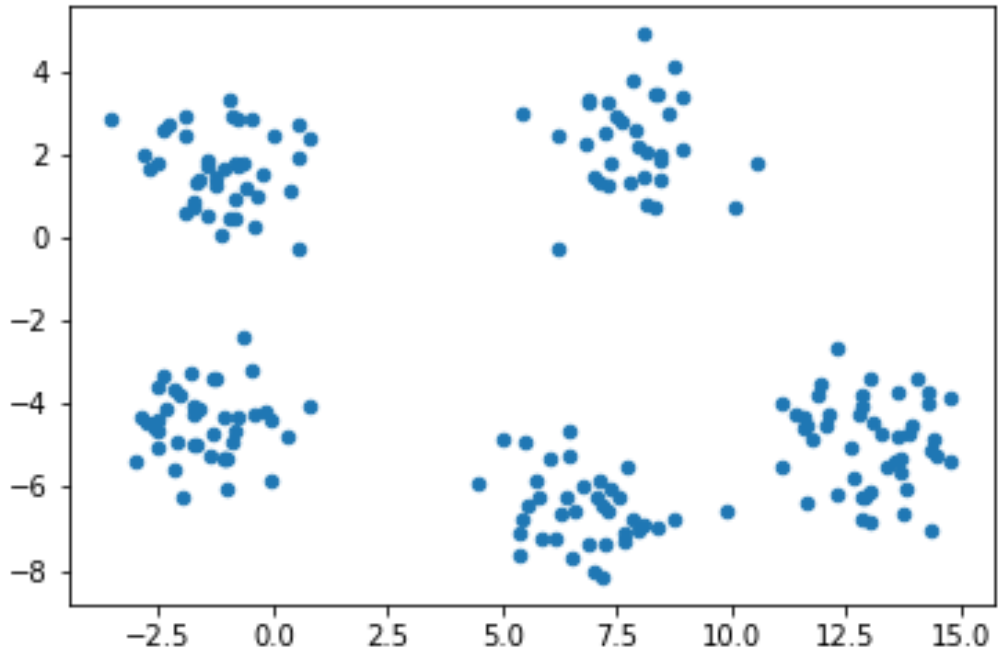
# INTRODUCTION À LA SCIENCE DES DONNÉES
# CLASSIFICATION NON SUPERVISÉE

V. Monbet

[1] Université de Rennes/UFR Mathématiques

# Outline

# Exemple d'un jeu de donnée



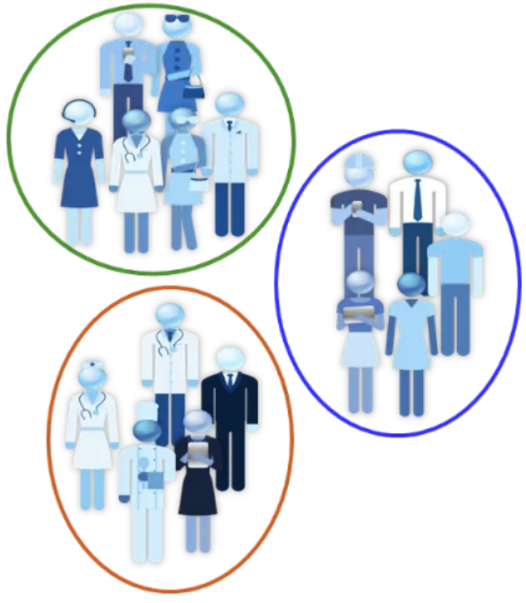

Que pouvons nous remarquer ?
Comment pourrions nous synthétiser l'information ?

# Pouvons-nous constituer des groupes d'individus ?

source

Nous allons devoir identifier

- - le nombre de classes
- - l'appartenance à une classe pour chaque individu



source de l’image

Ici, on a choisi de retenir 3 classes.

# Pourquoi chercher à regrouper des individus ou observations

- Mettre en évidence des régimes météorologiques, des sous familles de plantes ou d'animaux, ...
- Mieux traiter des patients en identifiant leur appartenance à certain sous-type de pathologie
- lancer des campagnes marketing
- etc

# Classification "non supervisée"

La **classification non supervisée** (ou clustering) est la **recherche d'une partition** ou d'une répartition des individus en **classes** homogènes, de sorte à ce que deux individus qui appartiennent à la même classe soient aussi semblables que possible et deux individus qui appartiennent à deux classes différent l'un de l'autre.

Définition des termes

- **Classification** : Regrouper des individus en groupes ou classes d'individus proches en un certain sens.
- **Non supervisée** : dont on ne connait pas la réalité, on n'a pas d'information a priori sur les classes.

- Attention : en anglais, *classification* $\neq$ *clustering*.

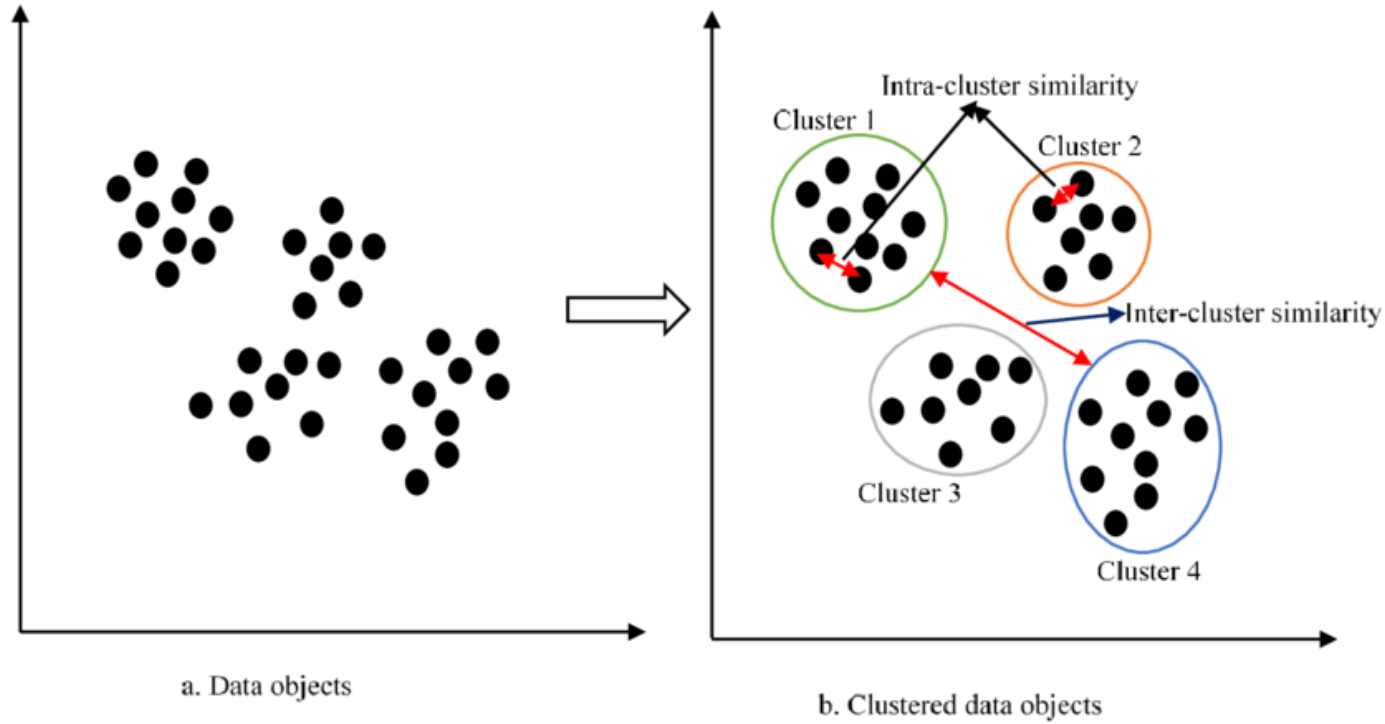
# De quoi avons nous besoin ?

Pour construire des algorithmes de classification non supervisée, nous avons besoin d'un critère à optimiser qui prenne en compte

- les distances ou les similarités entre les individus : *les individus d'une même classe sont plus proches les uns des autres que des individus de classes différentes* ;
- les distances entre les classes : *les classes sont "éloignées" les unes des autres*

On ajoute éventuellement à déplacer

- Un critère d'homogénéité des classes : *deux individus qui appartiennent à la même classe sont aussi semblables que possible*




source de l'image

# Différents algorithmes de classifications non supervisée

Il existe de nombreux algorithmes de classifications non supervisées. Les plus connus sont les kmeans et la classification hiérarchique ascendante. Un des plus utilisés en apprentissage machine est DBSCAN.

- Les **Kmeans**, Kmedoïd, ... sont des algorithmes dans lesquels on optimise un critère global (ie défini pour tout l'échantillon observé) par un algorithme itératif.

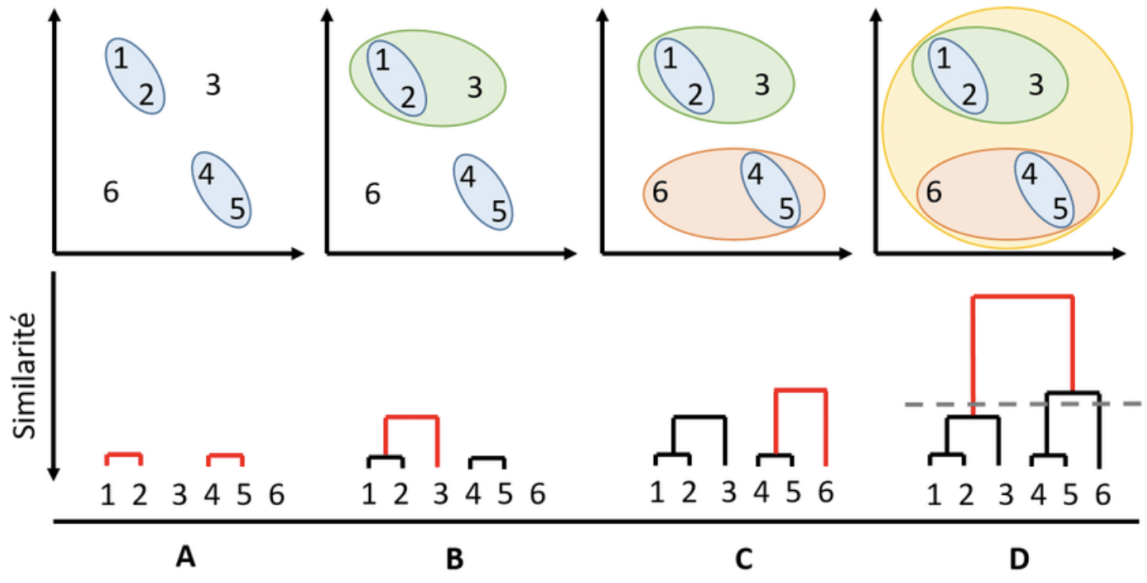
- La **classification hiérarchique ascendante**, **DBSCAN**, ... sont des algorithmes qui optimise un critère local (ie ne portant sur sur quelques individus et leurs voisins) à chaque étape de l'algorithme.

Nous reviendrons plus loin sur les avantages et inconvénients des différentes méthodes.

# Outline

# Classification Hiérarchique Ascendante

La classification hiérarchique ascendante (CAH) est une méthode de regroupement des objets en clusters de manière itérative, où chaque objet commence (seul) dans son propre groupe et des groupes de plus en plus grands sont formés en fusionnant les deux groupes les plus similaires à chaque étape.



# Outline

# Mesure d'éloignement

Pour mettre en oeuvre la classification hiérarchique ascendante on doit choisir d'abord choisir une mesure de (dis)similarité entre individus, puis une distance entre classes.

# Dissimilarité entre individus

Dans le cas de **données quantitatives**, on utilise le plus souvent une distance.

Soient, $x$ et $y$ deux vecteurs de $\mathbb{R}^p$, correspondant à deux observations.

- distance euclidienne (ou $L^2$)

$$||x - y||_2 = \sqrt{\sum_{j=1}^{p} (x_i - y_i)^2}$$

- distance Manhattan (ou $L^1$)

$$||x - y||_1 = \sum_{j=1}^{p} |x_i - y_i|$$

a privilégier en grande dimension ou pour des données comportant beaucoup de 0 ou si les données sont organisées en grille

- distance cosinus

$$\sqrt{2(1 - S_C(x, y))} \text{ avec } S_C(x, y) = \frac{xy}{||x||||y||}$$

pour des données compositionnelles, l'analyse de données textuelles, des expressions de gènes, ...

Dans le cas de **données qualitatives**, on utilise le plus souvent des indices.
En particulier, si les données sont binaires, $x$, $y$ dans $\{0, 1\}$.

Pour deux individus $i$ et $j$

$a_{ij}$ le nombre de caractères communs entre i et j

$b_{ij}$ le nombre de caractères de i que n'a pas j

$c_{ij}$ le nombre de caractères de j que n'a pas i

$d_{ij}$ le nombre de caractères que n'ont ni i ni j

avec $a_{ij} + b_{ij} + c_{ij} + d_{ij} = p$

Les indices de ressemblance les plus courants sont

- Concordance

$$\frac{a_{ij} + d_{ij}}{p}$$

- Indice de Jaccard

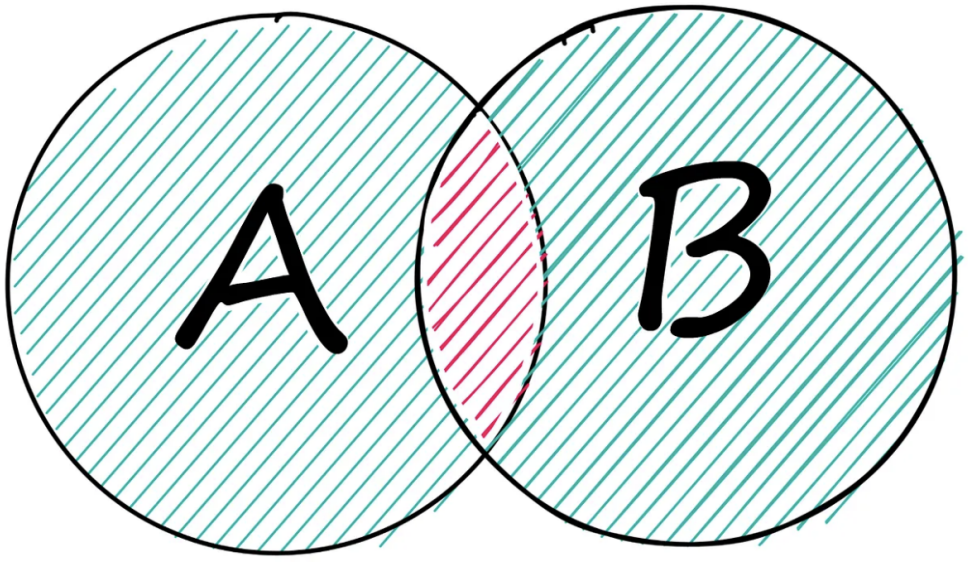
$$\frac{a_{ij}}{a_{ij} + b_{ij} + c_{ij}}$$

- Indice de Dice

$$\frac{2a_{ij}}{2a_{ij} + b_{ij} + c_{ij}}$$

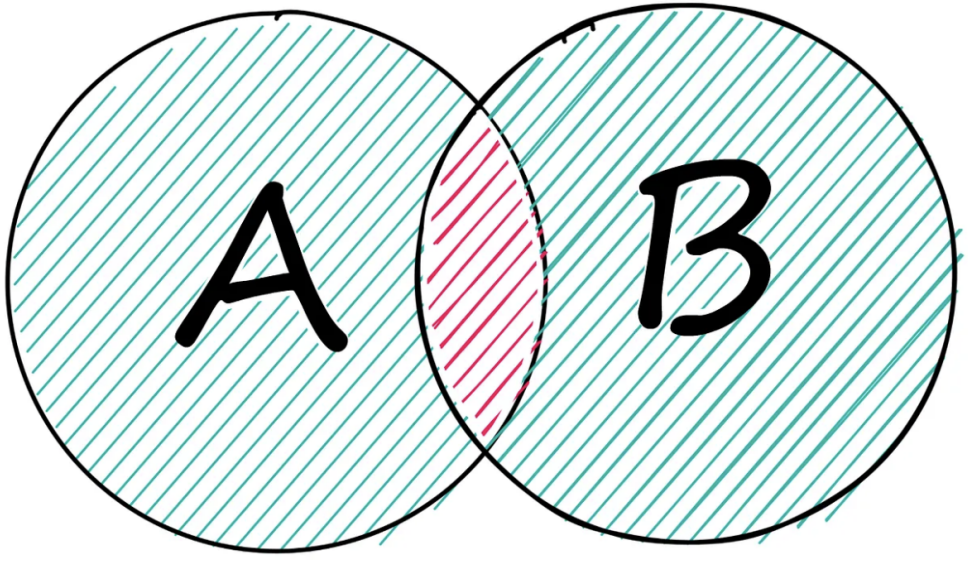
Par exemple, calculer l'indice de Jaccard entre les ensembles
Set A = {"Lion" , "Tiger" , "Cheetah" , "Leopard" , "Rhino"}
Set B = {"Lion" , "Monkey" , "Cheetah" , "Cat" , "Dog" }

$$Jaccard = \frac{intersection(A, B)}{union(A, B)}$$



Par exemple, calculer l'indice de Jaccard entre les ensembles
Set A = {"Lion" , "Tiger" , "Cheetah" , "Leopard" , "Rhino"}
Set B = {"Lion" , "Monkey" , "Cheetah" , "Cat" , "Dog" }

$$Jaccard = \frac{intersection(A, B)}{union(A, B)}$$



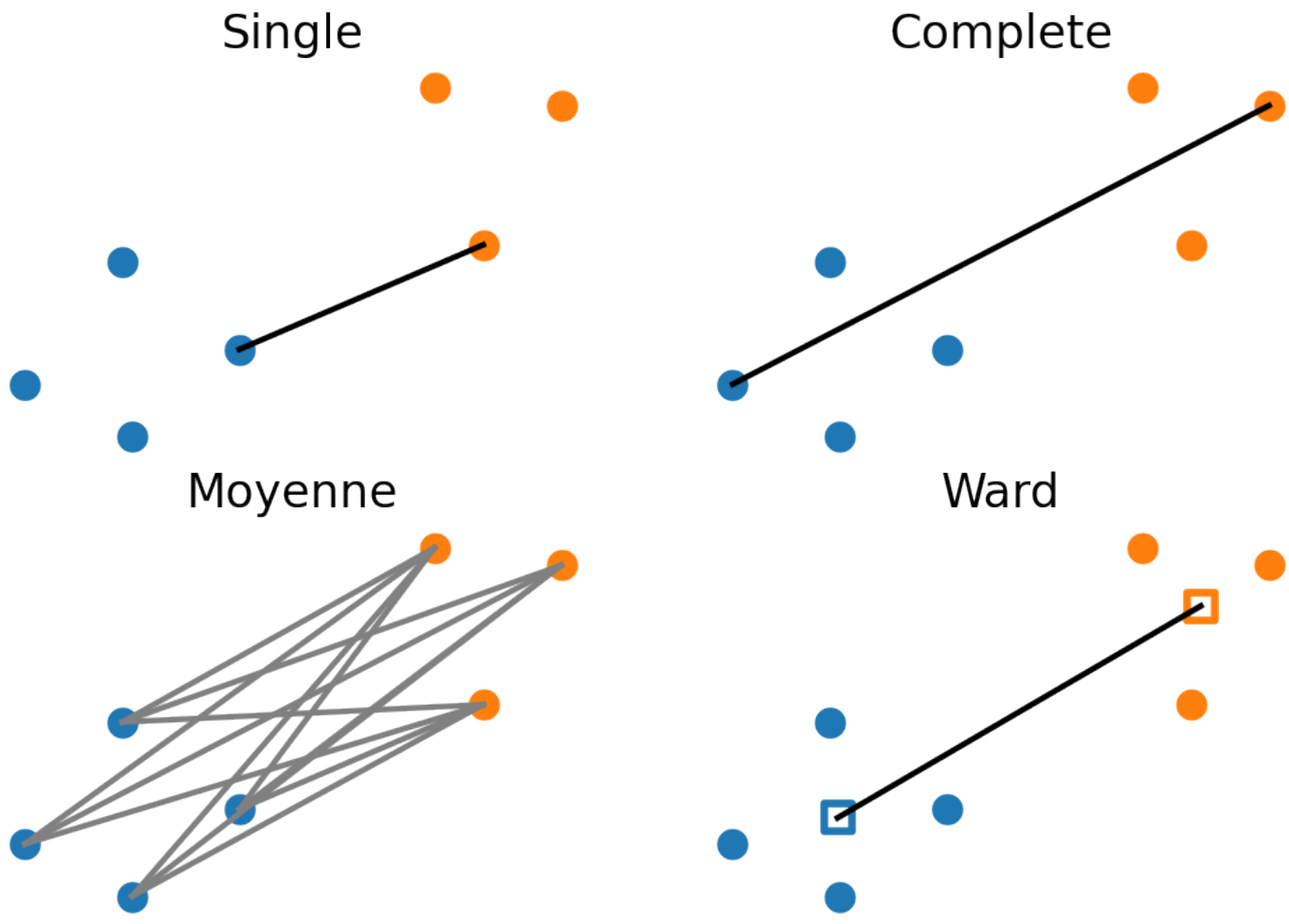

On remarque que
$(A \cap B)$ = {"Lion" , "Cheetah"} = 2
$(A \cup B)$ = {"Lion" , "Tiger" , "Cheetah" , "Leopard" , "Rhino" , "Monkey" , "Cat" , "Dog"} = 8
Ainsi

$$J(A, B) = \frac{[(A \cap B|}{|(A \cup B|} = 0.25$$

**Cas mixte** :

- Se ramener au *cas qualitatif* : séparer les valeurs d'une variable quantitative dans des classes (généralement définies par des quantiles)
- Se ramener au *cas quantitatif* : utiliser la mesure d'éloignement générale sur les variables qualitatives, réduire en classes les variables quantitatives, puis refaire la méthode générale sur l'ensemble des variables.

# Mesure d'éloignement **entre classes**



# Mesure d'éloignement **entre classes**

Quelques exemples dans le cas de données quantitatives

Cas d'une dissimilarité :

- *Single linkage* : $d(A,B) = \min_{x \in A, Y \in B} d(x,y)$
- *Complete linkage* : $d(A,B) = \max_{x \in A, Y \in B} d(x,y)$
- *Group Average linkage* : $d(A,B) = \frac{1}{|A||B|} \sum_{x \in A, Y \in B} d(x,y)$

Cas d'une distance euclidienne :

- *Centroïd, Distance des barycentres* : $d(g_A, g_B)$
- *Distance de Ward* : $\frac{|A||B|}{|A|+|B|} d(g_A, g_B)$

# Outline

# Classification ascendante hiérarchique

## Algorithme

- Initialiser les classes par les individus
- Calculer la matrice des distance deux é deux
- Répéter jusqu'à agrégation en une seule classe :
  - Fusionner les deux classes les plus proches
  - Calculer la matrice des distances avec les nouvelles classes

On obtient alors un arbre de classification, appelé **dendrogramme**.

# Exemple

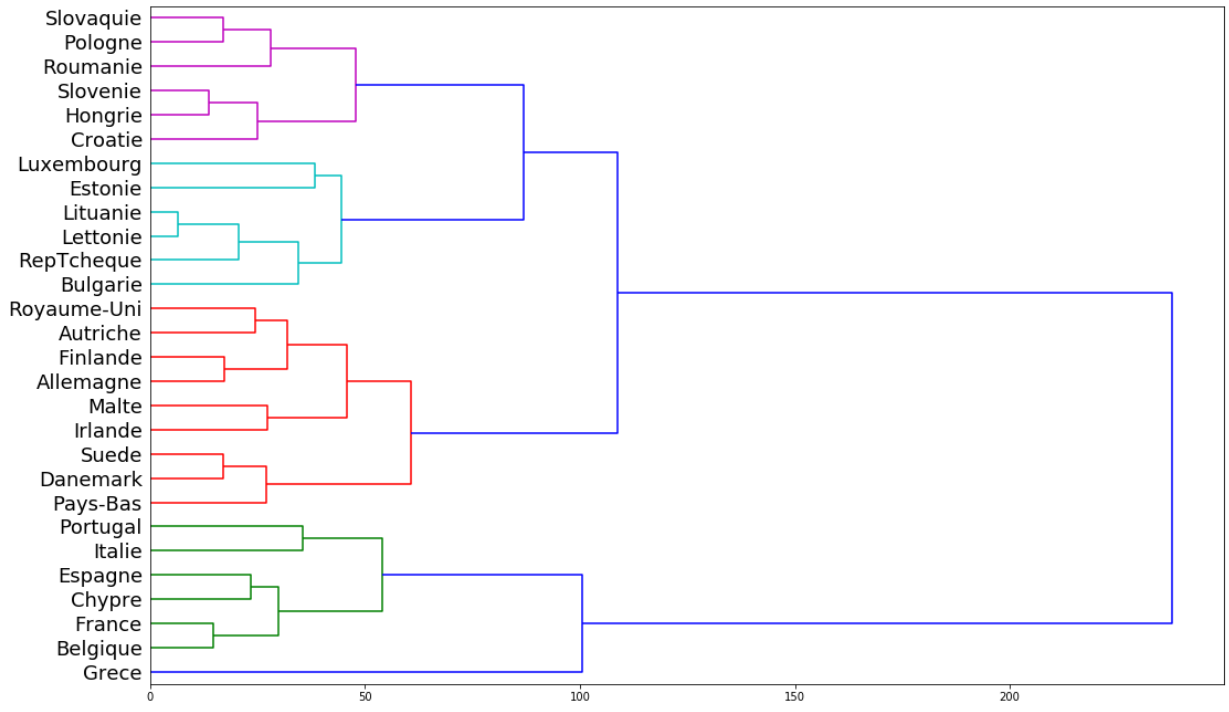



Figure – Dendrogramme des pays européens en fonction de plusieurs variables (taux de chémage, utilisation et accés é internet, argent public...), obtenu par la mesure de dissimilarité de Ward. Les variables sont quantitatives.

On représente en abscisse la **dissimilarité** entre les deux classes fusionnées.

# Exemple - Utilité de la standardisation des variables

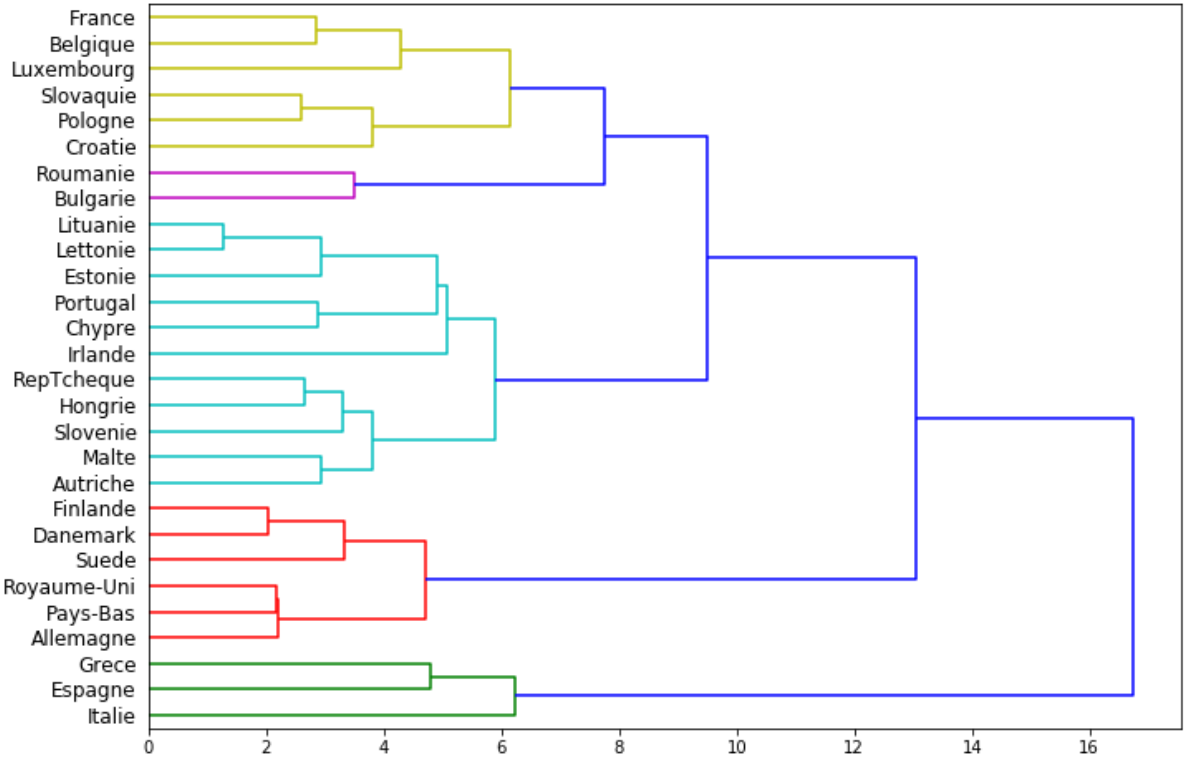



Figure – Dendrogramme des pays européens en fonction de plusieurs variables (taux de chémage, utilisation et accés é internet, argent public...), obtenu par la mesure de dissimilarité de Ward. Les variables sont quantitatives et centrées réduites.

# Outline

# Choix du nombre de classes

Il est maintenant nécessaire de choisir le **nombre de classes** qui nous intéresse.
Pour cela, il existe plusieurs méthodes...

# Choix du nombre de classes -Variance intra/interclasse

Rappel sur la variance :
Variance de E un ensemble de point est

$$V(E) = \frac{1}{|E|} \parallel x - g_E \parallel^2, \qquad g_E = \frac{1}{|E|} \sum_{x \in E} x$$

Soit une partition $A = A_1 \cup \cdots \cup A_q$, la variance vaut :

$$V(E) = \sum_i p_i \parallel g_{A_i} - g_E \parallel^2 + \sum_i p_i V(A_i), \qquad p_i = \frac{|A_i|}{|E|}$$

Variance totale = Variance interclasse + Variance intraclasse

En notant $\forall X \in A_i, \widehat{x} = g_{A_i}$,

$$\sum_x \parallel x - g_E \parallel^2 = \sum_x \parallel \widehat{x} - g_E \parallel^2 + \sum_x \parallel x - \widehat{x} \parallel^2$$

# Choix du nombre de classes - Variance intra/interclasse

On peut représenter le ratio de la variance intraclasse sur la variance interclasse, en fonction du nombre de classes.

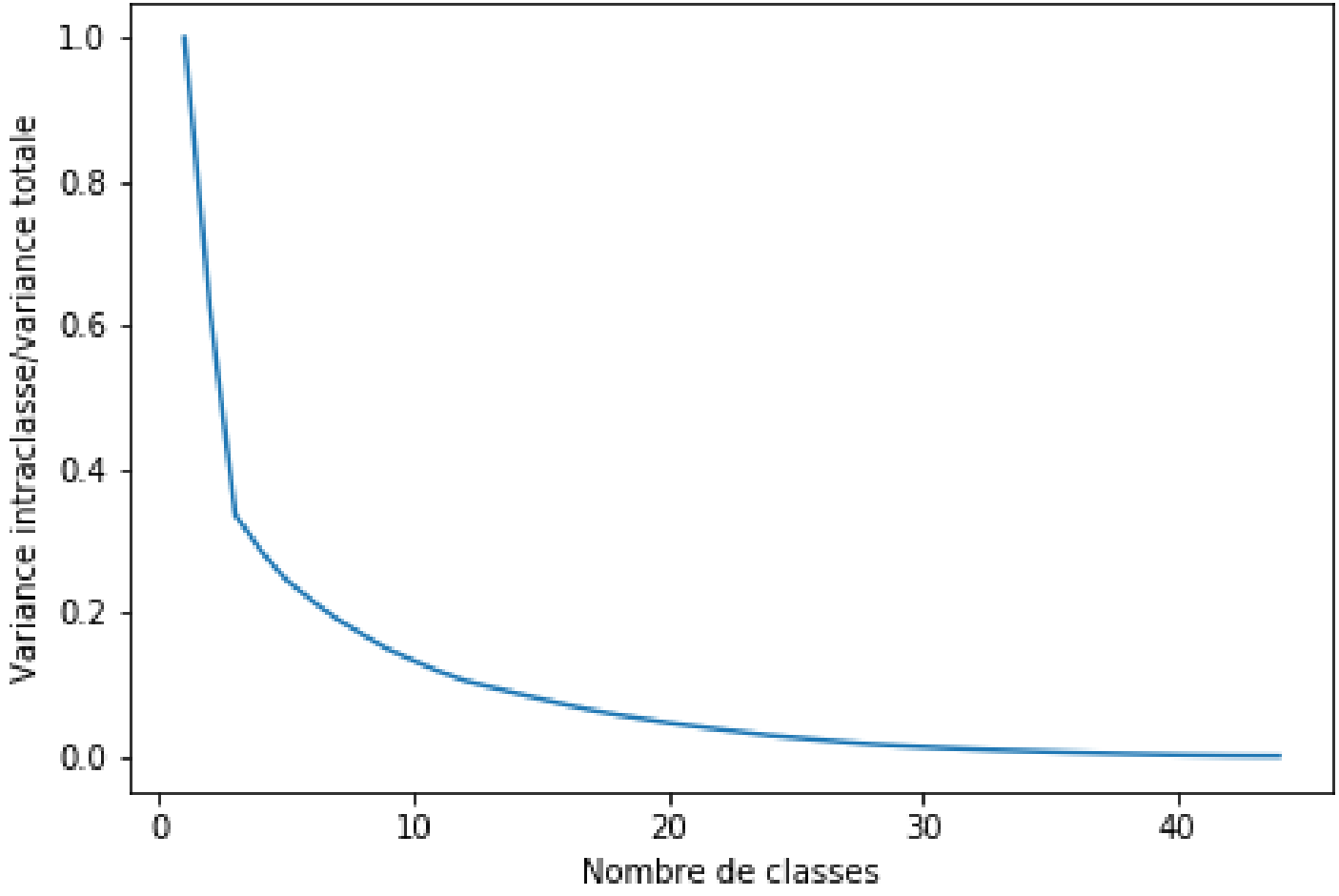



Figure – Ratio Variance intraclasse sur variance interclasse pour l'exemple précédent.

# Interprétation de la méthode de Ward

Par un rapide calcul, la distance de Ward peut se réécrire

$$\begin{aligned} d_{Ward}(A,B) &= \frac{|A||B|}{|A|+|B|} d(g_A, g_B) \\ &= \sum_{x \in A \cup B} \| x - g_{A\cup B} \|^2 - \sum_{x \in A} \| x - g_A \|^2 - \sum_{x \in B} \| x - g_B \|^2 \end{aligned}$$

Réunir deux classes avec $d_{Ward}(A,B)$ minimum revient é faire la fusion qui **augmente le moins la variance intraclasse**, car cette valeur correspond la différence avant et aprés fusion.

Ainsi, avec la distance de Ward, l'algorithme diminue le nombre de classe en augmentant le moins possible la variance intraclasse.

# Choix du nombre de classes - Interprétation par visualisation

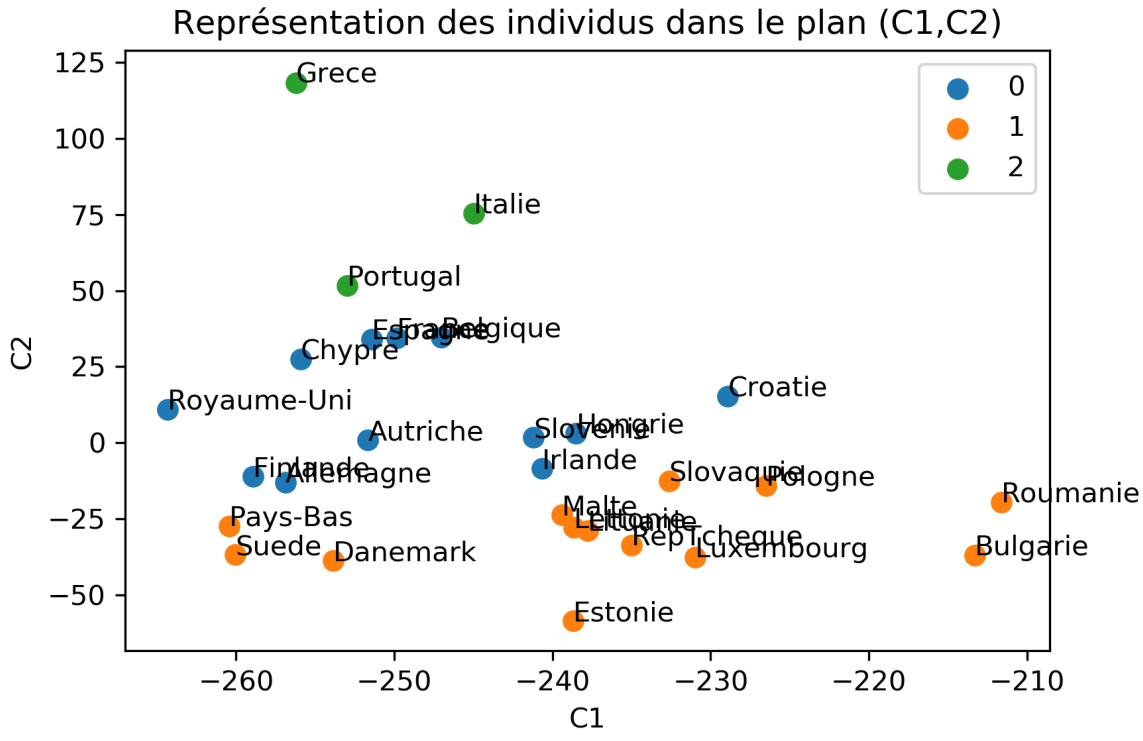



Figure – ACP é deux dimensions sur notre exemple précédent.

Il est possible aussi de regarder les boxplots par classes.

# Choix du nombre de classes - *Adjusted Rand Index*

Soient X et Y deux partitions d'un jeu de données, avec respectivement $r$ et $s$ classes.
On note $n_{ij} = |X_i \cap Y_j|$, $a_i = \sum_j n_{ij}$, $b_j = \sum_i n_{ij}$.

| | $Y_1$ | $\dots$ | $Y_s$ | *sums* |
|---|---|---|---|---|
| $X_1$ | $n_{11}$ | $\dots$ | $n_{1s}$ | $a_1$ |
| $\vdots$ | $\vdots$ | $\ddots$ | $\vdots$ | $\vdots$ |
| $X_r$ | $n_{r1}$ | $\dots$ | $n_{rs}$ | $a_r$ |
| *sums* | $b_1$ | $\dots$ | $b_s$ | |

On définit alors

$$ARI = \frac{\binom{n}{2}\sum_{ij}\binom{n_{ij}}{2} - \left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right]}{\frac{1}{2}\binom{n}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right]}$$

# A retenir

- ▶ La CAH est un algorithme glouton qui renvoie une suite de répartition en classes de plus en plus grandes.
- ✓ L'algorithme est très simple et ne nécessite pas qu'on lui précise le nombre de clusters à trouver.
- ✓ La CAH peut être utilisée avec n'importe quelle distance ou dissimmilarité.
- ✓ La CAH peut être utilisée pour des données qualitative et quantitatives.
- ✓ Les clusters n'ont pas pour obligation d'être linéairement séparables (tout comme pour l'algorithme des k-moyennes par exemple).
- ✓ La représentation sous forme de dendrogramme permet de bien visualiser les regroupement. La longueur des branches décrit le gain en variance intraclasse sur variance totale.
- ✗ La principale limite de la CAH est que l'algorithme est couteux pour les données comportant un grand nombre d'individus.