Clustering :
Classification ascendante hiérarchique

Effectué par :
Ait Boulahcen Aouatif
Khattari Fouad
Zizouan Aziza

Sous la supervision de :

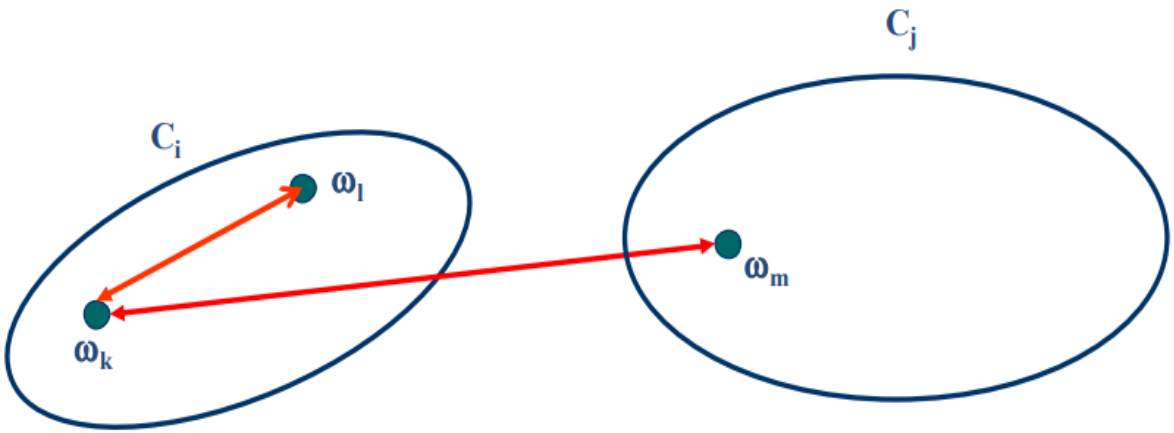
Pr. Ourdou Amal

2022-2023

Plan

Clustering

- Méthode d'apprentissage automatique qui consiste à regrouper des points de données par similarité ou par distance.
- Méthode d'apprentissage non supervisée et une technique populaire d'analyse statistique des données.
- Le clustering a pour objectif de séparer un ensemble d'observations en groupes homogènes.



Quelques applications du clustering

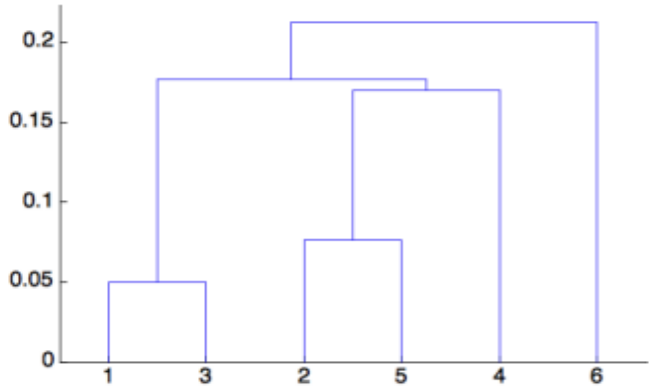
- Reconnaissance de formes
- Traitement d'images
- Clustering des usages du web pour découvrir des groupes d'accès similaires
- Recherche d'information : Catégorisation de documents ou de termes

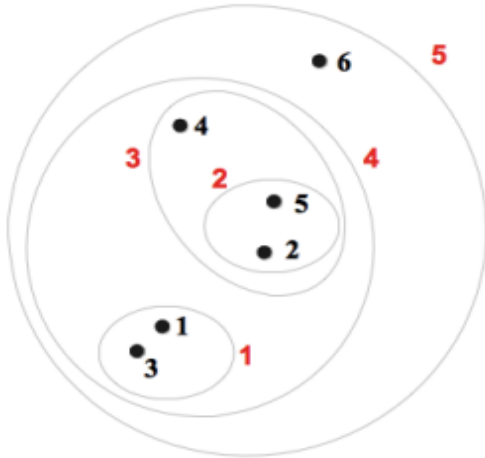
Clustering hiérarchique

Principe

- Chaque individu représente un groupe.
- Trouver les deux groupes les plus proches.
- Grouper les deux groupes en un nouveau groupe.
- Itérer jusqu'à N groupes

Visualisation sous la forme d'un dendogramme





Clustering hiérarchique

- **Classification hiérarchique ascendante**
  - Commencer avec les points en tant que clusters individuels.
  - A chaque étape, grouper les clusters les plus proches jusqu'à obtenir 1 seul ou k clusters.
- **Classification hiérarchique descendante**
  - Commencer avec 1 seul cluster comprenant tous les points.
  - A chaque étape, diviser un cluster jusqu''a obtenir des clusters ne contenant qu'un point ou jusqu'à obtenir k clusters.

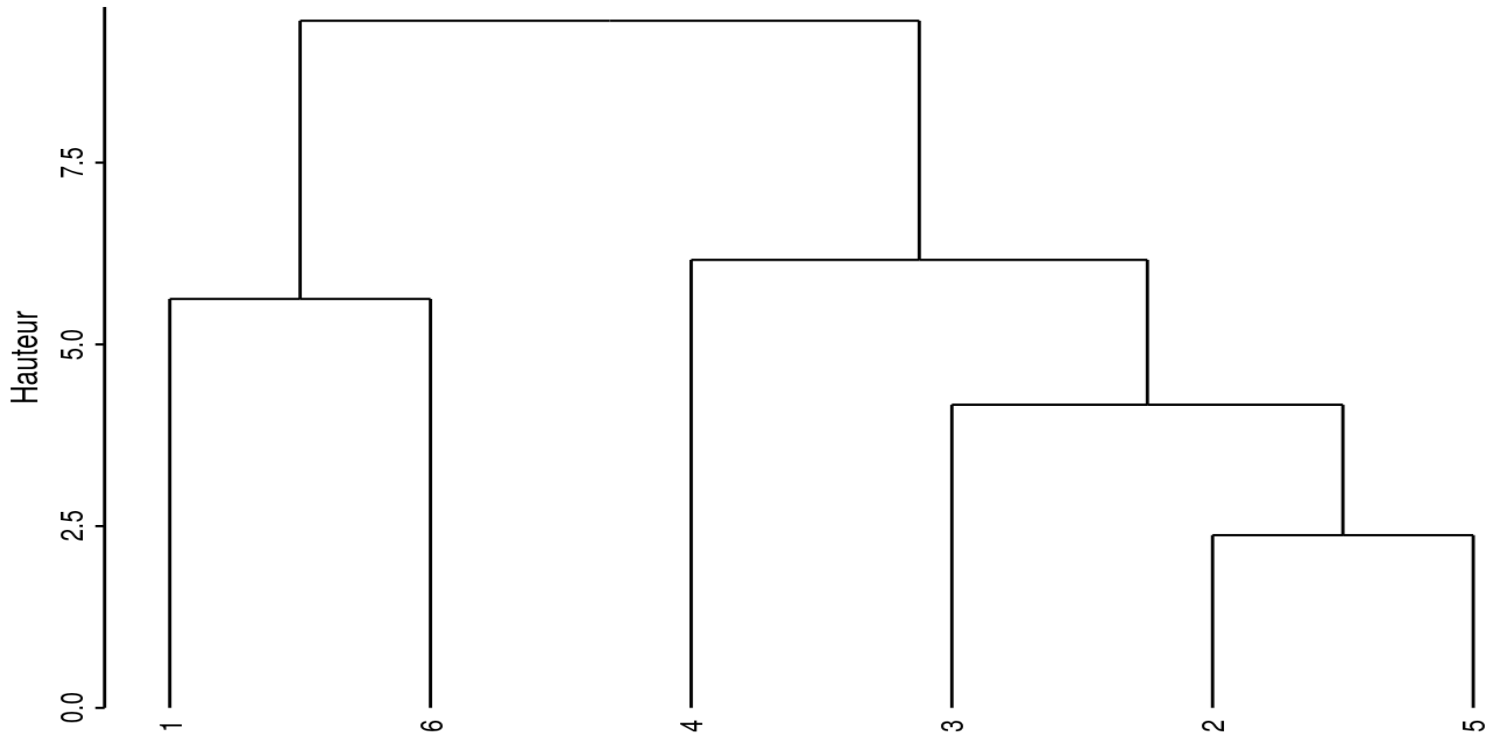
Défintion

- Elle consiste à regrouper progressivement les individus dans un groupe
- Il faut d'abord mettre les individus les plus proches ensemble ensuite rejeter les individus les plus éloignés. Cette méthode convient plus au cas des variables explicatives de type quantitatives
- L'etat de rapprochement ou d'éloignment entre les individus est mesuré souvent par la distance euclidienne

Algorithme

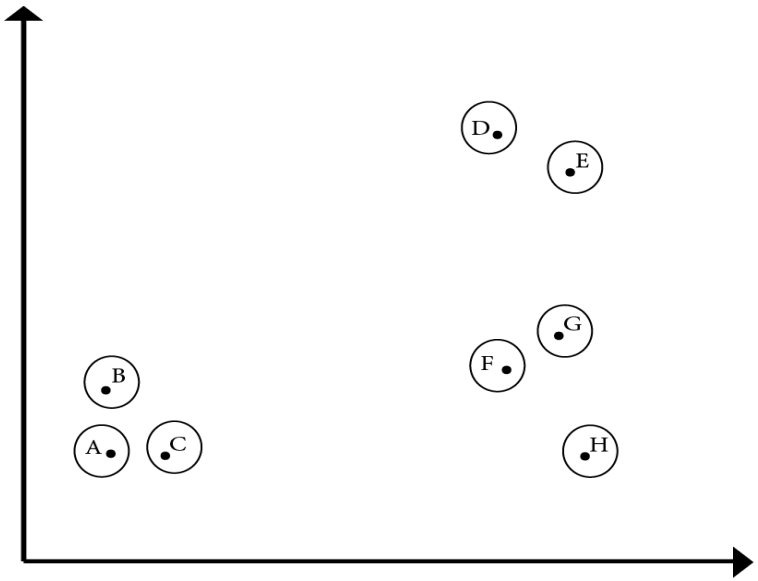
- Initialisation
  - Chaque individu est placé dans son cluster global
  - Calcul de la matrice de ressemblance M entre chaque couple de clusters
    NB: Où les coefficients de la matrice M sont les distances entre les clusters
- Répétition
  - Sélection dans M les deux clusters les plus proches C_i et C_j .
  - Fusion de C_i et C_j pour former un cluster C_g.
  - Mise à jour de M en calculant la ressemblance entre C_g et les clusters existants.
- Jusqu'a fusion des 2 derniers clusters.

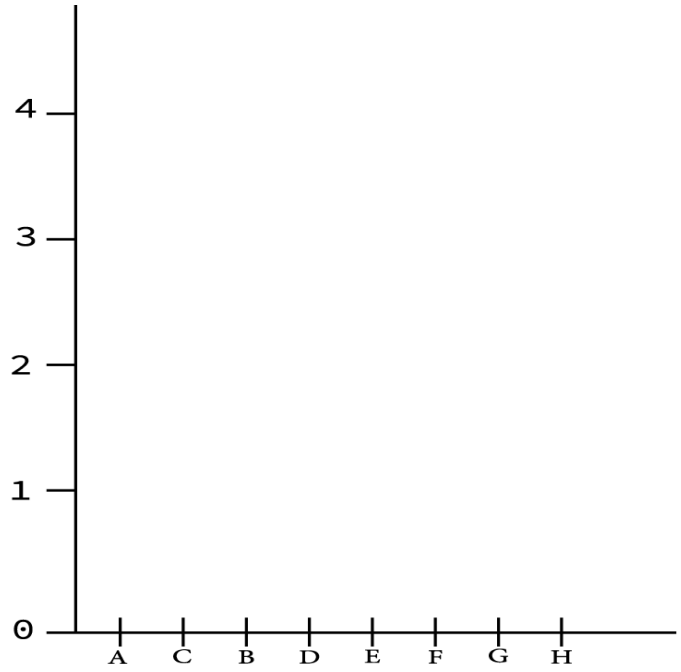
Les liens hiérarchiques apparaissent sur un dendrogramme tel que celui présenté ci-dessous :



Un **Dendrogamme** est un diagramme de regroupement hiérarchique, permettant d'organiser des données en arborescence en fonction de leurs similitudes

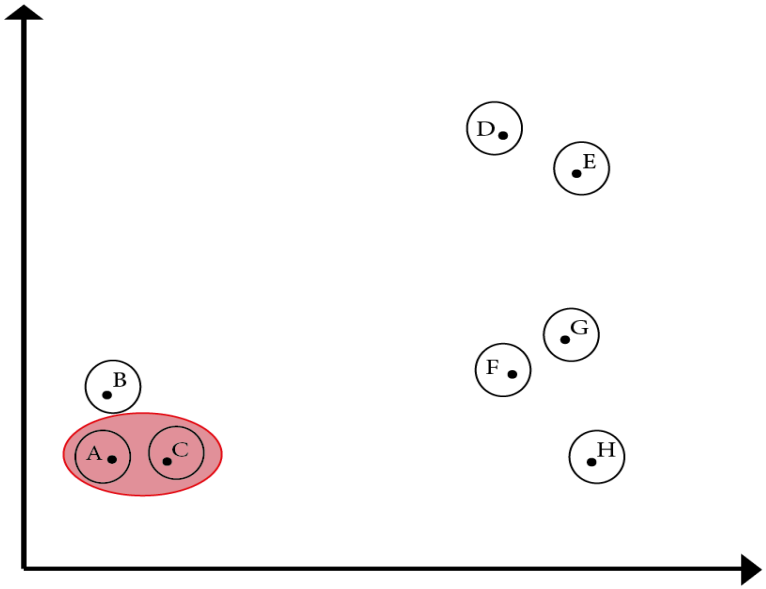
Etape 0

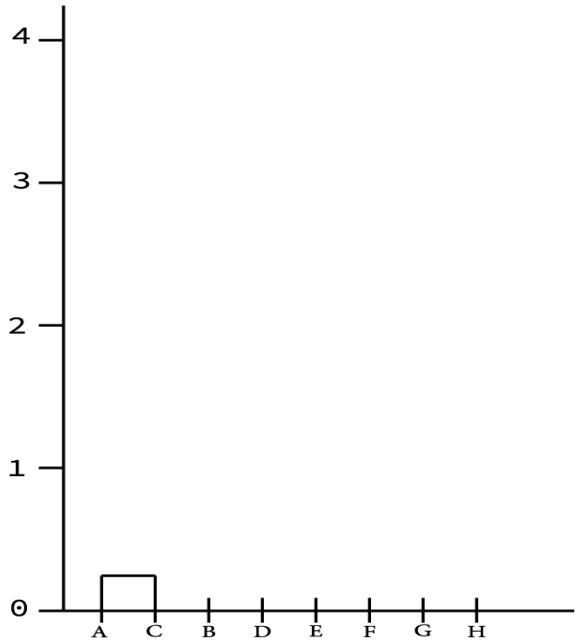




	A	B	C	D	E	F	G	H
A	0							
B	0.5	0						
C	0.25	0.56	0					
D	5	4.72	4.80	0				
E	4.78	5.55	5.57	1	0			
F	4.32	4.23	4.07	2.01	2.06	0		
G	4.92	4.84	4.68	2.06	1.81	0.61	0	
H	5	5.02	4.75	3.16	2.90	1.28	1.12	0

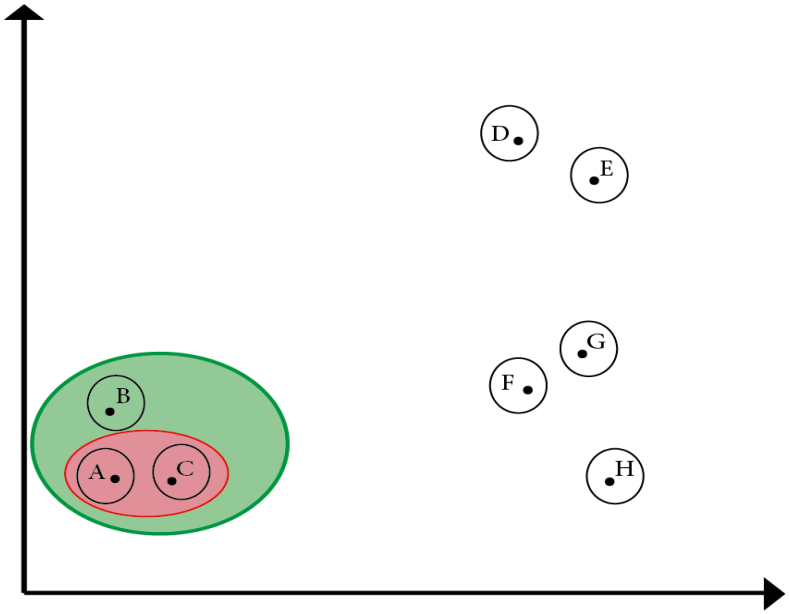
Etape 1

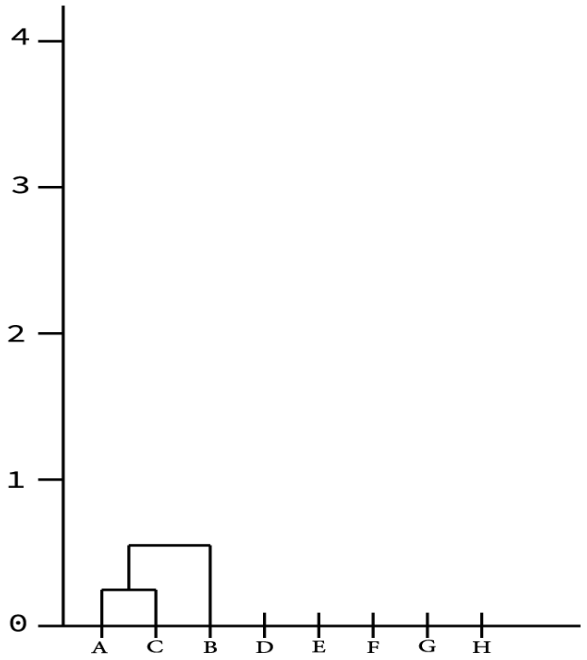




	AC	B	D	E	F	G	H
AC	0						
B	0.5	0					
D	4.8	4.72	0				
E	5.57	5.55	1	0			
F	4.07	4.23	2.01	2.06	0		
G	4.68	4.84	2.06	1.81	0.61	0	
H	4.75	5.02	3.16	2.90	1.28	1.12	0

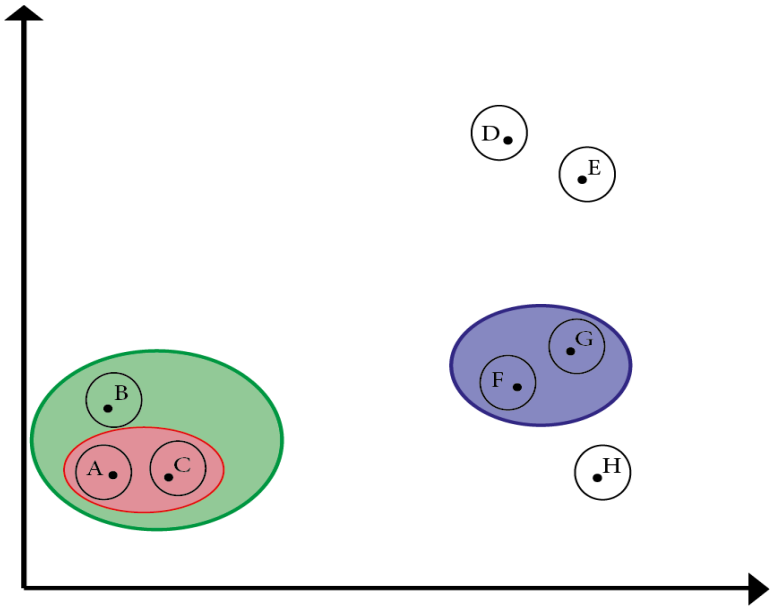
Etape 2

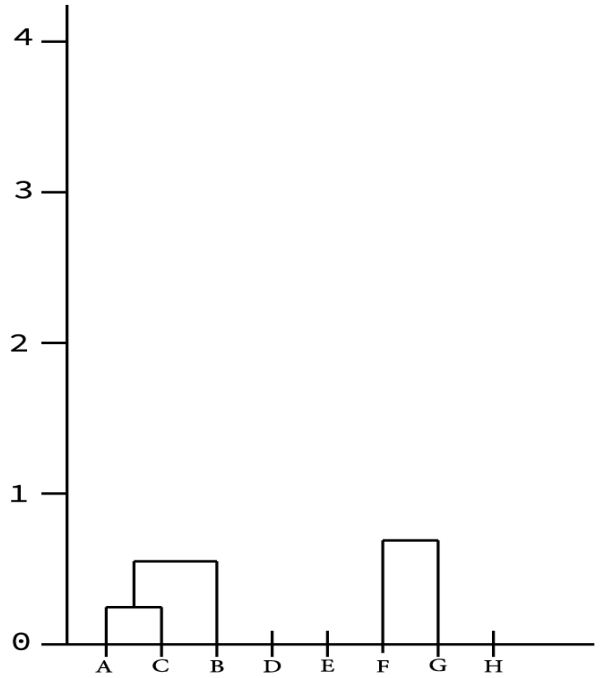




	ACB	D	E	F	G	H
ACB	0					
D	4.72	0				
E	5.55	1	0			
F	4.07	2.01	2.06	0		
G	4.68	2.06	1.81	0.61	0	
H	4.75	3.16	2.90	1.28	1.12	0

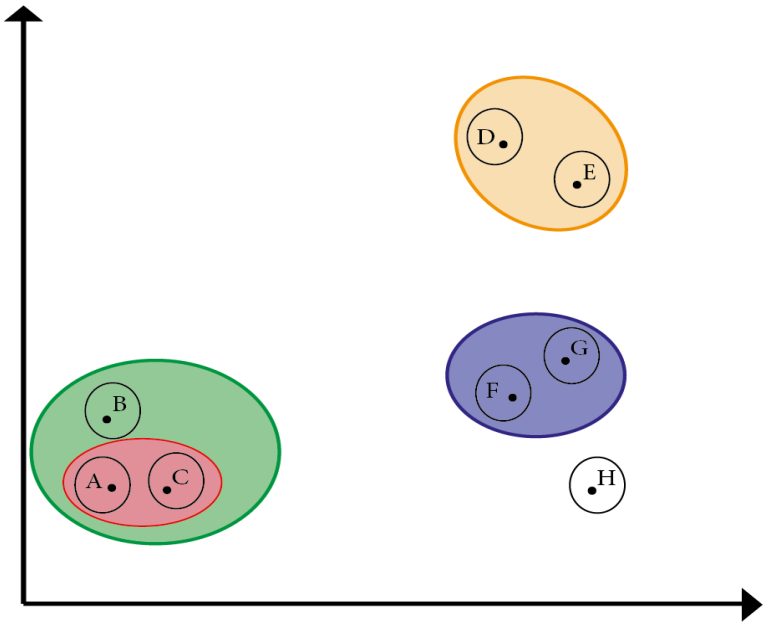
Etape 3

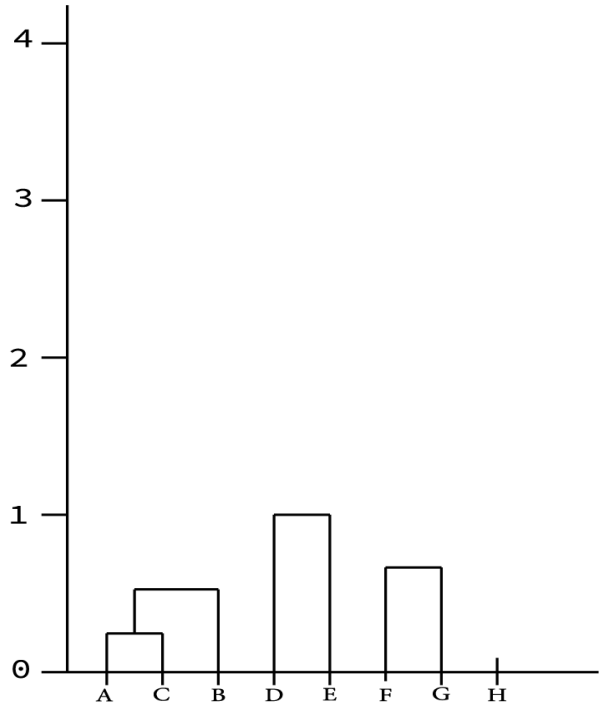




	ACB	D	E	FG	H
ACB	0				
D	4.72	0			
E	5.55	1	0		
FG	4.07	2.01	1.81	0	
H	4.75	3.16	2.90	1.12	0

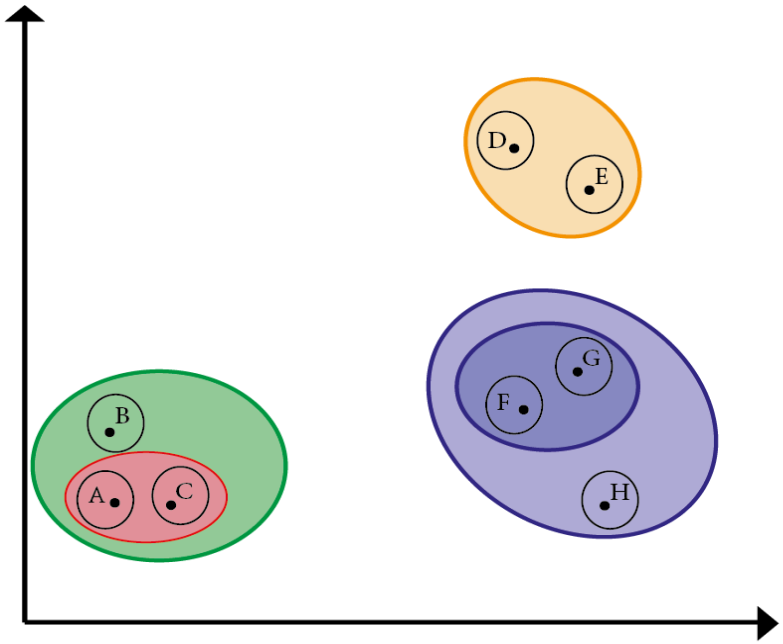
Etape 4

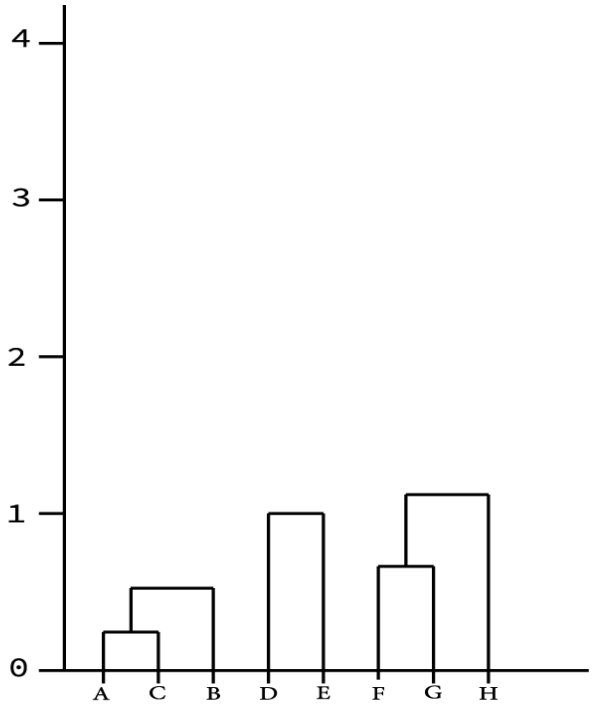




	ACB	DE	FG
DE	4.72	0	
FG	4.23	1.81	0
H	4.07	2.90	1.12

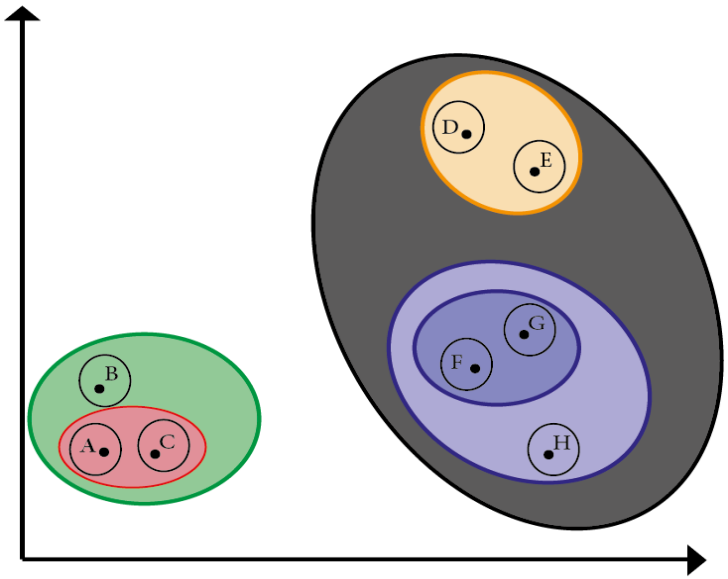
Etape 5

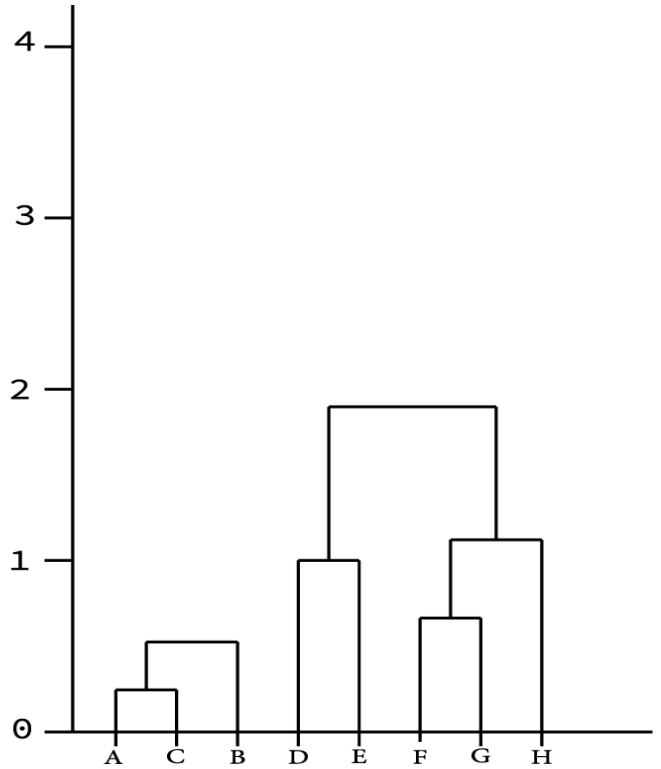




	ACB	DE
DE	4.72	0
FGH	4.01	(1.81)

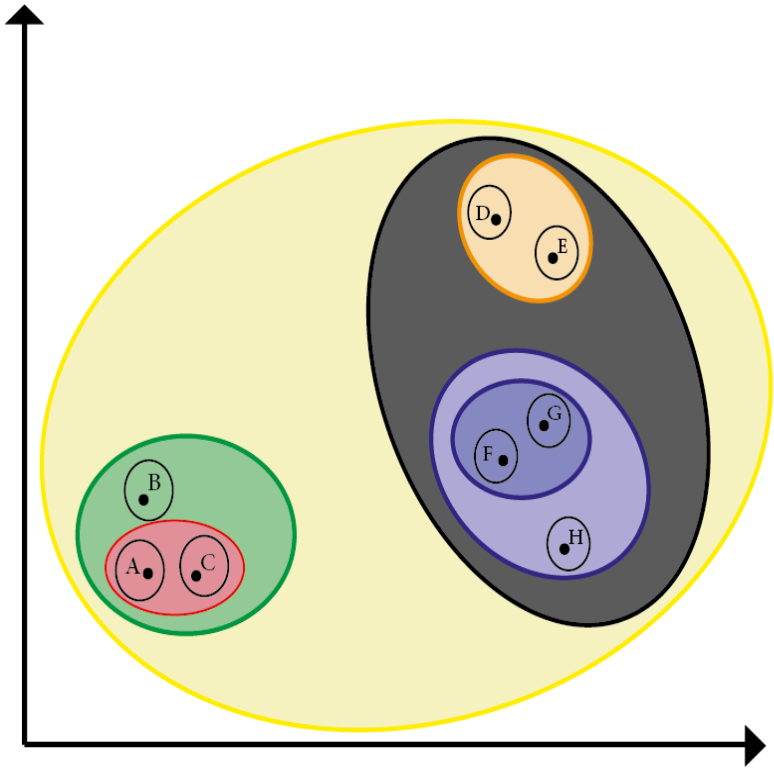
Etape 6

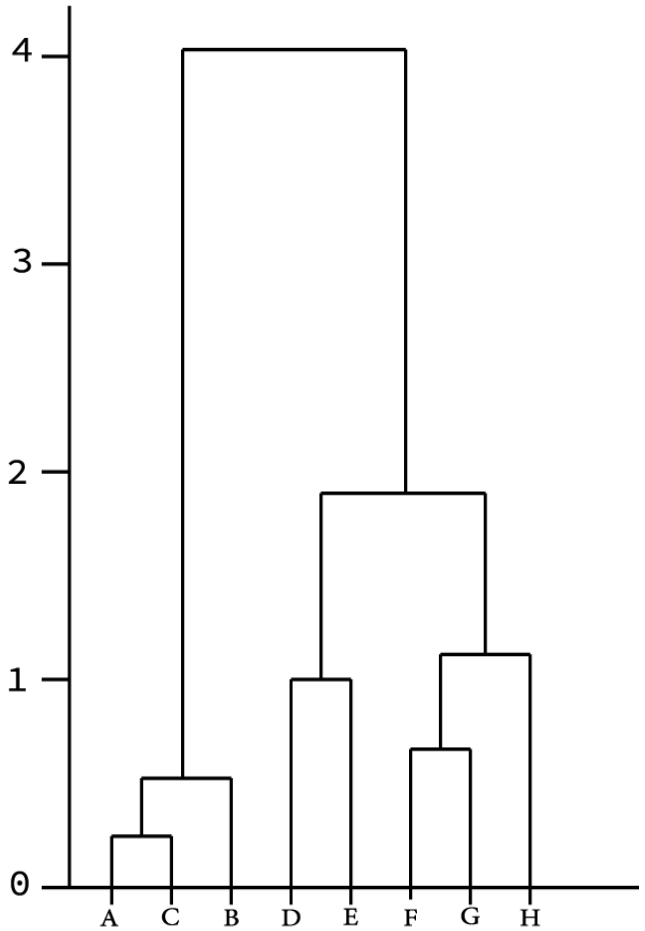




	ACB
DEFGH	4.07

Etape 7





Les avantages et les inconvénients

Avantages

- Aide au choix du nombre de Cluster
- Hiérarchique : La construction des groupes se fait etape par etape.
- Facile à utiliser : La CAH est relativement facile à comprendre et à appliquer

Les inconvénients

- Sensibilité à la taille des données
- Elle est exigeante en termes de temps de calcul.
- Une fois que deux individus sont groupés, ils ne seront jamais séparés.

L'algorithme CAH est une méthode flexible pour explorer la structure des données et détecter des groupes similaires. Il est important de bien préparer les données et de choisir le nombre de groupes avec soin pour obtenir des résultats interprétables.