Chapitre 5:

Méthodes analytiques avancées

Clustering

Clustering - Définition

- Le clustering ou la classification automatique vise à regrouper un ensemble de données en groupes de sorte que :
  - Les objets d'un même groupe soient aussi similaires que possible
  - Les objets de groupes différents soient aussi dissemblables que possible

$$I_{intraclass} = \frac{1}{N} \sum_{i=1}^{k} \sum_{j=1}^{n_i} d^2(O_{ij}, G_i)$$

$$I_{interclass} = \sum_{i=1}^{k} \frac{n_i}{N} d^2(G_i, G)$$

  - Minimiser l'inertie intra-classes.
  - Maximiser l'inertie inter-classes.

Similarité VS Dissimilarité

- L'indice de similarité
  - $\forall x \in \Omega, s(x, x) = 1$
  - s est symétrique : $s(x1, x2) = s(x2, x1)$
  - $s(x1, x1) = s(x2, x2) > s(x1, x2)$

- L'indice de dissimilarité:
  - $\forall x \in \Omega, d(x, x) = 0$
  - d est symétrique: $d(x1, x2) = d(x2, x1)$
  - $d(x1, x1) = d(x2, x2) < d(x1, x2)$

Distance

- La distance : est un indice de dissimilarité qui vérifie :
- $d(x1, x2) = 0$ if $x1 = x2$
- La symétrie: $d(x1, x2) = d(x2, x1)$
- Pour tout x1, x2, x3 de Ω,
- $d(x1, x2) \leq d(x1, x3) + d(x3, x2)$ -- inégalité triangulaire

Exemples de mesures de distance

- Distance Euclidienne

$$d(X,Y) = \sqrt{\sum_{i=1}^{p}(x_i - y_i)^2}$$

- Distance Euclidienne Carrée

$$d(X,Y) = \sum_{i=1}^{p}(x_i - y_i)^2$$

- Distance de Chebychev

$$d(X,Y) = \max_{i=\{1,\ldots,p\}} |x_i - y_i|$$

Exemples de mesures de distance

- Distance de Manhattan

$$d(X,Y) = \sum_{i=1}^{p} |x_i - y_i|$$

- Distance de Minkowski

$$d(X,Y) = \sqrt[\lambda]{\sum_{i=1}^{p} |x_i - y_i|^{\lambda}}$$

  - Pour $\lambda = 1$, nous avons la somme des valeurs absolues (Distance de Manhattan)
  - For $\lambda = 2$, Distance Euclidienne
  - For $\lambda \rightarrow +\infty$, Distance de Chebychev

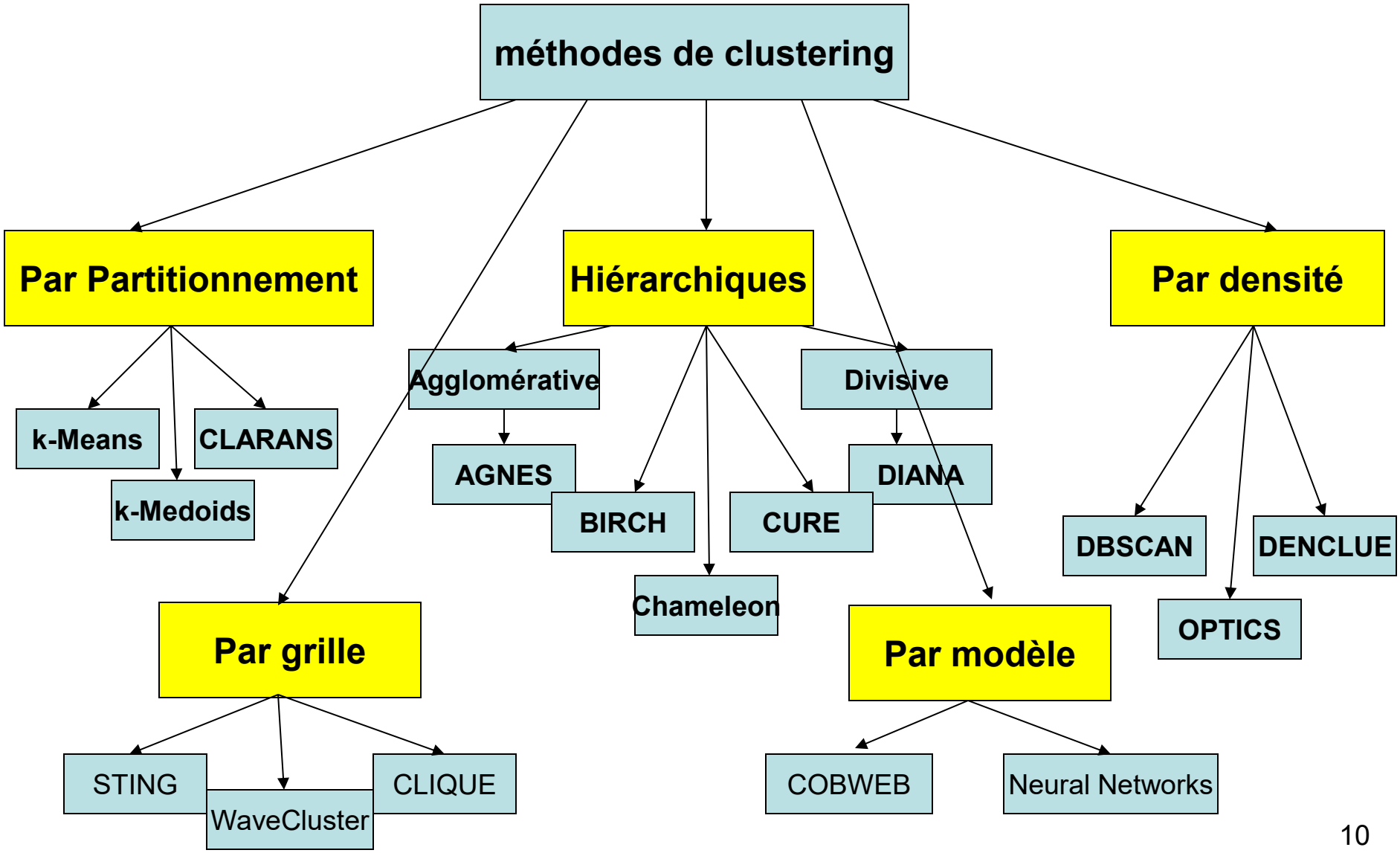
Distances (données binaires)

- On considère X et Y deux vecteurs binaires :
  - Soit a le nombre de fois où $X_j = Y_j = 1$
  - Soit b le nombre de fois où $X_j = 0$ et $Y_j = 1$
  - Soit c le nombre de fois où $X_j = 1$ et $Y_j = 0$
  - Soit d le nombre de fois où $X_j = Y_j = 0$
- Exemple de similarités souvent utilisées :
  - D1(X, Y) = a / (a+b+c+d)
  - D2(X, Y) = a / (a+b+c) *Mesure de Jaccard*
  - D3(X, Y) = 2a / (2a+b+c)
  - D4(X, Y) = a / (a+2(b+c))
  - D5(X, Y) = (a+d) / (a+b+c+d) *Mesure de Sokal & Michener*

Distances (données qualitatives)

- Similarités entre individus :
  - Codage disjonctif complet; permet de se ramener à un tableau de variables binaires (One-Hot Encoding)

Les méthodes de clustering



Principales méthodes de clustering

- Méthodes par partitionnement:
  - Construire k partitions et les corriger jusqu'à obtenir une similarité satisfaisante
    - k-means
    - k-medoids ou PAM (Partitionning Around Medoid)
    - CLARA (Clustering LARge Applications)
    - CLARANS (Clustering Large Applications based RANdomized Search)
- Méthodes hiérarchiques:
  - Créer une décomposition hiérarchique par agglomération ou division de groupes similaires ou dissimilaires
  - CAH, AGNES, DIANA, BIRCH, CURE, ROCK, …

Principales méthodes de clustering

- Méthodes par densité:
  - Grouper les objets tant que la densité de voisinage excède une certaine limite
  - DBSCAN, OPTICS, DENCLUE
- Méthodes par grille:
  - Diviser l'espace en cellules formant une grille multi-niveaux et grouper les cellules voisines en terme de distance
  - STING, WaveCluster, CLIQUE
- Méthodes par modèle:
  - Modéliser les groupes et utilise le modèle pour classer les points
  - COBWEB, Neural Networks

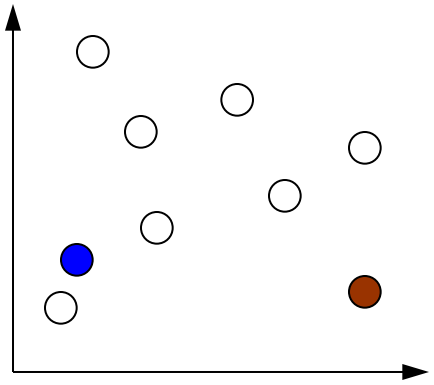
Principe du partitionnement

- ***N*** objets sont classés en k-partitions
  - Construire k partitions et les corriger jusqu'à obtenir une similarité satisfaisante
  - Optimisation d'une fonction d'objectif
    - similarité inter-classe
  - k-means, k-medoids (PAM)
  - CLARANS

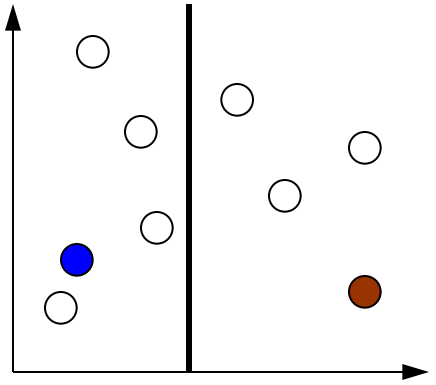
K-Means

- Méthode des K-moyennes (*MacQueen'67)*
  1. choisir K éléments initiaux "centres" des K groupes
  2. placer les objets dans le groupe de centre le plus proche
  3. recalculer le centre de gravité de chaque groupe
  4. itérer l'algorithme (répéter 2 et 3) jusqu'à ce que les objets ne changent plus de groupe
- Encore appelée méthode des centres mobiles
- C'est l'algorithme le plus utilisé

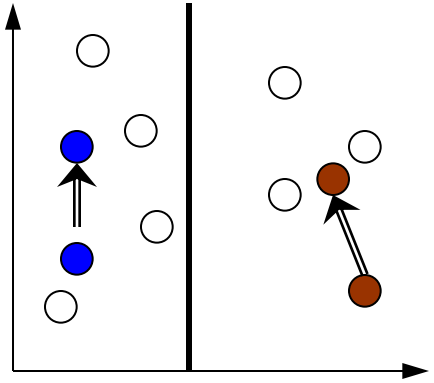
Exemple de K-means (k=2)



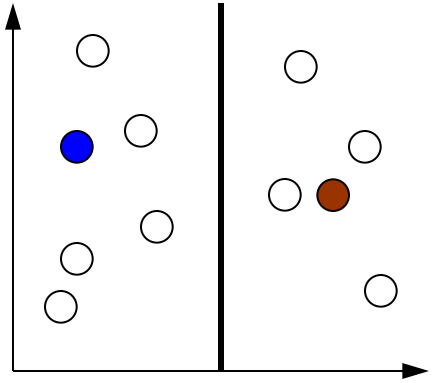
Choisir 2 centres



Assigner les objets

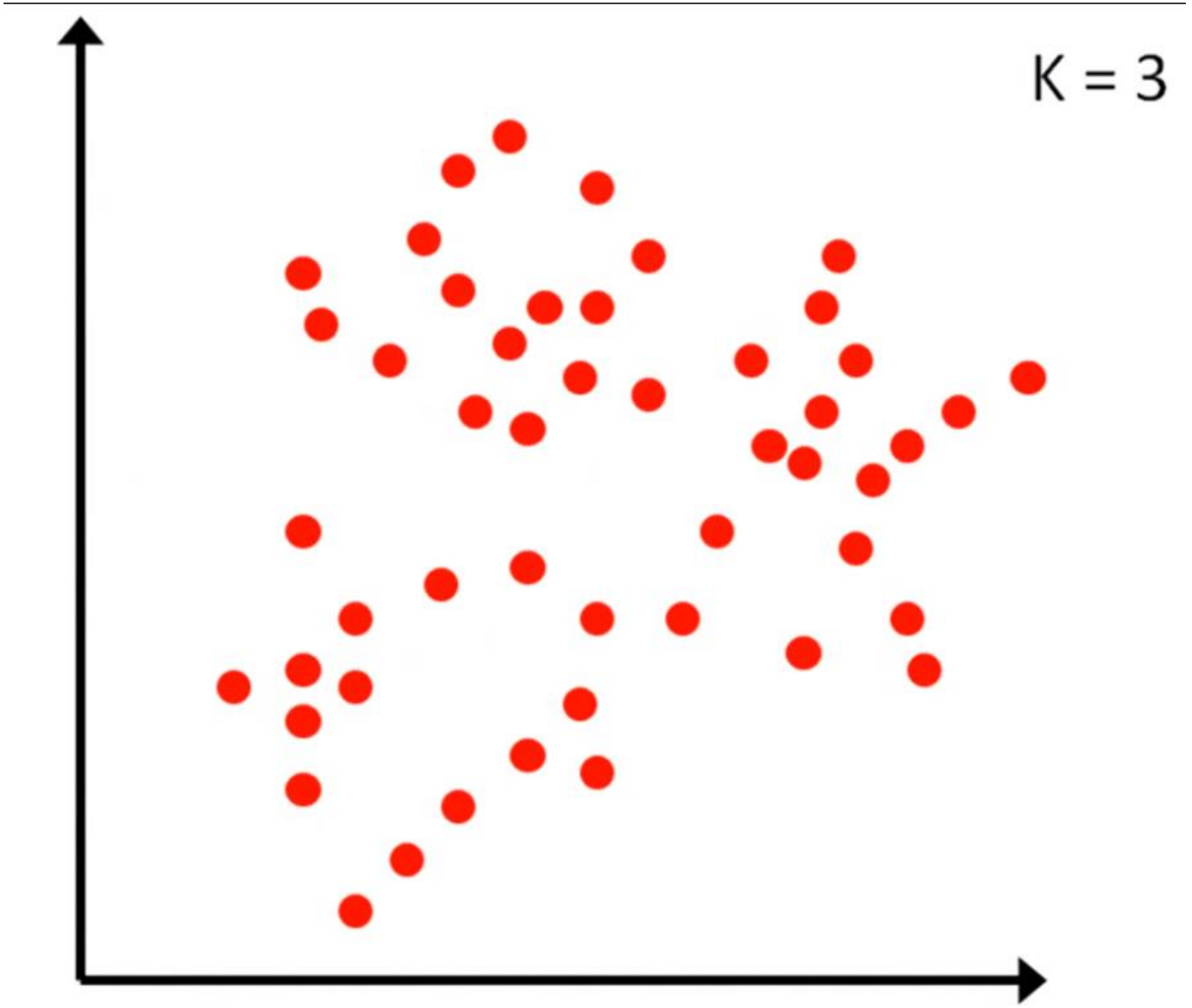


Recalculer les centres

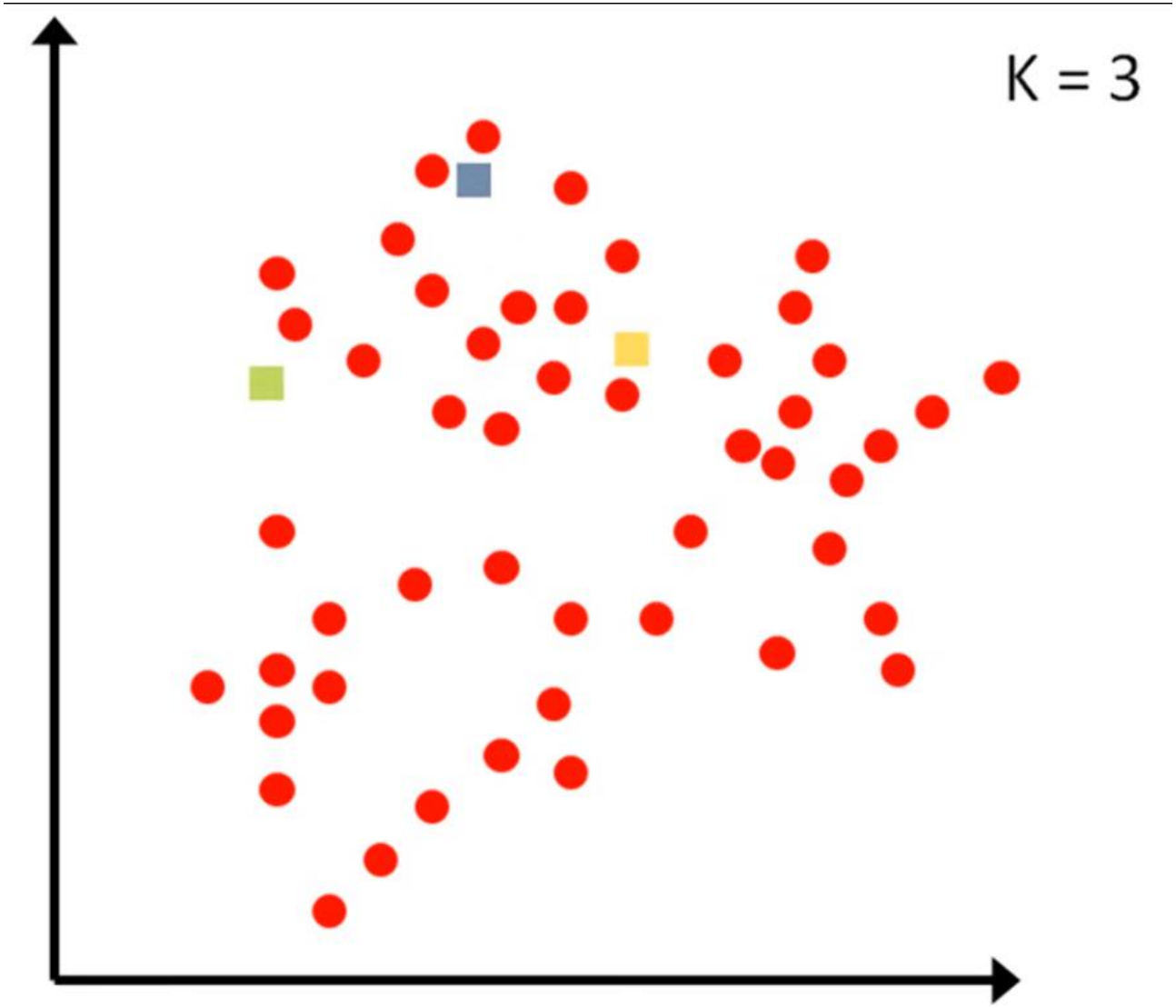


Réassigner les objets

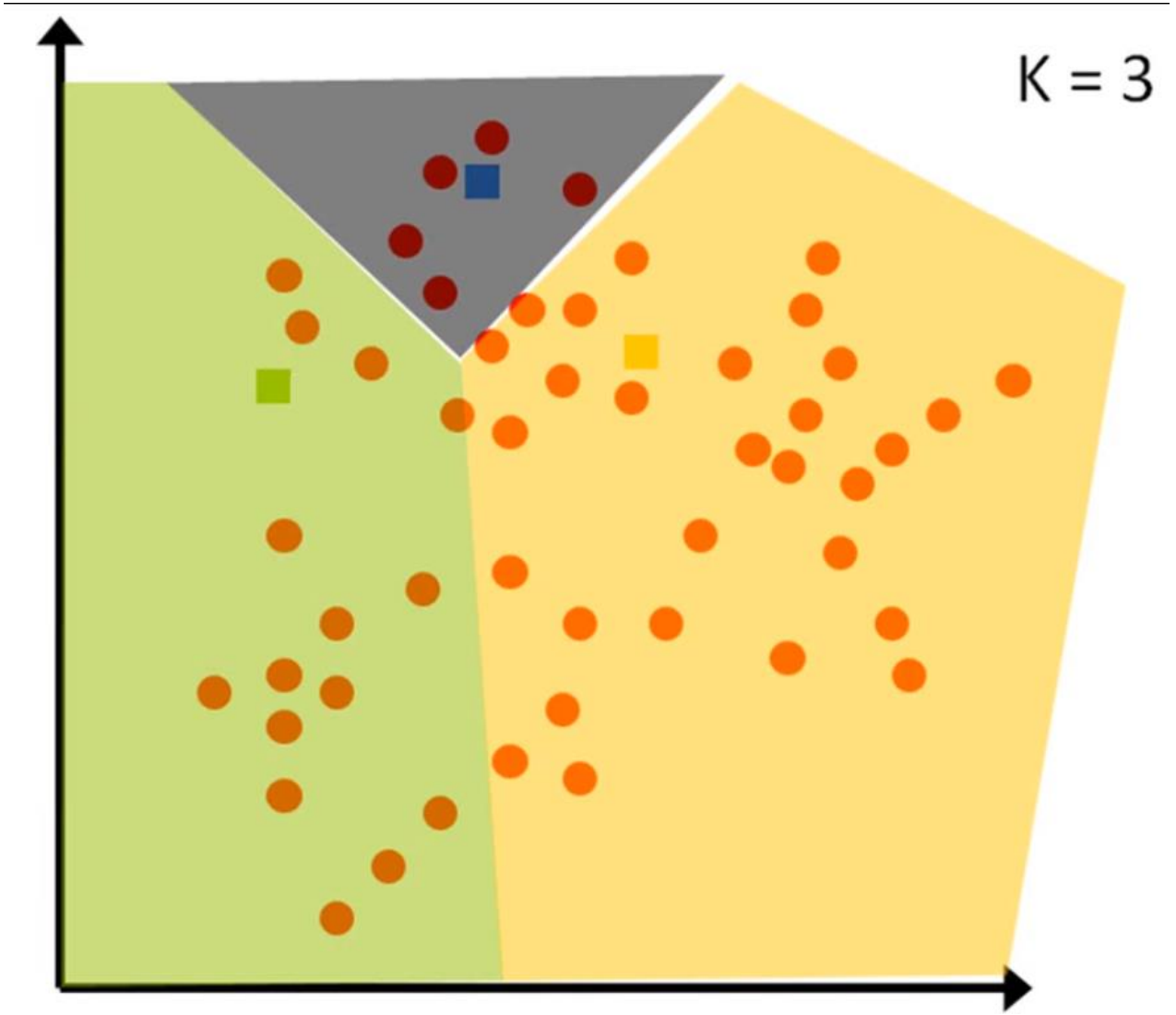
Exemple de K-means (k=3)



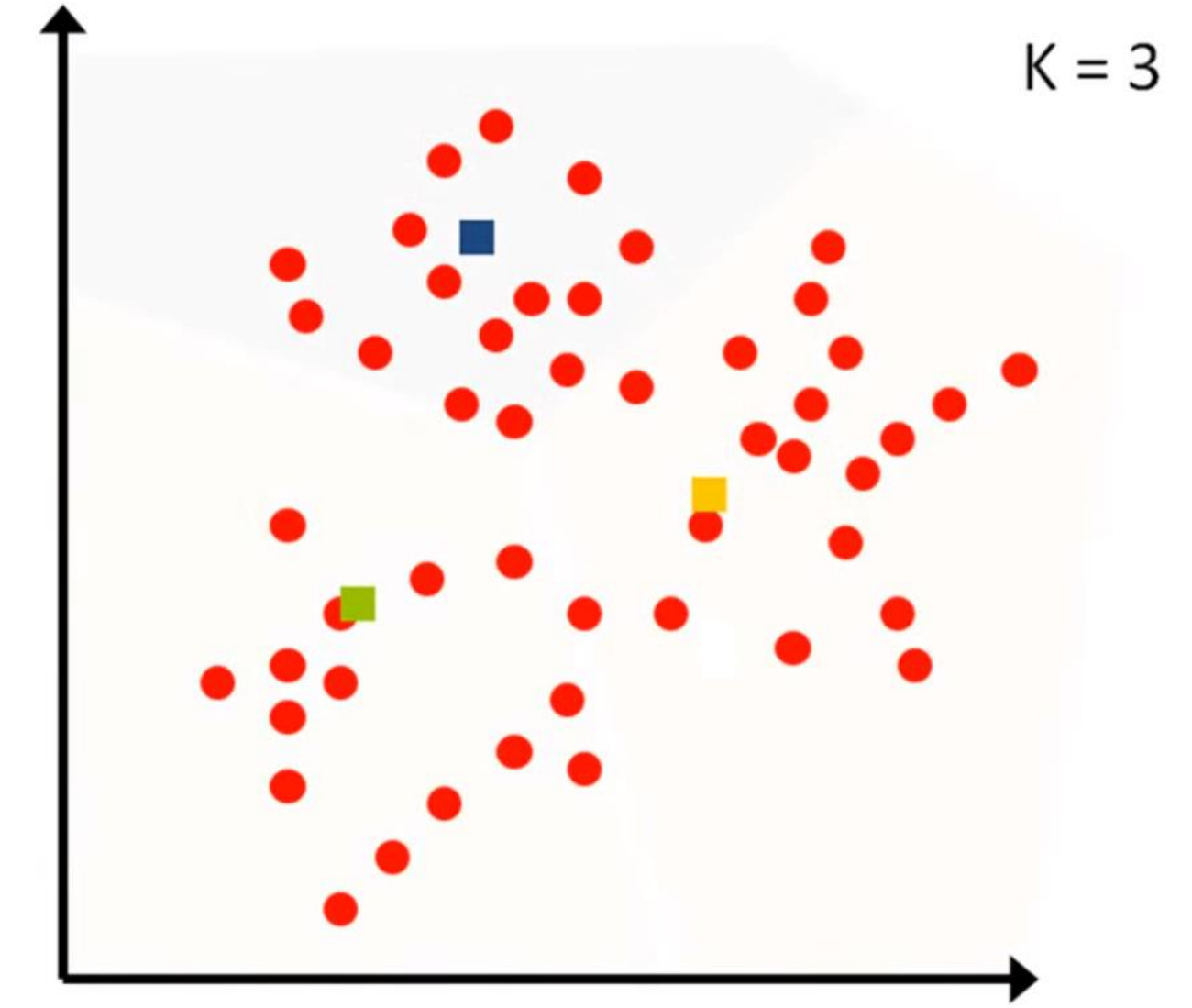
Exemple de K-means (k=3)



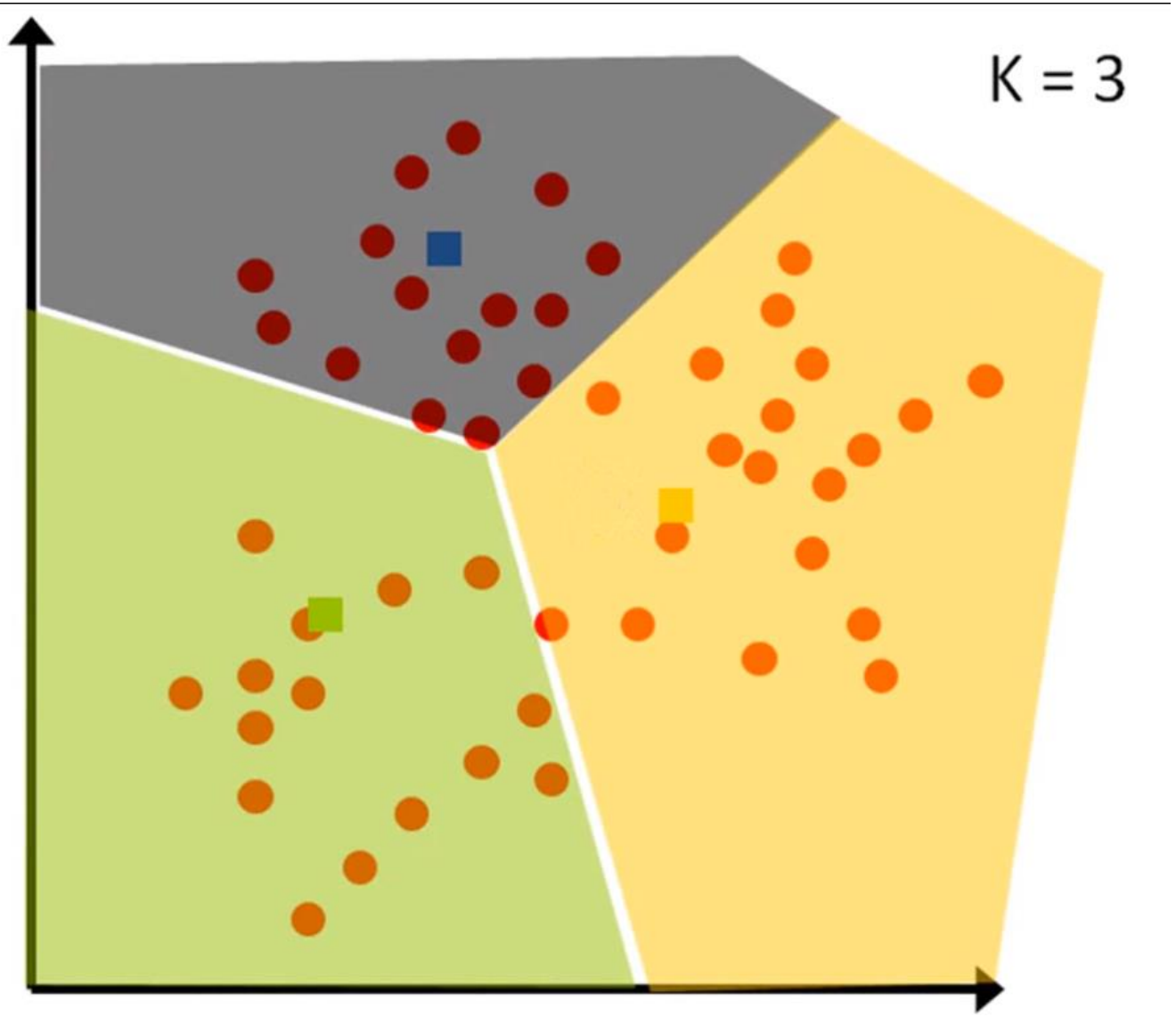
Exemple de K-means (k=3)



Exemple de K-means (k=3)



Exemple de K-means (k=3)



Faiblesse

- Mauvaise prise en compte des "outliers"
  - points extrêmes en dehors des groupes
  - faussent les moyennes et donc les centres
- Convergence plus ou moins rapide
- Amélioration:
  - utilisation de points centraux (médoïdes)

k-Médoïds ou PAM

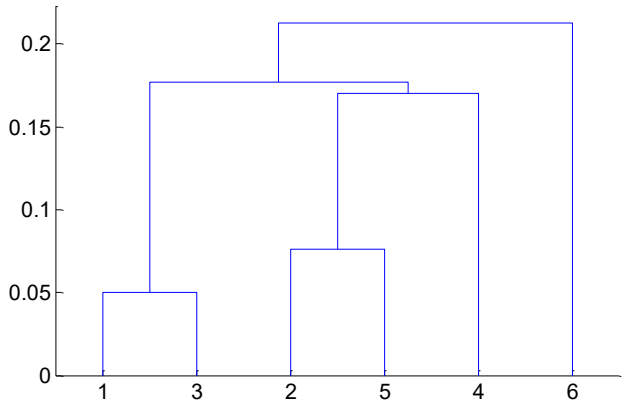
- Kaufman & Rousseeuw 1987
- Les centres sont des points effectifs
  - recherche de centres approchés des groupes
  - calculés par substitution aléatoire:
    - choix aléatoire d'un nouveau centre
    - calcul de la différence en distance des points
    - substitution si la différence est négative
  - essai de tous les couples (x,y) de chaque groupe
    - l'un est centre, l'autre non

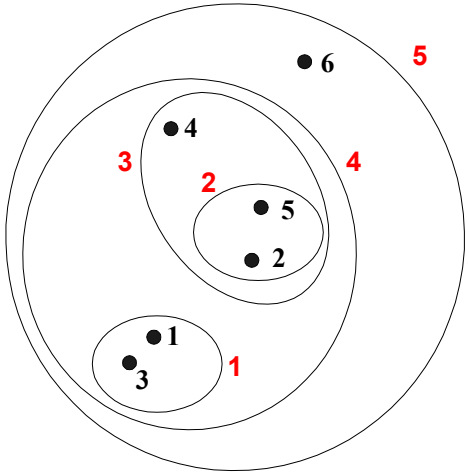
Forces et faiblesses

- Beaucoup plus coûteux que K-Means
  - Plus de calculs
- Plus robuste que k-means
  - Moins sensible aux "outliers"

La Classification Hiérarchique

- Produit un ensemble de groupes imbriqués organisés sous forme d'arbre hiérarchique.
- Impose une structure hiérarchique (par paires) à toutes les données.
- Souvent utile pour la visualisation : dendrogramme.
  - Un diagramme arborescent qui enregistre les séquences de fusions ou de divisions.





La Classification Hiérarchique

- Deux principaux types de classification hiérarchique :
  - Agglomérative (CAH: Classification Ascendante Hiérarchique)
    - Les points sont initialement considérés comme des clusters individuels.
    - À chaque étape, les clusters les plus proches sont fusionnés jusqu'à ce qu'il ne reste qu'un seul cluster (ou k clusters).
  - Divisive (CDH: Classification Descendante Hiérarchique)
    - Un seul cluster englobe tous les points.
    - À chaque étape, un cluster est divisé jusqu'à ce que chaque cluster contienne un point (ou qu'il y ait k clusters).

La Classification Hiérarchique

- Méthode de base (agglomérative) :
  1. Calculer toutes les distances entre toutes les paires.
  2. Combiner la paire d'individus les plus proches.
  3. Calculer la distance de cette paire à toutes les autres.
  4. Répéter à partir de l'étape 2 jusqu'à ce que toutes les paires soient combinées.

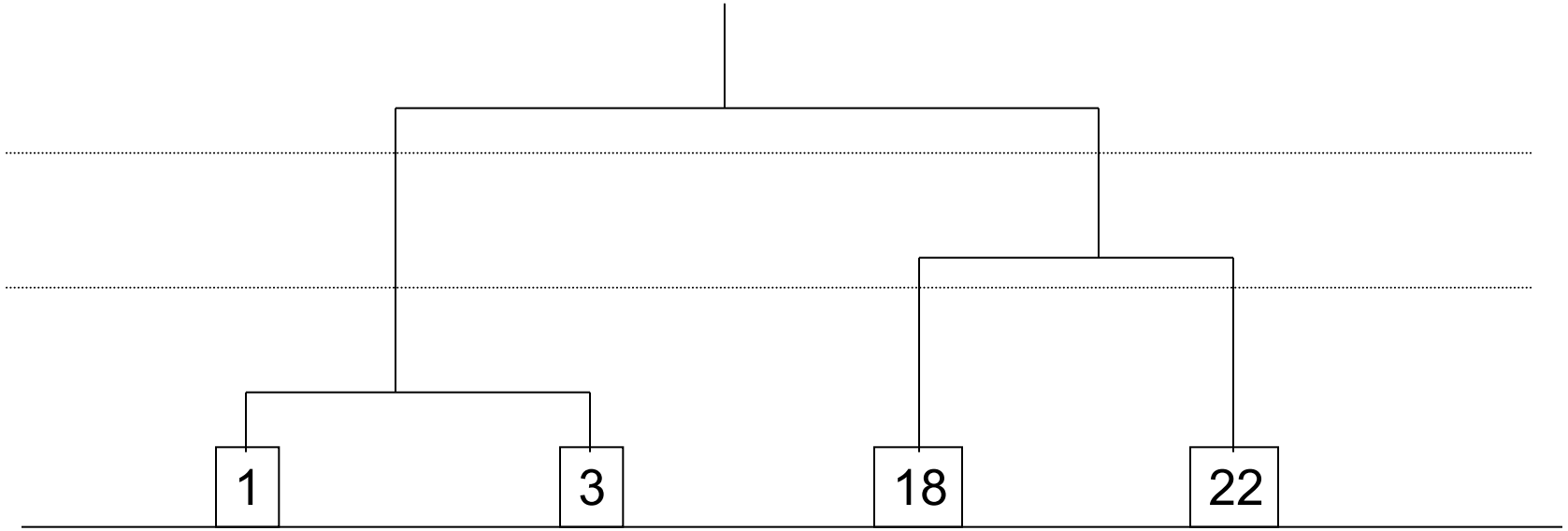
Agglomération hiérarchique - Principe

- Etapes :
  - Chaque individu représente un groupe
  - Trouver les deux groupes les plus proches
  - Grouper ces deux groupes en un nouveau groupe
  - Itérer jusqu'à obtenir tous les individus dans un seul groupe.

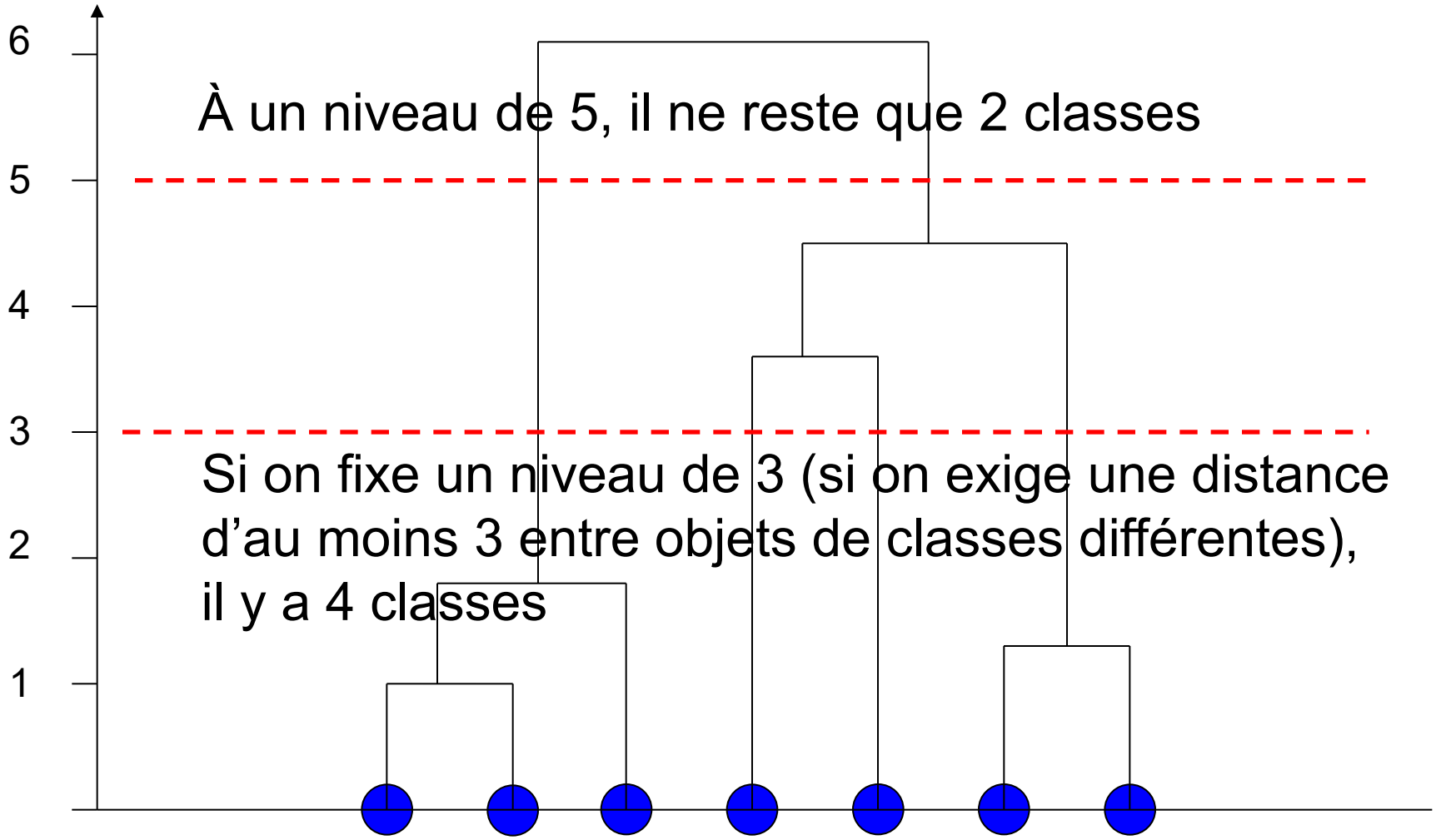
Agglomération - CAH

- Exemple de CAH (Classification Ascendante Hiérarchique)

- Dendrogramme

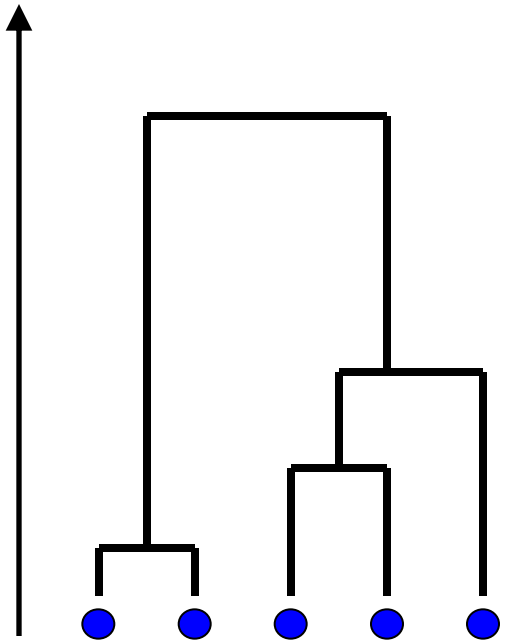


Exemple de dendrogramme

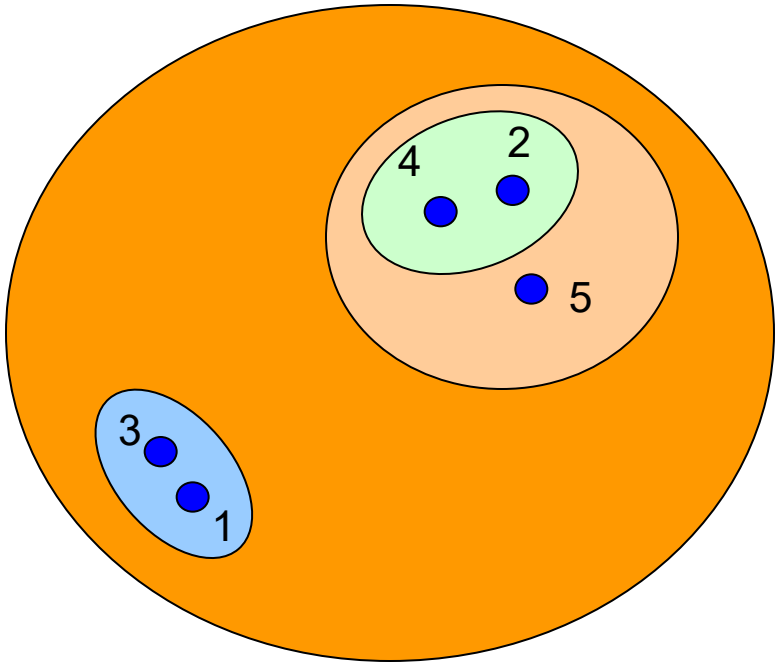


Exemple de dendrogramme

Distance entre les clusters joints



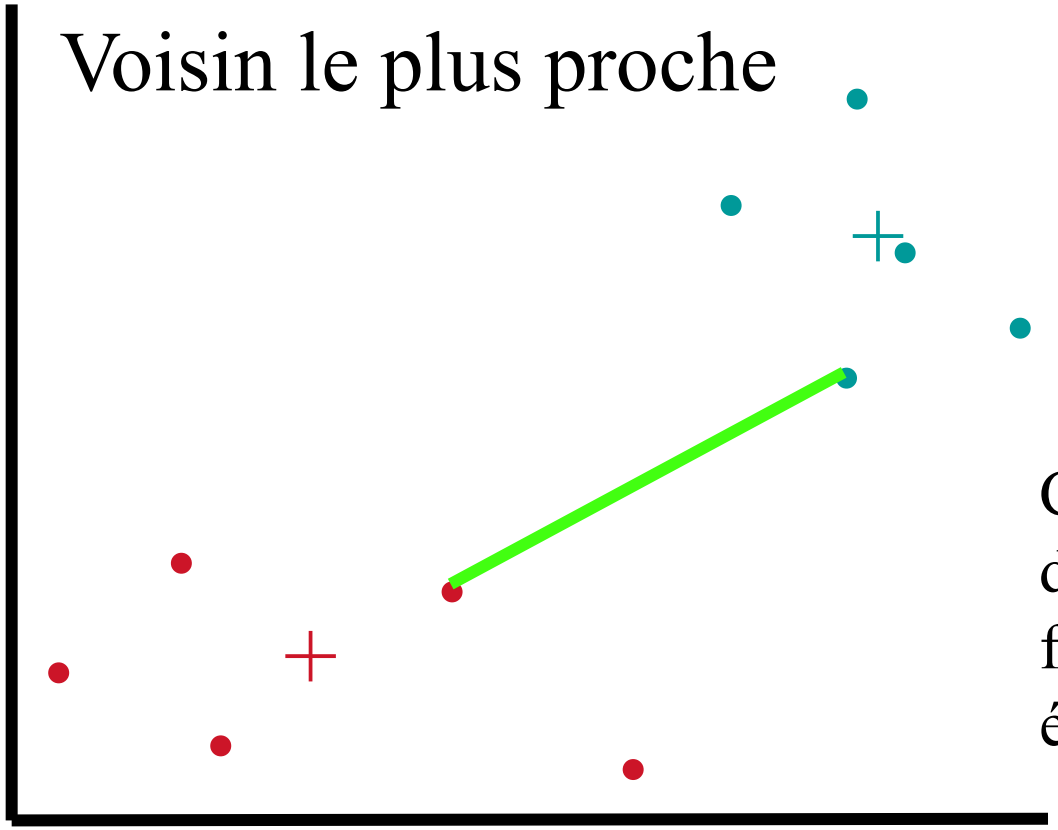
Dendrogramme



CAH: différentes méthodes d'agrégation

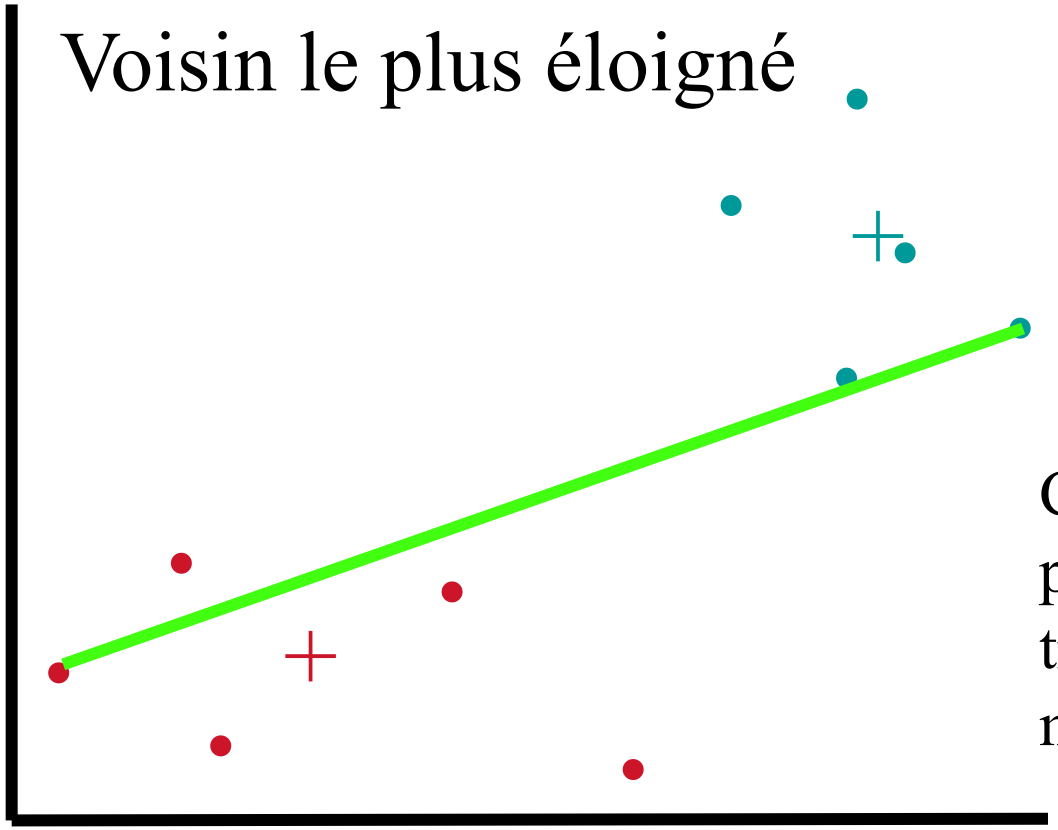
- Lien simple (single link)
- Lien complet (complete link)
- Lien moyen (average link)
- Lien centroïde (centroid link)

Lien simple (single link)



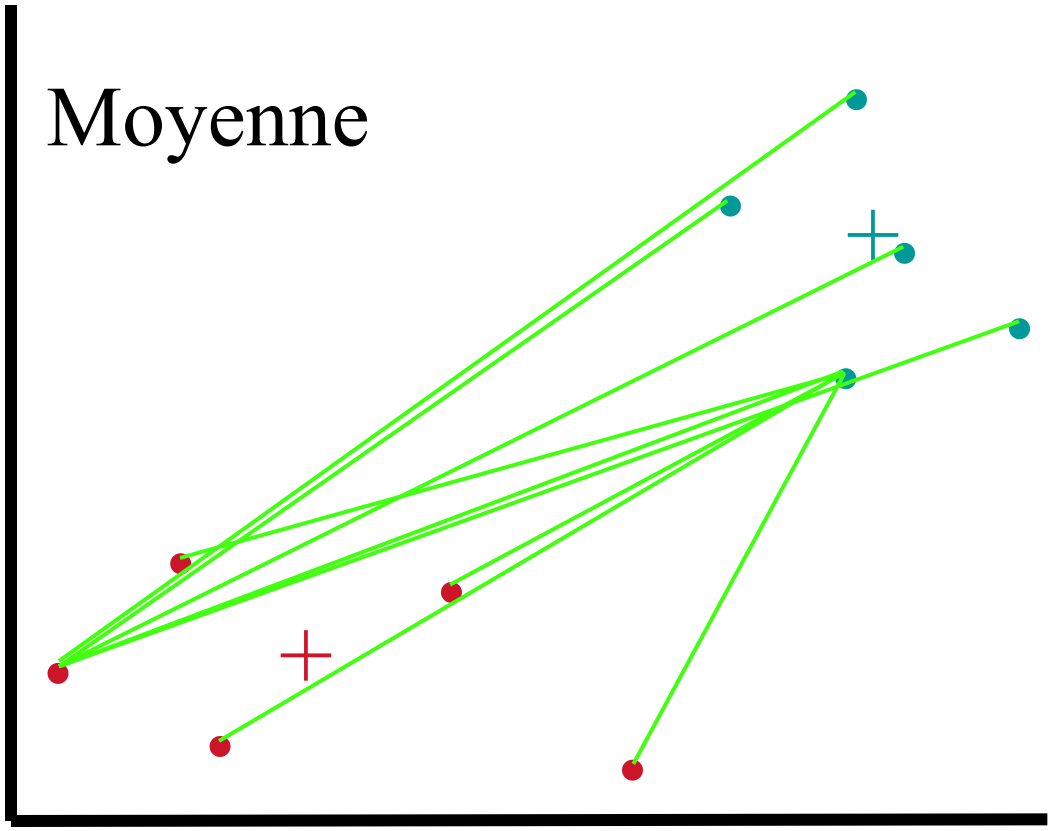
Cette méthode produit de longues chaînes qui forment des clusters épars.

Lien complet (complete link)

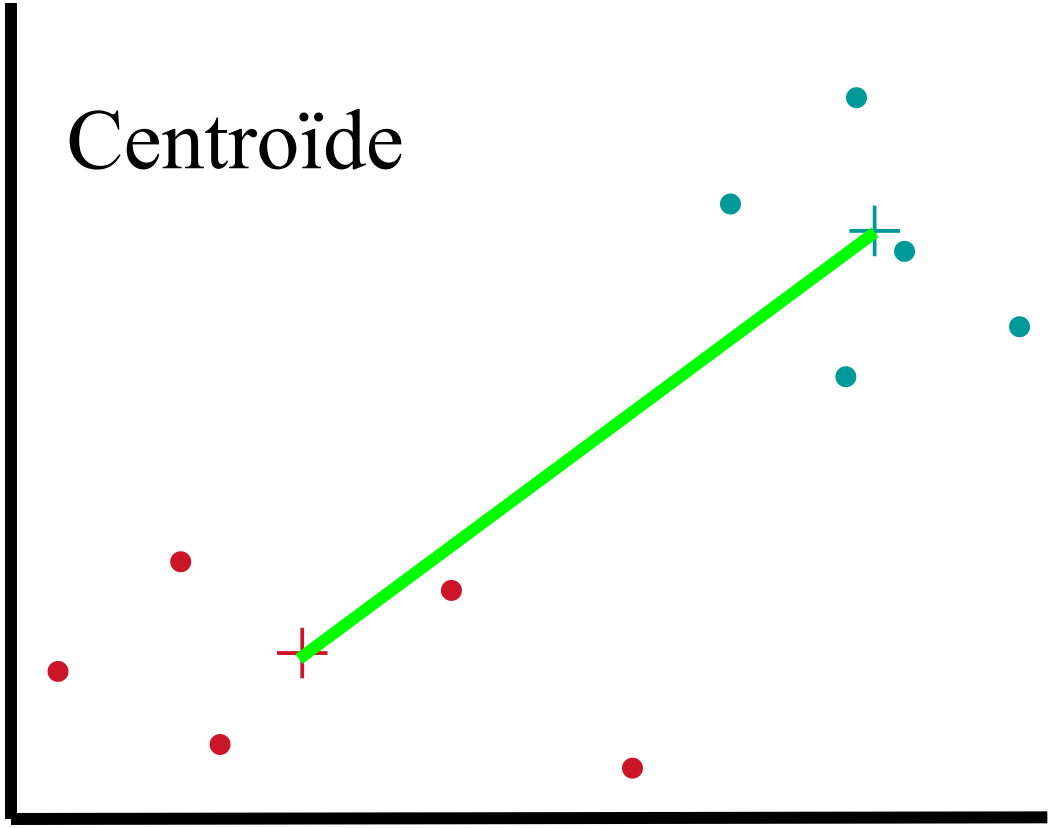


Cette méthode tend à produire des groupes très compacts de motifs similaires.

Lien moyen (average link)

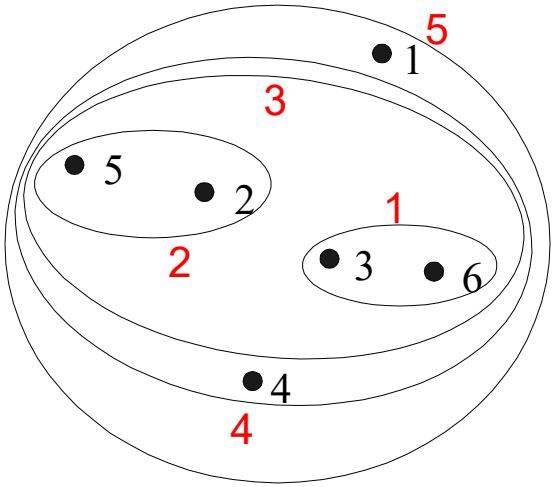


Lien centroïde (centroid link)



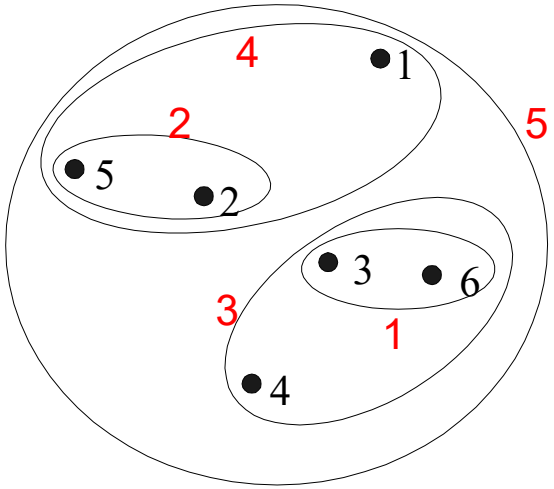
Les symboles « + » rouges et bleus indiquent les centroïdes des deux groupes.

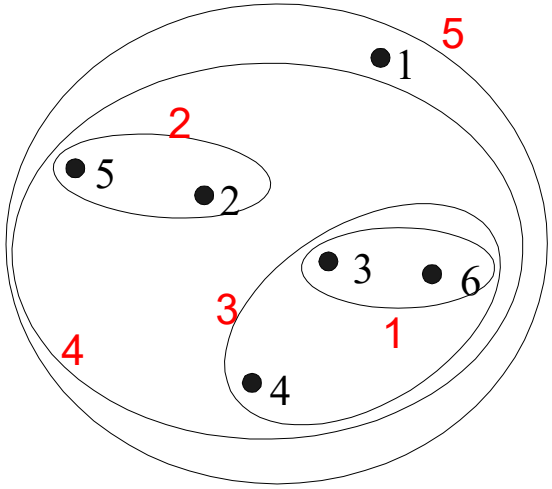
Classification hiérarchique : comparaison



MIN

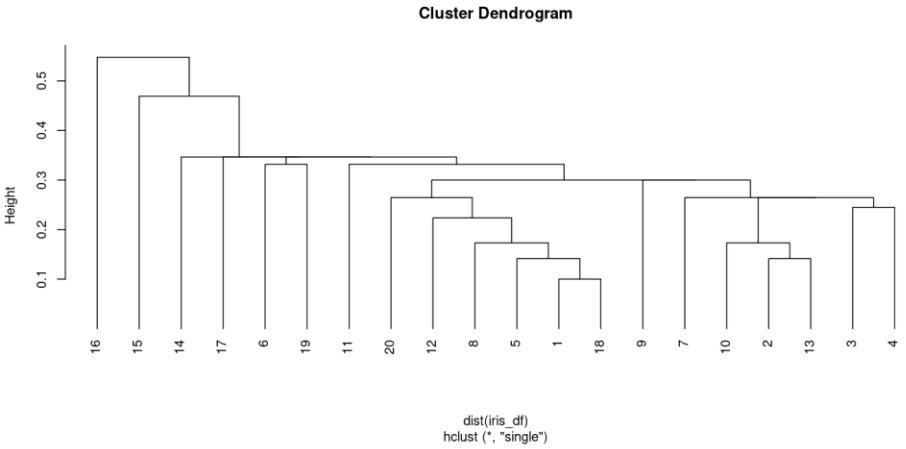
MAX

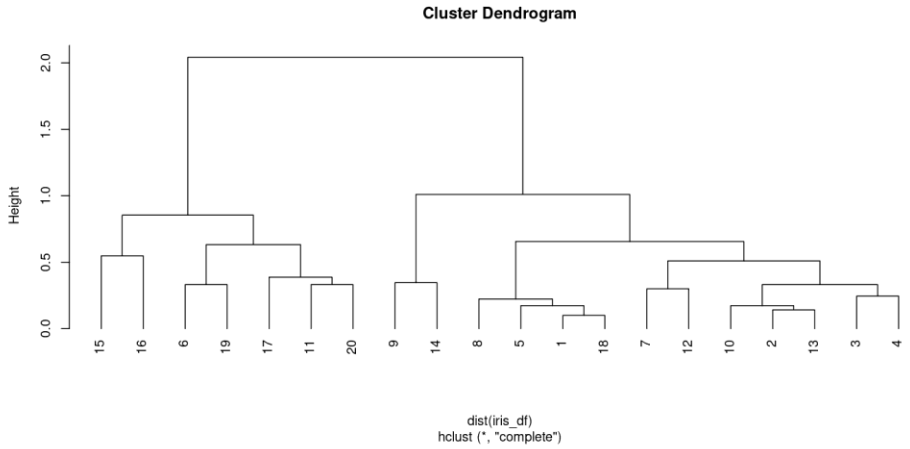


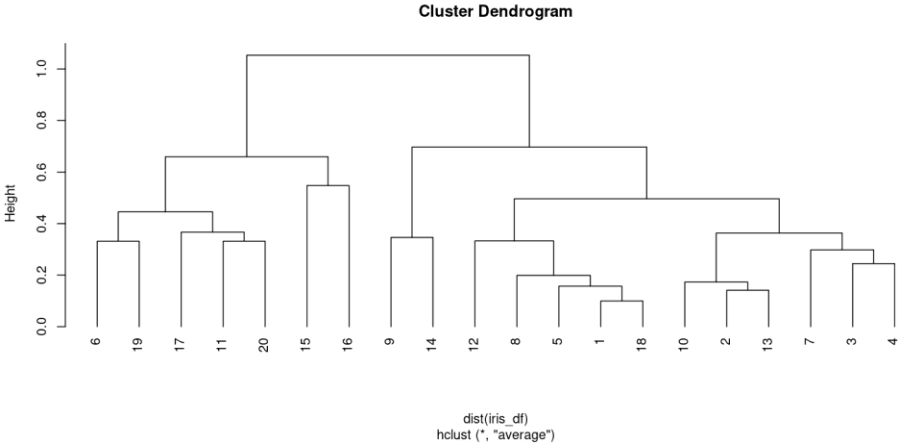


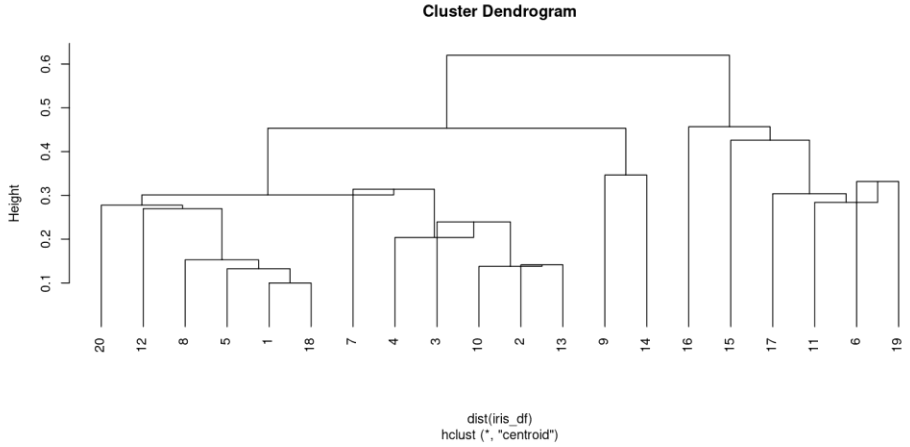
Average

Dendrogrammes avec différentes méthodes

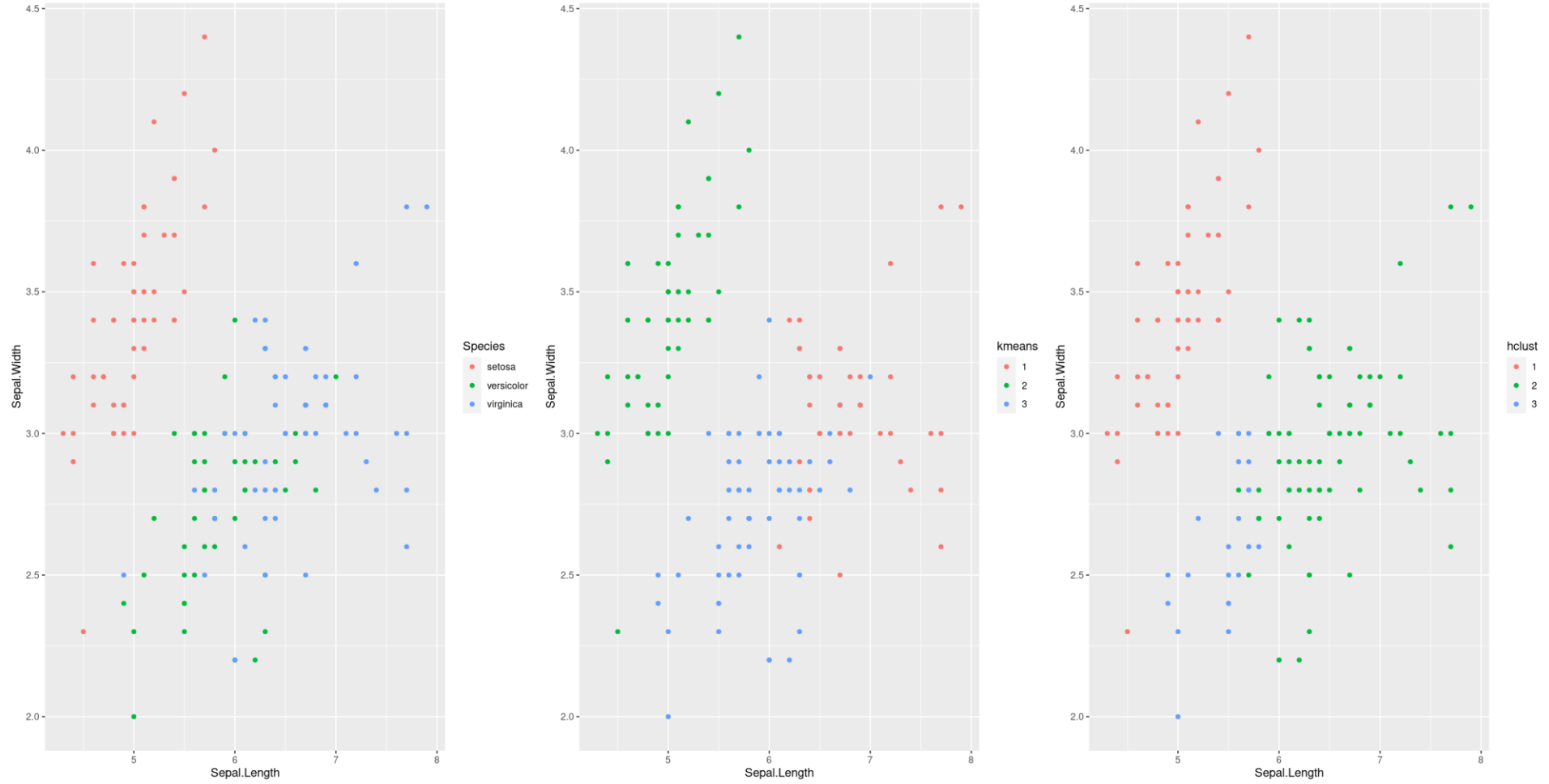




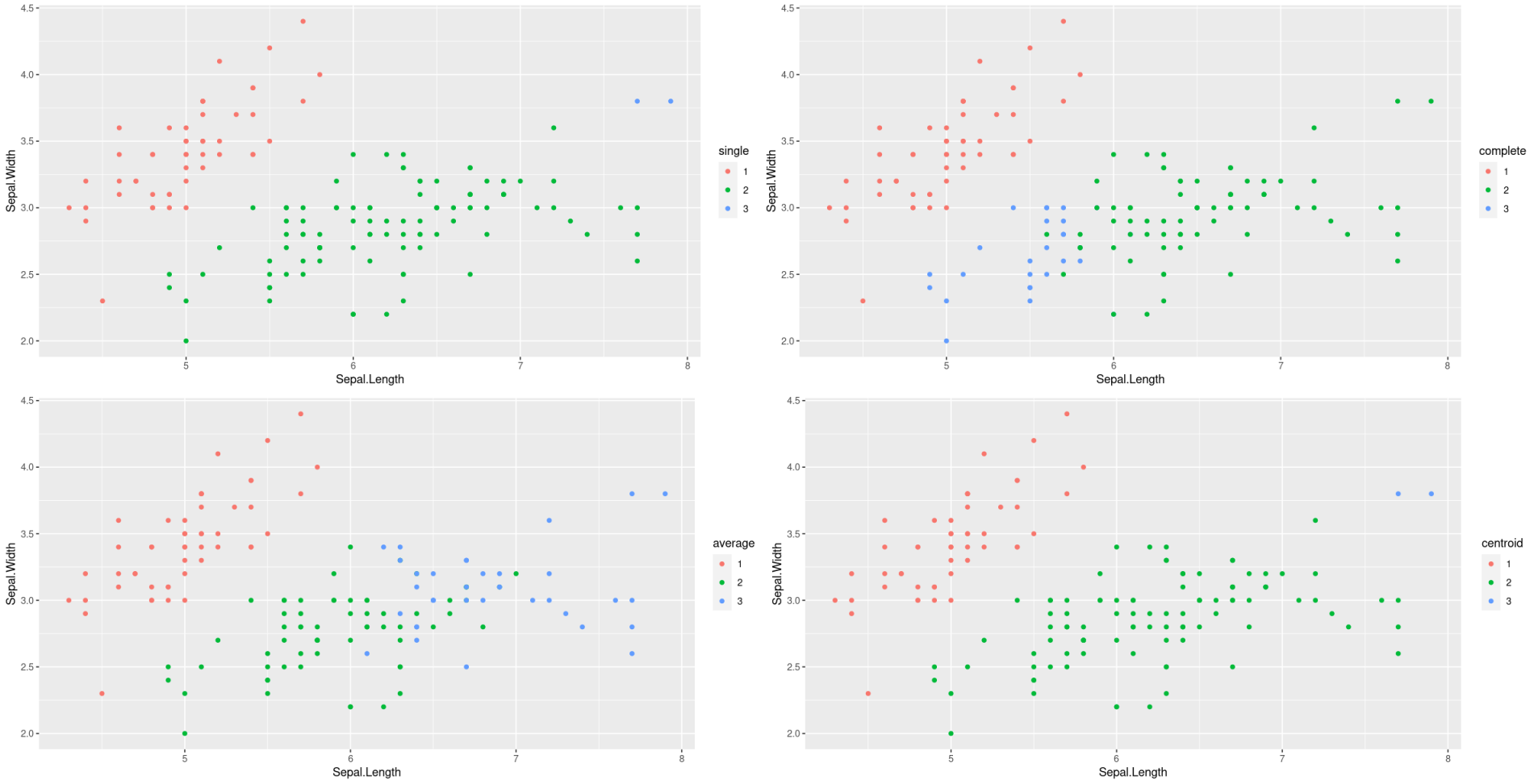




Reference vs KMeans vs Hclust (complete)

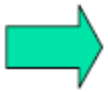


Single / Complete / Average / Centroid



Exemple: single link

$$\begin{array}{c|ccccc} & 1 & 2 & 3 & 4 & 5 \\ 1 & 0 & & & & \\ 2 & 2 & 0 & & & \\ 3 & 6 & 3 & 0 & & \\ 4 & 10 & 9 & 7 & 0 & \\ 5 & 9 & 8 & 5 & 4 & 0 \end{array}$$

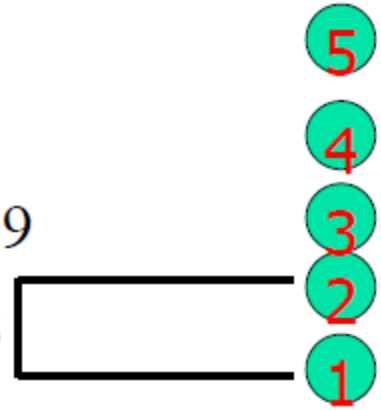


$$\begin{array}{c|cccc} & (1,2) & 3 & 4 & 5 \\ (1,2) & 0 & & & \\ 3 & 3 & 0 & & \\ 4 & 9 & 7 & 0 & \\ 5 & 8 & 5 & 4 & 0 \end{array}$$

$$d_{(1,2),3} = \min\{d_{1,3}, d_{2,3}\} = \min\{6,3\} = 3$$

$$d_{(1,2),4} = \min\{d_{1,4}, d_{2,4}\} = \min\{10,9\} = 9$$

$$d_{(1,2),5} = \min\{d_{1,5}, d_{2,5}\} = \min\{9,8\} = 8$$

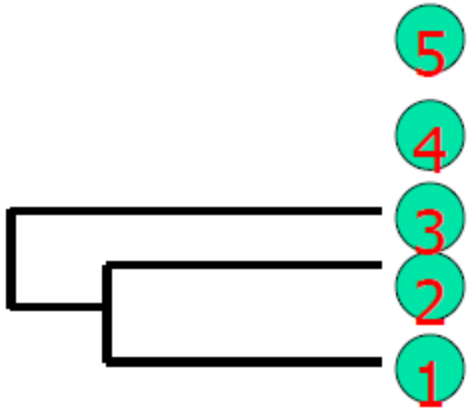


Exemple: single link

$$\begin{array}{c|ccccc} & 1 & 2 & 3 & 4 & 5 \\ 1 & 0 & & & & \\ 2 & 2 & 0 & & & \\ 3 & 6 & 3 & 0 & & \\ 4 & 10 & 9 & 7 & 0 & \\ 5 & 9 & 8 & 5 & 4 & 0 \end{array} \Rightarrow \begin{array}{c|cccc} & (1,2) & 3 & 4 & 5 \\ (1,2) & 0 & & & \\ 3 & 3 & 0 & & \\ 4 & 9 & 7 & 0 & \\ 5 & 8 & 5 & 4 & 0 \end{array} \Rightarrow \begin{array}{c|ccc} & (1,2,3) & 4 & 5 \\ (1,2,3) & 0 & & \\ 4 & 7 & 0 & \\ 5 & 5 & 4 & 0 \end{array}$$

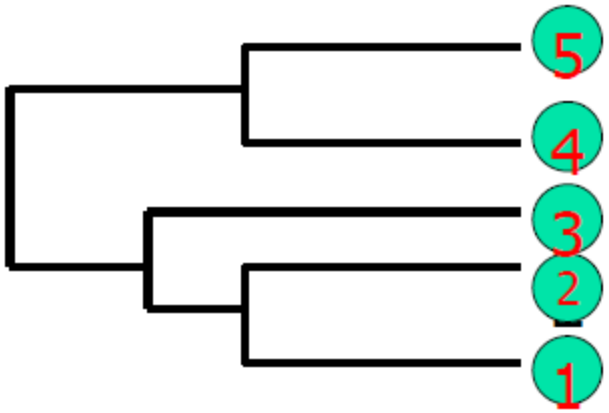
$$d_{(1,2,3),4} = \min\{d_{(1,2),4}, d_{3,4}\} = \min\{9,7\} = 7$$

$$d_{(1,2,3),5} = \min\{d_{(1,2),5}, d_{3,5}\} = \min\{8,5\} = 5$$



Exemple: single link

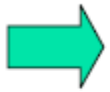
$$\begin{array}{c|ccccc} & 1 & 2 & 3 & 4 & 5 \\ 1 & 0 & & & & \\ 2 & 2 & 0 & & & \\ 3 & 6 & 3 & 0 & & \\ 4 & 10 & 9 & 7 & 0 & \\ 5 & 9 & 8 & 5 & 4 & 0 \end{array} \Rightarrow \begin{array}{c|cccc} & (1,2) & 3 & 4 & 5 \\ (1,2) & 0 & & & \\ 3 & 3 & 0 & & \\ 4 & 9 & 7 & 0 & \\ 5 & 8 & 5 & 4 & 0 \end{array} \Rightarrow \begin{array}{c|ccc} & (1,2,3) & 4 & 5 \\ (1,2,3) & 0 & & \\ 4 & 7 & 0 & \\ 5 & 5 & 4 & 0 \end{array}$$



$$d_{(1,2,3),(4,5)} = \min\{d_{(1,2,3),4}, d_{(1,2,3),5}\} = 5$$

Exemple: complete link

$$\begin{array}{c|ccccc} & 1 & 2 & 3 & 4 & 5 \\ 1 & 0 & & & & \\ 2 & 2 & 0 & & & \\ 3 & 6 & 3 & 0 & & \\ 4 & 10 & 9 & 7 & 0 & \\ 5 & 9 & 8 & 5 & 4 & 0 \end{array}$$

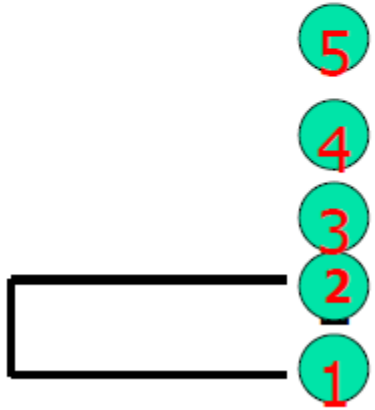


$$\begin{array}{c|cccc} & (1,2) & 3 & 4 & 5 \\ (1,2) & 0 & & & \\ 3 & 6 & 0 & & \\ 4 & 10 & 7 & 0 & \\ 5 & 9 & 5 & 4 & 0 \end{array}$$

$$d_{(1,2),3} = \max\{d_{1,3}, d_{2,3}\} = \max\{6,3\} = 6$$

$$d_{(1,2),4} = \max\{d_{1,4}, d_{2,4}\} = \max\{10,9\} = 10$$

$$d_{(1,2),5} = \max\{d_{1,5}, d_{2,5}\} = \max\{9,8\} = 9$$

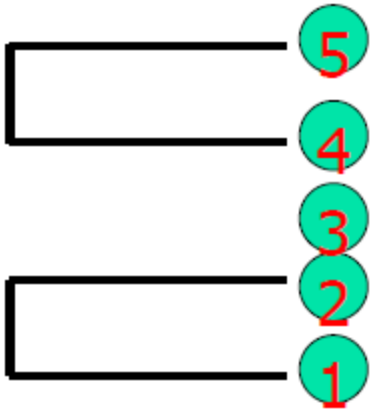


Exemple: complete link

$$\begin{array}{c|ccccc} & 1 & 2 & 3 & 4 & 5 \\ 1 & 0 & & & & \\ 2 & 2 & 0 & & & \\ 3 & 6 & 3 & 0 & & \\ 4 & 10 & 9 & 7 & 0 & \\ 5 & 9 & 8 & 5 & 4 & 0 \end{array} \Rightarrow \begin{array}{c|cccc} & (1,2) & 3 & 4 & 5 \\ (1,2) & 0 & & & \\ 3 & 6 & 0 & & \\ 4 & 10 & 7 & 0 & \\ 5 & 9 & 5 & 4 & 0 \end{array} \Rightarrow \begin{array}{c|ccc} & (1,2) & 3 & (4,5) \\ (1,2) & 0 & & \\ 3 & 6 & 0 & \\ (4,5) & 10 & 7 & 0 \end{array}$$

$$d_{(1,2),(4,5)} = \max\{d_{(1,2),4}, d_{(1,2),5}\} = \max\{10,9\} = 10$$

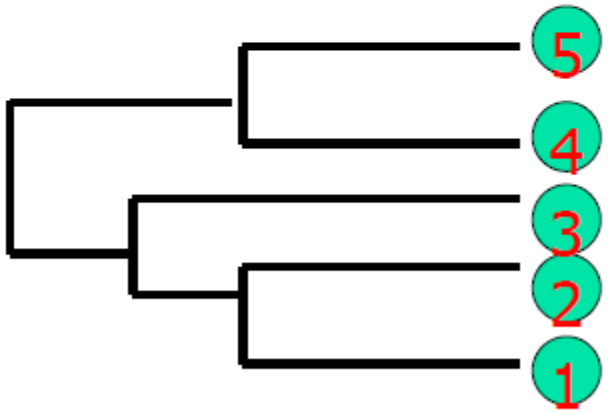
$$d_{3,(4,5)} = \max\{d_{3,4}, d_{3,5}\} = \max\{7,5\} = 7$$



Exemple: complete link

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array}\begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 0 & & & & \\ 2 & 0 & & & \\ 6 & 3 & 0 & & \\ 10 & 9 & 7 & 0 & \\ 9 & 8 & 5 & 4 & 0 \end{bmatrix} \end{array} \Rightarrow \begin{array}{c} \\ (1,2) \\ 3 \\ 4 \\ 5 \end{array}\begin{array}{c} \begin{array}{cccc} (1,2) & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 0 & & & \\ 6 & 0 & & \\ 10 & 7 & 0 & \\ 9 & 5 & 4 & 0 \end{bmatrix} \end{array} \Rightarrow \begin{array}{c} \\ (1,2) \\ 3 \\ (4,5) \end{array}\begin{array}{c} \begin{array}{ccc} (1,2) & 3 & (4,5) \end{array} \\ \begin{bmatrix} 0 & & \\ 6 & 0 & \\ 10 & 7 & 0 \end{bmatrix} \end{array}$$

$$d_{(1,2,3),(4,5)} = \max\{d_{(1,2),(4,5)}, d_{3,(4,5)}\} = 10$$



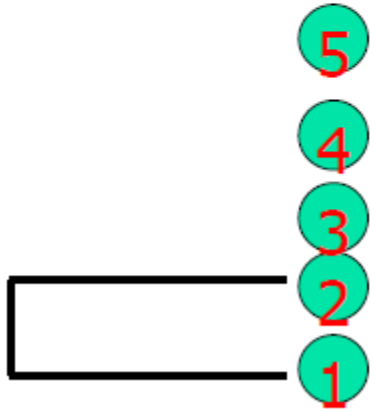
Exemple: average link

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array}\begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 0 & & & & \\ 2 & 0 & & & \\ 6 & 3 & 0 & & \\ 10 & 9 & 7 & 0 & \\ 9 & 8 & 5 & 4 & 0 \end{bmatrix} \end{array} \Rightarrow \begin{array}{c} \\ (1,2) \\ 3 \\ 4 \\ 5 \end{array}\begin{array}{c} \begin{array}{cccc} (1,2) & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 0 & & & \\ 4.5 & 0 & & \\ 9.5 & 7 & 0 & \\ 8.5 & 5 & 4 & 0 \end{bmatrix} \end{array}$$

$$d_{(1,2),3} = \frac{1}{2}(d_{1,3} + d_{2,3}) = \frac{6+3}{2} = 4.5$$

$$d_{(1,2),4} = \frac{1}{2}(d_{1,4} + d_{2,4}) = \frac{10+9}{2} = 9.5$$

$$d_{(1,2),5} = \frac{1}{2}(d_{1,5} + d_{2,5}) = \frac{9+8}{2} = 8.5$$

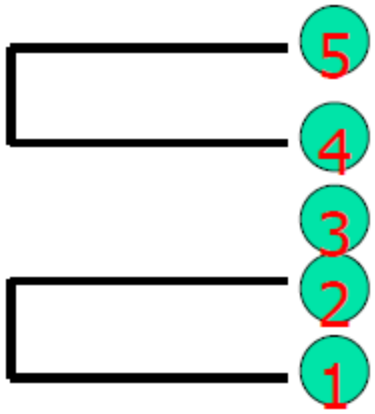


Exemple: average link

$$\begin{array}{c|ccccc} & 1 & 2 & 3 & 4 & 5 \\ 1 & 0 & & & & \\ 2 & 2 & 0 & & & \\ 3 & 6 & 3 & 0 & & \\ 4 & 10 & 9 & 7 & 0 & \\ 5 & 9 & 8 & 5 & 4 & 0 \end{array} \Rightarrow \begin{array}{c|cccc} & (1,2) & 3 & 4 & 5 \\ (1,2) & 0 & & & \\ 3 & 4.5 & 0 & & \\ 4 & 9.5 & 7 & 0 & \\ 5 & 8.5 & 5 & 4 & 0 \end{array} \Rightarrow \begin{array}{c|ccc} & (1,2) & 3 & (4,5) \\ (1,2) & 0 & & \\ 3 & 4.5 & 0 & \\ (4,5) & 9 & 6 & 0 \end{array}$$

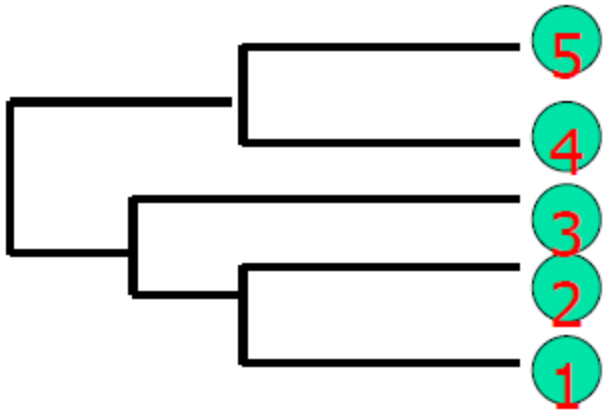
$$d_{(1,2),(4,5)} = \frac{1}{4}(d_{1,4} + d_{1,5} + d_{2,4} + d_{2,5}) = 9$$

$$d_{3,(4,5)} = \frac{1}{2}(d_{3,4} + d_{3,5}) = 6$$



Exemple: average link

$$\begin{array}{c|ccccc} & 1 & 2 & 3 & 4 & 5 \\ 1 & 0 & & & & \\ 2 & 2 & 0 & & & \\ 3 & 6 & 3 & 0 & & \\ 4 & 10 & 9 & 7 & 0 & \\ 5 & 9 & 8 & 5 & 4 & 0 \end{array} \Rightarrow \begin{array}{c|cccc} & (1,2) & 3 & 4 & 5 \\ (1,2) & 0 & & & \\ 3 & 4.5 & 0 & & \\ 4 & 9.5 & 7 & 0 & \\ 5 & 8.5 & 5 & 4 & 0 \end{array} \Rightarrow \begin{array}{c|ccc} & (1,2) & 3 & (4,5) \\ (1,2) & 0 & & \\ 3 & 4.5 & 0 & \\ (4,5) & 9 & 6 & 0 \end{array}$$



$$d_{(1,2,3),(4,5)} = \frac{1}{6}(d_{1,4} + d_{1,5} + d_{2,4} + d_{2,5} + d_{3,4} + d_{3,5}) = 8$$

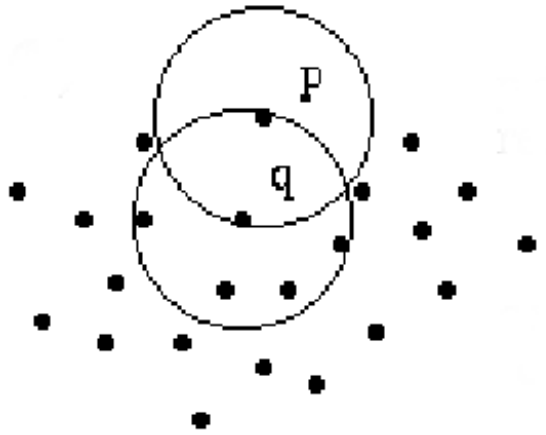
Exercice

- Classer les données selon les différents types d'agrégation

	1	2	3	4	5	6	7	8	9	10	11	12
1	0.00	0.65	0.68	0.16	0.49	0.14	0.04	0.46	0.35	0.30	0.19	0.81
2	0.65	0.00	0.03	0.49	0.16	0.78	0.62	0.19	0.30	0.35	0.84	0.16
3	0.68	0.03	0.00	0.51	0.19	0.81	0.65	0.22	0.32	0.38	0.86	0.14
4	0.16	0.49	0.51	0.00	0.32	0.30	0.14	0.30	0.19	0.14	0.35	0.65
5	0.49	0.16	0.19	0.32	0.00	0.62	0.46	0.02	0.14	0.19	0.68	0.32
6	0.14	0.78	0.81	0.30	0.62	0.00	0.16	0.59	0.49	0.43	0.05	0.95
7	0.04	0.62	0.65	0.14	0.46	0.16	0.00	0.43	0.32	0.27	0.22	0.78
8	0.46	0.19	0.22	0.30	0.02	0.59	0.43	0.00	0.11	0.16	0.65	0.35
9	0.35	0.30	0.32	0.19	0.14	0.49	0.32	0.11	0.00	0.05	0.54	0.46
10	0.30	0.35	0.38	0.14	0.19	0.43	0.27	0.16	0.05	0.00	0.49	0.51
11	0.19	0.84	0.86	0.35	0.68	0.05	0.22	0.65	0.54	0.49	0.00	1.00
12	0.81	0.16	0.14	0.65	0.32	0.95	0.78	0.35	0.46	0.51	1.00	0.00

Méthodes basées sur la densité

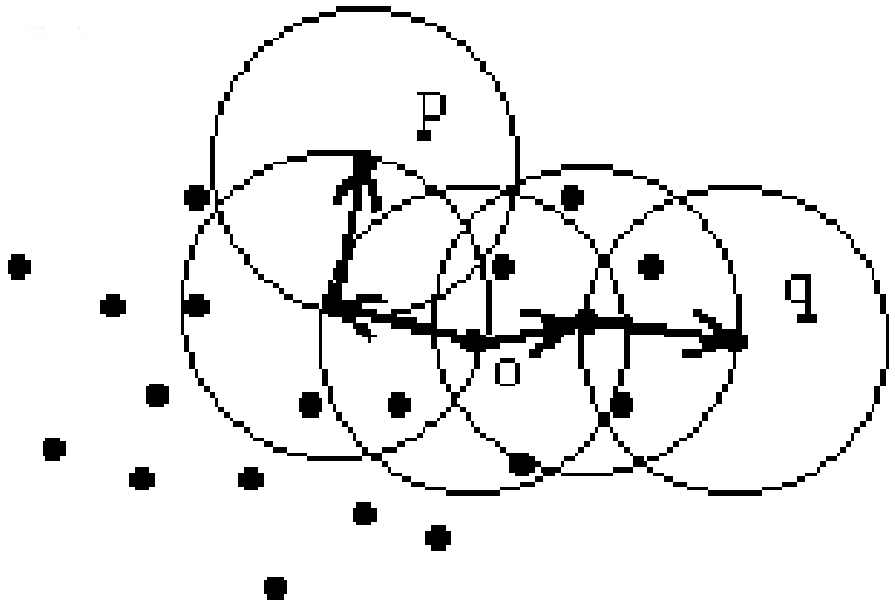
- DBSCAN
  - (Ester et al., 1995)
- Principe
  - Utilisation de la densité à la place de la distance
  - Un point est voisin d'un autre point s'il est à une distance inférieure à une valeur fixée (**`eps`**)
  - Un point est dense si le nombre de ses voisins dépasse un certain seuil (**`minPts`**)
- Exemple:
  - q est dense, mais pas p



DBSCAN

- La découverte d'un groupe se déroule en deux étapes
  - choisir aléatoirement un point dense
  - tous les points qui sont atteignables à partir de ce point, selon le critère de densité, forment un groupe
- Exemple

  on commence par *o*, ce cluster contient o, p, q, etc.

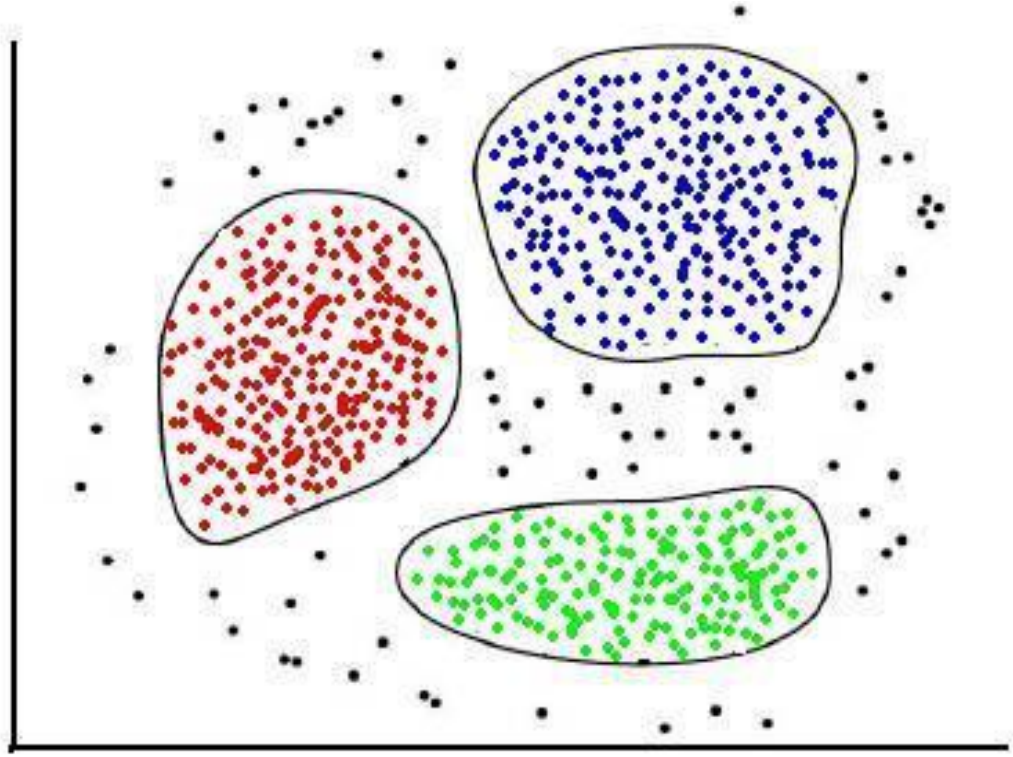


Algorithme DBSCAN

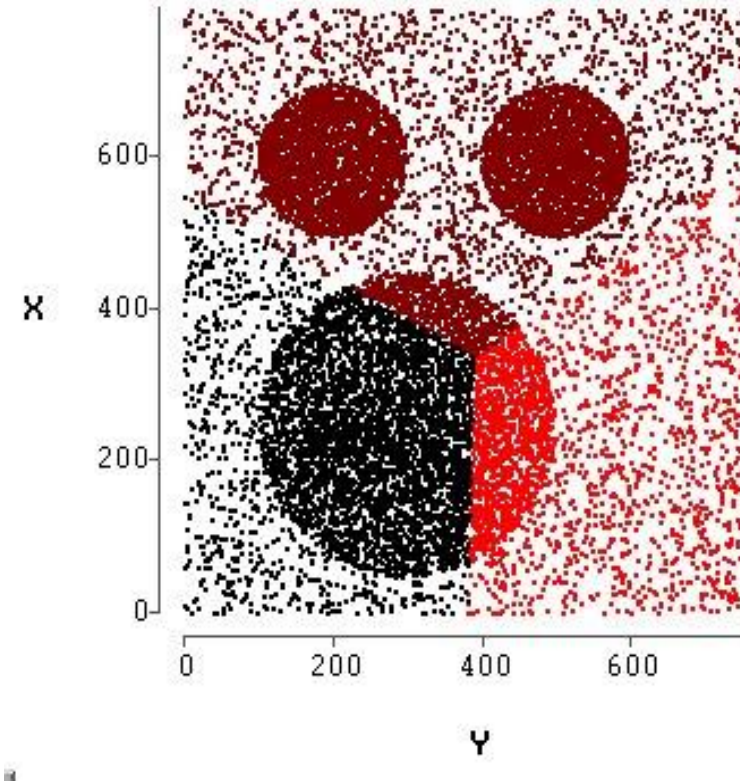
Input:
$D = \{t_1, t_2, ..., t_n\}$ //Set of elements.
$MinPts$ // Number of points in cluster.
Eps // Maximum distance for density measure.
Output:
$K = \{K_1, K_2, ..., K_k\}$ //Set of clusters.
DBSCAN Algorithm:
$k = 0$; // Initially there are no clusters.
for $i = 1$ *to* n **do**
  if t_i **is not in a cluster then**
    $X = \{t_j \mid t_j$ **is density-reachable from** $t_i\}$;
    if X **is a valid cluster then**
      $k = k + 1$;
      $K_k = X$;

DBSCAN

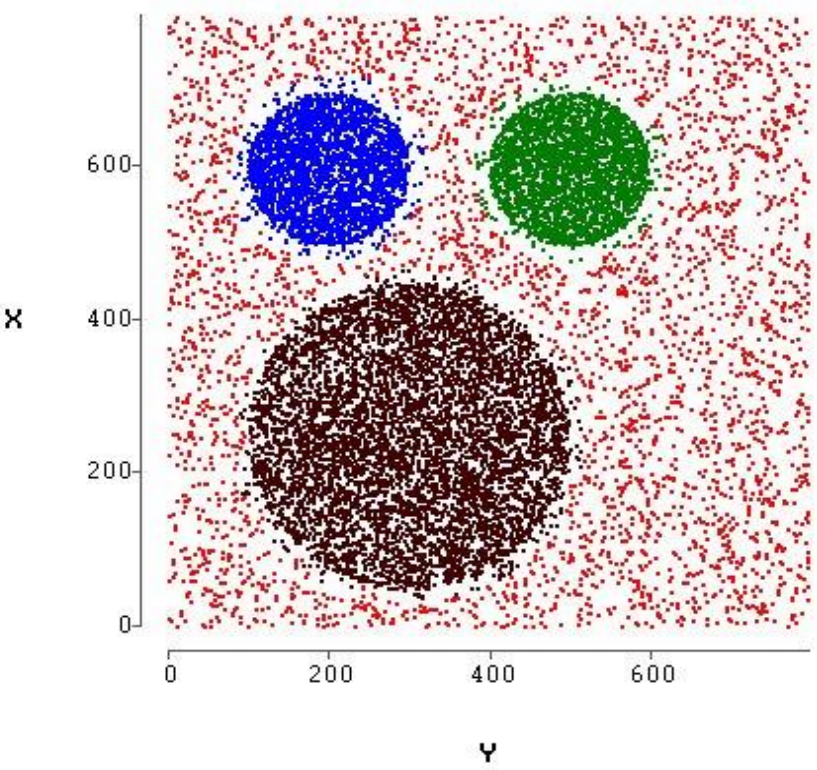
- Performances :
  - fait un seul parcours des objets du jeu de données.
  - découvre des clusters non convexes



K-means vs DBSCAN



Résultat de K-means



Résultat de DBSCAN