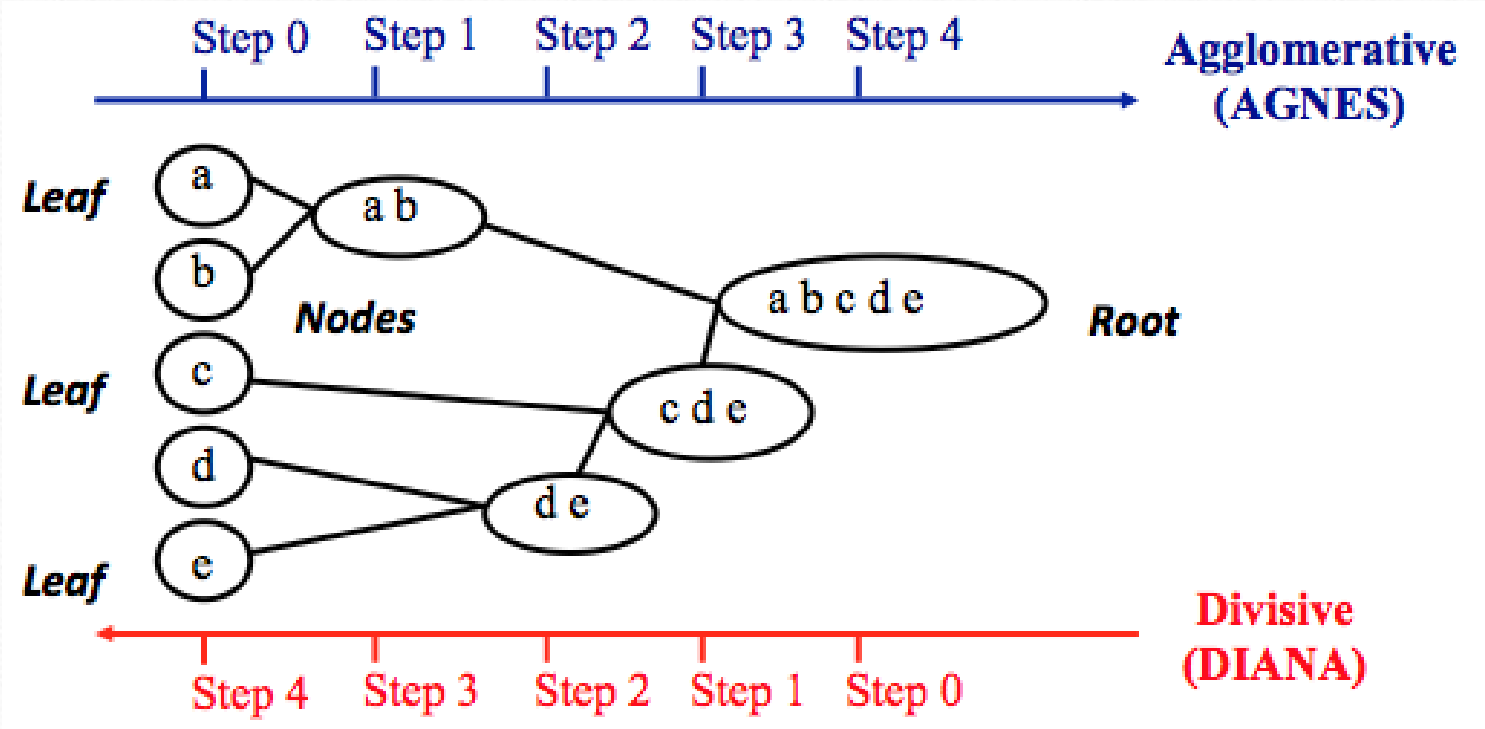
Le Clustering

Méthodes hiérarchiques

Clustering hiérarchique

- Le regroupement hiérarchique consiste à créer une décomposition hiérarchique des observations en fonction de certains critères.
- On distingue:
  - Regroupement Hiérarchique Ascendant (Agglomerative Hierarchical Clustering) CHA
  - Regroupement Hiérarchique Descendant
- Différentes méthodes : différentes définitions de la mesure de dissimilarité entre clusters

Clustering hiérarchique



AGNES (AGglomerative NESting): Algorithme

Initialisation

- Chaque individu est placé dans son propre cluster
- Calcul de la matrice de ressemblance M entre chaque couple de clusters

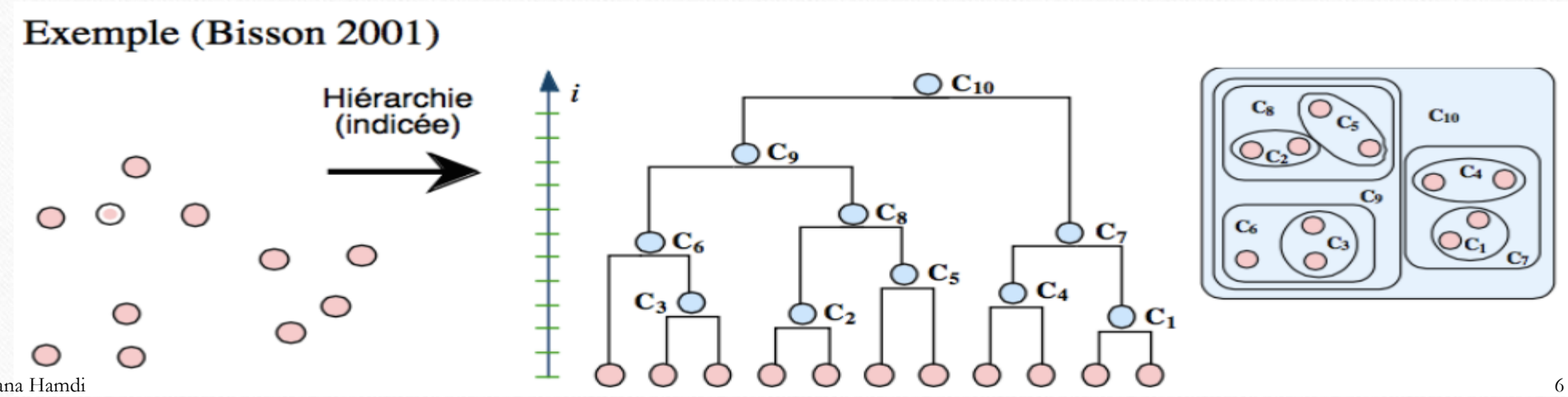
Répéter

- Sélection dans M des deux clusters les plus proches C_i et C_j
- Fusion de C_i et C_j par un cluster C_g plus général
- Mise à jour de M en calculant la ressemblance entre C_g et les clusters existants

Jusqu'à la fusion des 2 derniers clusters ou bien une condition d'arrêt est vérifiée (ex: obtention de *k* clusters)

Dendrogramme

- Représentation des fusions successives
- Un clustering est obtenu en coupant le dendrogramme au niveau choisi
- Hauteur d'un cluster dans un Dendrogramme représente la similarité des 2 clusters avant fusion



AGNES: Exemple

- Procédure adoptée :
  - regrouper ensemble tous les éléments dont la distance est inférieure ou égale à 3 et uniquement ceux-ci.

D	Xav	Yves	Ziad	Tania	Ute
Xav	0				
Yves	4,5	0			
Ziad	2	3,5	0		
Tania	5	6	4,5	0	
Ute	4	5,5	2,5	1,5	0

AGNES: Exemple

- **1ère étape:** La paire constituée des éléments les plus proches constituera le premier «agrégat».
- On regroupe *Tania* et *Ute,* on aura le nouvel ensemble {*Xav*, *Yves*, *Ziad*, (*Tania*, *Ute*)}
- On recalcule la matrice des distances en considérant que la distance entre un élément et un ensemble d'éléments est le **min des distances** à chaque élément.

D	Xav	Yves	Ziad	Tania,Ute
Xav	0			
Yves	4,5	0		
Ziad	2	3,5	0	
Tania,Ute	4	5,5	2,5	0

AGNES: Exemple

- **<u>2ème étape :</u>** on cherche la paire qui constituera le deuxième agrégat
- *Xav* et *Ziad* → nouvel ensemble {(*Xav*,*Ziad*),*Yves*, (*Tania*, *Ute*)}
- On recalcule la matrice des distances

D	Xav,Ziad	Yves	Tania,Ute
Xav, Ziad	0		
Yves	3,5	0	
Tania,Ute	2,5	5,5	0

AGNES: Exemple

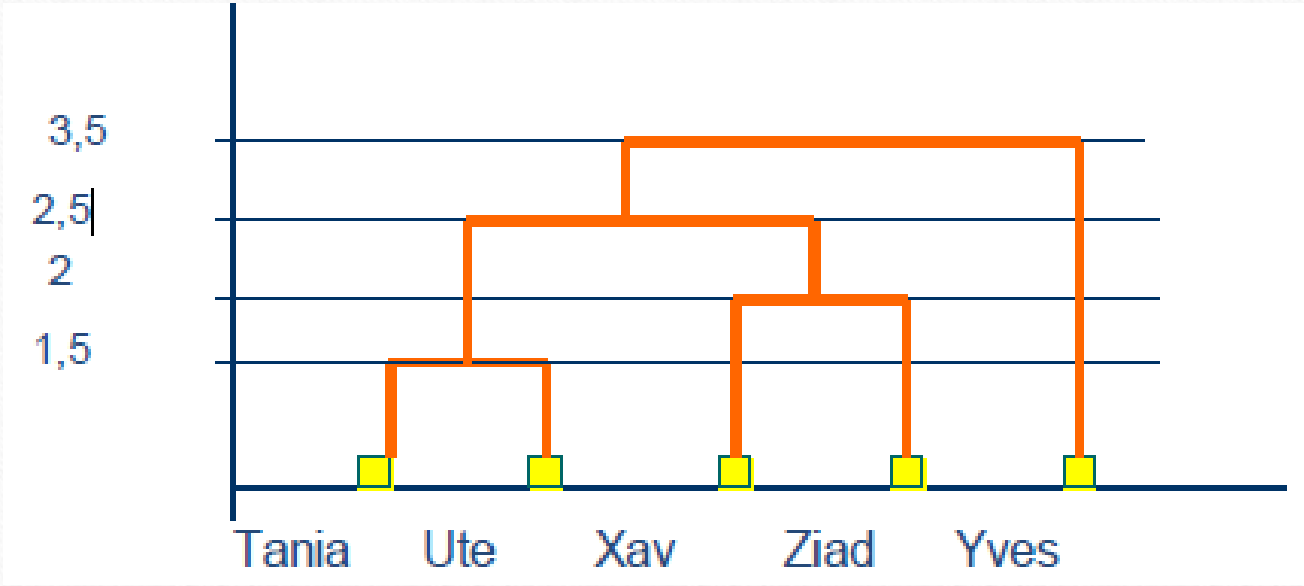
- **<u>3ème étape :</u>** on cherche la paire qui constituera le troisième agrégat
- les objets les plus proches sont (*Xav, Ziad*) et (*Tania*, *Ute*)→nouvel ensemble {(*Xav*, *Ziad*, *Tania*, *Ute*),*Yves*}
- On recalcule la matrice des distances

D	Xav, Ziad, Tania, Ute	Yves
Xav, Ziad, Tania,Ute	0	
Yves	3,5	0

AGNES: Exemple

- **<u>Conclusion :</u>**
- La dernière étape, symbolique, consiste à constater que la distance de *Yves* à (*Xav*, *Ziad*, *Tania*, *Ute*), est 3,5 > 3 → on arrête
- Si on veut distinguer trois groupes ou agrégats, le premier formé des jurés *Tania* et *Ute* distant de celui de *Xav* et *Ziad* de 2,5 unités et enfin au loin le juré *Yves* distant de 3,5 de ses voisins les plus proches.
- *Il semble que ce dernier devra être exclu du jury.*

Résultat : Dendrogramme



Autres stratégies

- La stratégie d'agrégation de l'exemple est une stratégie du **"minimum".**
- Peut-on en imaginer d'autres ? → On peut en envisager au moins 3 autres!

1. Saut minimal (single linkage) basé sur la distance Dmin(C1, C2) $d(A,B) = \min_{i \in A, j \in B} d(i,j)$

- fournit des classes générales

2. Saut maximal (complete linkage) basé sur la distance Dmax(C1, C2) $d(A,B) = \sup_{i \in A, j \in B} d(i,j)$

- fournit des classes spécifiques

3. Saut moyen basé sur la distance Dmoy(C1, C2) $d(A,B) = \frac{1}{card(A)card(B)} \sum_{i \in A, j \in B} d(i,j)$

- fournit des classes de variance proche

4. Barycentre basé sur la distance Dcg(C1, C2) $d(A,B) = d(g_A, g_B)$

- bonne résistance au bruit

DIANA (DIvisive ANAlysing) : Algorithme

Initialisation

- Tous les objets sont placés dans le même cluster

Répéter

- Dans la classe C présentant la plus grande dissimilarité entre deux objets, on sépare ses objets en deux classes A et B.
- Les objets de la classe C scindée en deux sont affectés à l'un ou l'autre des deux classes A ou B créées suivant l'algorithme suivant :
  - A est au départ constitué de tous les objets de C, B est vide
  - Pour chaque objet i de A, on calcule la dissimilarité moyenne aux autres objets de A. On affecte l'objet m ayant la plus forte dissimilarité moyenne dans le groupe B. On a alors A= A \{m} et B={m}
  - Pour chaque objet de A, on calcule la dissimilarité moyenne à A et à B. L'objet ayant la plus forte différence d(i,A)-d(i,B) est affecté au groupe B si la différence est positive sinon on s'arrête

Jusqu'à ce que chaque cluster contient un seul objet ou une condition d'arrêt est vérifiée (ex: obtention de *k* clusters)

DIANA: Exemple

D	Xav	Yves	Ziad	Tania	Ute
Xav	0				
Yves	4,5	0			
Ziad	2	3,5	0		
Tania	5	6	4,5	0	
Ute	4	5,5	2,5	1,5	0

Clustering hiérarchique: Avantages

- L'un des résultats est le dendrogramme:

→ Ce qui permet de visualiser le regroupement progressif des données

→ On peut alors se faire une idée d'un nombre adéquat de classes dans lesquelles les données peuvent être regroupées.

Clustering hiérarchique: Limites

- Résultats différents en fonction de la paramétrisation
  - Distances différentes
  - Choix d'agrégation différents
  - Lourdeur des calculs dès qu'on a un nombre de données important
- les regroupements sont définitifs, ce qui ne permet pas d'optimisation postérieure au clustering