LIRIS

UMR 5205 CNRS

Apprentissage non supervisé

Haytham Elghazel

Laboratoire d'InfoRmatique en Image et Systèmes d'information

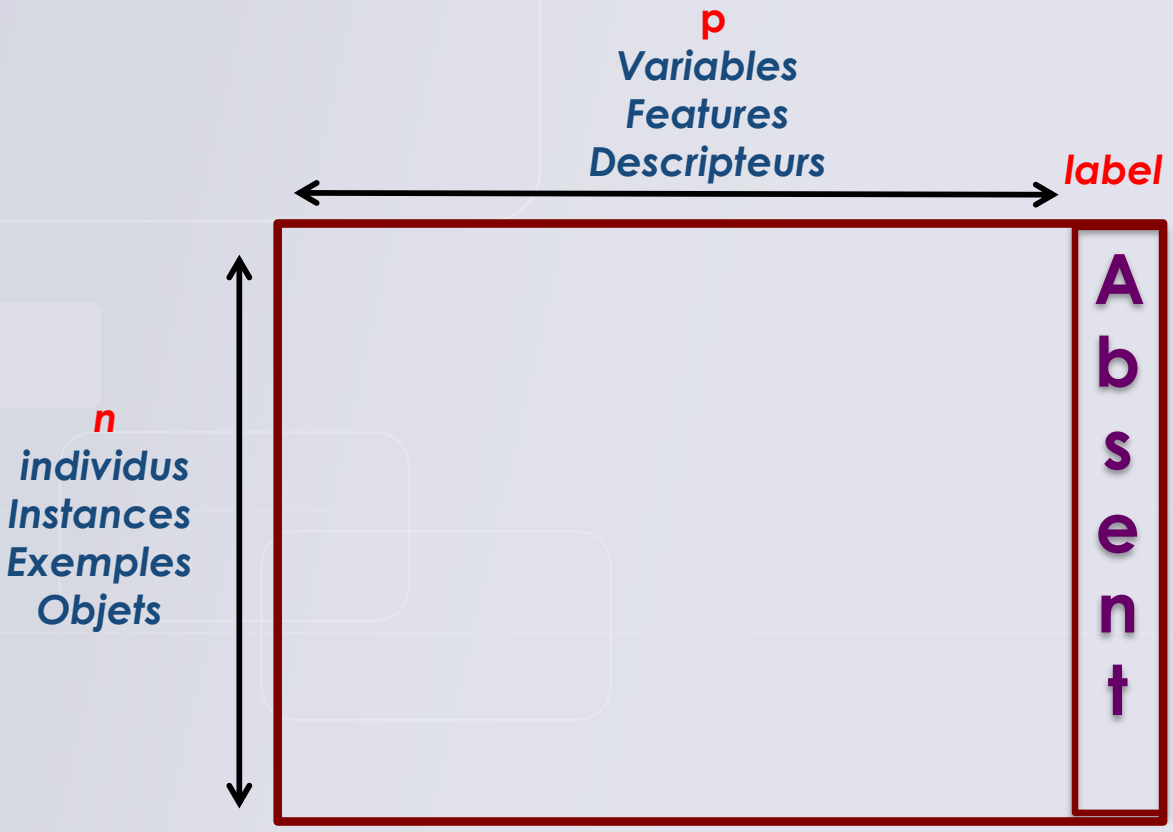
Pôle Data Science, Equipe DM2L

cnrs INSA Lyon 1 UNIVERSITÉ LUMIÈRE LYON 2 ÉCOLE CENTRALE LYON

Contexte : Apprentissage non supervisé

Consiste à inférer des connaissances sur les données

- Sur la seule base des échantillons d'apprentissage
- **Pas de cible** (pas le temps ni l'argent, pas de spécialistes), **recherche de structures naturelles dans les données**



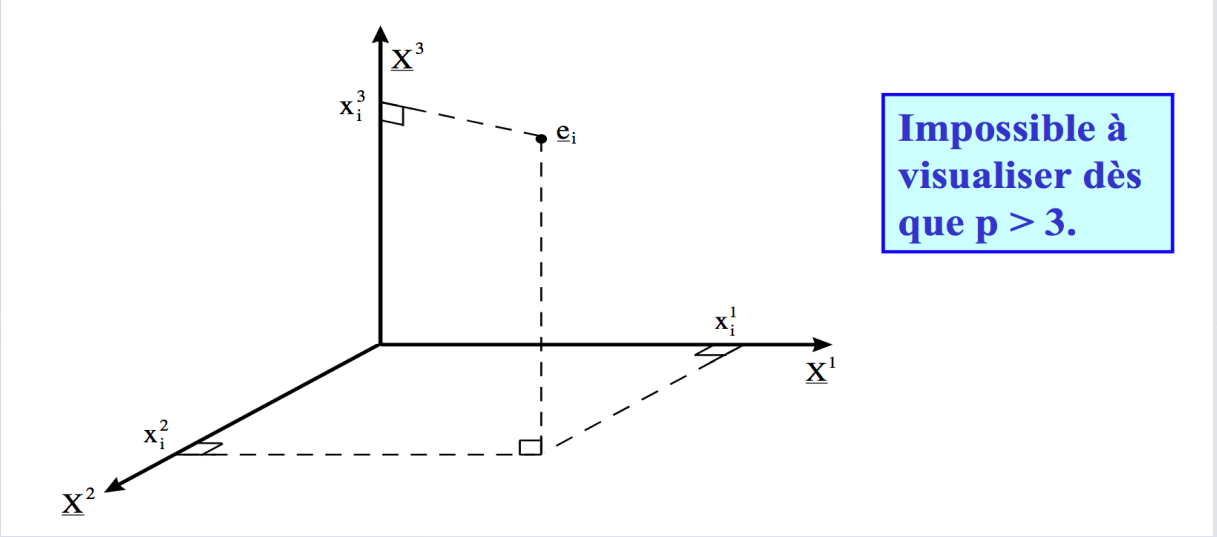
Contexte : Apprentissage non supervisé

Différentes tâches sont associées à l'apprentissage non supervisé

- **Réduction de dimensions** : une nouvelle projection de des données dans un nouvel espace de représentation réduit qui permet de synthétiser l'information
- **Custering** (segmentation, regroupement) : construire des classes automatiquement en fonction des exemples disponibles

Problématique : Réduction de dimensions

- **A chaque individu, on peut associer un point dans R^p.**
- **A chaque variable de la base est associé un axe de R^p**
- **Pourquoi veut-on réduite les dimensions ?**
  - Pour synthétiser l'information
  - Analyser les proximités entre les individus.
  - Une visualisation des données (le nuage des individus)



Impossible à visualiser dès que p > 3.

Réduction de dimensions : Exemple

$j : 1, \ldots, p$

$i : 1, \ldots, n$
Individus actifs

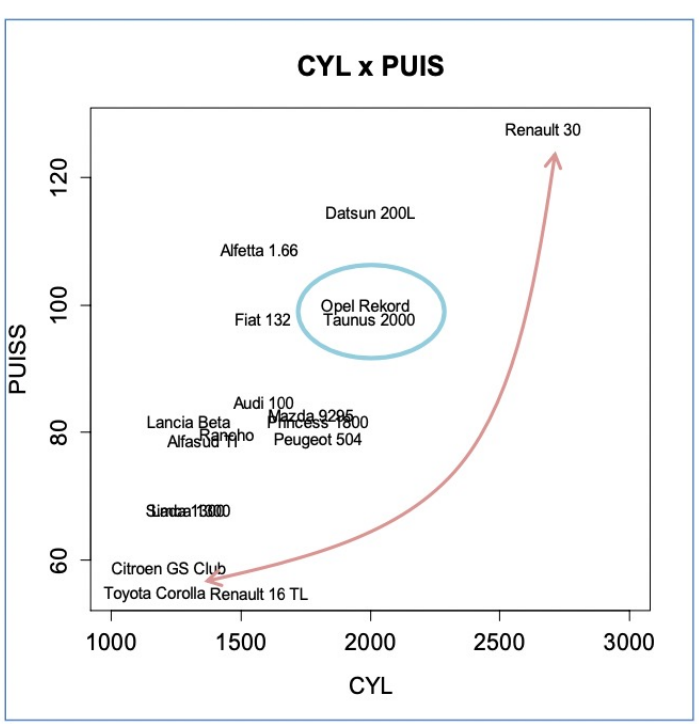
Modele	CYL	PUISS	LONG	LARG	POIDS	V.MAX
Alfasud TI	1350	79	393	161	870	165
Audi 100	1588	85	468	177	1110	160
Simca 1300	1294	68	424	168	1050	152
Citroen GS Club	1222	59	412	161	930	151
Fiat 132	1585	98	439	164	1105	165
Lancia Beta	1297	82	429	169	1080	160
Peugeot 504	1796	79	449	169	1160	154
Renault 16 TL	1565	55	424	163	1010	140
Renault 30	2664	128	452	173	1320	180
Toyota Corolla	1166	55	399	157	815	140
Alfetta 1.66	1570	109	428	162	1060	175
Princess 1800	1798	82	445	172	1160	158
Datsun 200L	1998	115	469	169	1370	160
Taunus 2000	1993	98	438	170	1080	167
Rancho	1442	80	431	166	1129	144
Mazda 9295	1769	83	440	165	1095	165
Opel Rekord	1979	100	459	173	1120	173
Lada 1300	1294	68	404	161	955	140

x_{ij}

- Quelles sont les véhicules qui se ressemblent ?
- Sur quelles variables sont fondées les similarités / différences ?
- Quelles sont les relations entre les variables ?

Réduction de dimensions : Exemple

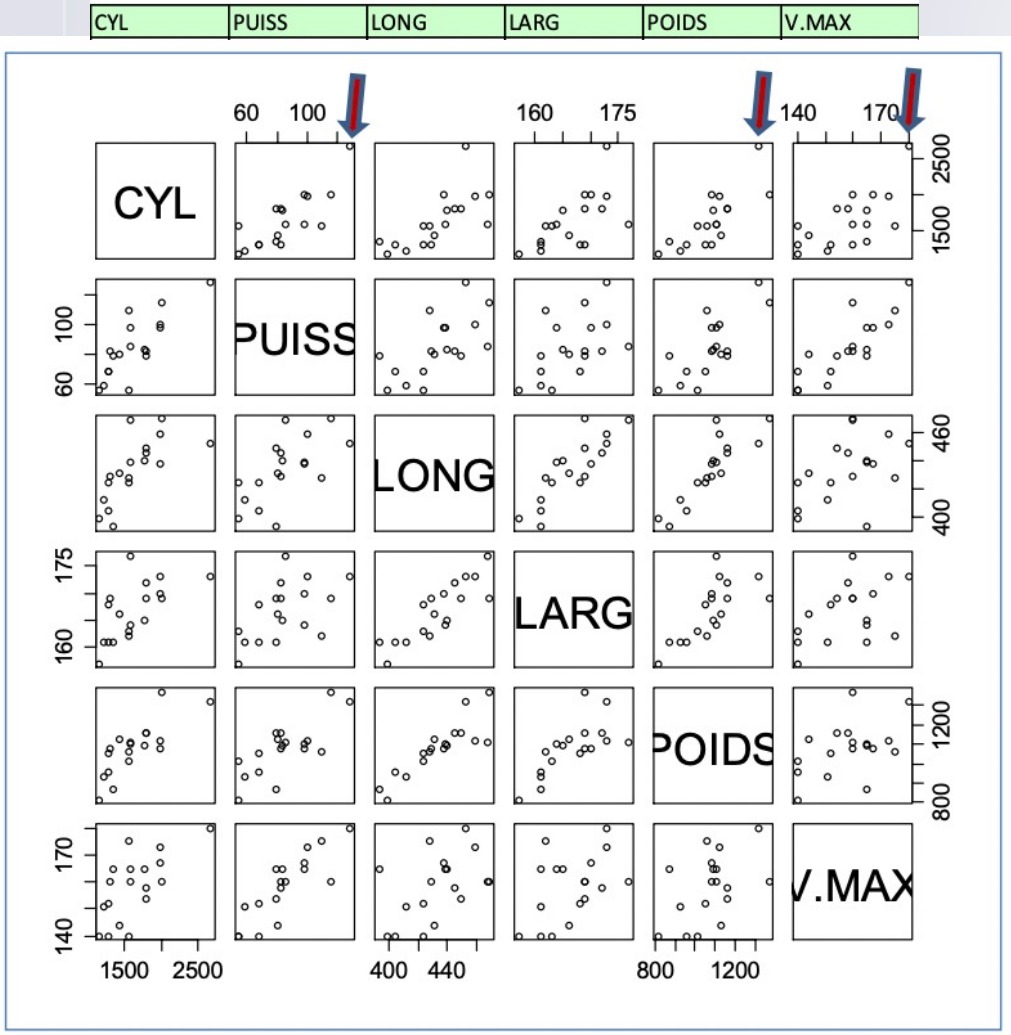
- **Si on prend en compte deux variables CYL et PUISS**
- **Ces deux variables sont liées (corrélées)**
- **On remarque que :**
  - l'Opel Reckord et la Ford Taunus 2000 ont le même profil
  - la Renault 30 et la Toyota Corolla ont des profils opposés



Modele	CYL	PUISS
Toyota Corolla	**1166**	**55**
Citroen GS Club	1222	59
Simca 1300	1294	68
Lada 1300	1294	68
Lancia Beta	1297	82
Alfasud TI	1350	79
Rancho	1442	80
Renault 16 TL	1565	55
Alfetta 1.66	1570	109
Fiat 132	1585	98
Audi 100	1588	85
Mazda 9295	1769	83
Peugeot 504	1796	79
Princess 1800	1798	82
Opel Rekord	**1979**	**100**
Taunus 2000	**1993**	**98**
Datsun 200L	1998	115
Renault 30	**2664**	**128**

Réduction de dimensions : Exemple

- **Si on prend en compte la totalité des 6 variables**
  - Très difficile de surveiller différents cadrans en même temps.
  - Peut être utilisé pour déceler des points atypiques (outliers, anomalies, ...)
    - **Exemple** : la **Renault 30**



Réduction de dimensions : Exemple

- **Solution :**
  - Construire un système de représentation de dimension réduite (q << p) qui préserve les distances entre les individus.

$$d^2(i,i') = \sum_{j=1}^{p} \left(x_{ij} - x_{i'j}\right)^2$$

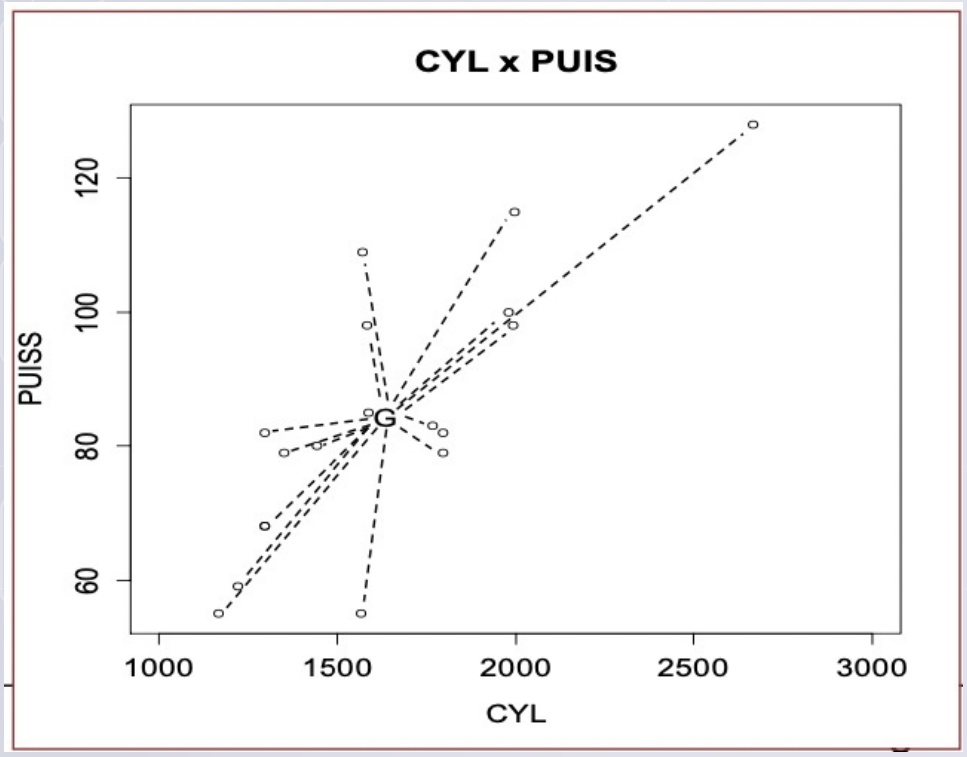
  - On peut la voir comme une compression avec perte contrôlée de l'information.
- **Un point clé :**
  - La distance entre l'ensemble des individus deux à deux,
    - *Dite aussi l'inertie du nuage de points dans l'espace original. Elle traduit la quantité d'information disponible.*

$$I_p = \frac{1}{2n^2} \sum_{i=1}^{n} \sum_{i'=1}^{n} d^2(i,i')$$

$$I_p = \frac{1}{n} \sum_{i=1}^{n} d^2(i,G)$$

Réduction de dimensions : Exemple

- L'inertie indique la dispersion autour du centre de gravité : c'est une variance multidimensionnelle (calculée sur toutes les dimensions)



Il faut préserver au mieux cette inertie (variance) sur le nouvelle espace à créer

La solution proposée : ACP

Analyse en composantes principales (ACP) :

- Outil puissant pour analyser les corrélations entre plusieurs variables.

$$r_{jm} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(x_{im} - \bar{x}_m)}{s_j \times s_m}$$

- **L'idée :** Les corrélations entre les variables originales sont aspirées par les nouvelles dimensions crées : *La compression*

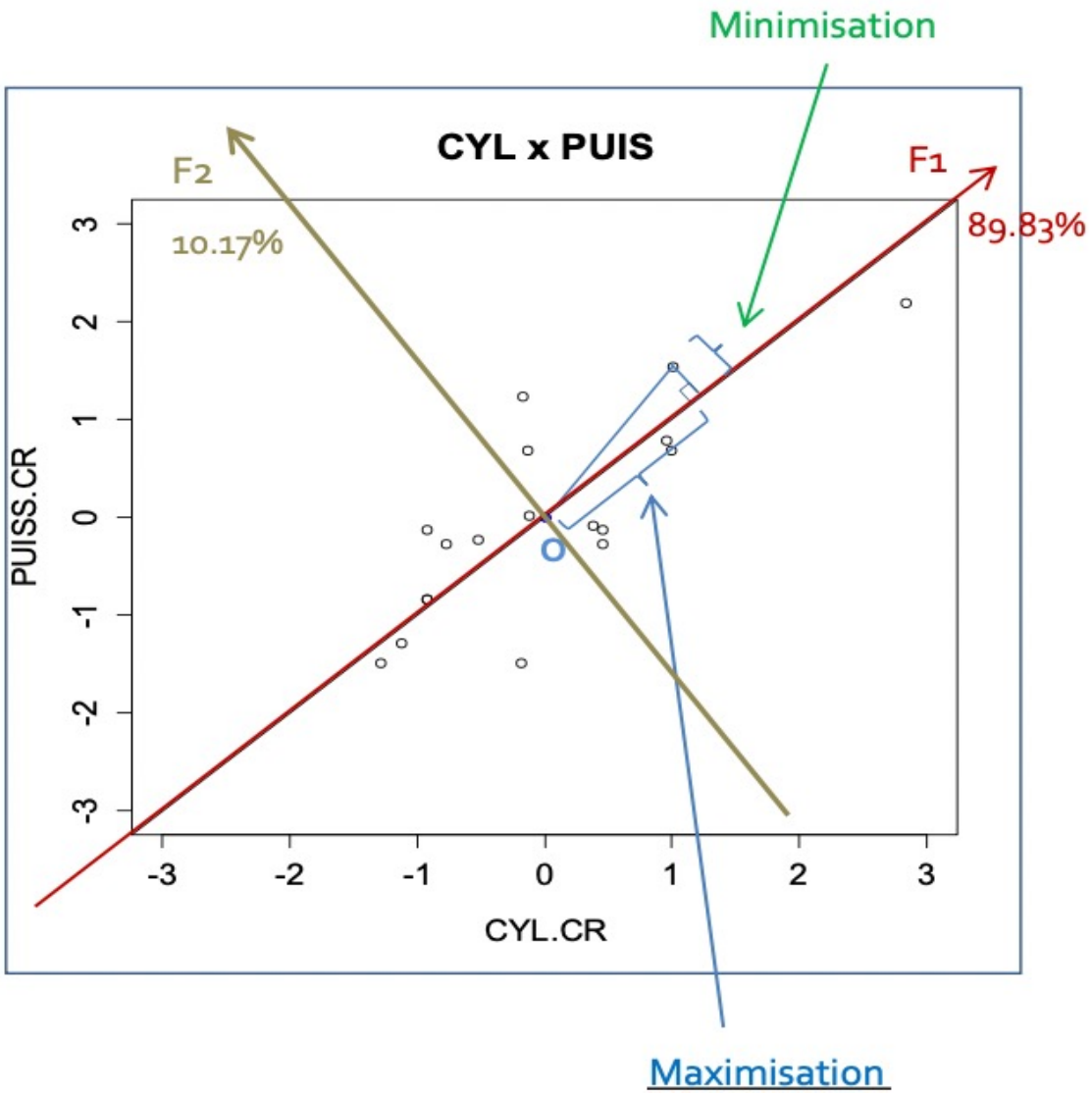
- Permet d'obtenir de nouvelles variables non corrélées à forte variance (*i.e.* à fort pouvoir informatif) par combinaison linéaire des *p* variables initiales.

- Ces variables seront appelées «*composantes principales*», les axes qu'elles déterminent : « axes principaux »

ACP : Cas particulier de deux variables

- **Données centrées réduites : ACP normée**
- **On prend l'exemple des variables CYL et PUISS**
  - **Inertie totale** I_p = p = 2
  - Après résolution du problème algébrique $\mathbf{det}(\boldsymbol{R} - \boldsymbol{\lambda} \times \boldsymbol{I}) = \mathbf{0}$, deux valeurs propres obtenues $\boldsymbol{\lambda 1 = 1.79662}$ et $\boldsymbol{\lambda 2 = 0.20337}$ sachant que $\boldsymbol{\lambda 1 + \lambda 2 = 2}$ dans le cas d'une ACP normée
  - Part d'inertie expliquée par l'axe 1 : $\frac{\lambda 1}{I_p} = \frac{1.79662}{2} = 89.83\%$
  - Part d'inertie expliquée par l'axe 2 : $\frac{\lambda 2}{I_p} = \frac{0.20337}{2} = 10.17\%$

ACP : Cas particulier de deux variables



L'axe 1 tout seul est bien représentatif

ACP : Cas particulier de deux variables

	Modele	CYL	PUISS
1	Toyota Corolla	-1.2814	-1.4953
2	Citroen GS Club	-1.1273	-1.2933
3	Simca 1300	-0.9292	-0.8389
4	Lada 1300	-0.9292	-0.8389
5	Lancia Beta	-0.9209	-0.1319
6	Alfasud TI	-0.7751	-0.2834
7	Rancho	-0.5219	-0.2329
8	Renault 16 TL	-0.1835	-1.4953
9	Alfetta 1.66	-0.1697	1.2316
10	Fiat 132	-0.1284	0.6761
11	Audi 100	-0.1202	0.0196
12	Mazda 9295	0.3779	-0.0814
13	Peugeot 504	0.4522	-0.2834
14	Princess 1800	0.4577	-0.1319
15	Opel Rekord	0.9558	0.7771
16	Taunus 2000	0.9943	0.6761
17	Datsun 200L	1.0081	1.5346
18	Renault 30	2.8408	2.1911

	Modele	F1 (89.83%)	F2 (10.17%)
1	Toyota Corolla	1.9635	0.1513
2	Citroen GS Club	1.7117	0.1174
3	Simca 1300	1.2502	-0.0639
4	Lada 1300	1.2502	-0.0639
5	Lancia Beta	0.7444	-0.5580
6	Alfasud TI	0.7484	-0.3477
7	Rancho	0.5337	-0.2044
8	Renault 16 TL	1.1871	0.9276
9	Alfetta 1.66	-0.7509	-0.9909
10	Fiat 132	-0.3873	-0.5689
11	Audi 100	0.0711	-0.0989
12	Mazda 9295	-0.2097	0.3248
13	Peugeot 504	-0.1194	0.5201
14	Princess 1800	-0.2304	0.4169
15	Opel Rekord	-1.2254	0.1263
16	Taunus 2000	-1.1812	0.2250
17	Datsun 200L	-1.7980	-0.3723
18	Renault 30	-3.5581	0.4594

$$d^2(1,2) = (-1.2814 - (-1.1273))^2 + (-1.4953 - (-1.2933))^2$$
$$= 0.06455$$
$$d^2(2,6) = 1.14415$$
$$d^2(1,6) = 1.72529$$

$$d^2_{\{F_1,F_2\}}(1,2) = (1.9635 - 1.7117)^2 + (0.1513 - 0.1174)^2$$
$$= 0.06455$$
$$d^2_{\{F_1,F_2\}}(2,6) = 1.14415$$
$$d^2_{\{F_1,F_2\}}(1,6) = 1.72529$$

ACP : Cas particulier de deux variables

$$d^2_{\{F_1,F_2\}}(1,2) = (1.9635 - 1.7117)^2 + (0.1513 - 0.1174)^2$$
$$= 0.06455$$
$$d^2_{\{F_1,F_2\}}(2,6) = 1.14415$$
$$d^2_{\{F_1,F_2\}}(1,6) = 1.72529$$

- Si on tient compte que du premier axe factoriel les proximités sont bien respectées

$$d^2_{\{F_1\}}(1,2) = (1.9335 - 1.7117)^2$$
$$= 0.06340$$
$$d^2_{\{F_1\}}(2,6) = 0.92783$$
$$d^2_{\{F_1\}}(1,6) = 1.147632$$

ACP : plusieurs variables

- Analyse de la relation entre les variables : Matrice de corrélation R

CORR	CYL	PUISS	LONG	LARG	POIDS	V.MAX
CYL	1	0.797	0.701	0.630	0.789	0.665
PUISS		1	0.641	0.521	0.765	0.844
LONG			1	0.849	0.868	0.476
LARG				1	0.717	0.473
POIDS					1	0.478
V.MAX						1

- ACP $\det(R - \lambda \times I) = 0$

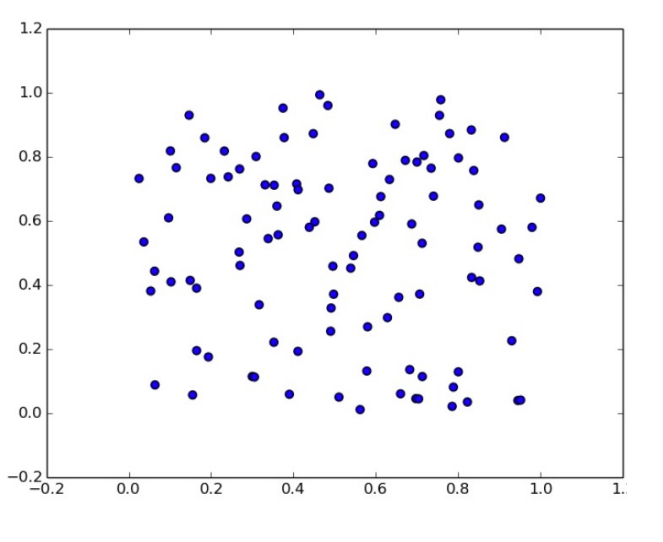
Axis	Eigen value	Proportion (%)	Cumulative (%)
1	4.421	73.68%	73.68%
2	0.856	14.27%	87.95%
3	0.373	6.22%	94.17%
4	0.214	3.57%	97.73%
5	0.093	1.55%	99.28%
6	0.043	0.72%	100.00%
Tot.	6	-	-

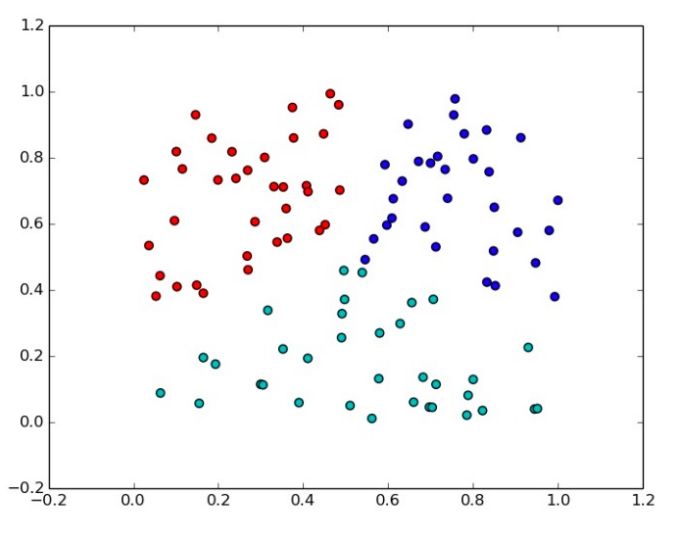
Les principales questions ici

- Comment choisir le nombre d'axes ?
  - Il faut obtenir une approximation suffisamment satisfaisante
- Comment interpréter les nouveaux axes ?

Problématique : Clustering

- **Les techniques de clustering cherchent à décomposer un ensemble d'individus en plusieurs sous ensembles les plus homogènes possibles**





- **Les méthodes sont très nombreuses**
  - Méthodes de partitionnement : k-means
  - Méthodes hiérarchiques : CAH

Problématique : Clustering

- **Défis du clustering**
  - On ne connaît pas la classe des exemples : nombre, forme, taille

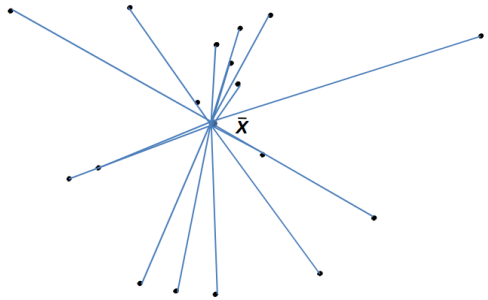
- **Principe de base**
  - Regrouper tout ce qui se ressemble,
  - Eloigner tout ce qui est franchement différent
  - Le point clé est le critère de similarité (fonction de distance)

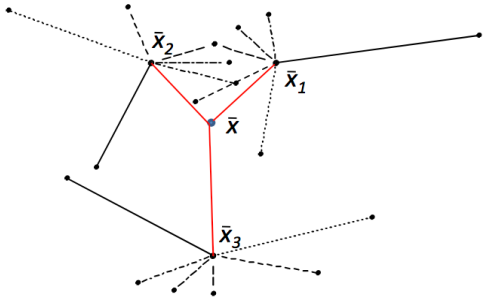
Problématique : Clustering

- **On cherche à minimiser *l'inertie intra-classe* et à maximiser *l'inertie inter-classe***
  - ***inertie intraclasse*** définie par la somme des distances des centres de gravités aux points de leurs classes
  - ***inertie interclasse*** définie par la *somme des* distances *entre les* centres de gravités des classes et le centre de gravité de la population totale

$\bar{x}_k$ moyenne de x_k, $\bar{x}_{qk}$ moyenne de x_k dans la classe q

$$\underbrace{\sum_{k=1}^{K}\sum_{q=1}^{Q}\sum_{i=1}^{I}(x_{iqk}-\bar{x}_k)^2}_{\text{Inertie totale}} = \underbrace{\sum_{k=1}^{K}\sum_{q=1}^{Q}\sum_{i=1}^{I}(x_{iqk}-\bar{x}_{qk})^2}_{\text{Inertie intra}} + \underbrace{\sum_{k=1}^{K}\sum_{q=1}^{Q}\sum_{i=1}^{I}(\bar{x}_{qk}-\bar{x}_k)^2}_{\text{Inertie inter}}$$



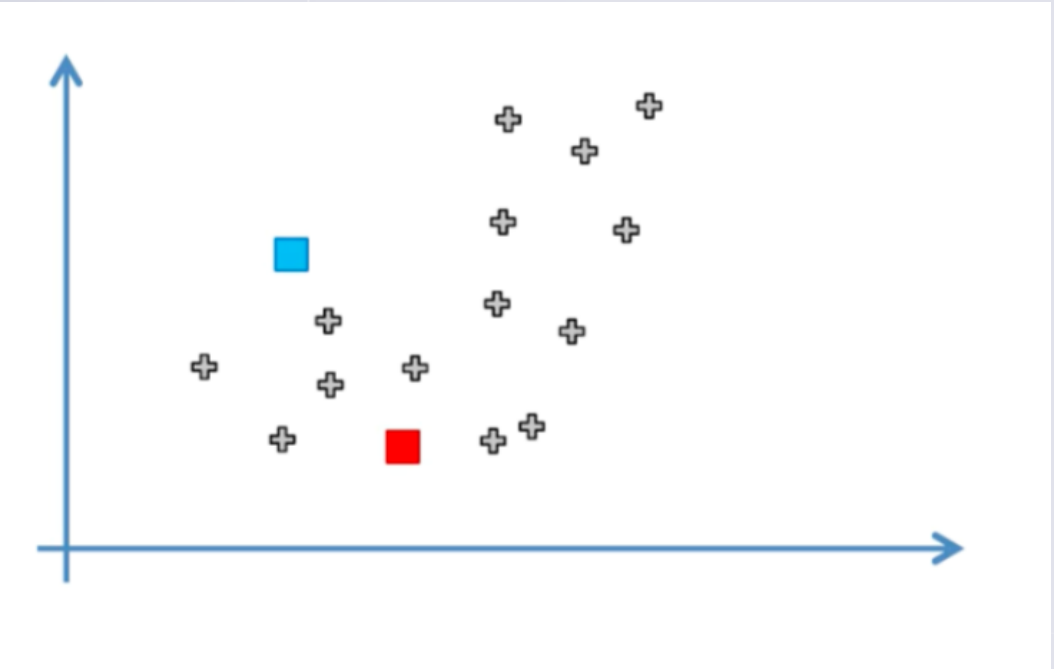


Clustering : K-means

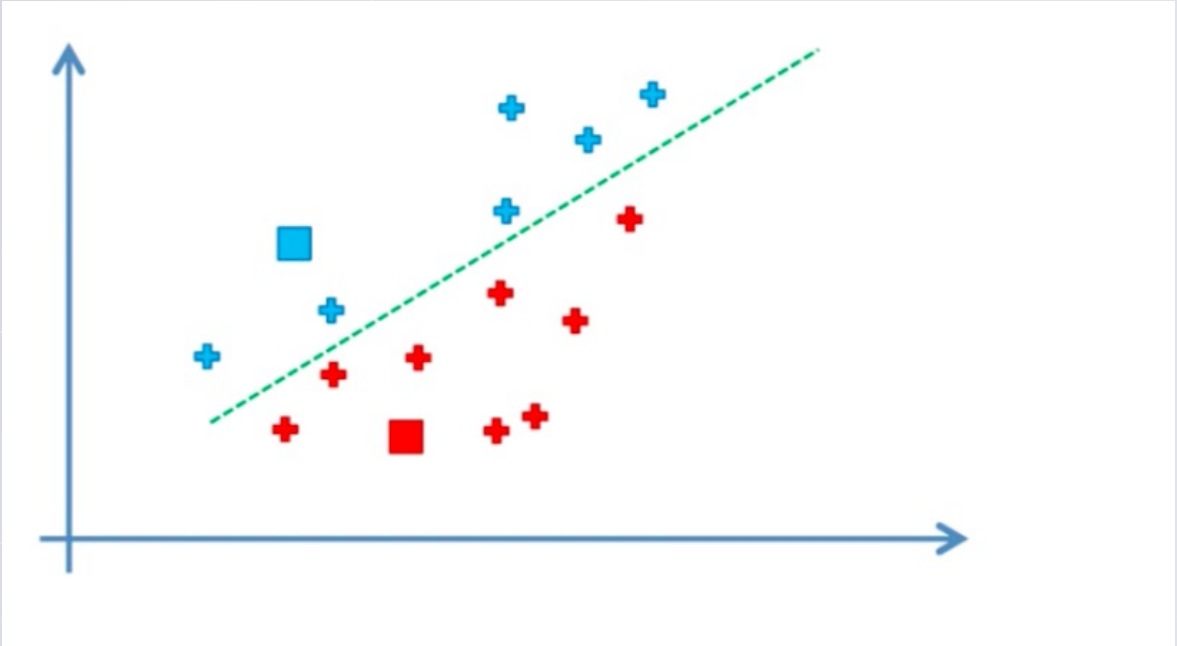
- Algorithme en deux étapes
  - Affecter chaque point à un cluster en fonction des centres de gravité.
  - Ré-estimer le centre gravité de chaque cluster en fonction de la nouvelle répartition
  - Et boucler jusqu'à convergence.

Clustering : K-means

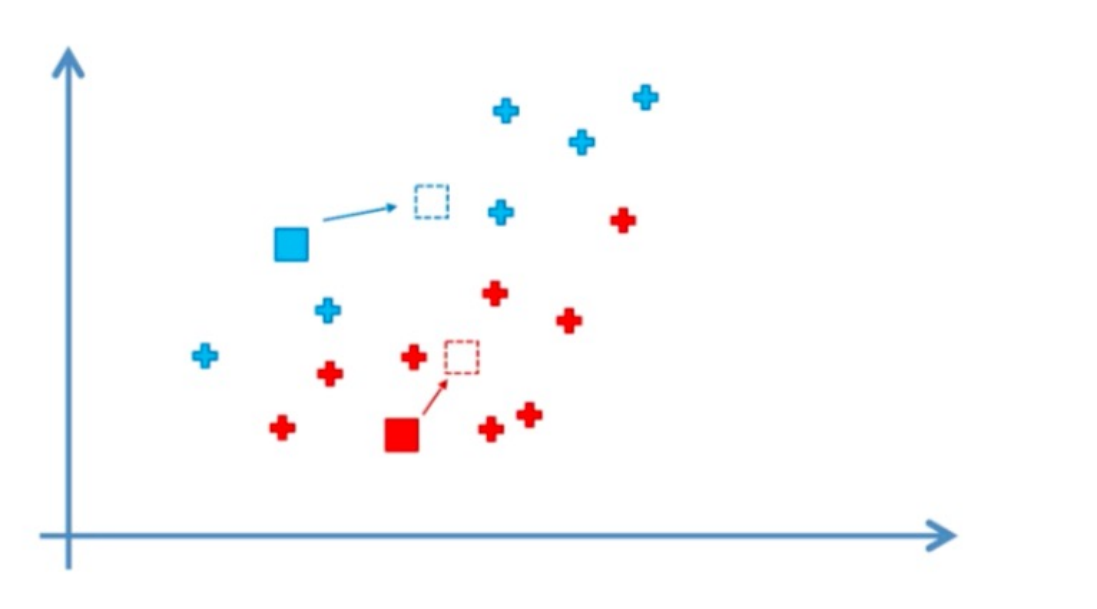
Clustering : K-means



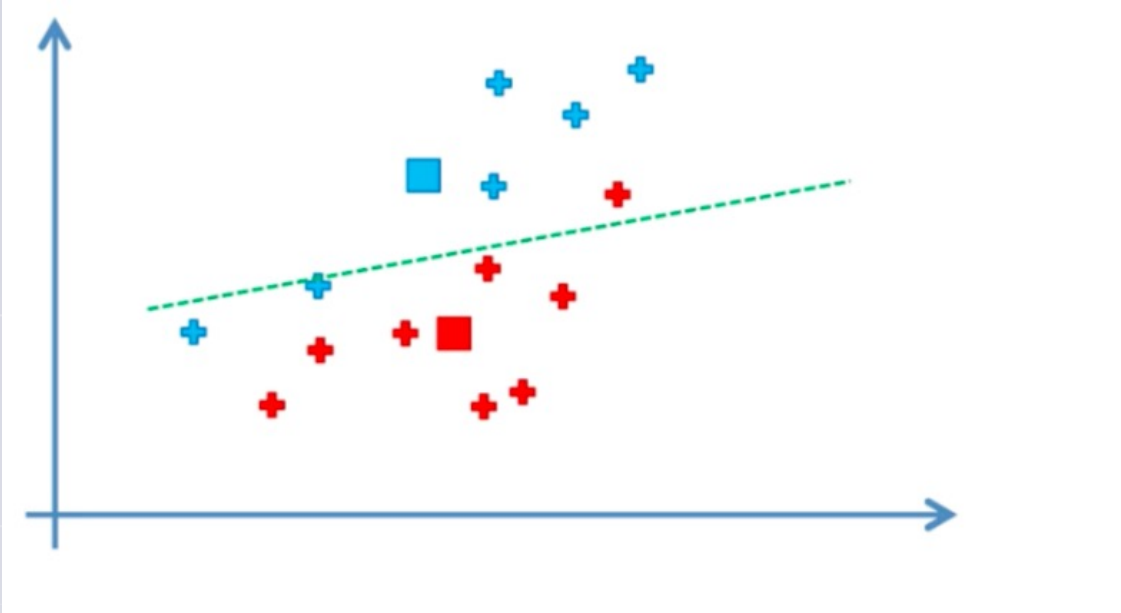
Clustering : K-means



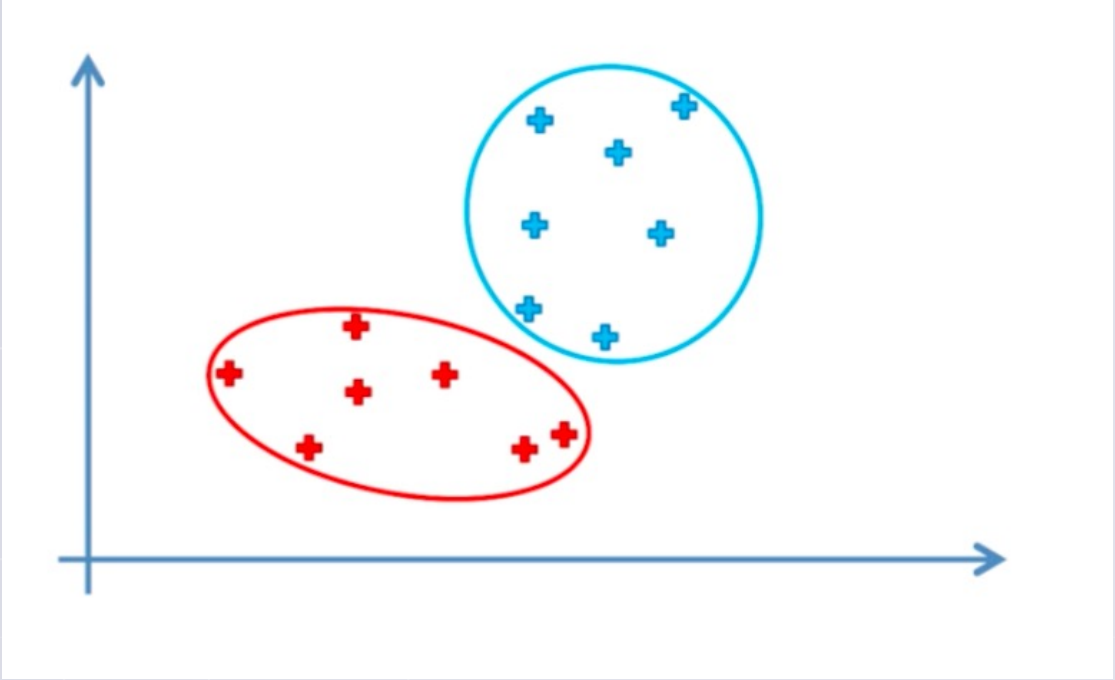
Clustering : K-means



Clustering : K-means



Clustering : K-means

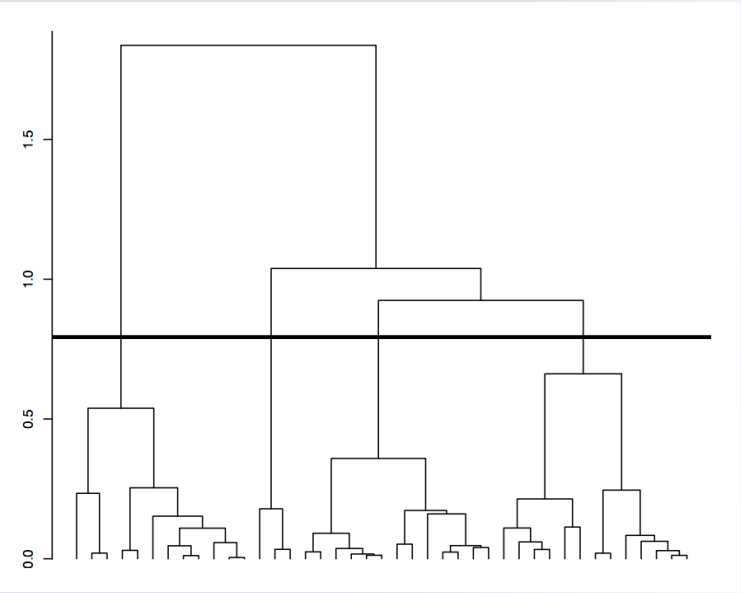


Clustering : CAH

- **Algorithme glouton**
  - Fusionner les instances les plus similaires dans un même cluster
  - Construire incrémentalement des clusters plus larges en fusionnant les 2 clusters les plus proches
  - S'arrêter lorsqu'il n'y a plus qu'un cluster.

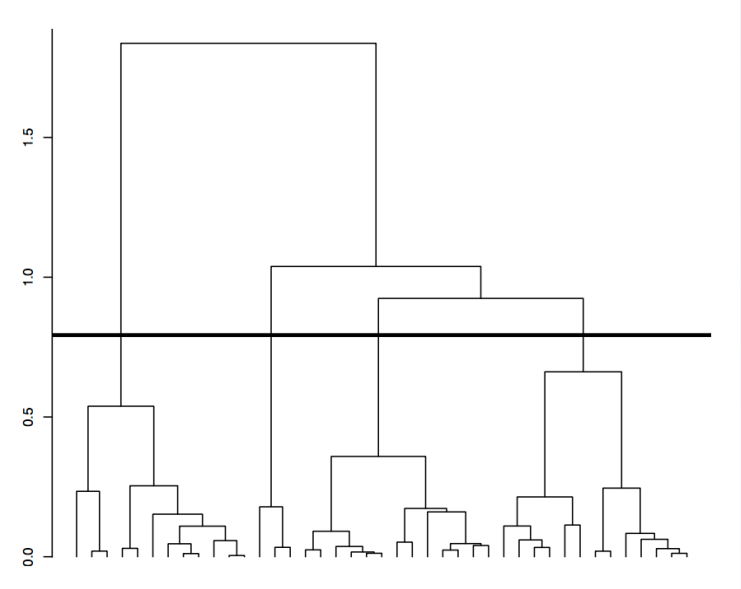
⇒ Construit un arbre de partitionnement

un *dendrogramme*.

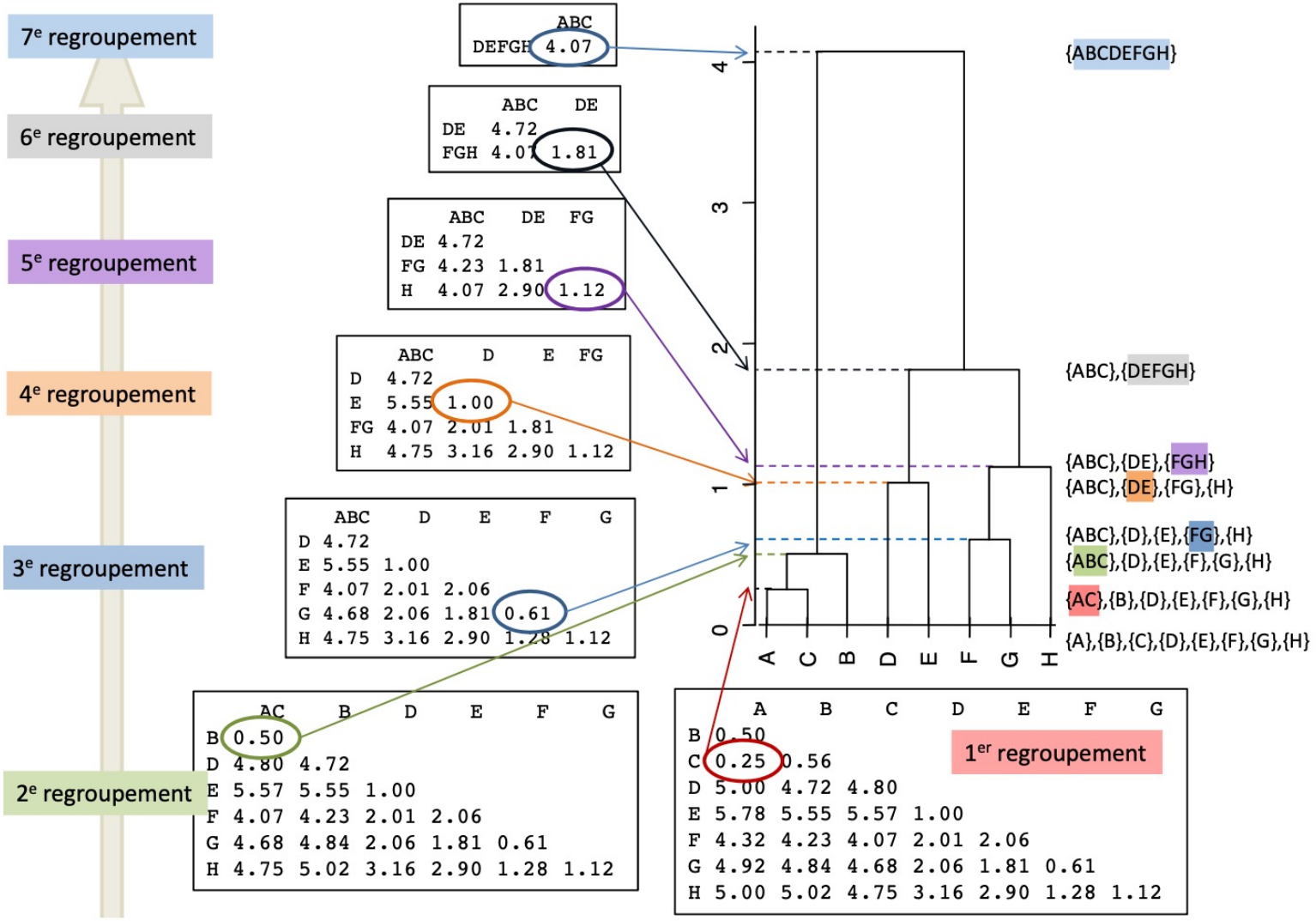


Clustering : CAH

- Algorithme glouton
  - Fusionner les instances les plus similaires dans un même cluster
  - Construire incrémentalement des clusters plus larges en fusionnant les 2 clusters les plus proches
  - S'arrêter lorsqu'il n'y a plus qu'un cluster.



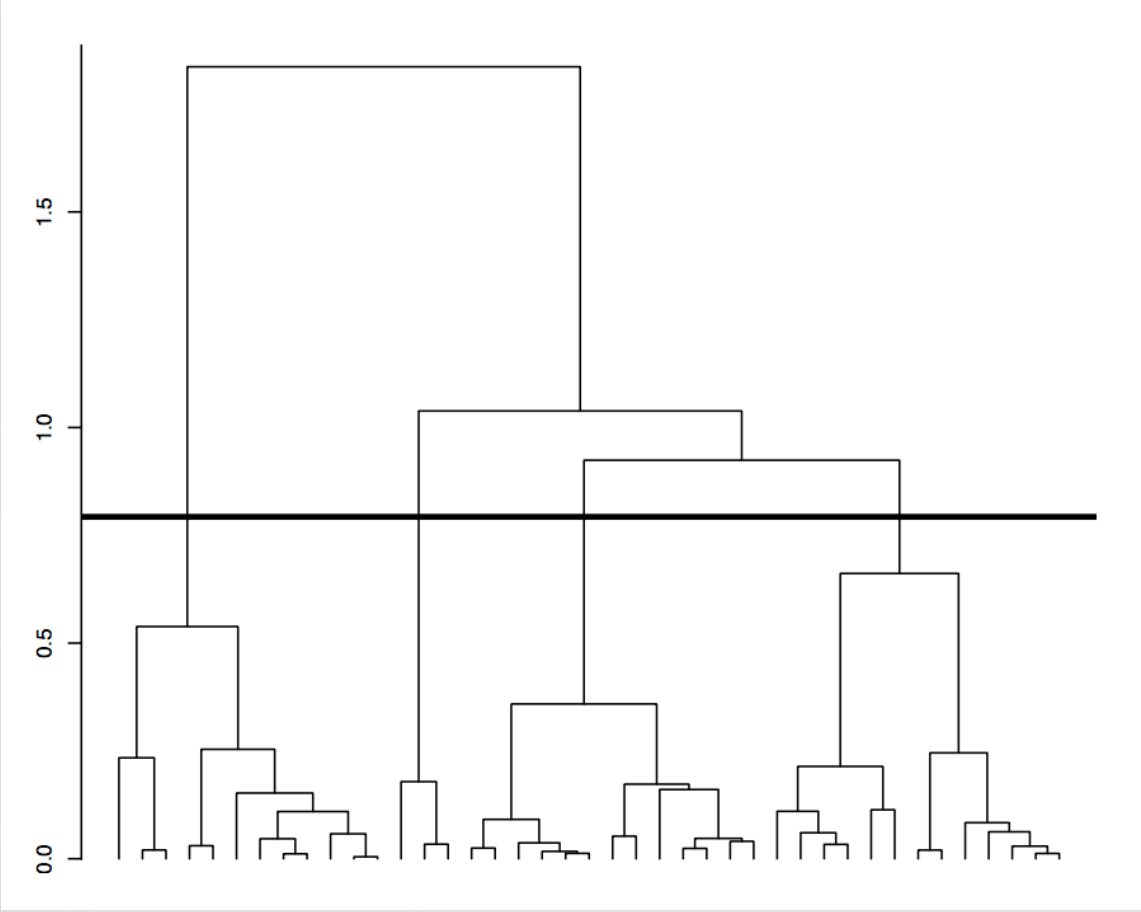
Clustering : CAH



Construction d'un arbre de partitionnement

***dendrogramme*.**

Clustering : CAH



Les principales questions ici

- Comment définir la similarité ?
- Combien de clusters ?
- Quelle méthode de clustering choisir ?
- Peut-on interpréter les clusters obtenus ?