APPRENTISSAGE NON SUPERVISÉ

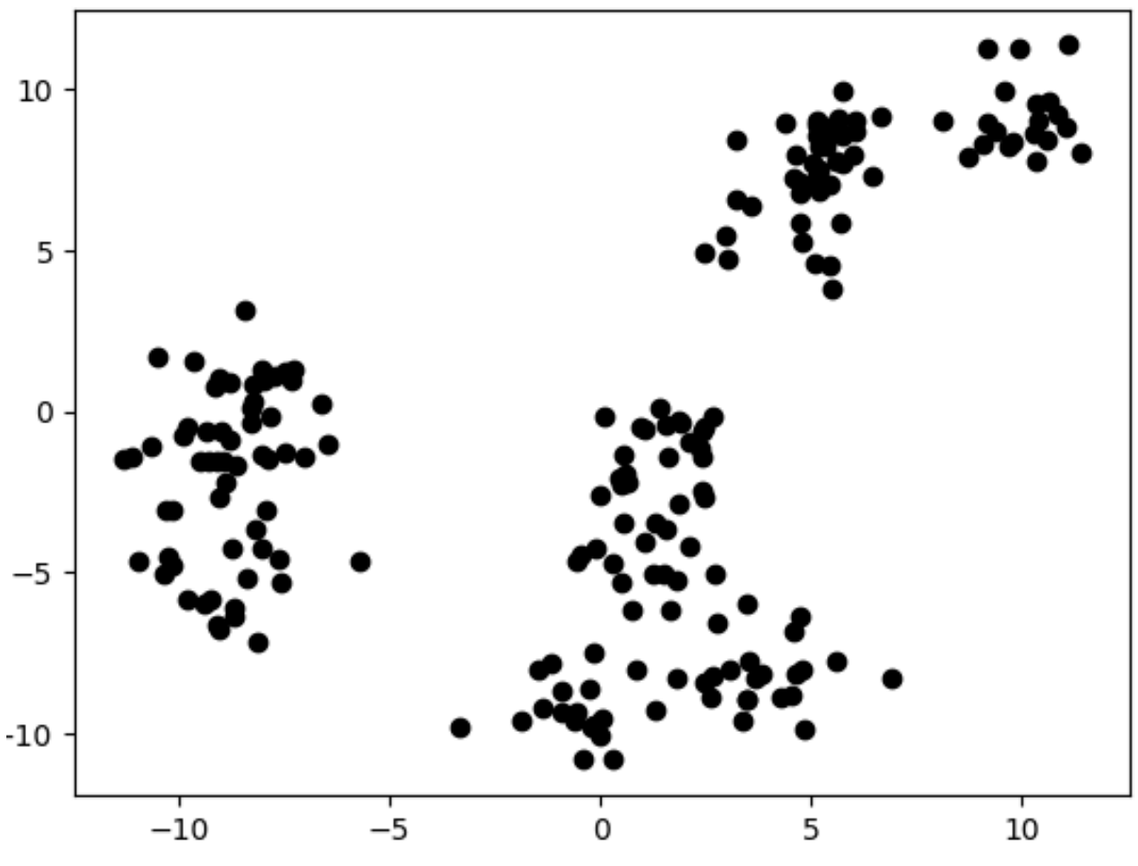
Cours 3

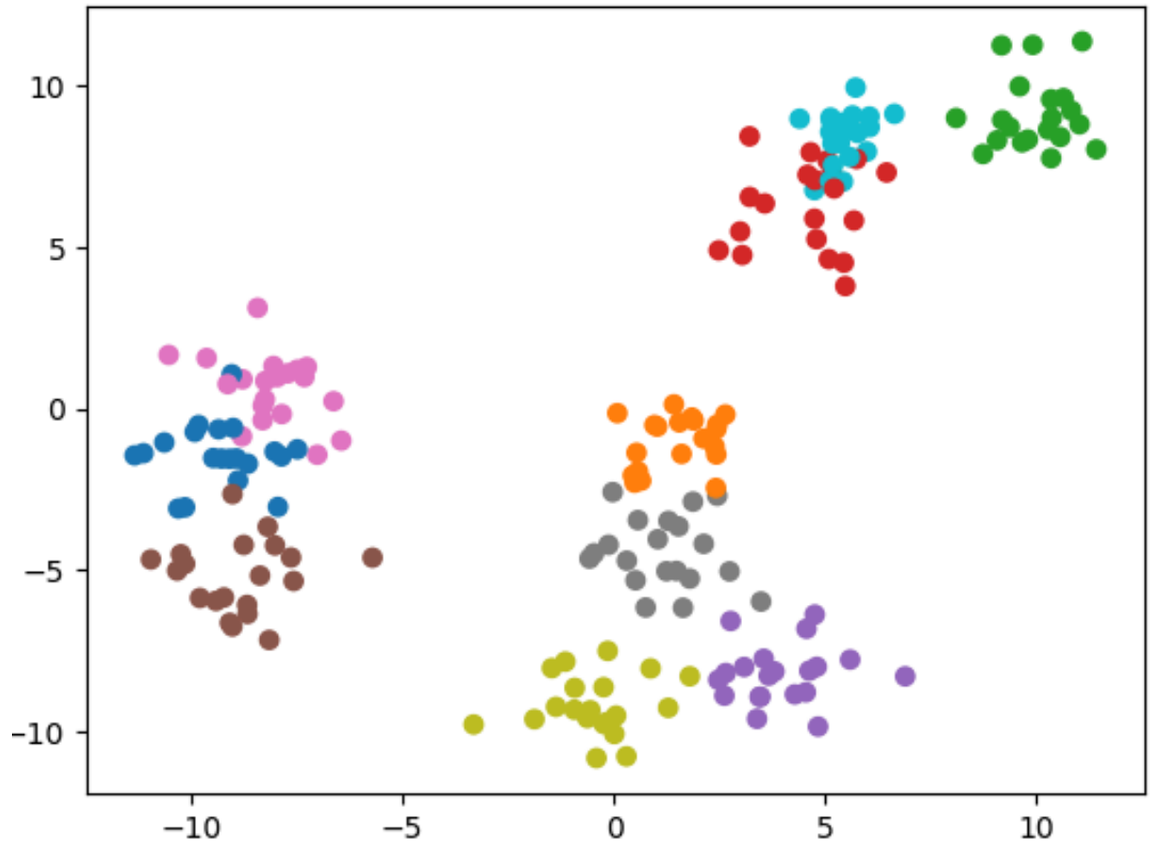
Caio Corro

APPRENTISSAGE NON SUPERVISÉ

Clustering ou partitionnement des données

- Recherche de la structure sous-jacente des données
- Regrouper les points similaires dans les mêmes clusters / groupes / classes





EXEMPLE : TYPOLOGIE LINGUISTIQUE

Peut-on regrouper les langues par rapport à leur structures syntaxiques/morphologiques communes ?

THE WORLD ATLAS OF LANGUAGE STRUCTURES ONLINE

Home | Features | Chapters | Languages | References | Authors

Family: Indo-European / Genus: Romance

Language French

WALS code: fre

Showing 1 to 158 of 158 entries

Fid	Value	Feature	Reference	Area
	Search	Search		
1A	Average	Consonant Inventories	Fougeron and Smith 1999; Sten 1963	Phonology
2A	Large (7-14)	Vowel Quality Inventories	Fougeron and Smith 1999; Sten 1963	Phonology
3A	Low	Consonant-Vowel Ratio	Fougeron and Smith 1999; Sten 1963	Phonology
4A	In both plosives and fricatives	Voicing in Plosives and Fricatives	Fougeron and Smith 1999; Sten 1963	Phonology
5A	None missing in /p t k b d g/	Voicing and Gaps in Plosive Systems	Fougeron and Smith 1999; Sten 1963	Phonology
6A	Uvular continuants only	Uvular Consonants	Fougeron and Smith 1999; Sten 1963	Phonology
7A	No glottalized consonants	Glottalized Consonants	Fougeron and Smith 1999; Sten 1963	Phonology
8A	/l/, no obstruent laterals	Lateral Consonants	Fougeron and Smith 1999; Sten 1963	Phonology
9A	No velar nasal	The Velar Nasal	Byrne et al. 1986	Phonology
10A	Contrast present	Vowel Nasalization	Harris 1988	Phonology
11A	High and mid	Front Rounded Vowels	Fougeron and Smith 1999; Sten 1963	Phonology
12A	Complex	Syllable Structure	Fougeron and Smith 1999; Sten 1963	Phonology
13A	No tones	Tone	Fougeron and Smith 1999; Sten 1963	Phonology

Glottocode: stan1290 ISO 639-3: fra

Coordinates WGS84 48°N, 2°E
48.00, 2.00

Spoken in: Switzerland, France

Alternative names

Ethnologue: French

Ruhlen: French

Sources

Byrne et al. 1986
A Comprehensive French Grammar
Dahl 1985
Tense and Aspect Systems
info at Google Books
Dell 1984
L'accentuation des phrases en Français
Dubois et al. 1955
Larousse's French-English English-French Dictionary
info at Google Books

EXEMPLE : TYPOLOGIE LINGUISTIQUE

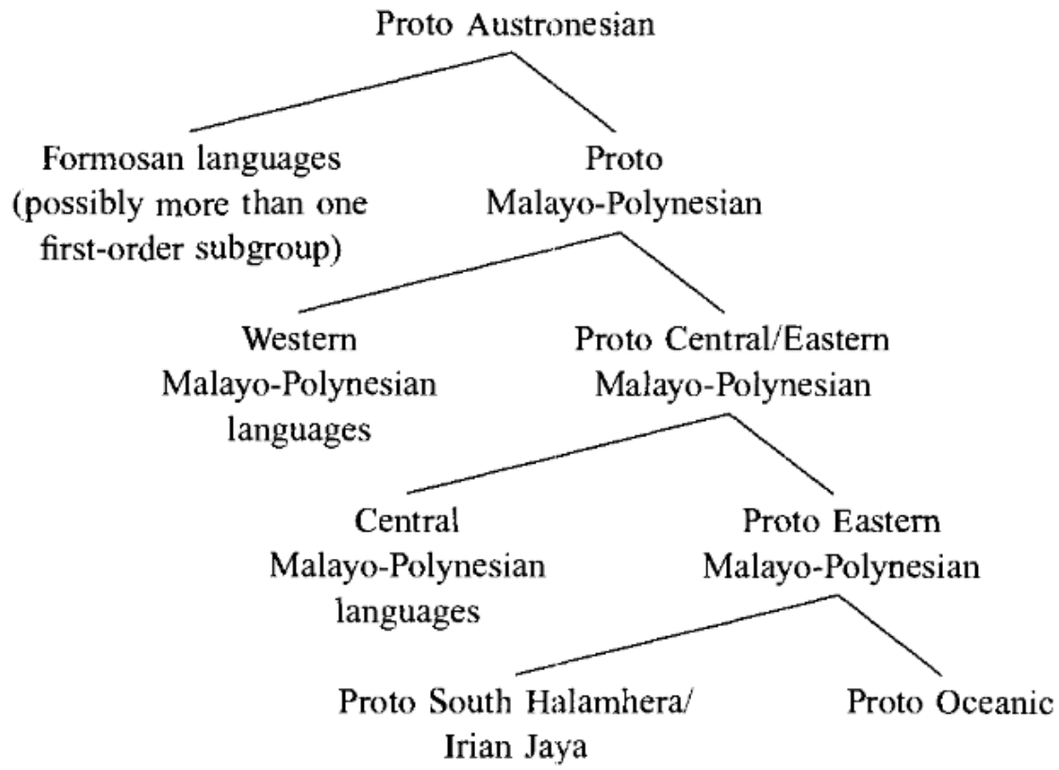


FIGURE 1.1 HIGHER-ORDER AUSTRONESIAN SUBGROUPS

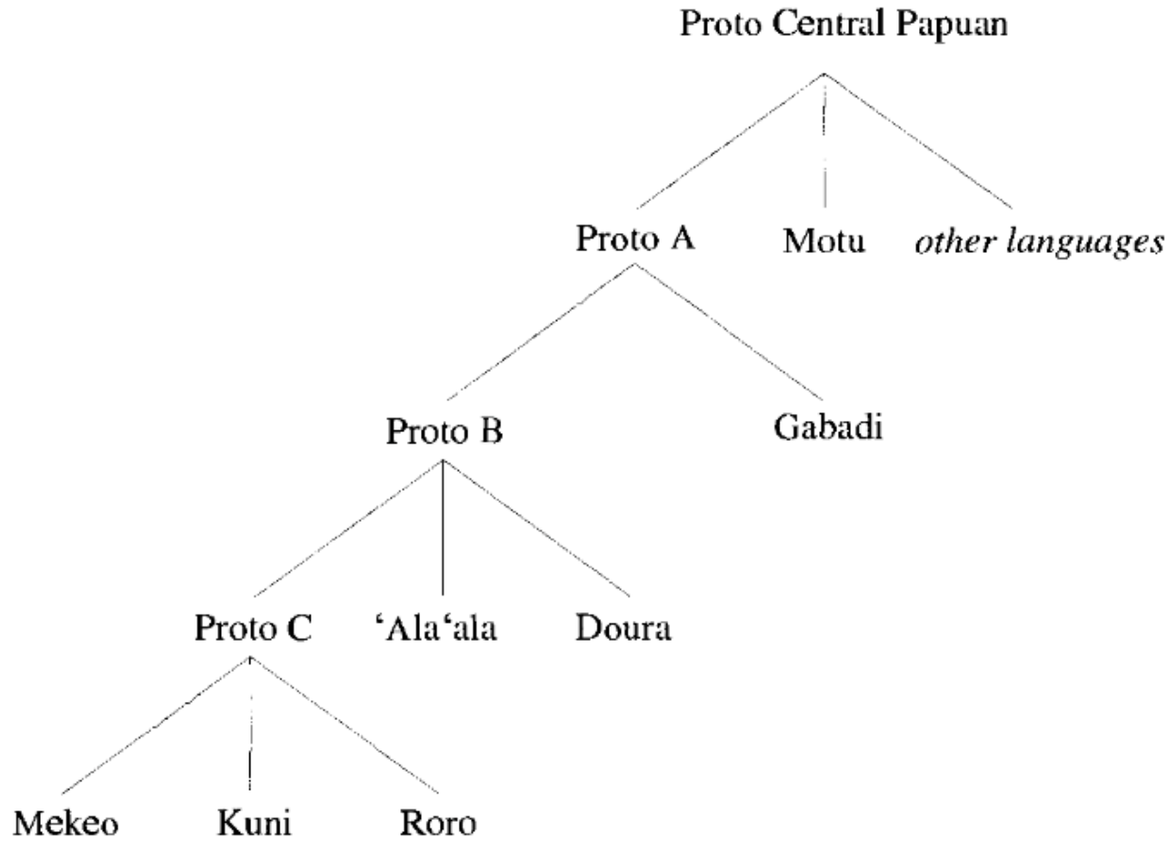


FIGURE 4.1 SUBGROUPING OF SOME OF THE CENTRAL PAPUAN LANGUAGES

The oceanic languages (Lynch et al.)

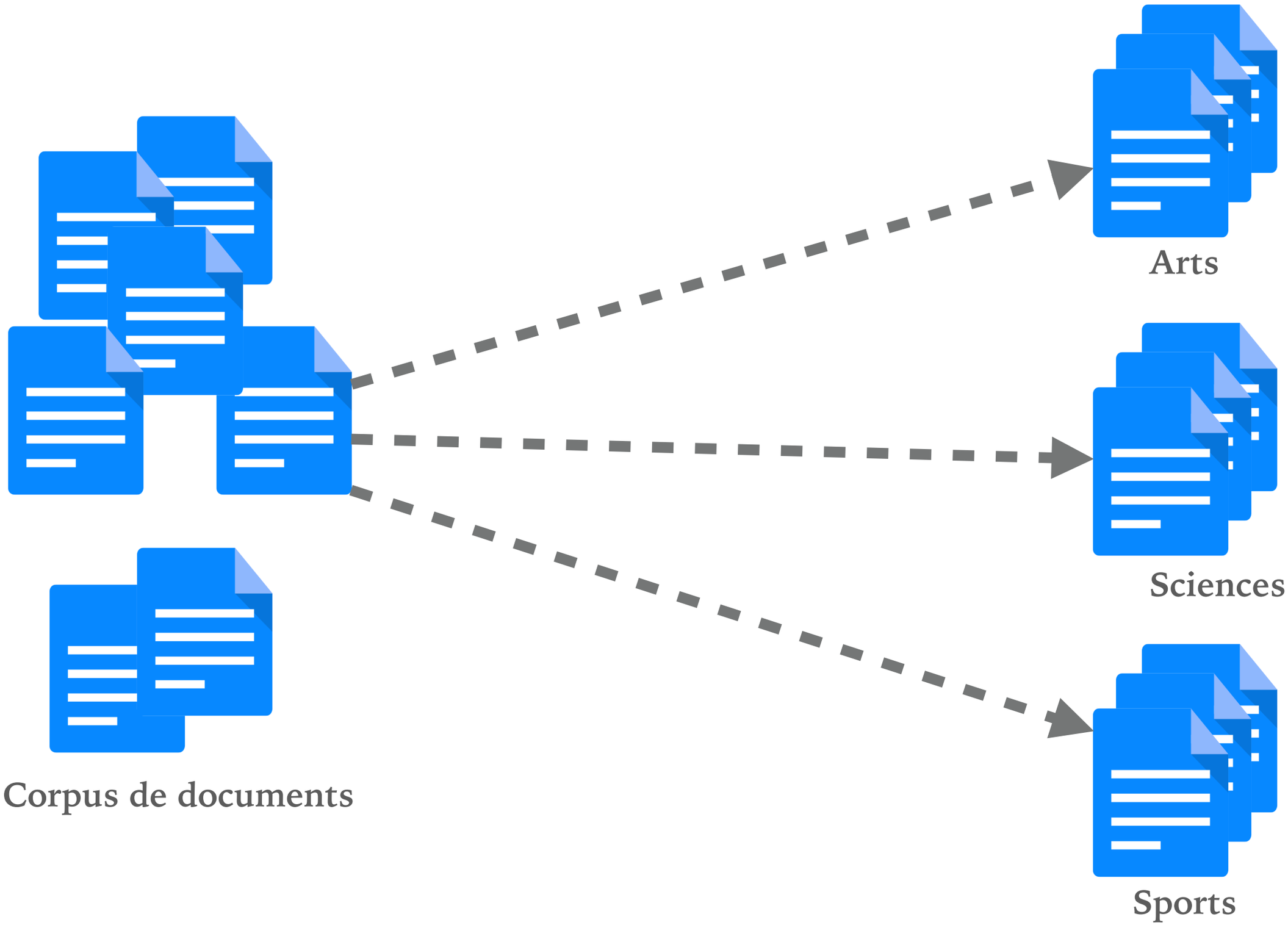
EXEMPLE : CLUSTERING D'ANIMAUX

Oiseaux

Mammifères

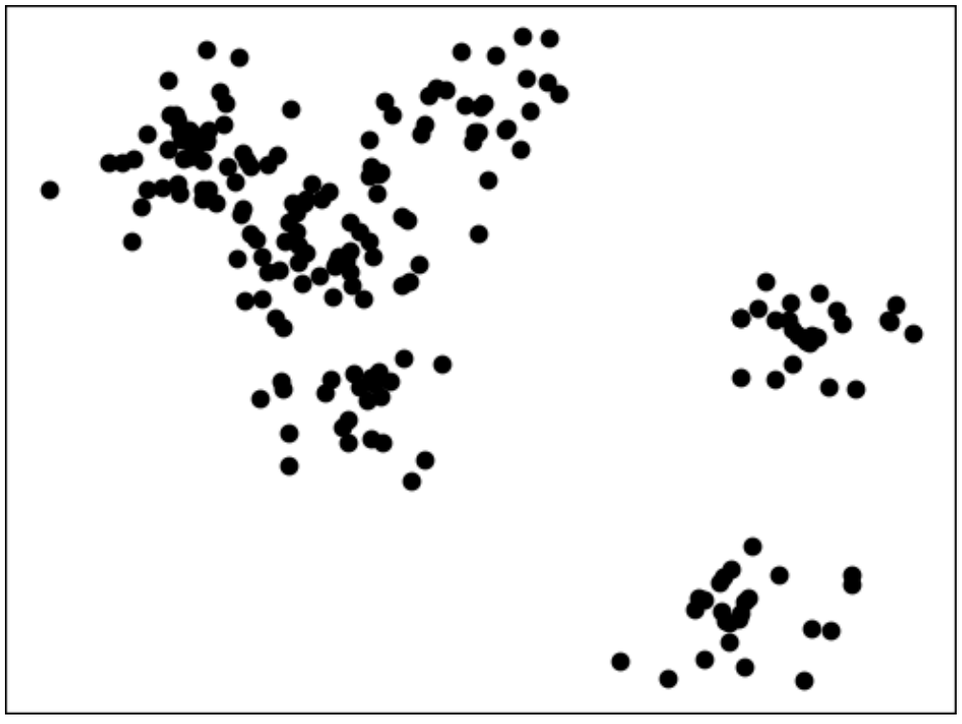
Reptiles

EXEMPLE : CLUSTERING DE DOCUMENTS



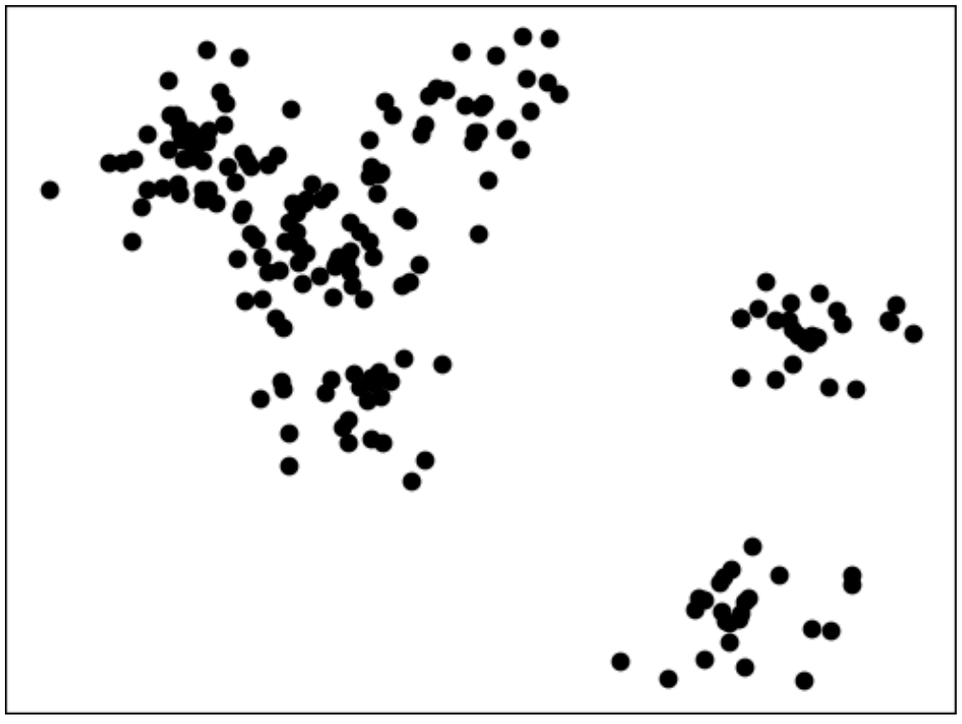
COMBIEN DE CLUSTERS ?

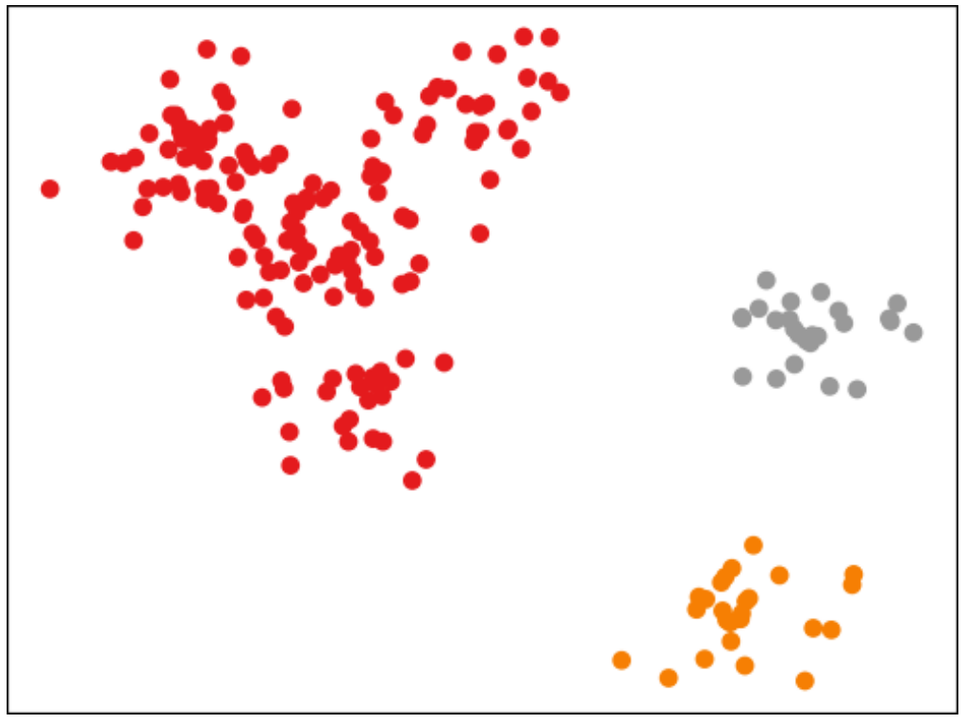
Combien de clusters/groupes y a-t-il dans ces données ?



COMBIEN DE CLUSTERS ?

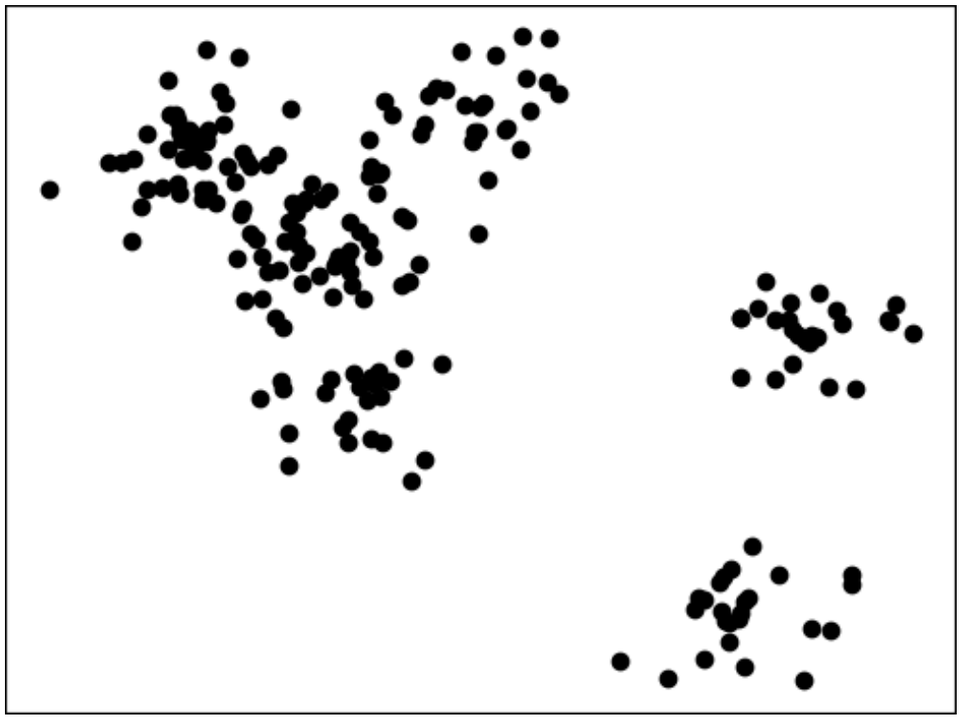
Combien de clusters/groupes y a-t-il dans ces données ?

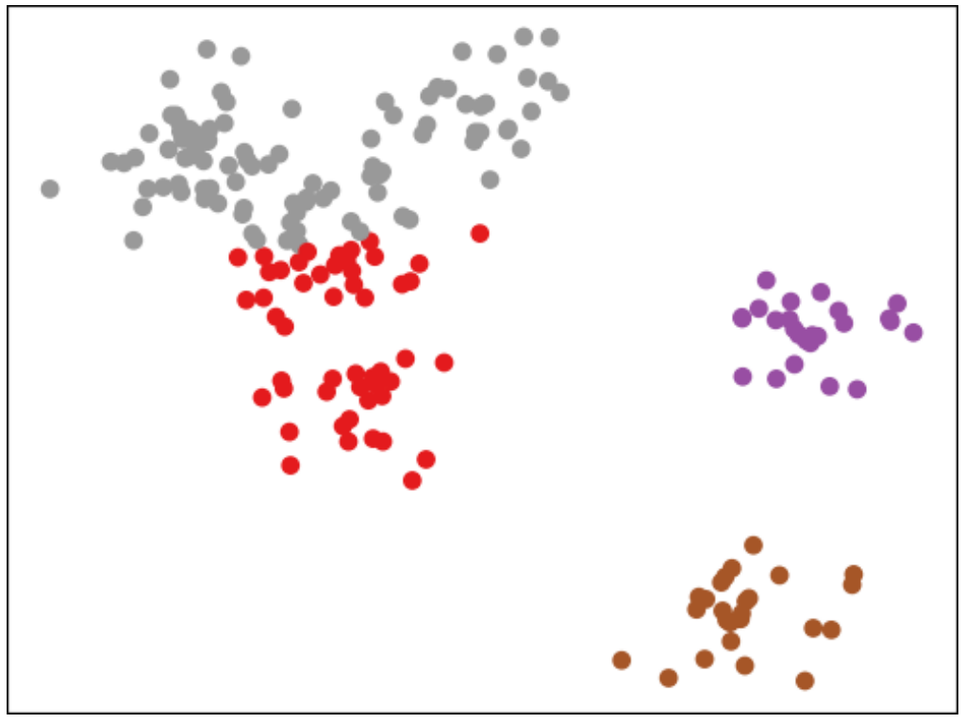




COMBIEN DE CLUSTERS ?

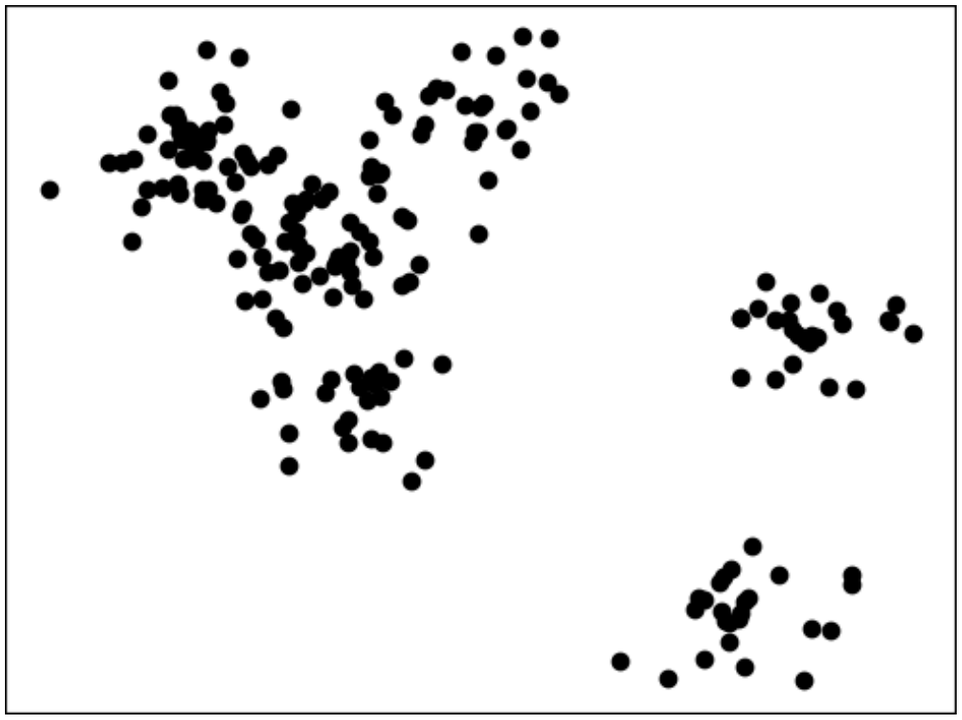
Combien de clusters/groupes y a-t-il dans ces données ?

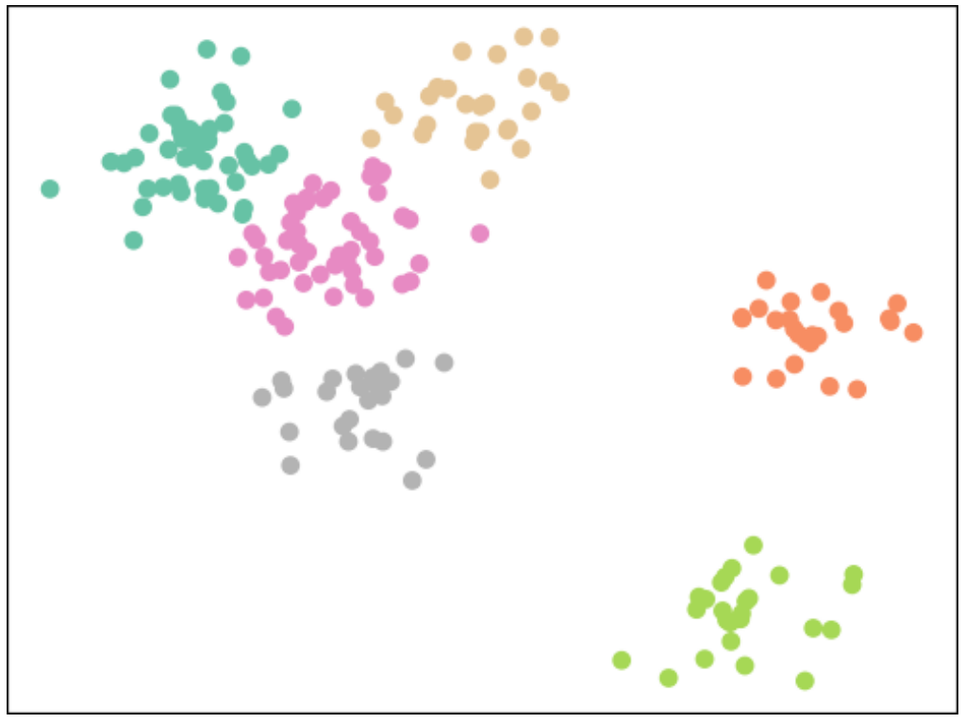




COMBIEN DE CLUSTERS ?

Combien de clusters/groupes y a-t-il dans ces données ?





Problème

- Très difficile de savoir à l'avance combien de clusters / groupes on veut trouver dans les données
- Cependant, très utile pour l'exploration des données
- Parfois on a des informations sur les données qui nous permettent de fixer le nombre de clusters

PARTITIONS

Vocabulaire : ensemble / multiensemble

Pour simplifier les notations, dans la suite on manipulera des ensembles,
mais cela devrait être des multiensembles si on voulait être rigoureux

Partition

Soit X un ensemble.

Un ensemble $\pi = \{C^{(1)}, \ldots C^{(k)}\}$, où chaque C(i) est nommé une partie ou un cluster, est une partition de P si et seulement si :

- Les parties/clusters de π sont non vides : $\forall i \quad : \quad C^{(i)} \neq \varnothing$
- L'union de toutes les parties/clusters de π est égale à D :

$$X = \bigcup_{C \in \pi} C \quad = \bigcup_{i=1}^{k} C^{(i)} \quad = C^{(1)} \cup C^{(2)} \cup \ldots \cup C^{(k)}$$

- Les parties/clusters de π sont deux à deux disjointes :

$$\forall i, j \quad s.t. \quad i \neq j \quad : \quad C^{(i)} \cap C^{(j)} = \varnothing$$

PARTITIONS

$$D = \{\ 1, 2, 3, 4, 5\ \}$$

Les ensembles suivant <u>sont</u> des partitions de D :

$$\pi = \left\{\ \{1\}, \{2\}, \{3\}, \{4\}, \{5\}\ \right\} \qquad \pi' = \left\{\ \{1, 2, 3, 4, 5\ \}\ \right\}$$

$$\pi'' = \left\{\ \{1,3,5\}, \{2,4\},\ \right\} \qquad \pi''' = \left\{\ \{1,3\}, \{2,4\}, \{\ 5\ \}\ \right\}$$

Les ensembles suivant ne <u>sont</u> des partitions de D :

$$\pi'''' = \left\{\ \{1\}, \{2\}, \{3\}, \{4,5\}, \{\ \}\ \right\} \qquad \pi''''' = \left\{\ \{1,3\}, \{2,4,5\}, \{\ 5\ \}\ \right\}$$

$$\pi'''''' = \left\{\ \{1,5\}, \{2,4\}\ \right\}$$

DISPERSION

Soit X un ensemble,

et $\pi = \{C^{(1)}, \dots C^{(k)}\}$ un partition de X.

Centroïde

On dénote m_i le centroïde du cluster $C^{(i)}$: $$m^{(i)} = \frac{1}{|\,C^{(i)}\,|} \sum_{x \in C^{(i)}} x$$

Dispersion intra-clusters

Dispersion des éléments qui composent chaque cluster autour de son centroïde :

$$\sum_{i=1}^{k} \sum_{x \in C^{(i)}} \|x - m^{(i)}\|_2^2$$

DISPERSION

Soit X un ensemble,

et $\pi = \{C^{(1)}, \dots C^{(k)}\}$ un partition de X.

Centroïde

On dénote m_i le centroïde du cluster $C^{(i)}$: $m^{(i)} = \frac{1}{|C^{(i)}|} \sum_{x \in C^{(i)}} x$

On dénote $\bar{x}$ le centroïde des données X : $\bar{x} = \frac{1}{|X|} \sum_{x \in X} x$

Dispersion inter-clusters

Dispersion des centroïdes des groupes autour des centroïdes des données :

$$\sum_{i=1}^{k} |C^{(i)}| \times \|m^{(i)} - \bar{x}\|_2^2$$

DISPERSION

Dispersion intra-clusters

Dispersion des éléments qui composent chaque cluster autour de son centroïde :

$$\sum_{i=1}^{k} \sum_{x \in C^{(i)}} \|x - m^{(i)}\|_2^2$$

Dispersion inter-clusters

Dispersion des centroïdes des groupes autour des centroïdes des données :

$$\sum_{i=1}^{k} |C^{(i)}| \times \|m^{(i)} - \bar{x}\|_2^2$$

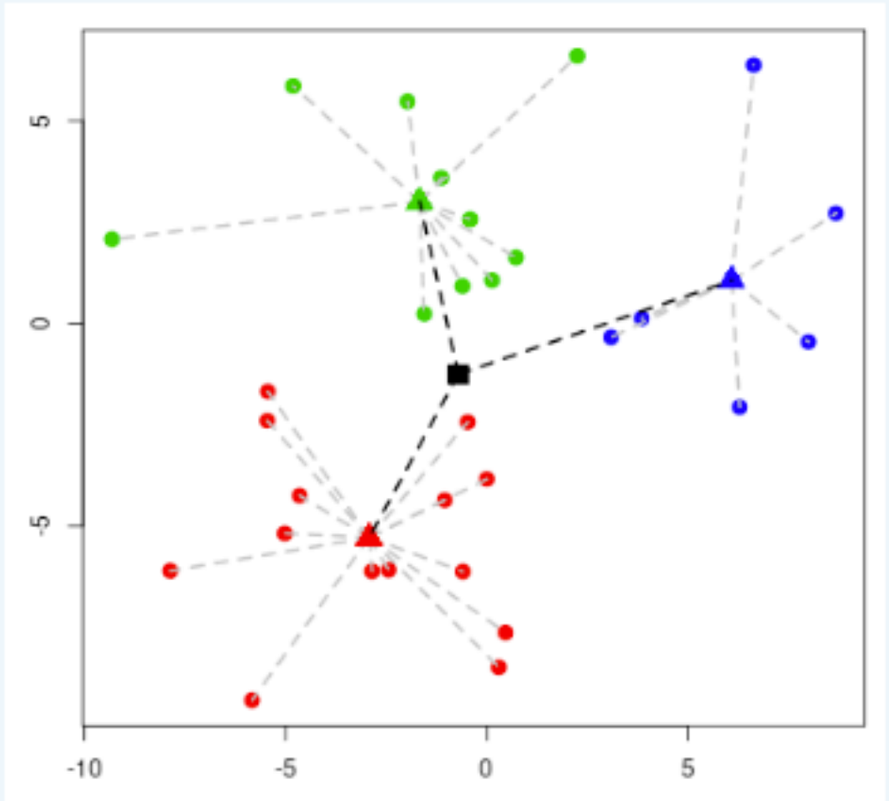
Que veux-t-on obtenir ?

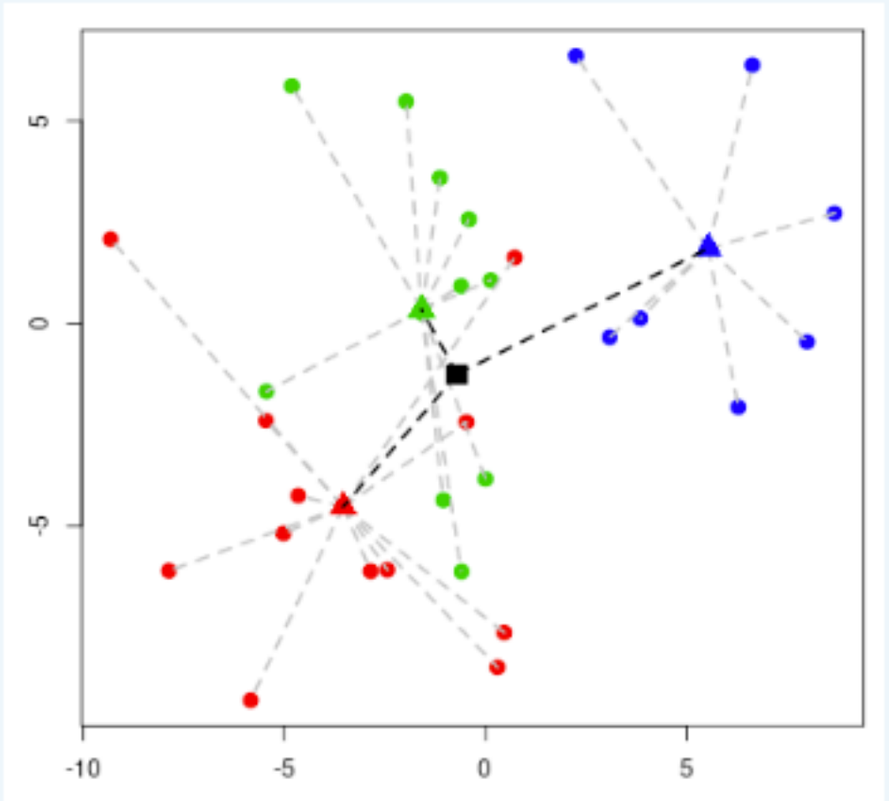
- Dispersion intra-clusters basse
  (c'est-à-dire que les clusters doivent contenir des éléments très similaires entre eux)
- Dispersion inter-clusters élevée
  (les clusters doivent être différents les un des autres)

DISPERSION

Inter-clusters : $\sum_{i=1}^{k} |C^{(i)}| \times \|m^{(i)} - \bar{x}\|_2^2$

Intra-clusters : $\sum_{i=1}^{k} \sum_{x \in C^{(i)}} \|x - m^{(i)}\|_2^2$





	Dispersion inter	Dispersion intra
Gauche	795	379
Droite	601	573

(Stéphane Robin)

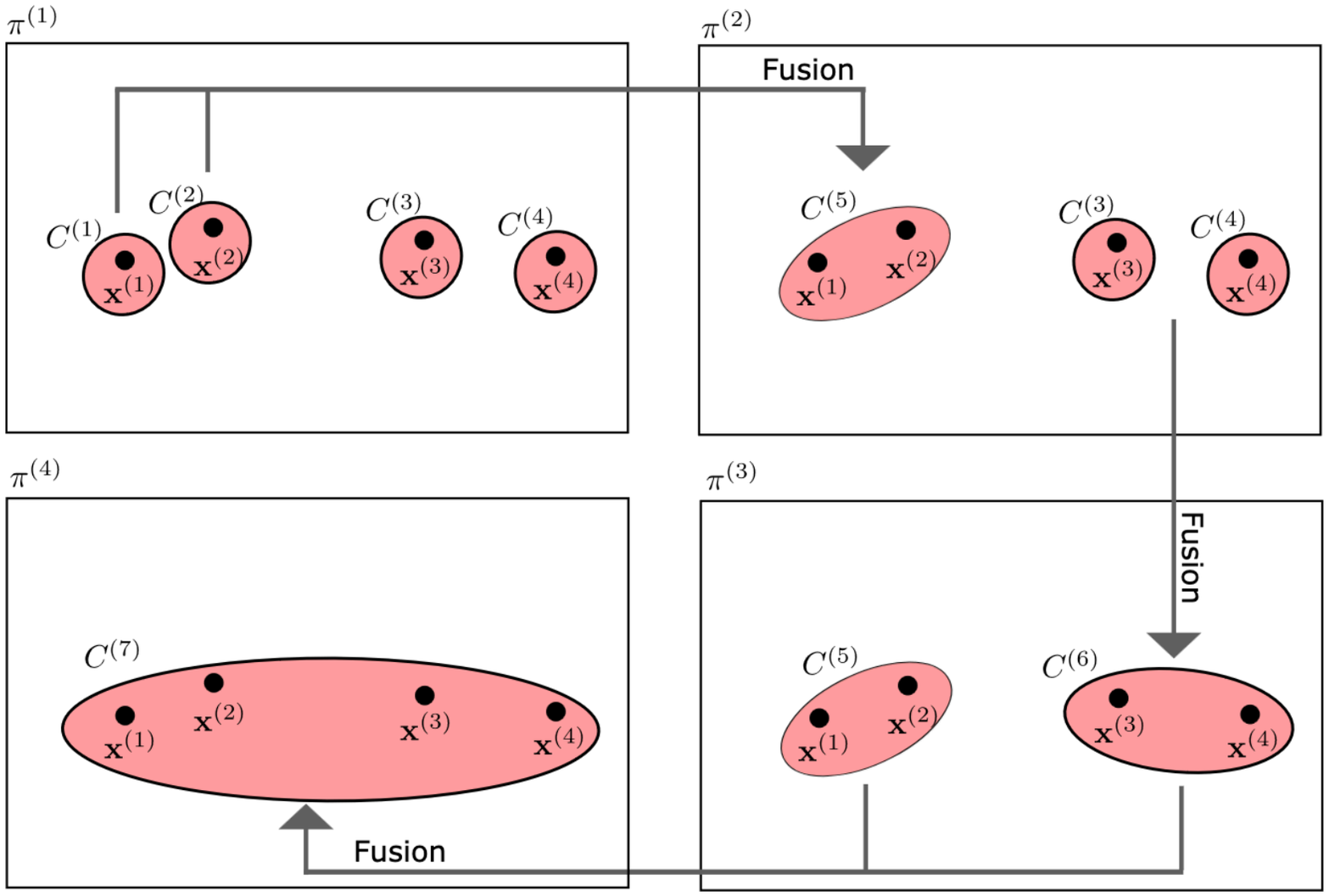
CLUSTERING HIÉRARCHIQUE ASCENDANT

CLUSTERING HIÉRARCHIQUE

Méthode ascendante (ou agglomérative)

Construction d'une hiérarchie de partitions des données de la façon suivante :

- On commence par assigner chaque point des données à son propre cluster
- On va itérativement fusionner les clusters deux à deux jusqu'à obtenir un cluster qui contient toutes les données



PSEUDO-CODE

Entrée : Ensemble X de points, $X = \{ x^{(i)} \}_{i=1}^{n}$ $\{ x^{(1)}, x^{(2)}, \ldots, x^{(n)} \}$

Sortie : Séquence de partitions $\pi^{(1)}$, $\pi^{(2)}$, $\pi^{(3)}$, etc
tel que $|\pi^{(1)}| = |X|$, $|\pi^{(2)}| = |X| - 1$, $|\pi^{(2)}| = |X| - 2$ ... $|\pi^{(|D|)}| = 1$

Algorithme :

1. Initialization : $\pi^{(1)} = \{ \{ x \} \}_{x \in X} = \{ \{ x^{(1)} \}, \ldots, \{ x^{(n)} \} \}$
2. Pour i = 2 à |X| :
    1. Trouver les deux parties/cluster de pi $\pi^{(i-1)}$ les plus proches
    2. Fusionner ces deux clusters pour créer $\pi^{(i)}$

CLUSTERING HIÉRARCHIQUE

Quand s'arrêter ?

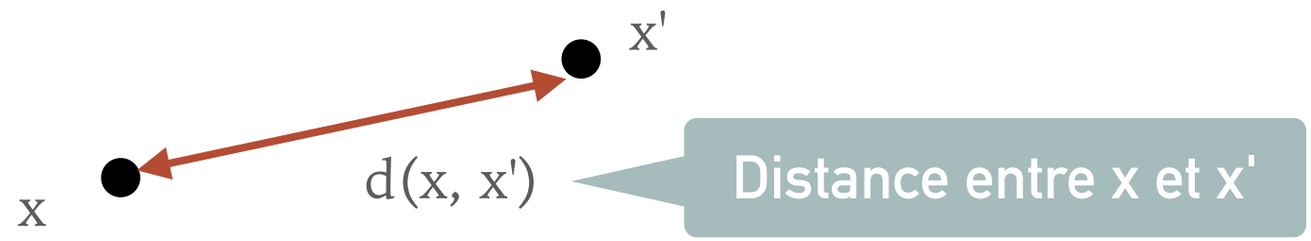
À chaque étape, on a un cluster en moins par rapport à l'étape précédente

- Si on connait le nombre de clusters visés, disons k, on s'arrête à l'étape où l'on a k clusters
- Sinon, on fait toutes les étapes jusqu'à avoir un cluster unique avec toutes les données, et on explore les résultats pour voir si certaines sous-structures sont intéressantes

Comment choisir quels clusters fusionner ?

- Fusionner les deux clusters les plus proches
- Mais comment définir la distance entre deux clusters ?!

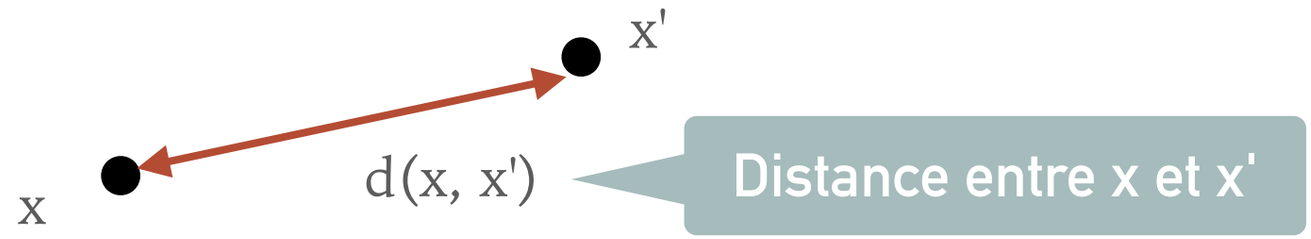
DISTANCES ENTRE DEUX POINTS



Exemples

- Distance euclidienne ou L2 : $d(x,\ x') = \|x - x'\|_2 = \sqrt{\sum_i (x_i - x_i')^2}$

DISTANCES ENTRE DEUX POINTS

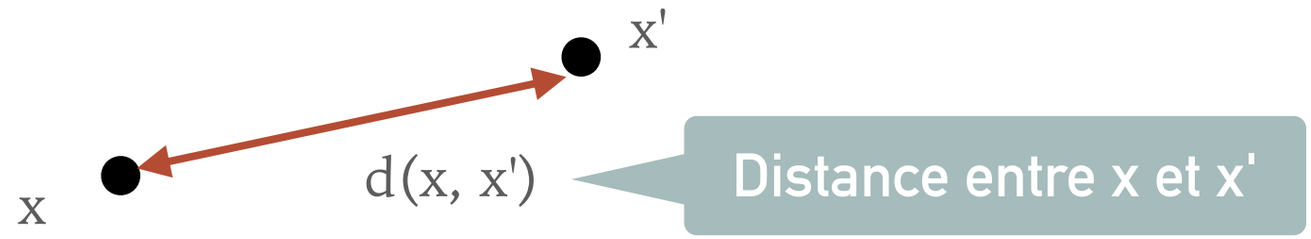


Exemples

- Distance euclidienne ou L2 : $d(x,\ x') = \|x - x'\|_2 = \sqrt{\sum_i (x_i - x_i')^2}$

- Distance de Manhattan ou L1 : $d(x,\ x') = \|x - x'\|_1 = \sum_i |x_i - x_i'|$

DISTANCES ENTRE DEUX POINTS



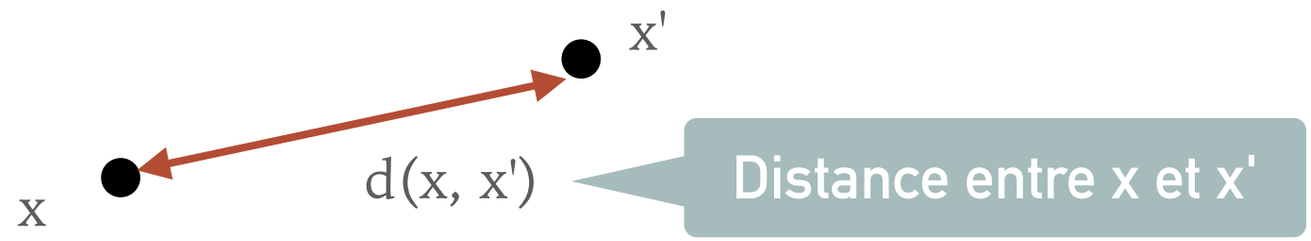
Exemples

- Distance euclidienne ou L2 : $d(x,\ x') = \|x - x'\|_2 = \sqrt{\sum_i (x_i - x'_i)^2}$

- Distance de Manhattan ou L1 : $d(x,\ x') = \|x - x'\|_1 = \sum_i |x_i - x'_i|$

- Norme L_∞ : $d(x,\ x') = \|x - x'\|_\infty = \max_i |x_i - x'_i|$

DISTANCES ENTRE DEUX POINTS



Exemples

- Distance euclidienne ou L2 : $d(x,\ x') = \|x - x'\|_2 = \sqrt{\sum_i (x_i - x'_i)^2}$

- Distance de Manhattan ou L1 : $d(x,\ x') = \|x - x'\|_1 = \sum_i |x_i - x'_i|$

- Norme L_∞ : $d(x,\ x') = \|x - x'\|_\infty = \max_i |x_i - x'_i|$

Distance L2 au carrée

On utilise souvent la distance euclidienne au carrée (pouvez-vous expliquer pourquoi ?)

$$d(x,\ x') = \|x - x'\|_2^2 = \sum_i (x_i - x'_i)^2$$

DISTANCES ENTRE CLUSTERS

Comment définir une distance entre deux clusters ?

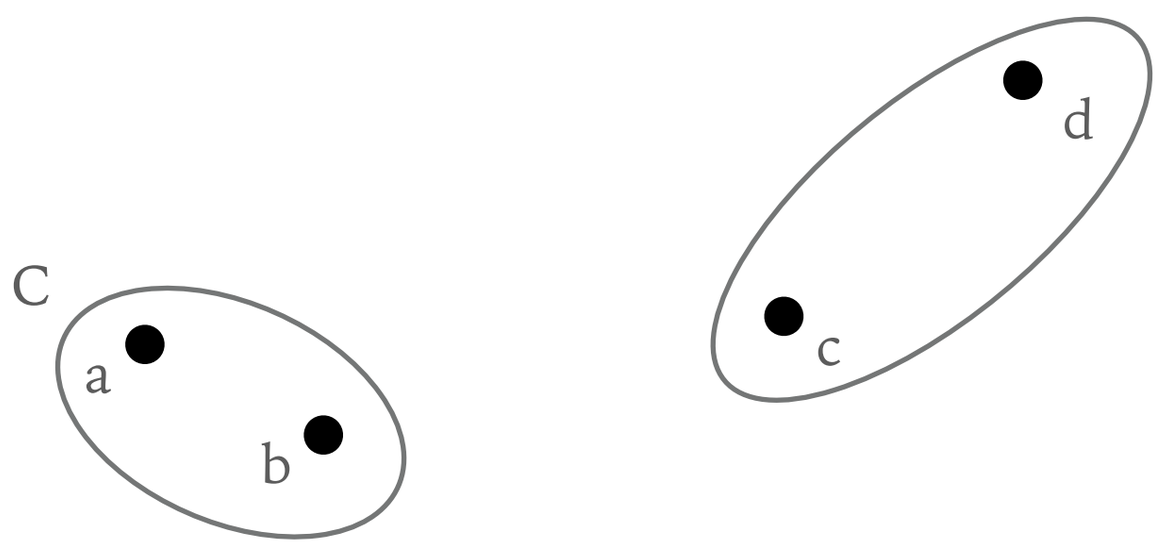
Il peut y avoir plusieurs points dans un cluster

- La distance doit-elle dépendre uniquement des poins les plus proches les un des autres ?
- Des points les plus éloignés ?
- De tous les points dans les deux clusters ?

Notation

Majuscule !

On écrira la distance entre les clusters C et C' comme D(C, C').

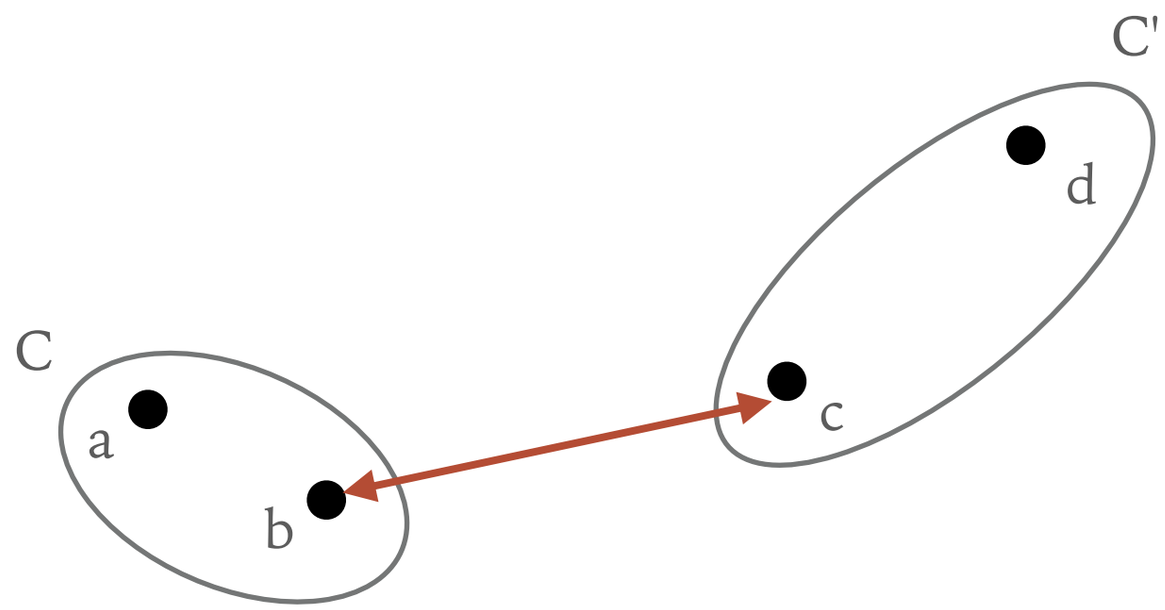


DISTANCES ENTRE CLUSTERS

Single linkage

La distance entre deux clusters est définie comme la distance **minimum** entre deux points respectifs de ces cluster

$$D_{min}(C, C') = \min_{\substack{x \in C, \\ x' \in C'}} d(x, x')$$

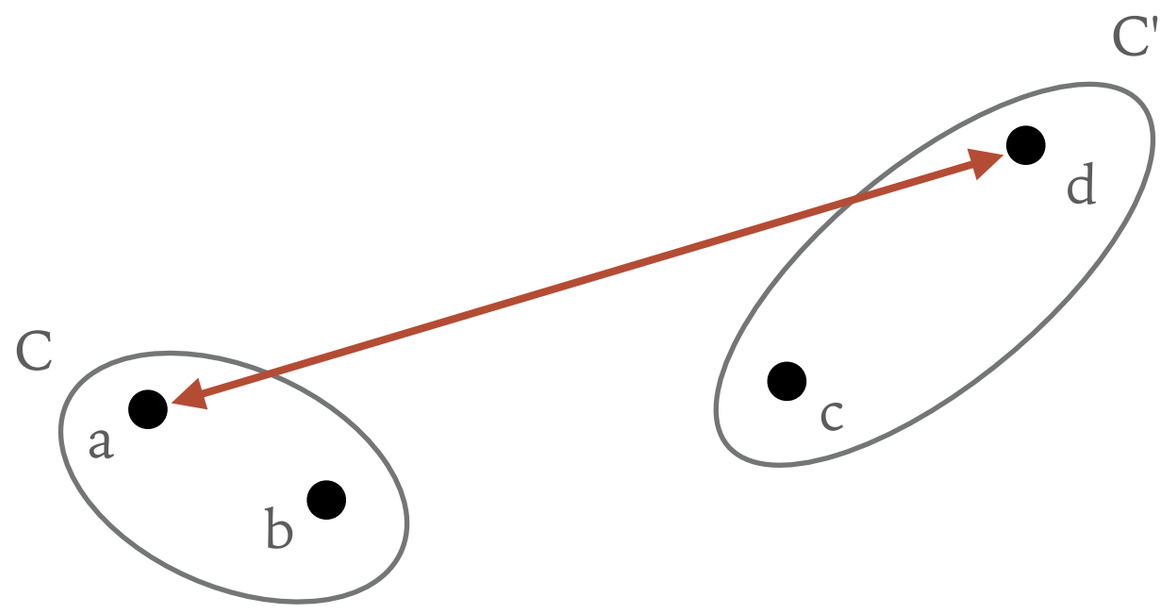


DISTANCES ENTRE CLUSTERS

Complete linkage

La distance entre deux clusters est définie comme la distance **maximum** entre deux points respectifs de ces cluster

$$D_{max}(C, C') = \max_{\substack{x \in C, \\ x' \in C'}} d(x, x')$$

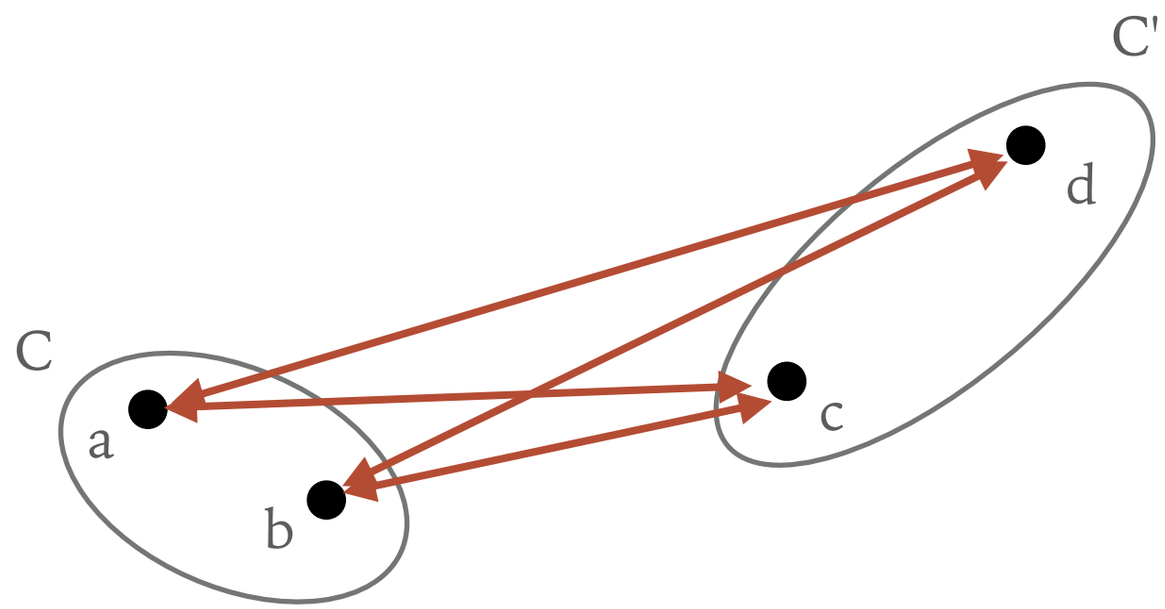


DISTANCES ENTRE CLUSTERS

Distance moyenne

La distance entre deux clusters est définie comme la distance **moyenne** entre deux points respectifs de ces cluster

$$D_{avg}(C, C') = \frac{1}{|C| \times |C'|} \sum_{x \in C} \sum_{x' \in C'} d(x, x')$$

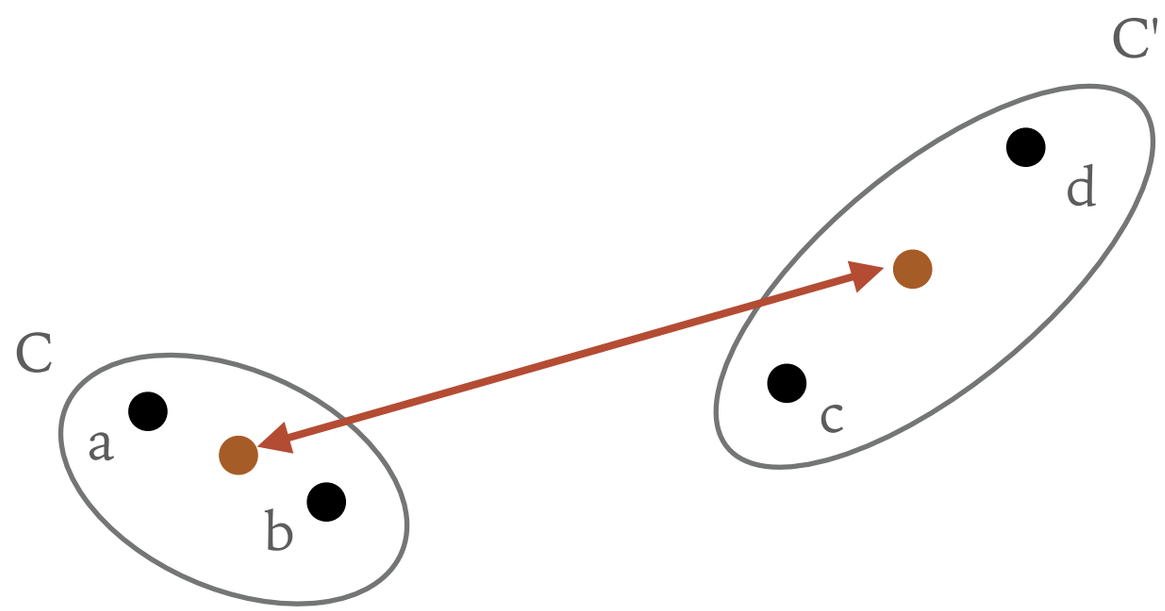


DISTANCES ENTRE CLUSTERS

Distance entre les centroïdes

La distance entre deux clusters est définie comme la distance entre leur centroides

$$D_{cent}(C, C') = d\left(\frac{1}{|C|}\sum_{x \in C} x,\ \frac{1}{|C'|}\sum_{x' \in C'} x'\right)$$



DISTANCES ENTRE CLUSTERS

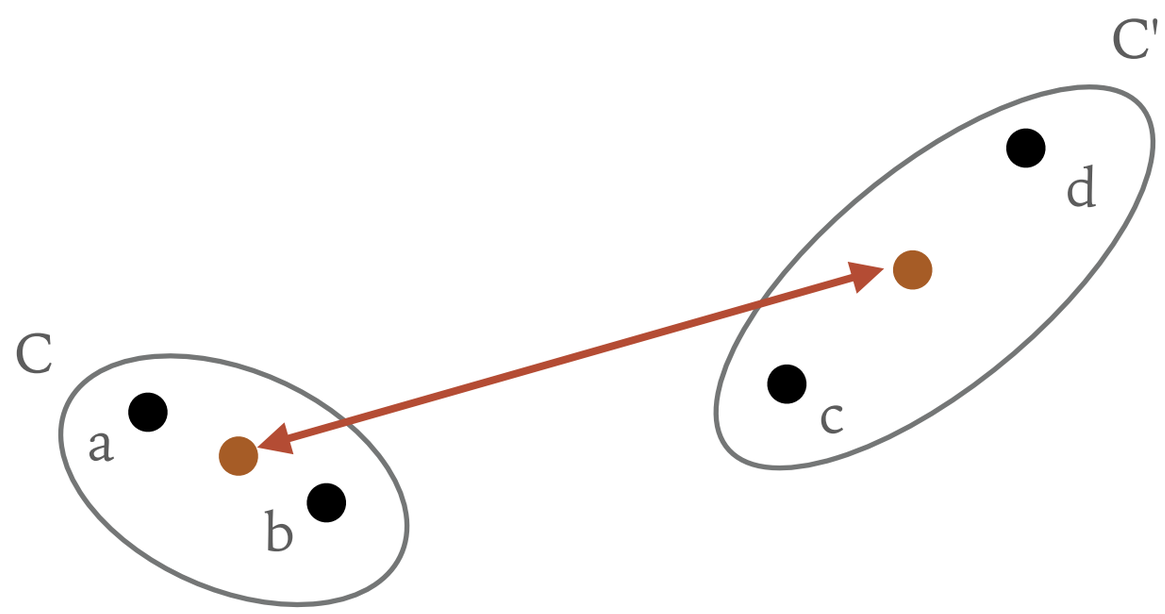
Distance de Ward

La distance entre deux clusters est définie comme la distance entre leur centroides mais :

- Il y a un terme de renormalisation supplémentaire
- La distance est la distance euclidienne au carré

(voir les exercices pour la motivation de cette formule)

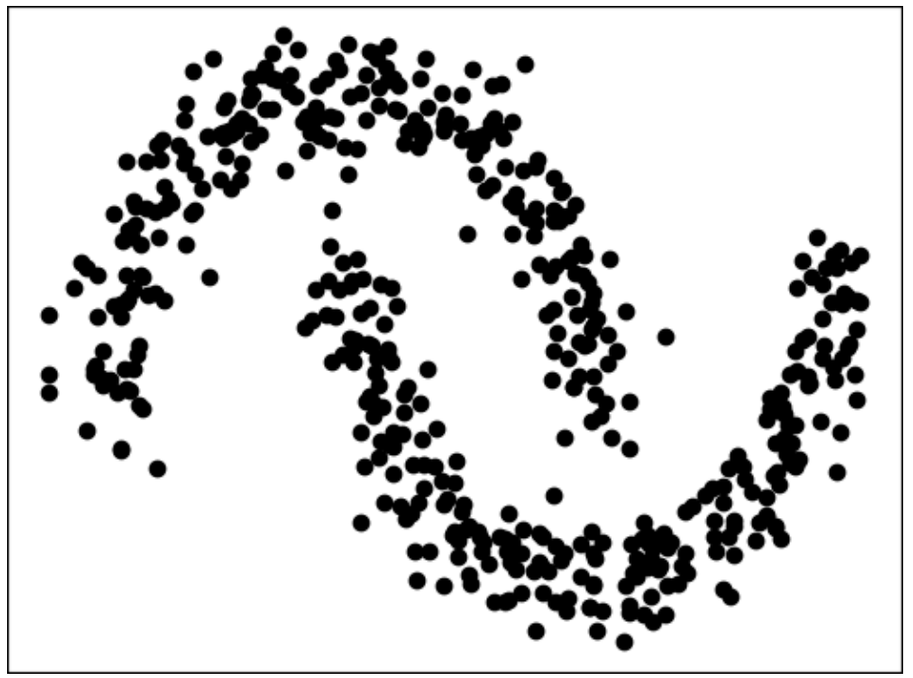
$$D_{ward}(C, C') = \frac{|C| \times |C'|}{|C| + |C'|} \times d\left(\frac{1}{|C|} \sum_{x \in C} x, \ \frac{1}{|C'|} \sum_{x' \in C'} x' \right)$$

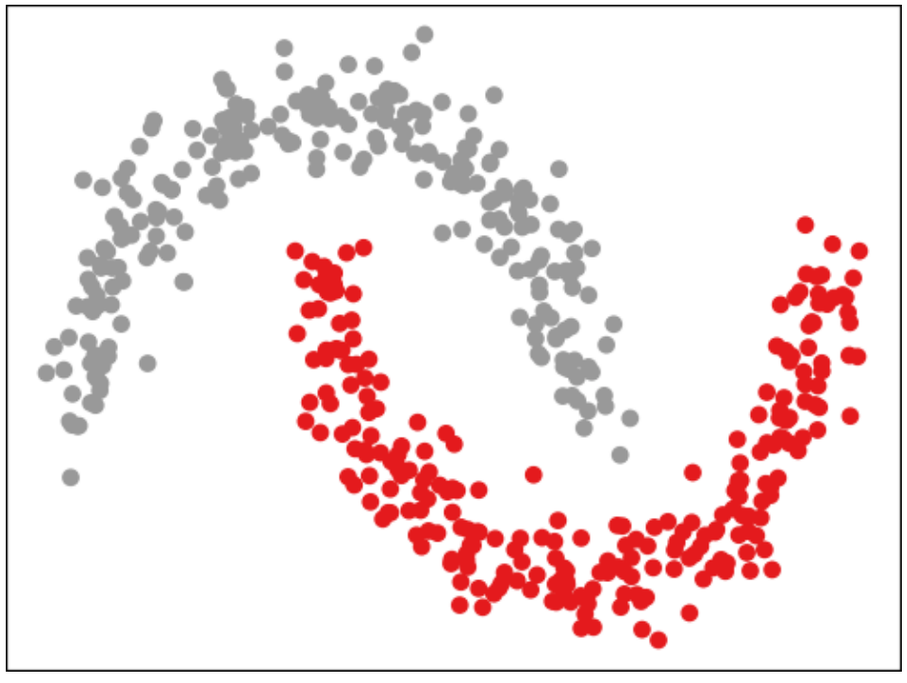


DIFFÉRENCE ENTRE LES DISTANCES

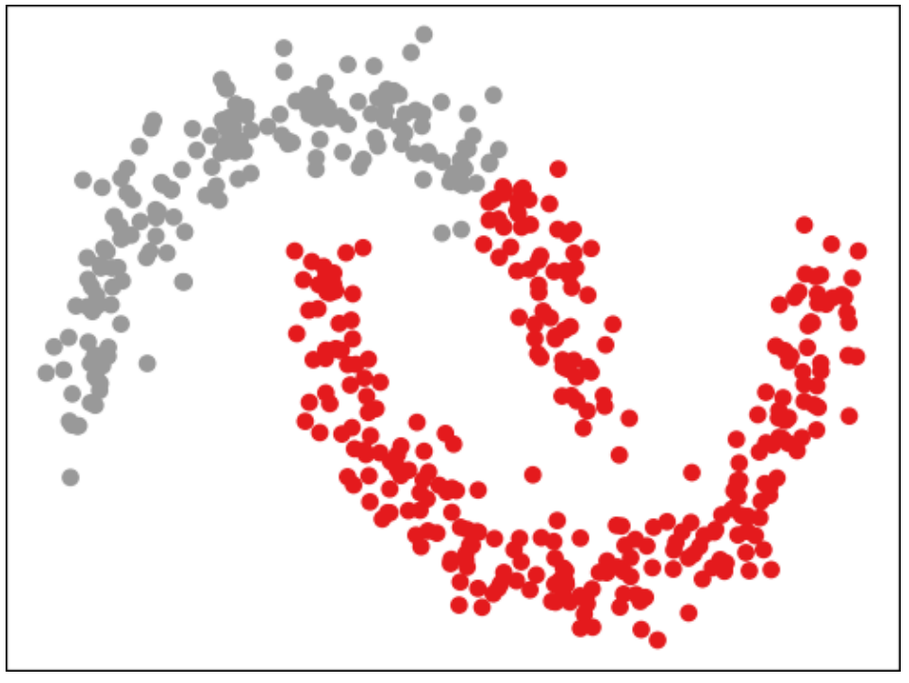
Différence en single et complete linkage

- Single : se focalise sur la cohérence locale, donc tendance à créer des clusters qui ressemblent à de longues chaînes
- Complete : se focalise sur la cohérence globale, la distance entre deux clusters est la distance entre leurs points respectifs les plus éloignés





(single) $D_{min}(C, C') = \min_{\substack{x \in C, \\ x' \in C'}} d(x, x')$

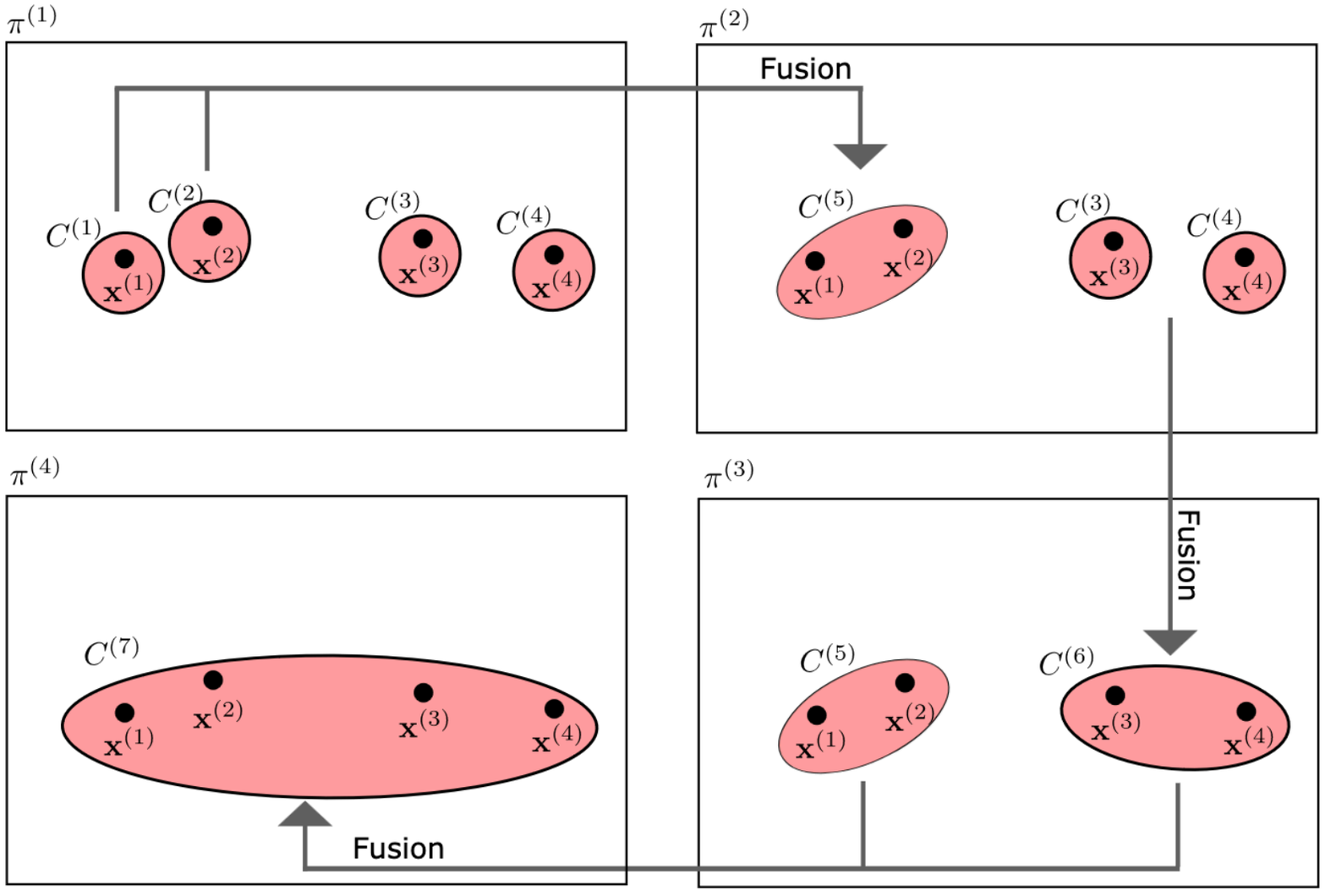


(complete) $D_{max}(C, C') = \max_{\substack{x \in C, \\ x' \in C'}} d(x, x')$

DENDROGRAMME

Objectif

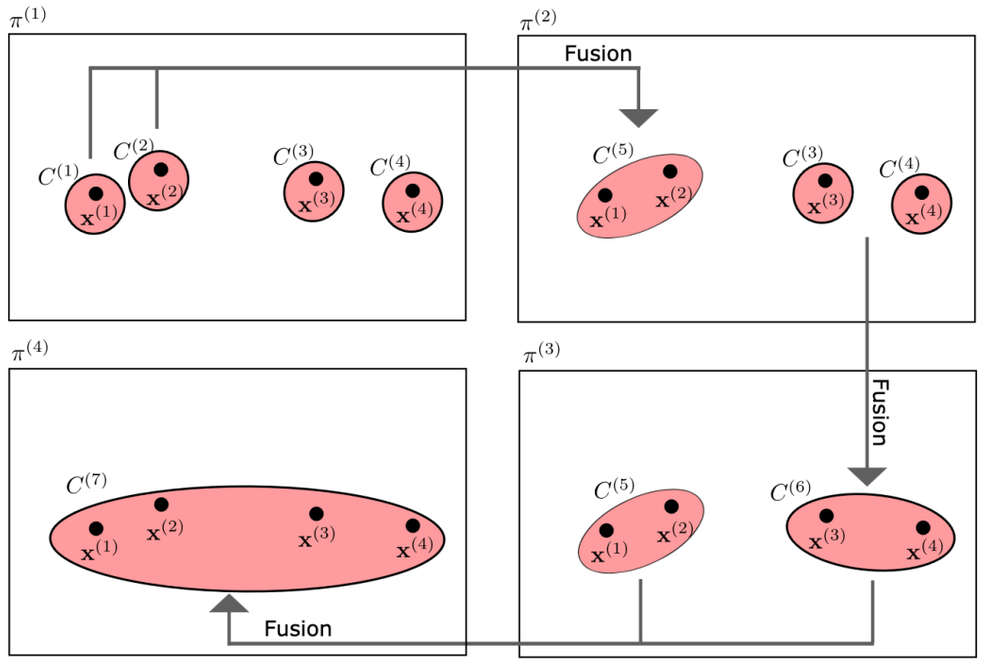
Résumer l'ensemble des fusions en une seule figure => un dendrogramme !



DENDROGRAMME

Structure d'un dendrogramme

- Illustre l'ordre des fusions
- La hauteur des fusions donne l'ordre dans lesquels elles sont réalisées

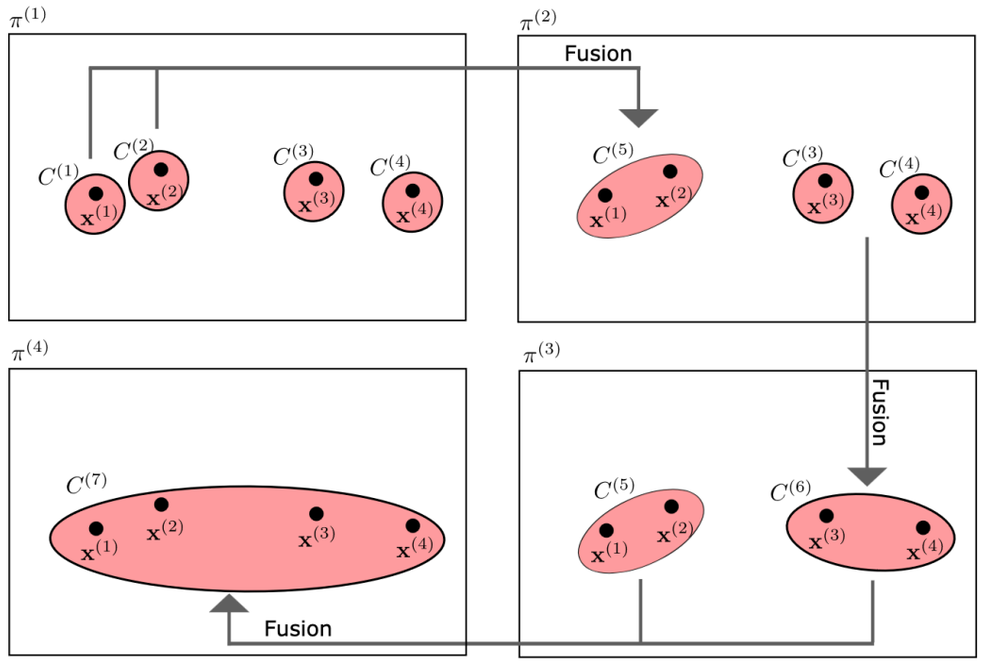


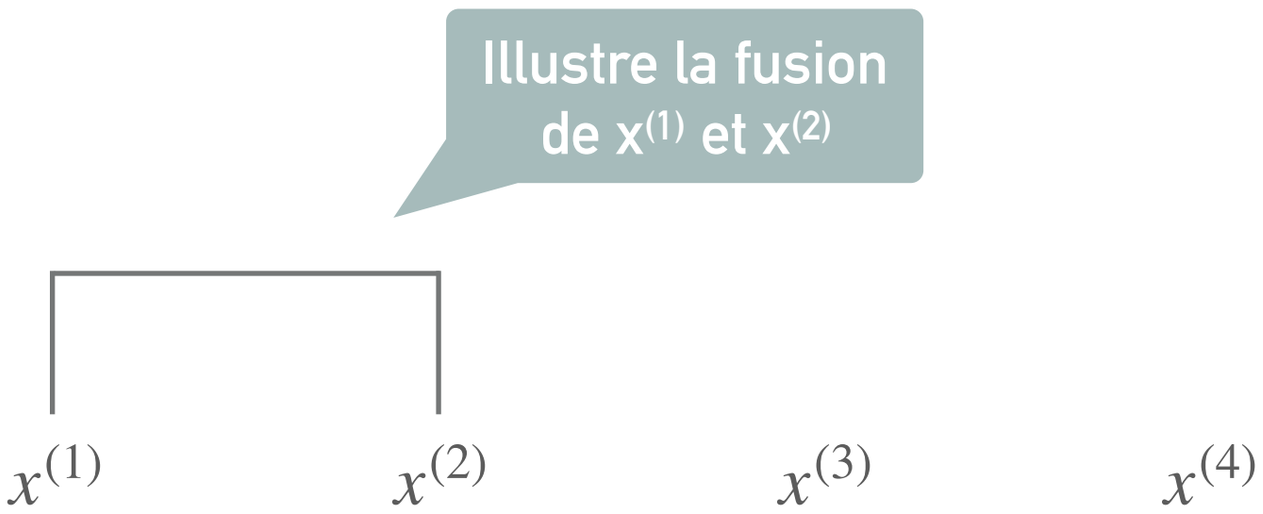
$x^{(1)}$ $x^{(2)}$ $x^{(3)}$ $x^{(4)}$

DENDROGRAMME

Structure d'un dendrogramme

- Illustre l'ordre des fusions
- La hauteur des fusions donne l'ordre dans lesquels elles sont réalisées

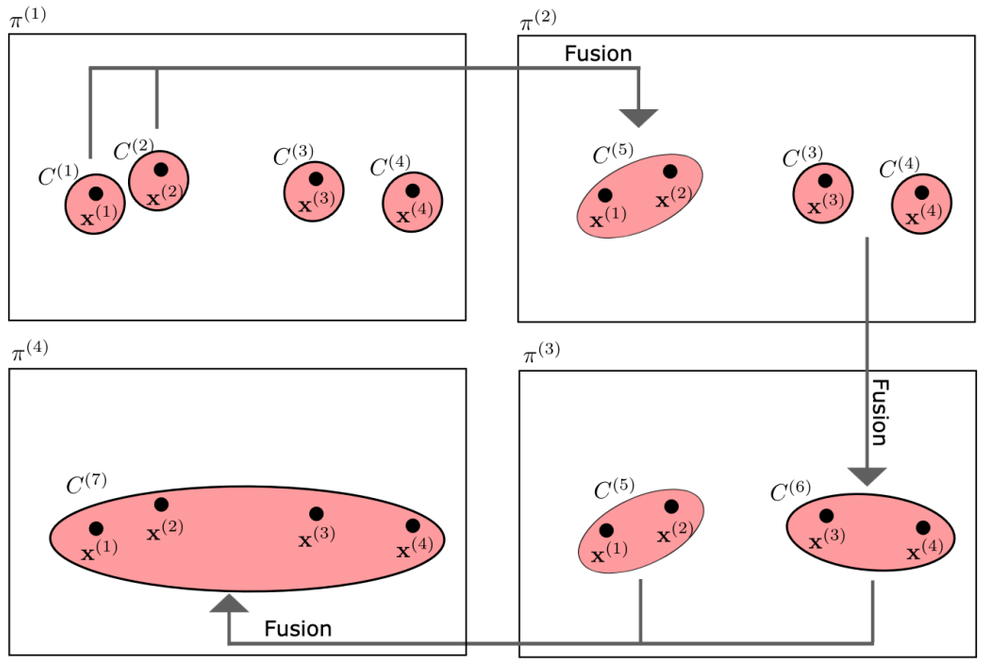


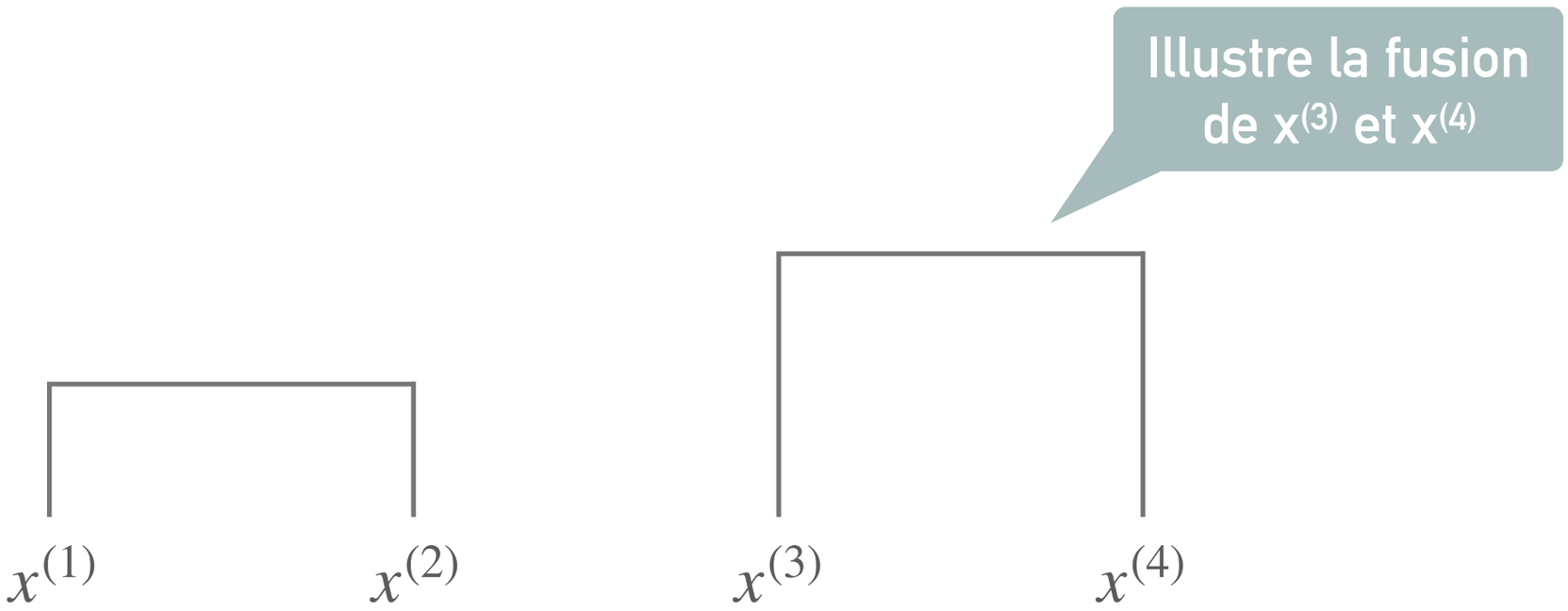


DENDROGRAMME

Structure d'un dendrogramme

- Illustre l'ordre des fusions
- La hauteur des fusions donne l'ordre dans lesquels elles sont réalisées

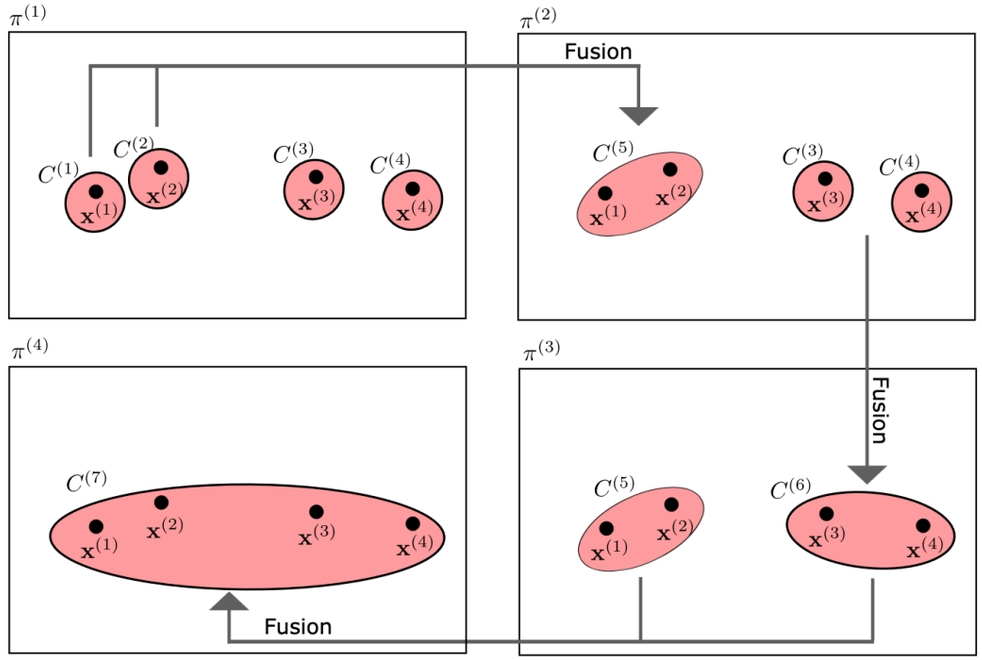




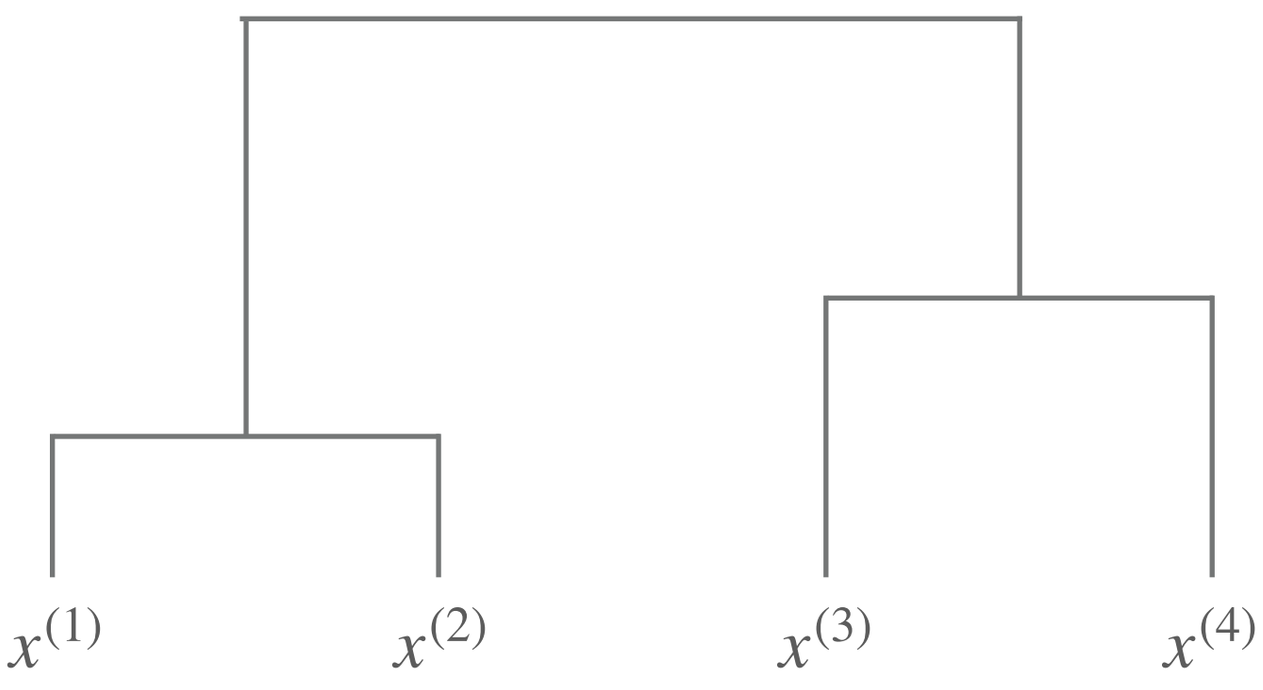
DENDROGRAMME

Structure d'un dendrogramme

- Illustre l'ordre des fusions
- La hauteur des fusions donne l'ordre dans lesquels elles sont réalisées



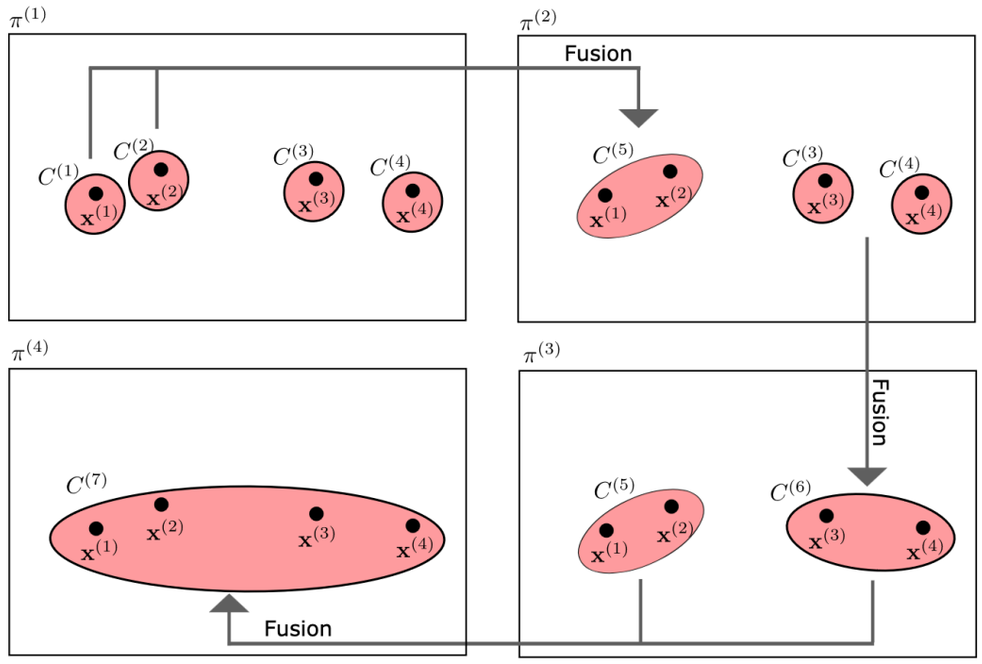
Illustre la fusion des deux clusters

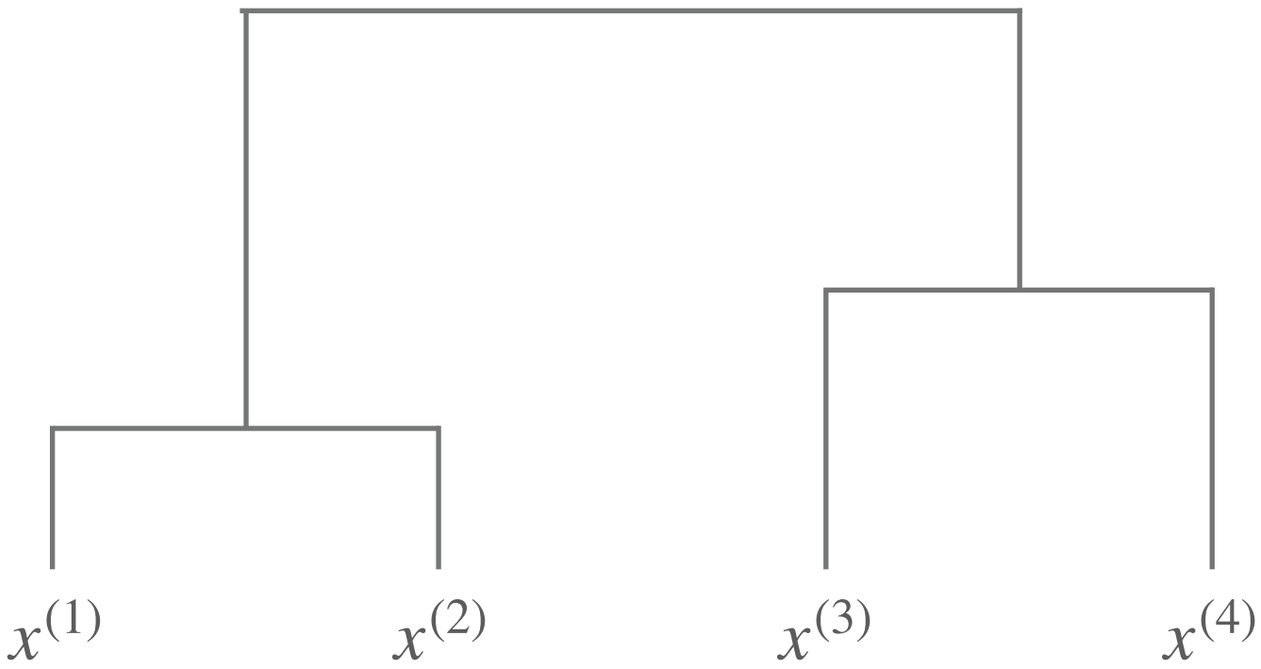


DENDROGRAMME

Structure d'un dendrogramme

- Illustre l'ordre des fusions
- La hauteur des fusions donne l'ordre dans lesquels elles sont réalisées





L'ordre des points est défini de telle sorte qu'il n'y ai pas de croisement dans le dendrogramme

PLONGEMENTS DE MOTS

PLONGEMENTS DE MOTS

Hypothèse distributionnelle

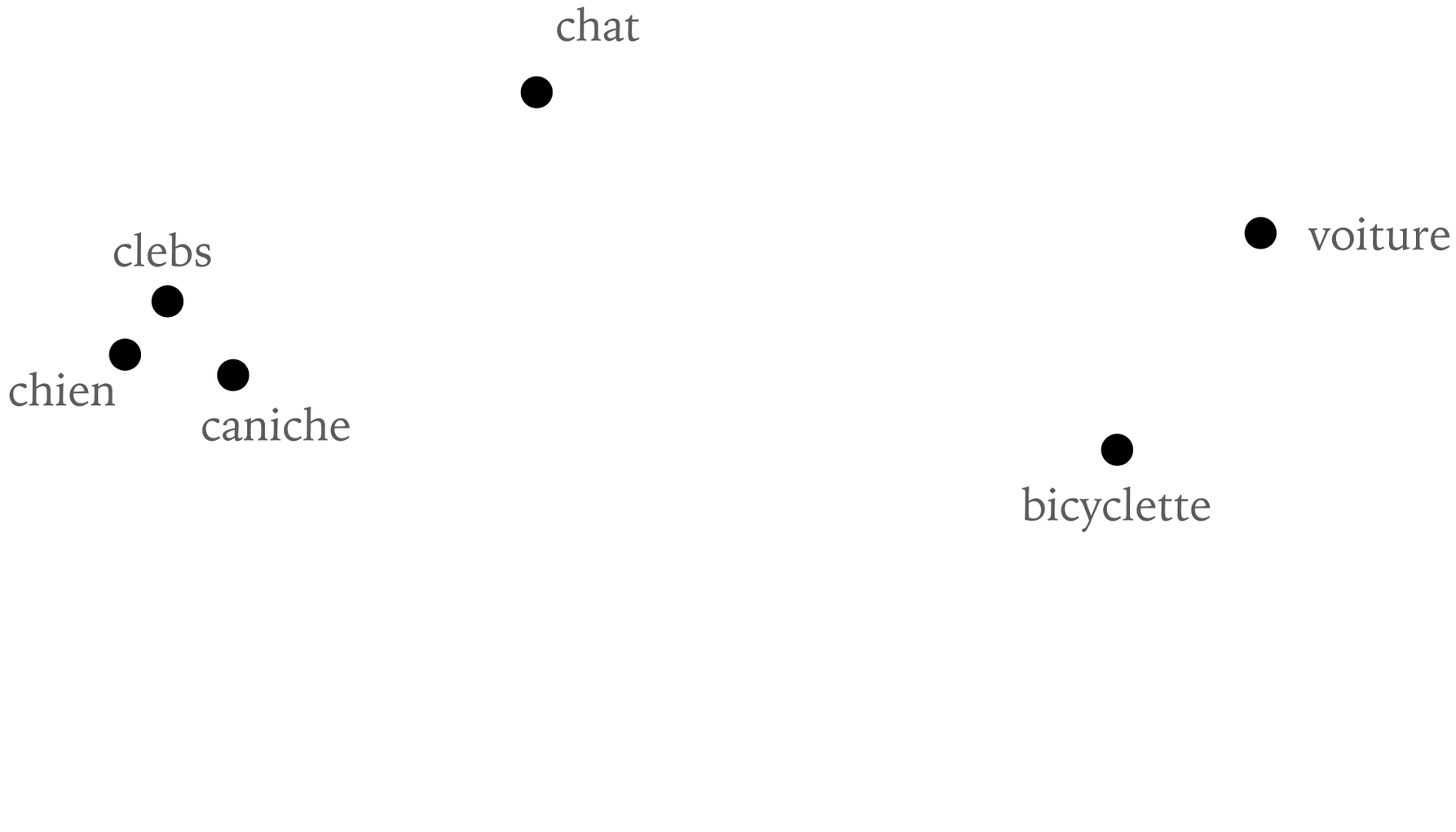
Le sens d'un mot peut être déduit du contexte dans lequel il est utilisé,
si deux mots sont synonymes, alors ils seront utilisés dans les mêmes contextes.

Approche géométrique du sens des mots

- Chaque mot est représenté par un vecteur
- La distance entre deux vecteurs mesure la distance sémantique entre deux mots (nulle ou quasi nulle pour des synonymes)
- Manipulation des vecteurs via des opérations algébriques

ILLUSTRATION

- Vecteurs en deux dimensions (juste pour l'illustration)
- Les mots ayant un sens "proche" sont aussi proche dans l'espace



CONSTRUCTION DES PLONGEMENTS DE MOTS

Table de co-occurences

- Une ligne par mot dans le vocabulaire (en Français)
- Une colonne par mot dans le vocabulaire (en Français)

Comment remplir le tableau ?

1. On initialise toutes les cellules du tableau à 0
2. On parcours toutes les phrases dans un grand corpus de text
   - Pour chaque couple de mots dans la phrase, on incrément la cellule correspondante
   - En général on ne regarde qu'une "fenêtre" autour d'un mot

	le	la	chien	mange	...
le	0	0	0	0	...
la	0	0	0	0	...
chien	0	0	0	0	...
mange	0	0	0	0	...
etc etc	...	...	..	..	..

CONSTRUCTION DES PLONGEMENTS DE MOTS

	le	la	chien	mange	dort	petit	dehors	intérieur	...
le	0	0	0	0	0	0	0	0	...
la	0	0	0	0	0	0	0	0	...
chien	0	0	0	0	0	0	0	0	...
mange	0	0	0	0	0	0	0	0	...
dort	0	0	0	0	0	0	0	0	...
petit	0	0	0	0	0	0	0	0	...
dehors	0	0	0	0	0	0	0	0	...
intérieu	0	0	0	0	0	0	0	0	...
etc etc	...	...	..	..	..	..	..	..	..

*Le petit **chien** dort dehors*

On se focalise sur le mot dort

CONSTRUCTION DES PLONGEMENTS DE MOTS

	le	la	chien	mange	dort	petit	dehors	intérieur	...
le	0	0	0	0	0	0	0	0	...
la	0	0	0	0	0	0	0	0	...
chien	0	0	0	0	0	0	0	0	...
mange	0	0	0	0	0	0	0	0	...
dort	0	0	0	0	0	0	0	0	...
petit	0	0	0	0	0	0	0	0	...
dehors	0	0	0	0	0	0	0	0	...
intérieu	0	0	0	0	0	0	0	0	...
etc etc	...	...	..	..	..	..	..	..	..

Fenêtre de taille 2

*Le petit **chien** dort dehors*

On se focalise sur le mot dort

CONSTRUCTION DES PLONGEMENTS DE MOTS

	le	la	chien	mange	dort	petit	dehors	intérieur	...
le	0	0	0	0	0	0	0	0	...
la	0	0	0	0	0	0	0	0	...
chien	+1	0	0	0	0	0	0	0	...
mange	0	0	0	0	0	0	0	0	...
dort	0	0	0	0	0	0	0	0	...
petit	0	0	0	0	0	0	0	0	...
dehors	0	0	0	0	0	0	0	0	...
intérieu	0	0	0	0	0	0	0	0	...
etc etc	...	...	..	..	..	..	..	..	..

Fenêtre de taille 2

*Le petit **chien** dort dehors*

On se focalise sur le mot dort

CONSTRUCTION DES PLONGEMENTS DE MOTS

	le	la	chien	mange	dort	petit	dehors	intérieur	…
le	0	0	0	0	0	0	0	0	…
la	0	0	0	0	0	0	0	0	…
chien	+1	0	0	0	0	+1	0	0	…
mange	0	0	0	0	0	0	0	0	…
dort	0	0	0	0	0	0	0	0	…
petit	0	0	0	0	0	0	0	0	…
dehors	0	0	0	0	0	0	0	0	…
intérieu	0	0	0	0	0	0	0	0	…
etc etc	…	…	..	..	..	..	..	..	..

Fenêtre de taille 2

*Le petit **chien** dort dehors*

On se focalise sur le mot dort

CONSTRUCTION DES PLONGEMENTS DE MOTS

	le	la	chien	mange	dort	petit	dehors	intérieur	…
le	0	0	0	0	0	0	0	0	…
la	0	0	0	0	0	0	0	0	…
chien	+1	0	0	0	+1	+1	0	0	…
mange	0	0	0	0	0	0	0	0	…
dort	0	0	0	0	0	0	0	0	…
petit	0	0	0	0	0	0	0	0	…
dehors	0	0	0	0	0	0	0	0	…
intérieu	0	0	0	0	0	0	0	0	…
etc etc	…	…	..	..	..	..	..	..	..

Fenêtre de taille 2

*Le petit **chien** dort dehors*

On se focalise sur le mot dort

CONSTRUCTION DES PLONGEMENTS DE MOTS

	le	la	chien	mange	dort	petit	dehors	intérieur	...
le	0	0	0	0	0	0	0	0	...
la	0	0	0	0	0	0	0	0	...
chien	+1	0	0	0	+1	+1	0	0	...
mange	0	0	0	0	0	0	0	0	...
dort	0	0	0	0	0	0	0	0	...
petit	0	0	0	0	0	0	0	0	...
dehors	0	0	0	0	0	0	0	0	...
intérieu	0	0	0	0	0	0	0	0	...
etc etc	...	...	..	..	..	..	..	..	..

Fenêtre de taille 2

*Le petit **chien** dort dehors*

On se focalise sur le mot dort

CONSTRUCTION DES PLONGEMENTS DE MOTS

	le	la	chien	mange	dort	petit	dehors	intérieur	...
le	0	0	0	0	0	0	0	0	...
la	0	0	0	0	0	0	0	0	...
chien	1	0	0	0	1	1	0	0	...
mange	0	0	0	0	0	0	0	0	...
dort	0	0	0	0	0	0	0	0	...
petit	0	0	0	0	0	0	0	0	...
dehors	0	0	0	0	0	0	0	0	...
intérieu	0	0	0	0	0	0	0	0	...
etc etc	...	...	..	..	..	..	..	..	..

*Le petit chien **dort** dehors*

On se focalise sur le mot dort

CONSTRUCTION DES PLONGEMENTS DE MOTS

	le	la	chien	mange	dort	petit	dehors	intérieur	...
le	0	0	0	0	0	0	0	0	...
la	0	0	0	0	0	0	0	0	...
chien	1	0	0	0	1	1	0	0	...
mange	0	0	0	0	0	0	0	0	...
dort	0	0	0	0	0	0	0	0	...
petit	0	0	0	0	0	0	0	0	...
dehors	0	0	0	0	0	0	0	0	...
intérieu	0	0	0	0	0	0	0	0	...
etc etc	...	...	..	..	..	..	..	..	..

*Le petit chien **dort** dehors*

Fenêtre de taille 2

On se focalise sur le mot dort

CONSTRUCTION DES PLONGEMENTS DE MOTS

	le	la	chien	mange	dort	petit	dehors	intérieur	...
le	0	0	0	0	0	0	0	0	...
la	0	0	0	0	0	0	0	0	...
chien	1	0	0	0	1	1	0	0	...
mange	0	0	0	0	0	0	0	0	...
dort	0	0	+1	0	0	0	0	0	...
petit	0	0	0	0	0	0	0	0	...
dehors	0	0	0	0	0	0	0	0	...
intérieu	0	0	0	0	0	0	0	0	...
etc etc	...	...	..	..	..	..	..	..	..

*Le petit chien **dort** dehors*

Fenêtre de taille 2

On se focalise sur le mot dort

CONSTRUCTION DES PLONGEMENTS DE MOTS

	le	la	chien	mange	dort	petit	dehors	intérieur	...
le	0	0	0	0	0	0	0	0	...
la	0	0	0	0	0	0	0	0	...
chien	1	0	0	0	1	1	0	0	...
mange	0	0	0	0	0	0	0	0	...
dort	0	0	+1	0	0	+1	0	0	...
petit	0	0	0	0	0	0	0	0	...
dehors	0	0	0	0	0	0	0	0	...
intérieu	0	0	0	0	0	0	0	0	...
etc etc	...	...	..	..	..	..	..	..	..

*Le petit chien **dort** dehors*

Fenêtre de taille 2

On se focalise sur le mot dort

CONSTRUCTION DES PLONGEMENTS DE MOTS

	le	la	chien	mange	dort	petit	dehors	intérieur	…
le	0	0	0	0	0	0	0	0	…
la	0	0	0	0	0	0	0	0	…
chien	1	0	0	0	1	1	0	0	…
mange	0	0	0	0	0	0	0	0	…
dort	0	0	+1	0	0	+1	+1	0	…
petit	0	0	0	0	0	0	0	0	…
dehors	0	0	0	0	0	0	0	0	…
intérieu	0	0	0	0	0	0	0	0	…
etc etc	…	…	..	..	..	..	..	..	..

*Le petit chien **dort** dehors*

Fenêtre de taille 2

On se focalise sur le mot dort

CONSTRUCTION DES PLONGEMENTS DE MOTS

	le	la	chien	mange	dort	petit	dehors	intérieur	…
le	0	0	0	0	0	0	0	0	…
la	0	0	0	0	0	0	0	0	…
chien	1	0	0	0	1	1	0	0	…
mange	0	0	0	0	0	0	0	0	…
dort	0	0	+1	0	0	+1	+1	0	…
petit	0	0	0	0	0	0	0	0	…
dehors	0	0	0	0	0	0	0	0	…
intérieu	0	0	0	0	0	0	0	0	…
etc etc	…	…	..	..	..	..	..	..	..

*Le petit chien **dort** dehors*

Fenêtre de taille 2

On se focalise sur le mot dort

CONSTRUCTION DES PLONGEMENTS DE MOTS

	le	la	chien	mange	dort	petit	dehors	intérieur	…
le	0	0	0	0	0	0	0	0	…
la	0	0	0	0	0	0	0	0	…
chien	1	0	0	0	1	1	0	0	…
mange	0	0	0	0	0	0	0	0	…
dort	0	0	+1	0	0	+1	+1	0	…
petit	0	0	0	0	0	0	0	0	…
dehors	0	0	0	0	0	0	0	0	…
intérieu	0	0	0	0	0	0	0	0	…
etc etc	…	…	..	..	..	..	..	..	..

*Le petit chien **dort** dehors*

Fenêtre de taille 2

On se focalise sur le mot dort

Et ainsi de suite pour tous les mots du corpus….

PLONGEMENTS DE MOTS

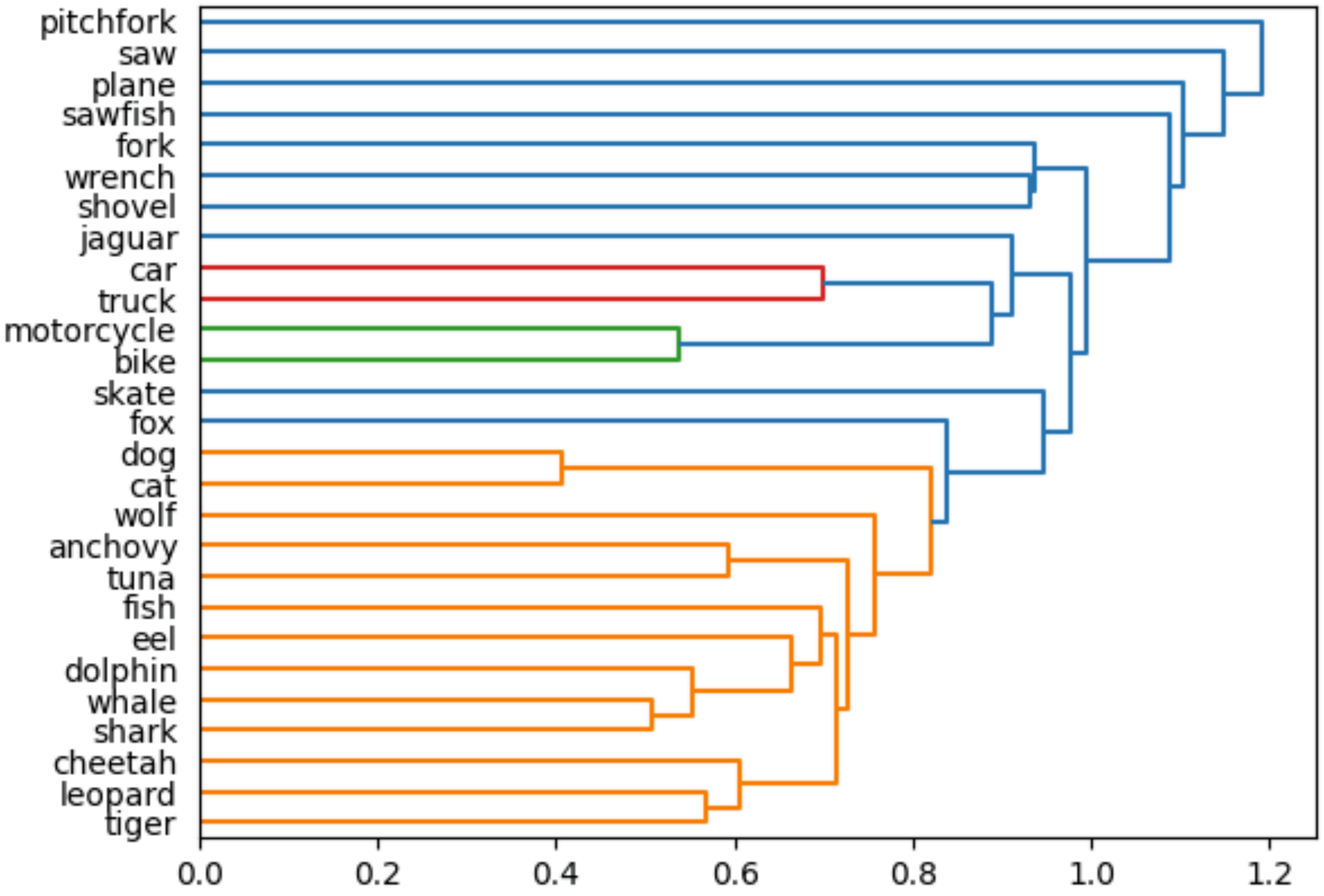
Traitement des tables de co-occurences

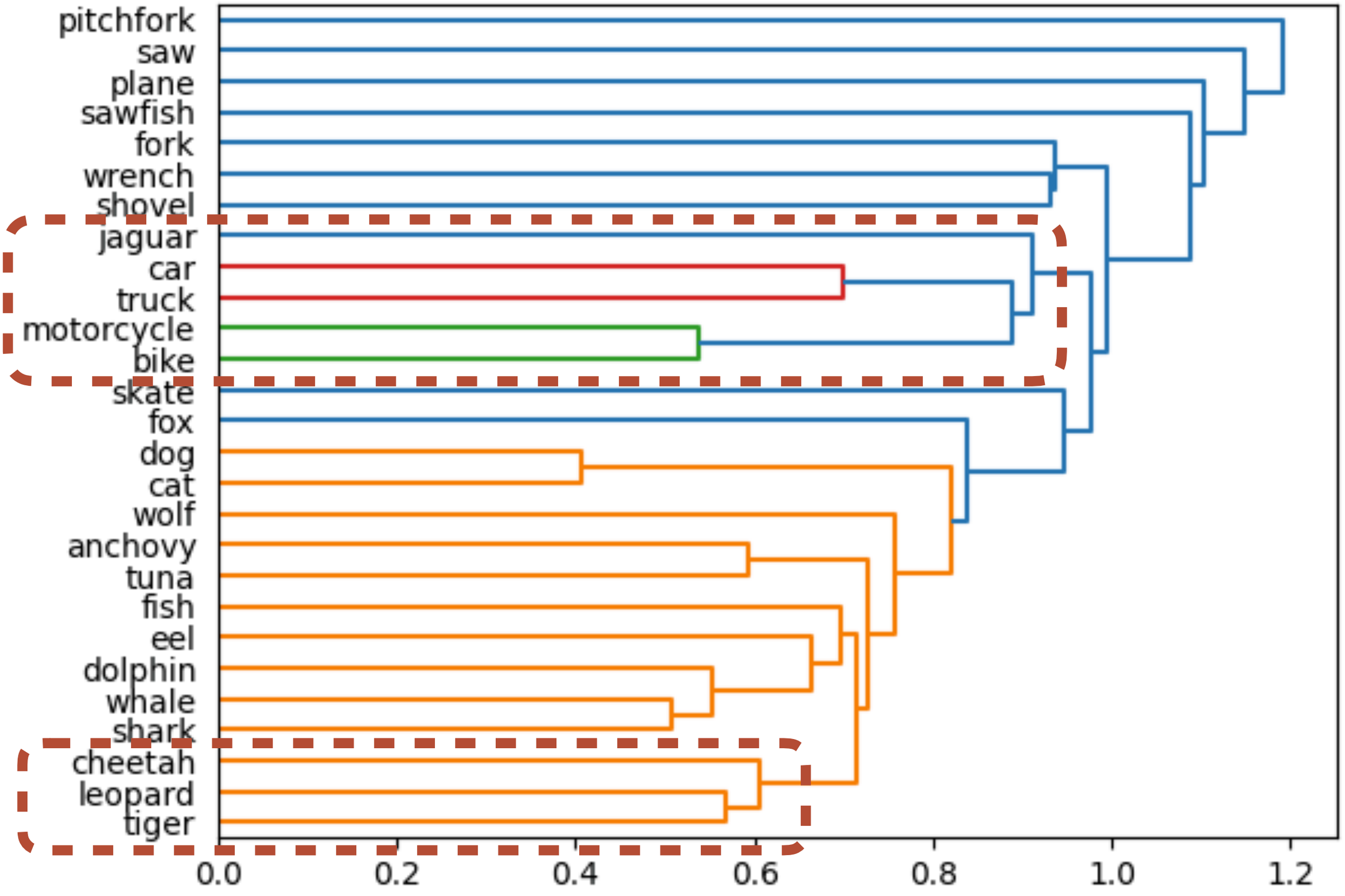
Les tables de co-occurences sont de très grandes dimensions (plusieurs centaines de milliers de lignes et de colonnes), et souvent très "bruitées

- Opération pour "nettoyer" le contenu,
  par exemple via (Positive) Pointwise Mutual Information (supprime des informations non pertinentes liée au fait que certains mots sont très fréquents dans le corpus, p. ex. les mots outils)
- Réduction de dimensions (voir chapitre précédent)

TP

Que donne le clustering de plongements de mots ? En fonction des distances utilisées ?


pitchfork
saw
plane
sawfish
fork
wrench
shovel
jaguar
car
truck
motorcycle
bike
skate
fox
dog
cat
wolf
anchovy
tuna
fish
eel
dolphin
whale
shark
cheetah
leopard
tiger
0.0
0.2
0.4
0.6
0.8
1.0
1.2


pitchfork
saw
plane
sawfish
fork
wrench
shovel
jaguar
car
truck
motorcycle
bike
skate
fox
dog
cat
wolf
anchovy
tuna
fish
eel
dolphin
whale
shark
cheetah
leopard
tiger
0.0
0.2
0.4
0.6
0.8
1.0
1.2