DATA MINING

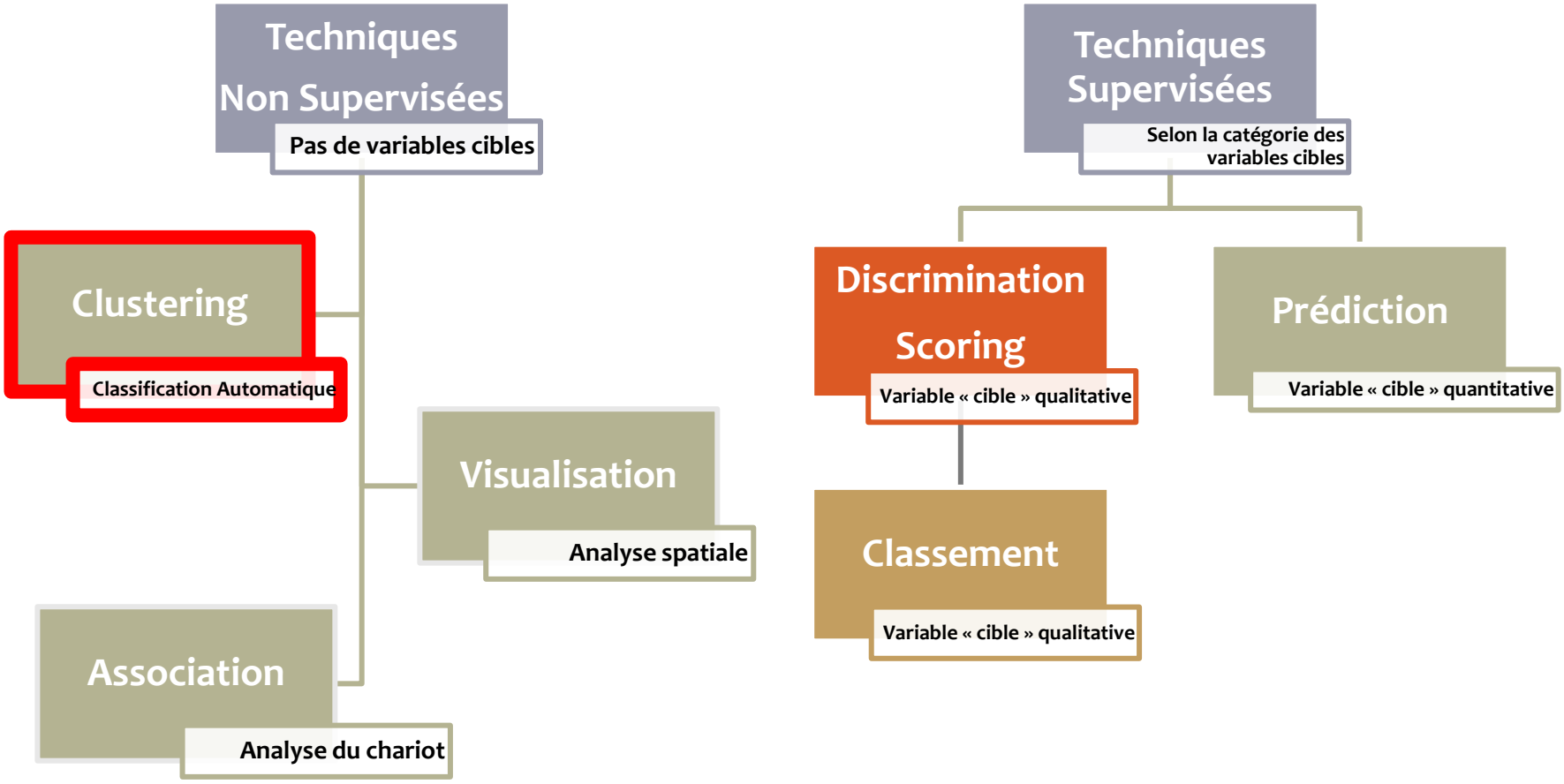
LA SEGMENTATION

CLUSTERING

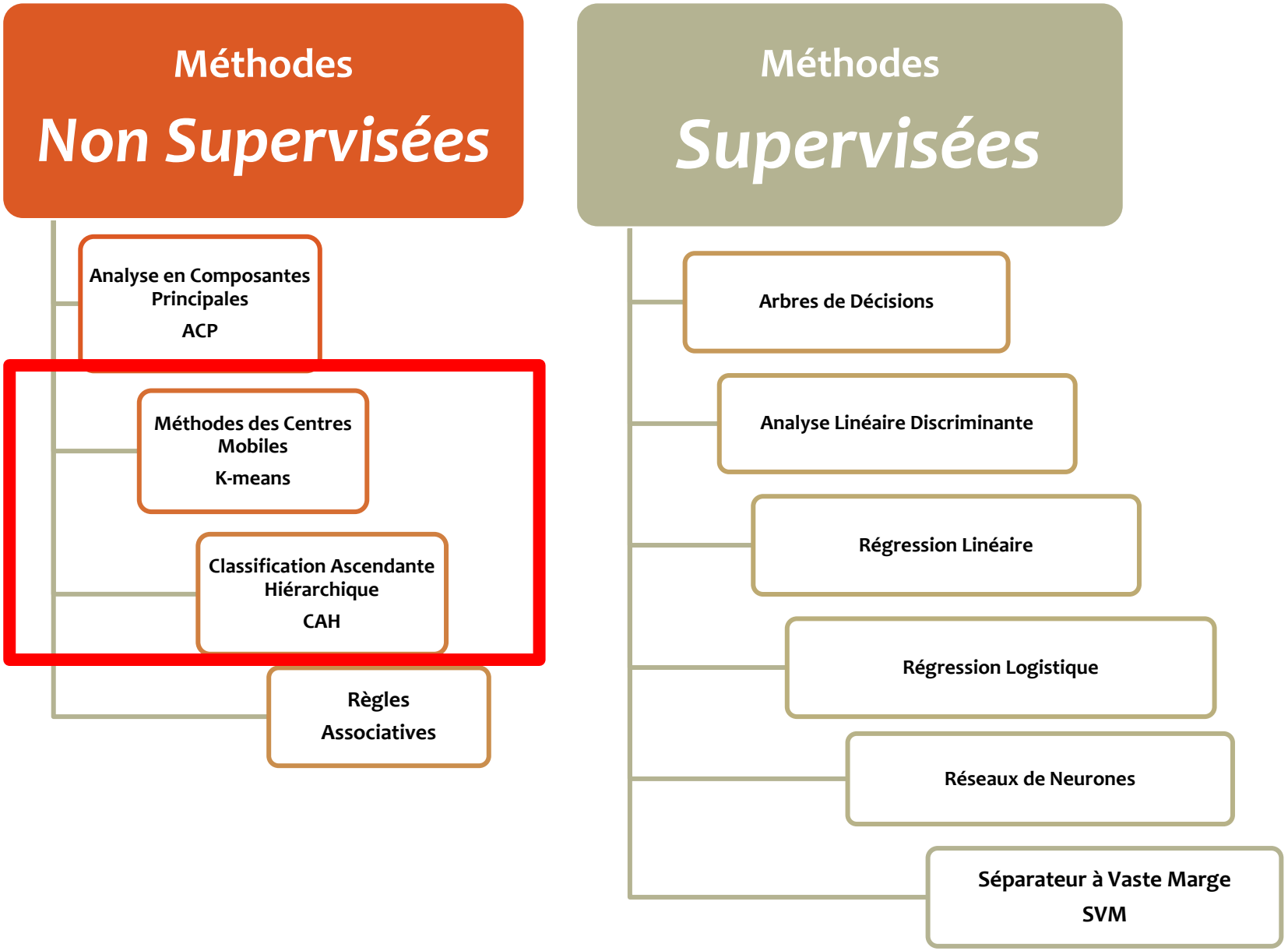
Equipe Data Mining

2023-2024

TYPES D'APPLICATIONS



DEUX FAMILLES DE TECHNIQUES



UTILITÉ DE LA SEGMENTATION

Banque / Assurance

- Catégoriser la clientèle : chercher un profil qui représente les membres de chaque classe
- Regrouper les clients selon des critères et caractéristiques communs : cibler « les mailing »

Médecine

- Déterminer des segments de patients susceptibles d'être soumis à des protocoles thérapeutiques déterminés, chaque segment regroupant tous les patients réagissant identiquement
- Retrouver les différents comportements similaires

Biologie – Zoologie – Ethologie – Sciences humaines

- Expliquer les relations entre espèces, races, genres, familles,
- Retrouver de nouvelles répartitions

- Profiling
- Analyse sémantique, sentimentale,
- Analyse et mesure de la tonalité d'un contenu textuel
- Catégorisation des concepts ou des entités nommés
- Construction d'agrégateur synthétique à partir des flux d'actualités

PROBLÉMATIQUE

En tant que data-mineurs votre client souhaite détecter des profils homogènes à partir de cette population.

PROBLÉMATIQUE

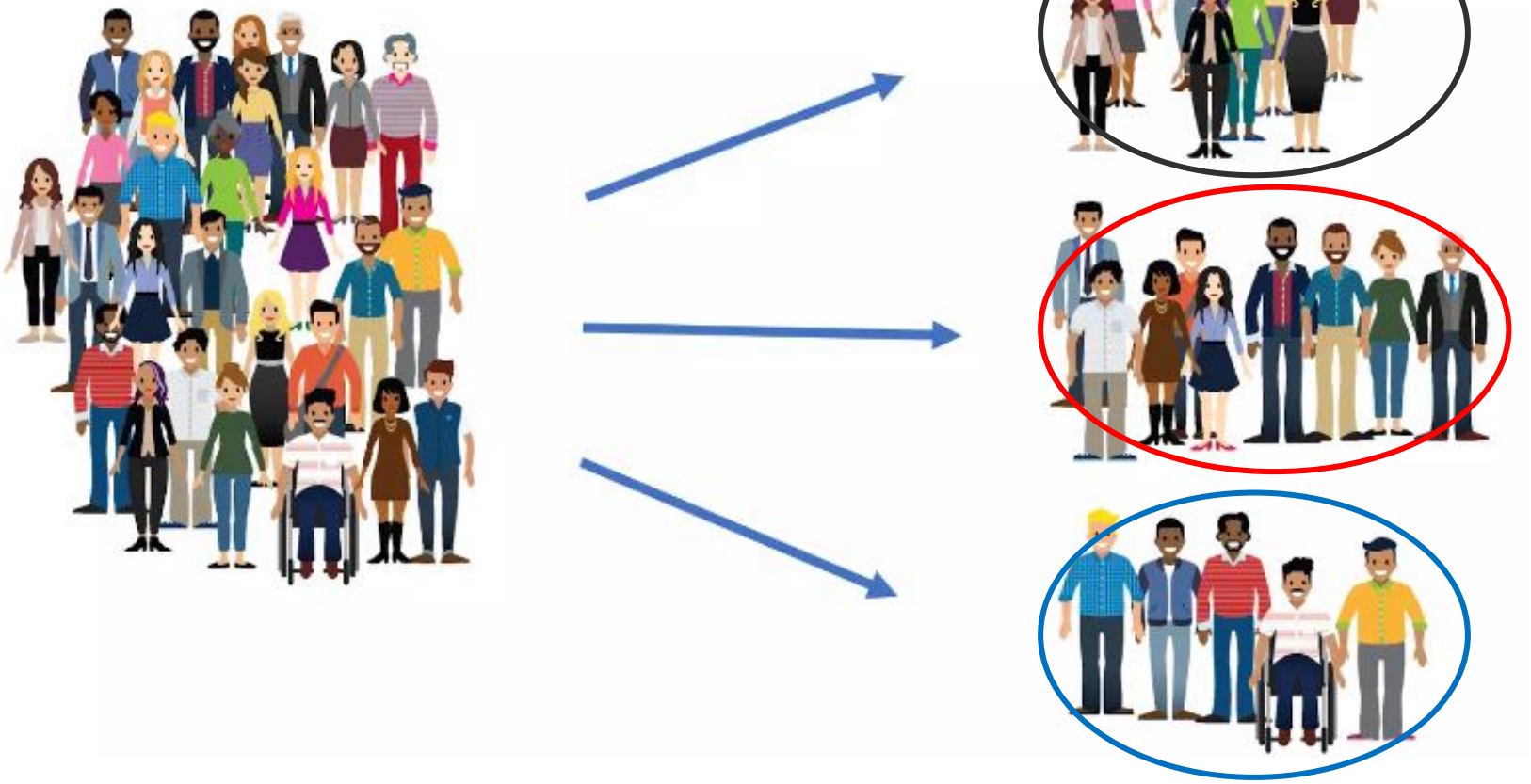
Segmentation démographique

	âge	situation familiale	genre	nationalité	revenus	éducation	etc
1							
2							

Segmentation comportementale

	Type page vues	Durée session	Fréquence utilisation	Type articles achetes	etc
1					
2					

PROBLÉMATIQUE



PROBLÉMATIQUE

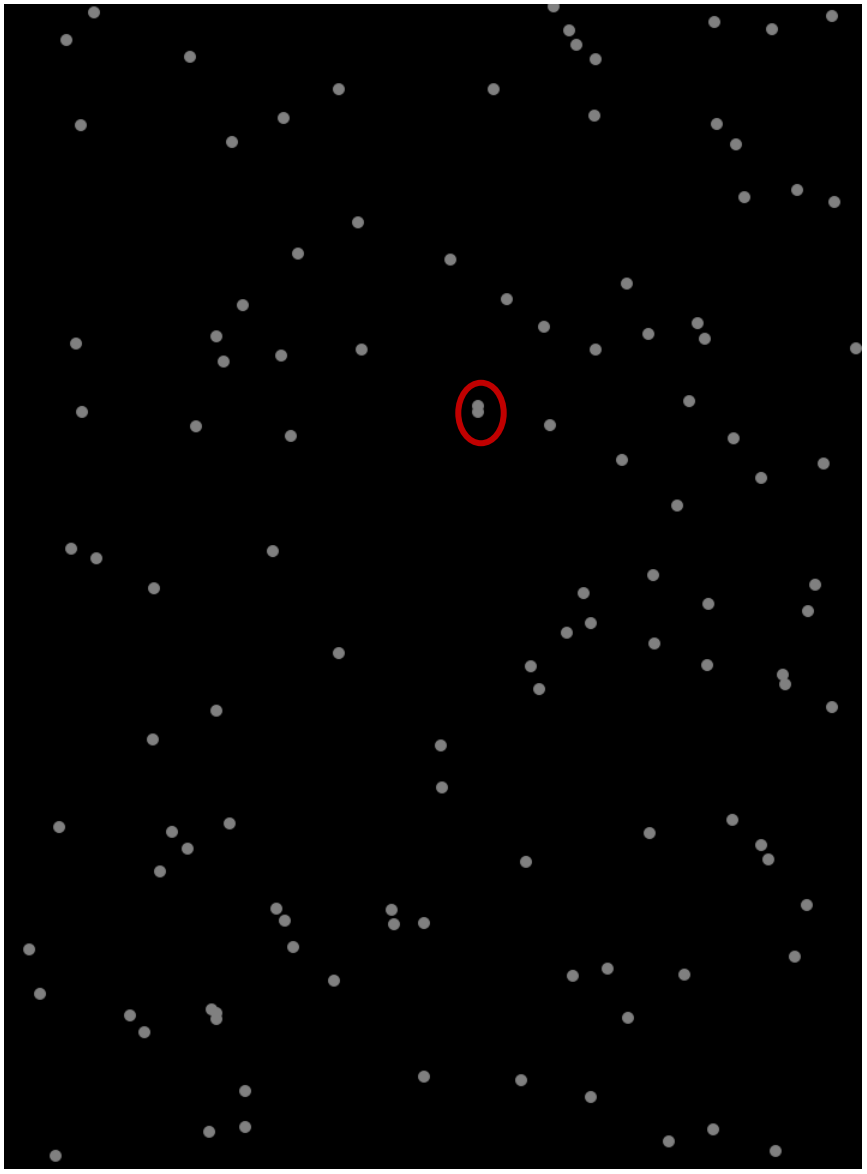
How many clusters?

Six Clusters

Two Clusters

Four Clusters

IDÉE DE BASE DE LA SEGMENTATION



Deux individus se ressemblent le plus

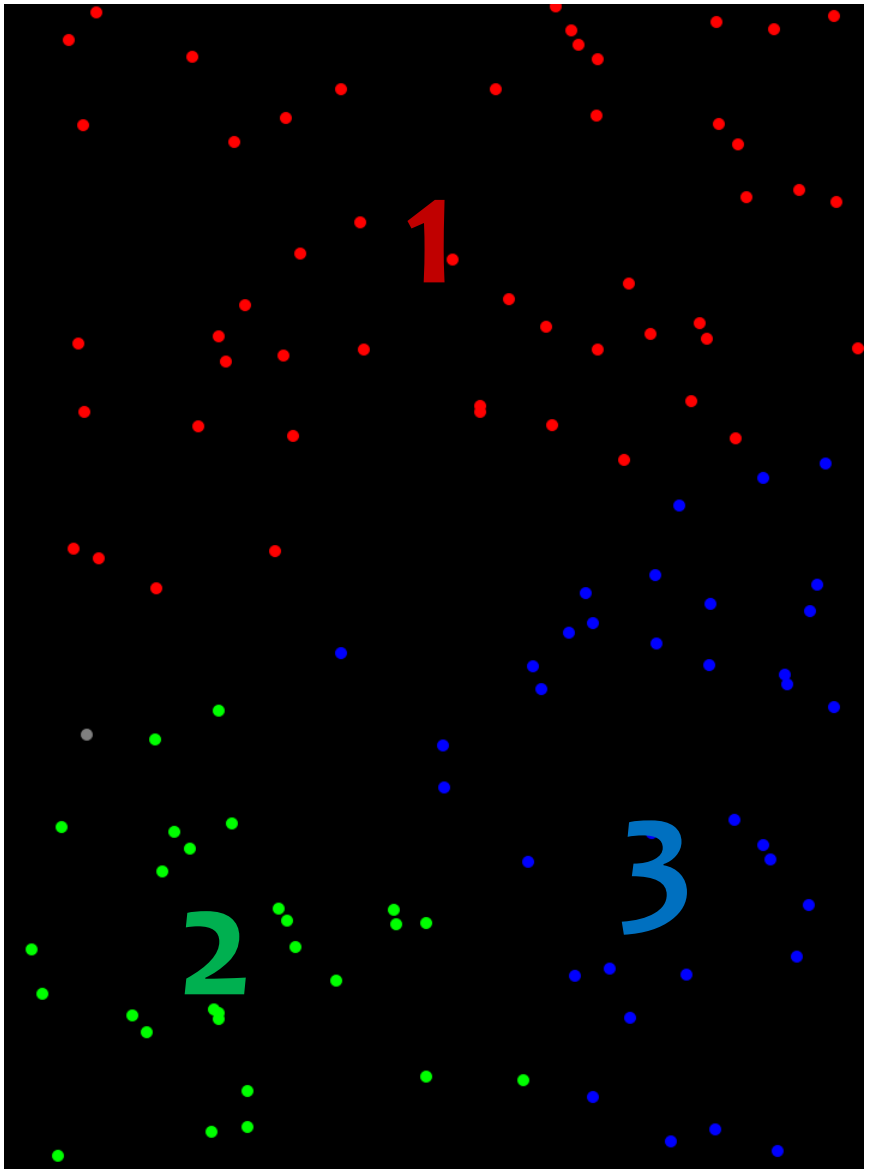
SI

les points qui les représentent dans le nuage sont les plus proches

Nécessité d'une métrique de la distance

Distance Euclidienne
Distance de Mahalanobis
Distance de Manhattan
Distance de Ward
….

IDÉE DE BASE DE LA SEGMENTATION



Deux individus se ressemblent le plus

SI

les points qui les représentent dans le nuage sont les plus proches

Nécessité d'une métrique de la distance

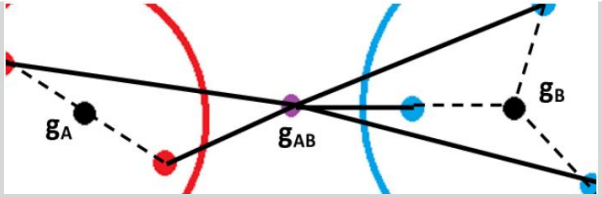
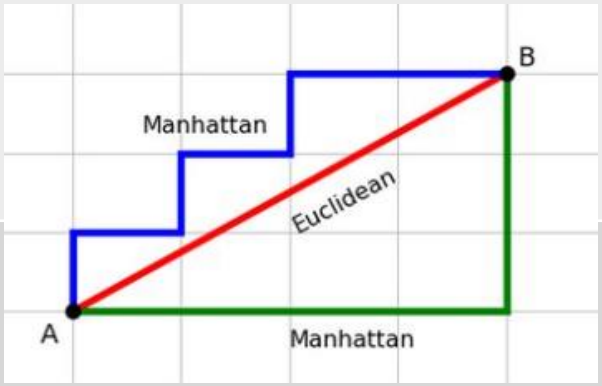
Distance Euclidienne
Distance de Mahalanobis
Distance de Manhattan
Distance de Ward
…..

Un critère d'évaluation d'une classification

$$I_{tot} = I_{inter} + I_{intra}$$

IDÉE DE BASE DE LA SEGMENTATION

Métriques de la distance

Nom	Fonction	Illustration
Distance de watt	$D_{KL} = \frac{\left\|\overline{x_K} - \overline{x_L}\right\|^2}{\frac{1}{N_K} + \frac{1}{N_L}}$	
distance euclidienne	$\sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$	
distance de Manhattan	$\sum_{i=1}^{n} \left\|x_i - y_i\right\|$	

IDÉE DE BASE DE LA SEGMENTATION

Inertie

Si $G = \{e_i \ : \ i = \{1 : n\}\}$ est un groupe d'individus, de centre de gravité g , partitionné en k classes d'effectifs $n_1, \ n_2, \ .., \ n_k$ qu'on appellera $G_1, \ G_2, \ .., \ G_k$ qui ont pour centres de gravité $g_1, \ g_2, \ .., \ g_k$ alors [i 1]

l'inertie totale du nuage est égale à : $I_t = \frac{1}{n} \sum_{i=1}^{n} d(e_i, g)^2$ où *d* est une distance

l'inertie interclasse est égale à : $I_e = \frac{1}{n} \sum_{i=1}^{k} n_i \times d(g_i, g)^2$

l'inertie intraclasse est égale à : $I_a = \frac{1}{n} \sum_{i=1}^{k} \sum_{e \in Gi} d(e, g_i)^2$

MÉTHODOLOGIES

Partitionnement

Kmeans

Hiérarchique

CAH

K-MEANS

MÉTHODE DES CENTRES MOBILES

Mohamed Heny SELMI

PRÉSENTATION DU K-MEANS

- ✓ L'algorithme des K-moyennes est un algorithme qui permet de trouver des classes dans des données.
- ✓ les classes qu'il construit n'entretiennent jamais de relations hiérarchiques: une classe n'est jamais incluse dans une autre classe
- ✓ L'algorithme fonctionne en précisant le nombre de classes attendues.
- ✓ L'algorithme calcule les distances **Intra-Classe** et **Inter-Classe.**
- ✓ Il travaille sur des **variables continues**.

PRINCIPE ALGORITHMIQUE

Algorithme K-Means

Entrée : *k le nombre de groupes cherchés*

DEBUT

Choisir aléatoirement les centres des groupes

REPETER

i. ***Affecter chaque cas au groupe dont il est le plus proche au son centre***

ii. ***Recalculer le centre de chaque groupe***

JUSQU'A *(stabilisation des* ***centres****)*

OU *(nombre d'itérations =* ***t****)*

OU *(stabilisation de* ***l'inertie totale*** *de la population)*

FIN

STABILISATION DE L'INERTIE TOTALE DE LA POPULATION

Inertie totale I_{tot} : somme de l'inertie intraclasse I_A et de l'inertie interclasse I_c

$$I_{tot} = I_A + I_c$$

Inertie intraclasse I_A : somme des inerties totales de chaque classe

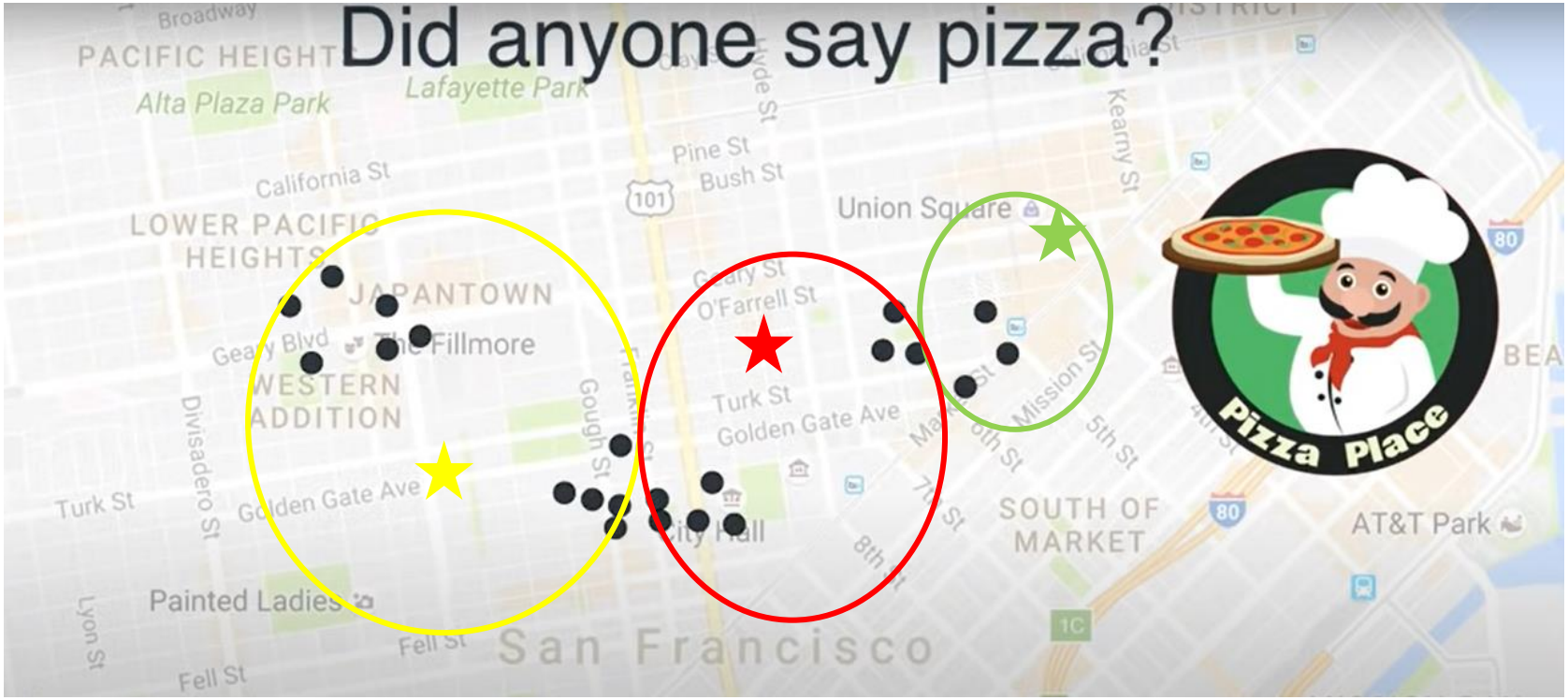
Inertie interclasse I_c : moyenne (pondérée par la somme des poids de chaque classe) des carrés des distances des barycentres de chaque classe au barycentre global

EXEMPLE K-MEANS

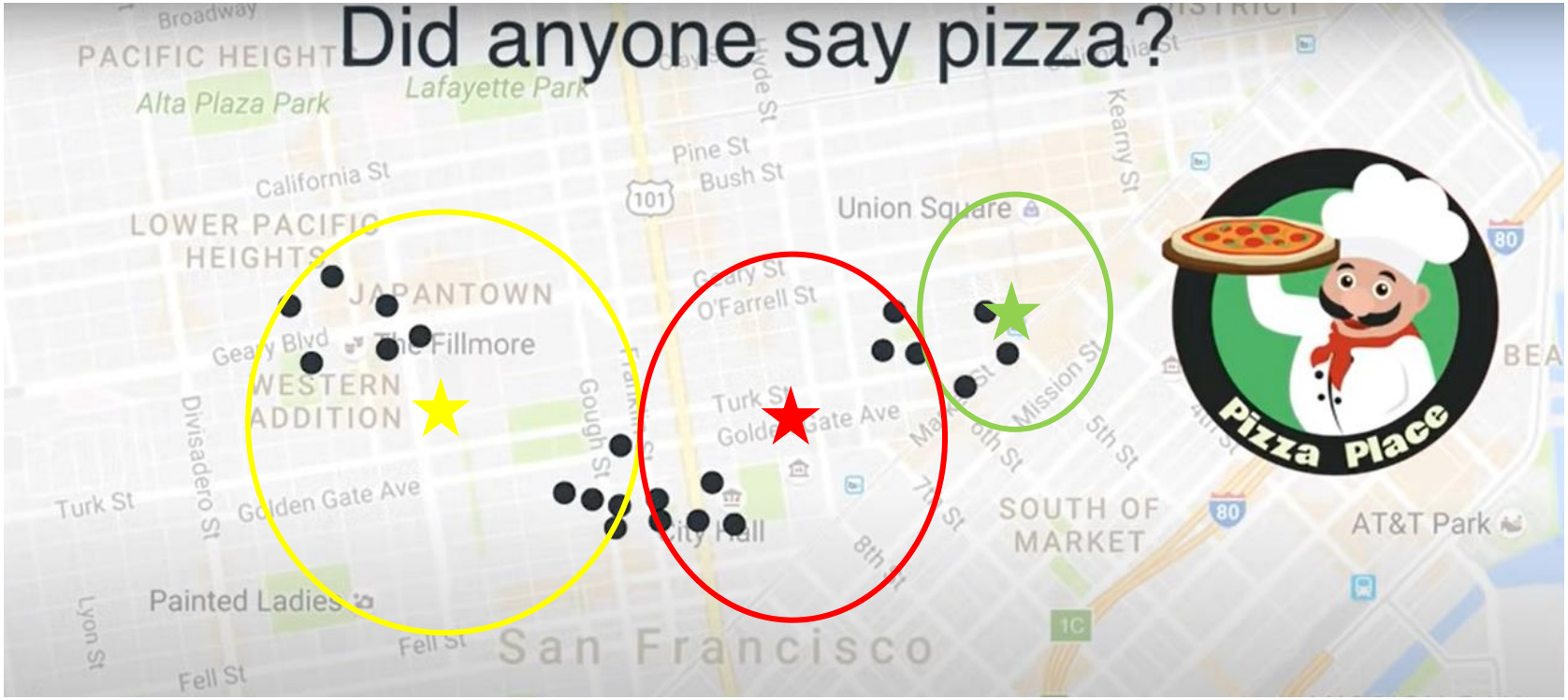
EXEMPLE K-MEANS

EXEMPLE K-MEANS

EXEMPLE K-MEANS

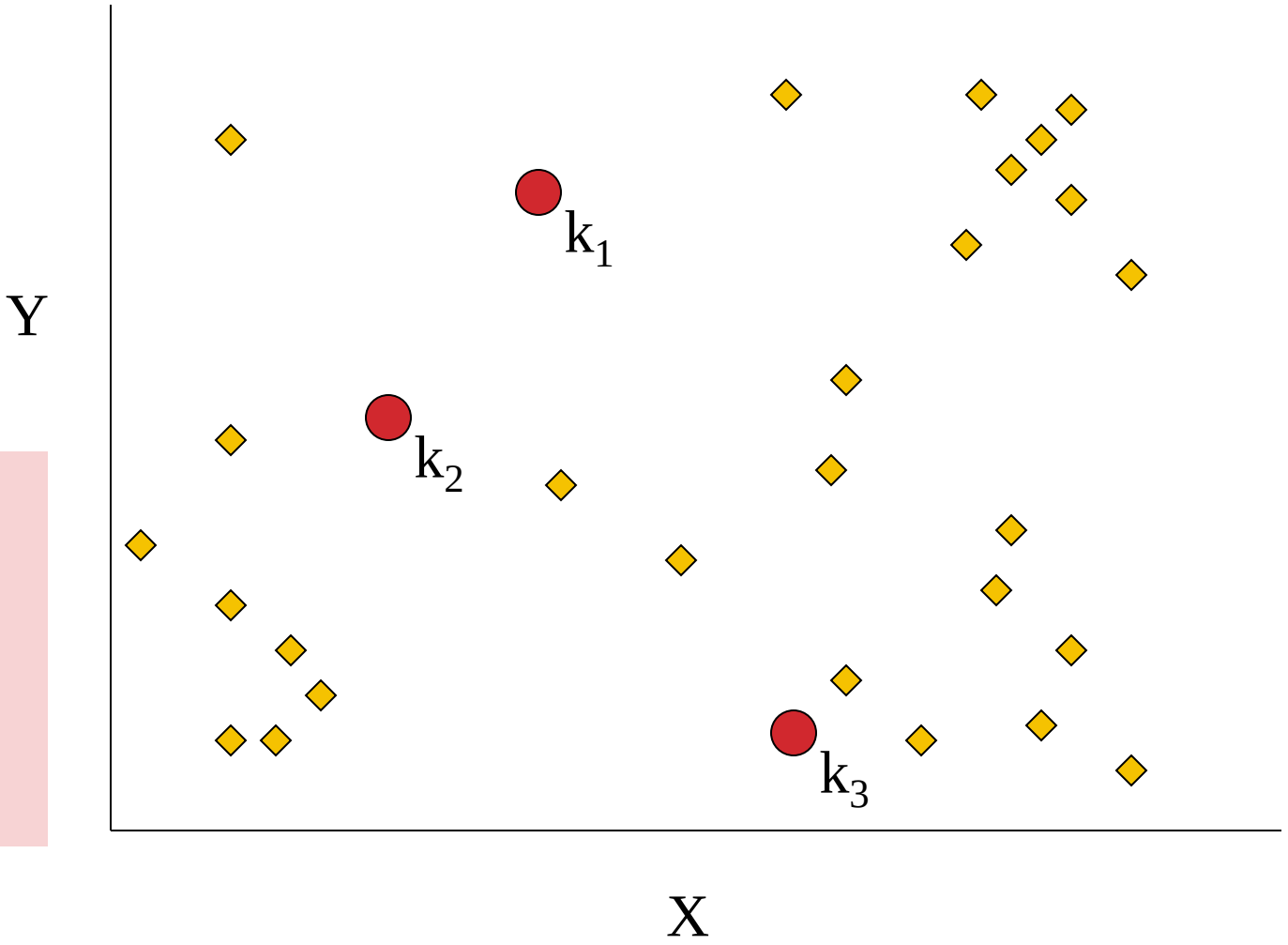


EXEMPLE K-MEANS

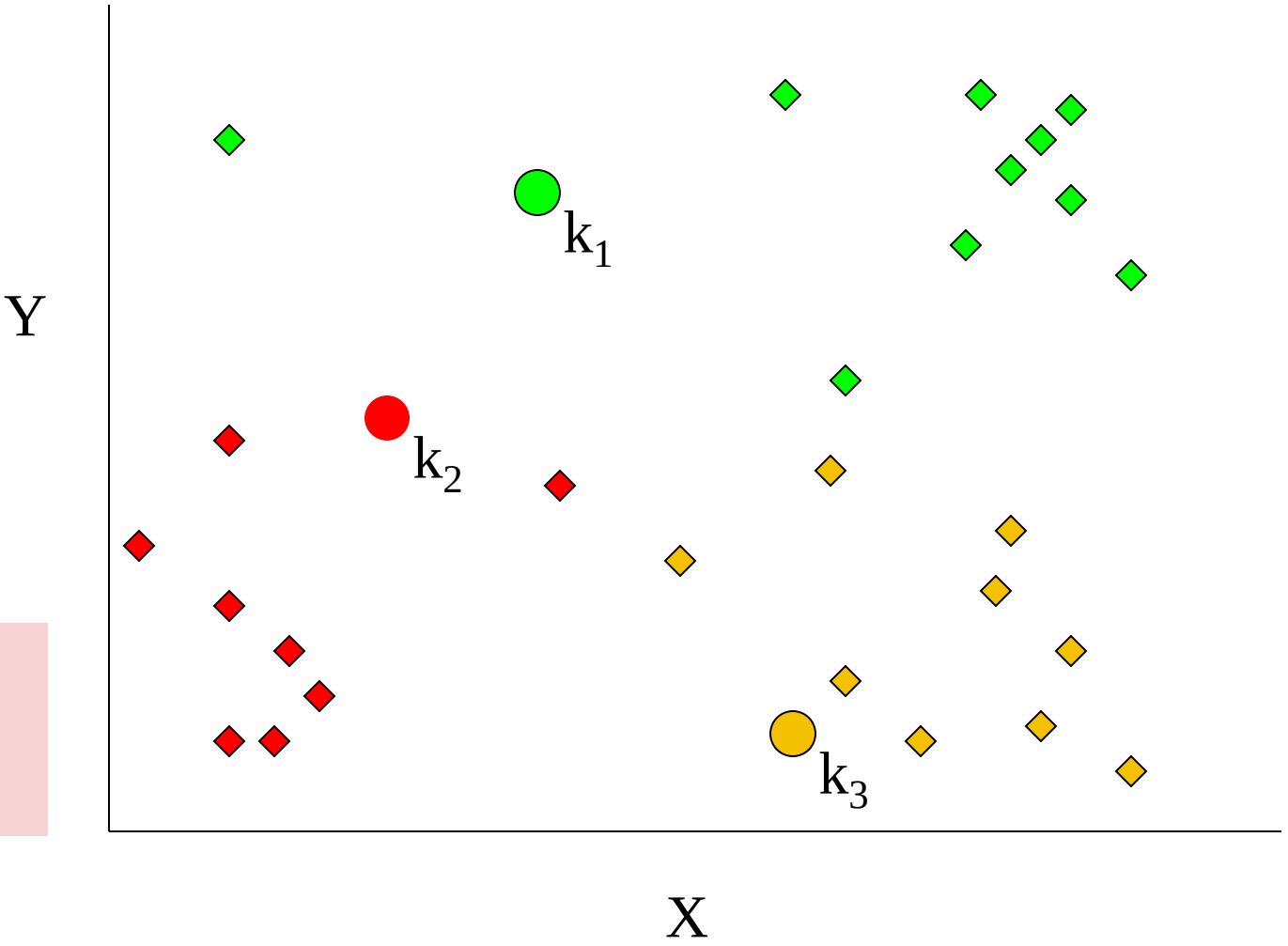


SIMULATION DU K-MEANS 1

Choisir **3** Centres de classes (au hasard)

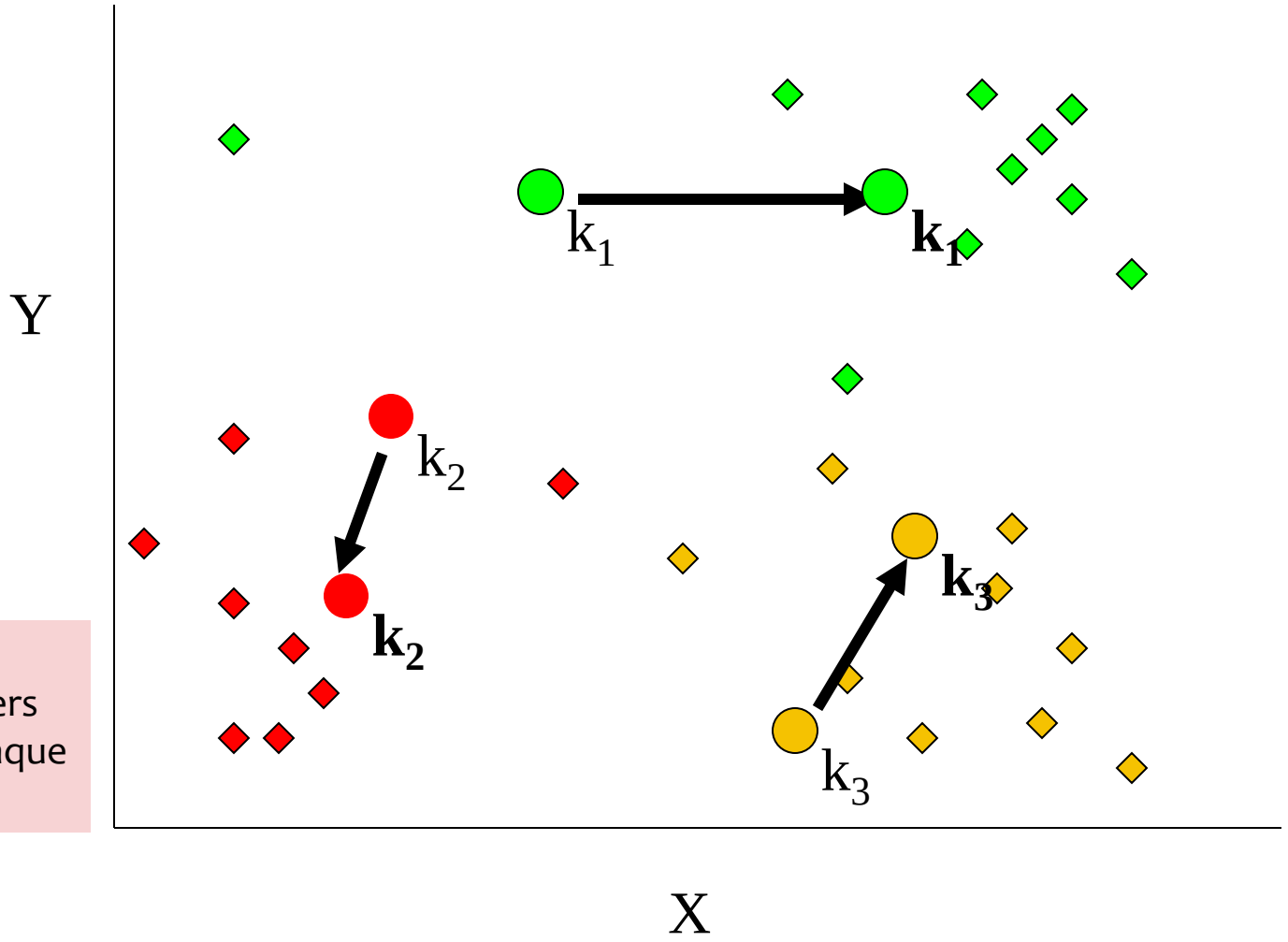


SIMULATION DU K-MEANS 2



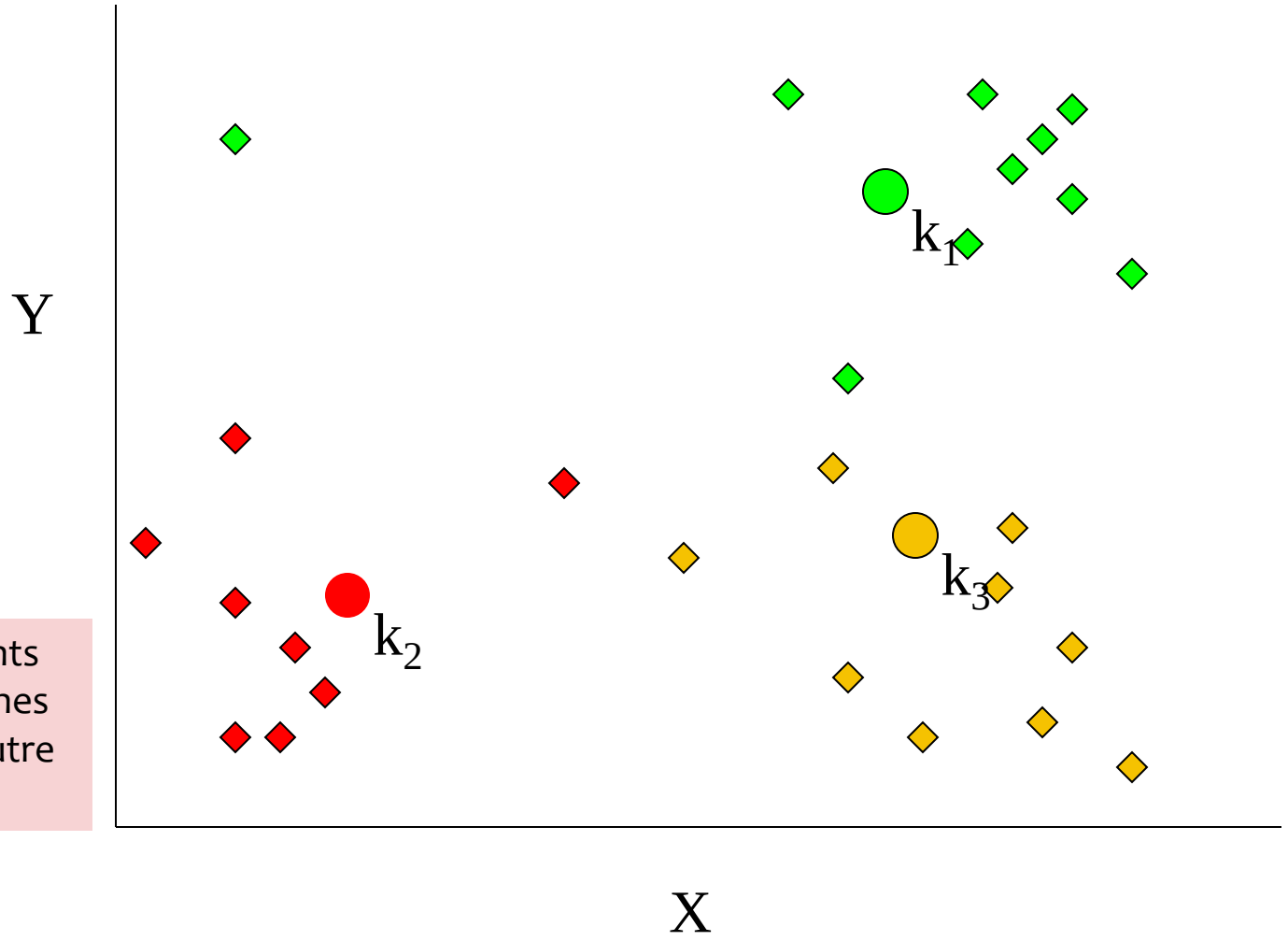
Affecter chaque point à la classe dont le centre est le plus proche

SIMULATION DU K-MEANS 3



Déplacer chaque centre de classe vers la moyenne de chaque classe

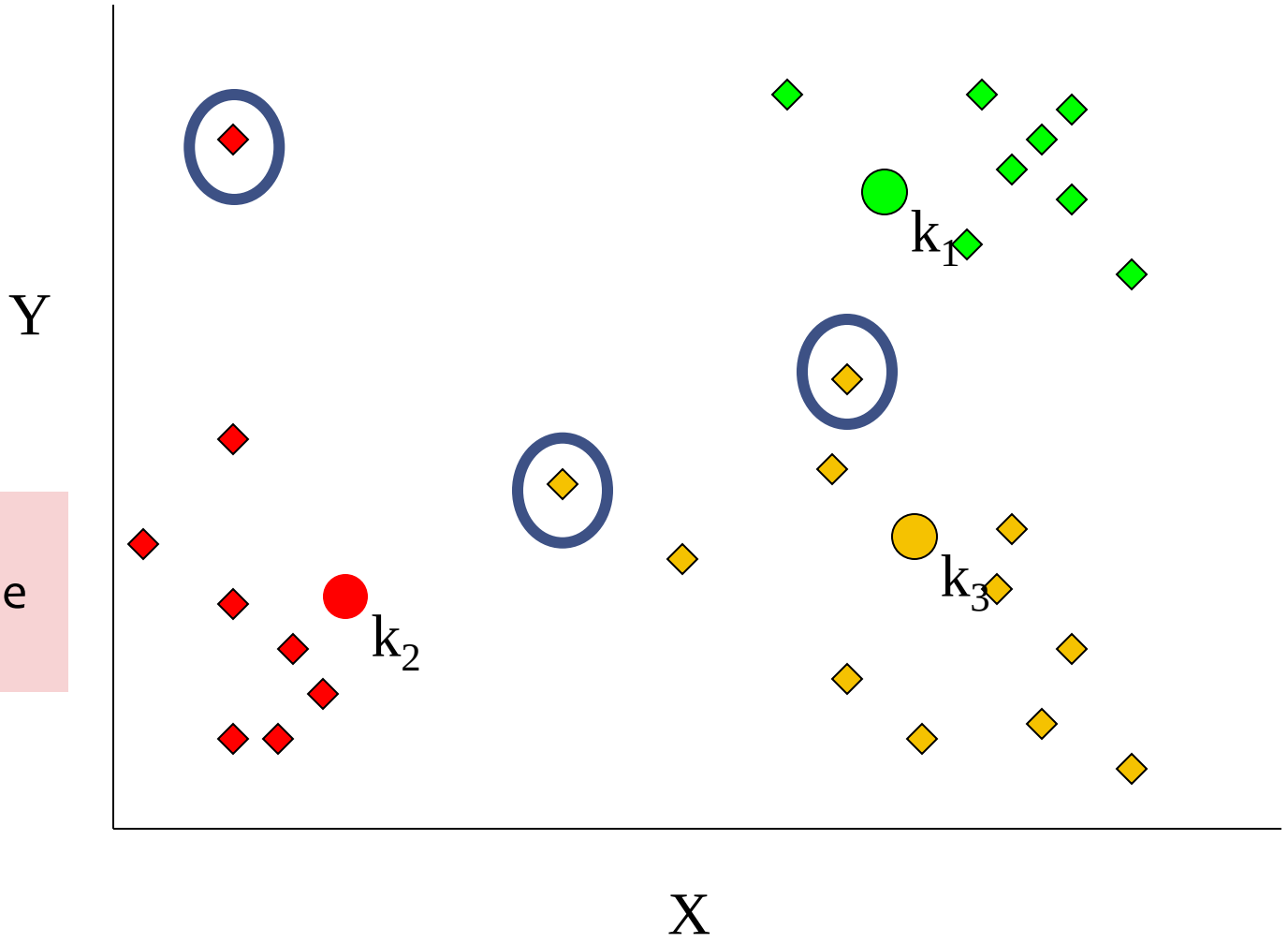
SIMULATION DU K-MEANS 4



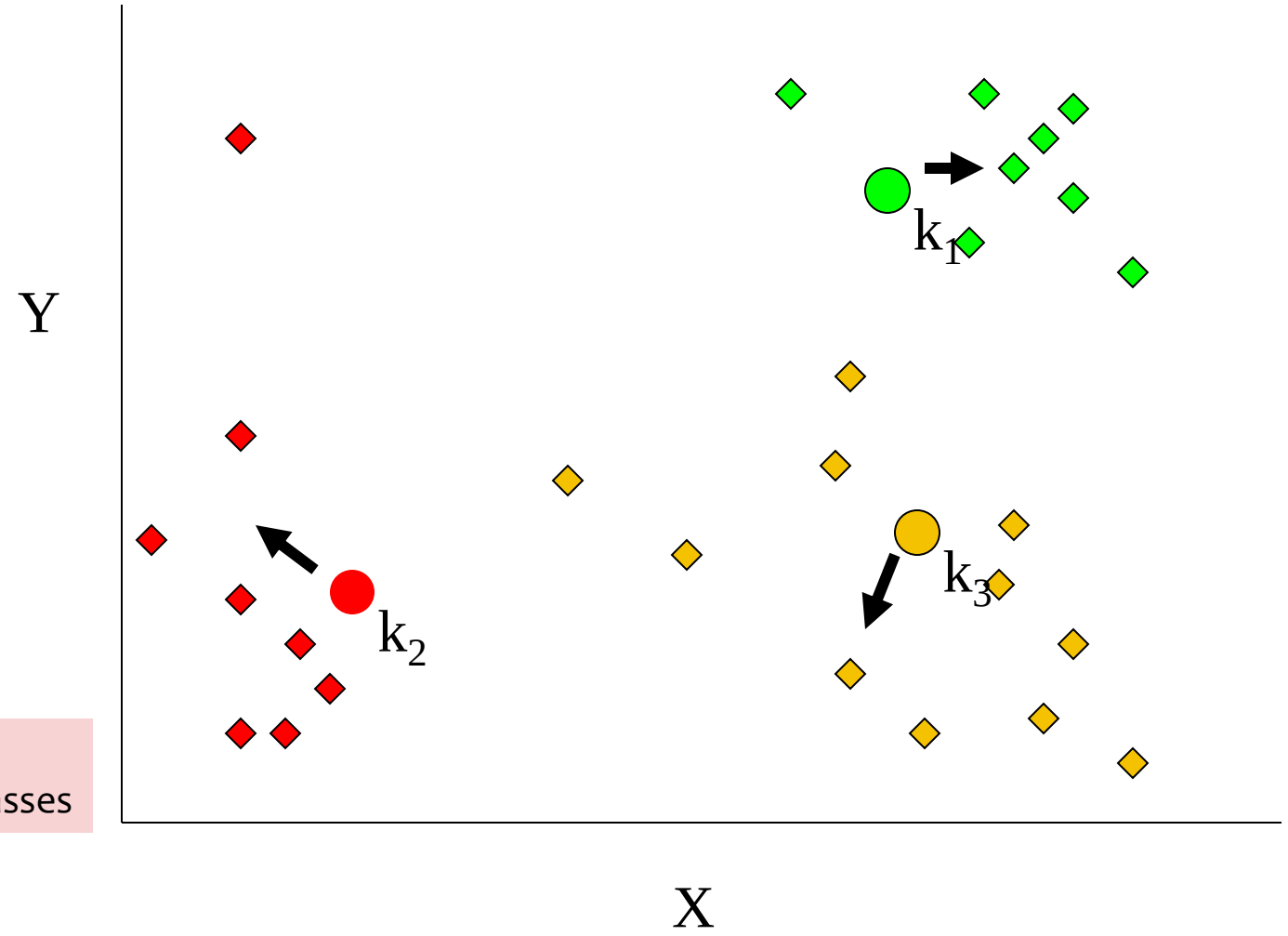
Réaffecter les points qui sont plus proches du centre d'une autre classe

SIMULATION DU K-MEANS 5

les trois points qui changent de classe

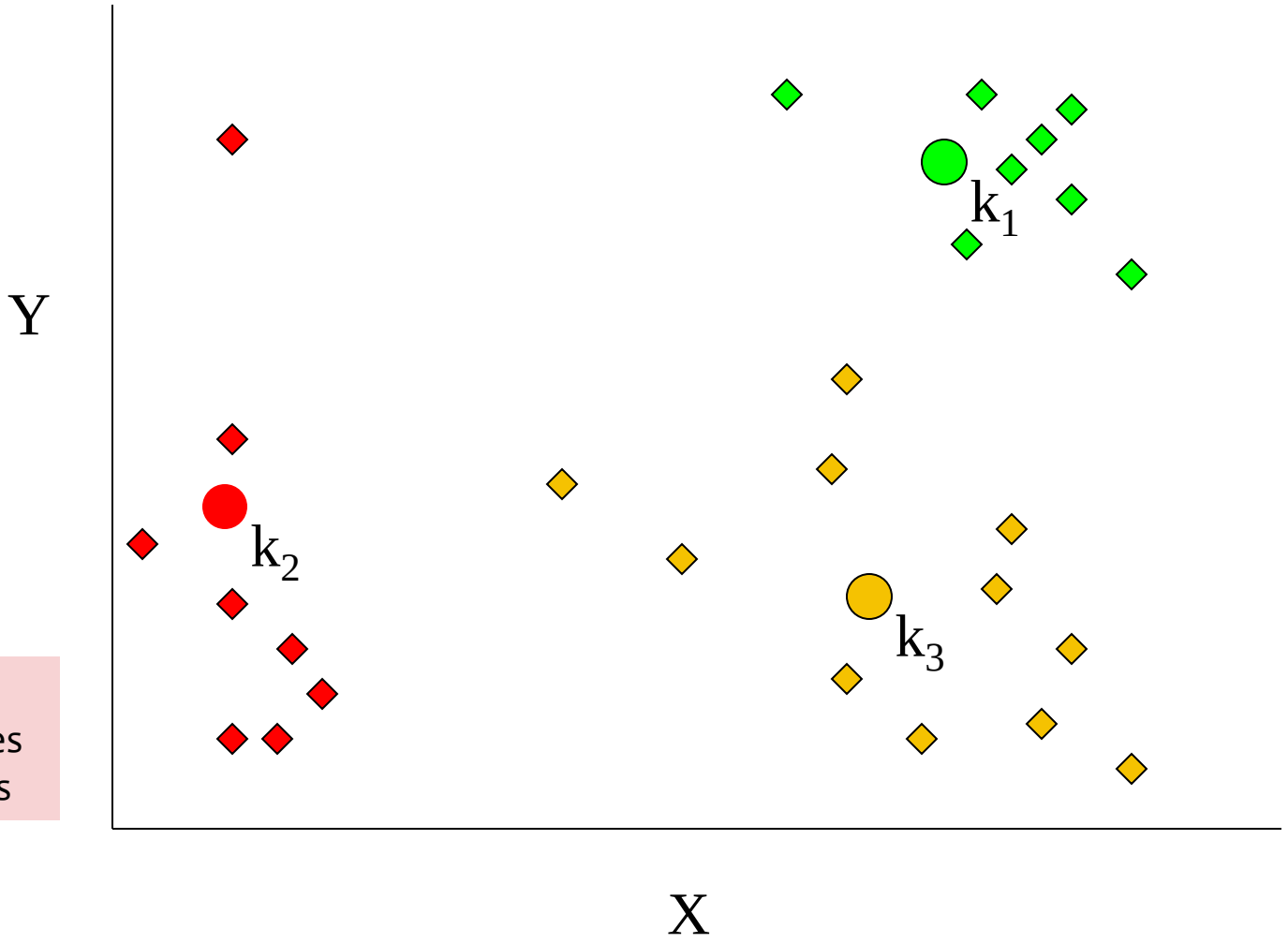


SIMULATION DU K-MEANS 6



Re-calculer les moyennes des classes

SIMULATION DU K-MEANS 7



Déplacer les centres des classes vers les moyennes

ILLUSTRATION K-MEANS

- Soit le tableau1 de **7** individus caractérisés par **2** variables. Tab.1
- On souhaite construire deux groupes homogènes à partir de ces individus.
- On propose de commencer la construction à partir des deux groupes du tableau 2.
- Continuer la construction des groupes en utilisant la distance euclidienne pour mesurer la similarité entre individus.

Subject	A	B
1	1.0	1.0
2	1.5	2.0
3	3.0	4.0
4	5.0	7.0
5	3.5	5.0
6	4.5	5.0
7	3.5	4.5

Tab.1

	Individual	Mean Vector (centroid)
Group 1	1	(1.0, 1.0)
Group 2	4	(5.0, 7.0)

Tab.2

$$d(i,j) = \sqrt{(|x_{i1} - x_{j1}|^2 + |x_{i2} - x_{j2}|^2 + ... + |x_{ip} - x_{jp}|^2)}$$

ILLUSTRATION K-MEANS 1

- Soit le tableau1 de **7** individus caractérisés par **2** variables. Tab.1
- On souhaite construire deux groupes homogènes à partir de ces individus.
- On propose de commencer la construction à partir des deux groupes du tableau 2.
- Continuer la construction des groupes en utilisant la distance euclidienne pour mesurer la similarité entre individus.

	Cluster 1		Cluster 2	
Step	Individual	Mean Vector (centroid)	Individual	Mean Vector (centroid)
1	1	(1.0, 1.0)	4	(5.0, 7.0)
2	1, 2	(1.2, 1.5)	4	(5.0, 7.0)
3	1, 2, 3	(1.8, 2.3)	4	(5.0, 7.0)
4	1, 2, 3	(1.8, 2.3)	4, 5	(4.2, 6.0)
5	1, 2, 3	(1.8, 2.3)	4, 5, 6	(4.3, 5.7)
6	1, 2, 3	(1.8, 2.3)	4, 5, 6, 7	(4.1, 5.4)

	Individual	Mean Vector (centroid)
Cluster 1	1, 2, 3	(1.8, 2.3)
Cluster 2	4, 5, 6, 7	(4.1, 5.4)

$$d(i,j) = \sqrt{(|x_{i1} - x_{j1}|^2 + |x_{i2} - x_{j2}|^2 + ... + |x_{ip} - x_{jp}|^2)}$$

ILLUSTRATION K-MEANS 2

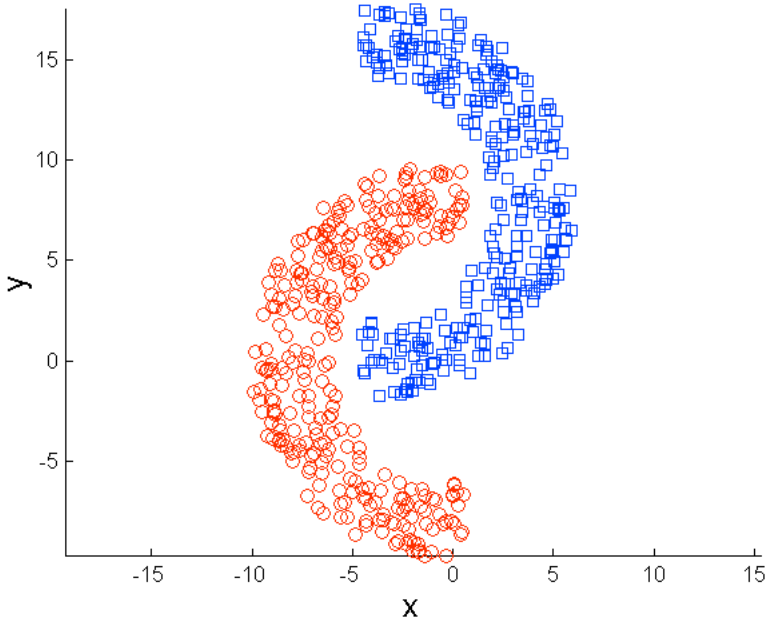
- Soit le tableau1 de **7** individus caractérisés par **2** variables.
- On souhaite construire deux groupes homogènes à partir de ces individus.
- On propose de commencer la construction à partir des deux groupes du tableau 2.
- Continuer la construction des groupes en utilisant la distance euclidienne pour mesurer la similarité entre individus.

Individual	Distance to mean (centroid) of Cluster 1	Distance to mean (centroid) of Cluster 2
1	1.5	5.4
2	0.4	4.3
3	2.1	1.8
4	5.7	1.8
5	3.2	0.7
6	3.8	0.6
7	2.8	1.1

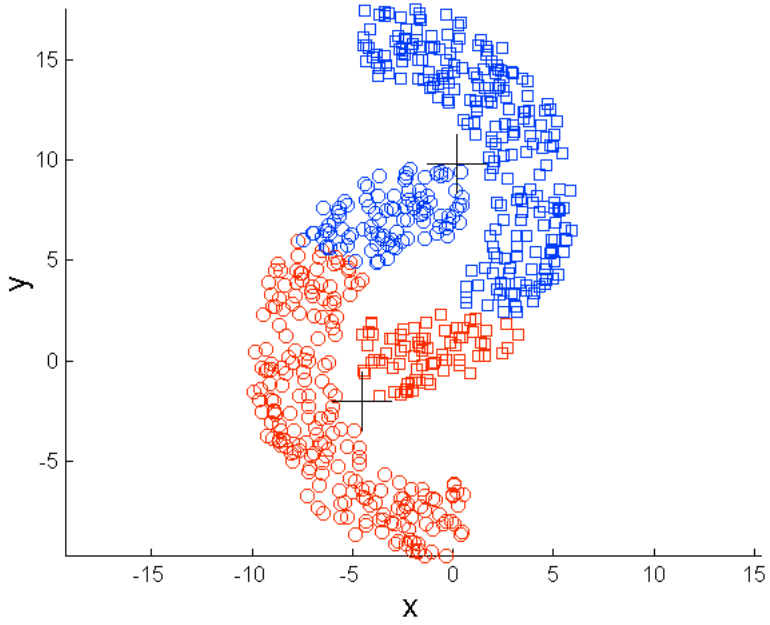
	Individual	Mean Vector (centroid)
Cluster 1	1, 2	(1.3, 1.5)
Cluster 2	3, 4, 5, 6, 7	(3.9, 5.1)

$$d(i,j) = \sqrt{(|x_{i1} - x_{j1}|^2 + |x_{i2} - x_{j2}|^2 + ... + |x_{ip} - x_{jp}|^2)}$$

POINTS FAIBLES DE K-MEANS



Original Points



K-means (2 Clusters)

POINTS FAIBLES DE K-MEANS

- Le choix du nombre de groupes est subjectif dans le cas où le nombre de classes est inconnu au sein de l'échantillon.
- L'algorithme du K-Means ne trouve pas nécessairement la configuration la optimale correspondant à la fonction objective minimale.
- Les résultats de l'algorithme du K-Means sont sensibles à l'initialisation aléatoires des centres.

CAH

CLASSIFICATION ASCENDANTE HIÉRARCHIQUE

PRINCIPE ALGORITHMIQUE

i. Créer à chaque étape une partition obtenue en agrégeant 2 à 2 les éléments les plus proches ! **Eléments :** individus ou groupe d'individus

ii. L'algorithme fournit une hiérarchie de partitions : arbre contenant l'historique de la classification et permettant de retrouver **n-1 partitions**.

iii. Nécessité de se munir d'une **métrique** (distance euclidienne, chi2, Ward…)

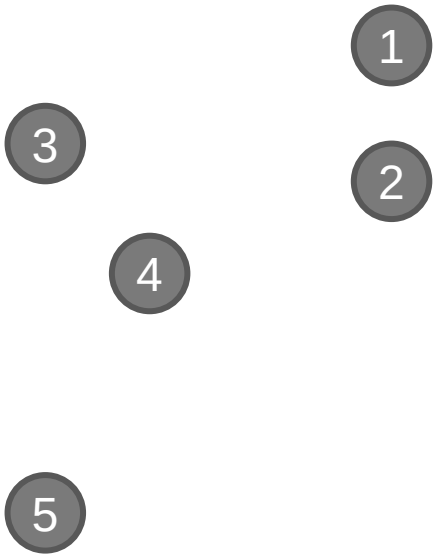
iv. Nécessité de fixer une règle pour agréger un individu et un groupe d'individus (ou bien 2 groupes d'individus)

LE DENDROGRAMME

- Durant les étapes d'un algorithmes de classification hiérarchique, on est en train de construire un dendrogramme.
- Le dendrogramme indique les objets et classes qui ont été fusionnées à chaque itération.
- Le dendrogramme indique aussi la valeur du critère choisi pour chaque partition rencontrée
- Il donne un résumé de la classification hiérarchique
- Chaque palier correspond à une fusion de classes
- Le niveau d'un palier donne une indication sur la qualité de la fusion correspondante
- Toute coupure horizontale correspond à une partition

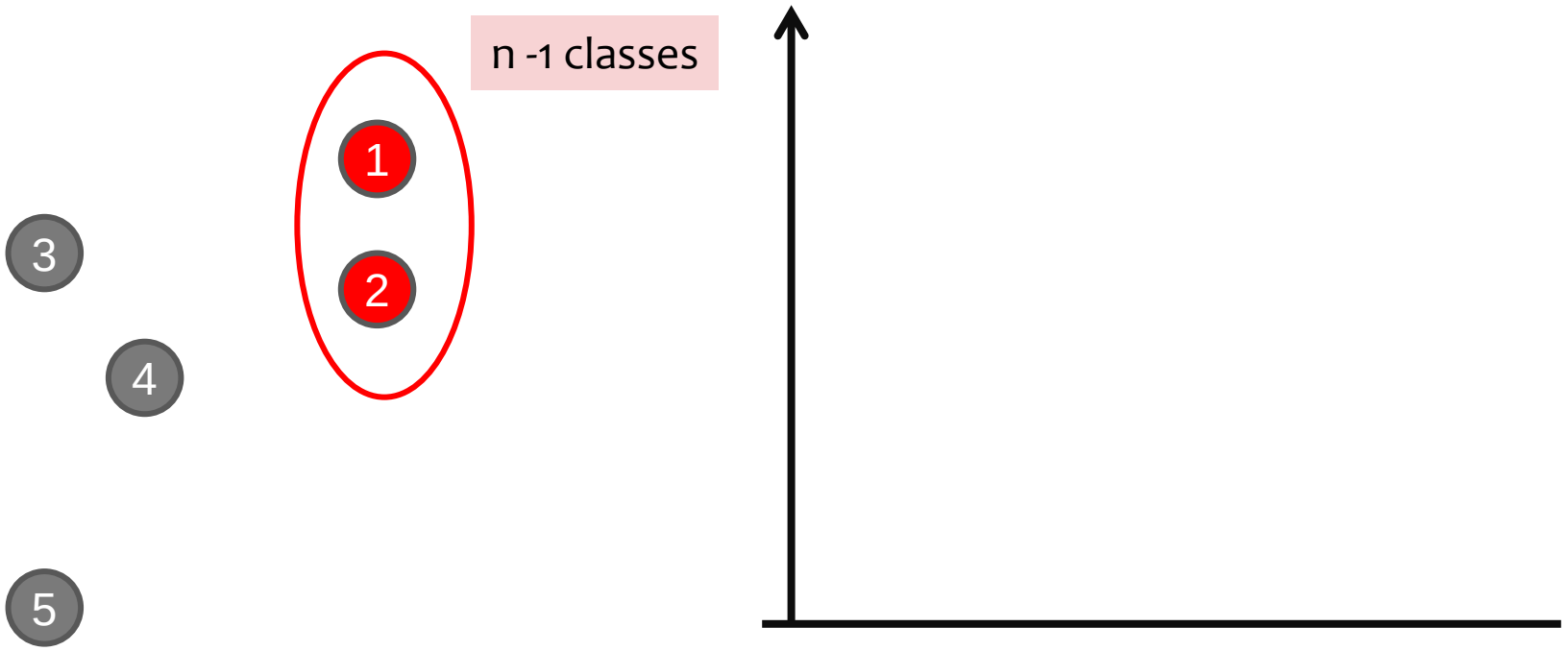
SIMULATION DU CAH 1

n individus / n classes

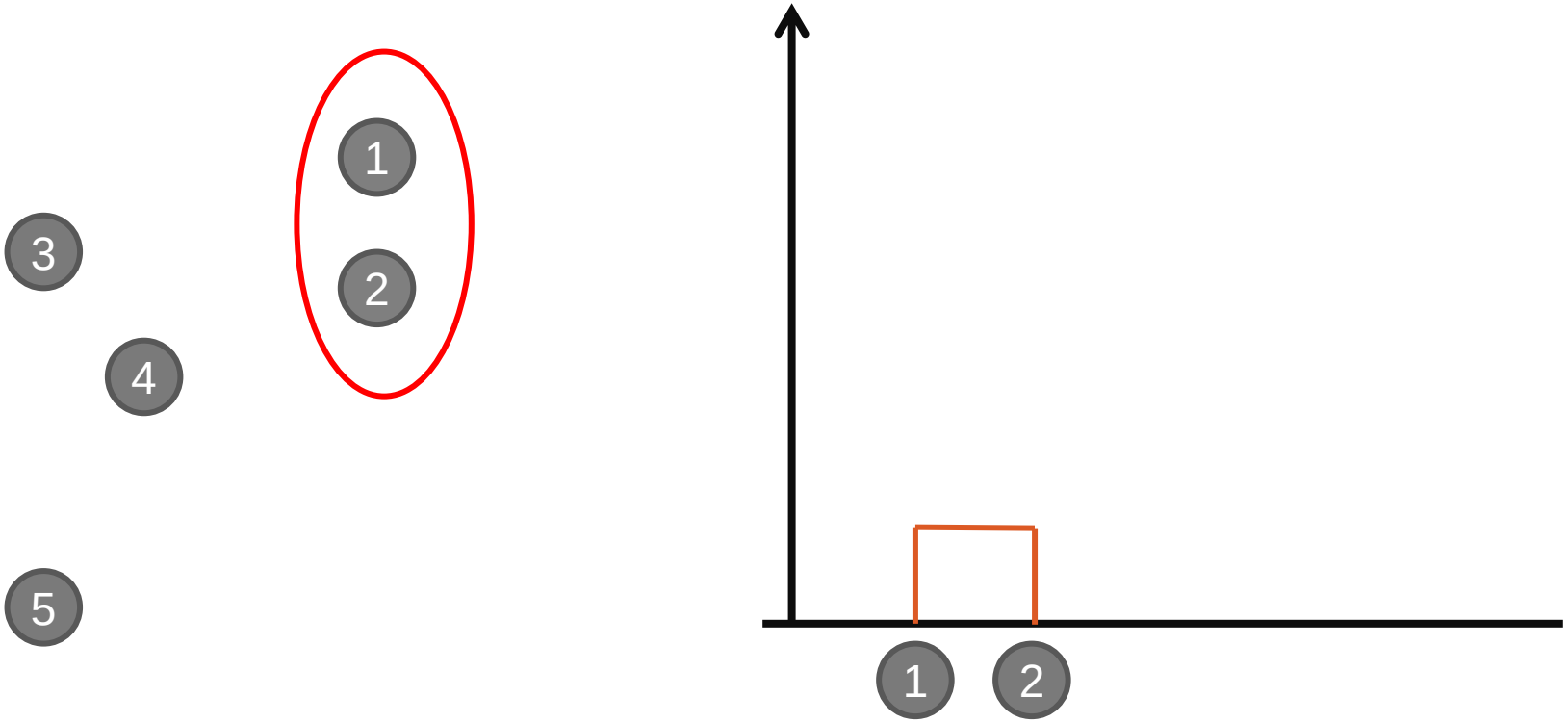


On construit la matrice de distance entre les n éléments
et on regroupe les 2 éléments les plus proches

SIMULATION DU CAH 2



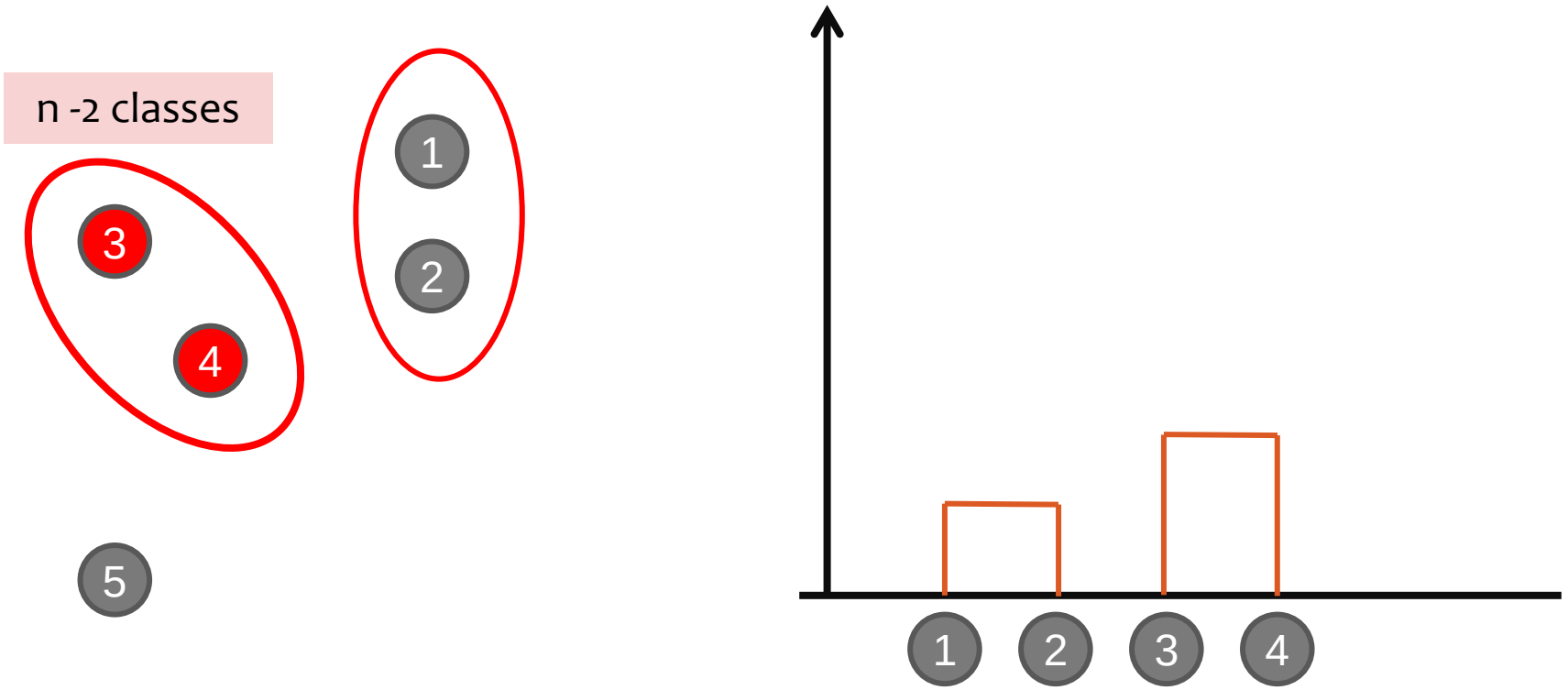
SIMULATION DU CAH 3



Comment mesurer la distance entre une classe et un élément individuel ?

Critères des centres de gravité, de la distance minimale, maximale, critère de Ward…

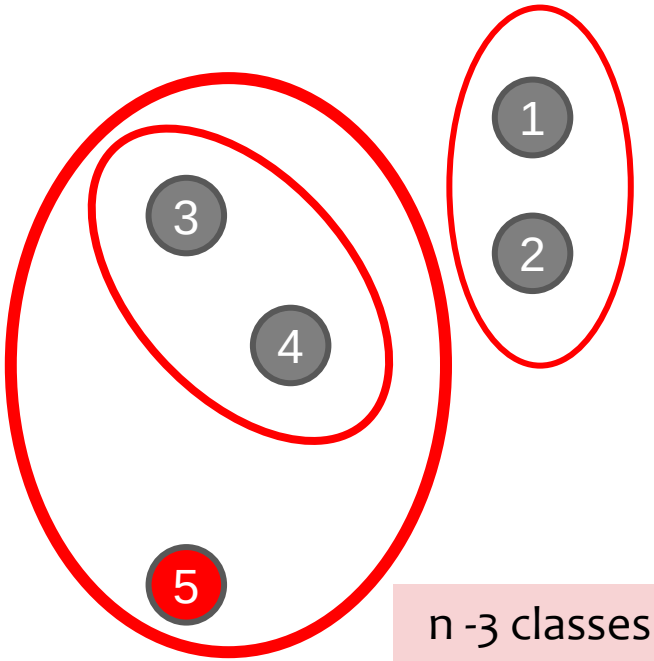
SIMULATION DU CAH 4



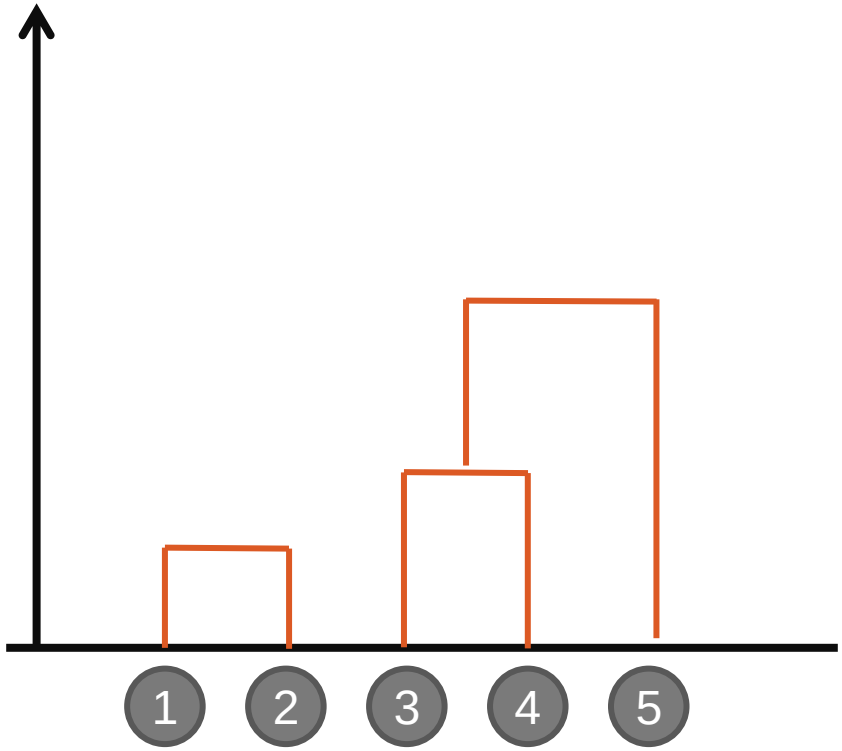
Comment mesurer la distance entre une classe et un élément individuel ?

Critères des centres de gravité, de la distance minimale, maximale, critère de Ward…

SIMULATION DU CAH 5



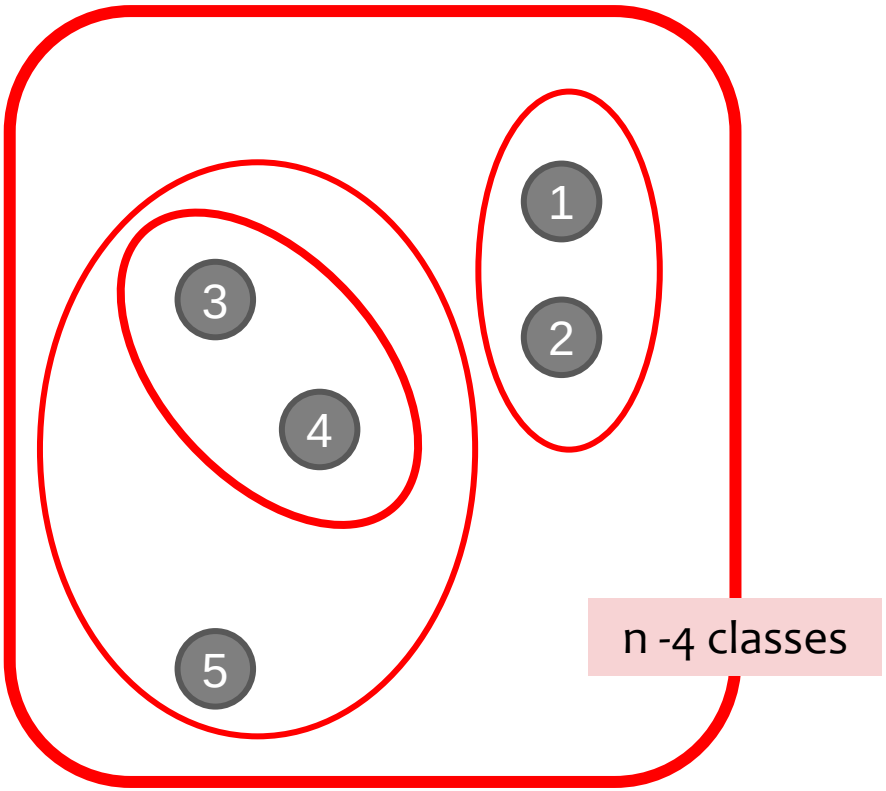
n -3 classes



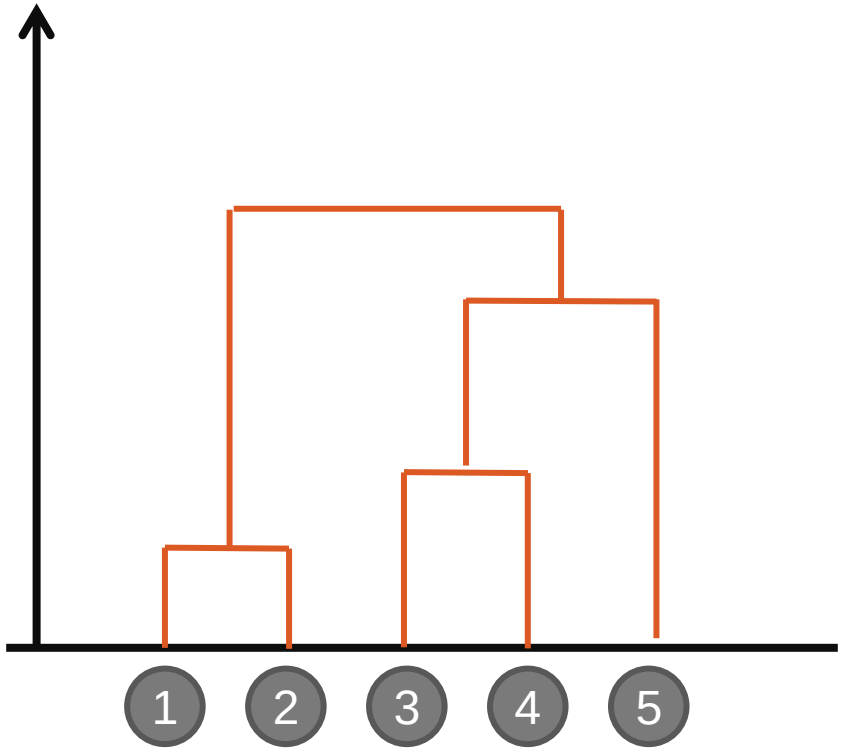
Comment mesurer la distance entre une classe et un élément individuel ?

Critères des centres de gravité, de la distance minimale, maximale, critère de Ward…

SIMULATION DU CAH 6



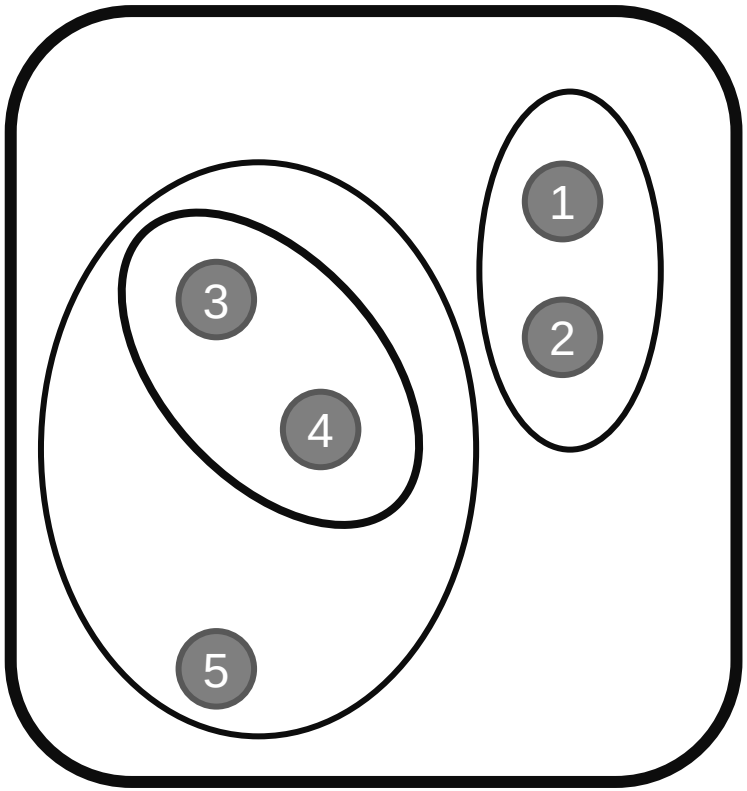
n -4 classes

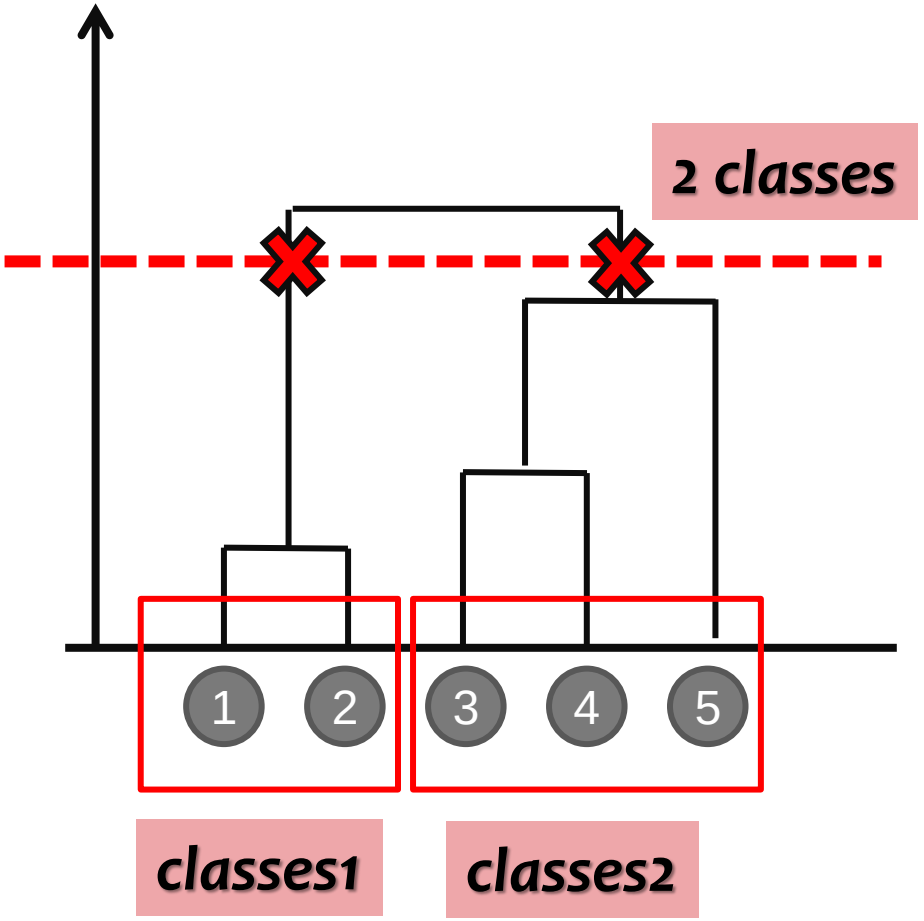


Comment mesurer la distance entre une classe et un élément individuel ?

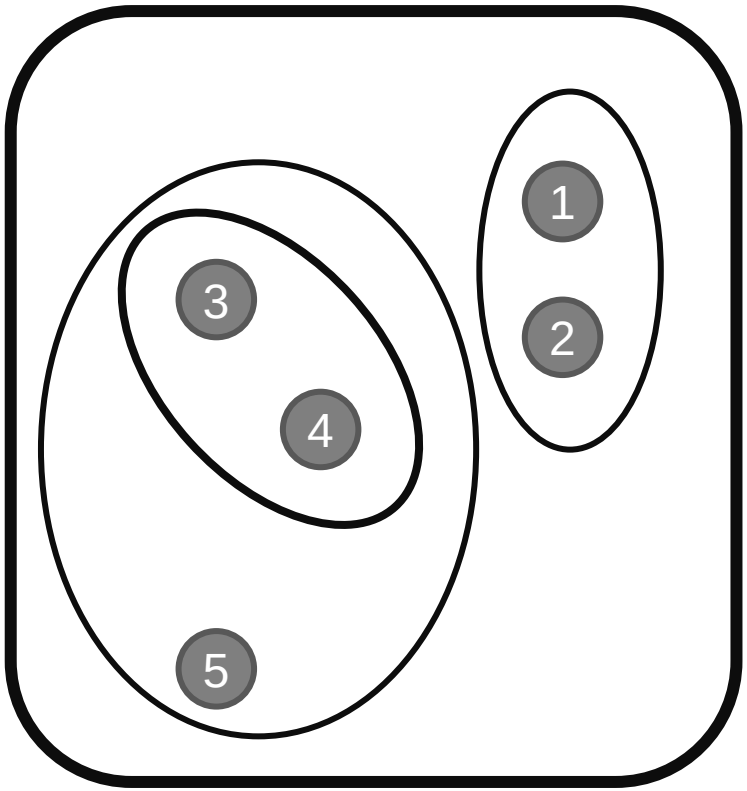
Critères des centres de gravité, de la distance minimale, maximale, critère de Ward…

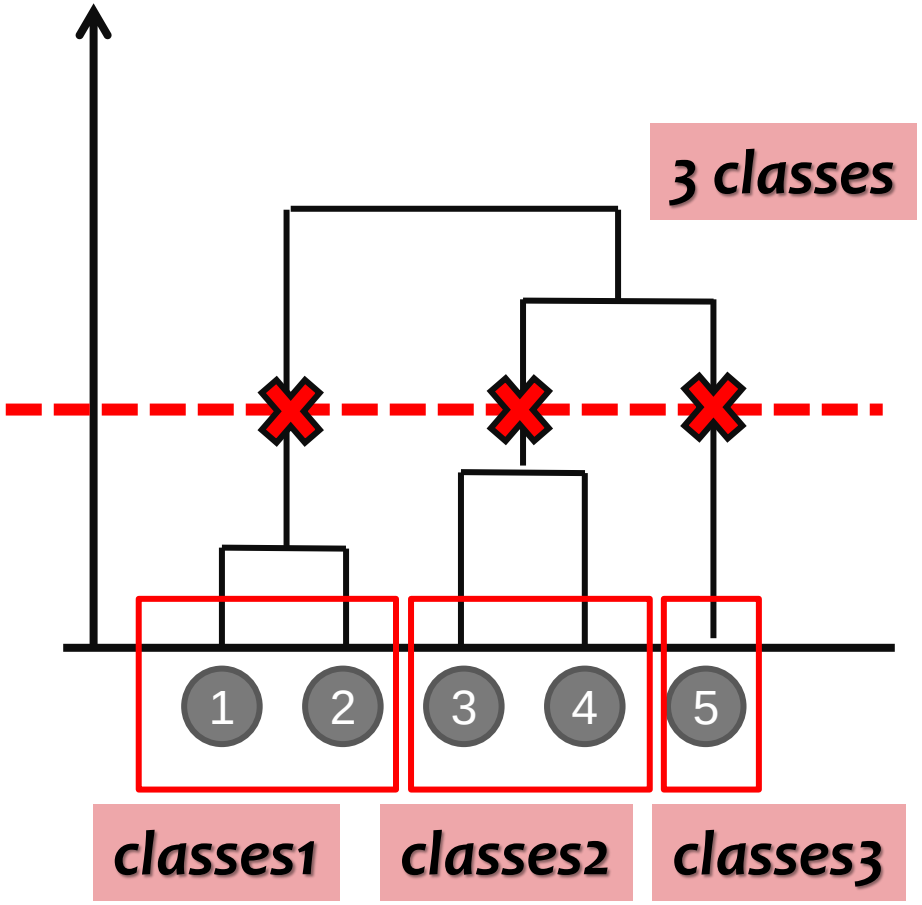
SIMULATION DES CLASSES





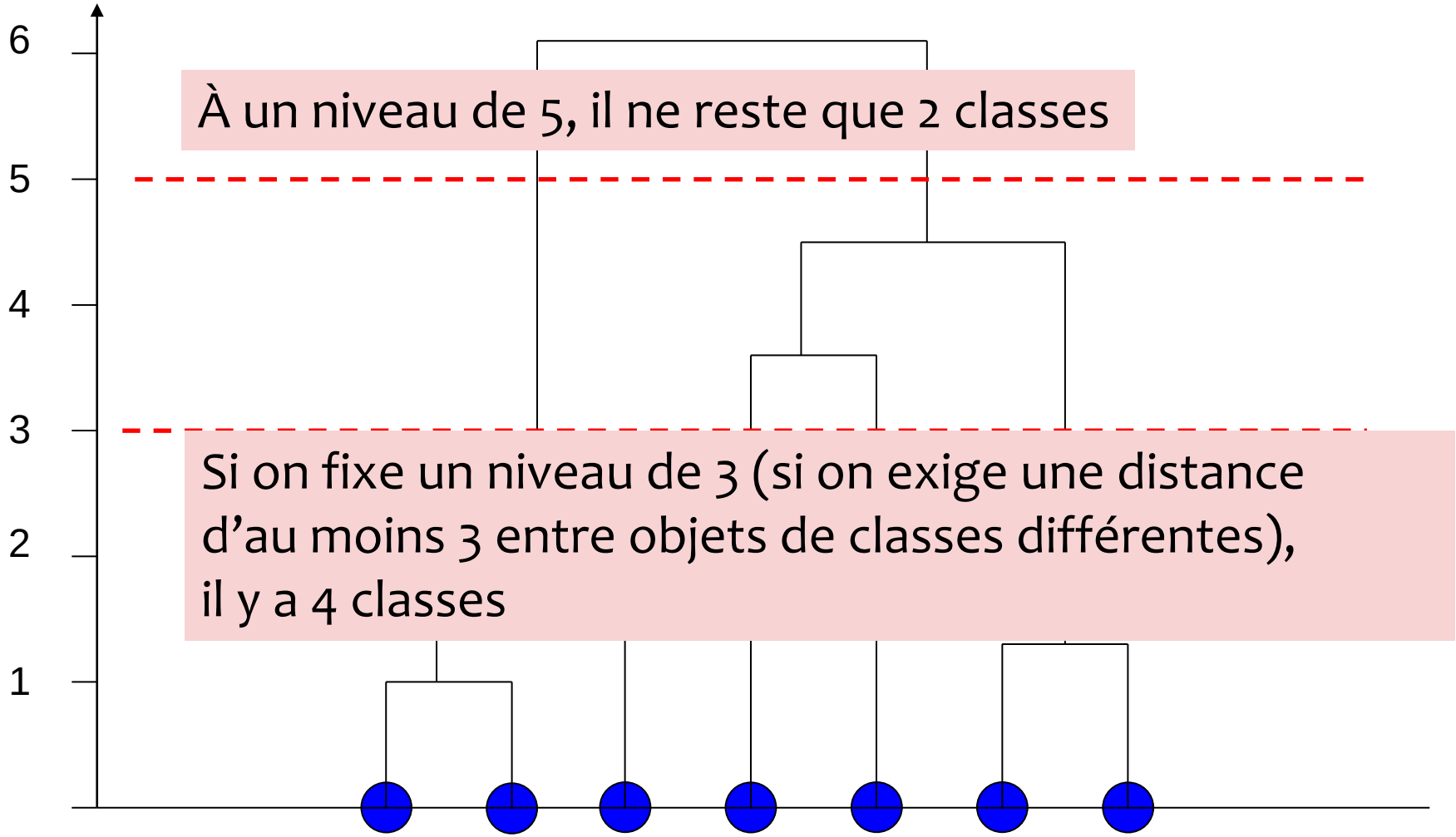
SIMULATION DES CLASSES





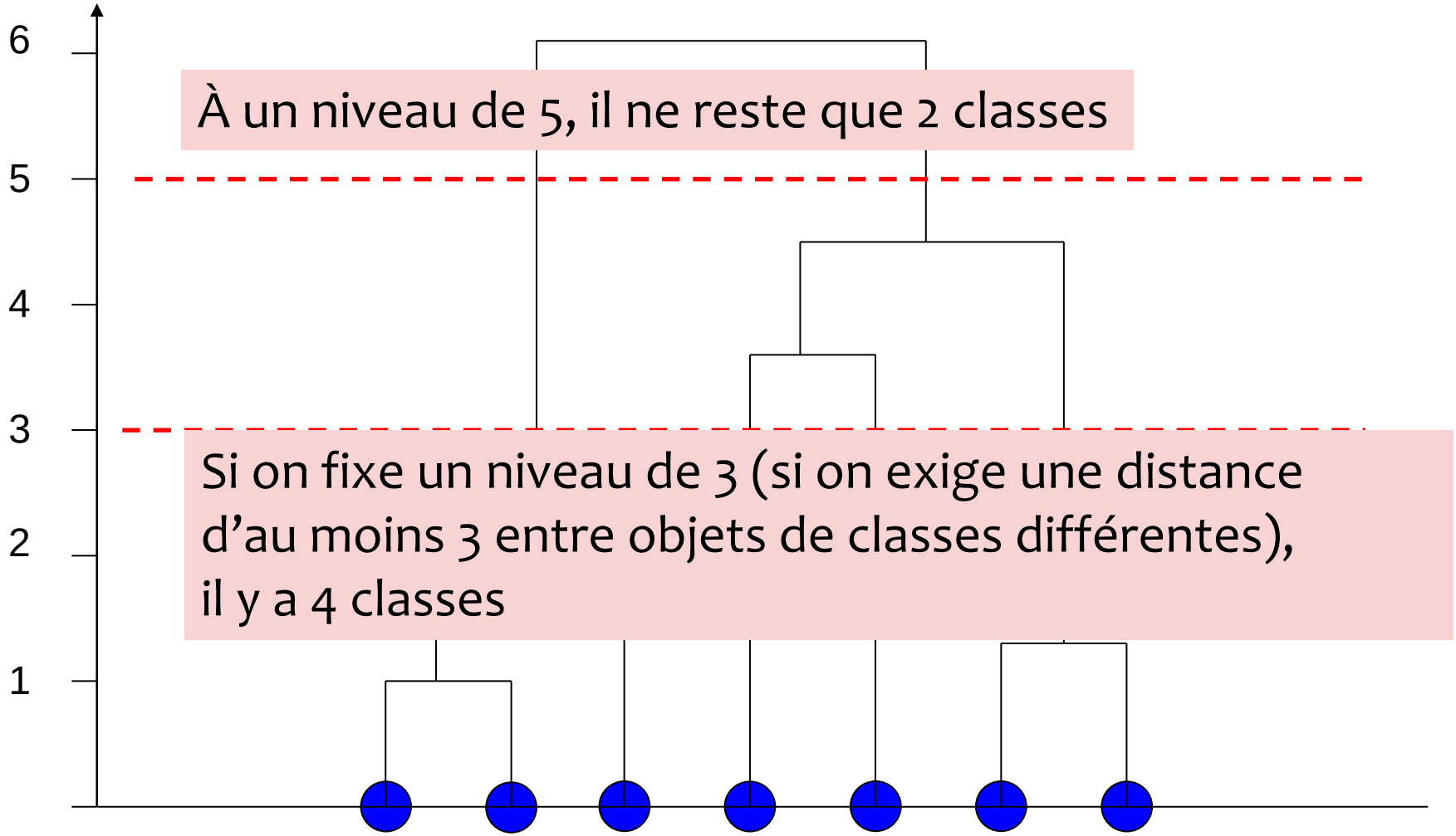
EXEMPLE DE DENDROGRAMME

On « coupe » l'arbre là où les branches sont longues

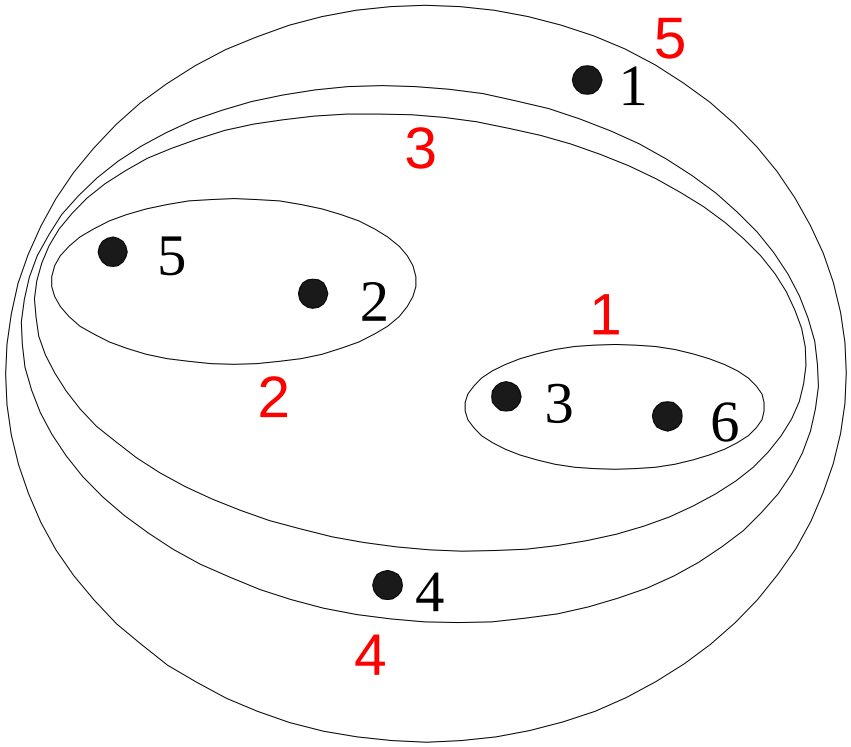


EXEMPLE DE DENDROGRAMME

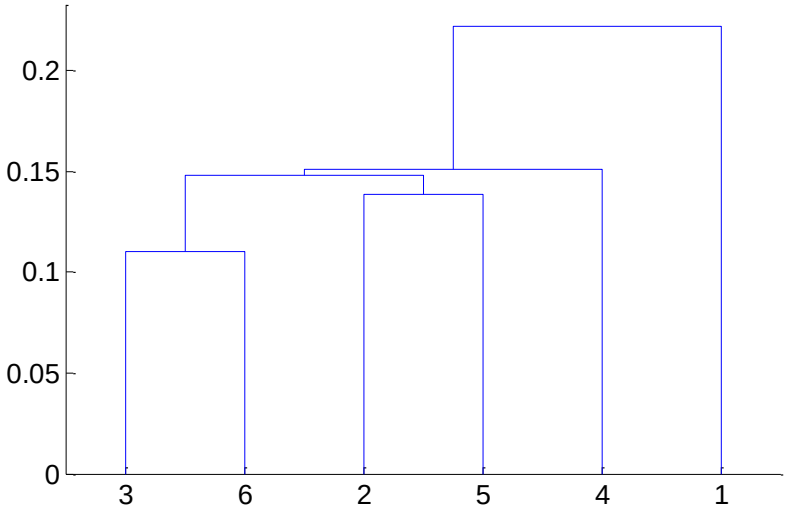
la hauteur d'une branche est proportionnelle à la perte d'inertie interclasse



DENDROGRAMME -> CLUSTERS ?



Nested Clusters



Dendrogram

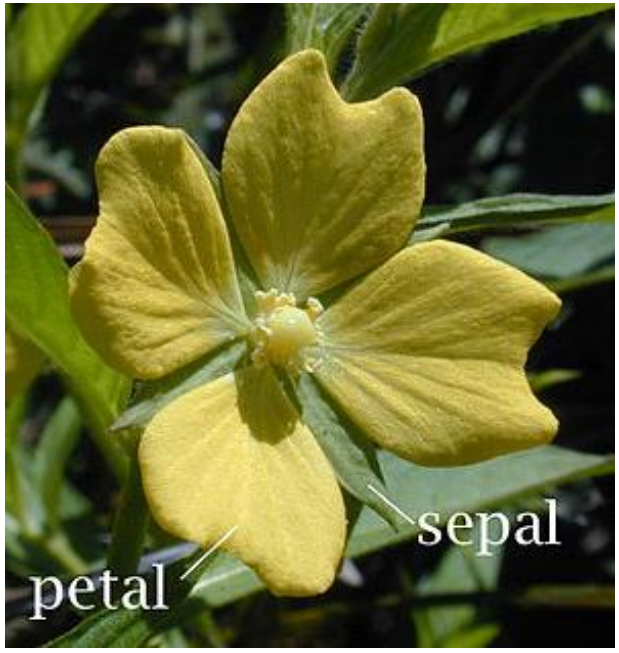
APPLICATION PRATIQUE

SIMULATION KMEANS ET CAH SUR R

APPLICATION K-MEANS « IRIS »

Etudier la qualité des résultats de K-means dans la construction de groupes de fleurs selon leurs caractéristiques.

> iris



RGui - [R Console]

Fichier Edition Voir Misc Packages Fenêtres Aide

```
> iris
    Sepal.Length Sepal.Width Petal.Length Petal.Width   Species
1            5.1         3.5          1.4         0.2    setosa
2            4.9         3.0          1.4         0.2    setosa
3            4.7         3.2          1.3         0.2    setosa
4            4.6         3.1          1.5         0.2    setosa
5            5.0         3.6          1.4         0.2    setosa
6            5.4         3.9          1.7         0.4    setosa
7            4.6         3.4          1.4         0.3    setosa
8            5.0         3.4          1.5         0.2    setosa
9            4.4         2.9          1.4         0.2    setosa
10           4.9         3.1          1.5         0.1    setosa
11           5.4         3.7          1.5         0.2    setosa
12           4.8         3.4          1.6         0.2    setosa
13           4.8         3.0          1.4         0.1    setosa
14           4.3         3.0          1.1         0.1    setosa
15           5.8         4.0          1.2         0.2    setosa
16           5.7         4.4          1.5         0.4    setosa
17           5.4         3.9          1.3         0.4    setosa
18           5.1         3.5          1.4         0.3    setosa
19           5.7         3.8          1.7         0.3    setosa
20           5.1         3.8          1.5         0.3    setosa
21           5.4         3.4          1.7         0.2    setosa
22           5.1         3.7          1.5         0.4    setosa
23           4.6         3.6          1.0         0.2    setosa
24           5.1         3.3          1.7         0.5    setosa
25           4.8         3.4          1.9         0.2    setosa
```

APPLICATION K-MEANS « IRIS »

> iris_for_kmeans<-iris[,1:4]

> km <- kmeans(iris_for_kmeans, 3)

```
> km <- kmeans(iris[,1:4], 3)
> km
K-means clustering with 3 clusters of sizes 62, 50, 38

Cluster means:
  Sepal.Length Sepal.Width Petal.Length Petal.Width
1     5.901613    2.748387     4.393548    1.433871
2     5.006000    3.428000     1.462000    0.246000
3     6.850000    3.073684     5.742105    2.071053

Clustering vector:
  [1] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 1 3 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [76] 1 1 3 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 3 1 3 3 3 3 1 3 3 3 3 3 3 1 1 3 3 3 3 1 3 1 3 1 3 3 1 1 3 3 3 3 3 1 3 3 3 3 1 3 3 3 1 3 3 3 1 3 3 1

Within cluster sum of squares by cluster:
[1] 39.82097 15.15100 23.87947
 (between_SS / total_SS =  88.4 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss" "betweenss"    "size"
> 
```

APPLICATION K-MEANS « IRIS »

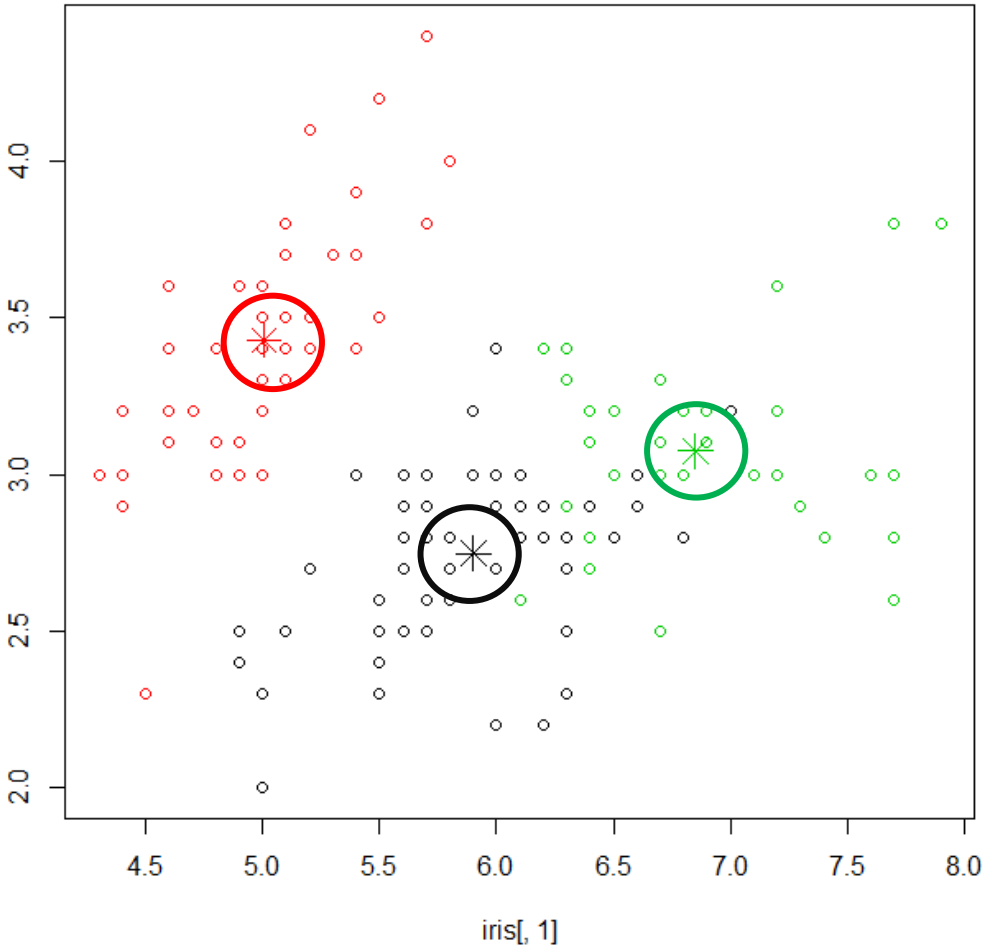
```
> plot(iris[,1], iris[,2], col=km$cluster)
```

```
> points(km$centers[,c(1,)], col=1:3, pch=8, cex=2)
```

```
> table(km$cluster, iris$Species)
```

	setosa	versicolor	virginica
1	0	48	14
2	50	0	0
3	0	2	36

	setosa	versicolor	virginica
Taux de classification	100%	96%	72%
% individus « mal classés »	0%	4%	28%
	10,67 %		



CAH SUR IRIS

Application de la CAH sur la base IRIS en utilisant la distance euclidienne et les 4 variables de longueur et largeur des pétales et des sépales.

1. Calcul de la matrice des distances sur les colonnes de 1 à 4

```
> d_euc = dist(iris[,1:4], method = 'euc')
```

2. Application de la fonction hclust

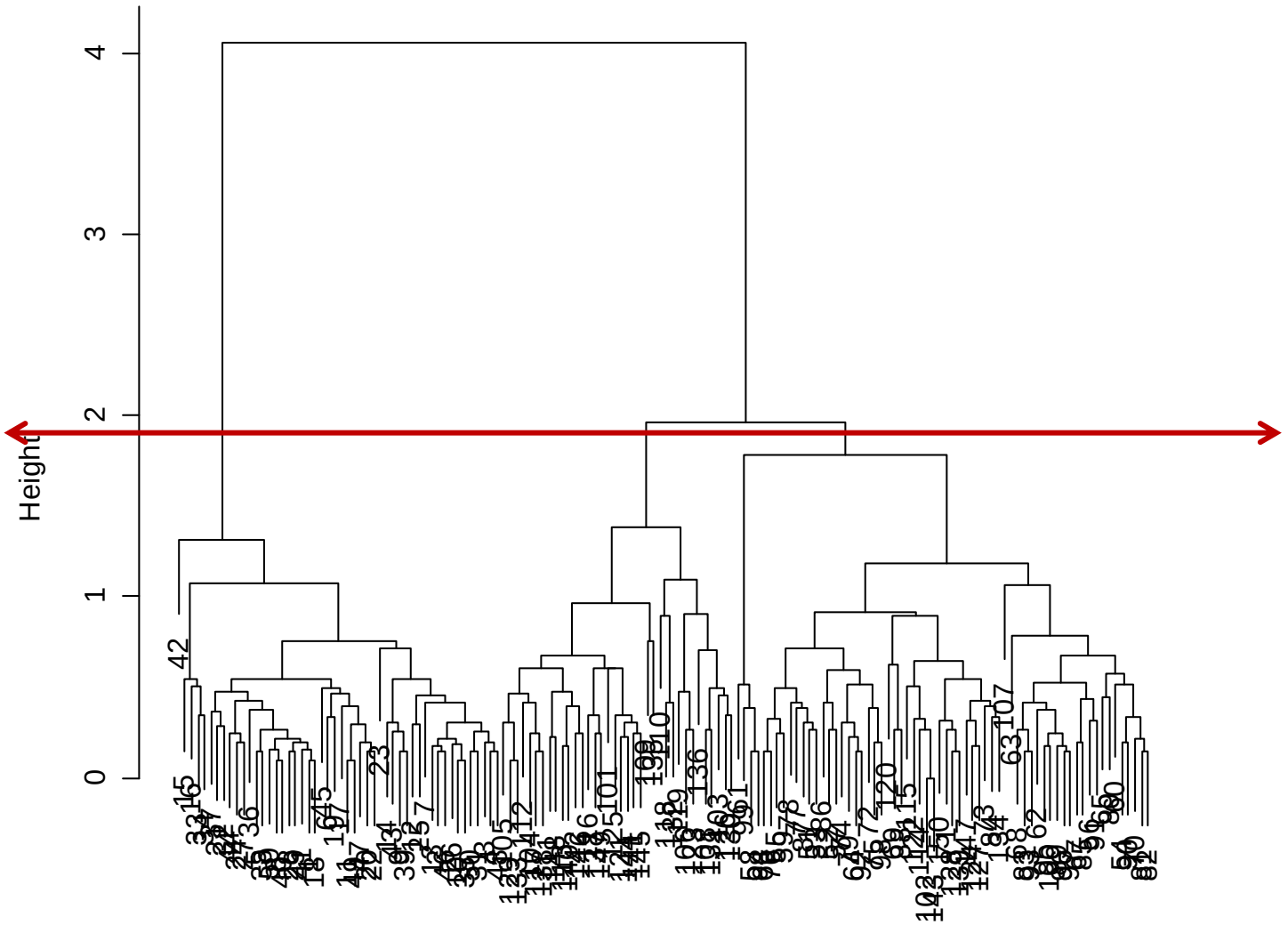
```
> hc = hclust(d_euc,method ='ave')
```

```
> plot (hc)
```

3. Extraire – à partir du dendrogramme – la classification en 3 groupes :

```
> classe<-cutree(hc,3)
```

Cluster Dendrogram



CAH SUR IRIS

Application de la CAH sur la base IRIS en utilisant la distance euclidienne et les 4 variables de longueur et largeur des pétales et des sépales.

1. Calcul de la matrice des distances sur les colonnes de 1 à 4

```
> d_euc = dist(iris[,1:4], method = 'euc')
```

2. Application de la fonction hclust

```
> hc = hclust(d_euc,method ='ave')
```

```
> plot (hc)
```

3. Extraire – à partir du dendrogramme – la classification en 3 groupes :

```
> classe<-cutree(hc,3)
```

APPLICATION CAH « IRIS »

> table(classe , iris$Species)

	setosa	versicolor	virginica
1	0	50	14
2	50	0	0
3	0	0	36

	setosa	versicolor	virginica
Taux de classification	100%	100%	72%
% individus « mal classés »	0%	0%	28%
	9,33 %		

CAH

	setosa	versicolor	virginica
1	0	48	14
2	50	0	0
3	0	2	36

	setosa	versicolor	virginica
Taux de classification	100%	96%	72%
% individus « mal classés »	0%	4%	28%
	10,67 %		

Kmeans

AVANTAGES DE LA *CAH*

- Permet de classer : des individus, des variables, des moyennes de classes obtenues en sortie d'un algorithme des centres mobiles
- si on classe des moyennes, on améliore les résultats si on connaît non seulement les moyennes des classes, mais aussi les inerties intraclasse et les effectifs des classes
- S'adapte aux diverses formes de classes, par le choix de la distance
- Permet de choisir le nombre de classes de façon optimale, grâce à des indicateurs de qualité de la classification en fonction du nombre de classes

INCONVÉNIENTS DE LA CAH

- Complexité algorithmique non linéaire (en n^2 ou n^3, parfois $n^2\log(n)$)

- Deux observations placées dans des classes différentes ne sont jamais plus comparées

OBJECTIFS DES TECHNIQUES DESCRIPTIVES

visent à mettre en évidence des informations présentes mais cachées par le volume des données

il n'y a pas de variable « cible » à prédire

projection du nuage de points sur un espace de dimension inférieure pour obtenir une visualisation de l'ensemble des liaisons entre : Individus, Variables… tout en minimisant la perte d'information

trouver dans l'espace de travail des groupes homogènes d'individus ou de variables

détection d'associations entre des objets