Ecole Supérieure en Sciences Appliquées de Tlemcen

Analyse de données

Chapitre 6: Classification hiérarchiques et Non hiérarchiques

Présentée par:

Dr Imane NEDJAR

Introduction

Analyses de données

Les modèles statistiques

Sont utilisés pour nettoyer les données au début par l'élimination des valeurs aberrantes, et aussi de visualiser les données, afin de construire l'ensemble initial d'exemples.

Les modèles factorielles

Cherchent à réduire le nombre de variables en les résumant par un petit nombre de composantes synthétiques en utilisant essentiellement des outils de l’algèbre linéaire.

Les modèles classification

Construisent des règles et des modèles prédictifs pour synthétiser et structurer l'information contenue dans des données.

Introduction

•**Les modèles de classification** basent sur la constitution des groupes (classes, clusters) de manière à ce que les individus dans un même groupe se ressemblent, et les individus dans des groupes différents soient dissemblables.

Terminologie

- **Classification:** terme employé par les auteurs français
- **Clustering:** terme ango-saxon
- **Segmentation:** terme employé en marketting (les segments de clientèle)
- **Taxinomie** (biologie)

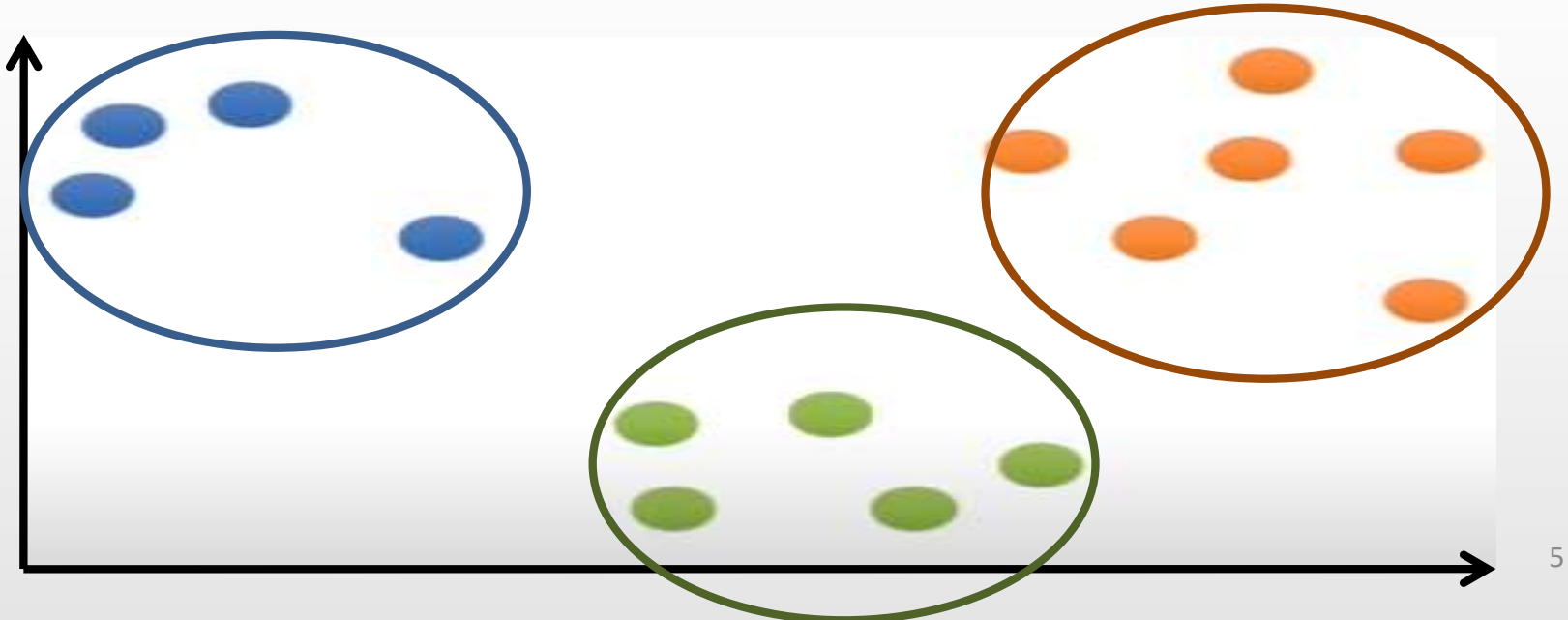
Introduction

Les méthodes de classification visent toutes à répartir **N** individus, caractérisés par **P** variables **X1, X2, ..., Xp** en un certain nombre **M** de sous-groupes aussi homogènes que possible, chaque groupe étant bien différencié des autres.

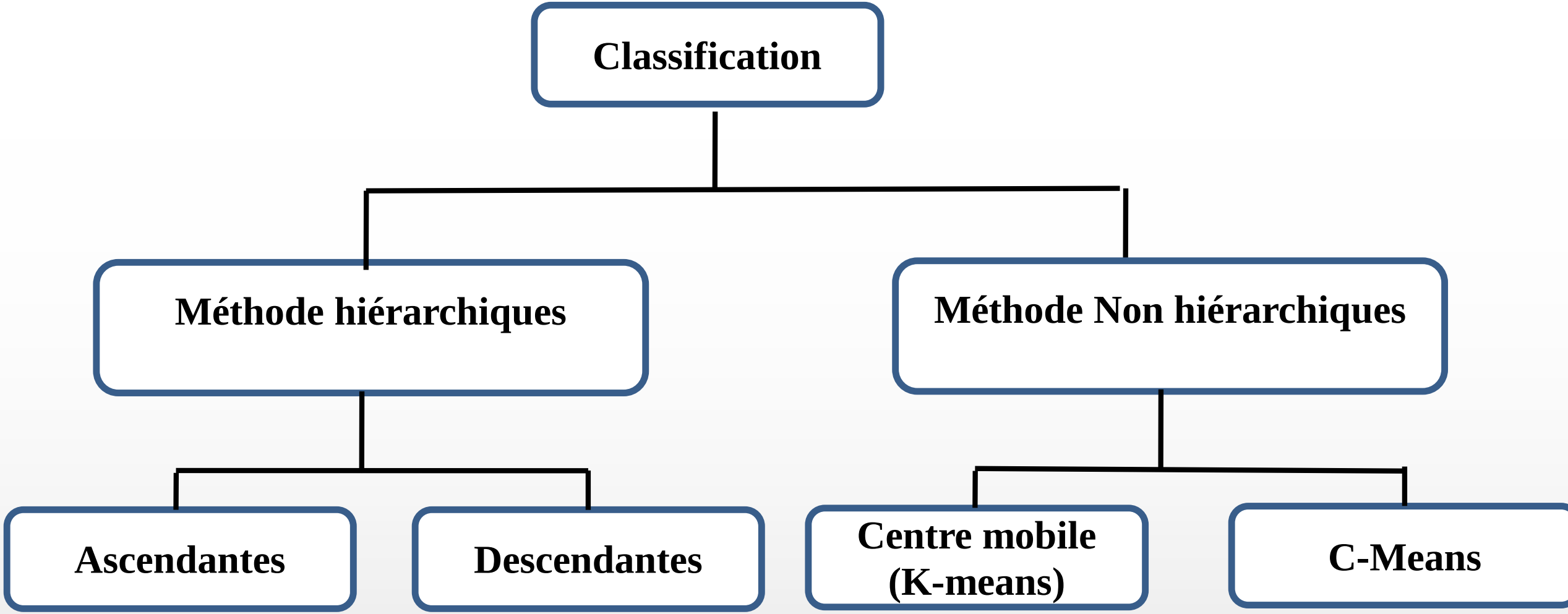
X1, X2, ..., Xp

X1, X2, ..., Xp

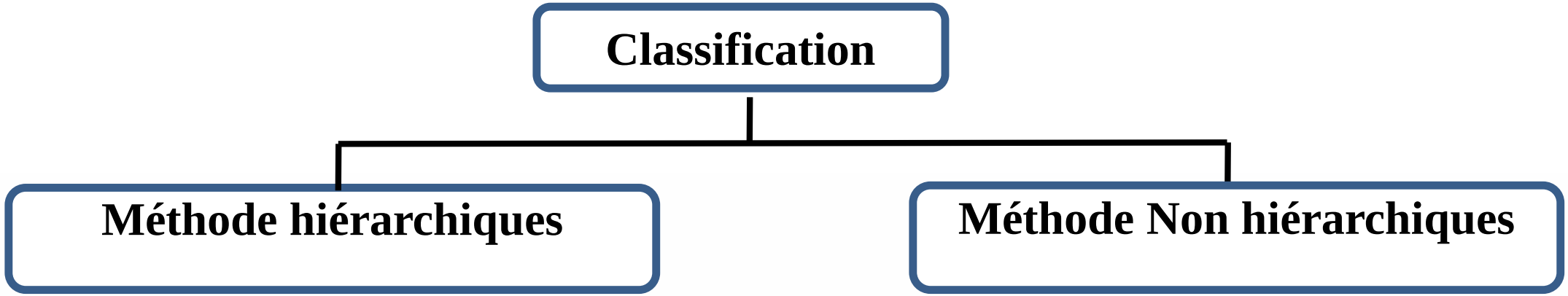
X1, X2, ..., Xp



Méthodes de classification



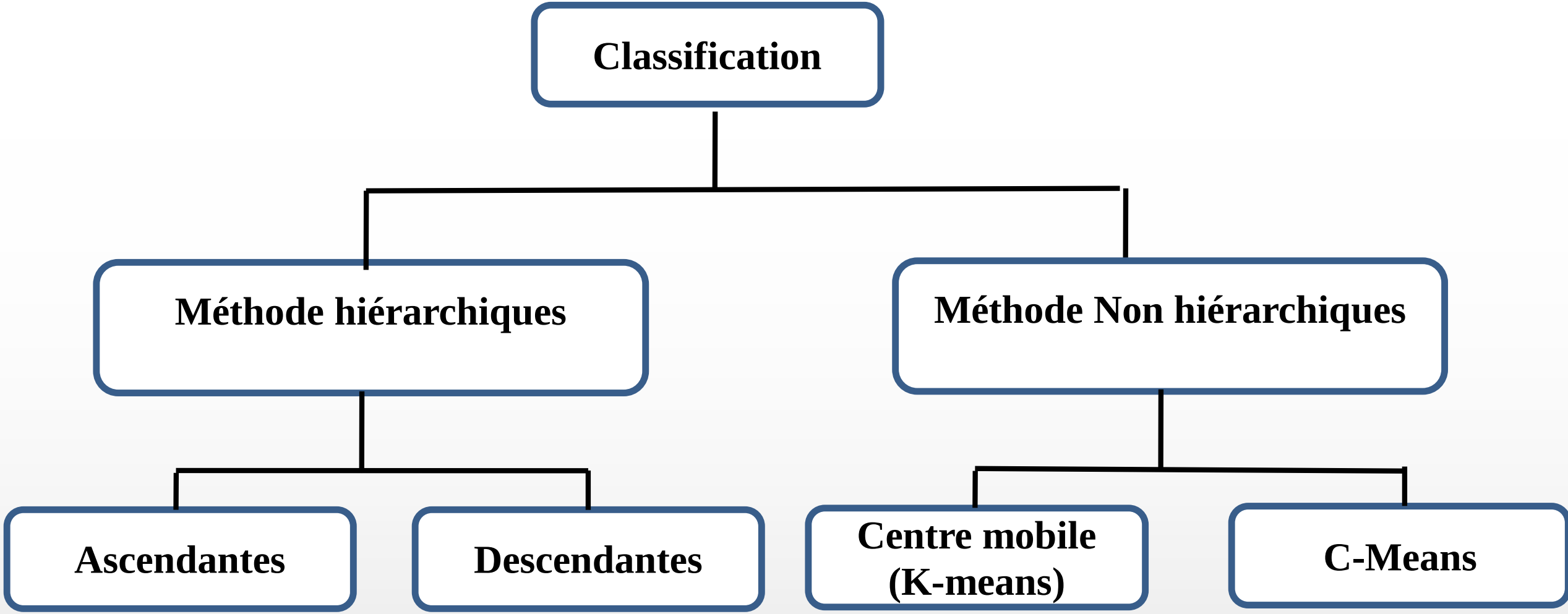
Méthodes de classification



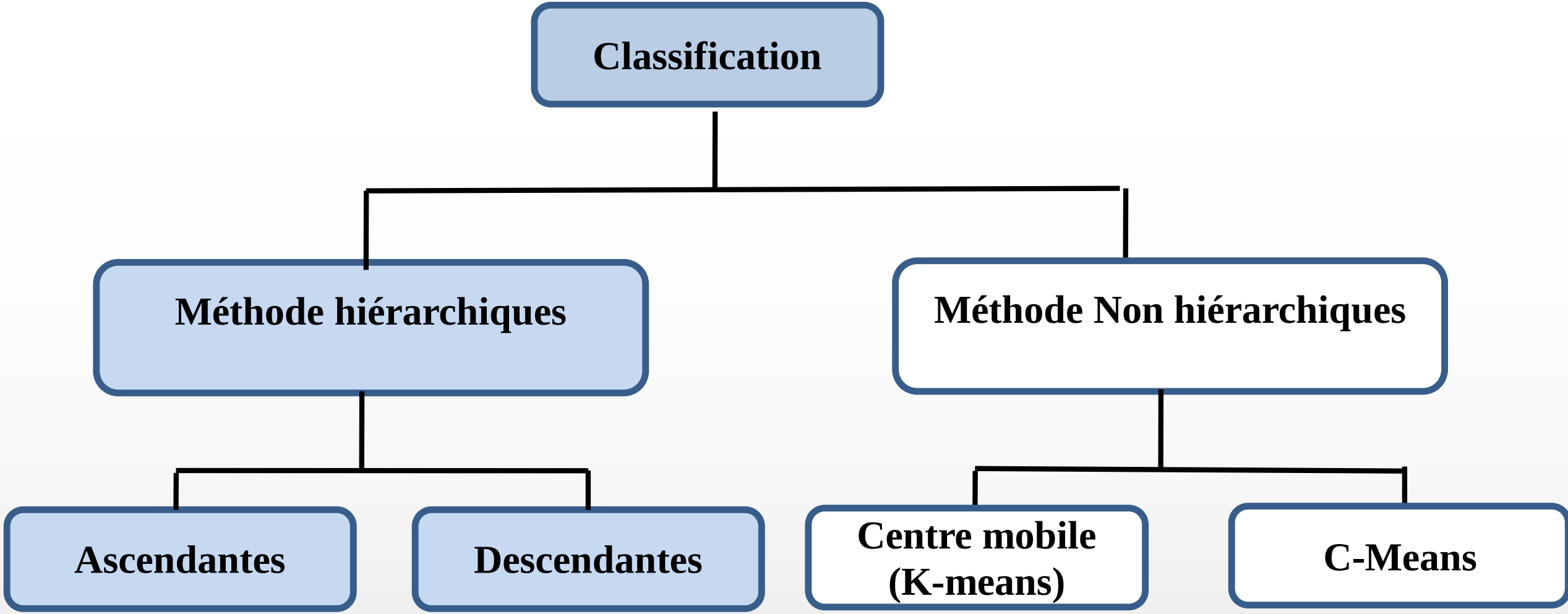
- Consiste à créer des clusters dans un ordre prédéfini
- On ne définit pas a priori le nombre de classes
- N'est pas adapter pour un grand nombre d'individus
- Le nombre d'individus doit être supérieur au nombre de variables

- Consiste a formé de nouveaux clusters en fusionnant ou en divisant les clusters au lieu de suivre un ordre hiérarchique
- Nécessite la fixation préalable du nombre de classes
- Plus intéressante si le nombre d'individu est assez important

Méthodes de classification



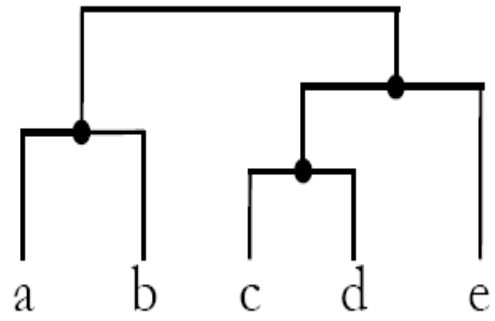
Méthodes de classification



Classification hiérarchiques

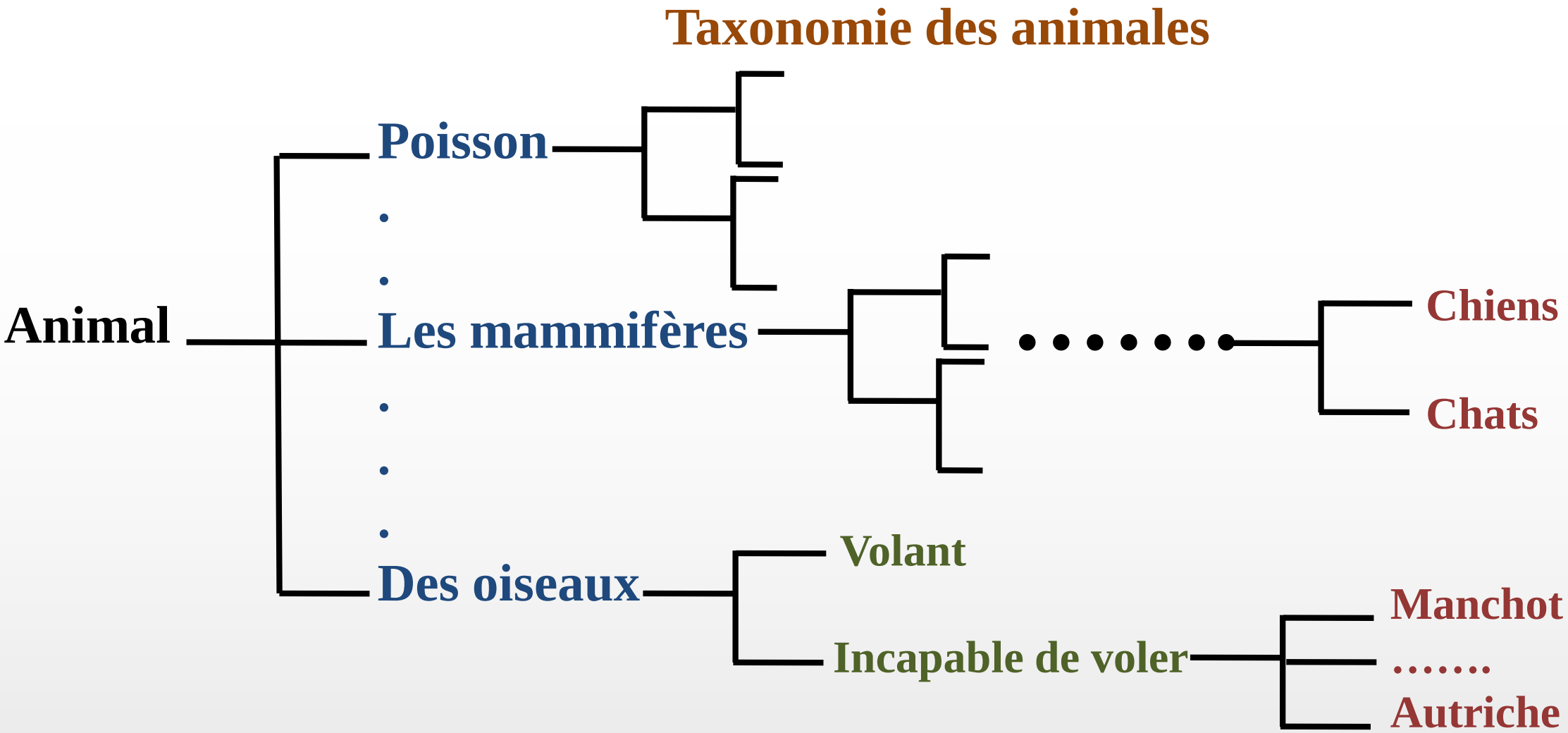
Classification hiérarchiques est un type de modèle de classification,

Permet de fournir un ensemble de partitions de moins en moins fines obtenus par regroupement successifs de parties



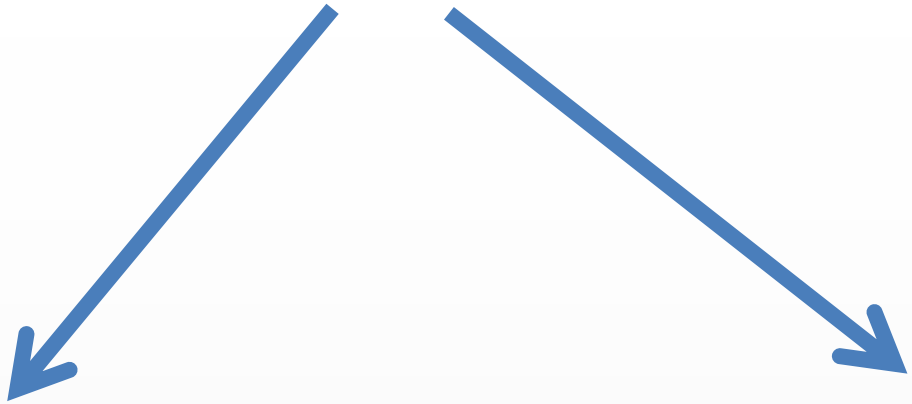
Arbre de classification
ou dendrogramme

Classification hiérarchiques



Classification hiérarchiques

Classification hiérarchiques



Descendantes-Divisible
(De haut en bas)

Ascendantes-Agglomératives
(De bas en haut)

Classification hiérarchiques

Descendantes-Divisible (De haut en bas)

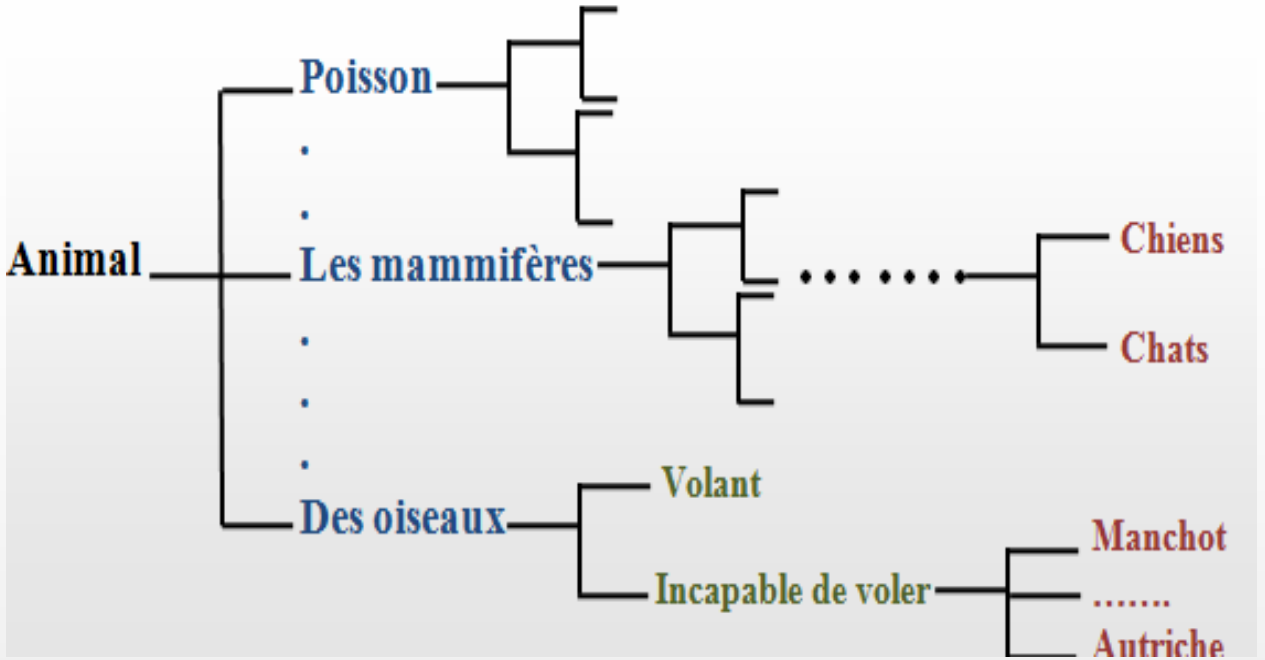
- Tout les objets constituent un unique cluster
- Séparer les objets (clusters) les plus dissimilaires (grande distance)
- Tous les objets sont des concepts feuilles

Ascendantes-Agglomératives (De bas en haut)

- Chaque objet constitue un cluster
- Regrouper les objets (clusters) les plus proches
- Jusqu'a arriver à au sommet

Classification hiérarchiques

Divisible
(De haut en bas)



Agglomérant
(De bas en haut)

Classification hiérarchiques

Descendantes-Divisible

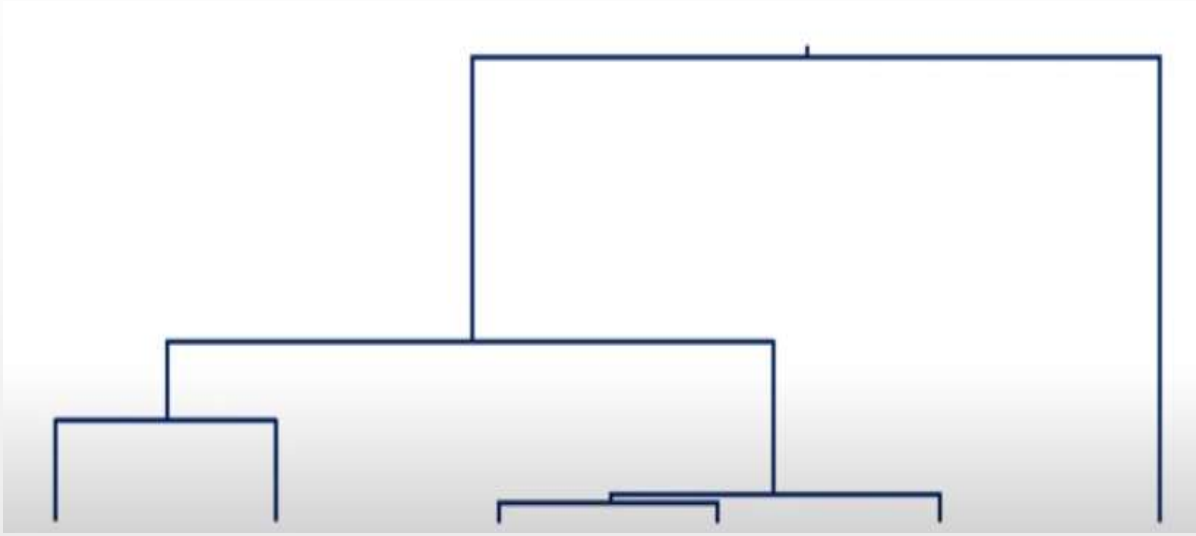
- Les données commencent comme un cluster combiné.
- Le cluster se divise en deux parties distinctes, selon un **certain degré de similitude.**
- Les clusters se divisent en deux encore et encore jusqu'à ce que les clusters ne contiennent qu'un seul point de données. Le clustering divisif par exemple : **Arbres de décision, Foret aléatoire** .

Classification hiérarchiques

Ascendantes

Dendrogramme

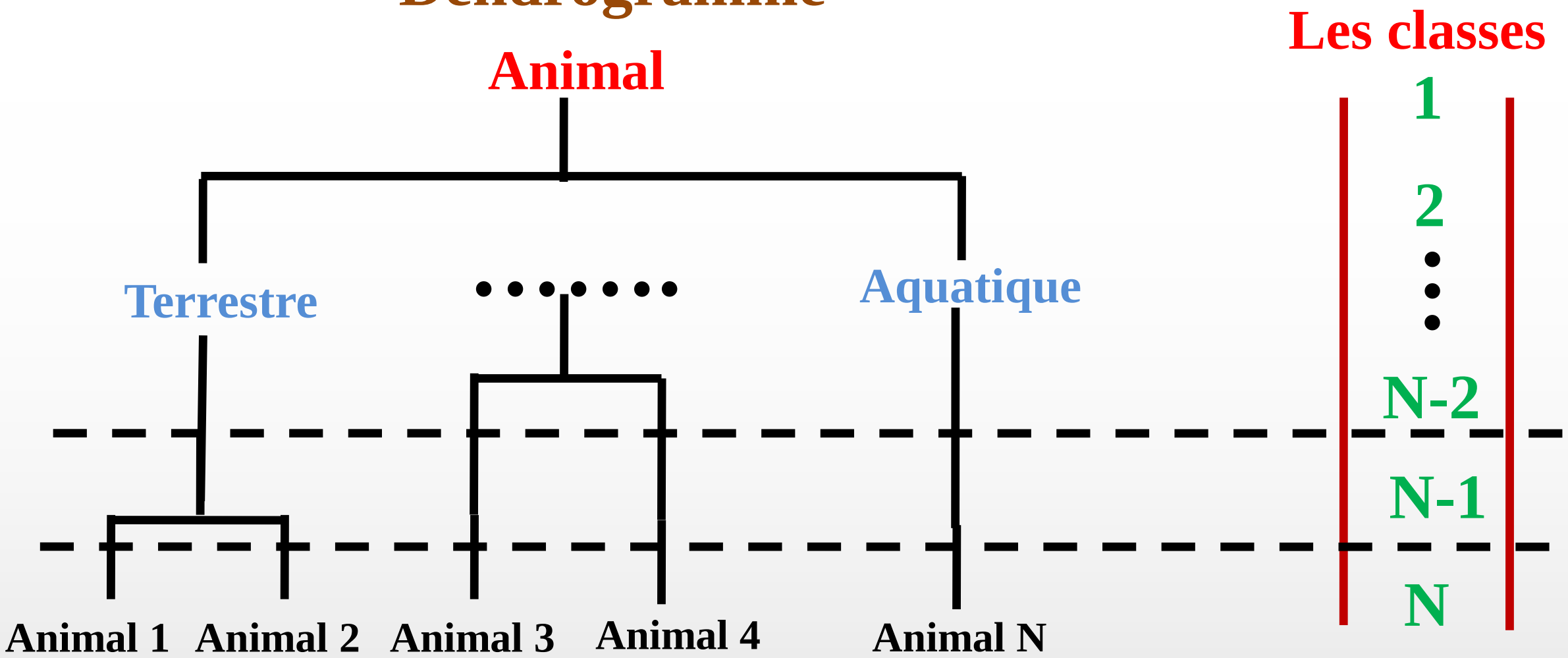
Est un diagramme fréquemment utilisé pour illustrer l'arrangement de groupe générés par un regroupement hiérarchique



Classification hiérarchiques

Ascendantes

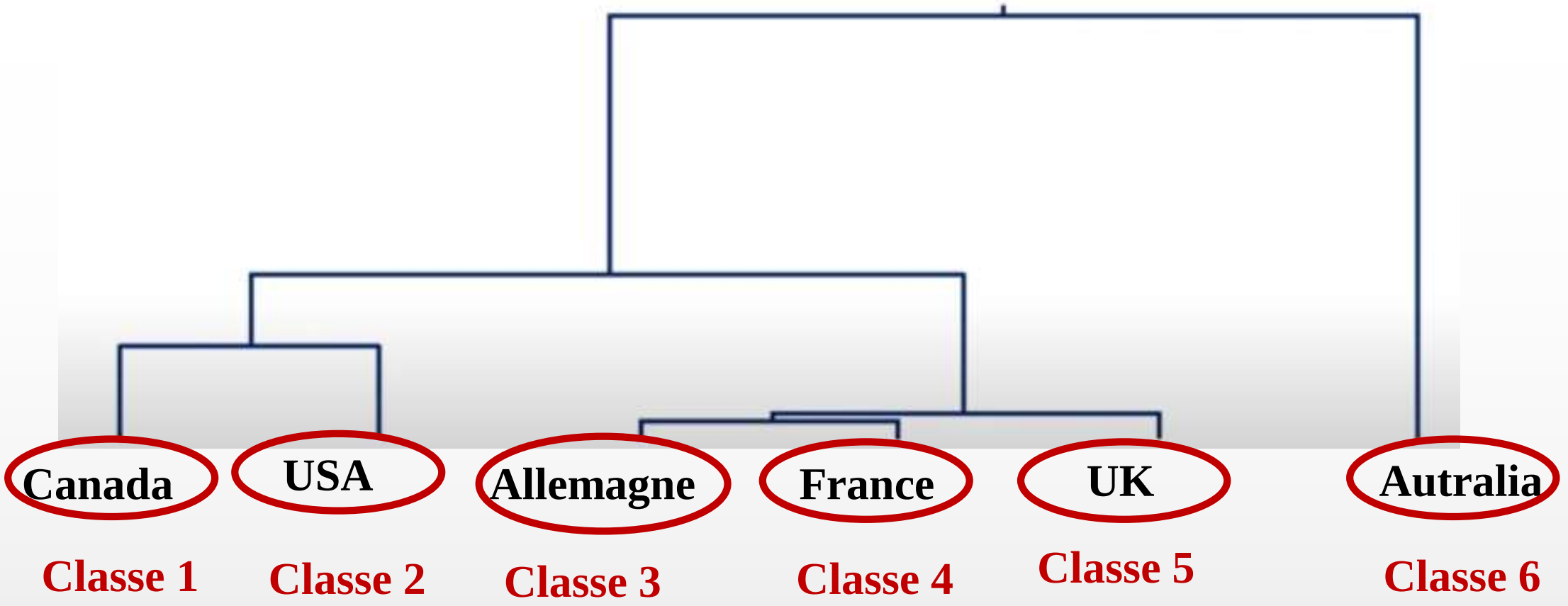
Dendrogramme



Classification hiérarchiques

Ascendantes

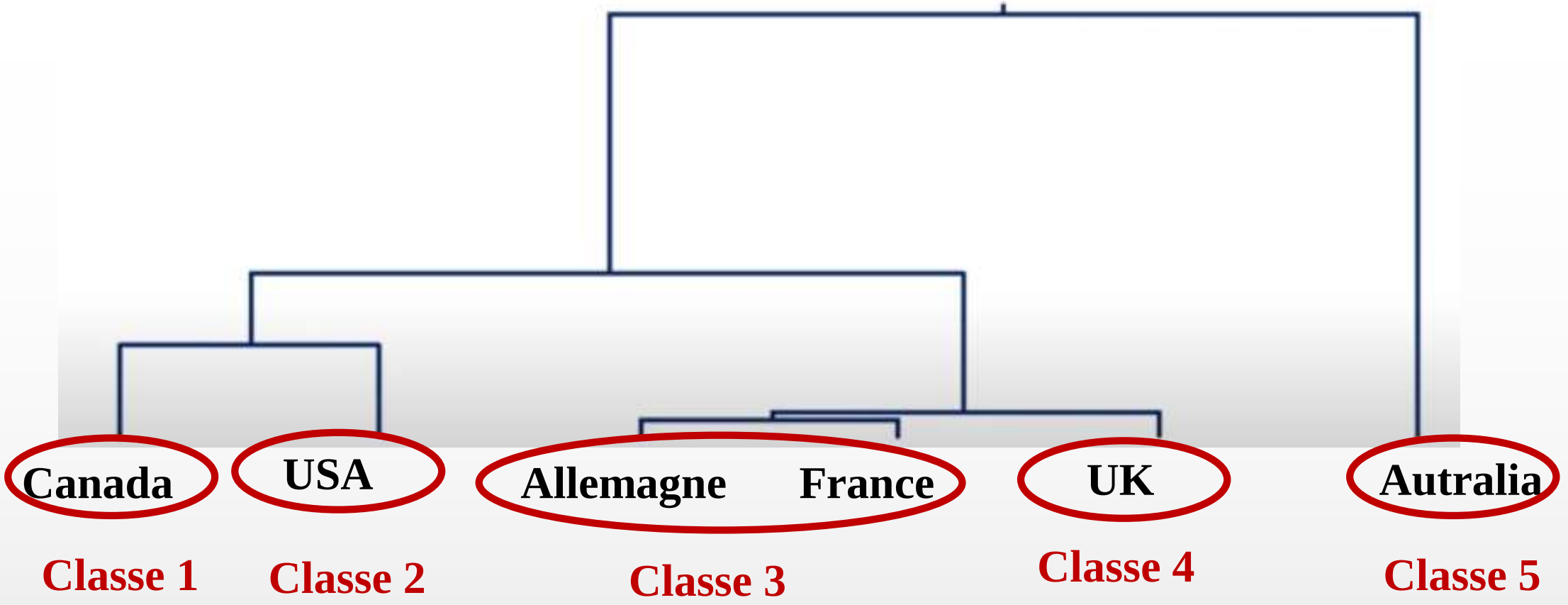
Dendrogramme



Classification hiérarchiques

Ascendantes

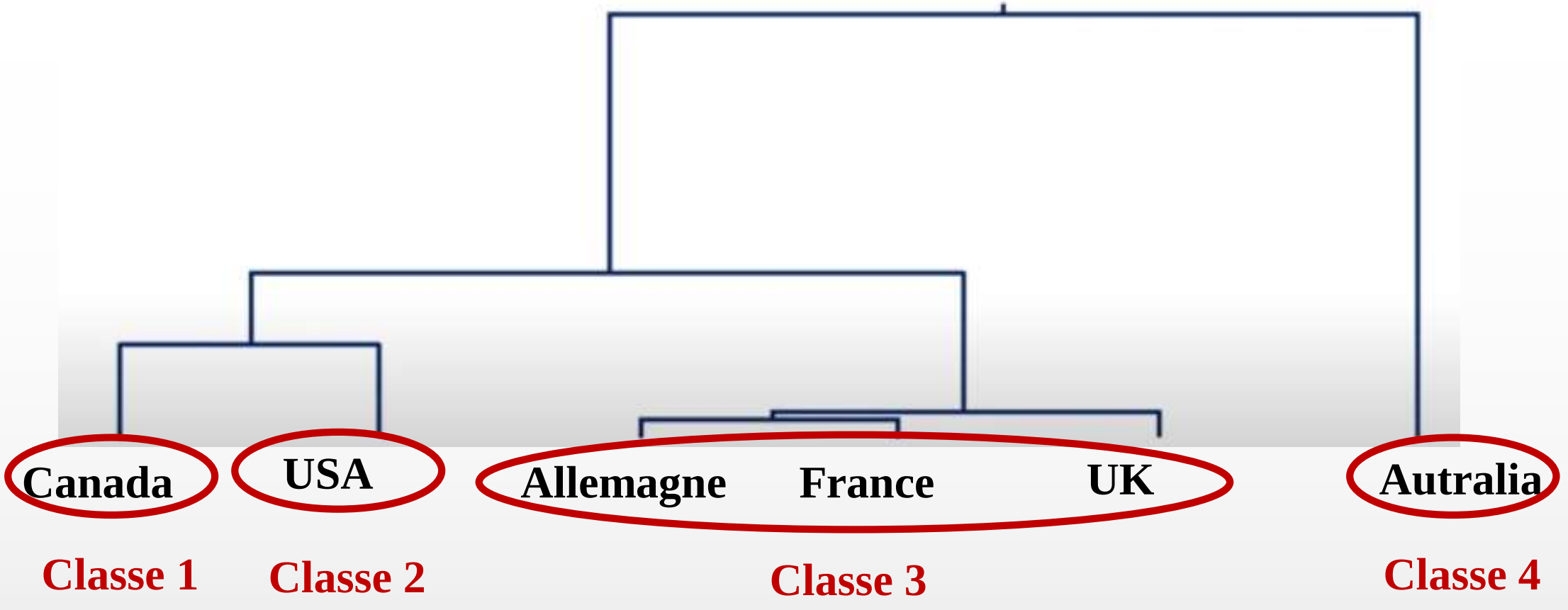
Dendrogramme



Classification hiérarchiques

Ascendantes

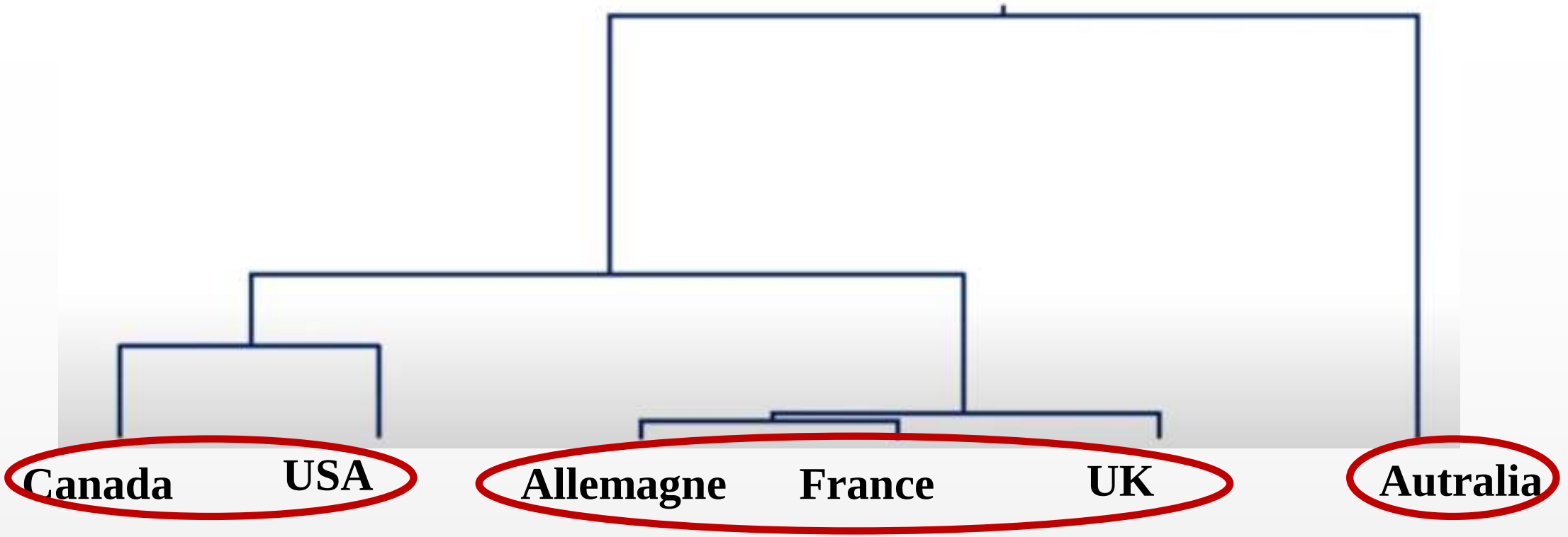
Dendrogramme



Classification hiérarchiques

Ascendantes

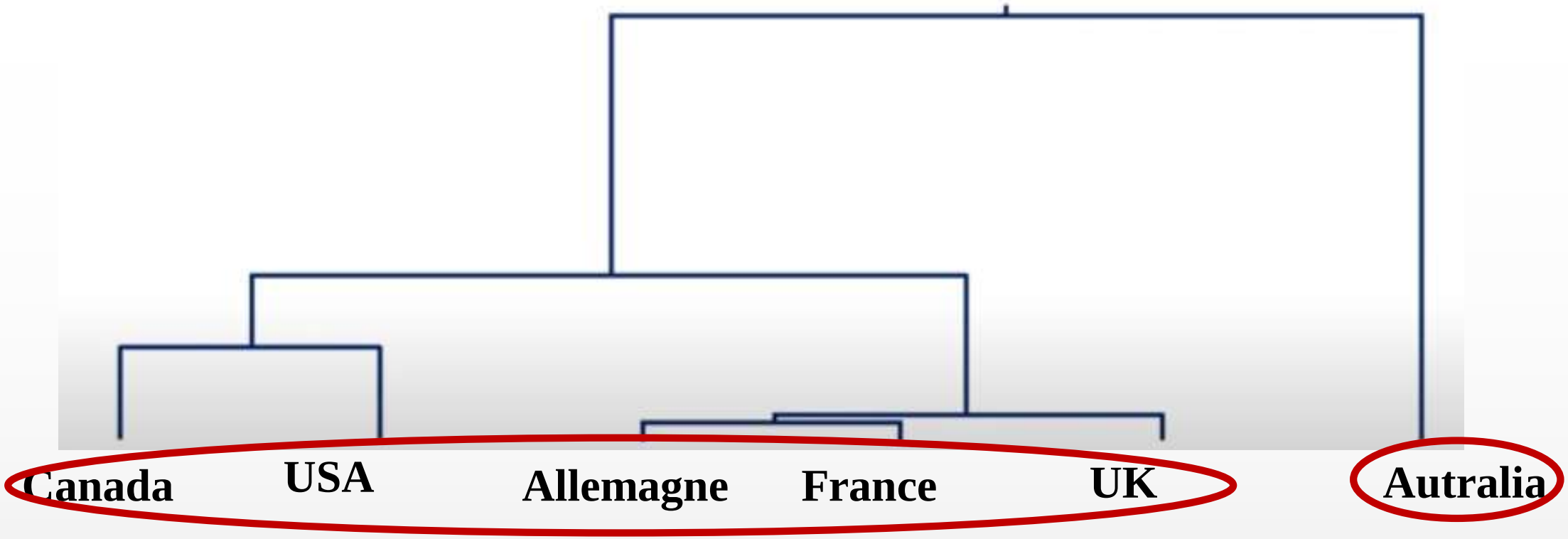
Dendrogramme



Classification hiérarchiques

Ascendantes

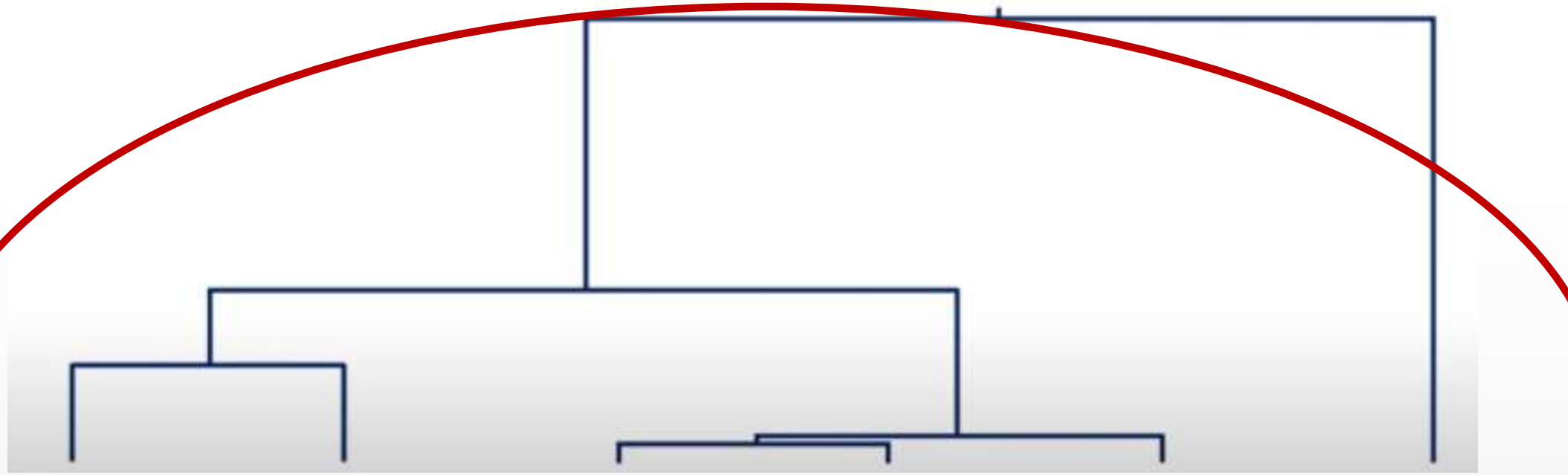
Dendrogramme



Classification hiérarchiques

Ascendantes

Dendrogramme

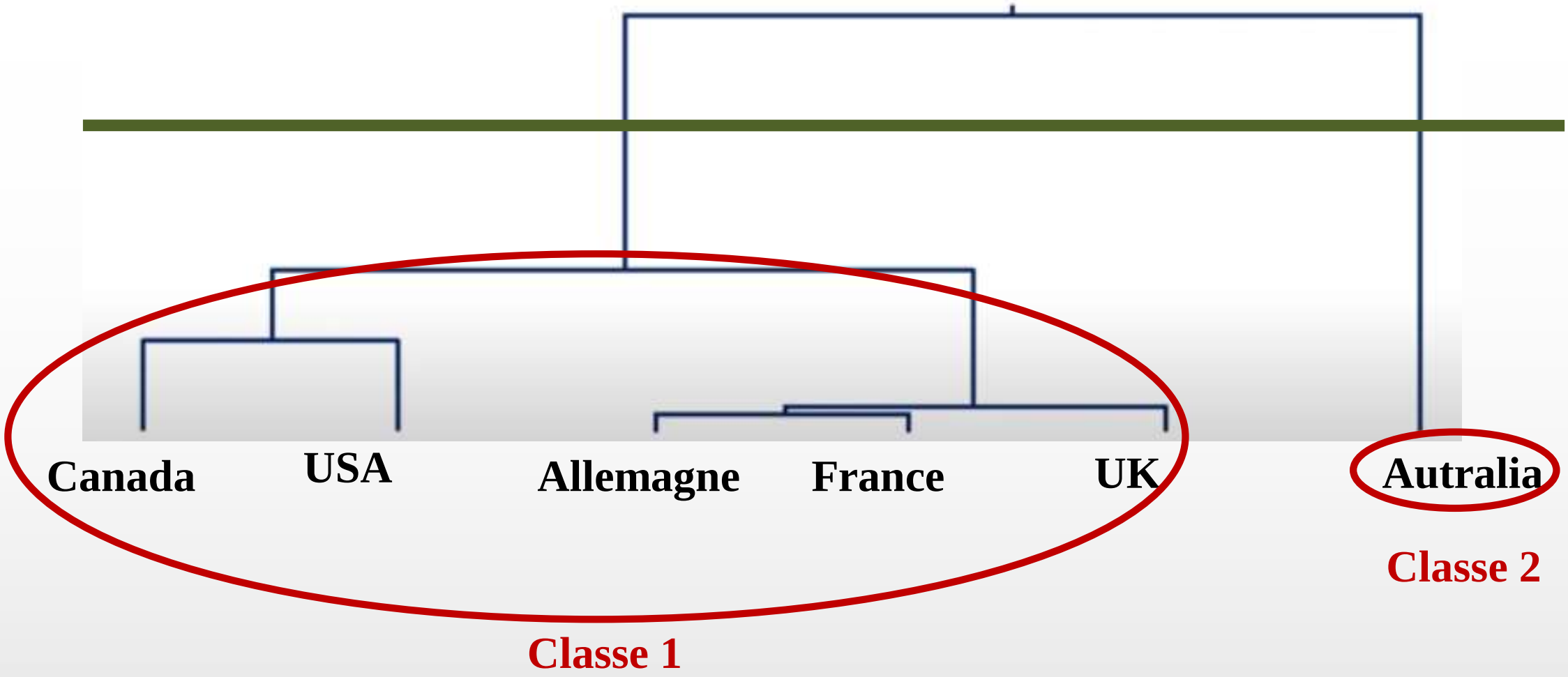


Classe 1

Classification hiérarchiques

Ascendantes

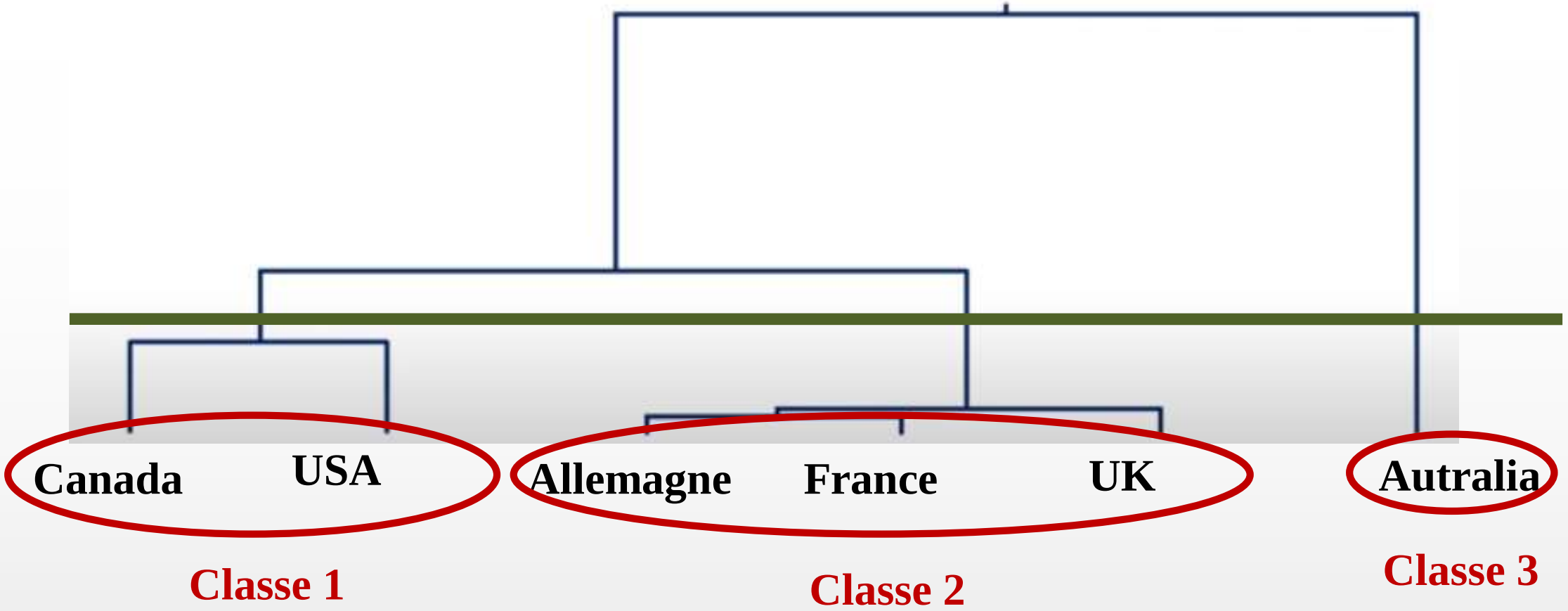
Dendrogramme



Classification hiérarchiques

Ascendantes

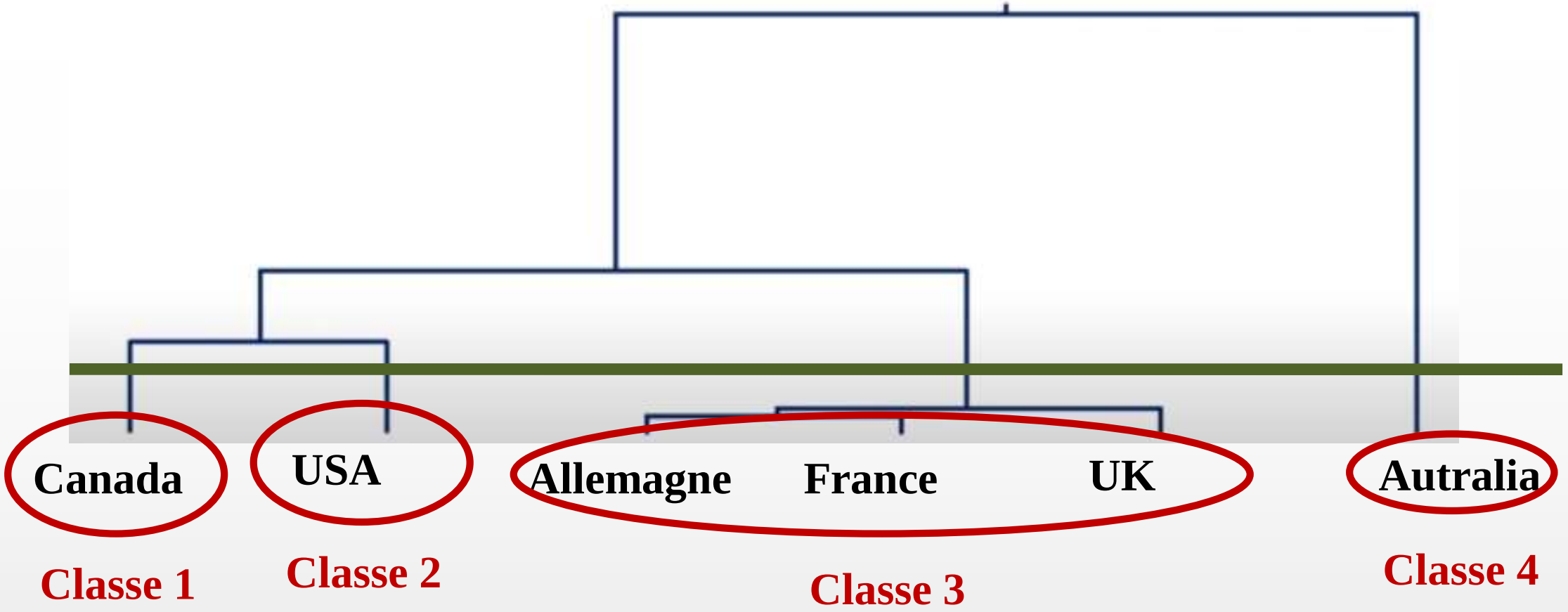
Dendrogramme



Classification hiérarchiques

Ascendantes

Dendrogramme



Classification hiérarchiques

Ascendantes

Algorithme

•Initialisation : les classes initiales sont les n individus. Calculer le tableau de leurs distances deux à deux

•Itérer les deux étapes suivantes jusqu'à l'agrégation en une seule classe :

- Regrouper les deux éléments (classes) les plus proches au sens de la distance entre groupes choisie
- Mettre à jour le tableau de distances en remplaçant les deux classes regroupées par la nouvelle et en calculant sa distance avec chacune des autres classes

Classification hiérarchiques

Ascendantes

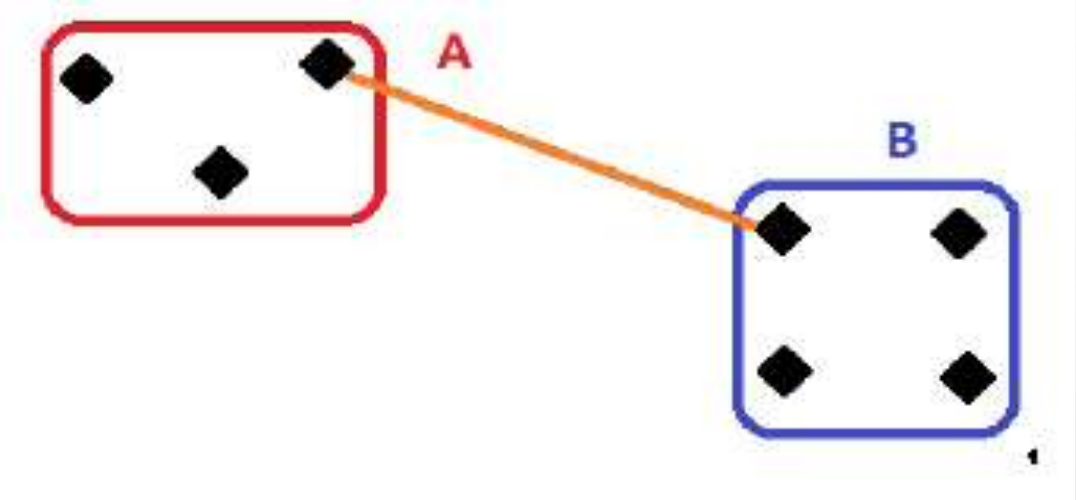
✓ **Nécessité de définir une distance entre groupes d'individus (appelé stratégie d'agrégation)**

✓ **Nécessite de choisir le nombre de classes à retenir**

Stratégies d'agrégation :

Stratégie du saut minimum ou single linkage (la distance entre les parties est représenté par la plus petite distance entre éléments des deux parties)

$$D(A,B) = \min_{i \in A, j \in B} d(i,j)$$

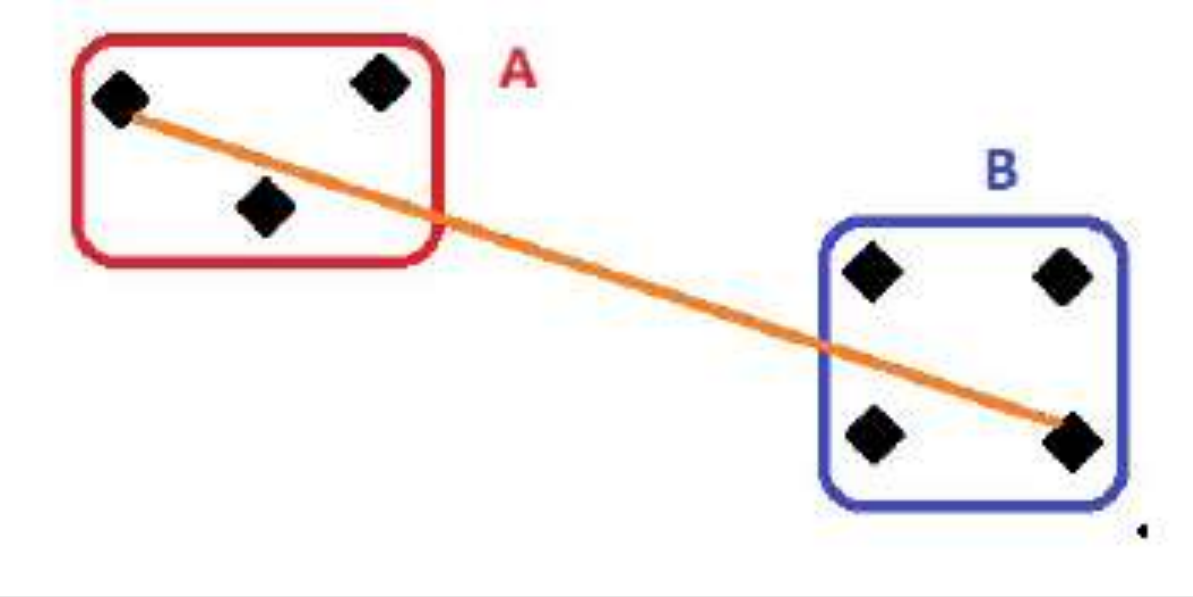


Ascendantes

Stratégies d'agrégation :

Stratégie du saut maximum ou du diamètre ou complète linkage (la distance entre parties est la plus grande distance entre éléments des deux parties)

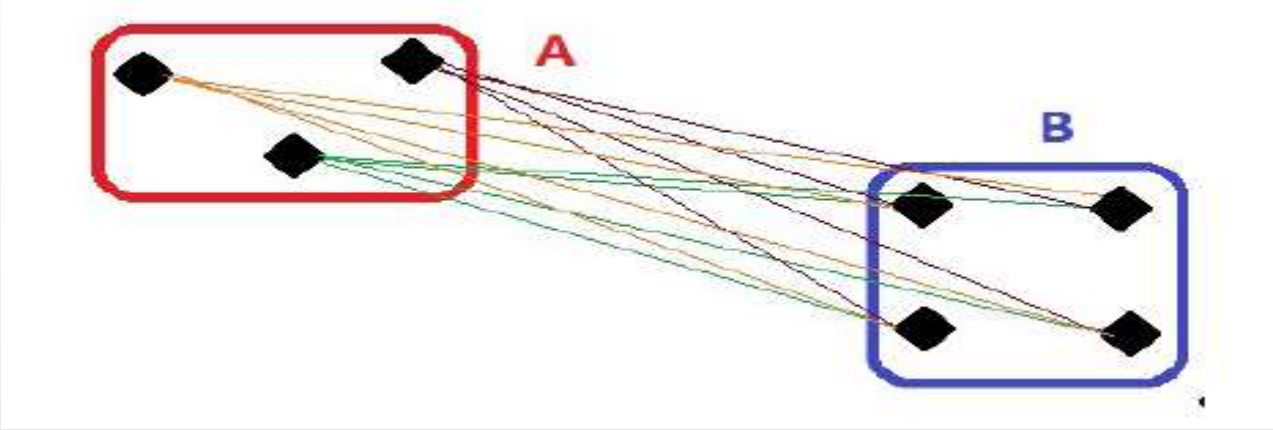
$$D(A,B) = \max_{i \in A,\, j \in B} d(i,j)$$



Stratégies d'agrégation :

Stratégie d'écart moyen: c'est l'écart entre deux groupes A et B est caractérisé par la distance moyenne entre les points de A et B où n_A est le nombre d'individus dans A, et n_B le nombre d'individus dans B.

$$D(A, B) = \frac{1}{n_A n_B} \sum_{i \in A} \sum_{j \in B} d(i, j)$$



Stratégies d'agrégation :

Méthode du saut Ward (en espace euclidien)

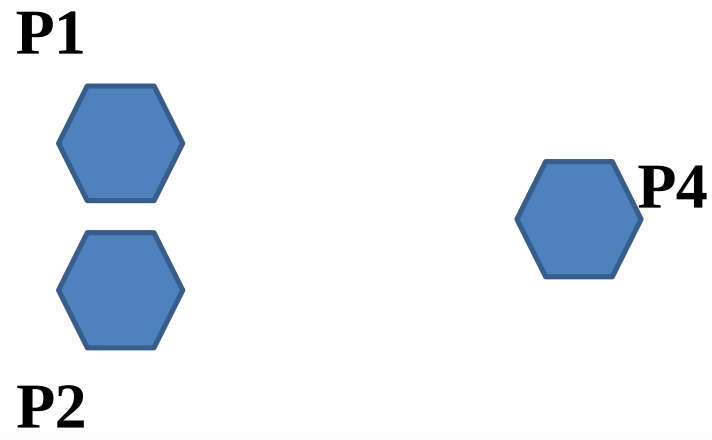
$$D(A,B) = \frac{n_A n_B}{n_A + n_B} d^2(G_A, G_B)$$

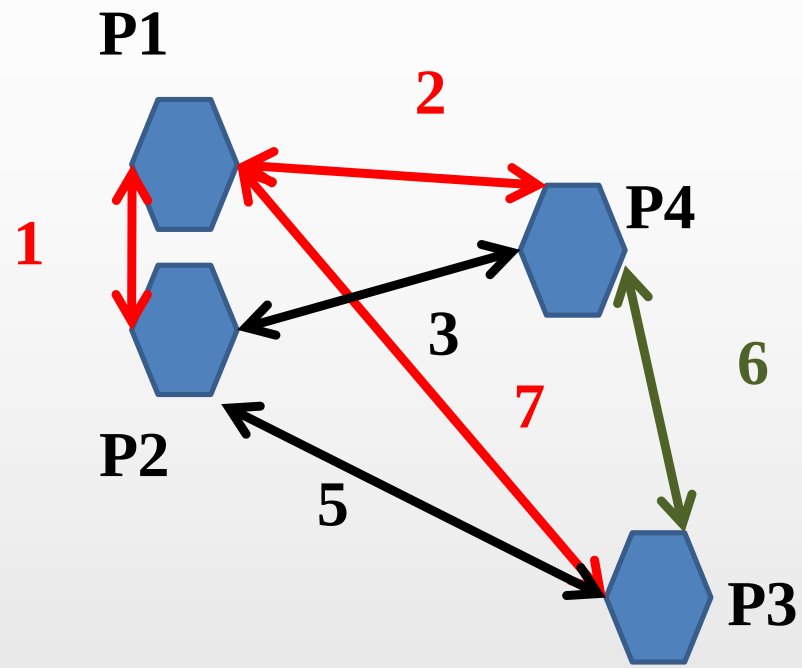
où G_A est le centre de gravité de A, et G_B celui de B Cette méthode prend en compte à la fois la dispersion à l'intérieur d'un groupe et la dispersion entre les groupes. Elle est utilisée par défaut dans la plupart des programmes informatiques.

Classification hiérarchiques

Ascendantes

Exemple





	P1	P2	P3	P4
P1	0			
P2	1	0		
P3	7	5	0	
P4	2	3	6	0

Classification hiérarchiques

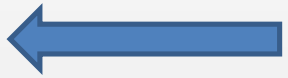
Ascendantes

Exemple avec du saut minimum

	P1	P2	P3	P4
P1	0			
P2	1	0		
P3	7	5	0	
P4	2	3	6	0

	{P1,P2}	P3	P4
{P1, P2}	0		
P3	5	0	
P4	2	6	0

	{P1, P2,P4}	P3
{P1, P2,P4}	0	
P3	5	0



	{P1, P2,P4,P3}
{P1, P2,P4,P3}	0

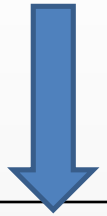
Classification hiérarchiques

Ascendantes

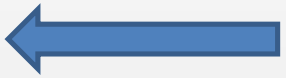
Exemple avec du saut maximum

	V1	V2	V3	V4
V1	0			
V2	8	0		
V3	2	3	0	
V4	7	10	6	0

	{V2,V4}	V3	V4
{V2, V4}	0		
V1	8	0	
V3	6	0	0



	{V1, V2,V4}	V3
{V1, V2,V4}	0	
V3	6	0



	{V1, V2,V4,V3}
{V1, V2,V4,V3}	0

Classification Non hiérarchiques

Vise à regrouper une population en k classes. Cela se fait de manière automatique; il n'y a pas de lien hiérarchique dans les regroupements contrairement à l'algorithme CAH.

Classification Non hiérarchiques est le mieux adapté aux très grands tableaux de données.

Classification Non hiérarchiques

Méthode des centres mobiles

- Partition en k classes
- **Avantages** : Permettent la classification d'ensembles volumineux.
- **Inconvénients** : On impose au départ le nombre de classes.

Algorithme

- **1ère étape** : choix de centres c_i (les c_i sont choisis au hasard).
  - La classe E_{ci} est formée de tous les points plus proches de c_i que de tout les autres centres.

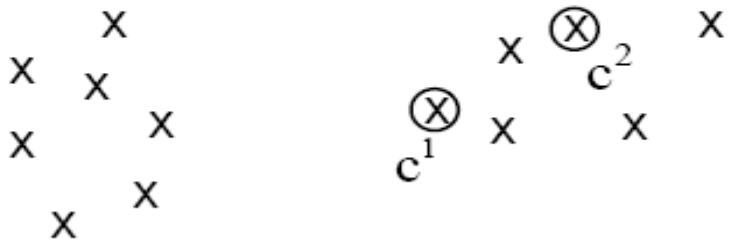
- **2ème étape** : calcul les centres de gravité de chaque classe

  → définition d'une nouvelle partition.
- Itération de la 2ème étape jusqu'à la stabilité des classes.

Classification Non hiérarchiques

Méthode des centres mobiles

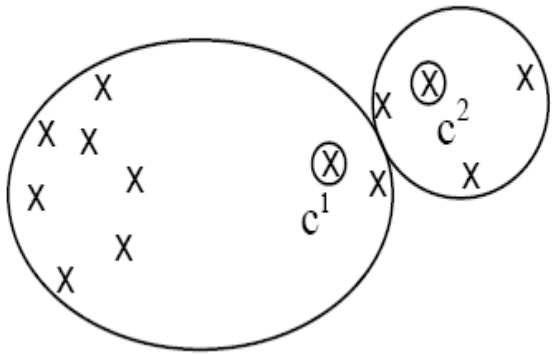
La classification des individus dépend du choix des centres initiaux



Etape 0

Choix des centres

c_1 c_2



Etape 1

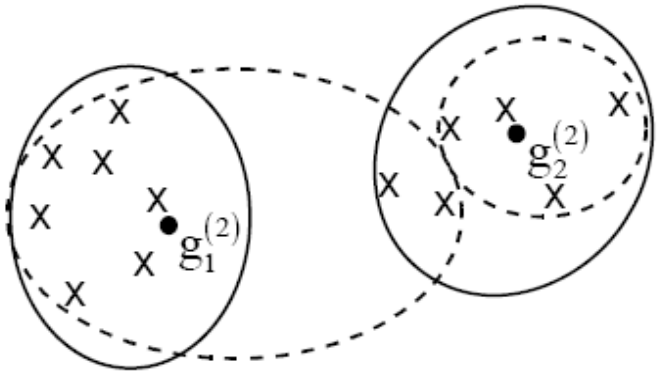
Constitution de classes autour des centres c_1 et c_2

Classe 1 : points plus proches de c_1 que de c_2

Classe 2 : points plus proches de c_2 que de c_1

Classification Non hiérarchiques

Méthode des centres mobiles

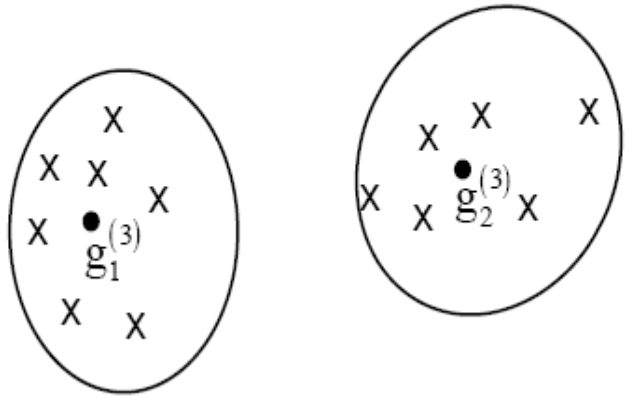


Etape 2

- Calcul des centres de gravité des 2 classes formées à l'étape 1 : g_1 g_2

\+

- Définition de nouvelles classes autour des centres de gravité



Etape 3

Calcul des centres de gravité des classes formées à l'étape 2.
Nouvelle définition des classes autour de ces centres → STABILITE

Conclusion

L'analyse par classification s'agit d'une classe de techniques permettant de regrouper les cas relativement homogènes en eux-mêmes et hétérogènes entre les autres sur la base d'un ensemble défini de variables.

Cas d'application

Segmentation du marché, regroupement des consommateurs en fonction de leurs préférences afin de faire la commercialisation des produits.