Clustering

Hierarchical clustering et Kmeans

Anne Badel, Frédéric Guyon & Jacques van Helden

2020-03-10

Anne Badel,
Frédéric
Guyon &
Jacques van
Helden

Anne Badel, Frédéric Guyon & Jacques van Helden

Les données

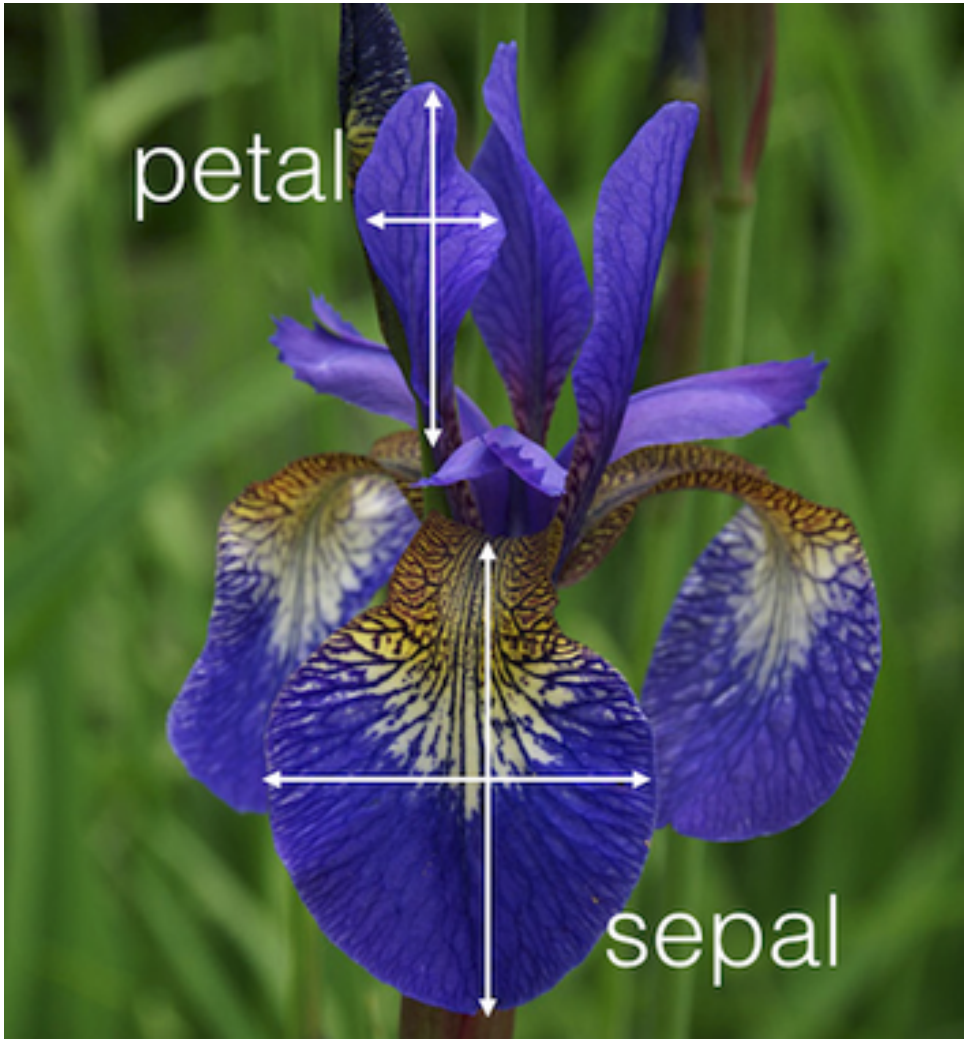
- Comment sont représentées les données dans l'ordinateur ?
- Comment représenter les données dans l'espace ?
- Comment découvrir des "clusters" dans les données ?
  - classification hiérarchique
  - kmeans
- comment déterminer le nombre de groupe optimal ?
- comment comparer deux classifications ?

Anne Badel, Frédéric Guyon & Jacques van Helden

Les données dans l'ordinateur (1)

Les iris de Fisher

Ces données sont un classique des méthodes d'apprentissage



Les données dans l'ordinateur (2)

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
1	5.1	3.5	1.4	0.2
2	4.9	3.0	1.4	0.2
3	4.7	3.2	1.3	0.2
4	4.6	3.1	1.5	0.2
5	5.0	3.6	1.4	0.2
6	5.4	3.9	1.7	0.4

Anne Badel, Frédéric Guyon & Jacques van Helden

Les données dans l'ordinateur (2)

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
1	5.1	3.5	1.4	0.2
2	4.9	3.0	1.4	0.2
3	4.7	3.2	1.3	0.2
4	4.6	3.1	1.5	0.2
5	5.0	3.6	1.4	0.2
6	5.4	3.9	1.7	0.4

- 1 ligne = 1 fleur = 1 vecteur
- 1 colonne = 1 variable = 1 vecteur
- l'ensemble des données = 1 échantillon = 1 data.frame

Représentons ces données : une fleur (1)

```
mes.iris[1,]
```

```
  Sepal.Length Sepal.Width Petal.Length Petal.Width
1          5.1         3.5          1.4         0.2
```

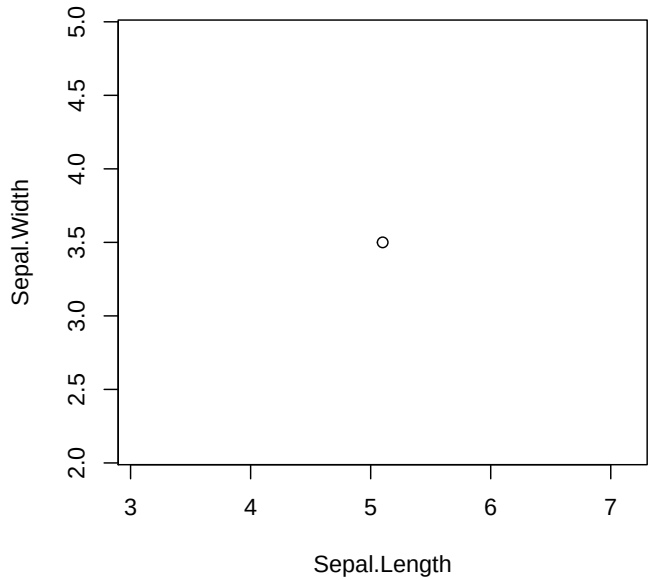
Comment représenter cette fleur ?

- par un point !

Dans quel espace de réprésentation ?

Représentons ces données : une fleur (2)

```
plot(mes.iris[1,1:2])
```



Dans le plan, un point de coordonnées :

- x = 5.1
- y = 3.5

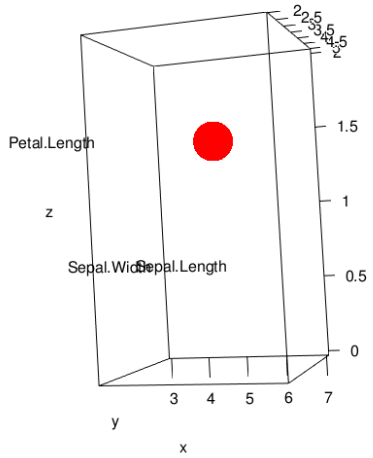
représenté par un vecteur $v2 = (\ 5.1\ ,\ 3.5)$ dans $\mathbb{R}^2$

Anne Badel, Frédéric Guyon & Jacques van Helden

Représentons ces données : une fleur (3)

Dans l'espace, un point de coordonnées :

- x = 5.1
- y = 3.5
- z = 1.4



représenté par un vecteur $v3 = (\ 5.1\ ,\ 3.5\ ,\ 1.4)$ dans $\mathbb{R}^3$

Anne Badel, Frédéric Guyon & Jacques van Helden

Représentons ces données : toutes les fleurs (4)

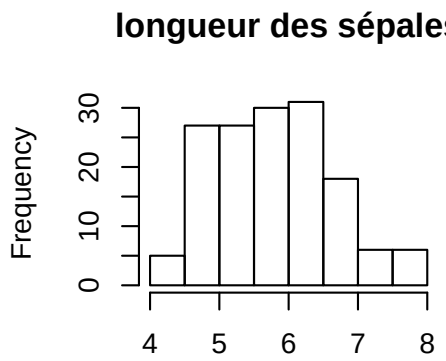
= un nuage de points dans un espace à 4 dimensions

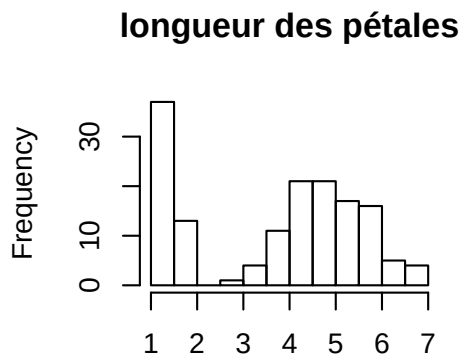
- chaque point est représenté par un vecteur dans $\mathbb{R}^4$
- le nuage de points est représenté par une matrice à n et p (= 4 dimensions)
  - n = nombre de lignes = nombre d'individus = taille de l'échantillon
  - p = nombre de colonnes = nombre de variables décrivant l'échantillon

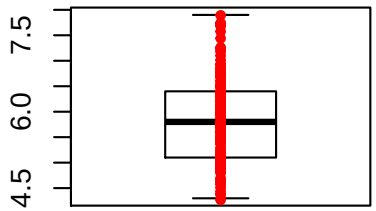
= PAS de représentation possible (pour l'instant)

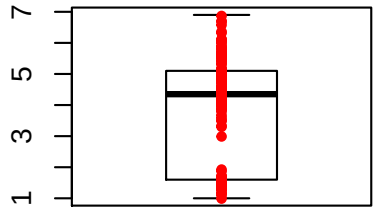
Anne Badel, Frédéric Guyon & Jacques van Helden

Représentons ces données : une variable à la fois (1)



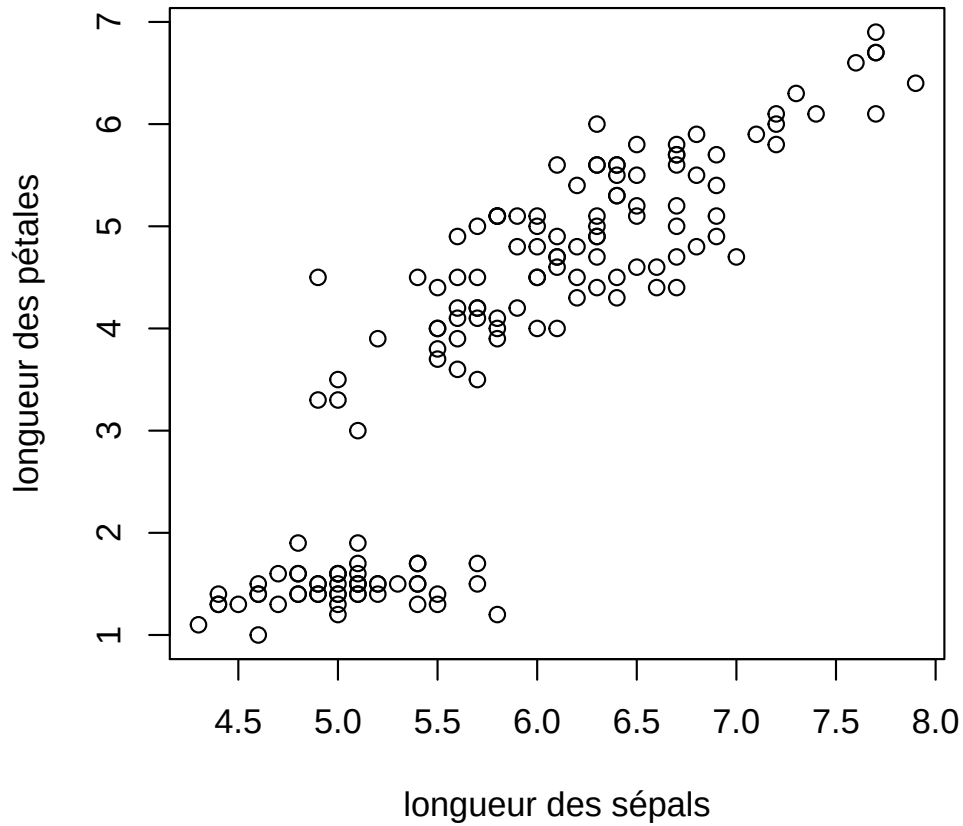






Anne Badel, Frédéric Guyon & Jacques van Helden

Représentons ces données : deux variables à la fois (2)



Anne Badel, Frédéric Guyon & Jacques van Helden

Il faut tenir compte de toutes les dimensions

c'est à dire de toutes les variables à notre disposition

Clustering et classification (termes anglais)

On a une **information** sur nos données

- variables quantitatives = vecteur de réels

Clustering : on cherche à mettre en évidence des groupes dans les données

- le clustering appartient aux méthodes dites **non supervisées**, ou descriptives

Anne Badel, Frédéric Guyon & Jacques van Helden

Clustering et classification (termes anglais)

On a une **information** sur nos données

Clustering : on cherche à mettre en évidence des groupes dans les données

Classification :

- on connaît le partitionnement de notre jeu de données
  - variables quantitatives = vecteur de réels
  - ET
  - variable qualitative = groupe (cluster) d'appartenance = vecteurs de entiers / niveau d'un facteur
  - on cherche à prédire le groupe (la classe) de nouvelles données
- la classification appartient aux méthodes dites **supervisées**, ou prédictives

Anne Badel, Frédéric Guyon & Jacques van Helden

Clustering

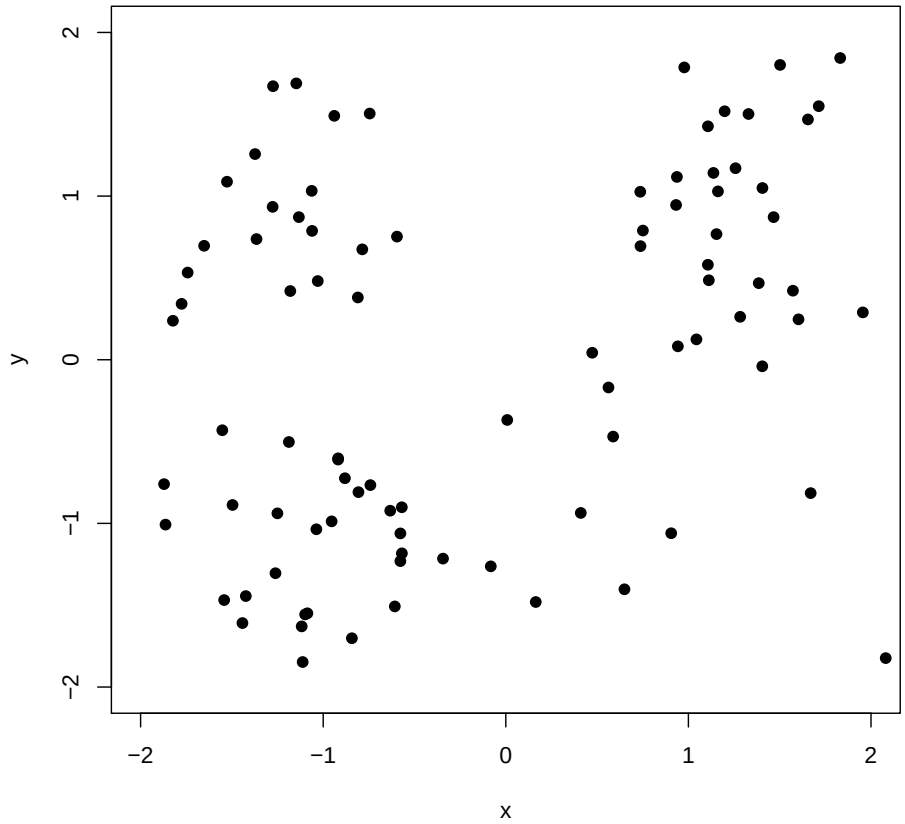


Figure 1: Y a-t-il des groupes ?

Méthodes

Trois grands principes de méthodes basées sur:

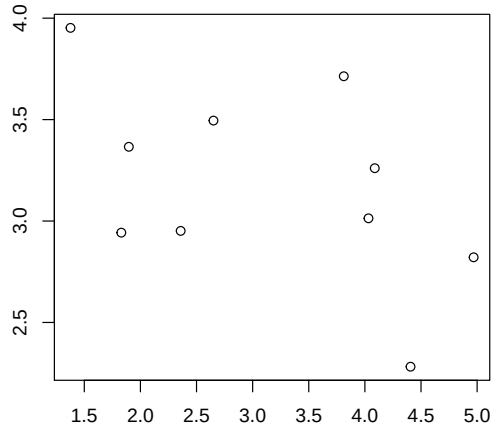
- La géométrie
- Les probabilités (statistique)
- Les graphes

En fait, trois façons de voir les mêmes algorithmes

Anne Badel, Frédéric Guyon & Jacques van Helden

Géométrie et distances (1)

On considère les données comme des points de $\mathbb{R}^n$



$\mathbb{R}^n$: espace Euclidien à n dimensions, où

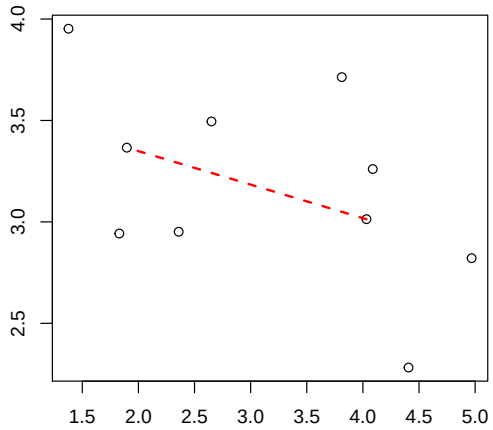
- chaque dimension représente une des variables observées;
- un individu est décrit comme un vecteur à n valeurs, qui correspond à un point dans cet espace.

Anne Badel, Frédéric Guyon & Jacques van Helden

Géométrie et distances (2)

On considère les données comme des points de R^n (*)

- géométrie donnée par distances
- distances = dissimilarités imposées par le problème
- dissimilarités $\longrightarrow$ permettent visualisation de l'ensemble des points



Anne Badel, Frédéric Guyon & Jacques van Helden

Géométrie et distances (3)

Sur la base d'une distance (souvent euclidienne)

- Clustering :
  - Méthode agglomérative ou hierarchical clustering
  - Moyennes mobiles ou K-means : séparation optimale des groupes connaissant le nombre de groupes

Distances

Définition d'une distance : fonction positive de deux variables

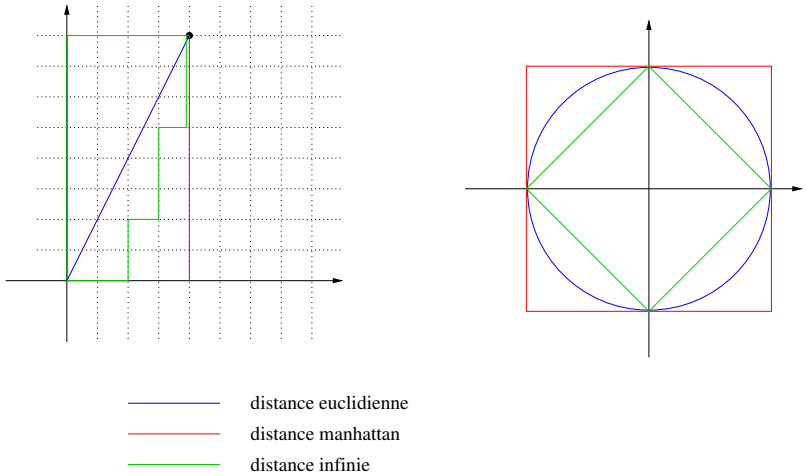
1. $d(x, y) \geq 0$
2. $d(x, y) = d(y, x)$
3. $d(x, y) = 0 \iff x = y$
4. **Inégalité triangulaire :** $d(x, z) \leq$ d(x,y)+d(y,z)

Si 1,2,3 : dissimilarité

Anne Badel, Frédéric Guyon & Jacques van Helden

Distances utilisées dans R (1)

- distance euclidienne ou distance L_2:
  $d(x, y) = \sqrt{\sum_i (x_i - y_i)^2}$
- distance de manahattan ou distance L_1:
  $d(x, y) = \sum_i |x_i - y_i|$
- distance du maximum ou L-infinis, L_∞:
  $d(x, y) = \max_i |x_i - y_i|$



Distances utilisées dans R (2)

- distance de Minkowski l_p:

$$d(x,y) = \sqrt[p]{\sum_i (|x_i - y_i|^p}$$

- distance de Canberra (x et y valeurs positives):

$$d(x,y) = \sum_i \frac{x_i - y_i}{x_i + y_i}$$

- distance binaire ou distance de Jaccard ou Tanimoto: proportion de propriétés communes

Autres distances non géométriques (pour information)

Utilisées en bio-informatique:

- Distance de **Hamming**: nombre de remplacements de caractères (substitutions)
- Distance de **Levenshtein**: nombre de substitutions, insertions, deletions entre deux chaînes de caractères

$$d(\text{"}BONJOUR\text{"}, \text{"}BONSOIR\text{"}) = 2$$

- Distance d'**alignements**: distances de Levenshtein avec poids (par ex. matrices BLOSSUM)
- Distances d'**arbre** (Neighbor Joining)
- Distances **ultra-métriques** (phylogénie UPGMA)

Distances plus classiques en génomique

Il existe d'autres mesures de distances, plus ou moins adaptées à chaque problématique :

- **Jaccard** (comparaison d'ensembles): $J_D = \frac{A \cap B}{A \cup B}$
- Distance du χ^2 (comparaison de tableau d'effectifs)

Ne sont pas des distances, mais indices de dissimilarité :

- **Bray-Curtis** (en écologie, comparaison d'abondance d'espèces)
- **Jensen-Shannon** (comparaison de distributions)

Note : lors du TP, sur les données d'expression RNA-seq, nous utiliserons le **coefficient de corrélation de Spearman** et la distance dérivée, $d_c = 1 - r$ ou $d_c = \sqrt{2 \times (1 - r)}$

Avec R (1)

- on utilise la fonction `dist()` avec l'option `method = "euclidean", "manhattan", ...`

2.36	4.03	1.38	1.90	2.65
2.95	3.01	3.95	3.37	3.50

distance euclidienne : 4.69

distance de manhattan = 12.81

Avec R (2)

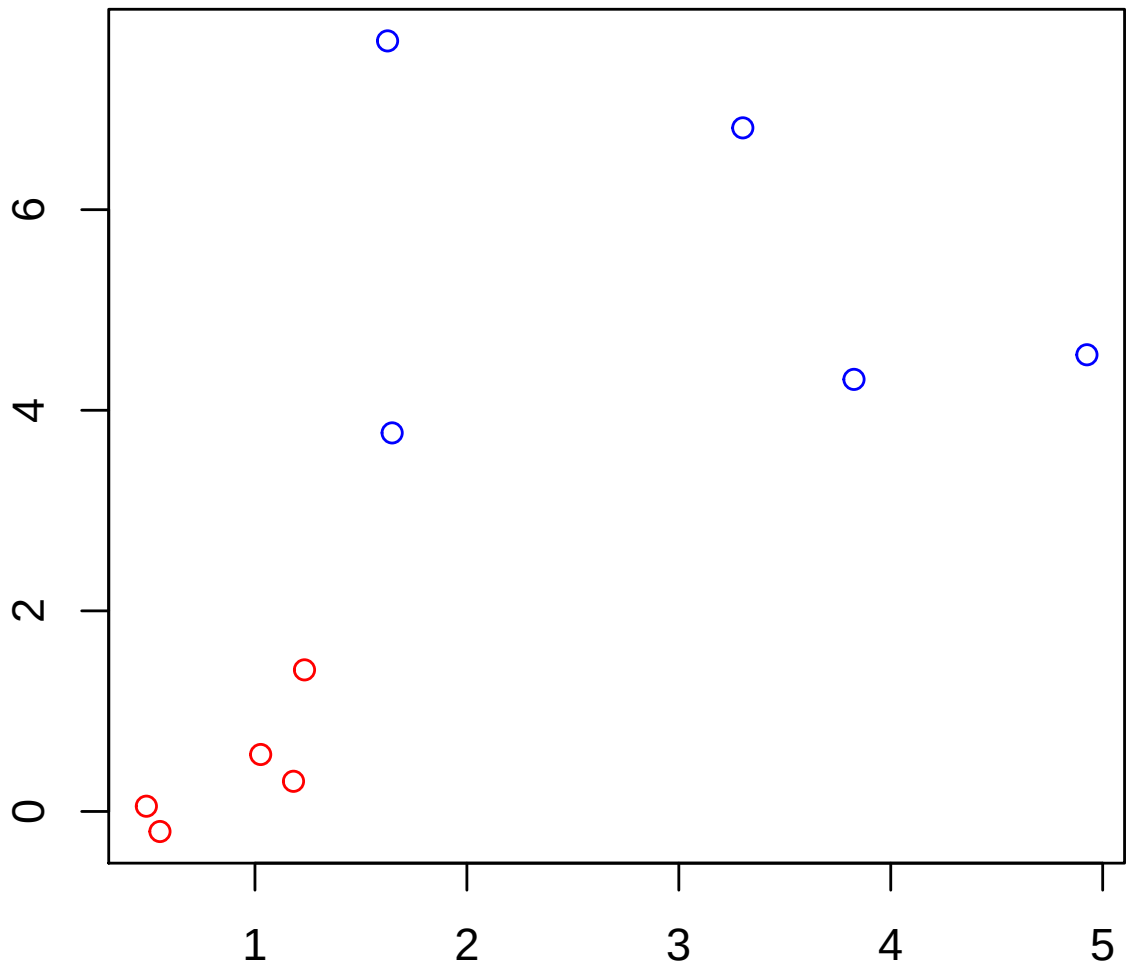
- ou pour des distances particulières, par exemple l'indice de Jaccard :

v.a	0	1	0	0	0	0	0
v.b	0	1	0	0	0	1	0
v.c	0	1	0	0	0	0	0

```
          v.a       v.b
v.b 0.3333333
v.c 0.0000000 0.3333333
```

Anne Badel, Frédéric Guyon & Jacques van Helden

Distances entre groupes (1)



Anne Badel, Frédéric Guyon & Jacques van Helden

Distances entre groupes (2)

- **Single linkage** : élements les plus proches des 2 groupes

$$D(C_1, C_2) = \min_{i \in C_1, j \in C_2} D(x_i, x_j)$$

- **Complete linkage** : éléments les plus éloignés des 2 groupes

$$D(C_1, C_2) = \max_{i \in C_1, j \in C_2} D(x_i, x_j)$$

Distances entre groupes (3)

- **Group average** : distance moyenne

$$D(C_1, C_2) = \frac{1}{N_1 N_2} \sum_{i \in C_1, j \in C_2} D(x_i, x_j)$$

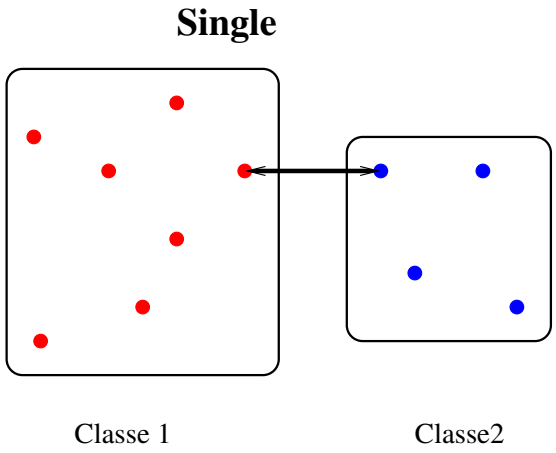
- **Ward**

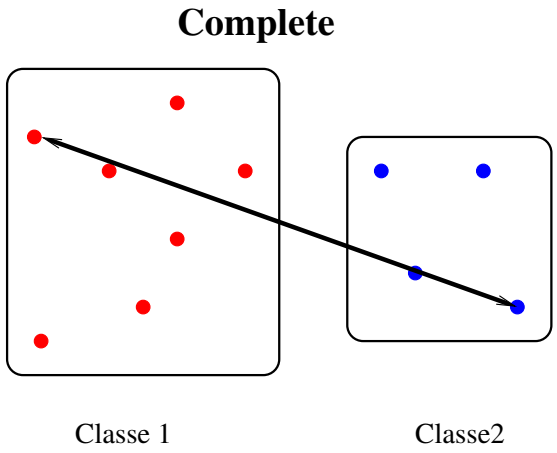
$$d^2(C_i, C_j) = I_{intra}(C_i \cup C_j) - I_{intra}(C_i) - I_{intra}(C_j)$$

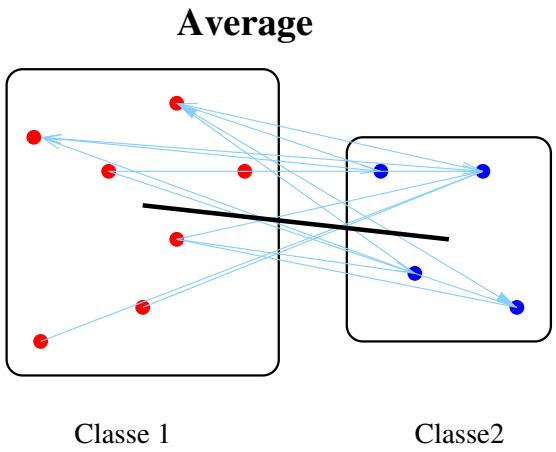
$$D(C_1, C_2) = \sqrt{\frac{N_1 N_2}{N_1 + N_2}} \|m_1 - m_2\|$$

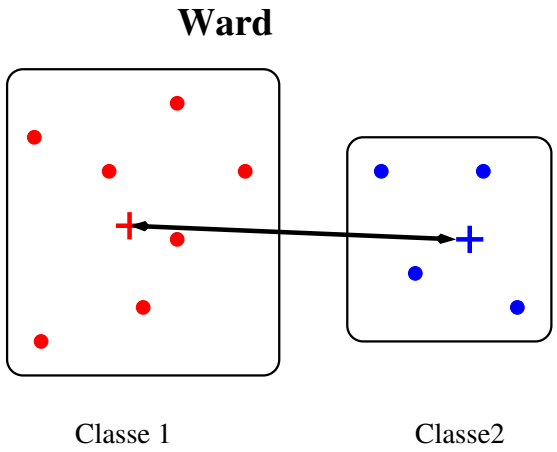
Anne Badel, Frédéric Guyon & Jacques van Helden

Distances entre groupes (4)









Les données

Ces données sont un classique des méthodes d'apprentissage

Dans un premier temps, regardons les données

```
dim(mes.iris)
```

```
[1] 150   4
```

```
head(mes.iris)
```

```
  Sepal.Length Sepal.Width Petal.Length Petal.Width
1          5.1         3.5          1.4         0.2
2          4.9         3.0          1.4         0.2
3          4.7         3.2          1.3         0.2
4          4.6         3.1          1.5         0.2
5          5.0         3.6          1.4         0.2
6          5.4         3.9          1.7         0.4
```

Anne Badel, Frédéric Guyon & Jacques van Helden

```
str(mes.iris)
```

```
'data.frame':   150 obs. of  4 variables:
 $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4
 $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2
 $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5
 $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2
```

```
summary(mes.iris)
```

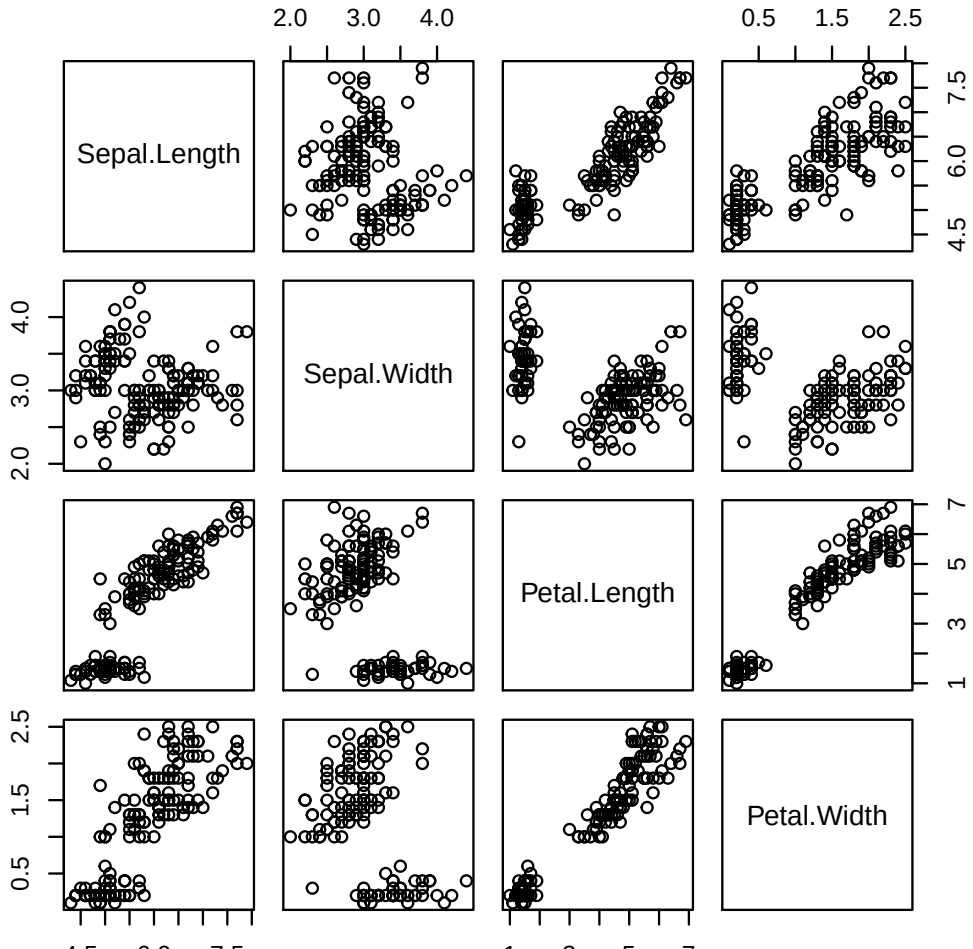
```
  Sepal.Length    Sepal.Width     Petal.Length    Pet
 Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.
 1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st
 Median :5.800   Median :3.000   Median :4.350   Medi
 Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean
 3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd
 Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.
```

Visualisation des données

On peut ensuite essayer de visualiser les données

- par un `plot`

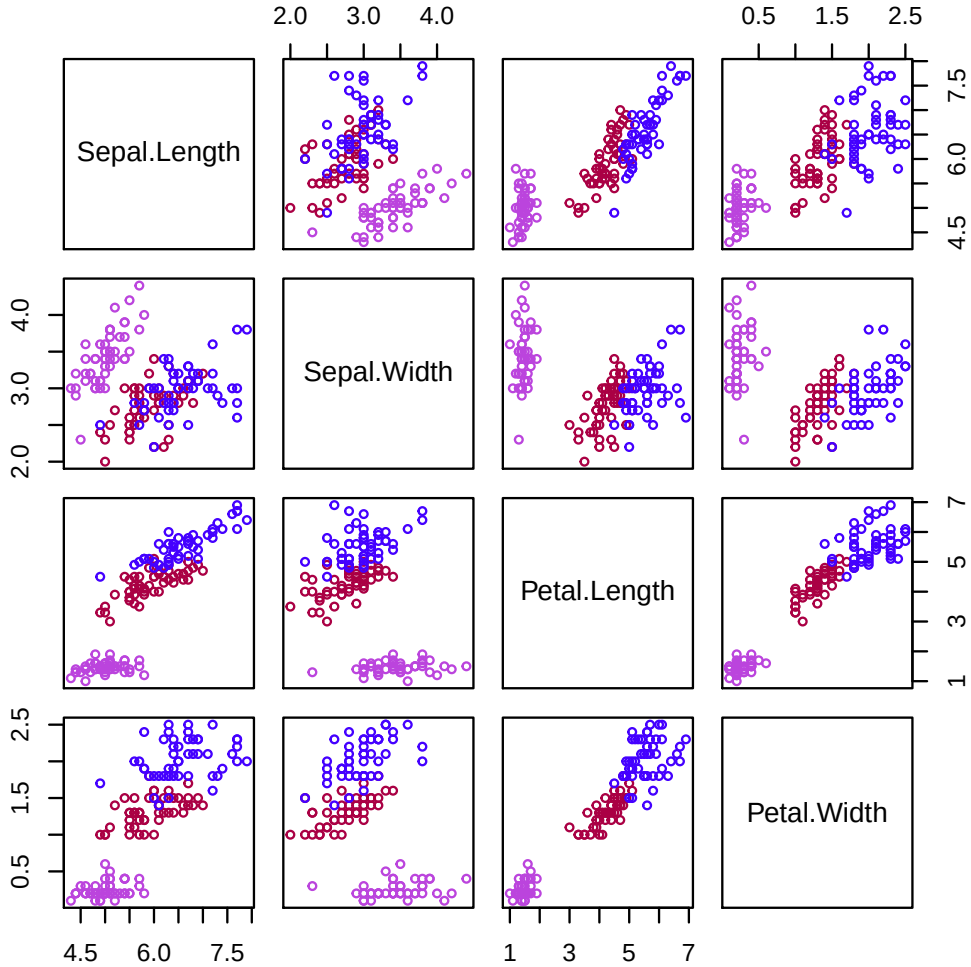
```
plot(mes.iris)
```



Anne Badel, Frédéric Guyon & Jacques van Helden

Visualisation des données - coloration par espèces

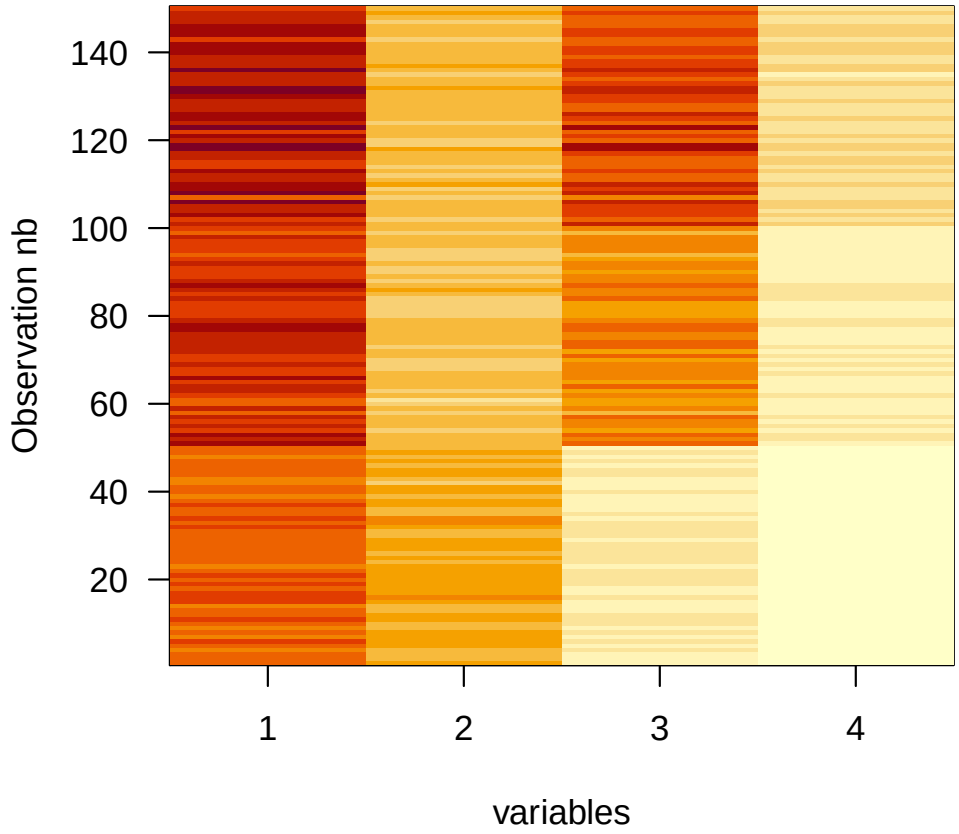
```
species.colors <- c(setosa = "#BB44DD", virginica = "
plot(mes.iris, col = species.colors[iris$Species], ce
```



Visualisation des données

- par la fonction `image()`

```
image(1:nb.var, 1:nb.iris ,t(as.matrix(mes.iris)), xl
```



Anne Badel, Frédéric Guyon & Jacques van Helden

Nettoyage des données (1) : données manquantes

Avant de commencer à travailler, il est nécessaire de commencer par vérifier que :

- il n'y a pas de données manquantes

```
sum(is.na(mes.iris))
```

```
[1] 0
```

Anne Badel, Frédéric Guyon & Jacques van Helden

Nettoyage des données (2) : variables constantes

- aucune variable n'est constante (aucune variable n'a une variance nulle)

```
iris.var <- apply(mes.iris, 2, var)
kable(iris.var, digits = 3, col.names = "Variance")
```

	Variance
Sepal.Length	0.686
Sepal.Width	0.190
Petal.Length	3.116
Petal.Width	0.581

```
sum(apply(mes.iris, 2, var) == 0)
```

```
[1] 0
```

Anne Badel, Frédéric Guyon & Jacques van Helden

Normalisation

Afin de pouvoir considérer que toutes les variables sont à la même échelle, il est parfois nécessaire de normaliser les données.

- soit
  - en centrant (ramener la moyenne de chaque variable à 0)

```
mes.iris.centre <- scale(mes.iris, center=TRUE, scale
```

- soit
  - en centrant (ramener la moyenne de chaque variable 0)
  - et mettant à l'échelle (ramener la variance de chaque variable à 1)

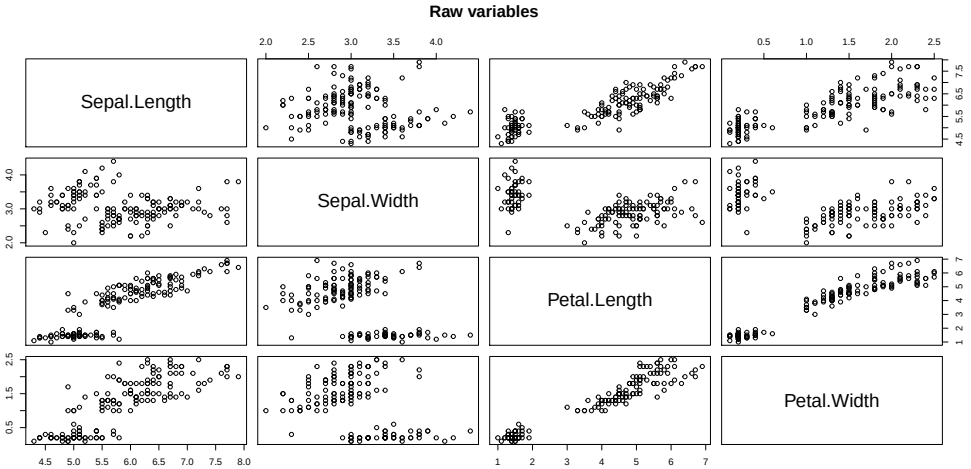
```
mes.iris.scaled <- scale(mes.iris, center=TRUE, scale
```

Anne Badel, Frédéric Guyon & Jacques van Helden

On peut visuellement regarder l'effet de la normalisation :

par un plot des données

```
plot(mes.iris, main = "Raw variables")
```

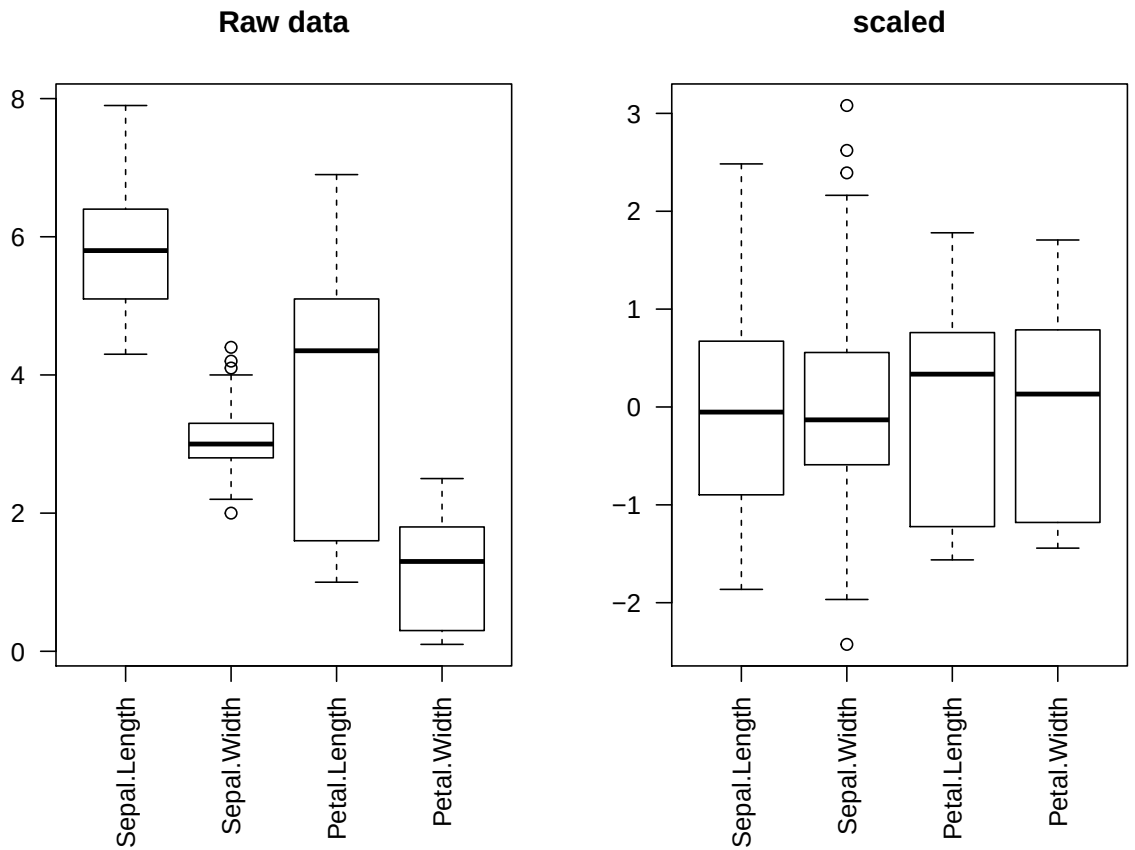


! ne pas faire si "grosses" données

Anne Badel, Frédéric Guyon & Jacques van Helden

... par une boîte à moustaches (boxplot)

```
par(mfrow = c(1,2))
par(mar = c(7, 4.1, 4.1, 1.1)) # adapt margin sizes f
boxplot(mes.iris, main = "Raw data", las = 2)
boxplot(mes.iris.scaled, main = "scaled", las = 2)
```

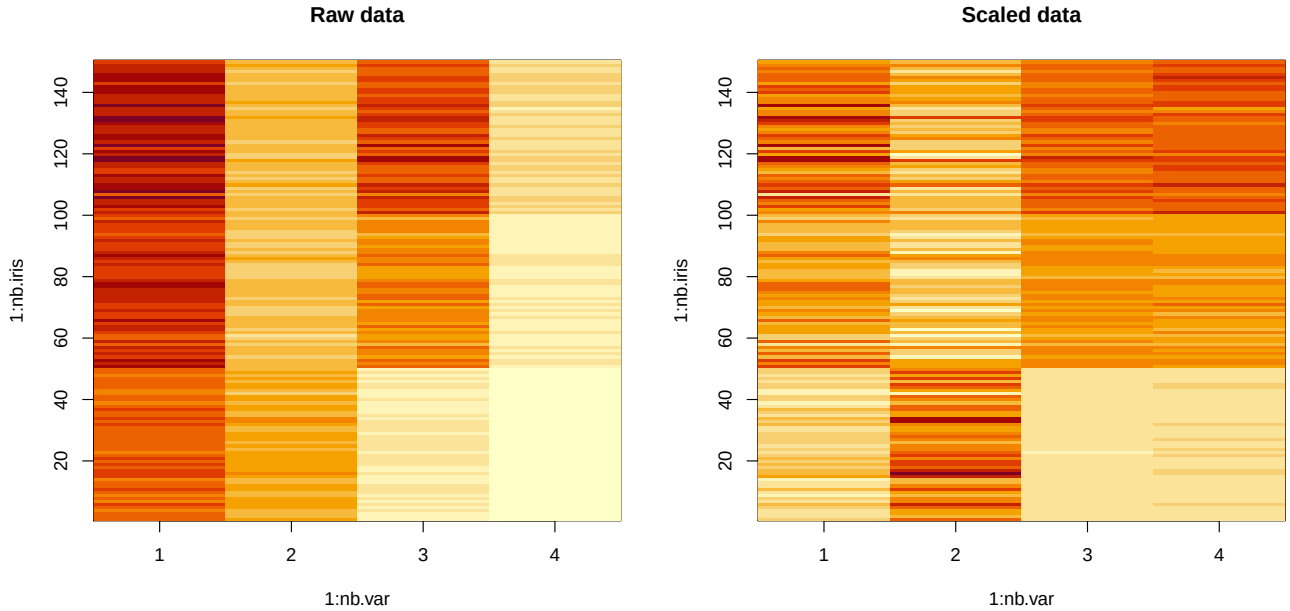



```
par(mar = c(5.1, 4.1, 4.1, 2.1)) # Restore original m
```

Anne Badel, Frédéric Guyon & Jacques van Helden

. . . par une image

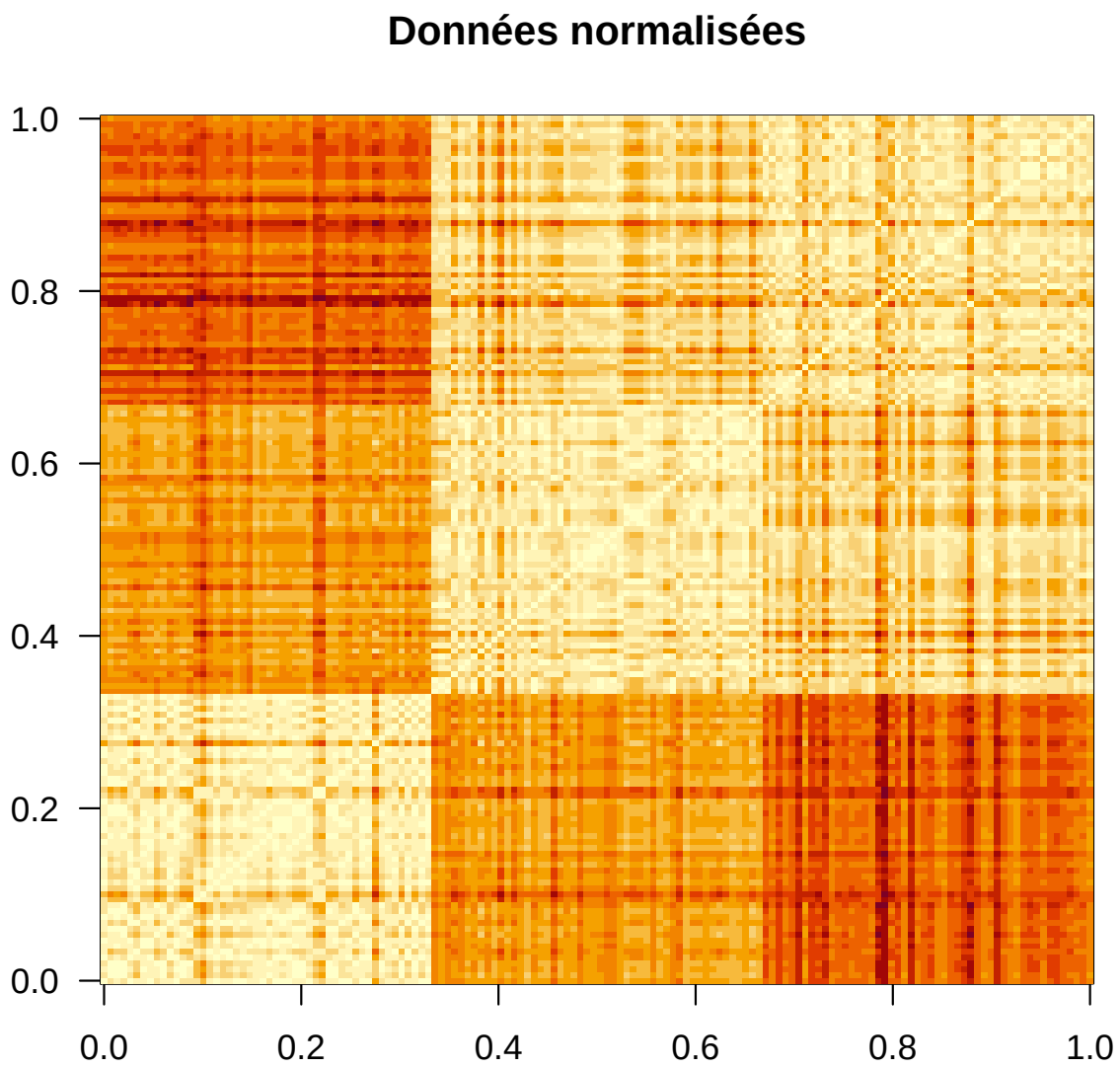
```
par(mfrow=c(1,2))
image(1:nb.var, 1:nb.iris, t(as.matrix(mes.iris)), ma
image(1:nb.var, 1:nb.iris, t(as.matrix(mes.iris.scale
```



Anne Badel, Frédéric Guyon & Jacques van Helden

La matrice de distances

Nous utilisons ici la distance euclidienne sur données **normalisées**.



Anne Badel, Frédéric Guyon & Jacques van Helden

La classification hiérarchique : principe

classification hiérarchique : mettre en évidence des liens hiérachiques entre les individus

- classification hiérarchique **ascendante** : partir des individus pour arriver à des classes / cluster
- classification hiérarchique **descendante** : partir d'un groupe qu'on subdivise en sous-groupes /clusters jusqu'à arriver à des individus.

Anne Badel, Frédéric Guyon & Jacques van Helden

Notion importante, cf distances

- ressemblance entre individus = distance
- ressemblance entre groupes d'invidus = critère d'aggrégation
  - lien simple
  - lien complet
  - lien moyen
  - critère de Ward

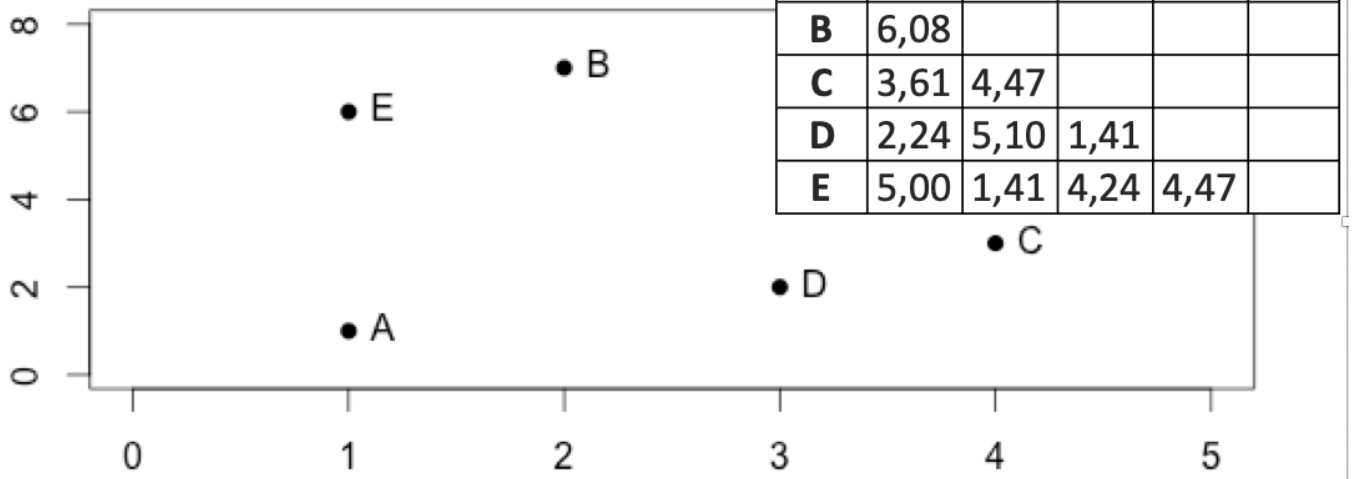
Anne Badel, Frédéric Guyon & Jacques van Helden

L'algorithme : étape 1

- départ : n individus = n clusters distincts
- calcul des distances entre tous les individus
  - choix de la métrique à utiliser en fonction du type de données
- regroupement des 2 individus les plus proches => (n-1) clusters

Anne Badel, Frédéric Guyon & Jacques van Helden

au départ

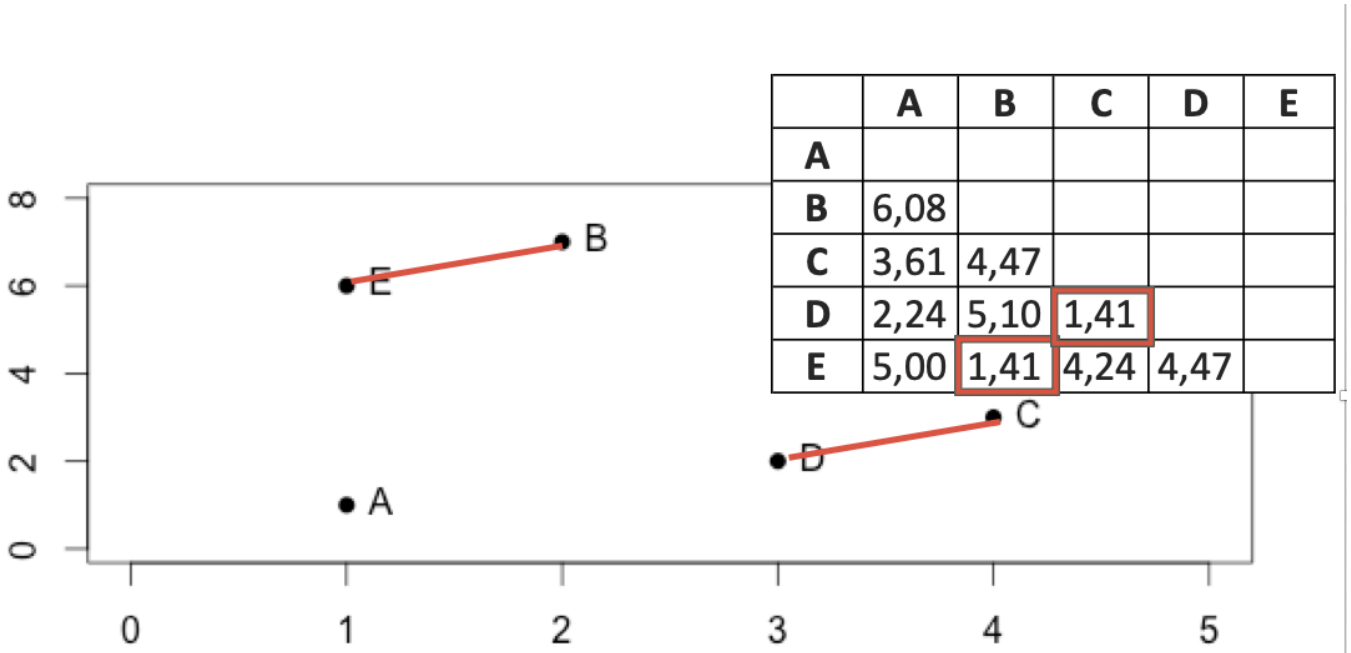


	A	B	C	D	E
A					
B	6,08				
C	3,61	4,47			
D	2,24	5,10	1,41		
E	5,00	1,41	4,24	4,47	

initialisation : (A), (B), (C), (D), (E)

Anne Badel, Frédéric Guyon & Jacques van Helden

identification des individus les plus proches



	A	B	C	D	E
A					
B	6,08				
C	3,61	4,47			
D	2,24	5,10	1,41		
E	5,00	1,41	4,24	4,47	

première partition : (A), (BE), (CD)

Anne Badel, Frédéric Guyon & Jacques van Helden

construction du dendrogramme

Anne Badel, Frédéric Guyon & Jacques van Helden

étape j :

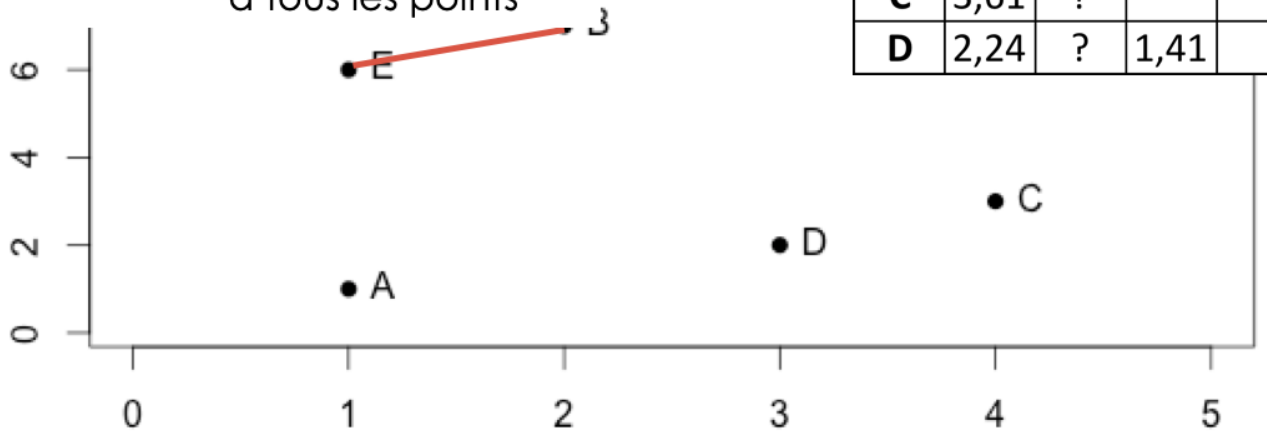
- calcul des dissemblances entre chaque groupe obtenu à l'étape $(j-1)$
- regroupement des deux groupes les plus proches $=>$ $(n-j)$ clusters

Anne Badel, Frédéric Guyon & Jacques van Helden

calcul des nouveaux représentants ‘BE’ et ‘CD’

par le lien moyen,
calcul de BE
calcul de la distance de BE
à tous les points

	A	BE	C	D
A				
BE	?			
C	3,61	?		
D	2,24	?	1,41	



par le lien moyen,
calcul de CD

Anne Badel, Frédéric Guyon & Jacques van Helden

calcul des distances de l'individu restant 'A' aux points moyens

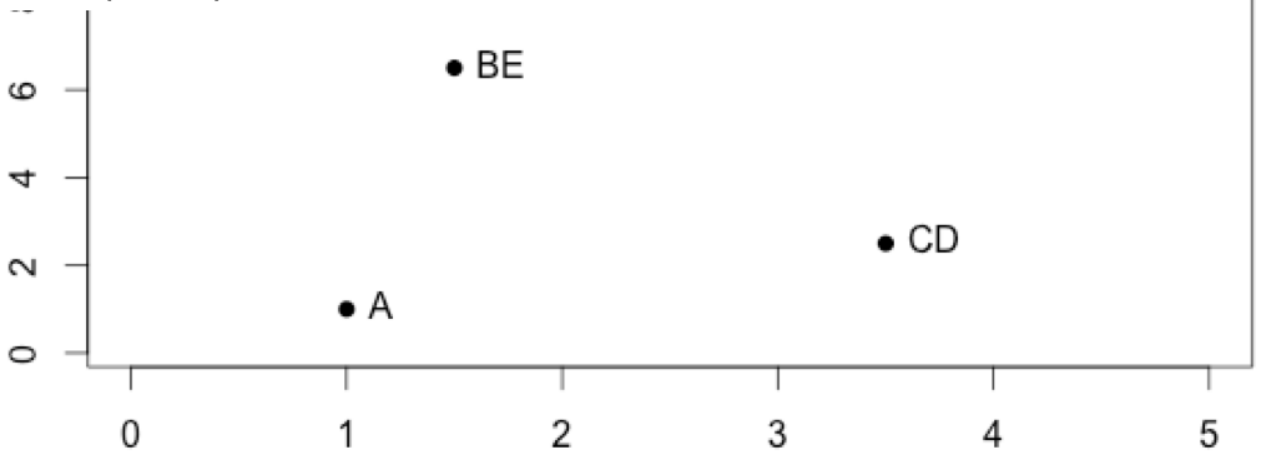
calcul des distances

d(BE, A) =

d(BE, CD) =

d(CD,A) =

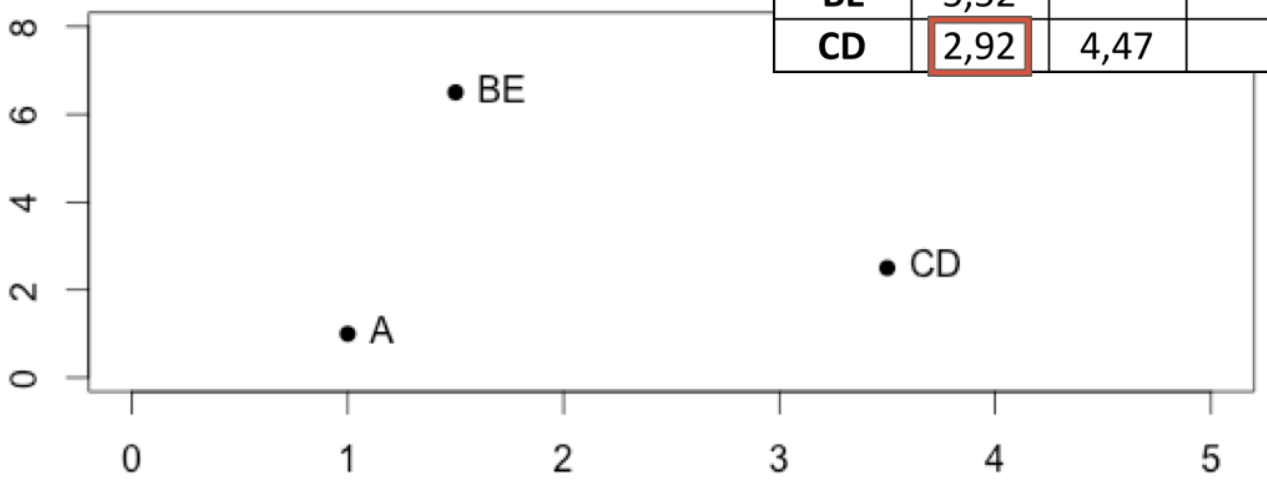
$$d(BE,A) = \frac{d(B,A) + d(E,A)}{2} = \frac{6.08 + 5}{2} = 5.52$$



Anne Badel, Frédéric Guyon & Jacques van Helden

A est plus proche de . . .

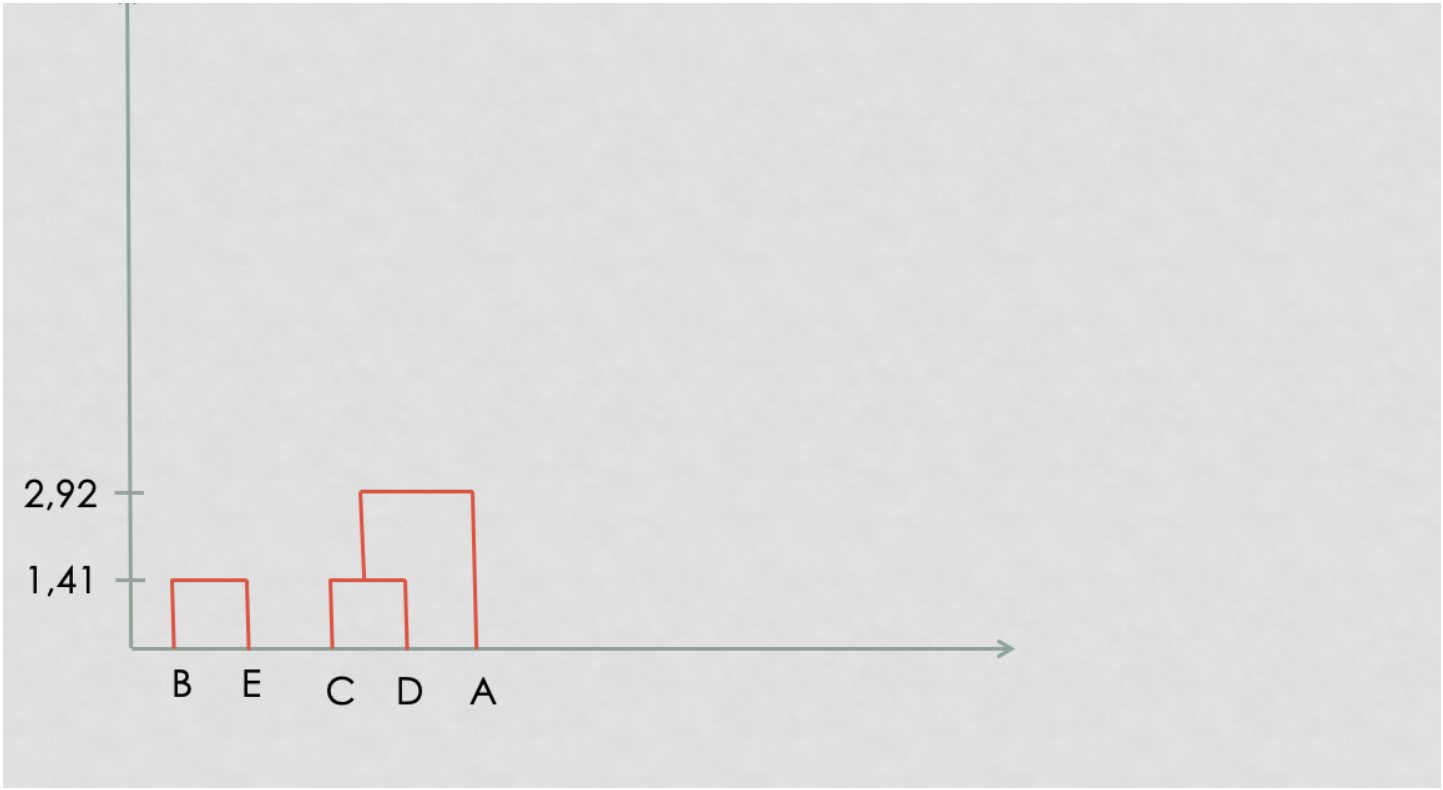
	A	BE	CD
A			
BE	5,52		
CD	2,92	4,47	



deuxième partition : (BE), (CDA)

Anne Badel, Frédéric Guyon & Jacques van Helden

dendrogramme

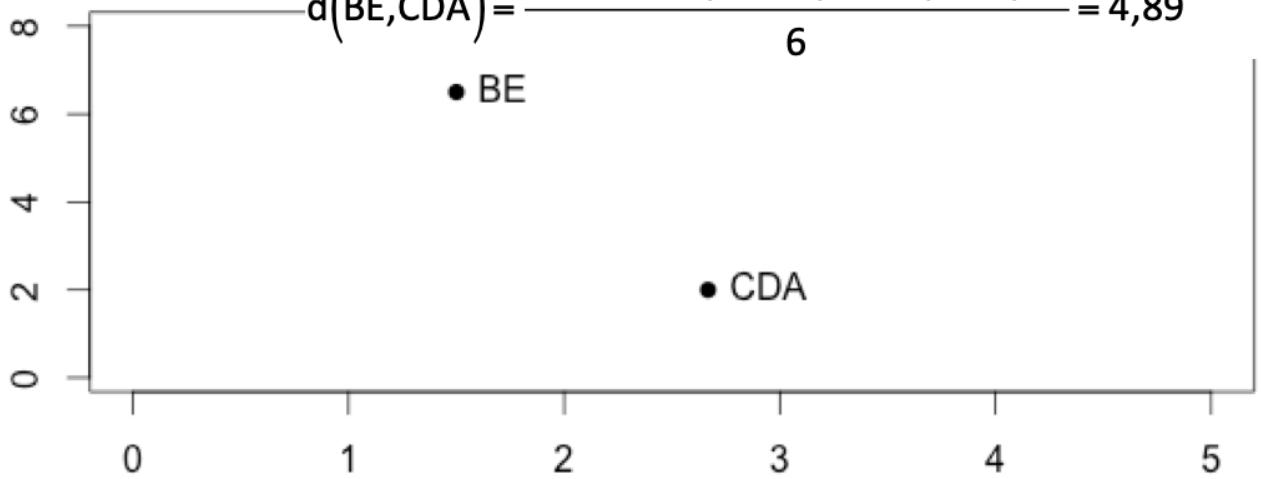


Anne Badel, Frédéric Guyon & Jacques van Helden

pour finir

$$d(BE,CDA) = \frac{d(B,C)+d(E,C)+d(B,D)+d(E,D)+d(B,A)+d(E,A)}{6}$$

$$d(BE,CDA) = \frac{6.08+5+4,47+4,24+5,1+4,47}{6} = 4,89$$

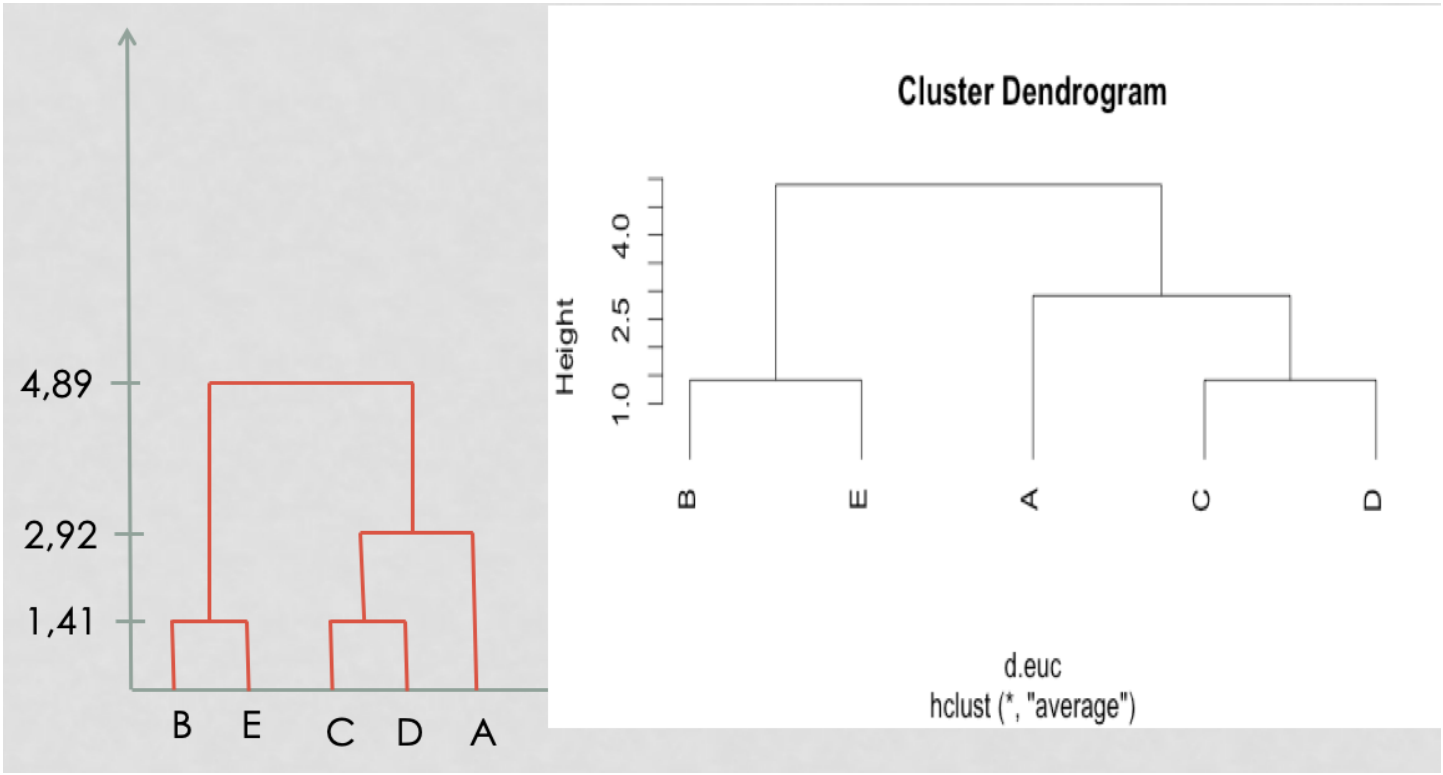


Anne Badel, Frédéric Guyon & Jacques van Helden

- à l'étape $(n-1)$, tous les individus sont regroupés dans un même cluster

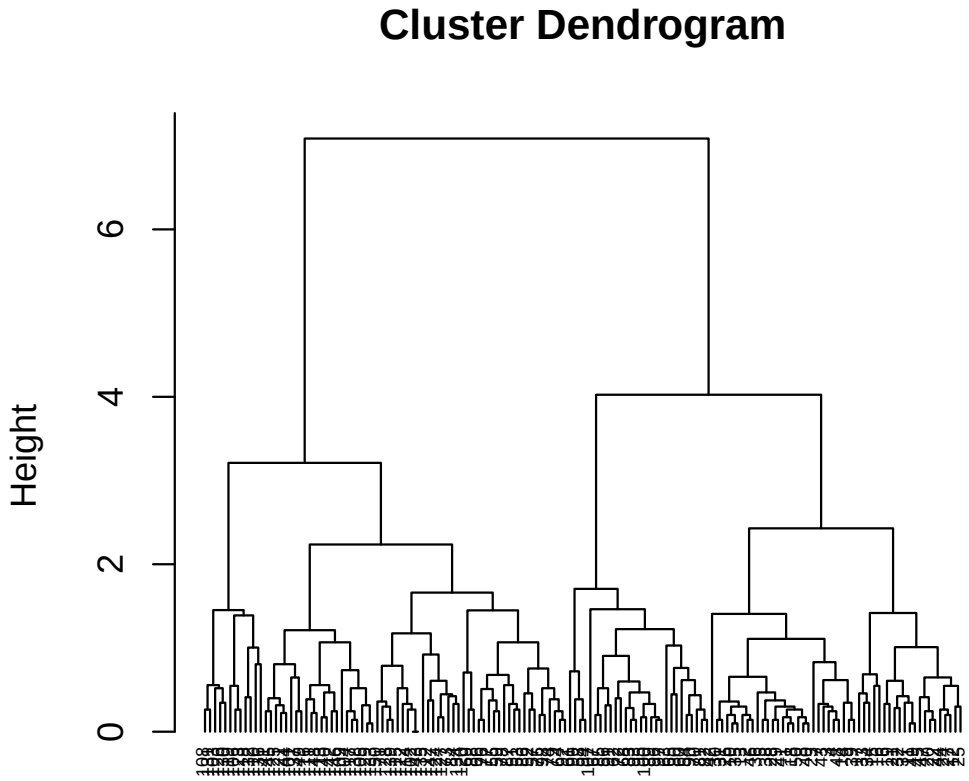
Anne Badel, Frédéric Guyon & Jacques van Helden

dendrogramme final



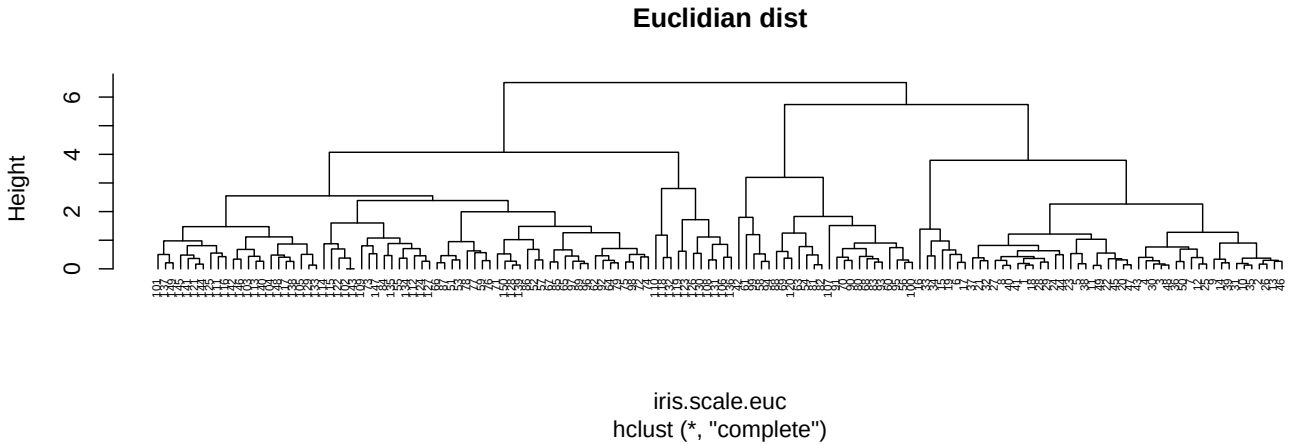
Je ne fais pas attention à ce que je fais . . .

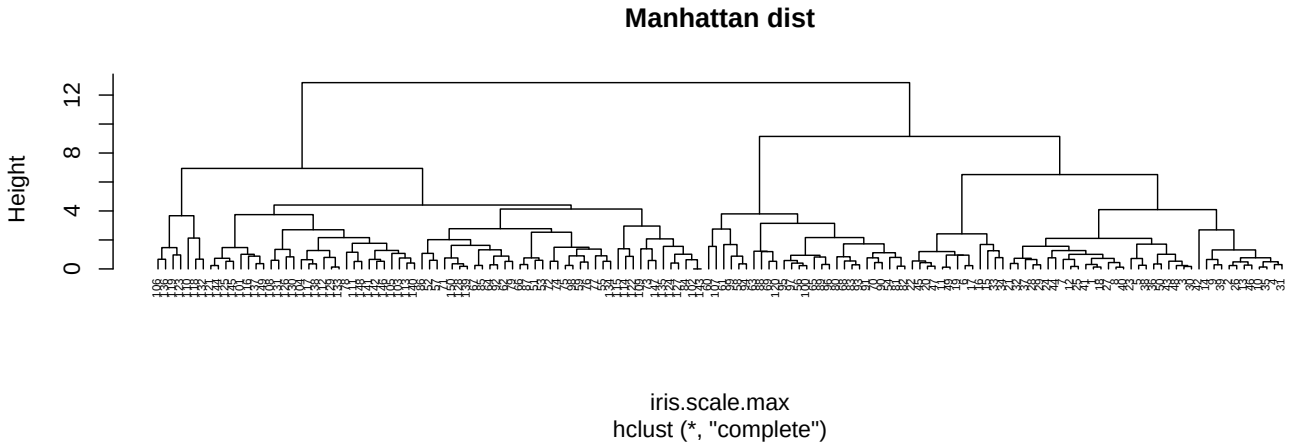
. . . c'est à dire aux options des fonctions `dist()` et `hclust()`



Anne Badel, Frédéric Guyon & Jacques van Helden

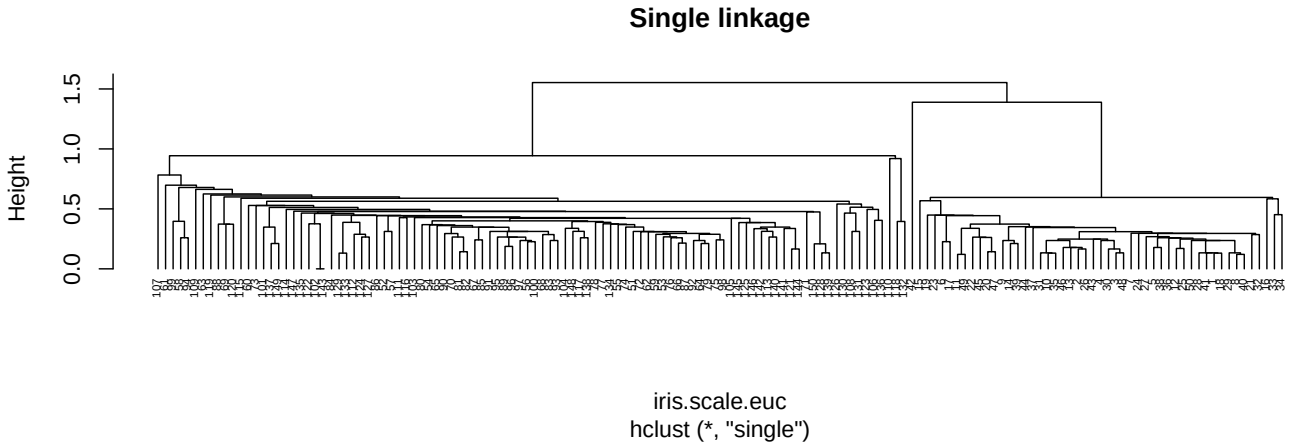
En utilisant une autre métrique

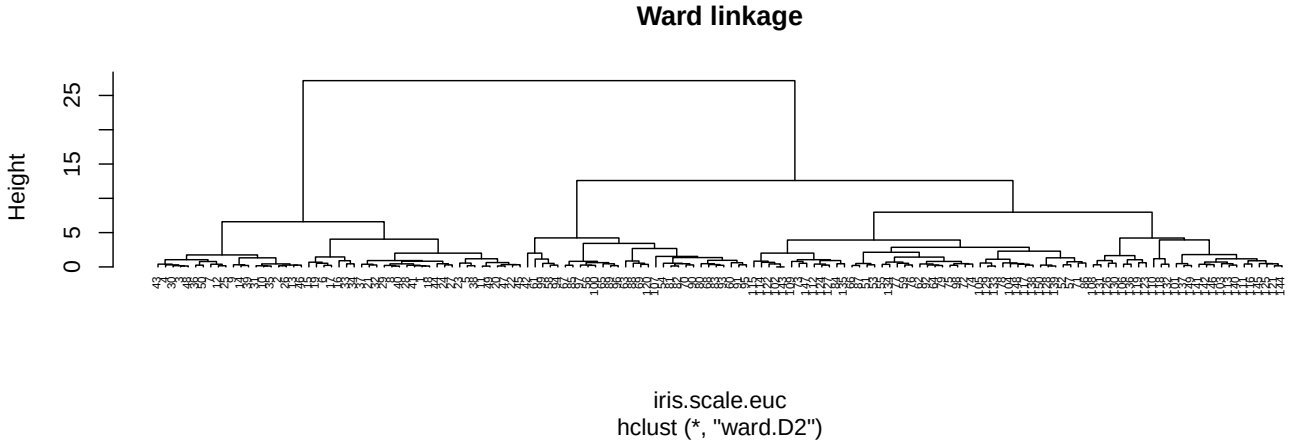




Anne Badel, Frédéric Guyon & Jacques van Helden

En utilisant un autre critère d'aggrégation





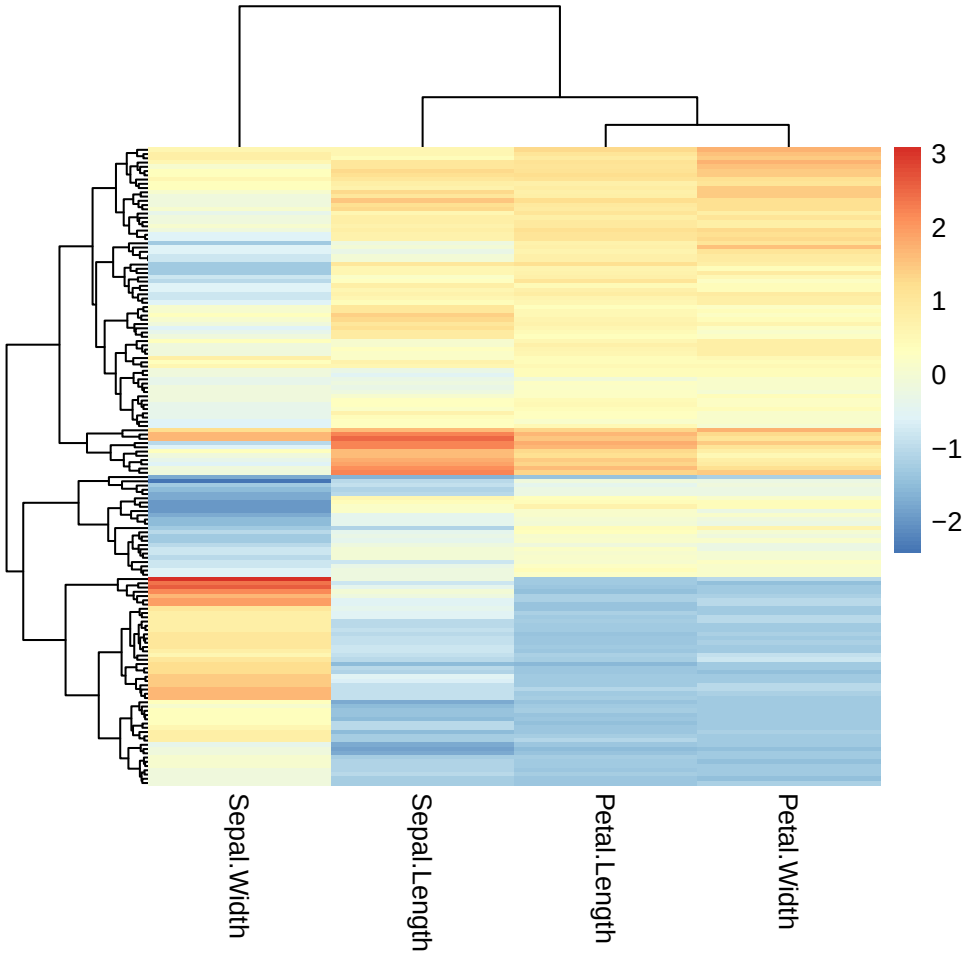
En conclusion

- Faire attention au données
  - données manquantes
  - données invariantes
  - données normalisées
- Choisir la distance et le critère d'aggrégation adaptés à nos données

Anne Badel, Frédéric Guyon & Jacques van Helden

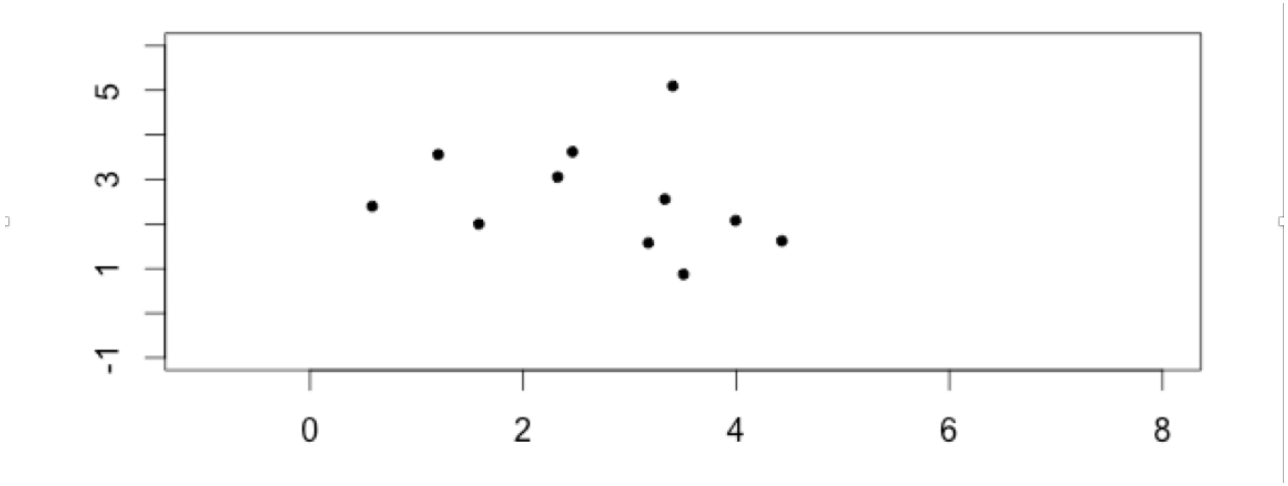
Les heatmap

```
pheatmap::pheatmap(mes.iris.scaled)
```



Les k-means

Les individus dans le plan



$=>$ faire apparaitres des classes / des clusters

L'algorithme

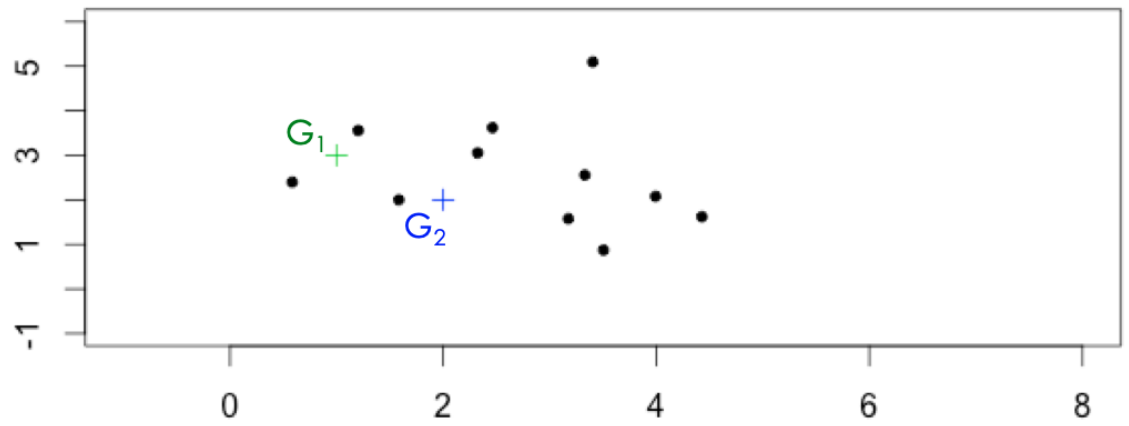
étape 1 :

- k centres provisoires tirés au hasard
- k clusters créés à partir des centres en regroupant les individus les plus proches de chaque centre
- obtention de la partition P_0

Anne Badel, Frédéric Guyon & Jacques van Helden

Choix des centres provisoires

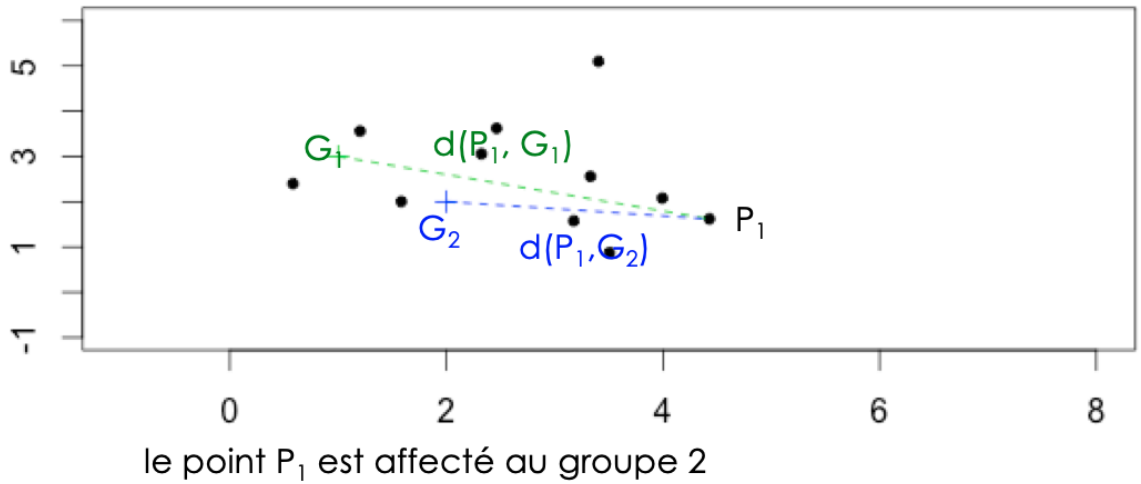
combien de cluster ?
les deux centres initiaux (G_1 et G_2) sont choisis au hasard



Anne Badel, Frédéric Guyon & Jacques van Helden

Calcul des distances aux centres provisoires

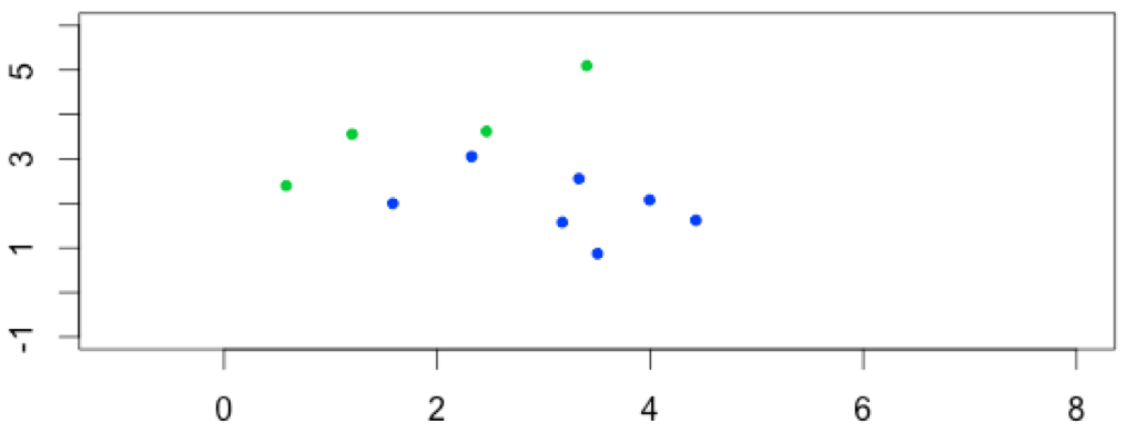
- calcul des distances de chaque point aux centres G_1 et G_2,



le point P_1 est affecté au groupe 2

Anne Badel, Frédéric Guyon & Jacques van Helden

Affectation à un cluster

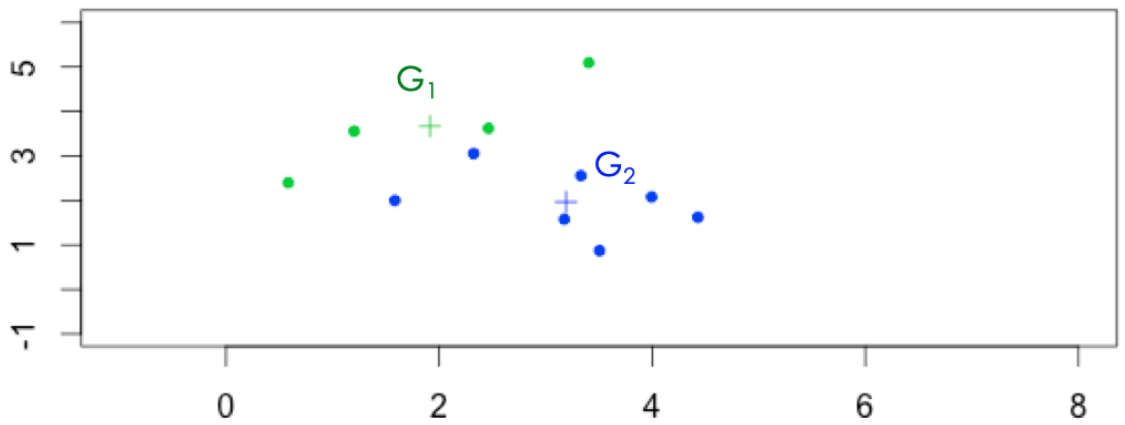


Anne Badel, Frédéric Guyon & Jacques van Helden

Calcul des nouveaux centres de classes

Etape j :

- construction des centres de gravité des k clusters construits à l'étape $(j-1)$
- k nouveaux clusters créés à partir des nouveaux centres suivant la même règle qu'à l'étape 0
- obtention de la partition P_j



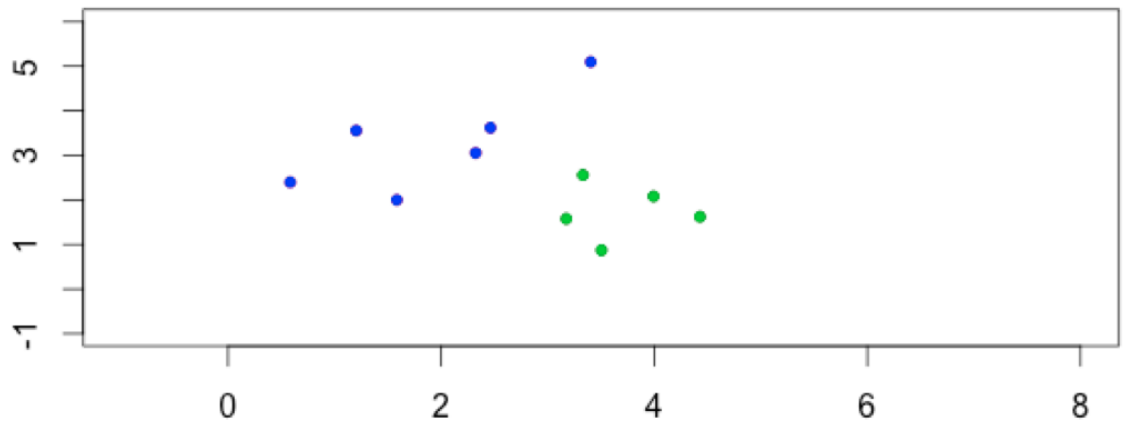
Anne Badel, Frédéric Guyon & Jacques van Helden

Fin :

- l'algorithme converge vers une partition stable

Arrêt :

- lorsque la partition reste la même, ou lorsque la variance intra-cluster ne décroit plus, ou lorsque le nombre maximal d'itérations est atteint.



Anne Badel, Frédéric Guyon & Jacques van Helden

Un premier k-means en 5 groupes

```
iris.scale.kmeans5 <- kmeans(mes.iris.scaled, center=
iris.scale.kmeans5
K-means clustering with 5 clusters of sizes 50, 34, 2

Cluster means:
  Sepal.Length Sepal.Width Petal.Length   Petal.Width
1    0.3558492  -0.3930869    0.5846038  0.5466361525
2   -1.1924784   0.4015443   -1.3090939 -1.2608902424
3   -0.3628650  -1.4097814    0.1074147  0.0008746178
4   -0.6259564   1.8042613   -1.2826445 -1.2290567253
5    1.3926646   0.2323817    1.1567451  1.2132759051

Clustering vector:
  [1] 2 2 2 2 4 4 2 2 2 2 4 2 2 2 4 4 4 2 4 4 2 4 2 2
[148] 1 5 1

Within cluster sum of squares by cluster:
[1] 29.59039 15.97485 11.95194  6.45758 26.89129
 (between_SS / total_SS =  84.8 %)
```

Anne Badel, Frédéric Guyon & Jacques van Helden

Comment déterminer le nombre de clusters ? (1)

Ces méthodes non supervisées, sont sans *a priori* sur la structure, le nombre de groupe, des données.

rappel : un cluster est composé

- d'individus qui se ressemblent
- d'individus très différents des individus de ceux des autres clusters

Anne Badel, Frédéric Guyon & Jacques van Helden

Comment déterminer le nombre de clusters ? (2)

- si les individus d'un même cluster sont proches
  - homogénéité maximale à l'intérieur de chaque cluster => variance intra faible
- si les individus de 2 clusters différents sont éloignés => variance inter forte
  - hétérogénéité maximale entre chaque cluster

Anne Badel, Frédéric Guyon & Jacques van Helden

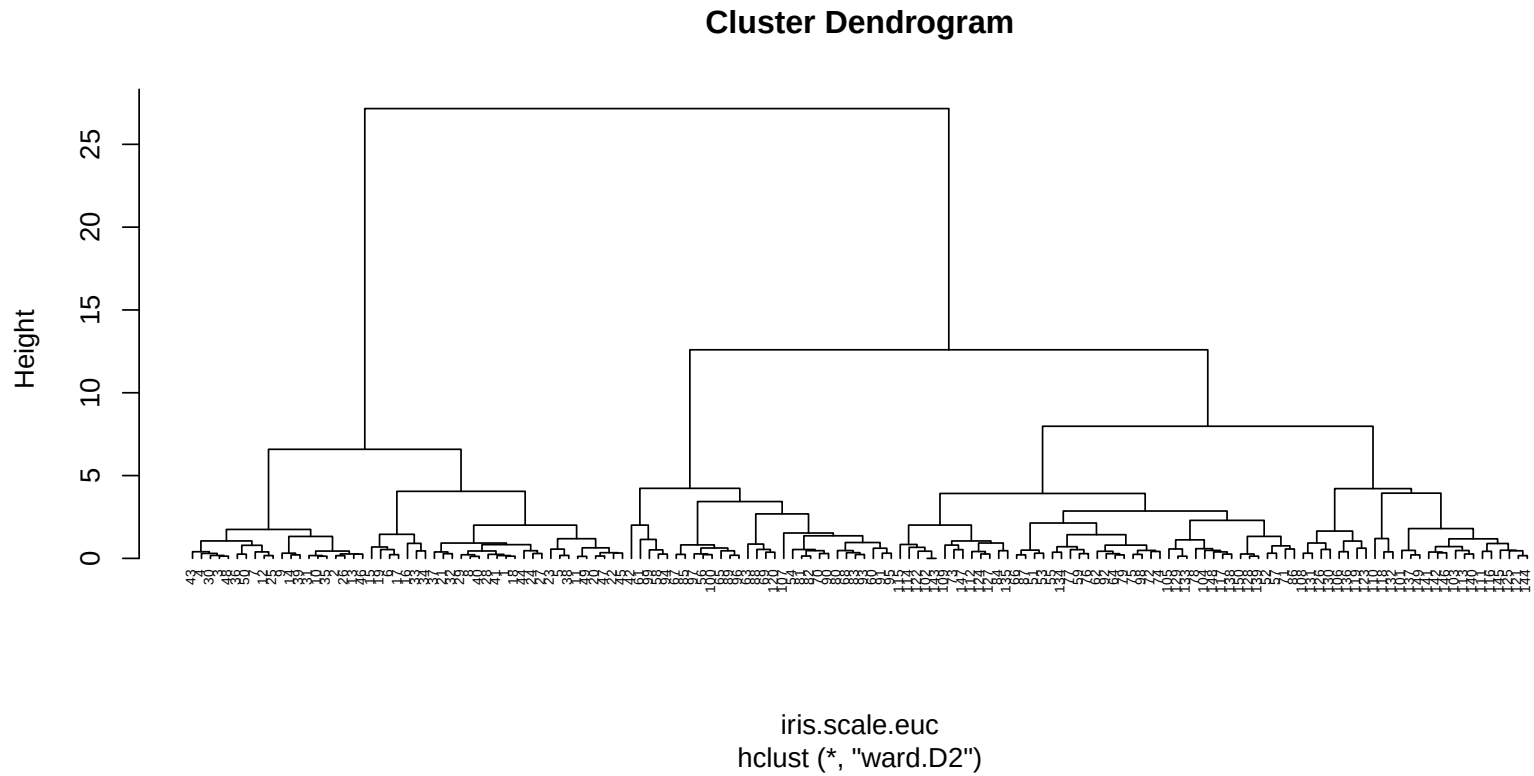
Comment déterminer le nombre de clusters ? avec la classification hiérarchique

La coupure de l'arbre à un niveau donné construit une partition. la coupure doit se faire :

- après les agrégations correspondant à des valeurs peu élevées de l'indice
- avant les agrégations correspondant à des niveaux élevés de l'indice, qui dissocient les groupes bien distincts dans la population.

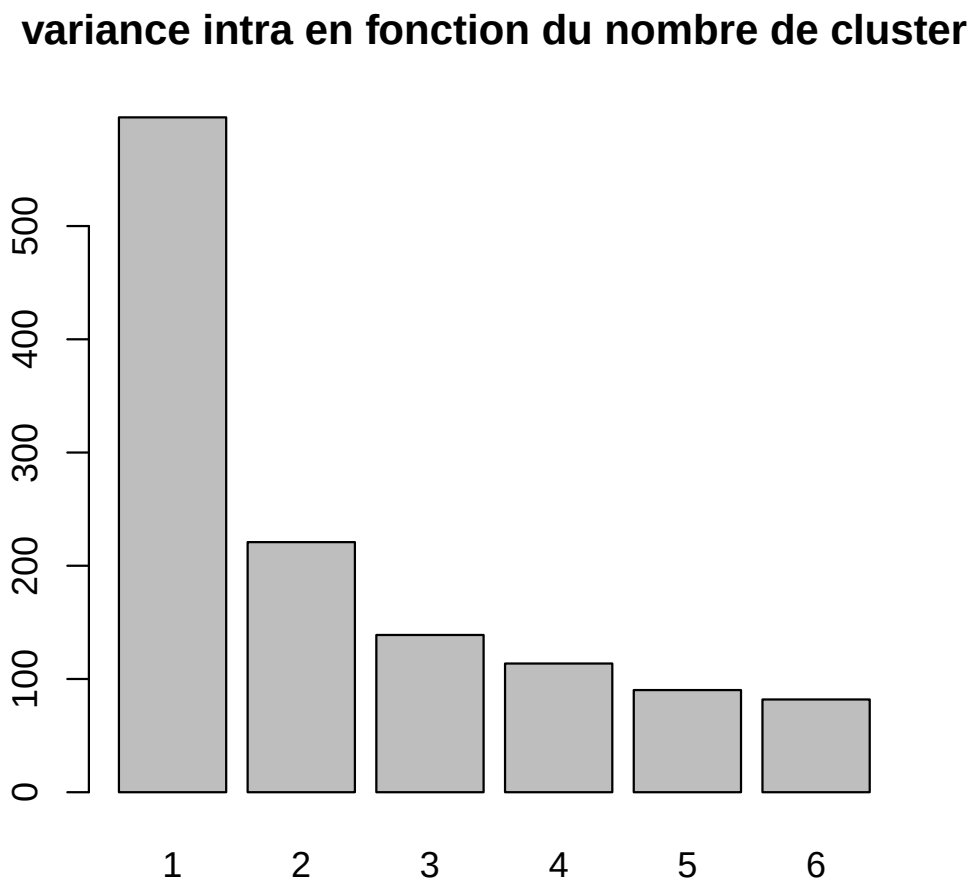
Anne Badel, Frédéric Guyon & Jacques van Helden

```
plot(iris.scale.hclust.ward, hang=-1, cex=0.5)
```



Anne Badel, Frédéric Guyon & Jacques van Helden

Comment déterminer le nombre de clusters ? avec les kmeans



Anne Badel, Frédéric Guyon & Jacques van Helden

Comparaison de clustering: Rand Index

Mesure de similarité entre deux clustering

à partir du nombre de fois que les classifications sont d'accord

$$R = \frac{m + s}{t}$$

- m=nombre de paires dans la même classe dans les deux classifications
- s=nombre de paires séparées dans les deux classifications
- t=nombre de paires totales

Anne Badel, Frédéric Guyon & Jacques van Helden

Comparaison de clustering: Adjusted Rand Index

$$ARI = \frac{RI - ExpectedRI}{MaxRI - ExpectedRI}$$

- ARI=RI normalisé
- Prend en compte la taille des classes
- ARI=1 pour classification identique
- ARI $\simeq$ 0 pour classification aléatoire (peut être $<$0)
- Adapté pour nombre de classe différent entre les deux classifications et taille de classe différente

Anne Badel, Frédéric Guyon & Jacques van Helden

Comparaison des résultats des deux clustering

- par une table de confusion

0	29	0
0	20	0
29	1	0
24	0	21
0	0	26

Anne Badel, Frédéric Guyon & Jacques van Helden

Comparaison de clustering: Rand Index

Mesure de similarité entre deux clustering

à partir du nombre de fois que les classifications sont d'accord

$$R = \frac{m + s}{t}$$

- m = nombre de paires dans la même classe dans les deux classifications
- s = nombre de paires séparées dans les deux classifications
- t = nombre de paires totales

Anne Badel, Frédéric Guyon & Jacques van Helden

Comparaison de clustering: Adjusted Rand Index

$$ARI = \frac{RI - ExpectedRI}{MaxRI - ExpectedRI}$$

- ARI=RI normalisé
- Prend en compte la taille des classes
- ARI=1 pour classification identique
- ARI $\simeq$ 0 pour classification aléatoire (peut être $<$0)
- Adapté pour nombre de classe différent entre les deux classifications et taille de classe différente

Anne Badel, Frédéric Guyon & Jacques van Helden

Comparaison des résultats des deux classifications

- rand index et adjusted rand index

```
clues::adjustedRand(cluster.hclust5, cluster.kmeans3)
```

```
     Rand         HA         MA         FM    Jaccard
0.7848770 0.4637776 0.4730527 0.6167001 0.4299265
```

Pros et cons des différents algorithmes

Algorithme	Pros	Cons
Hiérarchique	L'arbre reflète la nature imbriquée de tous les sous-clusters	Complexité quadratique (mémoire et temps de calcul) $\rightarrow$ quadruple chaque fois qu'on double le nombre d'individus
	Permet une visualisation couplée dendrogramme (groupes) + heatmap (profils individuels)	
	Choix a posteriori du nombre de clusters	

Algorithme	Pros	Cons
K-means	Rapide (linéaire en temps), peut traiter des jeux de données énormes (centaines de milliers de pics ChIP-seq)	Positions initiales des centres est aléatoire $\rightarrow$ résultats changent d'une exécution à l'autre Distance euclidienne (pas appropriée pour transcriptome par exemple)

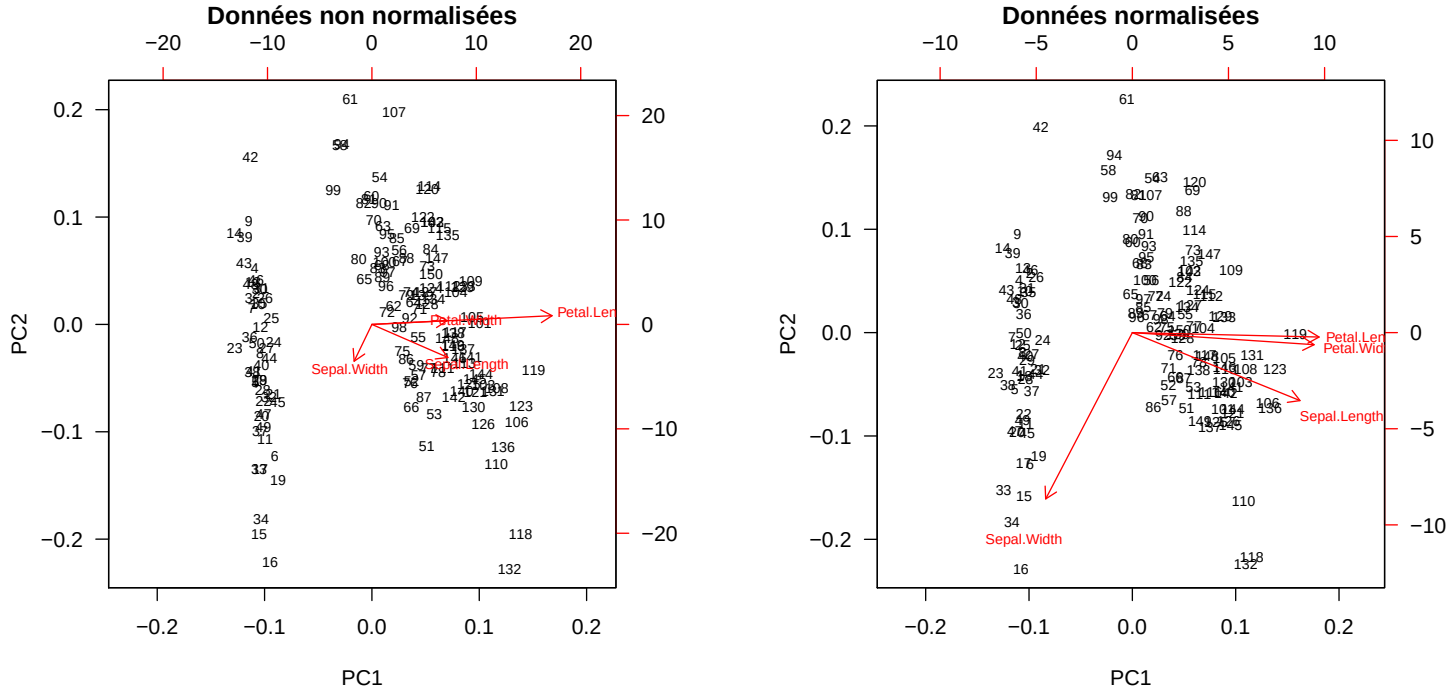
Anne Badel, Frédéric Guyon & Jacques van Helden

```
## Print the complete list of libraries + versions us
sessionInfo()

R version 3.6.1 (2019-07-05)
Platform: x86_64-apple-darwin15.6.0 (64-bit)
Running under: macOS Catalina 10.15.3

Matrix products: default
BLAS:   /Library/Frameworks/R.framework/Versions/3.6/
LAPACK: /Library/Frameworks/R.framework/Versions/3.6/

locale:
[1] fr_FR.UTF-8/fr_FR.UTF-8/fr_FR.UTF-8/C/fr_FR.UTF-8

attached base packages:
[1] stats     graphics  grDevices utils     datasets

other attached packages:
 [1] pheatmap_1.0.12    vegan_2.5-6        lattice_0.
```

Anne Badel, Frédéric Guyon & Jacques van Helden

. . . par une projection sur une ACP

```
par(mfrow = c(1,2))
biplot(prcomp(mes.iris), las = 1, cex = 0.7,
       main = "Données non normalisées")
biplot(prcomp(mes.iris, scale = TRUE), las = 1, cex =
       main = "Données normalisées")
```



Anne Badel, Frédéric Guyon & Jacques van Helden

Cas d'étude : TCGA Breast Invasive Cancer (BIC)

- Présentation du cas d'étude (Jacques van Helden A COMPLETER)

Anne Badel, Frédéric Guyon & Jacques van Helden

TP : analyse de données d'expression

- TP clustering : [html] [pdf] [Rmd]
- Première partie : chargement des données

Anne Badel, Frédéric Guyon & Jacques van Helden

Contact: anne.badel@univ-paris-diderot.fr