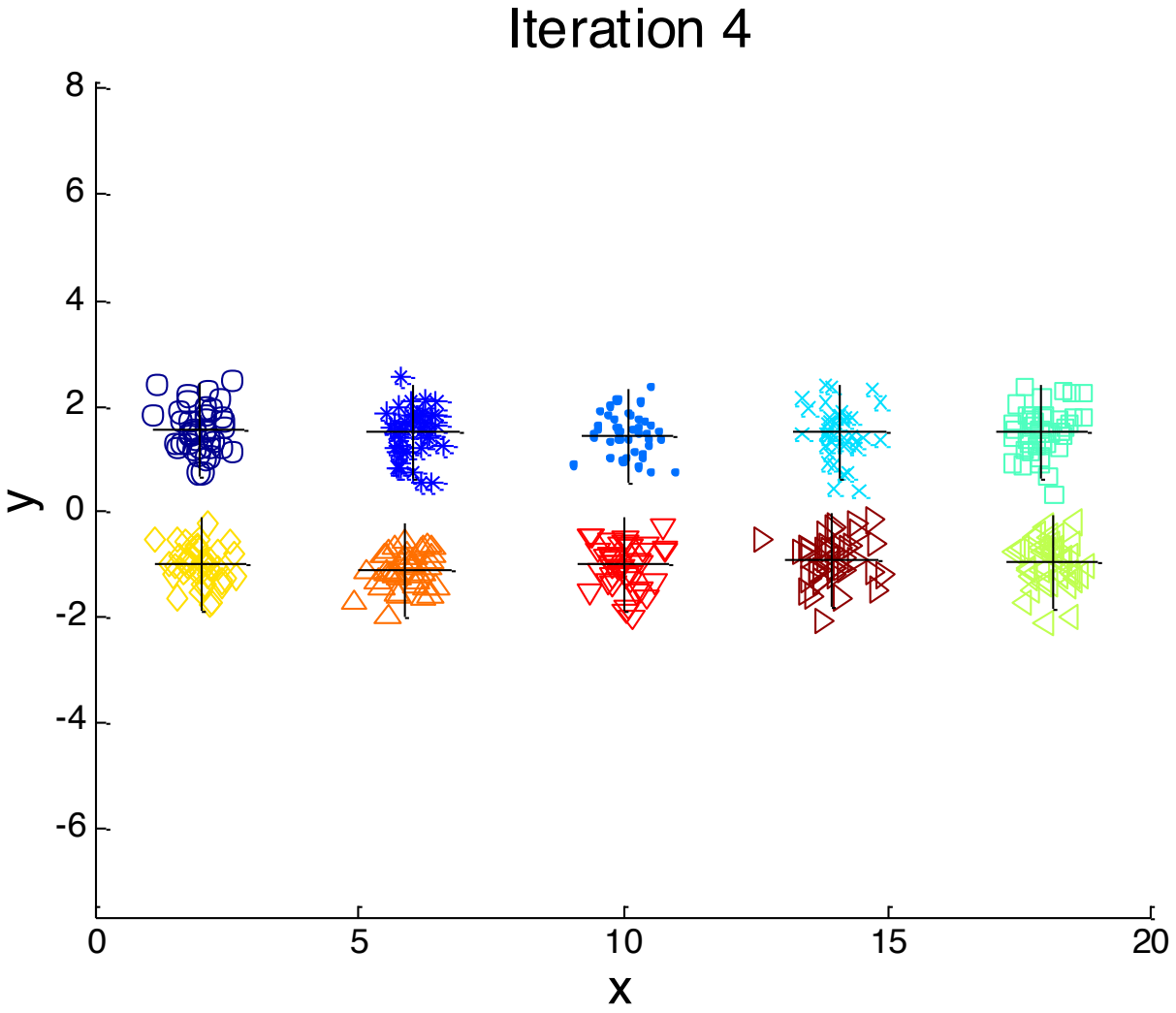
Problème de la sélection initiale des points

- Si il y a K clusters réels, alors la chance de sélectionner une centroïde dans chacun des clusters est faible
  - Chance relativement faible si K est élevé
  - Si les clusters ont la même taille, *n*, alors

$$P = \frac{\text{number of ways to select one centroid from each cluster}}{\text{number of ways to select } K \text{ centroids}} = \frac{K!n^K}{(Kn)^K} = \frac{K!}{K^K}$$

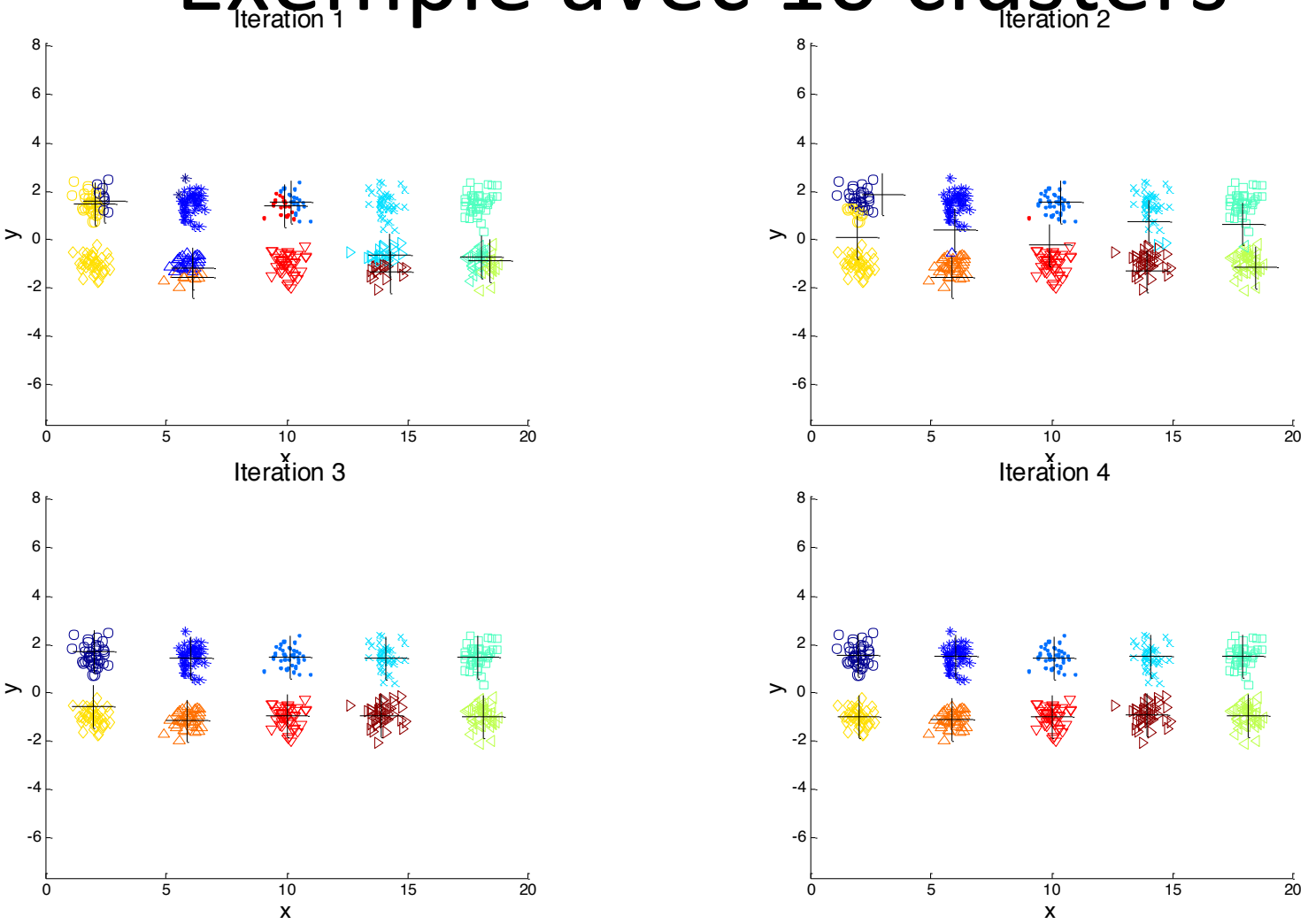
  - Par exemple, si K=10 la probabilité est 0.00036
  - Parfois, le choix de centroïdes initial se réajuste « bien », et parfois non
  - Exemple avec 5 paires de clusters

Exemple avec 10 clusters



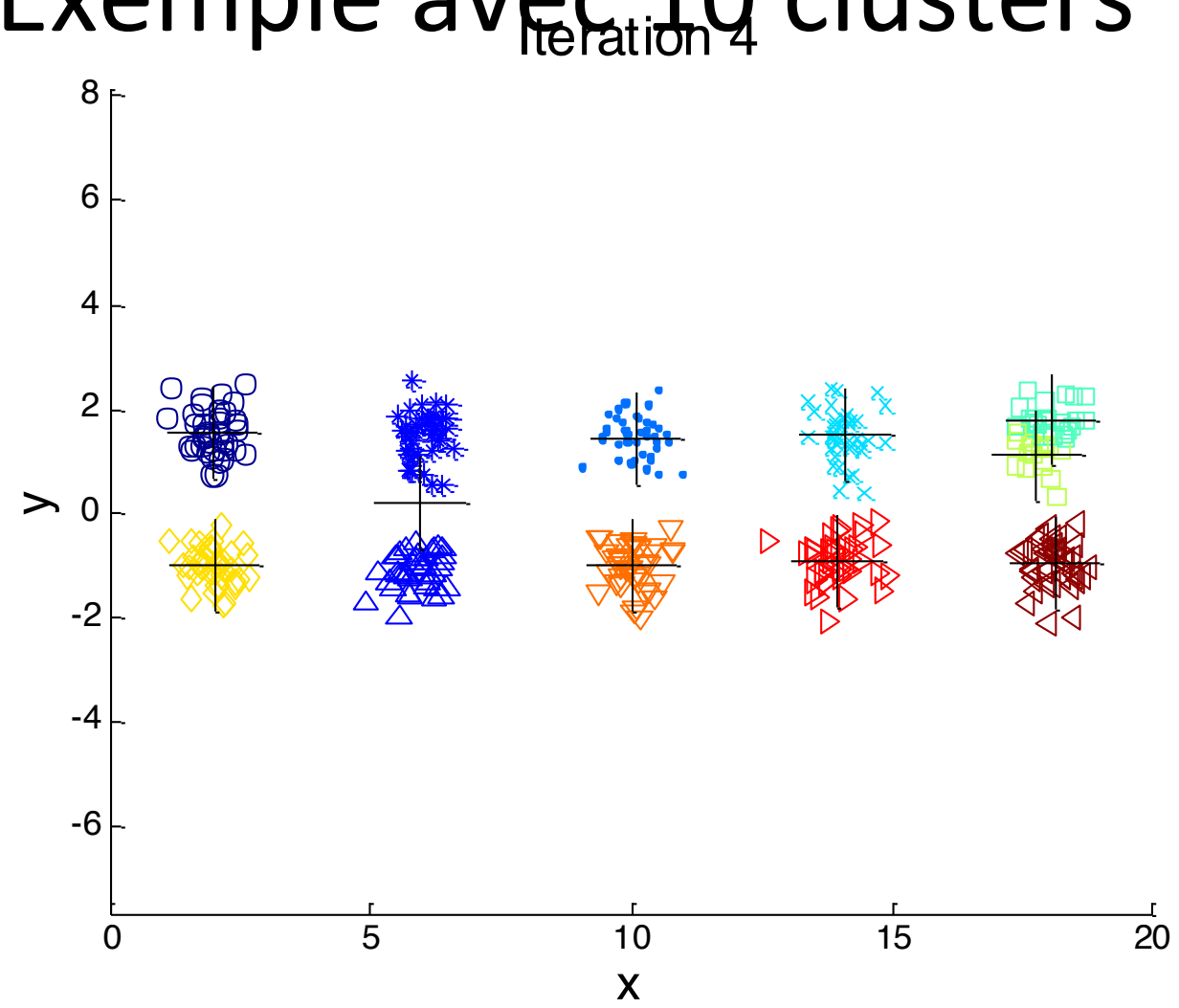
Démarrage avec 2 centroïdes initiales dans 1 cluster de chaque paire de clusters

Exemple avec 10 clusters



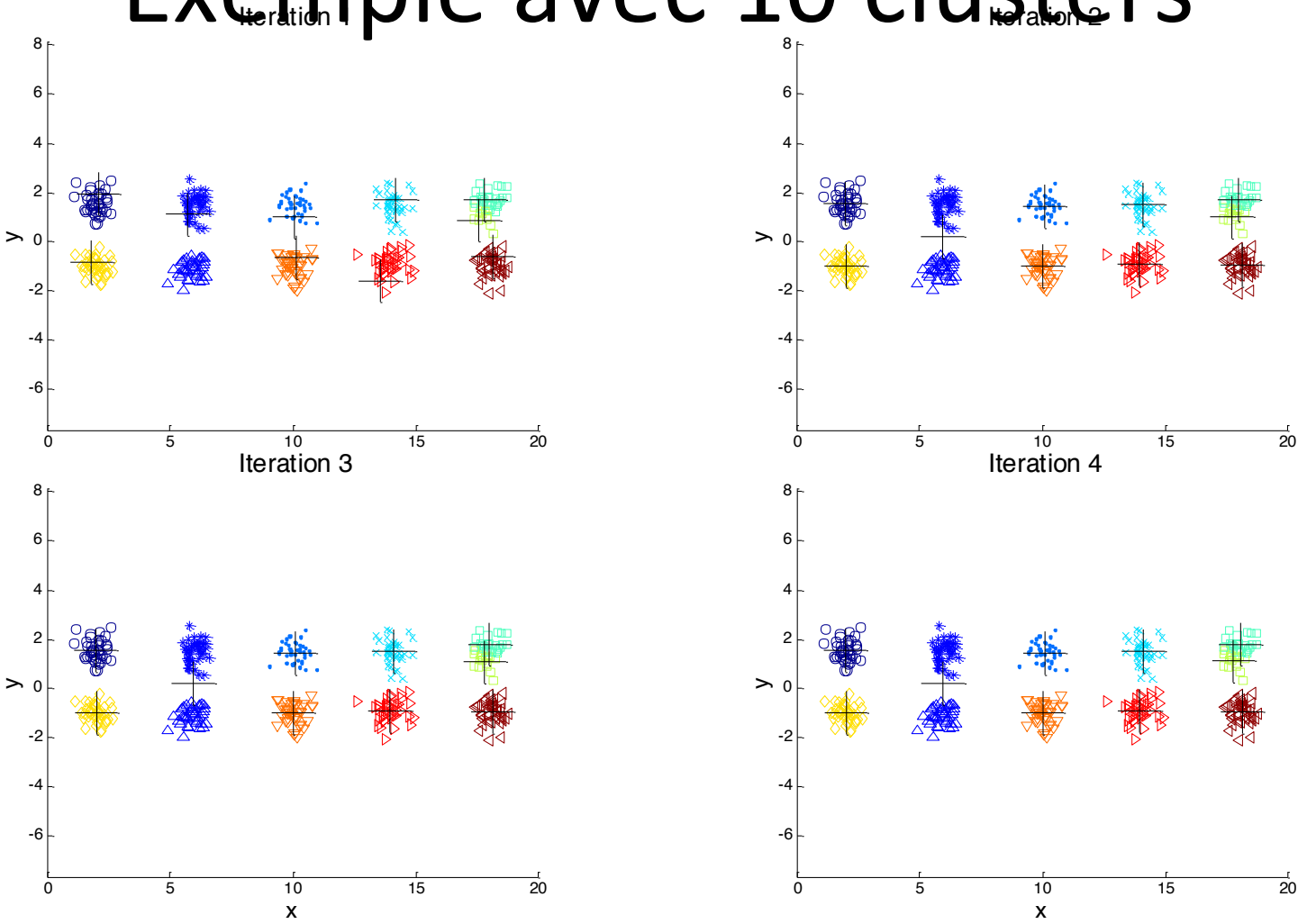
Démarrage avec 2 centroïdes initiales dans 1 cluster de chaque paire de clusters

Exemple avec 10 clusters



Démarrage avec 3 centroïdes dans certaines paires de clusters, 1 dans les autres

Exemple avec 10 clusters



Démarrage avec 3 centroïdes dans certaines paires de clusters, 1 dans les autres

Solutions au problème des centroïdes initiales

- Plusieurs exécutions
  - Aide, mais les probabilités ne sont pas en notre faveur
- Échantillonnage et clustering hiérarchique pour déterminer les centroïdes initiales
- Sélectionner plus de k centroïdes initiales puis sélectionner parmi ces centroïdes
  - Sélectionner les plus éloignées
- Bisection par K-means
  - Moins sensible aux problèmes d'initialisation

Gestion des clusters vides

- L'algorithme K-means « de base » peut générer des clusters vides
- Plusieurs stratégies
  - Choisir le point qui contribue le plus à la SSE
  - Choisir un point du cluster ayant la plus grande SSE
  - S'il y a plusieurs clusters vides, on peut répéter ces étapes

Mise à jour incrémentale des centres

- Dans l'algorithme K-means original, les centroïdes sont mises à jour après que les points soient assignées à un centroïd
- Une alternative est de mettre à jour après chaque assignation (approche incrémentale)
  - Chaque assignation met à jour 0 ou 2 centroïdes
  - Plus coûteux
  - Introduit un ordre de dépendances
  - Ne produit jamais de cluster vide
  - Peut utiliser des « poids » pour changer l'impact

Pré-processing et Post-processing

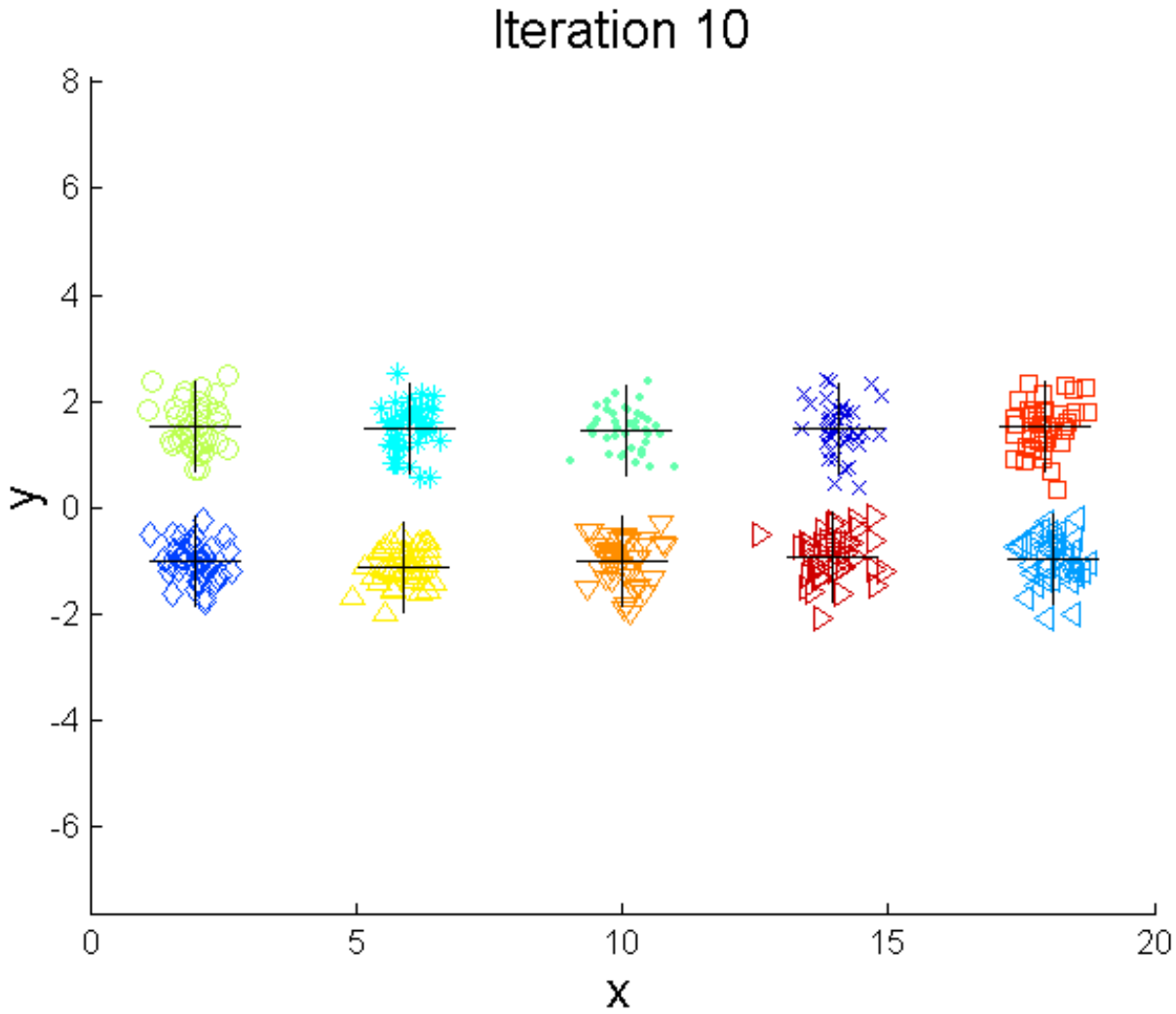
- Pré-processing
  - Normaliser les données
  - Éliminer les valeurs aberrantes
- Post-processing
  - Élimine les petits clusters qui pourraient être des aberrations
  - Divise les clusters « flottants », cad avec une SSE élevée
  - Fusionne les clusters qui sont « proches » et ont une SSE faible
  - On peut utiliser certaines de ces étapes pendant le clustering
    - ISODATA

Bisection par K-means

- Algorithme de bisection par K-means
  - Variante de K-means qui peut produire un clustering de partitions ou hiérarchique

1: Initialize the list of clusters to contain the cluster containing all points.
2: **repeat**
3: Select a cluster from the list of clusters
4: **for** $i = 1$ to *number_of_iterations* **do**
5: Bisect the selected cluster using basic K-means
6: **end for**
7: Add the two clusters from the bisection with the lowest SSE to the list of clusters.
8: **until** Until the list of clusters contains K clusters

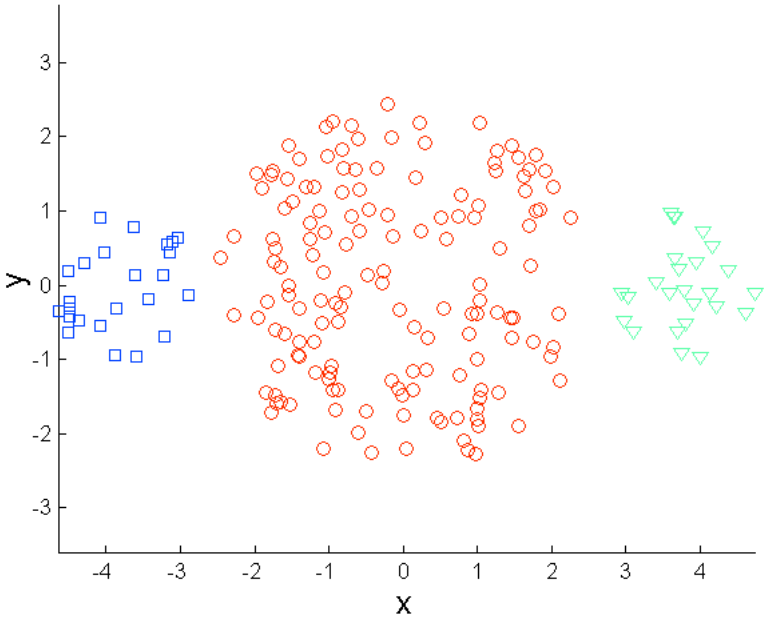
Exemple de bisection par K-means



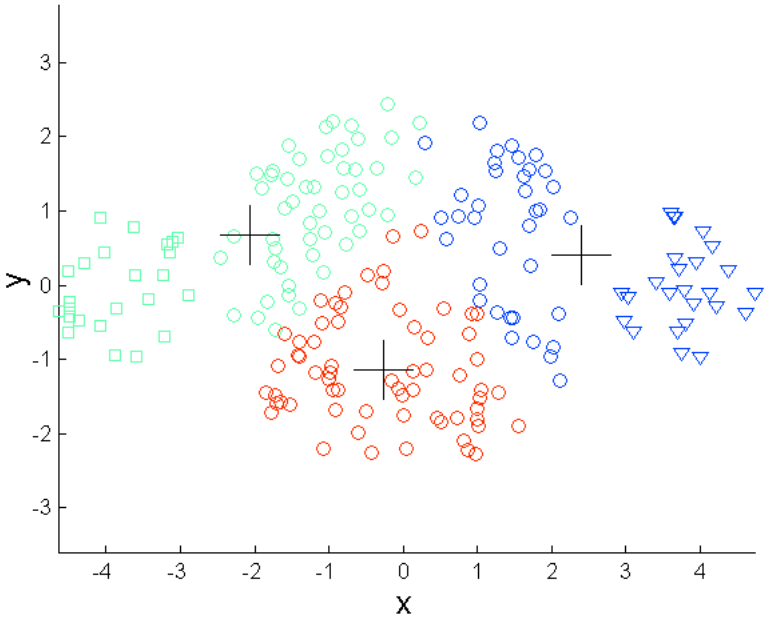
Limitations de K-means

- K-means a des problèmes lorsque les clusters sont
  - De différentes tailles
  - De différentes densités
  - De forme non sphérique
- K-means a des problèmes lorsque les donneés contiennent des valeurs aberrantes

Limitations de K-means: tailles différentes

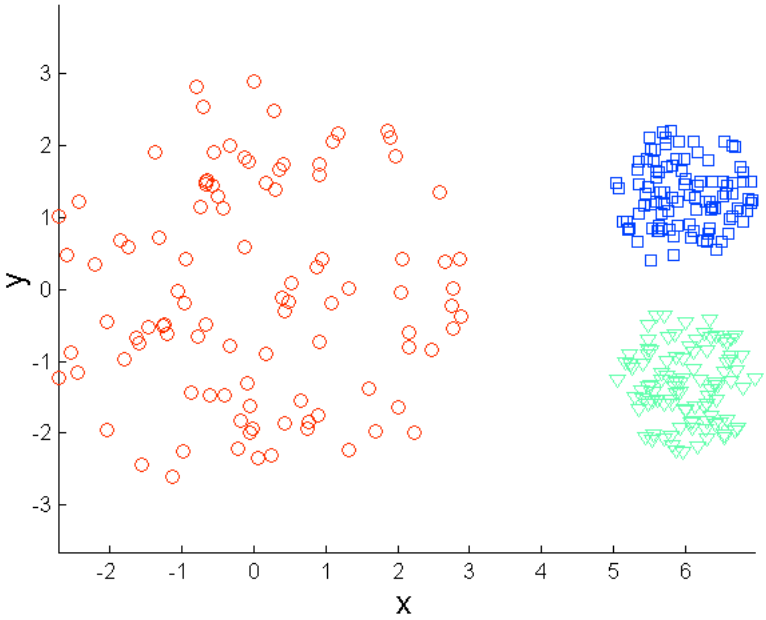


Original Points

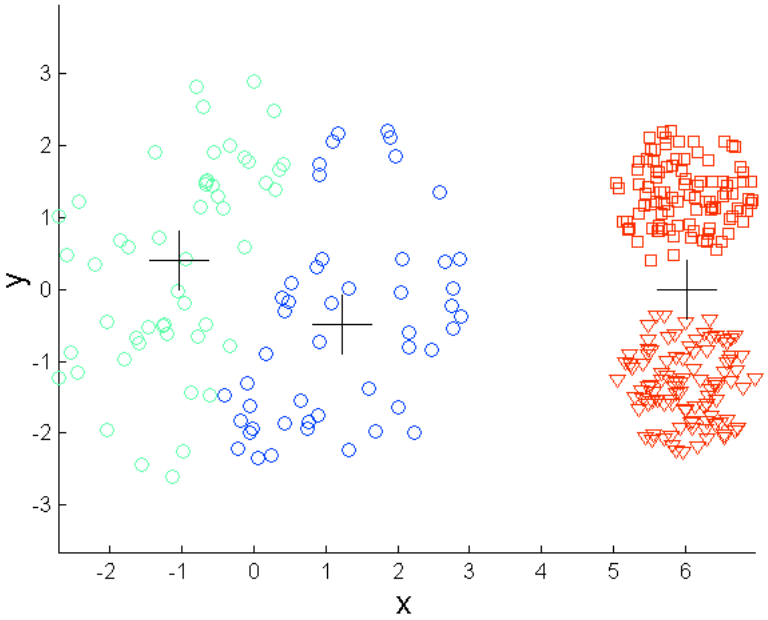


K-means (3 Clusters)

Limitations de K-means: différentes densités

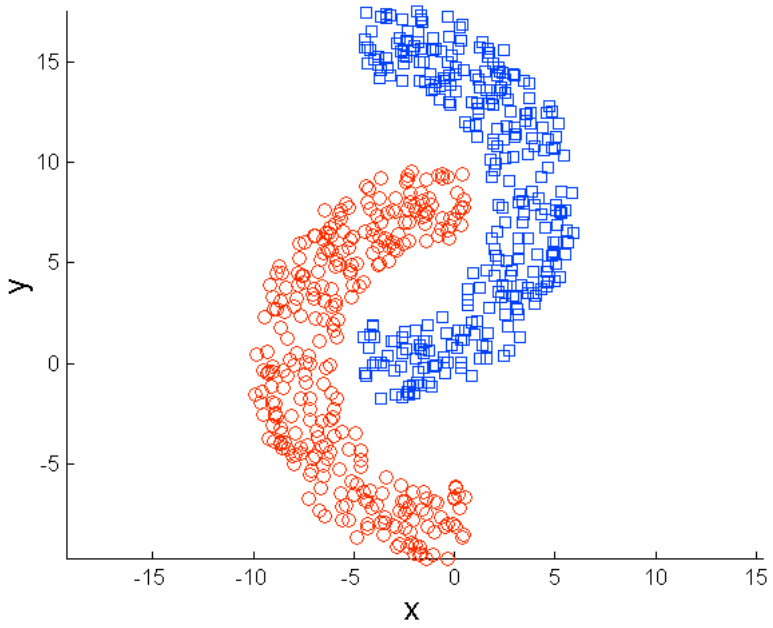


Original Points

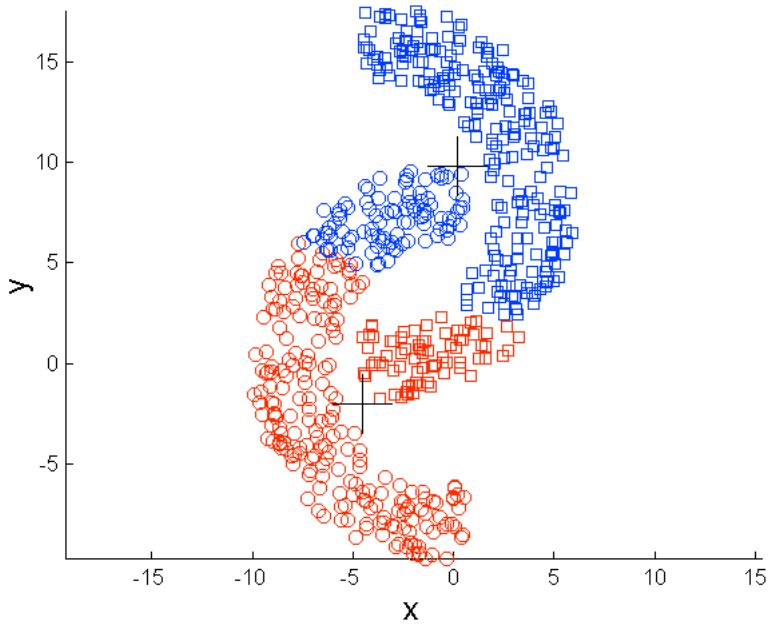


K-means (3 Clusters)

Limitations de K-means: formes non sphériques

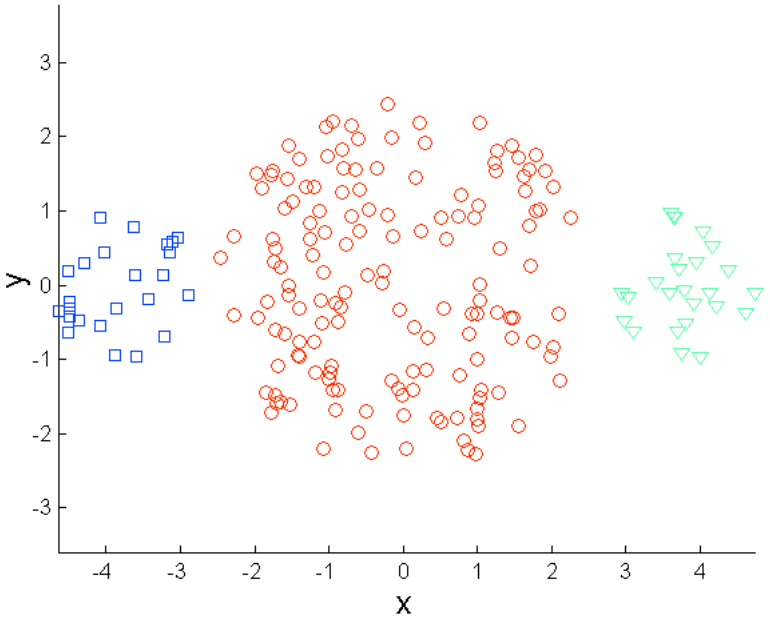


Original Points

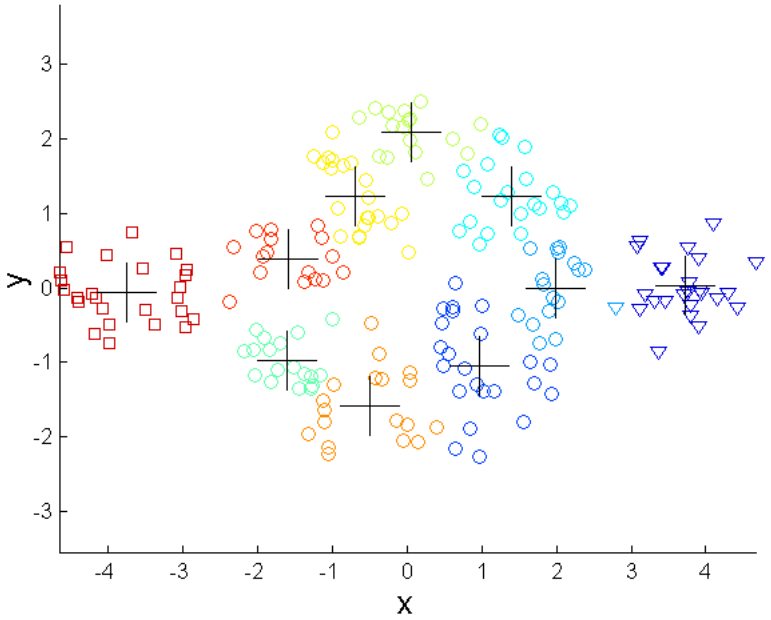


K-means (2 Clusters)

Surmonter les limitations de K-means



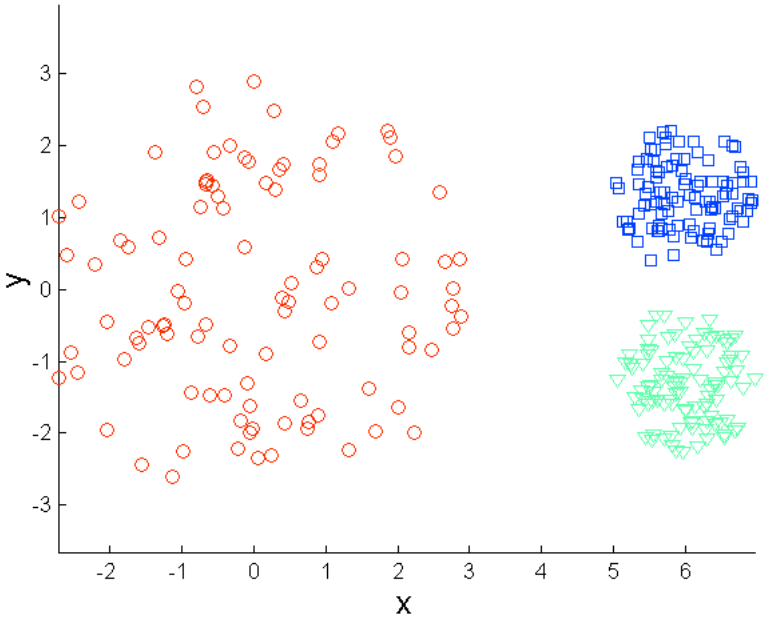
Original Points



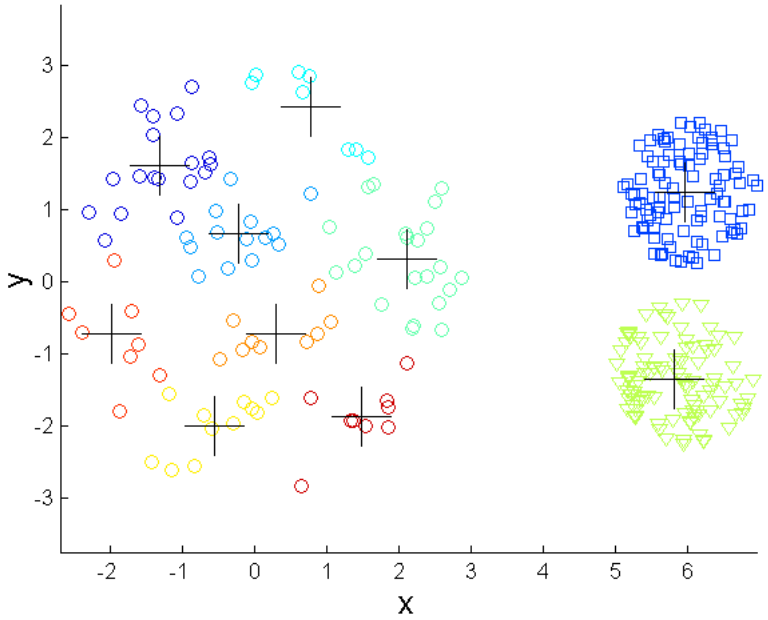
K-means Clusters

- Une solution est de construire beaucoup de clusters, cela forme des parties de clusters mais il faut les fusionner par la suite

Surmonter les limitations de K-means

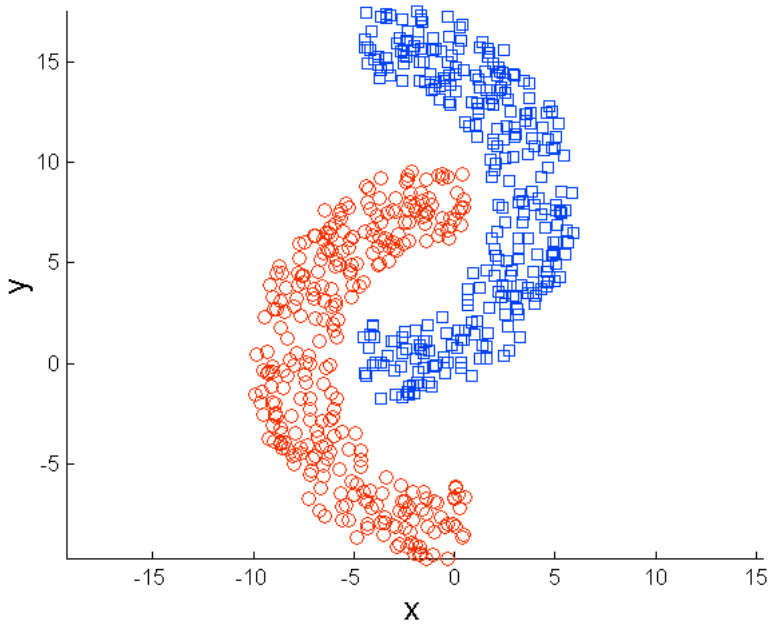


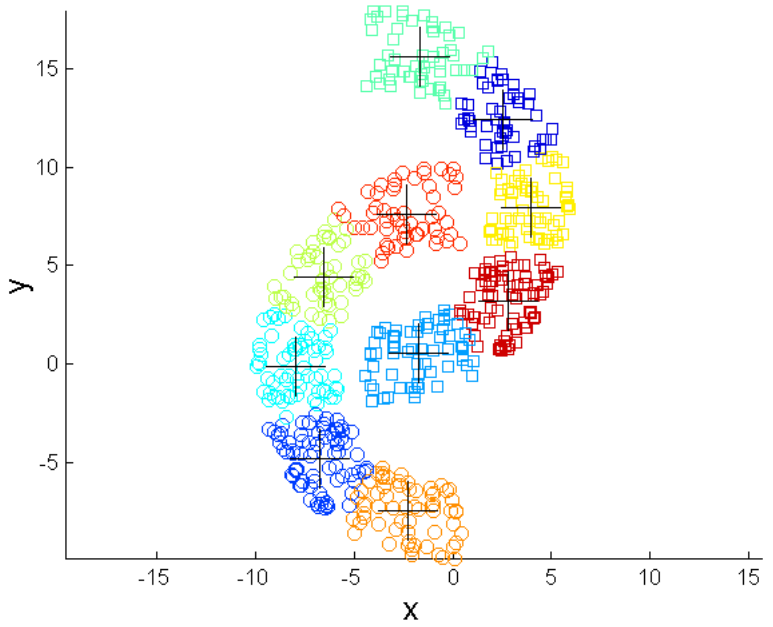
Original Points



K-means Clusters

Surmonter les limitations de K-means



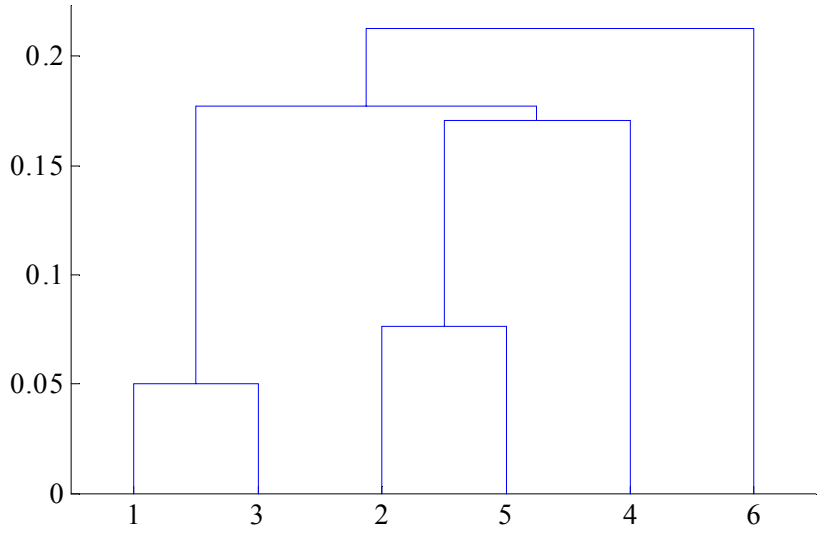


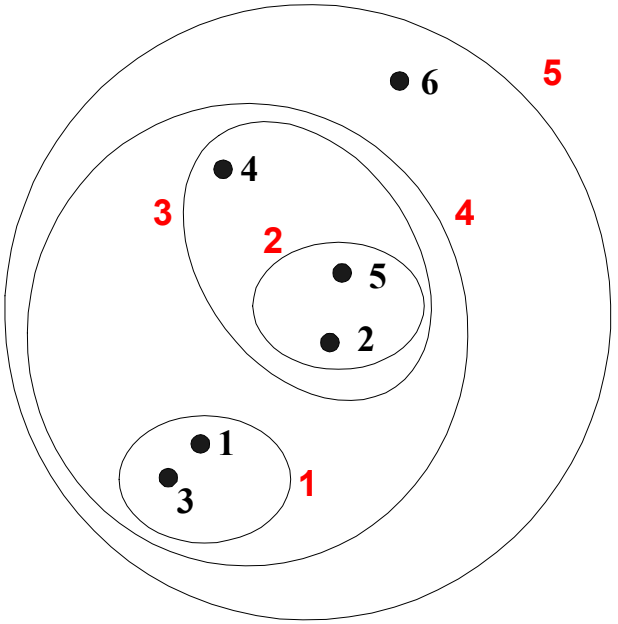
Original Points

K-means Clusters

Clustering Hiérarchique

- Produit un ensemble de clusters imbriqués organisés suivant un arbre hiérarchique
- Peut être visualisé sous forme de dendrogramme
  - Un diagramme sous forme d'arbre qui enregistre les divisions/fusions successives





Avantages du clustering hiérarchique

- On n'a pas besoin d'avoir un nombre de clusters en paramètre
  - On peut obtenir n'importe quel nombre de clusters en « coupant » le dendrogramme au niveau souhaité
- Cette hiérarchie peut correspondre à une taxonomie
  - Exemples en biologie …

Clustering hiérarchique

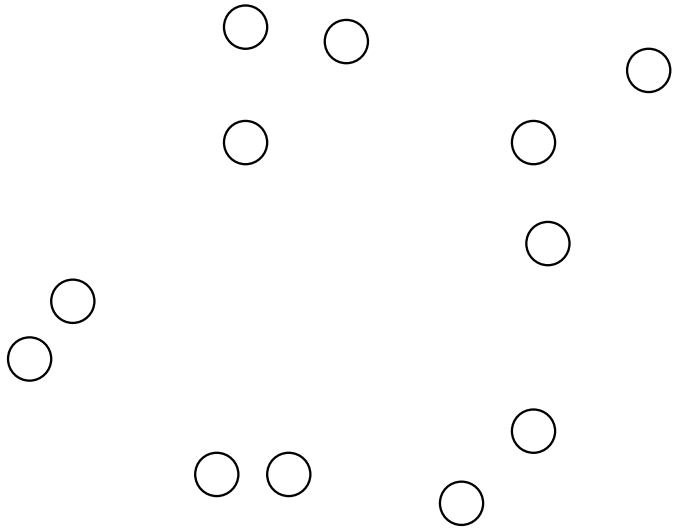
- Deux types principaux
  - Clustering agglomératif
    - On part de points individuels, un cluster par point
    - À chaque étape on fusionne la paire de clusters la plus proche jusqu'à ce qu'il ne reste qu'un (ou k) cluster(s)
  - Clustering par division
    - On part avec un seul cluster contenant tous les points
    - À chaque étape, on divise un cluster jusqu'à ce que chaque cluster ne contienne qu'un (ou k) point(s)
- Les algorithmes hiérarchiques classiques utilisent une matrice de similarité / distance
  - Division ou fusion d'un cluster à la fois

Algorithme de clustering agglomératif

- Méthode la plus courante pour le clustering hiérarchique
- L'algorithme est relativement simple
  1. Calculer la matrice de similarité
  2. Initialiser un cluster à chaque point
  3. Répéter
  4. Fusionner les 2 clusters les plus proches
  5. Mettre à jour la matrice de similarité
  6. Jusqu'à ce qu'il ne reste plus qu'un cluster
- L'opération clé est la mise à jour du calcul de similarité entre 2 clusters
  - Différentes approches pour définir la distance entre 2 clusters

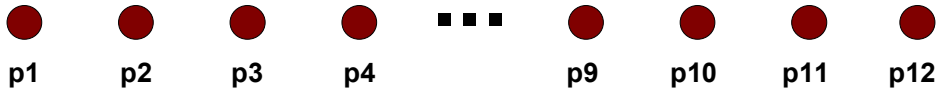
Situation de départ

- Démarrer avec des clusters pour les points individuels et une matrice de proximité



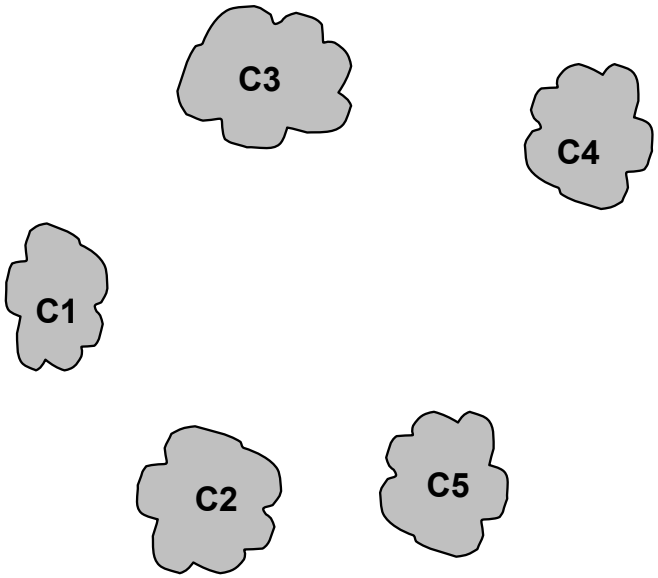
	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Proximity Matrix



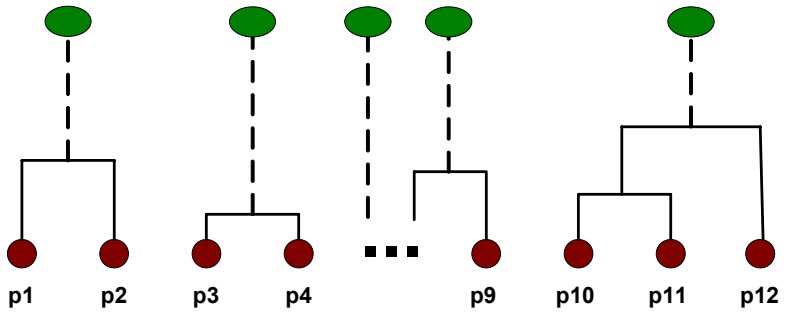
Étape intermédifaire

- Après plusieurs étapes de fusion, nous avons des clusters



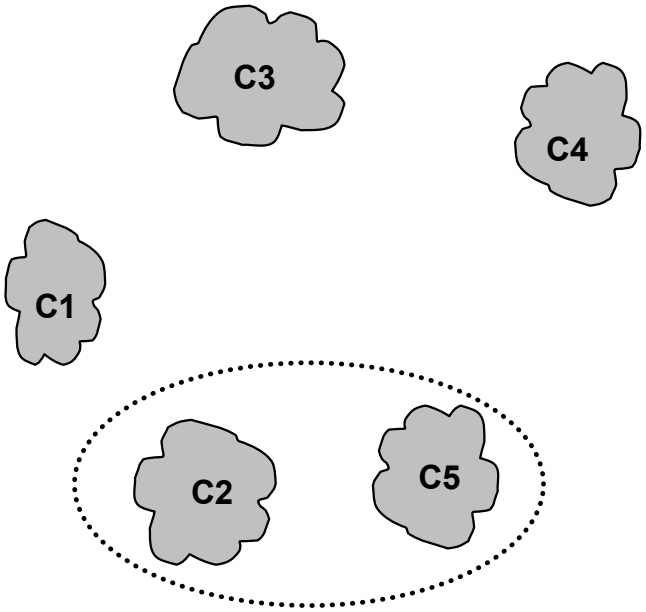
	C1	C2	C3	C4	C5
C1					
C2					
C3					
C4					
C5					

Proximity Matrix



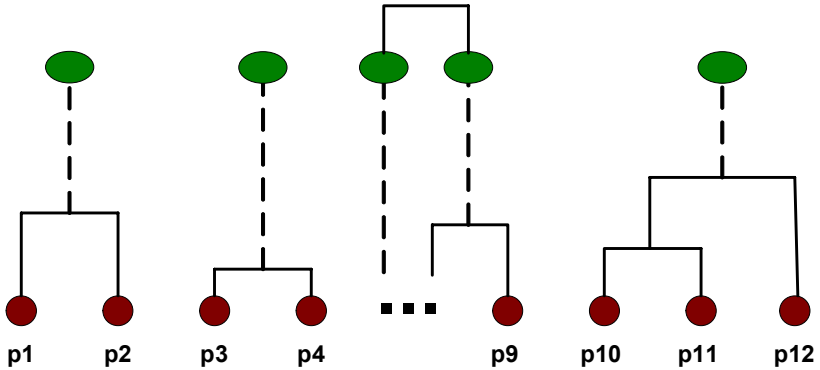
Situation intermédiaire

- On veut fusionner les 2 clusters (C2 et C5) et mettre à jour la matrice de proximité



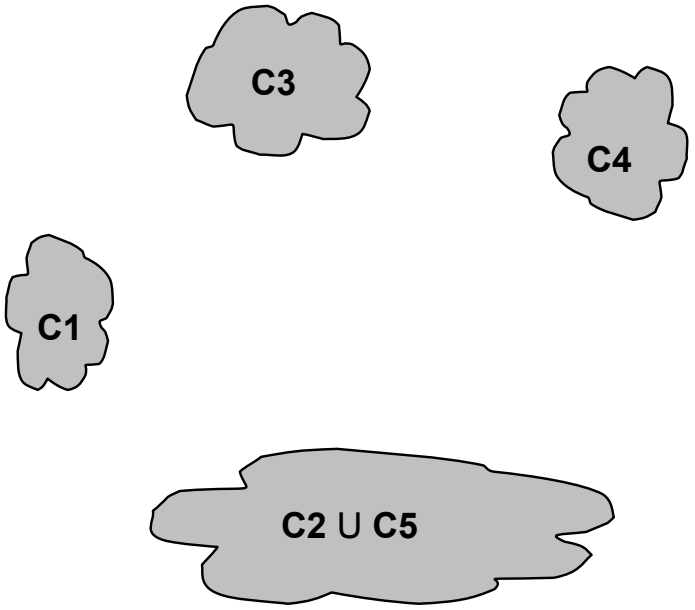
	C1	C2	C3	C4	C5
C1					
C2					
C3					
C4					
C5					

Proximity Matrix



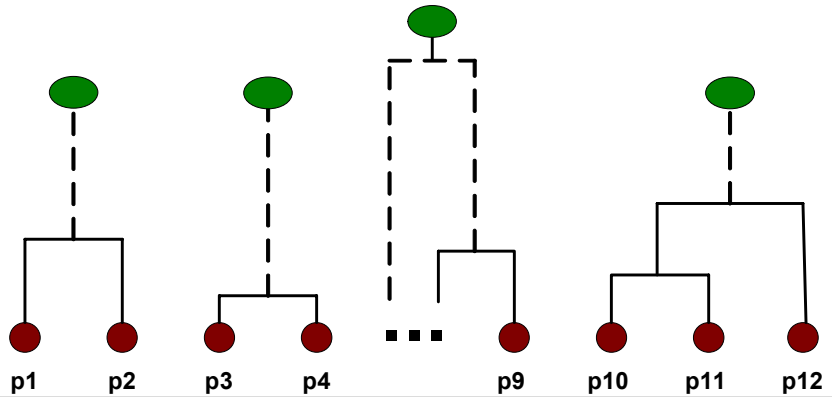
Après la fusion

- La question est comment mettre à jour la matrice de proximité ?

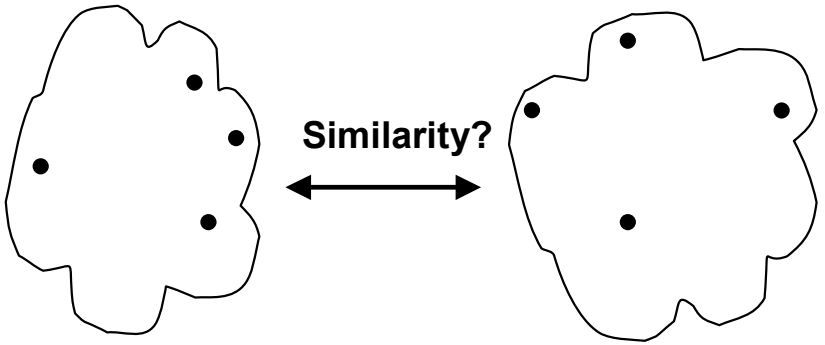


	C1	C2 U C5	C3	C4
C1		?		
C2 U C5	?	?	?	?
C3		?		
C4		?		

Proximity Matrix



Comment définir la similarité entre clusters

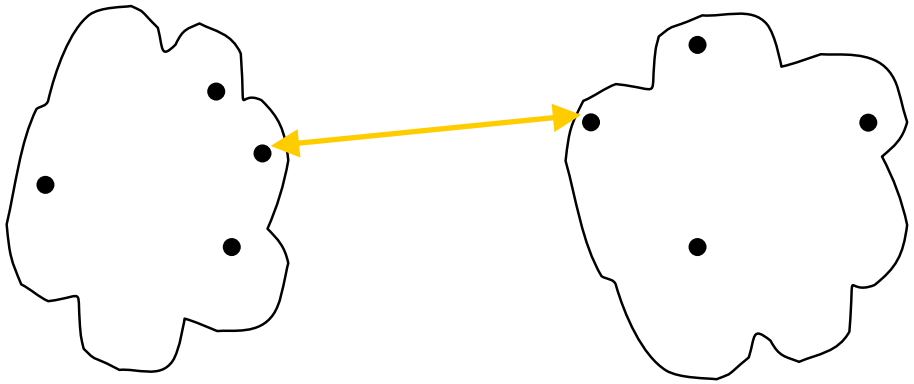


	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Proximity Matrix

- MIN
- MAX
- Moyenne sur le groupe
- Distance entre les centroïdes
- Autre méthode obtenue grâce à une fonction objectif
  - La méthode de Ward utilise l'erreur au carré

Comment définir la similarité entre clusters

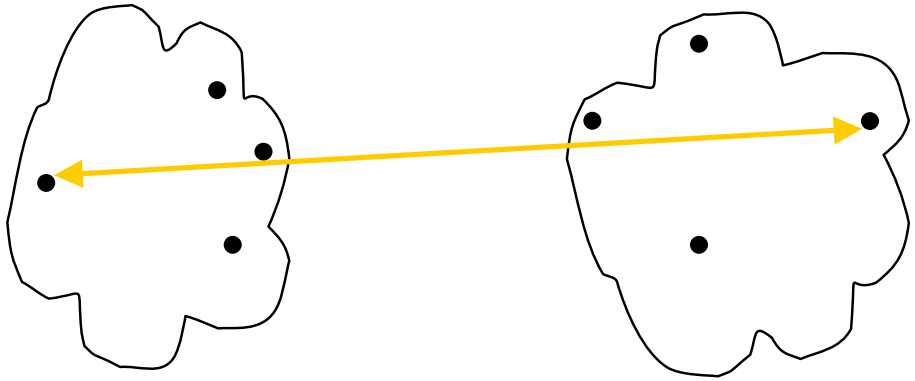


	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Proximity Matrix

- MIN
- MAX
- Moyenne sur le groupe
- Distance entre les centroïdes
- Autre méthode obtenue grâce à une fonction objectif
  - La méthode de Ward utilise l'erreur au carré

Comment définir la similarité entre clusters

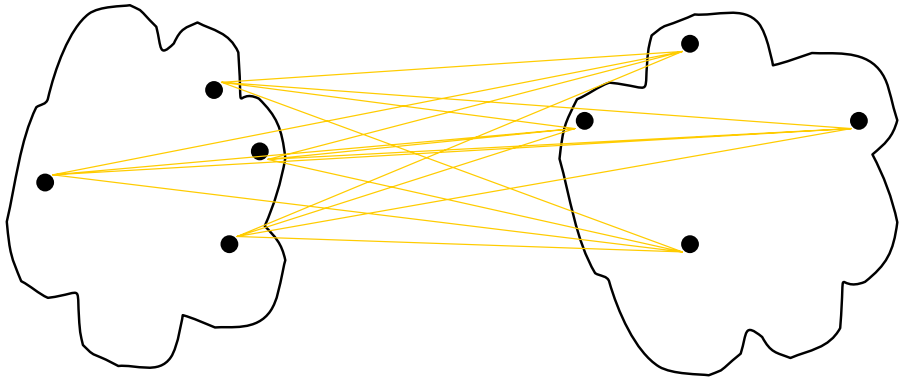


	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Proximity Matrix

- MIN
- MAX
- Moyenne sur le groupe
- Distance entre les centroïdes
- Autre méthode obtenue grâce à une fonction objectif
  - La méthode de Ward utilise l'erreur au carré

Comment définir la similarité entre clusters

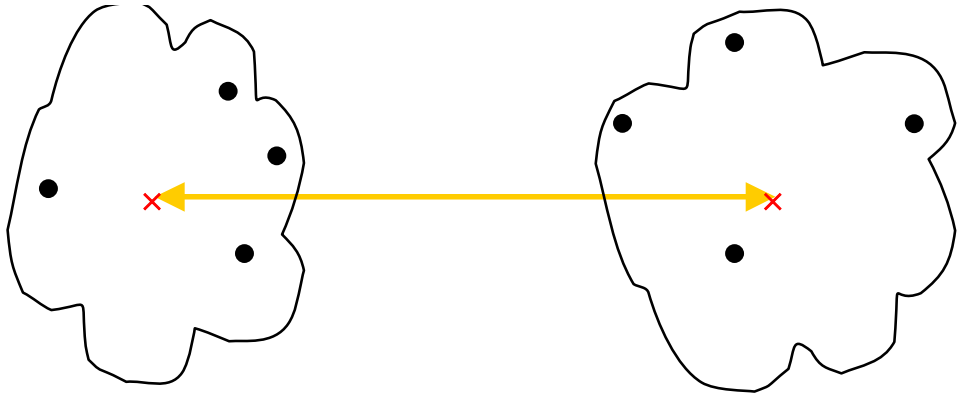


	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Proximity Matrix

- MIN
- MAX
- Moyenne sur le groupe
- Distance entre les centroïdes
- Autre méthode obtenue grâce à une fonction objectif
  - La méthode de Ward utilise l'erreur au carré

Comment définir la similarité entre clusters



	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

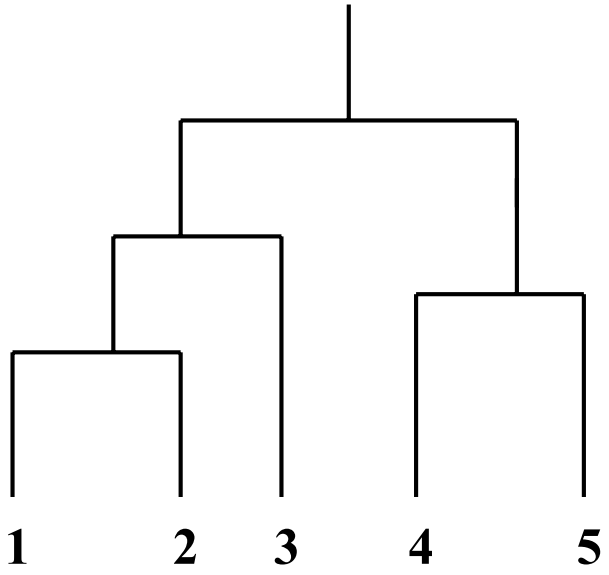
Proximity Matrix

- MIN
- MAX
- Moyenne sur le groupe
- Distance entre les centroïdes
- Autre méthode obtenue grâce à une fonction objectif
  - La méthode de Ward utilise l'erreur au carré

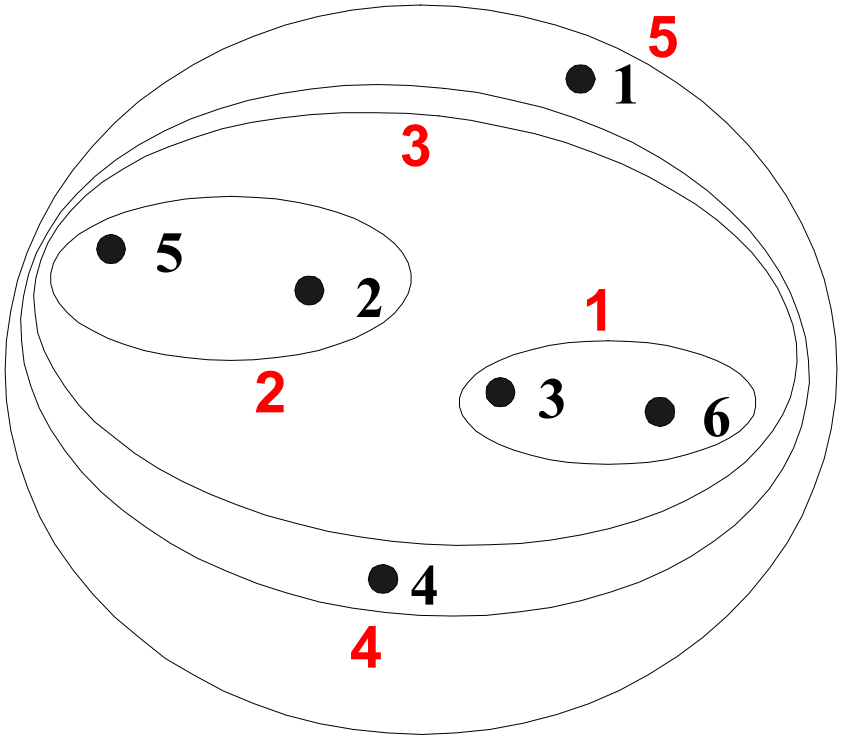
Similarité entre clusters: MIN

- La similarité entre 2 clusters est basée sur les 2 points les plus similaires entre les 2 clusters
  - Déterminé par une paire de points, i.e. par un lien dans le graphe de proximité

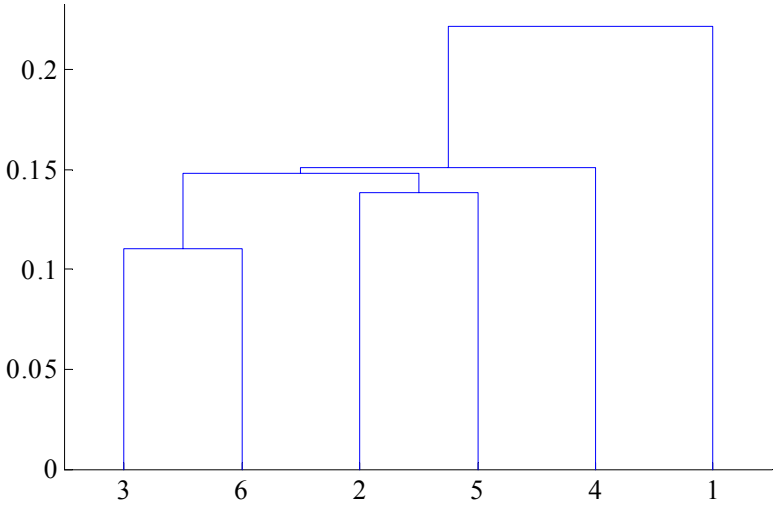
	I1	I2	I3	I4	I5
I1	1.00	0.90	0.10	0.65	0.20
I2	0.90	1.00	0.70	0.60	0.50
I3	0.10	0.70	1.00	0.40	0.30
I4	0.65	0.60	0.40	1.00	0.80
I5	0.20	0.50	0.30	0.80	1.00



Clustering hiérarchique: MIN



Nested Clusters



Dendrogram

Avantage de MIN

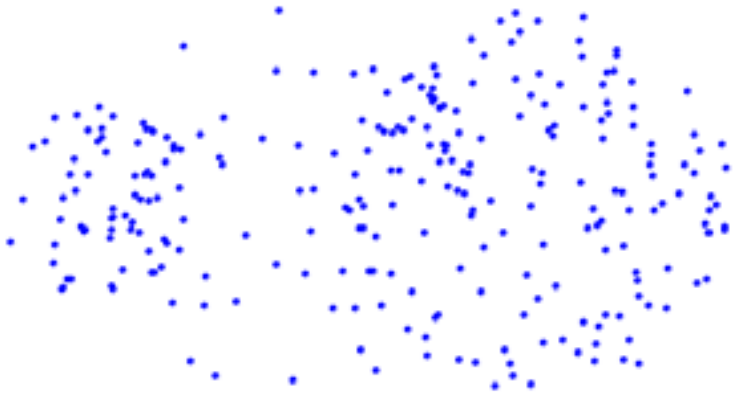
- Peut gérer les formes non elliptiques

Original Points

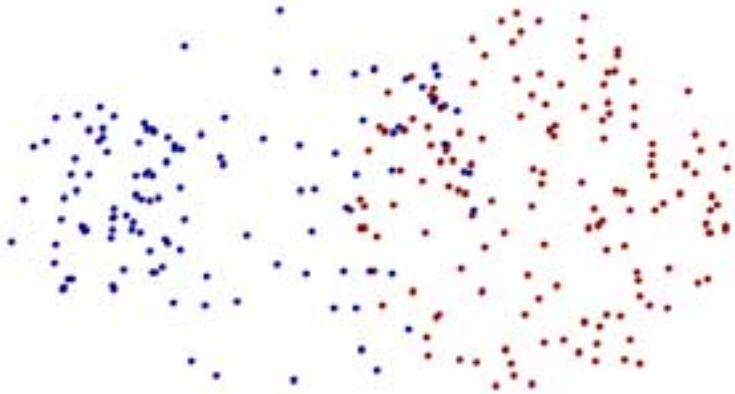
Two Clusters

Limite de MIN

- Sensible au bruit



Original Points

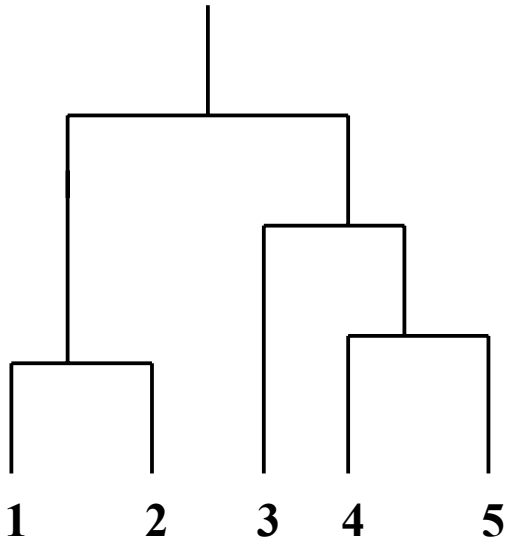


Two Clusters

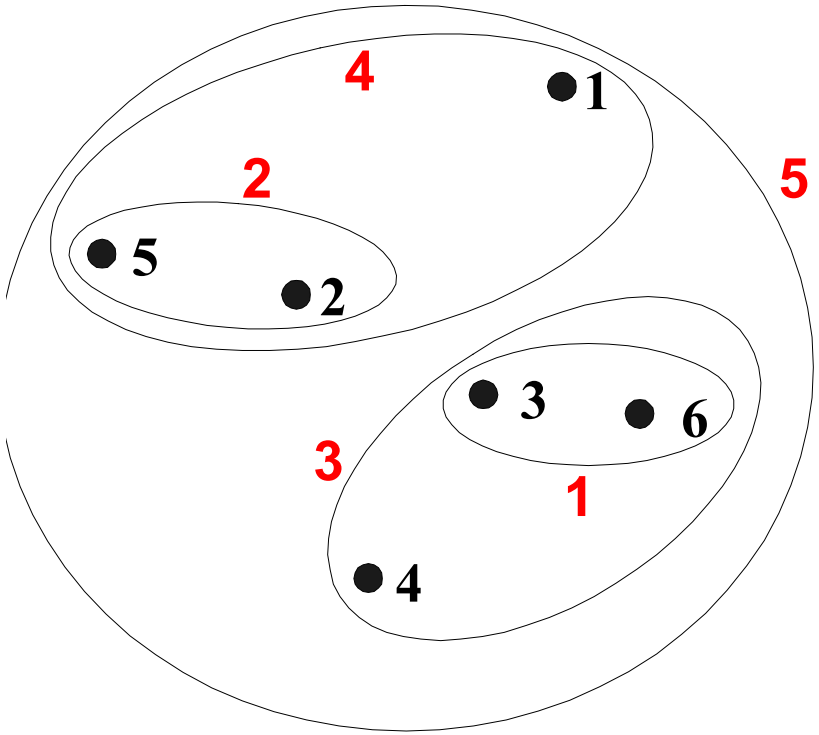
Similarité entre clusters : MAX

- La similarité entre 2 clusters est basée sur les 2 points les plus différents des clusters
  - Déterminé par toutes les paires de points entre les 2 clusters

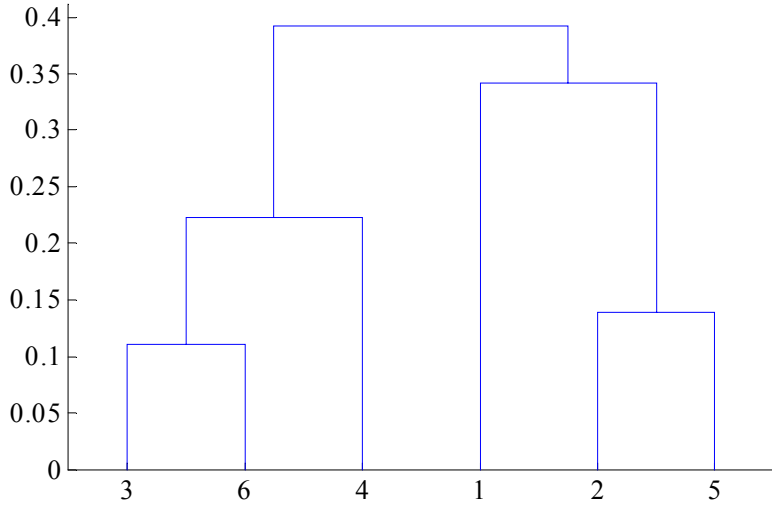
	I1	I2	I3	I4	I5
I1	1.00	0.90	0.10	0.65	0.20
I2	0.90	1.00	0.70	0.60	0.50
I3	0.10	0.70	1.00	0.40	0.30
I4	0.65	0.60	0.40	1.00	0.80
I5	0.20	0.50	0.30	0.80	1.00



Clustering hiérarchique: MAX



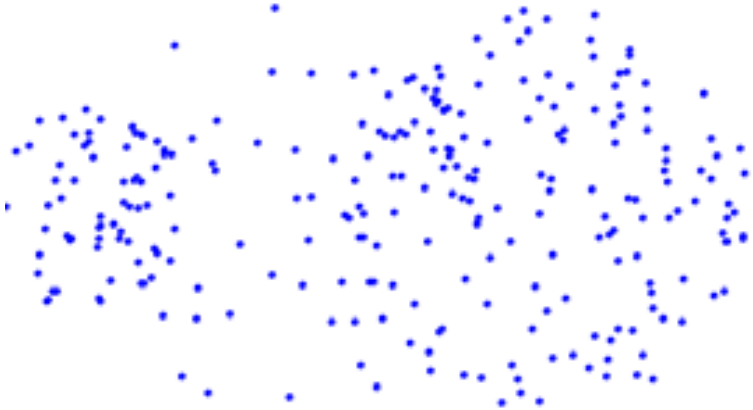
Nested Clusters



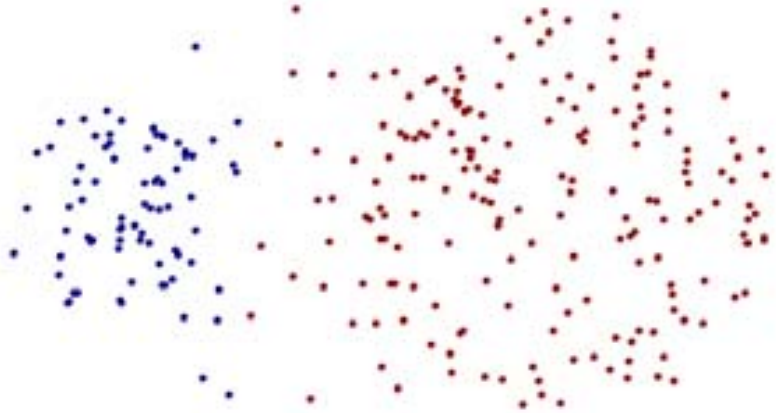
Dendrogram

Avantage de MAX

- Moins sensible au bruit



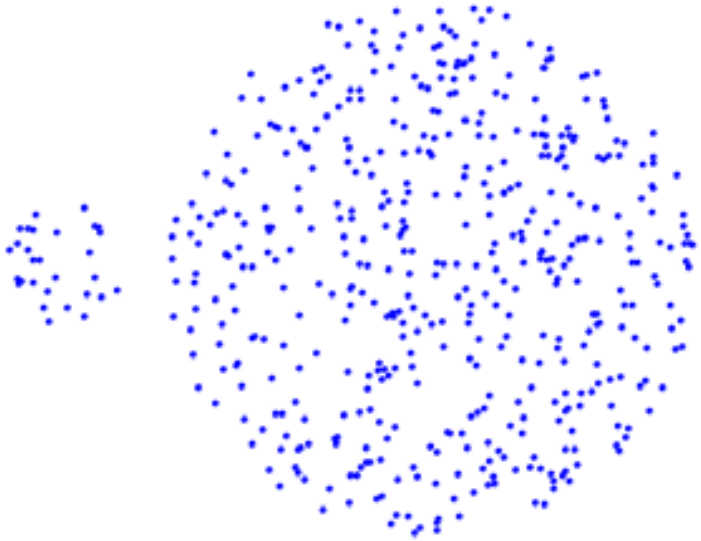
Original Points



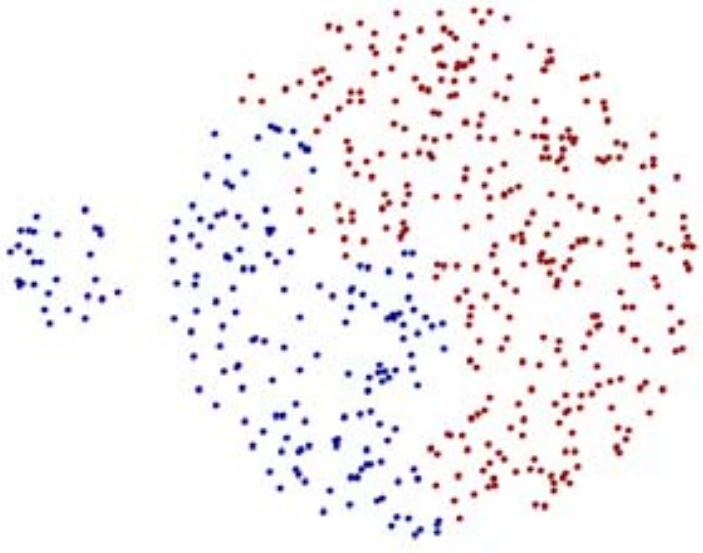
Two Clusters

Limite de MAX

- A tendance à casser les grands clusters
- Biaisé vers des clusters sphériques



Original Points

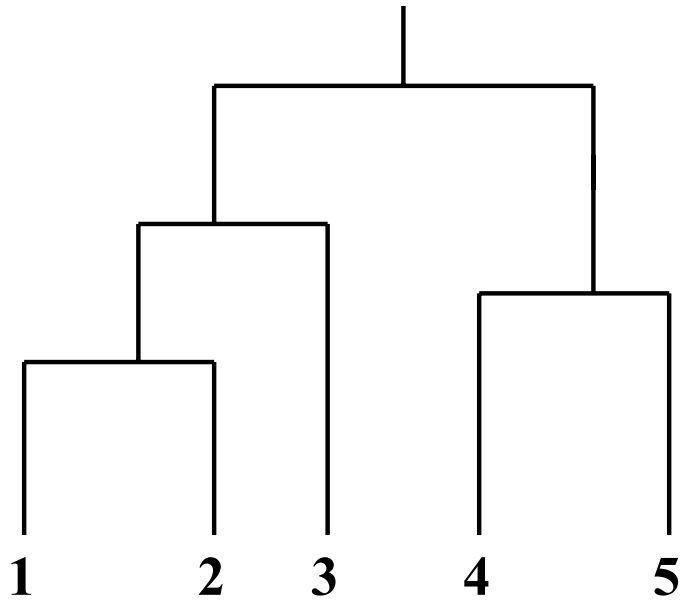


Two Clusters

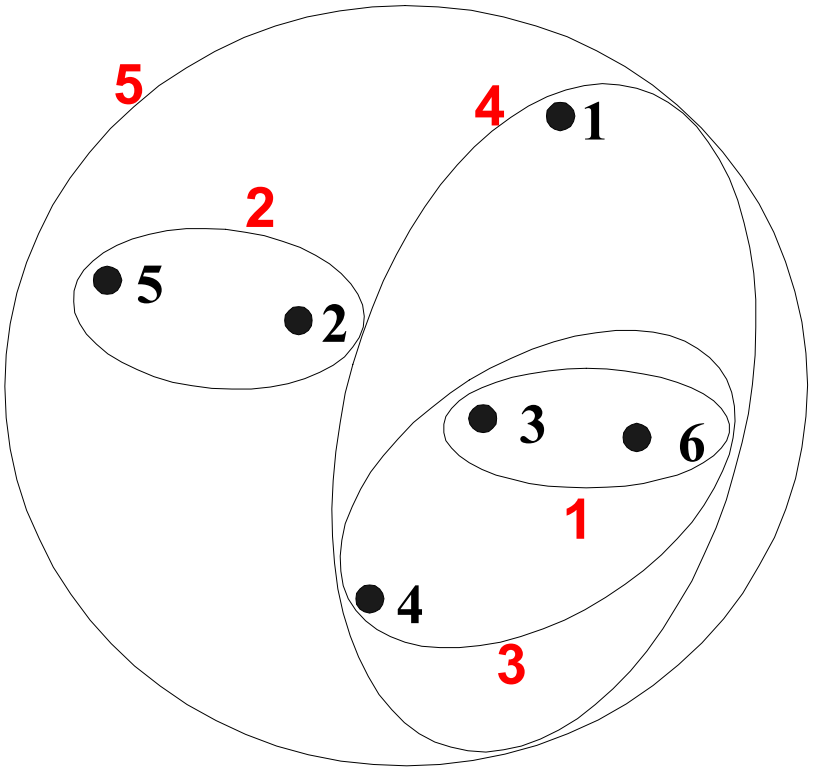
Similarité entre clusters : moyenne de groupe

- La proximité entre 2 clusters est la moyenne des distances entre les points de ces clusters

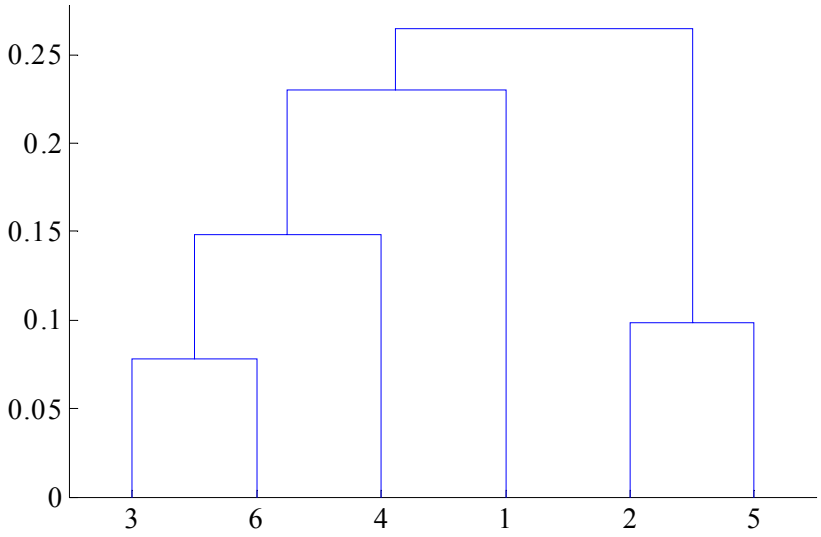
	I1	I2	I3	I4	I5
I1	1.00	0.90	0.10	0.65	0.20
I2	0.90	1.00	0.70	0.60	0.50
I3	0.10	0.70	1.00	0.40	0.30
I4	0.65	0.60	0.40	1.00	0.80
I5	0.20	0.50	0.30	0.80	1.00



Clustering hiérarchique : moyenne de groupe



Nested Clusters



Dendrogram

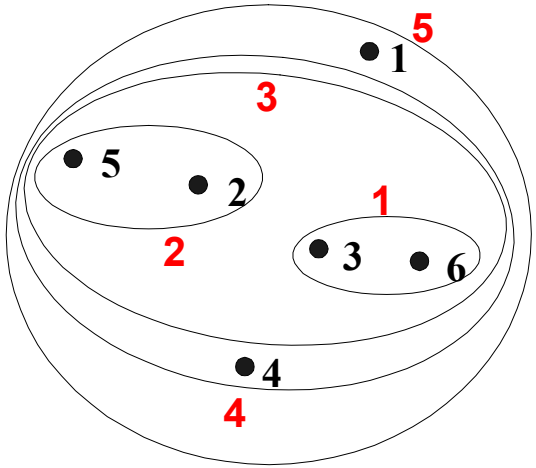
Clustering hiérarchique : moyenne de groupe

- Compromis entre Min et MAX
- Avantage
  - Moins susceptible au bruit
- Limitation
  - Biaisé vers les clusters sphériques

Similarité entre clusters : méthode de Ward

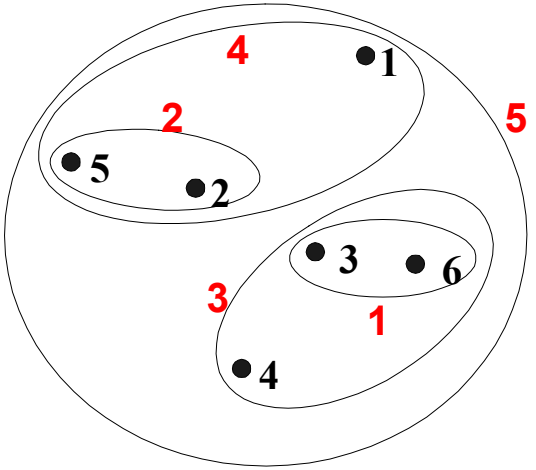
- La similarité entre 2 clusters est basée sur l'augmentation de l'erreur au carré lorsque les 2 clusters sont fusionnés
  - Similaire à la moyenne de groupe si la distance entre les points est mise au carré
- Moins susceptible au bruit
- Biaisé vers les clusters sphériques
- Équivalent hiérarchique de k-means
  - Peut être utilisé pour initialiser k-means

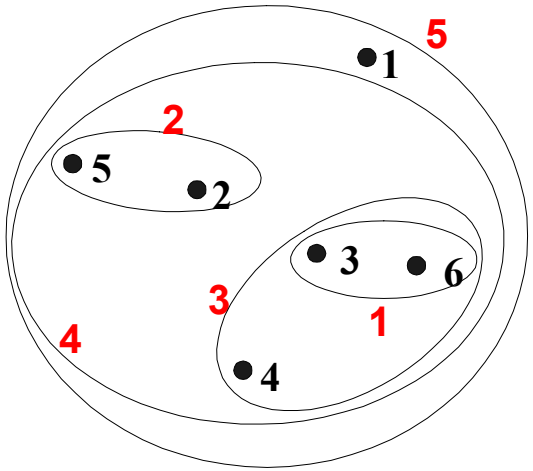
Clustering hiérarchique : comparaison



MIN

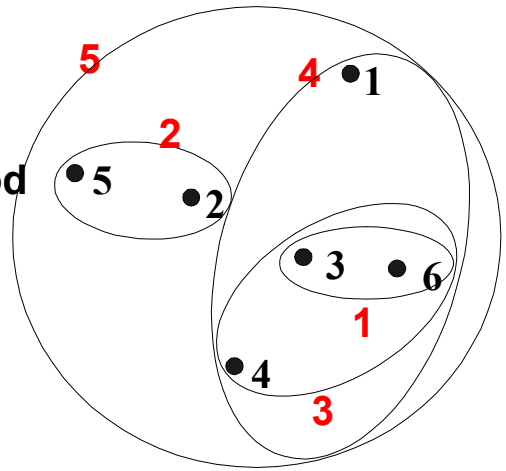
MAX





Group Average

Ward's Method



Clustering hiérarchique : complexité

- $O(N^2)$ en espace puisqu'on calcule la matrice de similarité
  - N = nombre de points
- $O(N^3)$ en temps dans beaucoup de cas
  - Il y a N étapes, et à chaque étape on met à jour/ recherche dans la matrice de similarité de taille N^2
  - Certaines approchent réduisent cette complexité à $O(N^2 \log(N))$

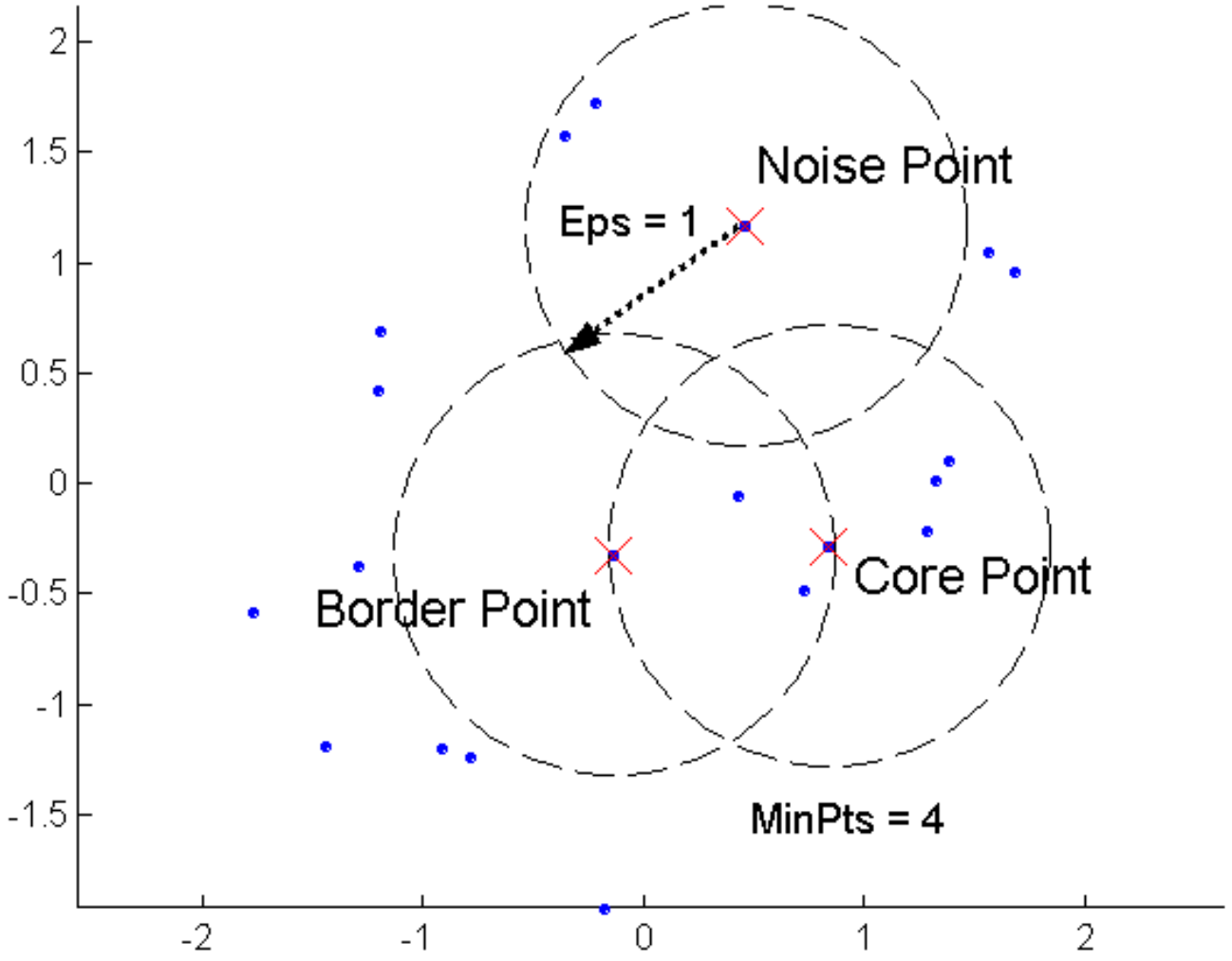
Clustering hiérarchique : Problèmes et limitations

- Lorsque la décision de fusionner 2 clusters est prise, on ne peut pas l'annuler
- Aucune fonction objectif n'est minimisée directement
- Différentes approches ont différents problèmes avec les aspects suivants
  - Sensibilité au bruit
  - Difficulté à gérer des clusters de taille différente ou de forme convexe
  - Division des grands clusters

DBSCAN : clustering par densité

- DBSCAN est un algorithme de clustering basé sur la densité
  - Densité = nombre de points dans un rayon donné (Eps)
  - Un point est un point central si il a plus d'un nombre spécifique de points (MinPts) dans un rayon Eps
  - Un point extérieur a moins de MinPts points dans un rayon Eps, mais est voisin d'un point central
  - Une valeur aberrante est un point qui n'est ni central ni extérieur

DBSCAN


Noise Point
Eps = 1
Border Point
Core Point
MinPts = 4
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-1
0
1
2

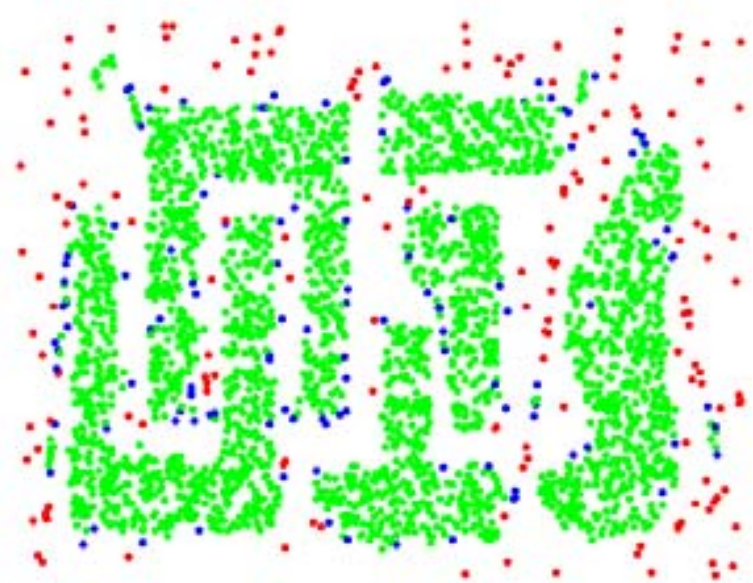
Algorithme DBSCAN

- Éliminer le bruit
- Faire le clustering sur les points restants

```
current_cluster_label ← 1
for all core points do
  if the core point has no cluster label then
    current_cluster_label ← current_cluster_label + 1
    Label the current core point with cluster label current_cluster_label
  end if
  for all points in the Eps-neighborhood, except i^th the point itself do
    if the point does not have a cluster label then
      Label the point with cluster label current_cluster_label
    end if
  end for
end for
```

DBSCAN

Original Points

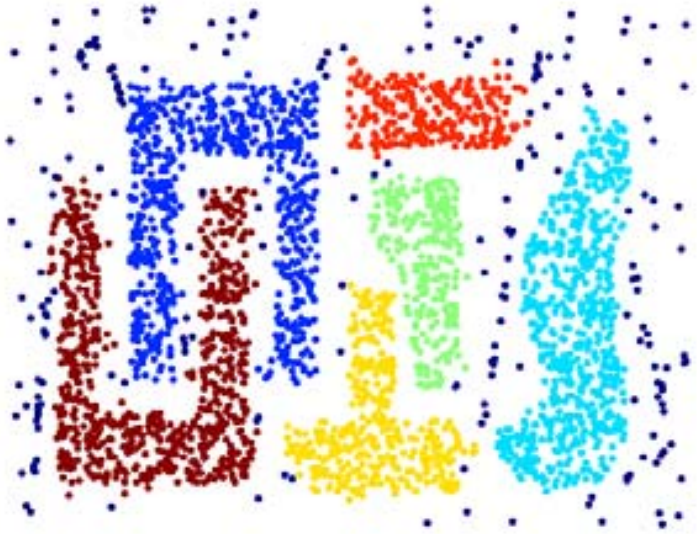


Point types: core, border and noise

Eps = 10, MinPts = 4

Quand DBSCAN marche bien

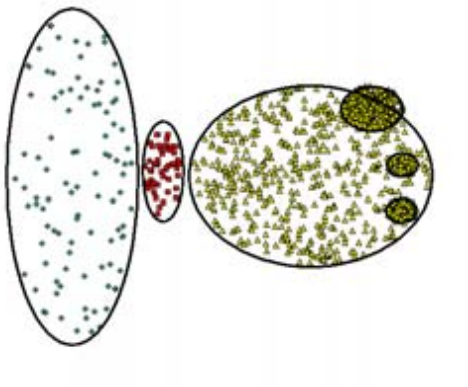
Original Points



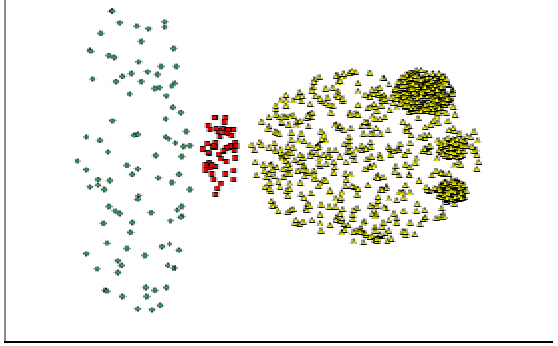
Clusters

- Résistant au bruit
- Peut gérer les clusters de différentes formes et tailles

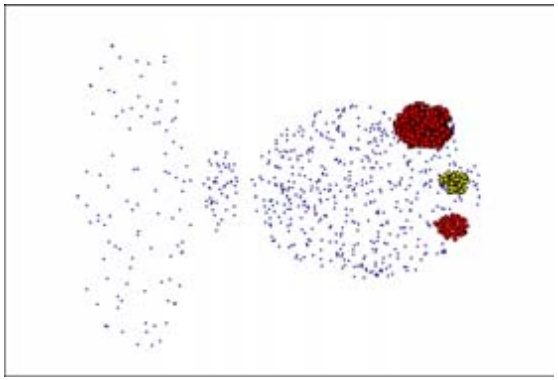
Quand DBSCAN ne marche PAS bien



Original Points



(MinPts=4, Eps=9.75).



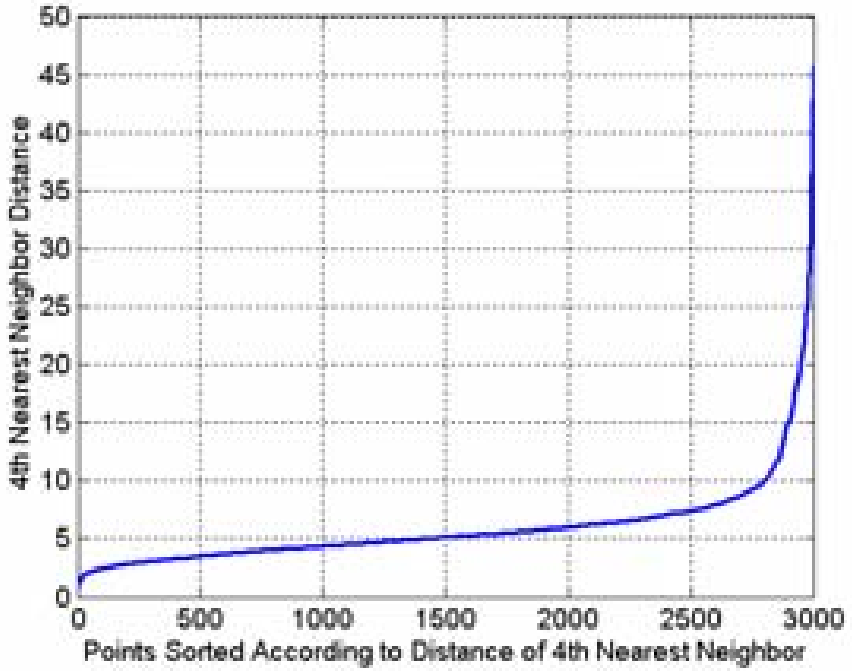
(MinPts=4, Eps=9.92)

- **Varying densities**
- **High-dimensional data**

- Densités variables
- Données en haute dimension

DBSCAN : déterminer Eps et MinPts

- L'idée est que pour les points d'un cluster, leur $k^{ième}$ voisin le plus proche est environ à la même distance
- Les aberrations ont un $k^{ième}$ voisin plus loin
- Donc, on affiche la distance de tous les points à leur $k^{ième}$ voisin

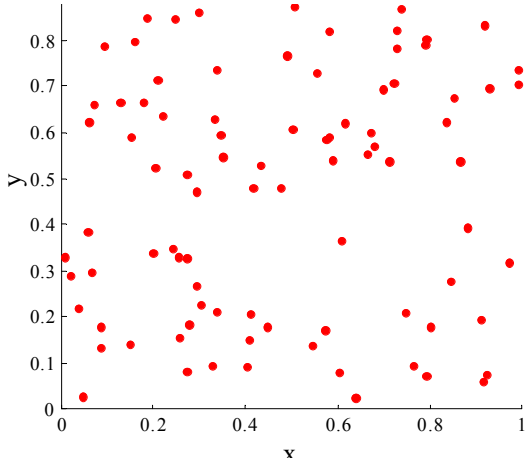


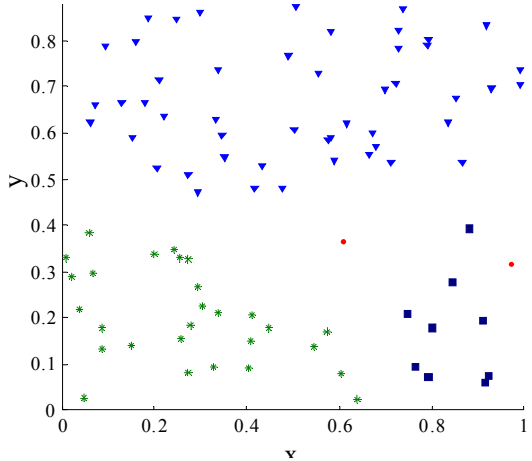
Validité d'un cluster

- Pour de la classification supervisée, on a plusieurs mesures pour vérifier la qualité d'un modèle
  - Précision, recall …
- Pour l'analyse de clusters, comment évaluer la qualité d'un clustering ?
  - Éviter de trouver des clusters dans du bruit
  - Comparer des algorithmes de clustering
  - Comparer 2 ensembles de clusters
  - Comparer 2 clusters

Clusters trouvés dans des données aléatoires

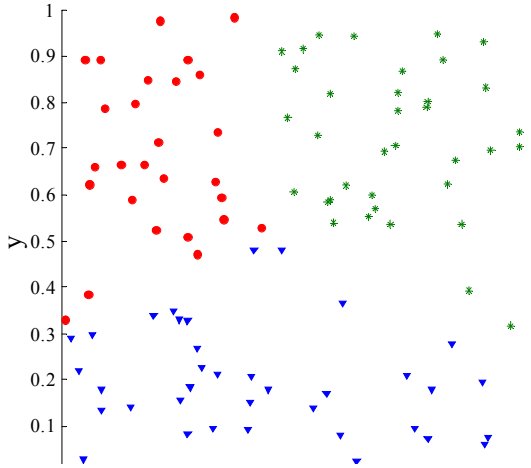
Random Points

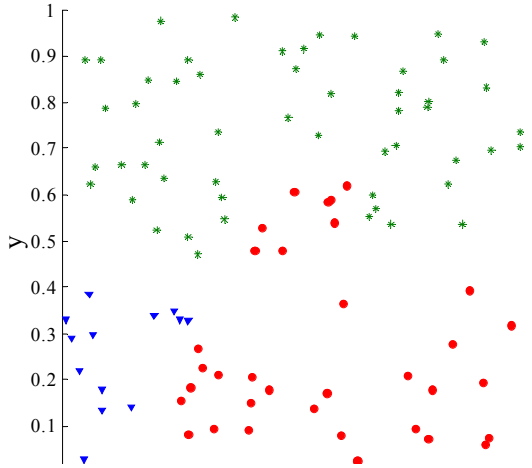




DBSCAN

K-means

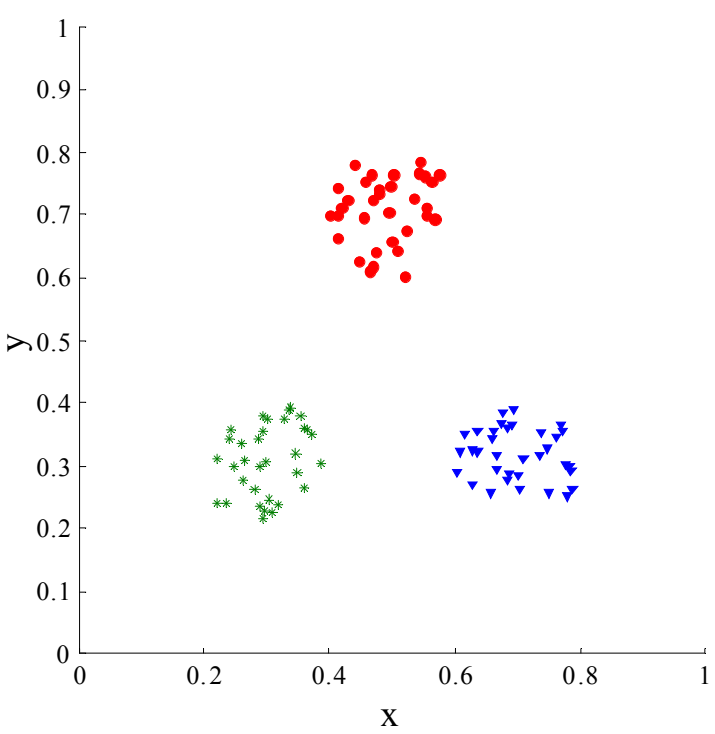


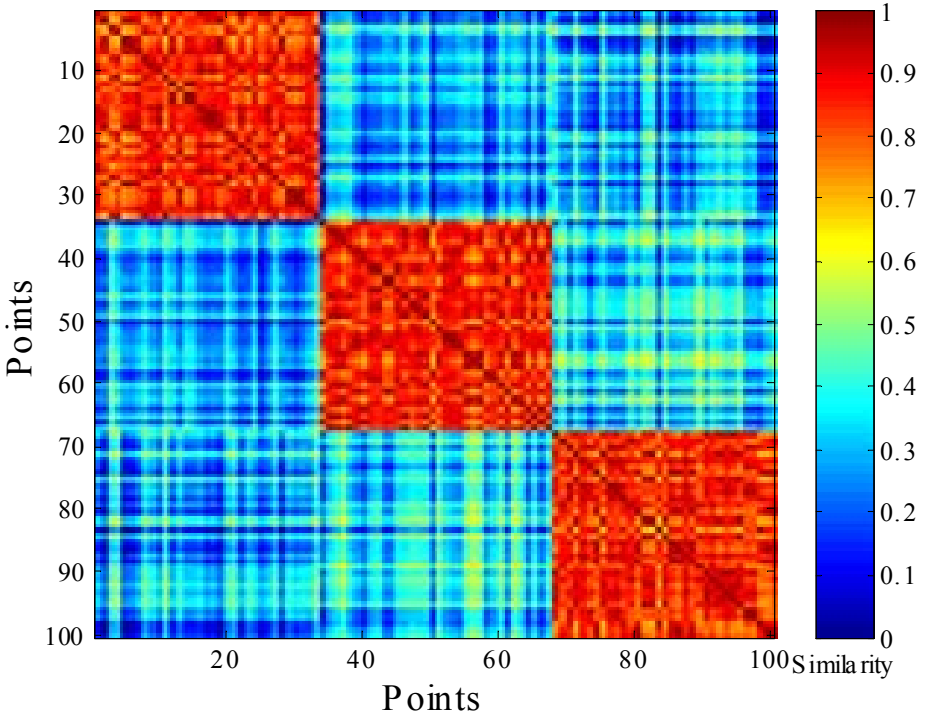


Complete Link

Utiliser la matrice de similarité pour valider les clusters

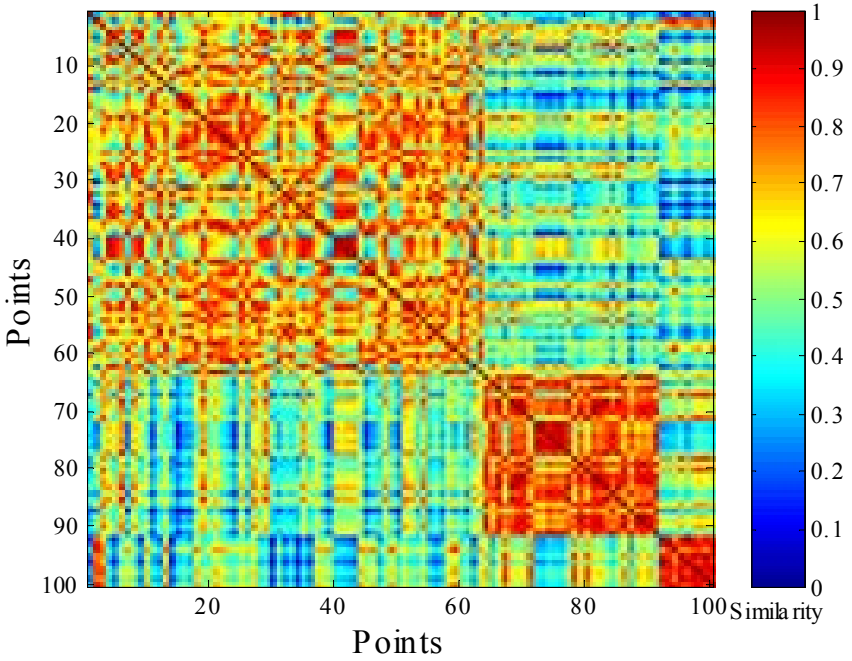
- Trier les données suivant les clusters et valider visuellement

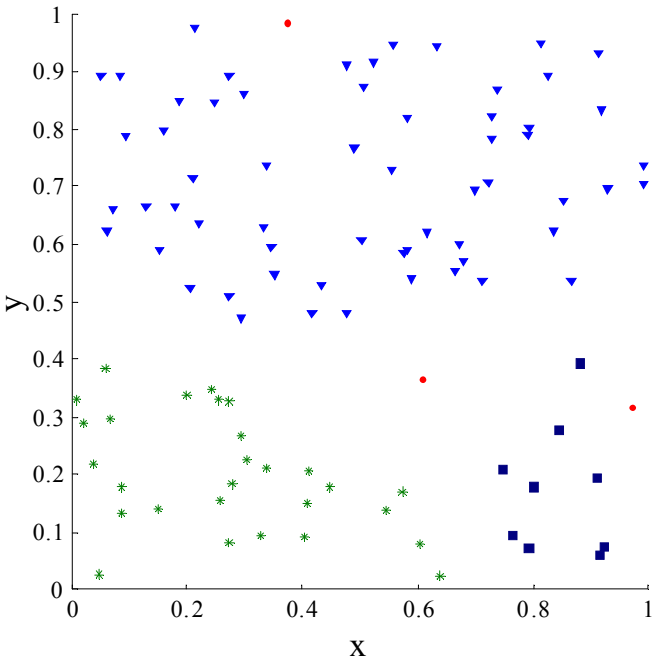




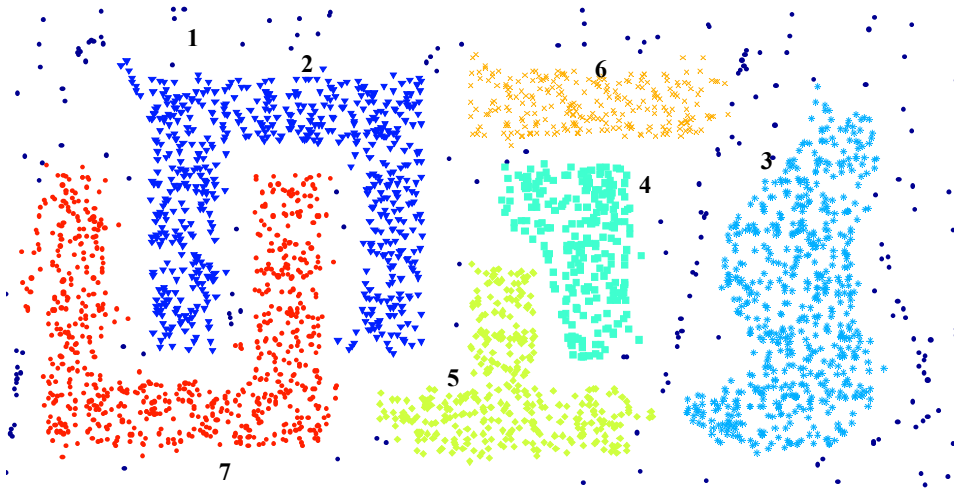
Utiliser la matrice de similarité pour valider les clusters

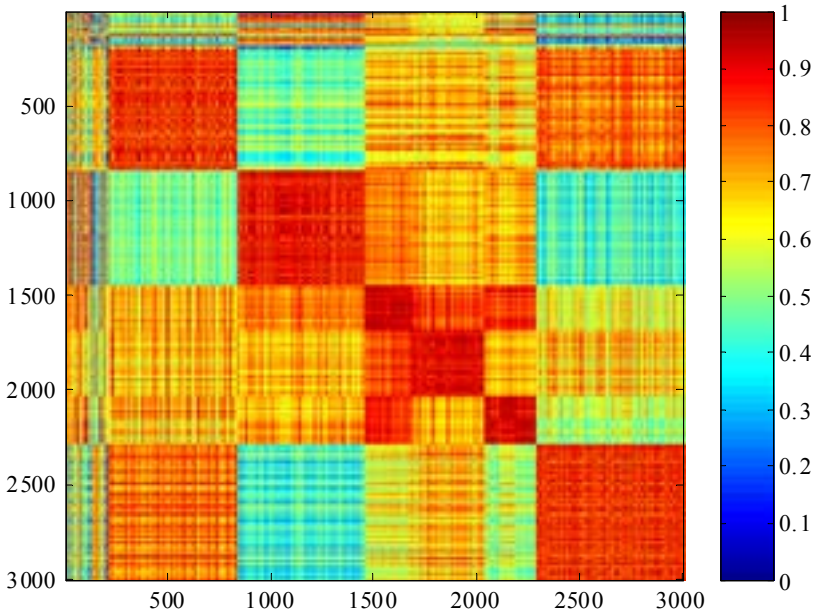
- Les clusters dans des données aléatoires ne sont pas nets





Utiliser la matrice de similarité pour valider les clusters





Commentaire final sur la validation de clusters

"The validation of clustering structures is the most difficult and frustrating part of cluster analysis.

Without a strong effort in this direction, cluster analysis will remain a black art accessible only to those true believers who have experience and great courage."

Algorithms for Clustering Data, Jain and Dubes