Clustering (Segmentation)

Problèmatique

- Soient *N* instances de données à *k* attributs,
- Trouver un partitionnement en *c* clusters (groupes) ayant un sens (*Similitude*)

- Affectation automatique de "labels" aux clusters
- *c* peut être donné, ou "découvert"
- Plus difficile que la classification car les classes ne sont pas connues à l'avance (non supervisé)
- Attributs
  - Numériques (distance bien définie)
  - Enumératifs ou mixtes (distance difficile à définir)

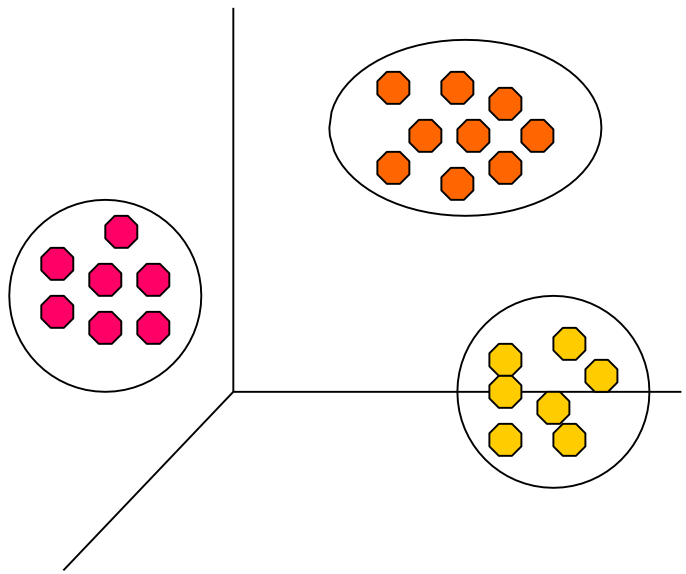
Qualité d'un clustering

- Une bonne méthode de clustering produira des clusters d'excellente qualité avec :
  - Similarité **intra-classe** importante
  - Similarité **inter-classe** faible
- La **qualité** d'un clustering dépend de :
  - La mesure de similarité utilisée
  - L'implémentation de la mesure de similarité
- La **qualité d'une méthode** de clustering est évaluée par son abilité à découvrir certains ou tous les "patterns" cachés.

Objectifs du clustering

Minimiser les distances intra-cluster

Maximiser les distances inter-clusters



Exemples d'applications

- **Marketing** : segmentation du marché en découvrant des groupes de clients distincts à partir de bases de doneées d'achats.
- **Environnement**: identification des zones terrestres similaires (en termes d'utilisation) dans une base de données d'observation de la terre.
- **Assurance**: identification de groupes d'assurés distincts associés à un nombre important de déclarations.
- **Planification de villes**: identification de groupes d'habitations suivant le type d'habitation, valeur, localisation géographique, …
- **Médecine**: Localisation de tumeurs dans le cerveau
  - Nuage de points du cerveau fournis par le neurologue
  - Identification des points définissant une tumeur

Mesure de la similarité

- Il n'y a pas de **définition unique** de la similarité entre objets
  - Différentes mesures de distances d(x,y)
- La définition de la similarité entre objets dépend de :
  - Le type des données considérées
  - Le type de similarité recherchée

Choix de la distance

- Propriétés d'une distance :

1. $d(x, y) \geq 0$
2. $d(x, y) = 0 \text{ iff } x = y$
3. $d(x, y) = d(y, x)$
4. $d(x, z) \leq d(x, y) + d(y, z)$

- Définir une distance sur chacun des champs
- Champs numériques : d(x,y) = |x-y|, d(x,y)= |x-y|/dmax (distance normalisée).
- Exemple : Age, taille, poids, …

Distance – Données numériques

- Combiner les distances : Soient x=(x1,…,xn) et y=(y1, …,yn)
- Exemples numériques :

- Distance euclidienne : $d(x,y)=\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2}$
- Distance de Manhattan : $d(x,y)=\sum_{i=1}^{n}|x_i-y_i|$
- Distance de Minkowski : $d(x,y)=\sqrt[q]{\sum_{i=1}^{n}|x_i-y_i|^q}$

q=1 : distance de Manhattan.

q=2 : distance euclidienne

Choix de la distance

- Champs discrets :
  - Données binaires : d(0,0)=d(1,1)=0, d(0,1)=d(1,0)=1
  - Donnée énumératives : distance nulle si les valeurs sont égales et 1 sinon.
  - Donnée énumératives ordonnées : idem. On peut définir une distance utilisant la relation d'ordre.
- Données de types complexes : textes, images, données génétiques, ...

Distance – Données binaires

Table de contingence (dissimilarité)

		Object *j*		
		1	0	*sum*
Object *i*	1	*a*	*b*	*a+b*
	0	*c*	*d*	*c+d*
	sum	*a+c*	*b+d*	*p*

- **Coefficient de correspondance simple** (similarité invariante, si la variable binaire est **symétrique**) :

$$d(i,j) = \frac{b+c}{a+b+c+d}$$

- **Coefficient de Jaccard** (similarité non invariante, si la variable binaire est **asymétrique**):

$$d(i,j) = \frac{b+c}{a+b+c}$$

Données énumératives

- *Généralisation* des variables binaires, **avec plus de 2 états**, e.g., rouge, jaune, bleu, vert

- **correpondance simple**
  - **m**: # de correspondances, **p**: # total de variables

$$d(i, j) = \frac{p - m}{p}$$

Distance – Données binaires

Exemple : dissimilarité entre variables binaires

- Table de patients

Nom	Sexe	Fièvre	Toux	Test-1	Test-2	Test-3	Test-4
Jack	M	Y	N	P	N	N	N
Mary	F	Y	N	P	N	P	N
Jim	M	Y	P	N	N	N	N

- 8 attributs, avec
  - Sexe un attribut symétrique, et
  - Les attributs restants sont asymétriques (test VIH, …)

Distance – Données binaires

- Les valeurs Y et P sont initialisées à 1, et la valeur N à 0.
- Calculer la distance entre patients, basée sur le coefficient de Jaccard.

$$d(jack, mary) = \frac{0+1}{2+0+1} = 0.33$$

$$d(jack, jim) = \frac{1+1}{1+1+1} = 0.67$$

$$d(jim, mary) = \frac{1+2}{1+1+2} = 0.75$$

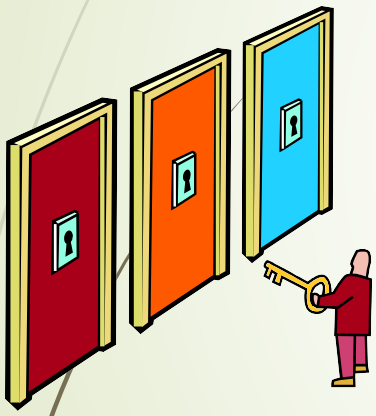
Distance – Données mixtes

- Exemple : (Age, Propriétaire résidence principale, montant des mensualités en cours)

- x=(30,1,1000), y=(40,0,2200), z=(45,1,4000)
- $d(x,y)=sqrt((10/15)^2 + 1^2 + (1200/3000)^2) = 1.27$
- $d(x,z)= sqrt((15/15)^2 + 0^2 + (3000/3000)^2) = 1.41$
- $d(y,z)= sqrt((5/15)^2 + 1^2 + (1800/3000)^2) = 1.21$
- plus proche voisin de x = y

- Distances normalisées.
- Sommation : $d(x,y)=d1(x1,y1) + \ldots + dn(xn,yn)$

Méthodes de Clustering



- Méthode de partitionnement (K-moyennes)
- Méthodes hiérarchiques (par agglomération)
- Méthode par voisinage dense

- Caractéristiques
  - Apprentissage non supervisé (classes inconnues)
  - Pb : interprétation des clusters identifiés

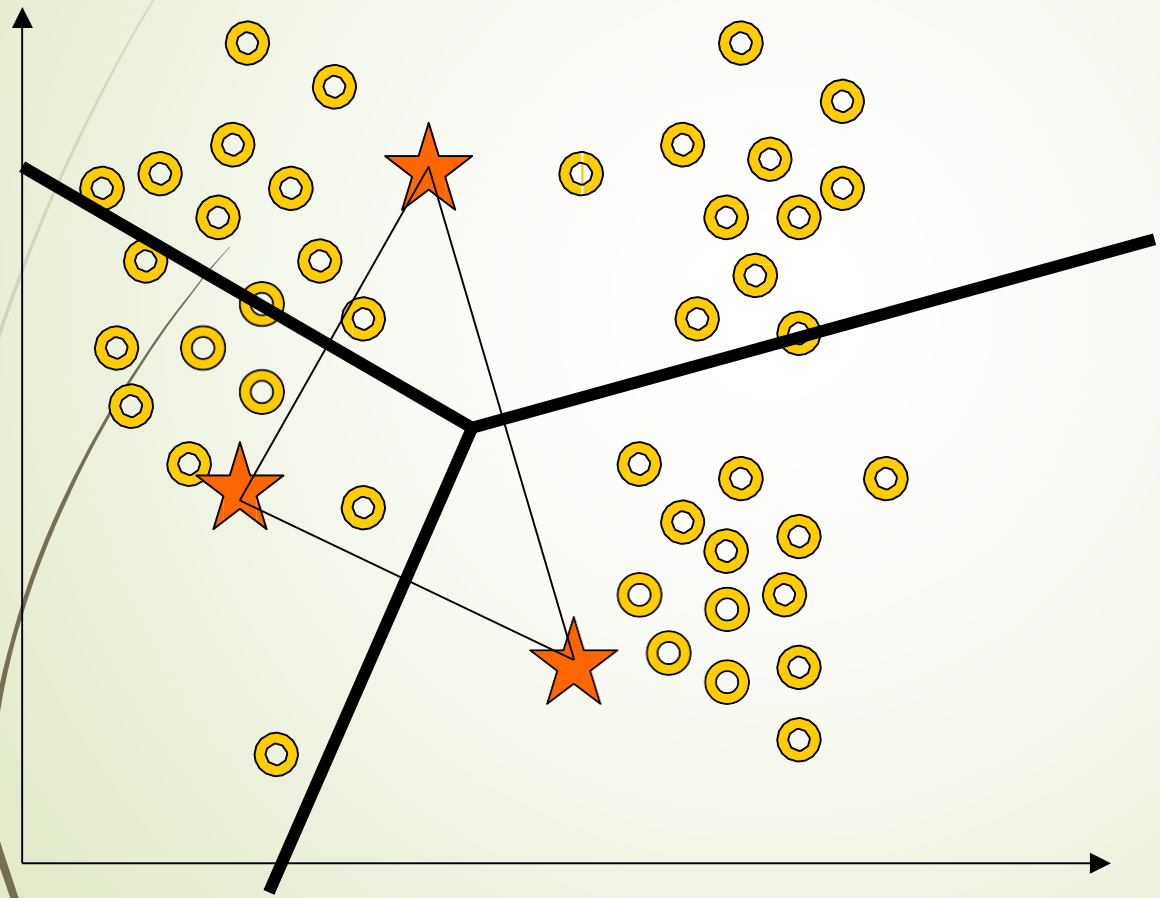
Méthodes de clustering

- Extensibilité
- Abilité à traiter différents types de données
- Découverte de clusters de différents formes
- Connaissances requises (paramètres de l'algorithme)
- Abilité à traiter les données bruitées et isolées.

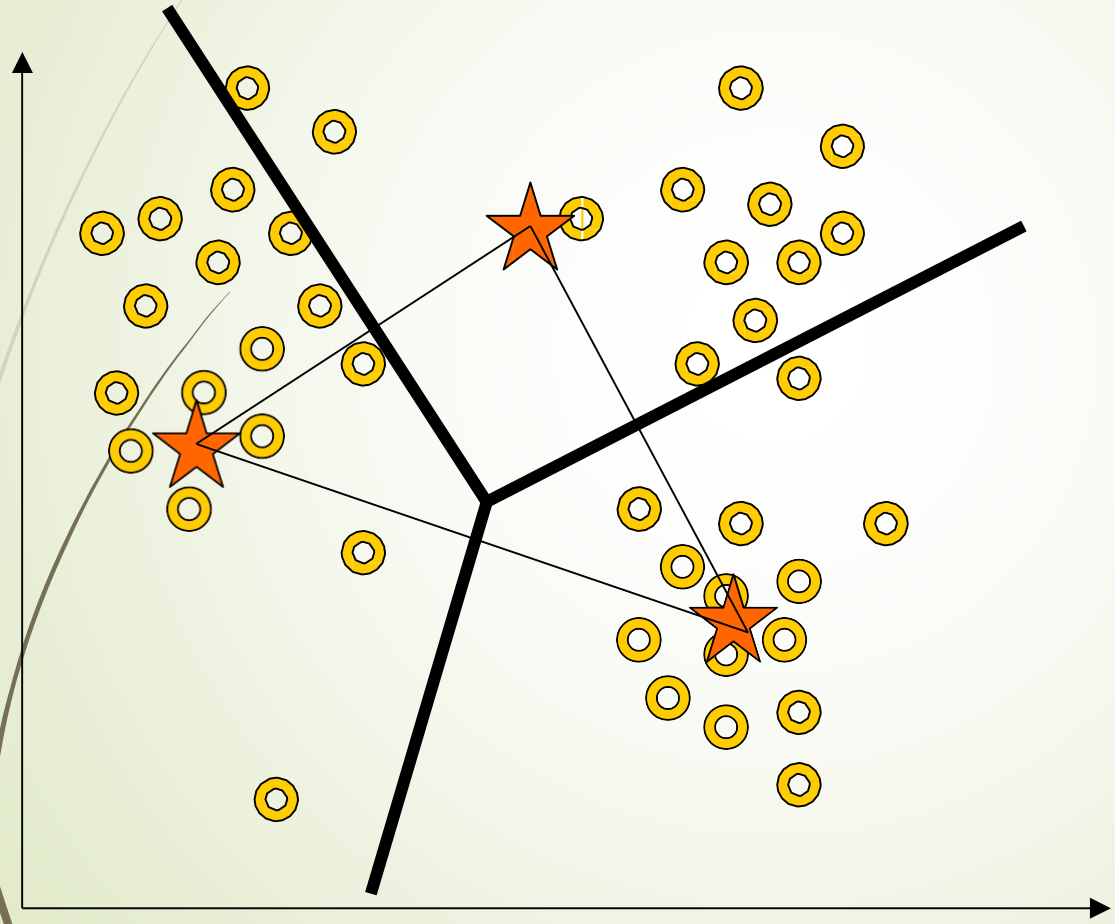
k-moyennes (K-means)

- Entrée : un échantillon de m enregistrements x1, …, xm
- 1. Choisir k centres initiaux c1, …, ck
- 2. Répartir chacun des m enregistrements dans le groupe i dont le centre ci est le plus proche.
- 3. Si aucun élément ne change de groupe alors arrêt et sortir les groupes
- 4. Calculer les nouveaux centres : pour tout i, ci est la moyenne des éléments du groupe i.
- Aller en 2.

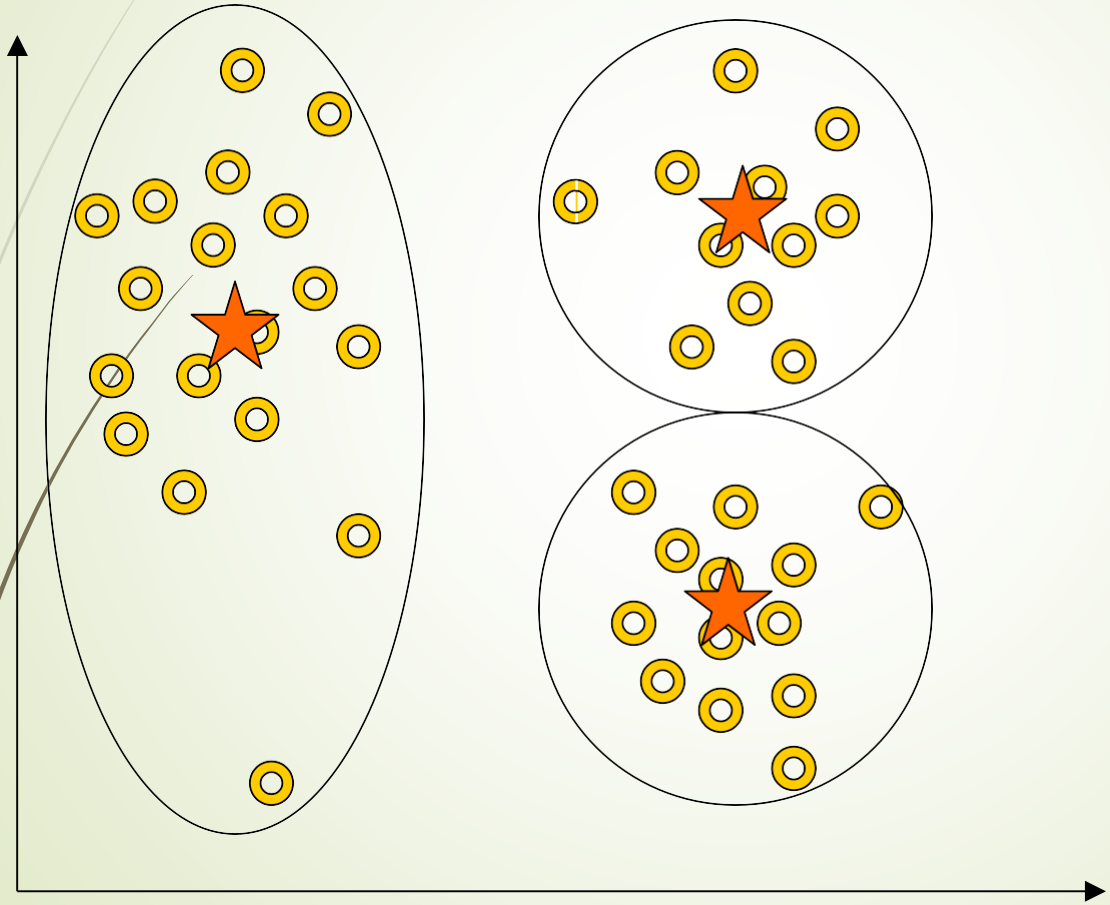
K-means (Illustration)



K-means (Illustration)



K-means (Illustration)



K-means : Avantages

- **Relativement extensible** dans le traitement d'ensembles de taille importante

- **Relativement efficace** : $O(t.k.n)$, où
- n représente nombre objets, k nombre clusters, et t nombre d'iterations.

- Produit généralement un **optimum local** ; un optimum global peut être obtenu en utilisant d'autres techniques telles que : algorithmes génétiques, …

K-means : Inconvénients

- **Applicable** seulement dans le cas où la moyenne des objets est définie
- **Besoin de spécifier** *k, le nombre de clusters*, a priori
- **Incapable** de traiter les données bruitées (noisy).
- **Non adapté** pour découvrir des clusters avec structures non-convexes, et des clusters de tailles différentes
- Les **points isolés** sont mal gérés (doivent-ils appartenir obligatoirement à un cluster ?) - probabiliste

K-moyennes : Variantes

- Sélection des centres initiaux
- Calcul des similarités
- Calcul des centres (K-medoids : [Kaufman & Rousseeuw'87])
- GMM : Variantes de *K-moyennes* basées sur les probabilités
- K-modes : données catégorielles [Huang'98]
- K-prototype : données mixtes (numériques et catégorielles)

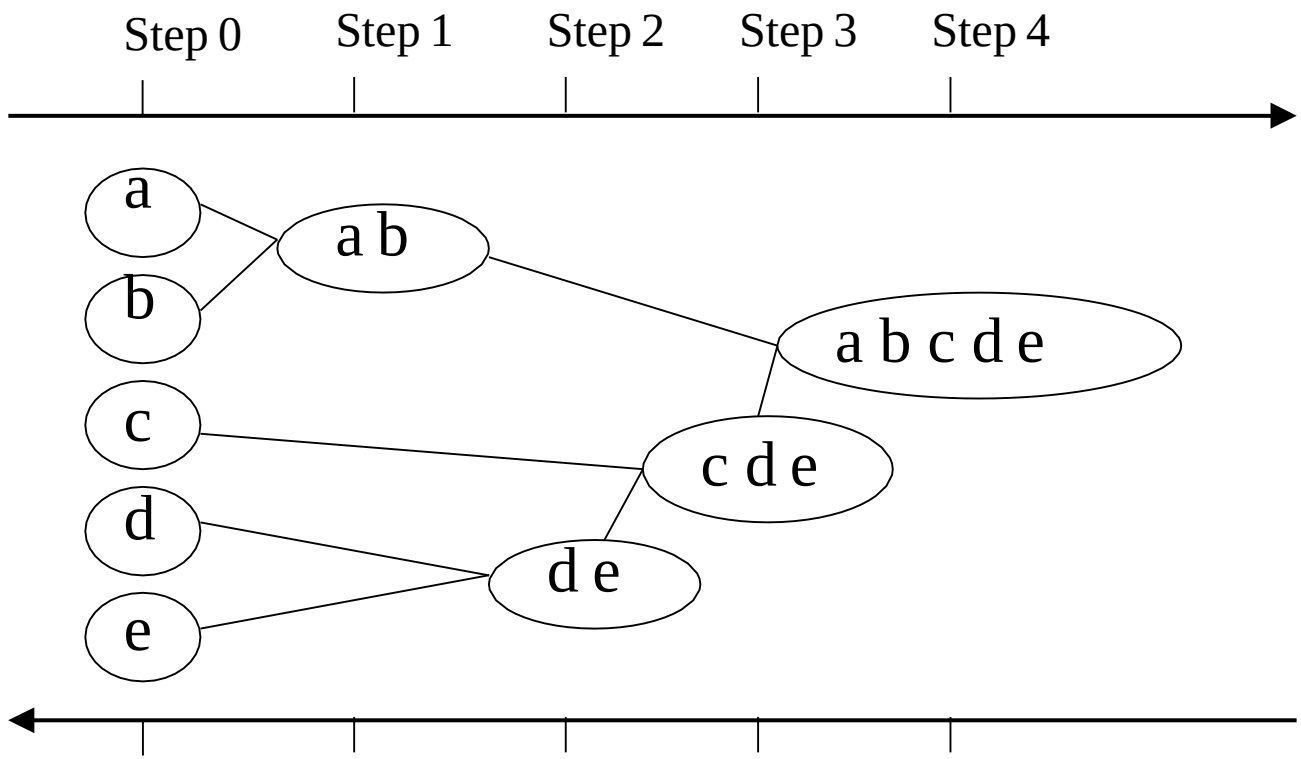
Méthodes hiérarchiques

- **Une méthode hiérarchique** : construit une hiérarchie de clusters, non seulement une partition unique des objets.
- Le nombre de clusters **k** n'est pas exigé comme donnée
- Utilise une **matrice de distances** comme critère de clustering
- Une **condition de terminaison** peut être utilisée (ex. Nombre de clusters)

Méthodes hiérarchiques

- Entrée : un échantillon de m enregistrements x1, …, xm
- 1. On commence avec m clusters (cluster = 1 enregistrement)
- 2. Grouper les deux clusters les plus « proches ».
- 3. S'arrêter lorsque tous les enregistrements sont membres d'un seul groupe
- 4. Aller en 2.

Exemple



Arbre de clusters

- Résultat : Graphe hiérarchique qui peut être coupé à un niveau de dissimilarité pour former une partition.
- La hiérarchie de clusters est représentée comme un **arbre de clusters**, appelé **dendrogramme**
- Les feuilles de l'arbre représentent les objets
- Les noeuds intermédiaires de l'arbre représentent les clusters

Distance entre clusters

- Distance entre les centres des clusters (Centroid Method)
- Distance minimale entre toutes les paires de données des 2 clusters (Single Link Method) $d(i, j) = \min_{x \in Ci, y \in Cj} \{ d(x, y) \}$
- Distance maximale entre toutes les paires de données des 2 clusters (Complete Link Method)

$$d(i, j) = \max_{x \in Ci, y \in Cj} \{ d(x, y) \}$$

- Distance moyenne entre toutes la paires d'enregistrements (Average Linkage) $d(i, j) = avg_{x \in Ci, y \in Cj} \{ d(x, y) \}$

Méthodes hiérarchiques : Avantages

- **Conceptuellement simple**
- **Propriétés théoriques** sont bien **connues**
- Quand les clusters sont groupés, **la décision est définitive => le nombre d'alternatives différentes à** à examiner est **réduit**

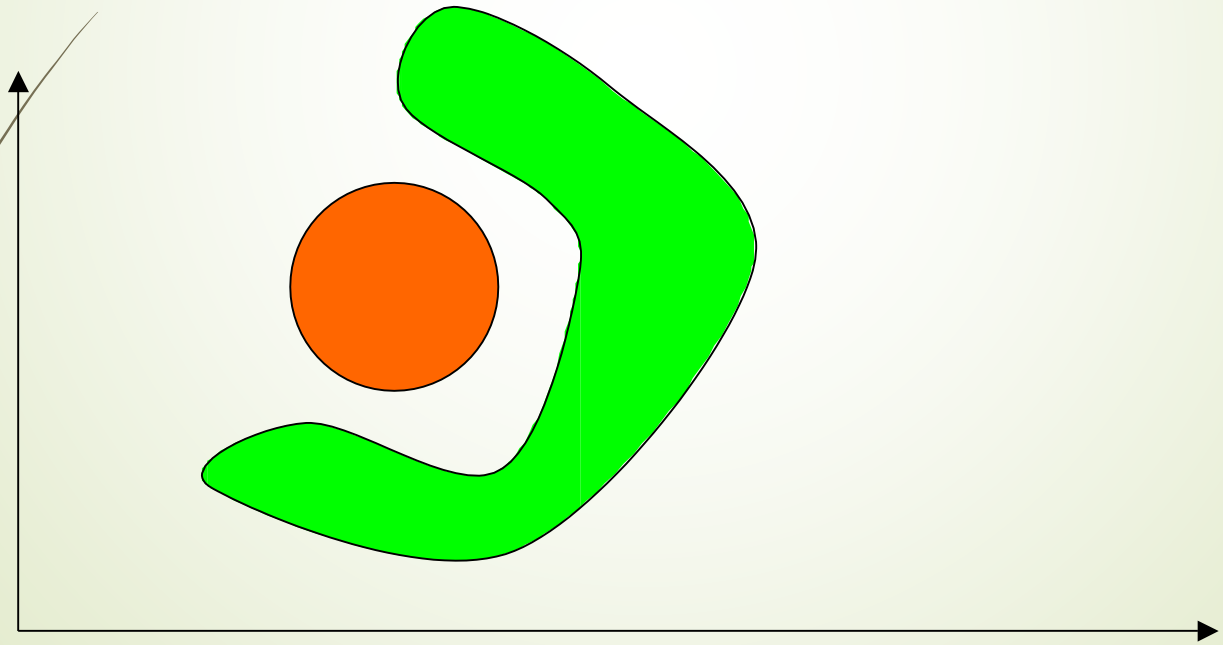
Méthodes hiérarchiques : Inconvénients

- **Groupement** de clusters est **définitif** => décisions erronnées sont **impossibles à modifier** ultérieurement

- Méthodes **non extensibles** pour des ensembles de données de grandes tailles

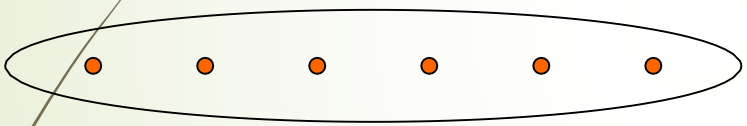
Méthodes basées sur la densité

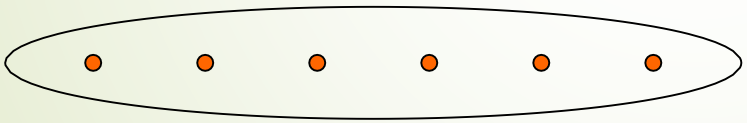
- Pour ce types de problèmes, l'utilisation de mesures de similarité (distance) est moins efficace que l'utilisation de densité de voisinage.

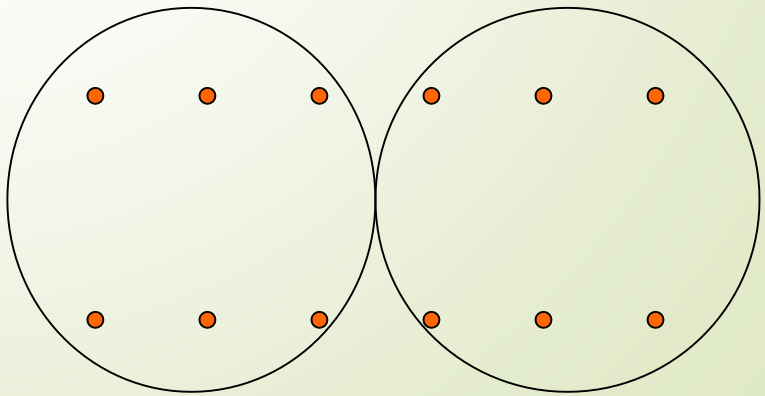


Méthodes basées sur la densité

- Minimiser la distance inter-clusters n'est pas toujours un bon critère pour reconnaître des «formes » (applications géographiques, reconnaissance de formes – tumeurs, …).







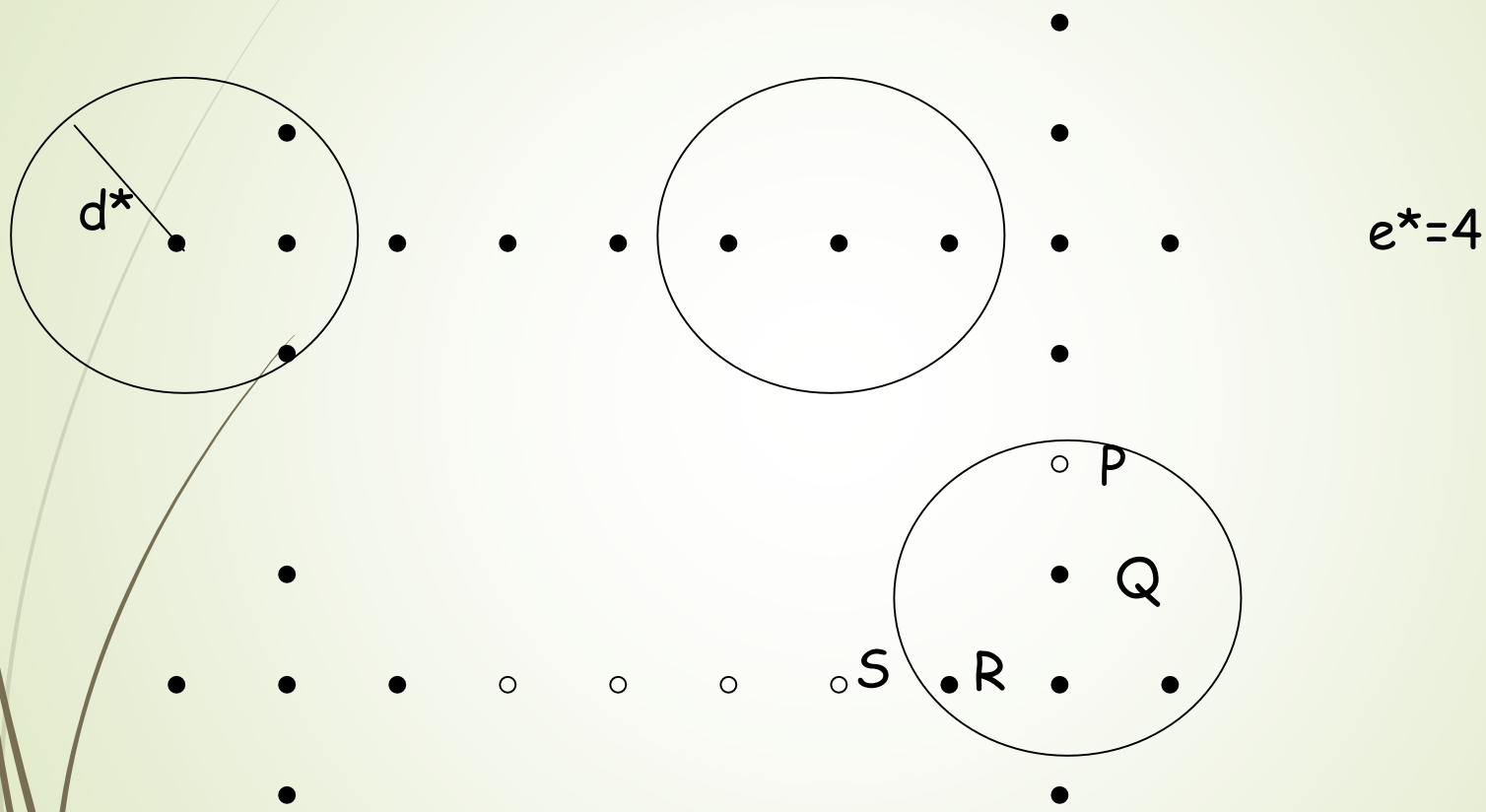
Méthodes basées sur la densité

- Soit d* un nombre réel positif
- Si d(P,Q)<=d*, Alors P et Q appartiennent au même cluster
- Si P et Q appartiennent au même cluster, et d(Q,R)<=d*, Alors P et R appartiennent au même cluster

Méthodes basées sur la densité

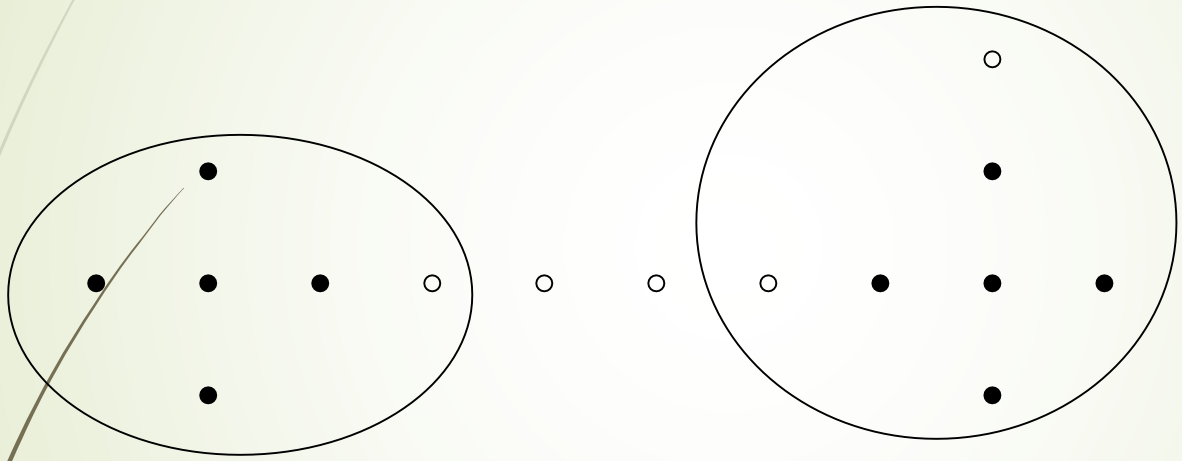
- Soit e* un nombre réel positif
- Un point P est dense ssi |{Q/d(P,Q)<=d*}|>=e*
- Si P et Q appartiennent au même cluster, et d(Q,R)<=d* et Q est dense, Alors P et R appartiennent au même cluster
- Les points non-denses sont appelés points de « bordure ».
- Les points en dehors des clusters sont appelés « bruits ».

Méthodes basées sur la densité



- Points noirs sont denses ; les autres ne sont pas denses
- Pour montrer que P et S appartiennent au même cluster, il suffit de montrer que P et R appartiennent au même cluster. Pour le montrer pour P et R, il suffit de le montrer pour P et Q …

Méthodes basées sur la densité



- Deux clusters sont trouvés
- Deux points sont des « bruits »
- Trois points sont des « bordures »

Clustering – Résumé (1)

- Le clustering groupe des objets en se basant sur leurs similarités.
- Le clustering possède plusieurs applications.
- La mesure de similarité peut être calculée pour différents types de données.
- La sélection de la mesure de similarité dépend des données utilisées et le type de similarité recherchée.

Clustering – Résumé (2)

- Les méthodes de clustering peuvent être classées en :
  - Méthodes de partitionnement,
  - Méthodes hiérarchiques,
  - Méthodes à densité de voisinage.
- Plusieurs travaux de recherche sur le clustering en cours et en perspective.
- Plusieurs applications en perspective : Génomique, Environnement, …