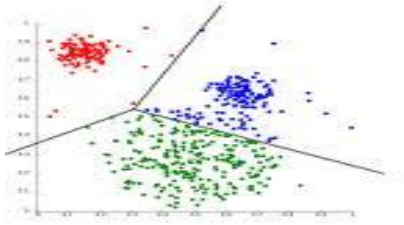
I. MACHINE LEARNING
DATA MINING : MÉTHODES DESCRIPTIVES

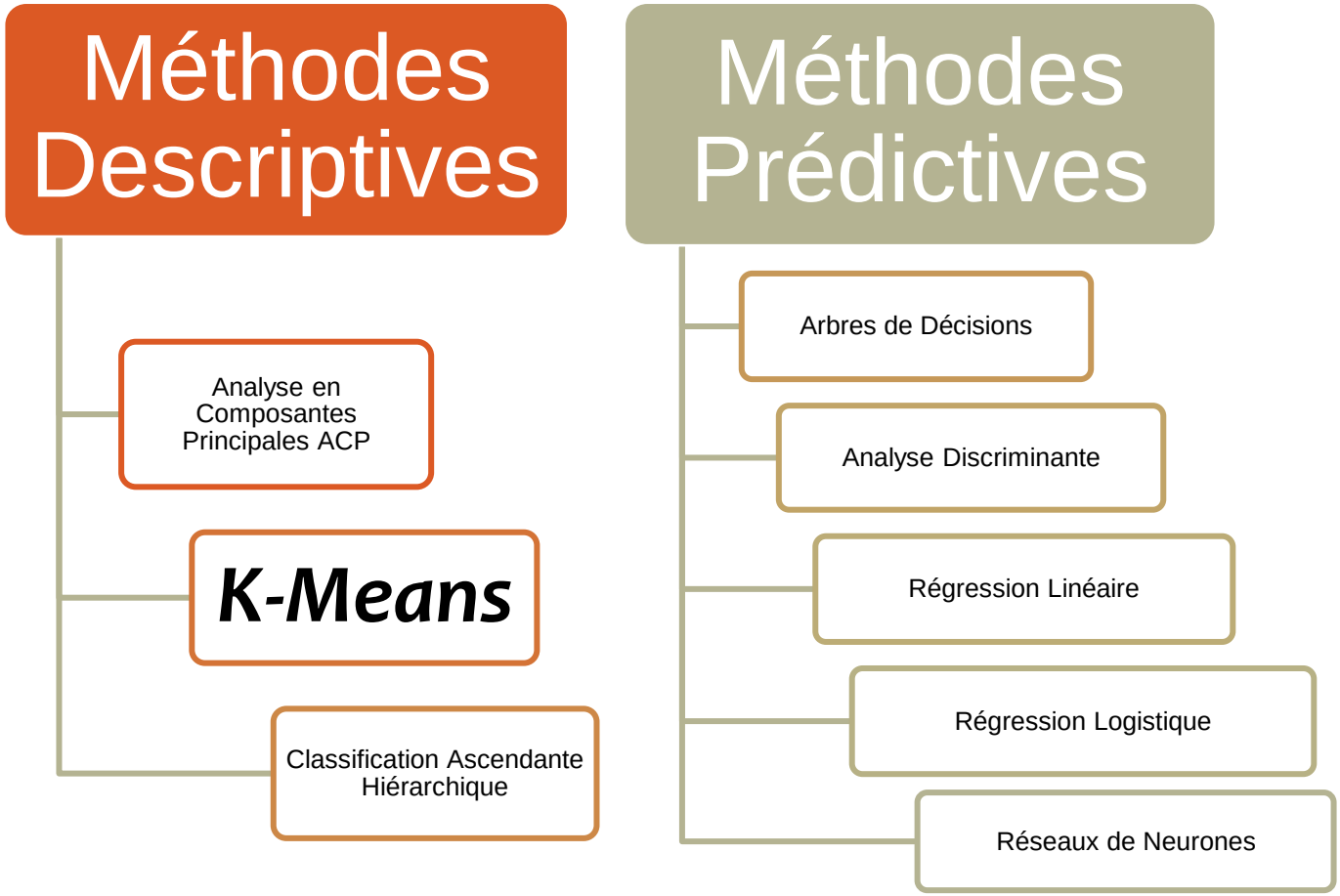
K-MEANS

MÉTHODE DES CENTRES MOBILES



Mohamed Heny SELMI

DEUX FAMILLES DE TECHNIQUES



OBJECTIFS DES TECHNIQUES DESCRIPTIVES DE CLASSIFICATION

- ✓ Distinguer des sous-ensembles (ou classes) distincts dans la population de départ.
- ✓ la classification se distingue du classement par le fait que les critères de classification ne sont pas connus avant étude de la population : C'est la population qui détermine les critères.
- ✓ La classification est le plus souvent un préalable à d'autres opérations de data mining.
- ✓ La classification permet de limiter le nombre de variables par sous-ensemble.
- ✓ La classification permet de rechercher des corrélations propres à chaque classe et donc plus précises.
- ✓ il n'existe pas une solution unique au problème de la classification. Autrement dit, il n'y a pas « LA » bonne classification, mais plusieurs classifications possibles.
- ✓ Regrouper les objets en groupes, classes, familles, segments, clusters, de sorte que :
  - Tous deux objets d'un même groupe se ressemblent le plus.
  - Tous deux objets de groupes différents se distinguent le plus.
  - Le nombre de groupes est parfois fixé.
- ✓ visent à synthétiser des informations présentes complexes mais cachées par le volume des données
- ✓ il n'y a pas de variable « cible » à prédire

EXEMPLE D'APPLICATION MARKETING

- L'intérêt de la classification en marketing est de définir les profils de client.
- Chaque classe « résume » une clientèle ce qui permet une communication spécifique, « one-to-one ».
- Les classes permettent de se constituer un échantillon représentatif permanent de personnes ou de classes de personnes que l'on interroge régulièrement s
- On parle de segmentation, de typologie, d'analyse typologique ou de Clustering à la place de classification.
- On parle de classe, de segment ou de cluster pour parler tant de l'extension (les individus) que de l'intension (les variables et leurs valeurs possibles) des sous-ensembles définis par la classification.
- On parle de typologie ou de type pour parler de l'intension (les variables et leurs valeurs possibles).

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

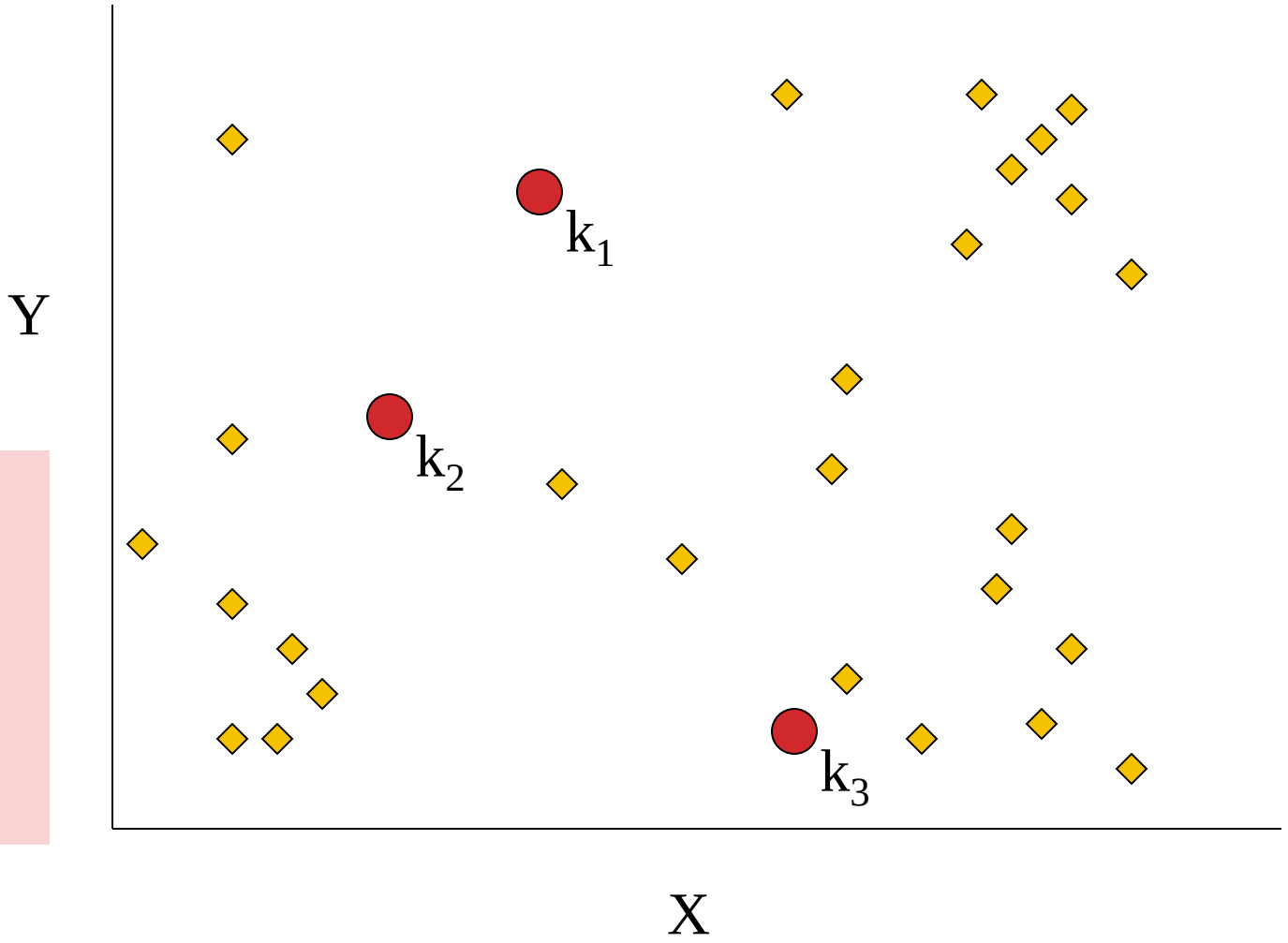
Inertie totale I_{tot} : somme de l'inertie intraclasse I_A et de l'inertie interclasse I_c

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

Inertie intraclasse I_A : somme des inerties totales de chaque classe

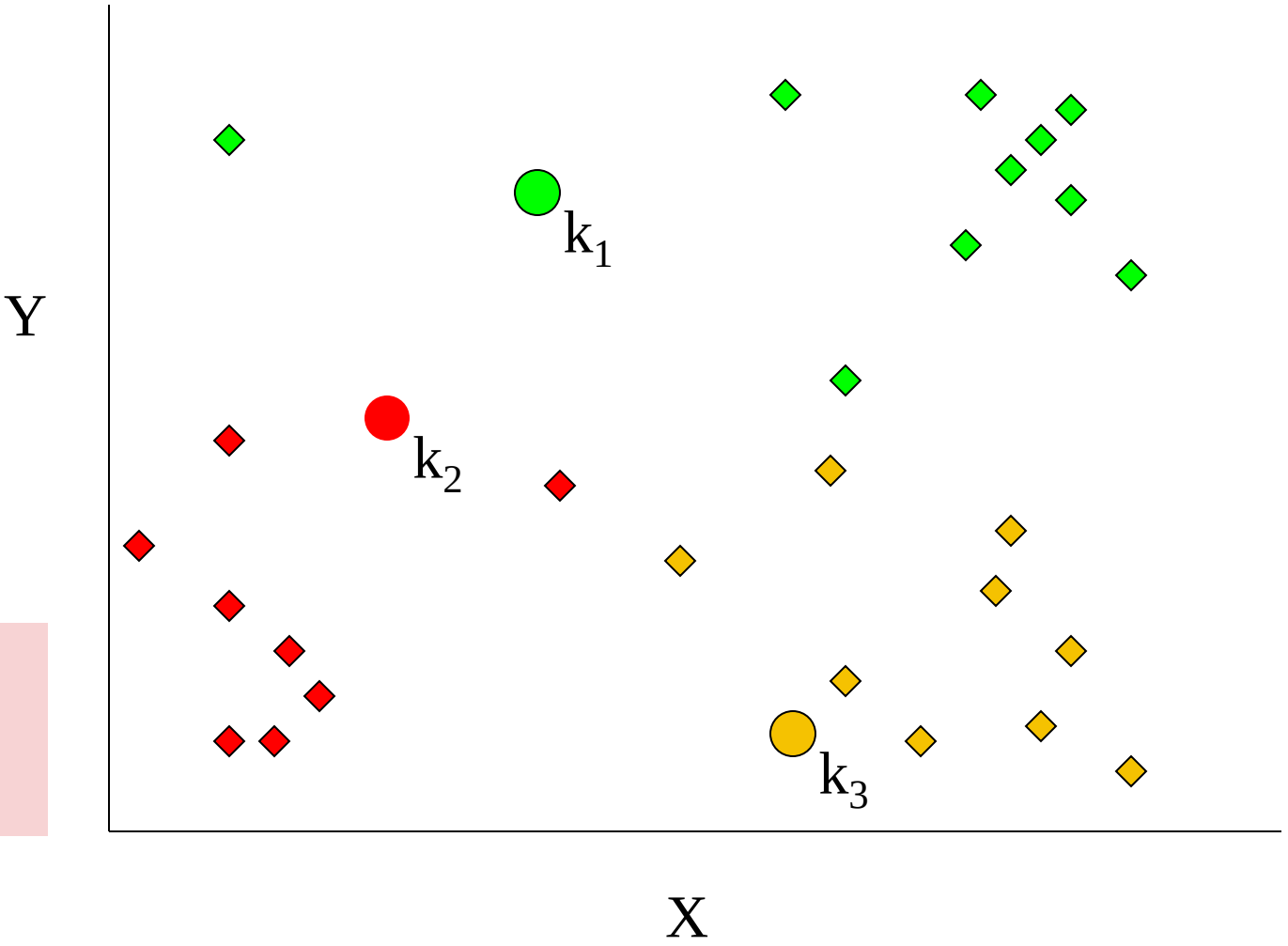
Inertie interclasse I_c : moyenne (pondérée par la somme des poids de chaque classe) des carrés des distances des barycentres de chaque classe au barycentre global

SIMULATION DU K-MEANS 1



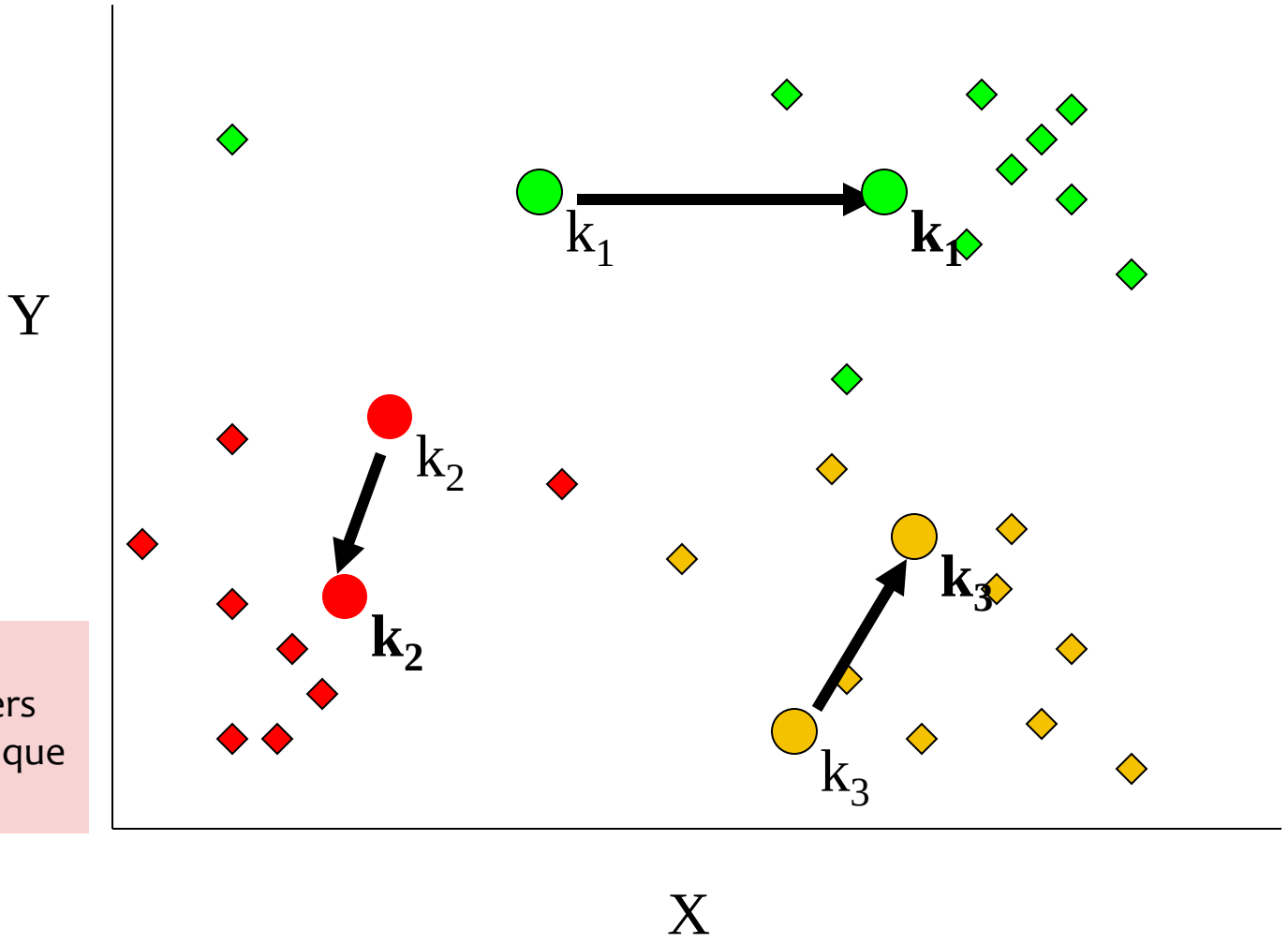
Choisir **3** Centres de classes (au hasard)

SIMULATION DU K-MEANS 2



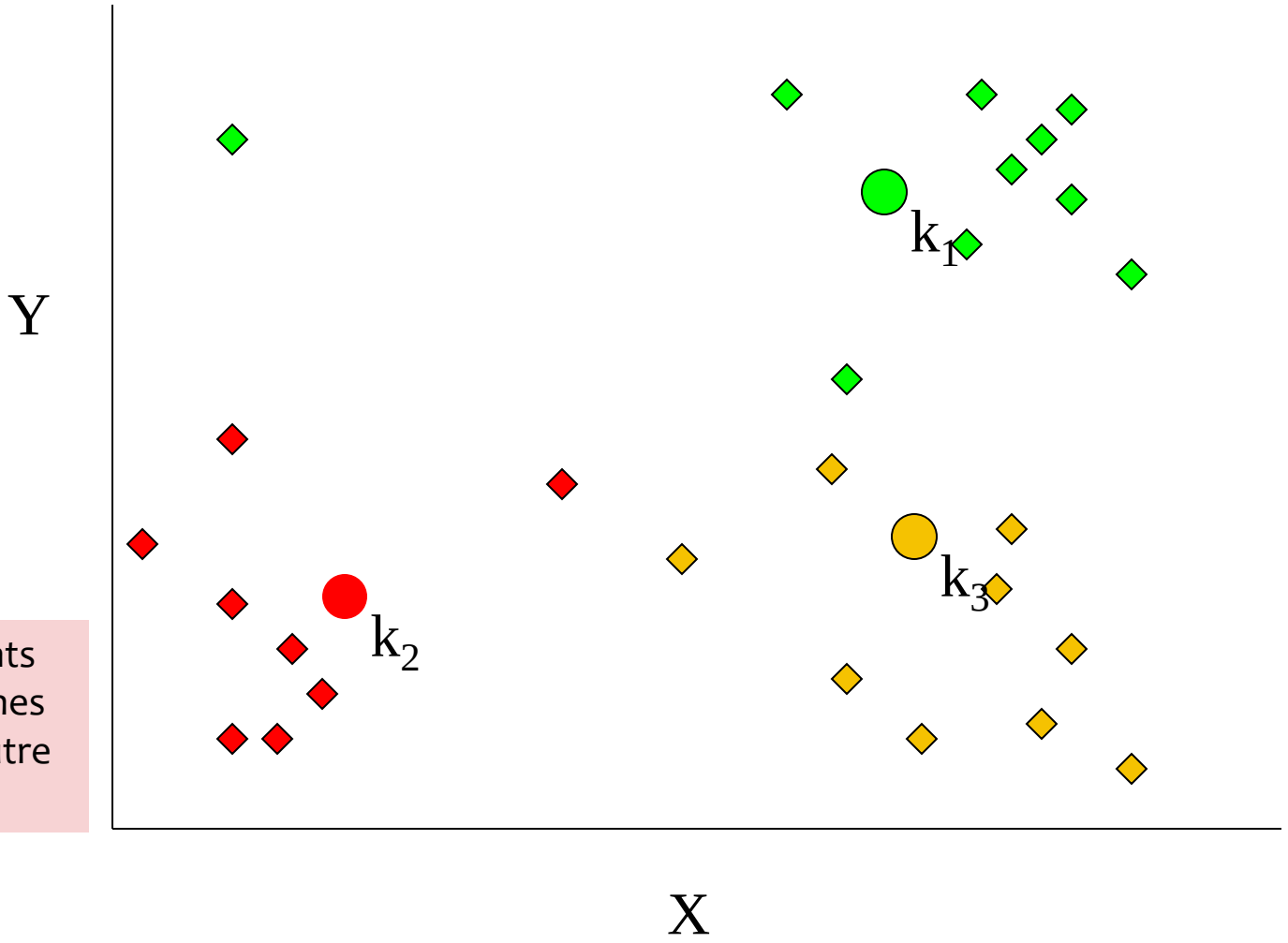
Affecter chaque point à la classe dont le centre est le plus proche

SIMULATION DU K-MEANS 3



Déplacer chaque centre de classe vers la moyenne de chaque classe

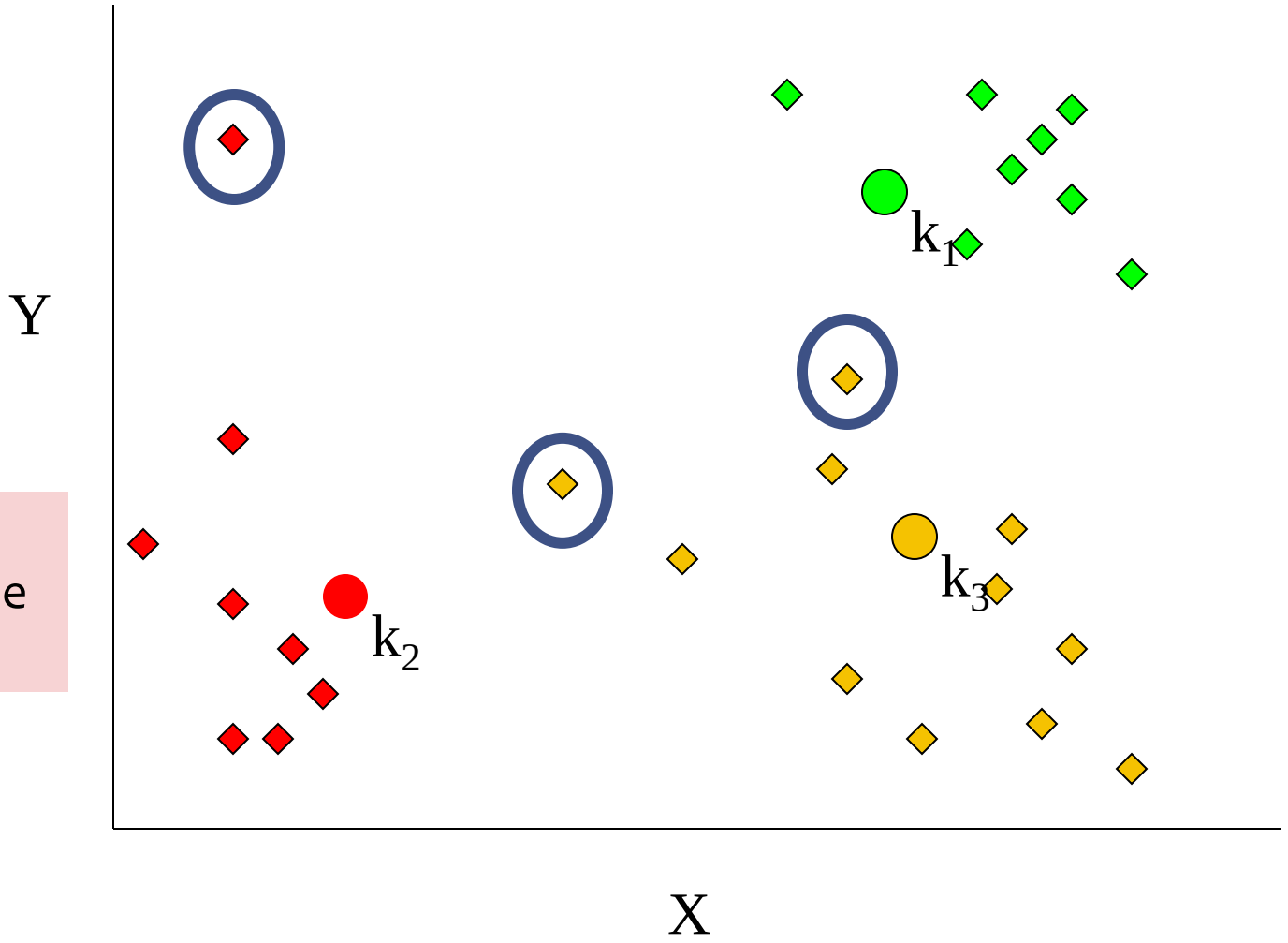
SIMULATION DU K-MEANS 4



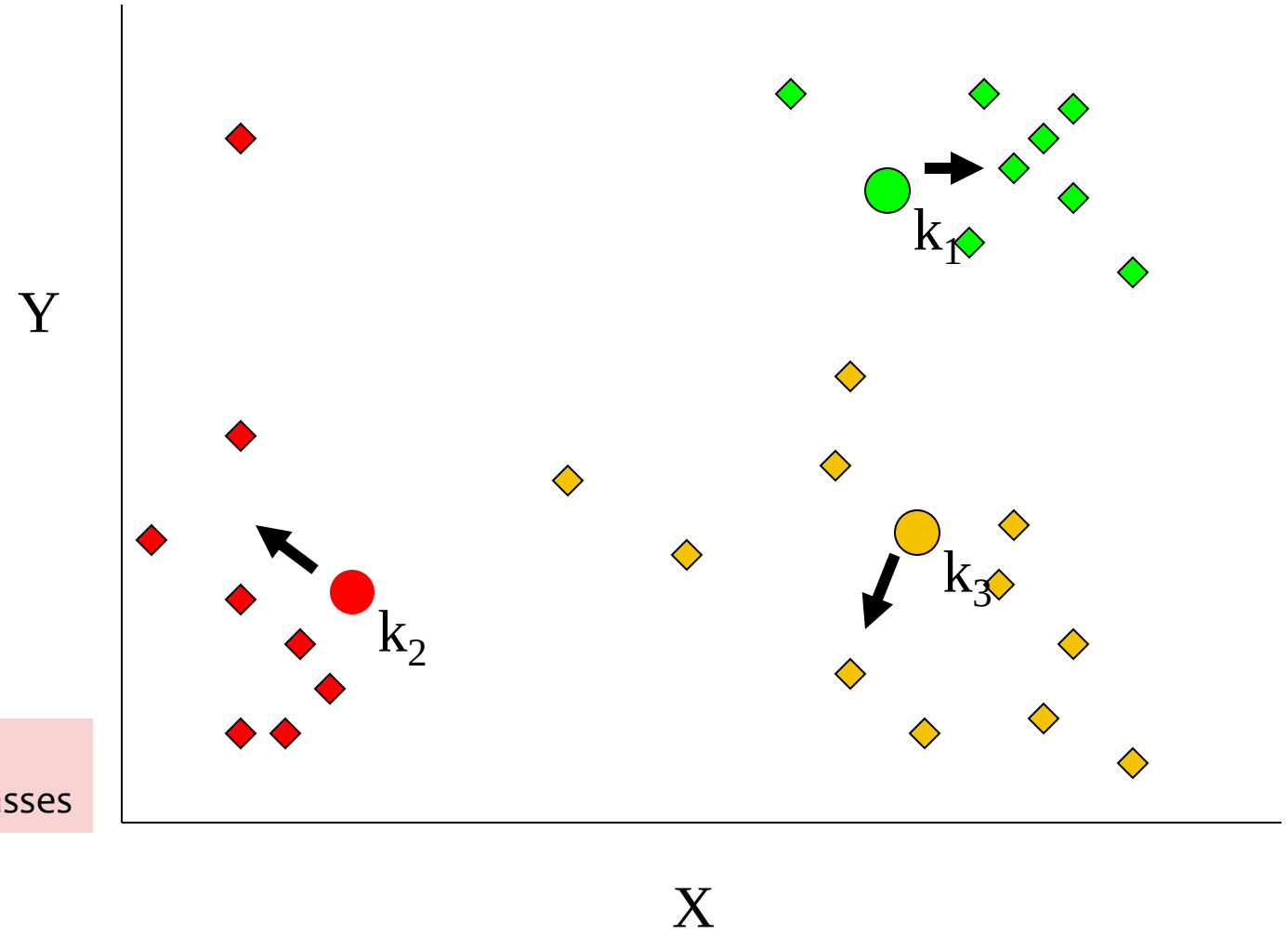
Réaffecter les points qui sont plus proches du centre d'une autre classe

SIMULATION DU K-MEANS 5

les trois points qui changent de classe

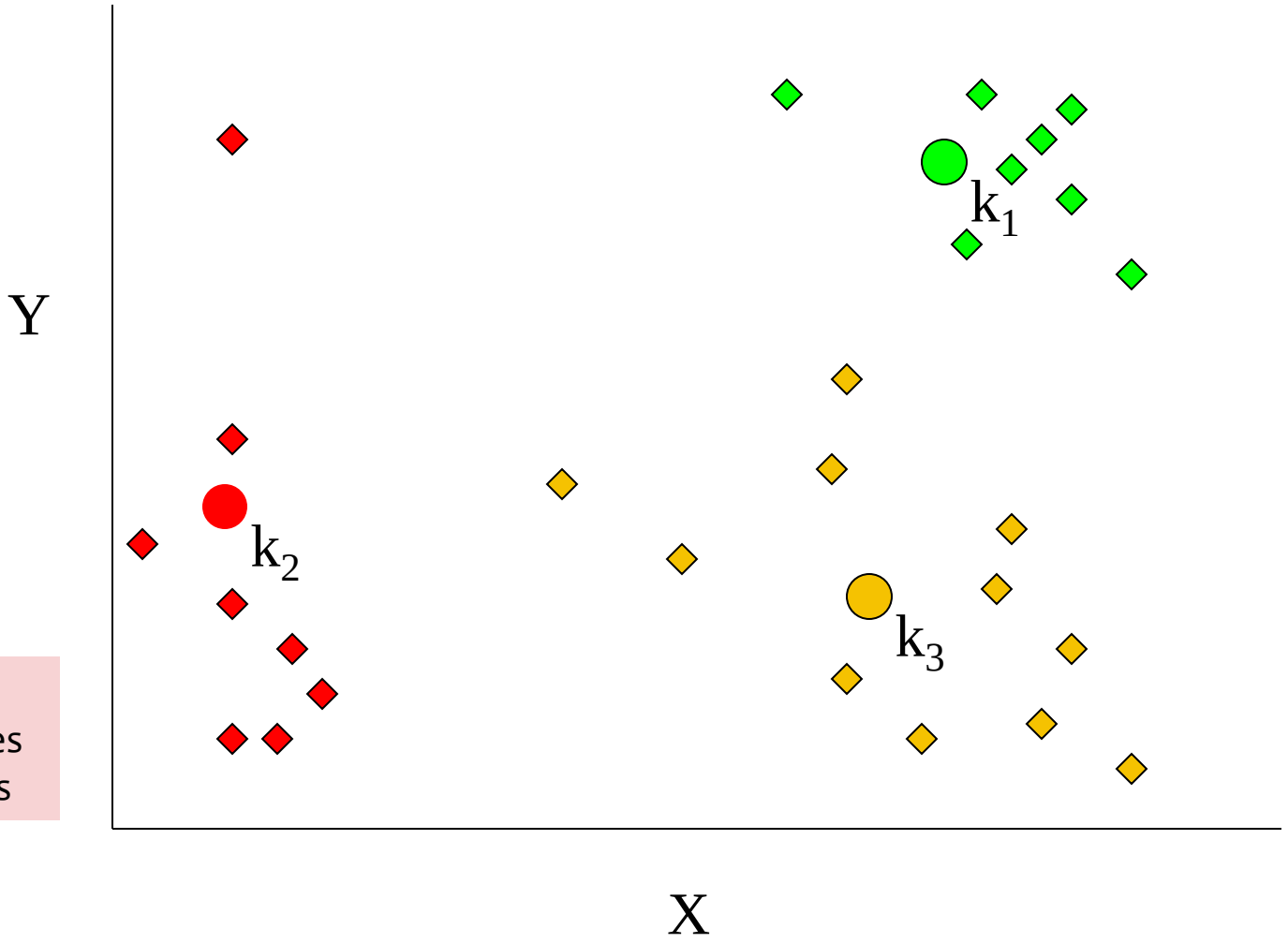


SIMULATION DU K-MEANS 6



Re-calculer les moyennes des classes

SIMULATION DU K-MEANS 7



Déplacer les centres des classes vers les moyennes

Mohamed Heny SELMI © ESPRIT 2012-2013

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

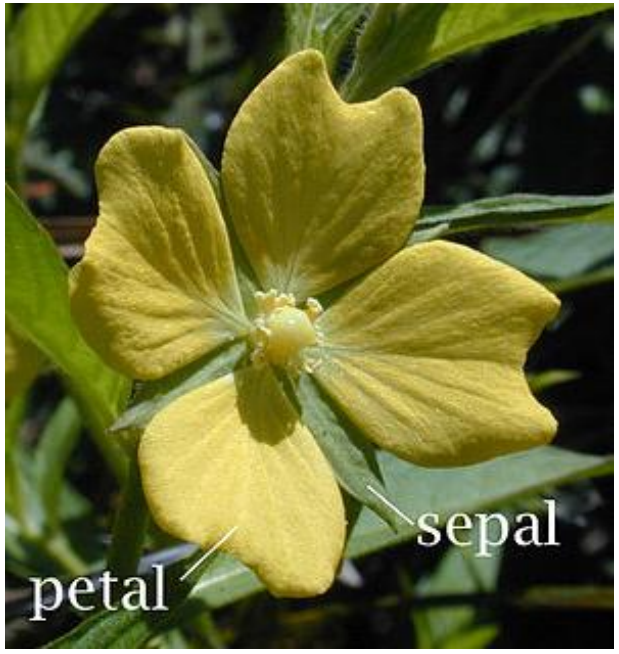
	Individual	Mean Vector (centroid)
Cluster 1	1, 2	(1.3, 1.5)
Cluster 2	3, 4, 5, 6, 7	(3.9, 5.1)

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

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
```

Mohamed Heny SELMI © ESPRIT 2012-2013

APPLICATION K-MEANS « IRIS »

> iris_for_kmeans<-iris[,1:4]

> km <- kmeans(iris_for_kmeans, 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 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
 [76] 1 1 3 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 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 »

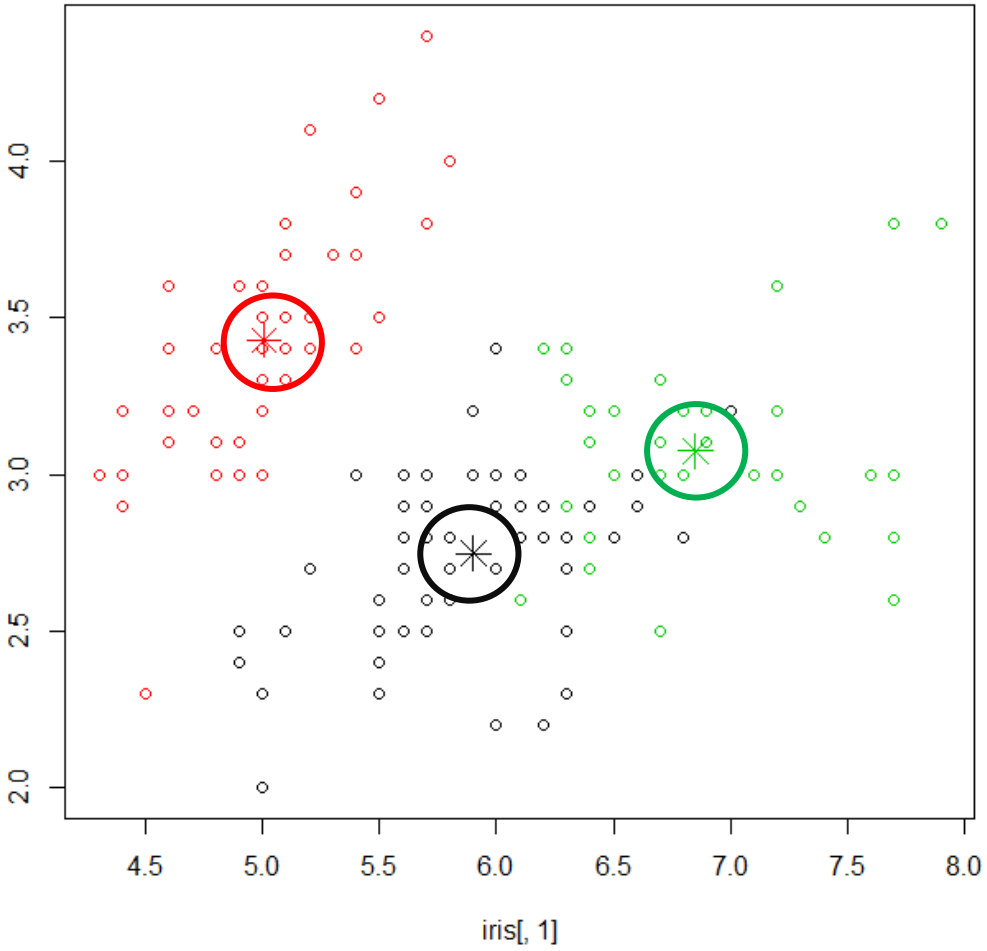
```
> plot(iris[,1], iris[,2], col=km$cluster)
```

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

```
> table(km$clster, iris$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 %		



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 plus optimale correspondant à la fonction objective minimale.
- Les résultats de l'algorithme du K-Means sont sensibles à l'initialisation aléatoires des centres.