# ANALYSE DES DONNEES
# PANORAMA DES METHODES

*Pierre-Louis GONZALEZ*

# ANALYSE DES DONNEES
# LES METHODES USUELLES

- **Analyse en Composantes Principales (A.C.P.)**
- **Analyse factorielle des correspondances simples (A.F.C.)**
- **Analyse factorielle des correspondances multiples (A.C.M.)**

- **Méthodes de classification automatique**

# ANALYSE DES DONNEES

Analyser des données, c'est extraire d'une **masse d'informations brutes**, des **éléments de réponse** aux **questions** qui résultent **des objectifs globaux poursuivis**.

# Analyse des données

## Traitement de données en masse

- Grand nombre d'individus
- Grand nombre de variables

## Développement parallèle à l'informatique

- Fichiers volumineux → demande de méthodes
- Capacité de calcul → méthodes praticables

# Analyse des données

**Principes mathématiques anciens mais utilisation nouvelle**

- A.C.P. (K. Pearson, 1901)
- A.F.C. (Hirchsfeld, 1936)
- A.F.C. multiple (Guttman, 1941)

**Auteur français** : J.P. Benzecri, 1967….

# Analyse des données

**Outil d'exploration, de description et d'analyse d'ensembles d'individus:**

- Représentations géométriques
- Création de nouvelles variables
- Typologie

**L'analyse de données est bien plus que la statistique descriptive:**

- Nouveau regard sur les données
- L'individu redevient le point d'intérêt central.

# Analyse des données

**Les outils mathématiques de l'analyse des données:**

- **Algèbre linéaire**
- **Calcul matriciel**

# Méthodes factorielles

- **Analyse en Composantes Principales (A.C.P.)**

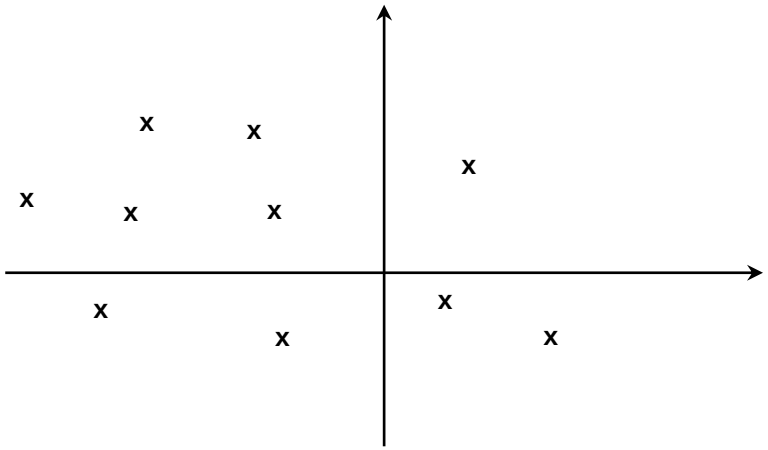
  analyse de variables quantitatives

- **Analyse factorielle des correspondances**

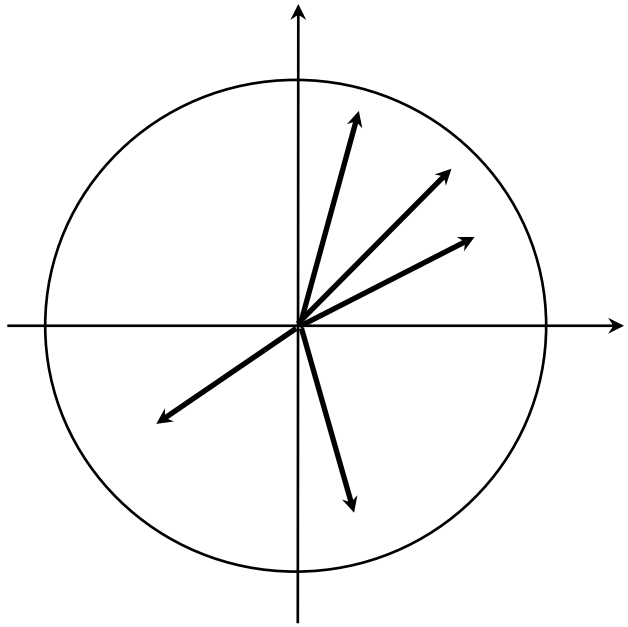
  analyse de variables qualitatives

  - **Correspondances simples (A.F.C.)** (Étude d'un tableau de contingence)
  - **Correspondances multiples (A.C.M.)** (Utile lors du dépouillement d'enquêtes)

# L'analyse en composantes principales



**Représentation des individus: Plan factoriel**



**Représentation des variables: Cercle des corrélations**

# Données AUTOS 2005 - 1

| NOM COURT | PUISSANCE | CYLINDREE | VITESSE | LONGUEUR | LARGEUR |
|---|---|---|---|---|---|
| ALFA 156 | 250 | 3179 | 250 | 443 | 175 |
| AUDIA3 | 102 | 1595 | 185 | 421 | 177 |
| AUDIA8 | 280 | 3697 | 250 | 506 | 203 |
| AVENSIS | 115 | 1995 | 195 | 463 | 176 |
| BMW X5 | 218 | 2993 | 210 | 467 | 188 |
| BMW530 | 231 | 2979 | 250 | 485 | 185 |
| CHRYS300 | 340 | 5654 | 250 | 502 | 188 |
| CITRONC2 | 61 | 1124 | 158 | 367 | 166 |
| CITRONC4 | 138 | 1997 | 207 | 426 | 178 |
| CITRONC5 | 210 | 2496 | 230 | 475 | 178 |
| CLIO | 100 | 1461 | 185 | 382 | 164 |
| CORSA | 70 | 1248 | 165 | 384 | 165 |
| CORVETTE | 404 | 5970 | 300 | 444 | 185 |
| FIESTA | 68 | 1399 | 164 | 392 | 168 |
| GOLF | 75 | 1968 | 163 | 421 | 176 |
| LAGUNA | 165 | 1998 | 218 | 458 | 178 |
| LANDCRUI | 204 | 4164 | 170 | 489 | 194 |
| MAZDARX8 | 231 | 1308 | 235 | 443 | 177 |
| MEGANECC | 165 | 1998 | 225 | 436 | 178 |
| MERC_A | 140 | 1991 | 201 | 384 | 177 |

# Données AUTOS 2005 - 2

| NOM COURT | HAUTEUR | POIDS | COFFRE | RESERVOIR | CONSO | CO2 | PRIX |
|---|---|---|---|---|---|---|---|
| ALFA 156 | 141 | 1410 | 378 | 63 | 12,1 | 287 | 40800 |
| AUDIA3 | 143 | 1205 | 350 | 55 | 7 | 168 | 21630 |
| AUDIA8 | 145 | 1770 | 500 | 90 | 11,7 | 281 | 78340 |
| AVENSIS | 148 | 1400 | 510 | 60 | 5,8 | 155 | 26400 |
| BMW X5 | 172 | 2095 | 465 | 93 | 8,6 | 229 | 52000 |
| BMW530 | 147 | 1495 | 520 | 70 | 9,5 | 231 | 46400 |
| CHRYS300 | 148 | 1835 | 442 | 72 | 12,2 | 291 | 54900 |
| CITRONC2 | 147 | 932 | 224 | 41 | 5,9 | 141 | 10700 |
| CITRONC4 | 146 | 1381 | 314 | 60 | 5,4 | 142 | 23400 |
| CITRONC5 | 148 | 1589 | 471 | 65 | 10 | 238 | 33000 |
| CLIO | 142 | 980 | 255 | 50 | 4,3 | 113 | 17600 |
| CORSA | 144 | 1035 | 260 | 45 | 4,7 | 127 | 13590 |
| CORVETTE | 125 | 1517 | 295 | 69 | 13 | 310 | 63350 |
| FIESTA | 144 | 1138 | 261 | 45 | 4,4 | 117 | 14150 |
| GOLF | 149 | 1217 | 350 | 55 | 5,4 | 143 | 19140 |
| LAGUNA | 143 | 1320 | 430 | 70 | 8,2 | 196 | 25350 |
| LANDCRUI | 185 | 2495 | 403 | 96 | 11,1 | 292 | 67100 |
| MAZDARX8 | 134 | 1390 | 287 | 61 | 11,4 | 284 | 34000 |
| MEGANECC | 141 | 1415 | 190 | 60 | 8 | 191 | 27800 |
| MERC_A | 160 | 1340 | 435 | 54 | 5,4 | 141 | 24550 |

# Données AUTOS 2005 - 3

| NOM COURT | PUISSANCE | CYLINDREE | VITESSE | LONGUEUR | LARGEUR |
|---|---|---|---|---|---|
| MERC_E | 204 | 3222 | 243 | 482 | 183 |
| MODUS | 113 | 1598 | 188 | 380 | 170 |
| MONDEO | 145 | 1999 | 215 | 474 | 194 |
| MURANO | 234 | 3498 | 200 | 477 | 188 |
| MUSA | 100 | 1910 | 179 | 399 | 170 |
| OUTLAND | 202 | 1997 | 220 | 455 | 178 |
| P1007 | 75 | 1360 | 165 | 374 | 169 |
| P307CC | 180 | 1997 | 225 | 435 | 176 |
| P407 | 136 | 1997 | 212 | 468 | 182 |
| P607 | 204 | 2721 | 230 | 491 | 184 |
| PANDA | 54 | 1108 | 150 | 354 | 159 |
| PASSAT | 150 | 1781 | 221 | 471 | 175 |
| PTCRUISER | 223 | 2429 | 200 | 429 | 171 |
| SANTA_FE | 125 | 1991 | 172 | 450 | 185 |
| TAHOE | 290 | 5327 | 170 | 506 | 223 |
| TWINGO | 60 | 1149 | 151 | 344 | 163 |
| VECTRA | 150 | 1910 | 217 | 460 | 180 |
| VELSATIS | 150 | 2188 | 200 | 486 | 186 |
| X-TRAIL | 136 | 2184 | 180 | 446 | 177 |
| YARIS | 65 | 998 | 155 | 364 | 166 |

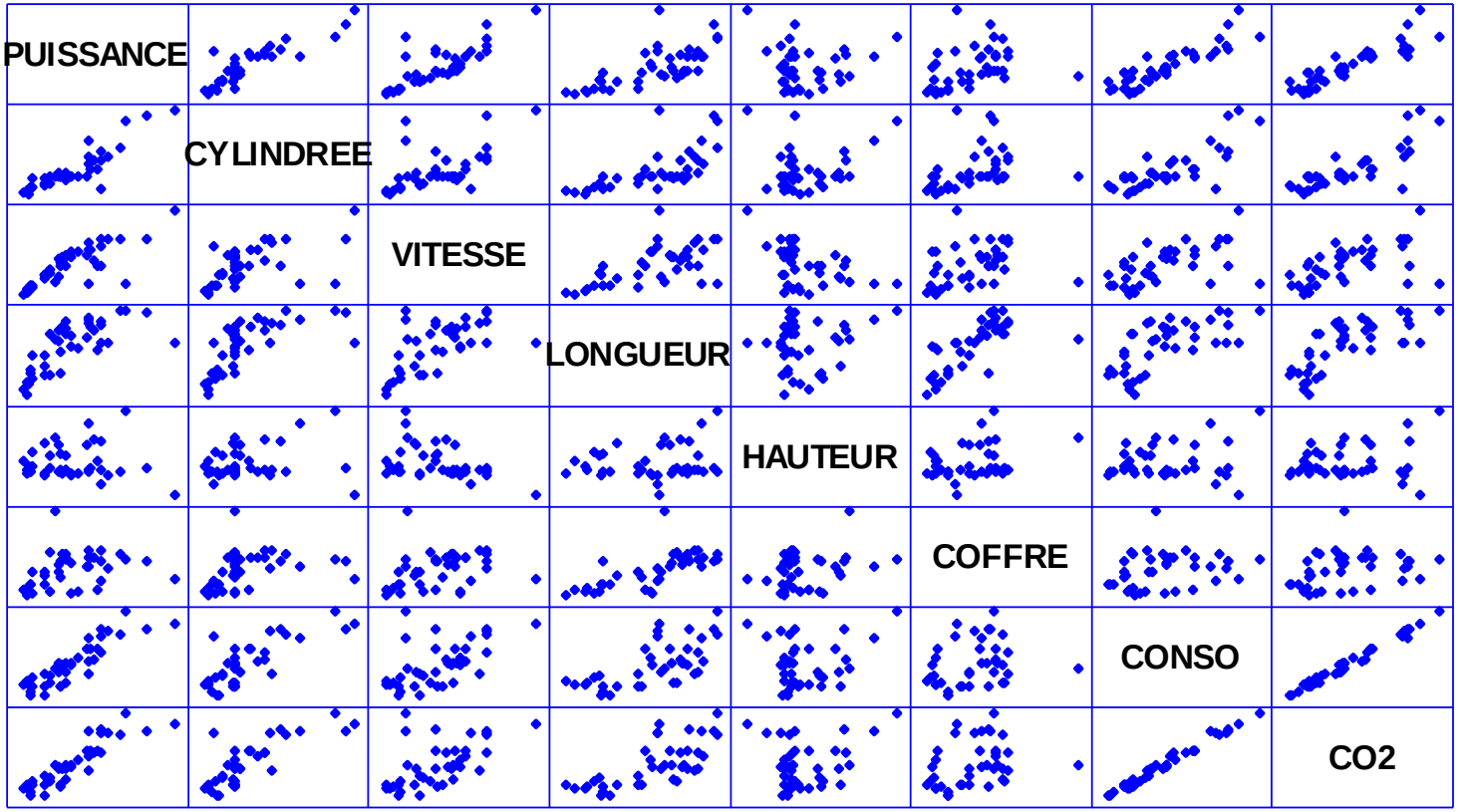
# Données AUTOS 2005 - 4

| NOM COURT | HAUTEUR | POIDS | COFFRE | RESERVOIR | CONSO | CO2 | PRIX |
|---|---|---|---|---|---|---|---|
| MERC_E | 146 | 1735 | 520 | 80 | 6,9 | 183 | 46450 |
| MODUS | 159 | 1170 | 198 | 49 | 6,8 | 163 | 16950 |
| MONDEO | 143 | 1378 | 500 | 59 | 7,9 | 189 | 23100 |
| MURANO | 171 | 1870 | 438 | 82 | 12,3 | 295 | 44000 |
| MUSA | 169 | 1275 | 320 | 47 | 5,5 | 146 | 17900 |
| OUTLAND | 167 | 1595 | 402 | 60 | 10 | 237 | 29990 |
| P1007 | 161 | 1181 | 178 | 50 | 6,4 | 153 | 13600 |
| P307CC | 143 | 1490 | 204 | 50 | 8,8 | 210 | 28850 |
| P407 | 145 | 1415 | 407 | 66 | 8,2 | 194 | 23400 |
| P607 | 145 | 1723 | 468 | 80 | 8,4 | 223 | 40550 |
| PANDA | 154 | 860 | 206 | 35 | 5,7 | 135 | 8070 |
| PASSAT | 147 | 1360 | 475 | 62 | 8,2 | 197 | 27740 |
| PTCRUISER | 154 | 1595 | 210 | 57 | 9,9 | 235 | 27400 |
| SANTA_FE | 173 | 1757 | 833 | 65 | 7,5 | 197 | 27990 |
| TAHOE | 196 | 2463 | 460 | 98 | 14,3 | 340 | 49600 |
| TWINGO | 143 | 840 | 168 | 40 | 6 | 143 | 8950 |
| VECTRA | 146 | 1428 | 500 | 61 | 5,9 | 159 | 26550 |
| VELSATIS | 158 | 1735 | 460 | 80 | 7,1 | 188 | 38250 |
| X-TRAIL | 168 | 1520 | 350 | 60 | 7,2 | 190 | 29700 |
| YARIS | 150 | 880 | 205 | 45 | 5,6 | 134 | 10450 |

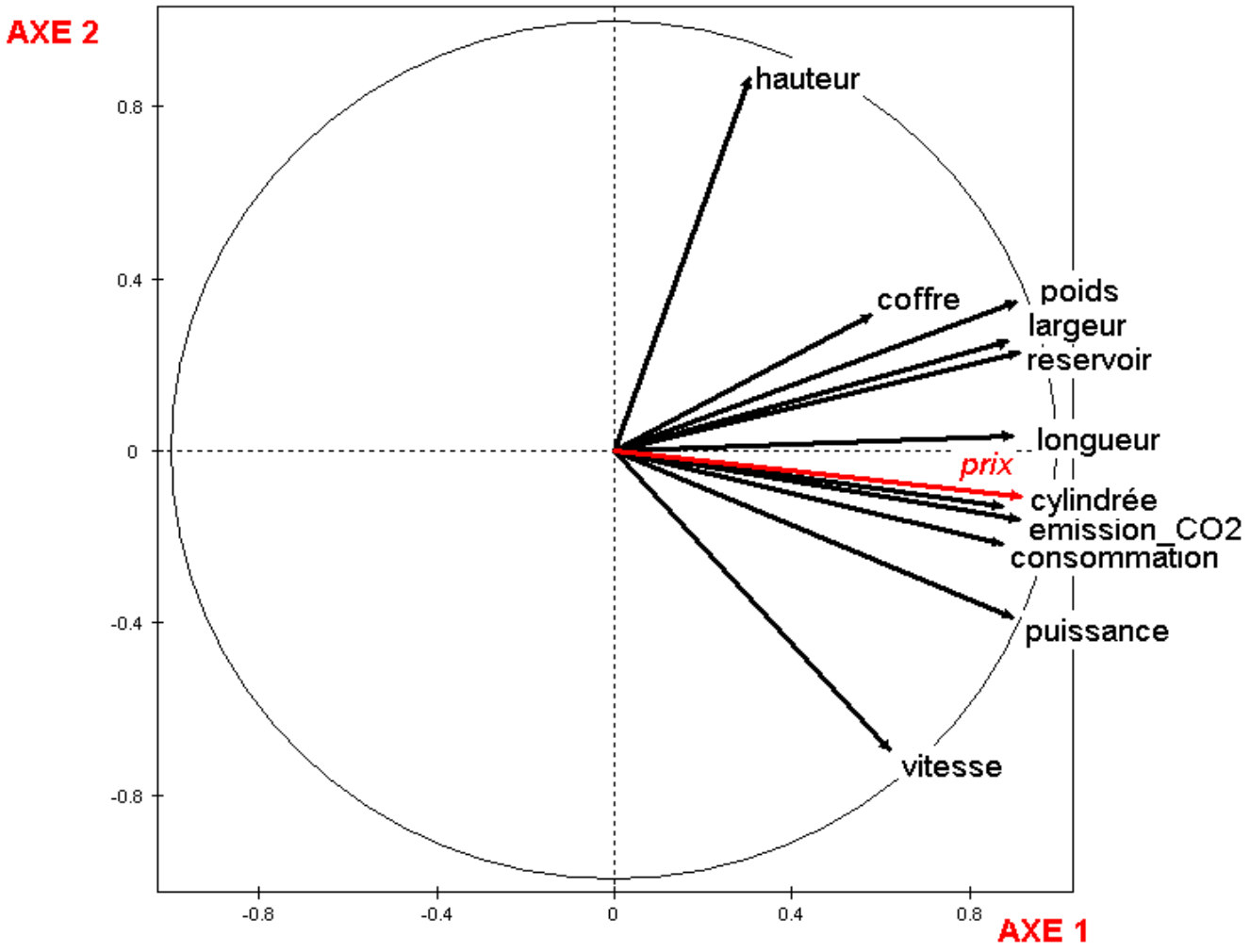
# MATRICE DES CORRELATIONS ENTRE VARIABLES

| | PUIS | CYLI | VITE | LONG | LARG | HAUT | POID | COFF | RESE | CONS | CO2 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| PUIS | 1.00 | | | | | | | | | | |
| CYLI | 0.89 | 1.00 | | | | | | | | | |
| VITE | 0.80 | 0.56 | 1.00 | | | | | | | | |
| LONG | 0.71 | 0.66 | 0.61 | 1.00 | | | | | | | |
| LARG | 0.66 | 0.73 | 0.36 | 0.82 | 1.00 | | | | | | |
| HAUT | 0.00 | 0.21 | -0.45 | 0.18 | 0.42 | 1.00 | | | | | |
| POID | 0.68 | 0.74 | 0.33 | 0.82 | 0.85 | 0.59 | 1.00 | | | | |
| COFF | 0.32 | 0.34 | 0.30 | 0.71 | 0.59 | 0.27 | 0.56 | 1.00 | | | |
| RESE | 0.70 | 0.73 | 0.43 | 0.86 | 0.87 | 0.41 | 0.92 | 0.60 | 1.00 | | |
| CONS | 0.89 | 0.79 | 0.58 | 0.66 | 0.69 | 0.18 | 0.69 | 0.26 | 0.67 | 1.00 | |
| CO2 | 0.90 | 0.81 | 0.58 | 0.71 | 0.73 | 0.23 | 0.76 | 0.32 | 0.74 | 0.99 | 1.00 |

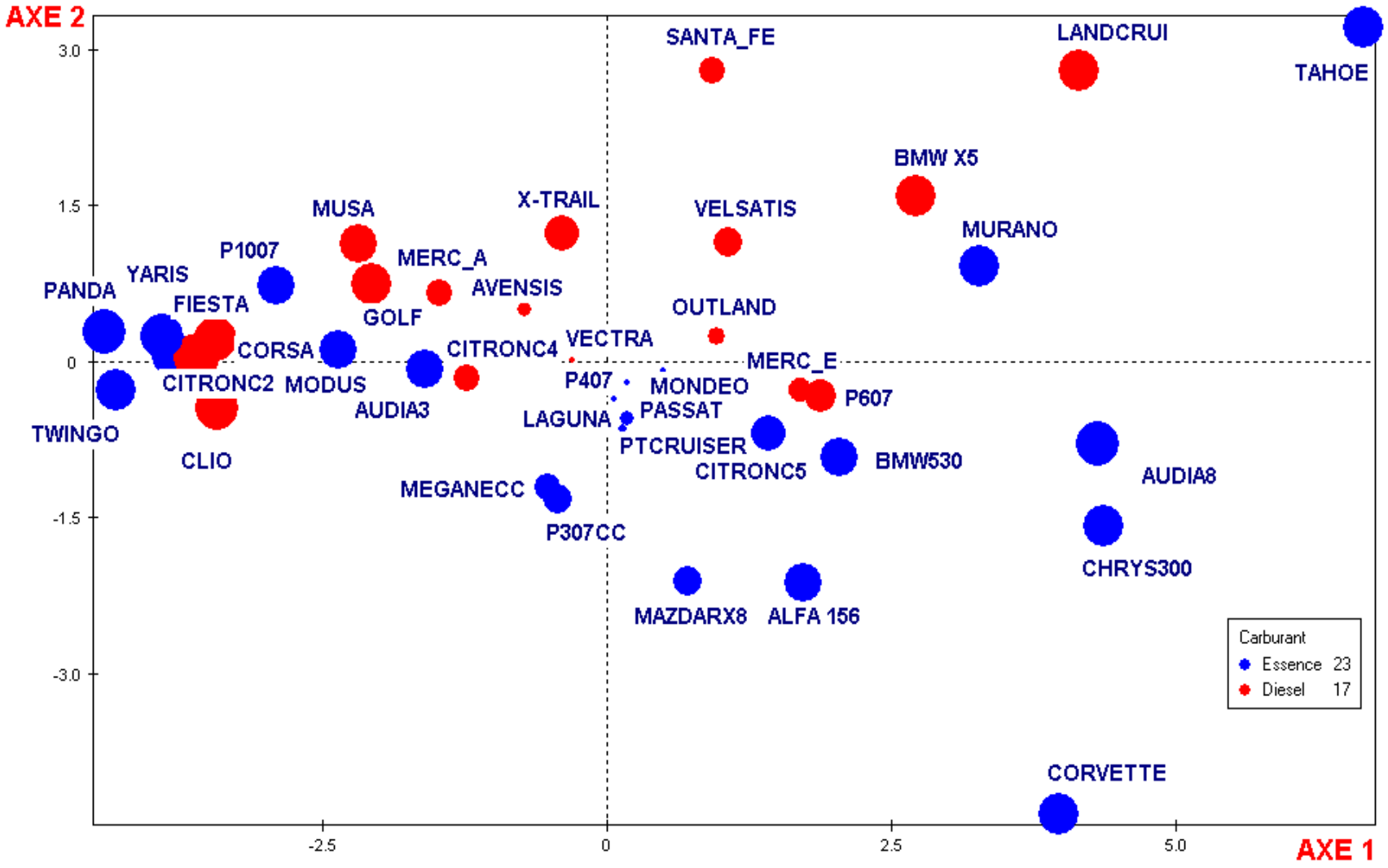
# Galerie de nuages de points



# A.C.P. Cercle des corrélations



# A.C.P. Représentation des individus dans le plan 1-2



# Méthodes de classification automatique

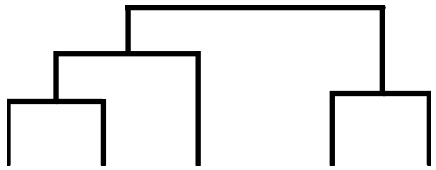
**Méthodes non hiérarchiques** ou méthodes à partitions :

- Centres mobiles
- Nuées dynamiques

**Méthodes hiérarchiques**

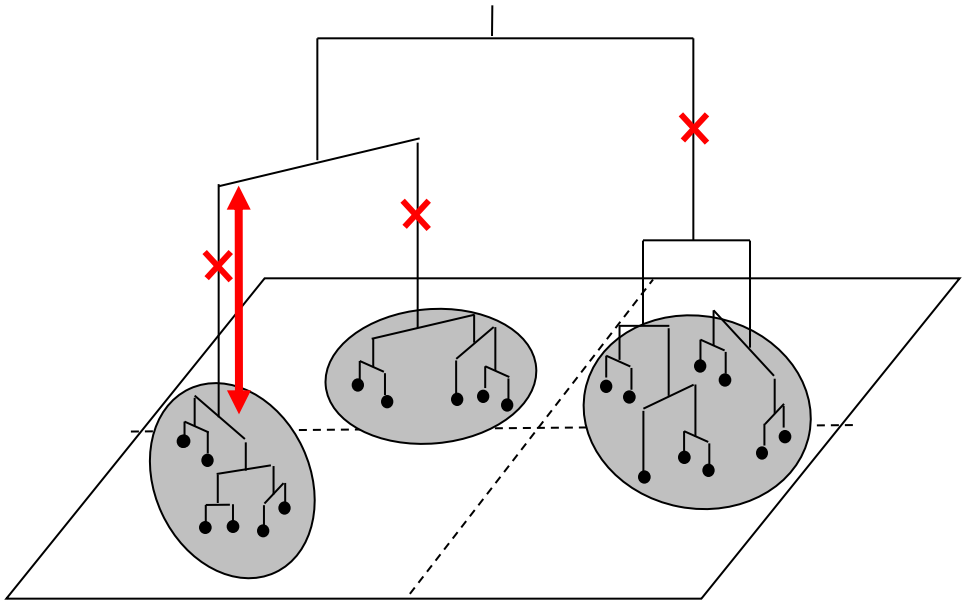
Représentations sous forme d'arbres

Méthode Ward



# Méthodes de classification

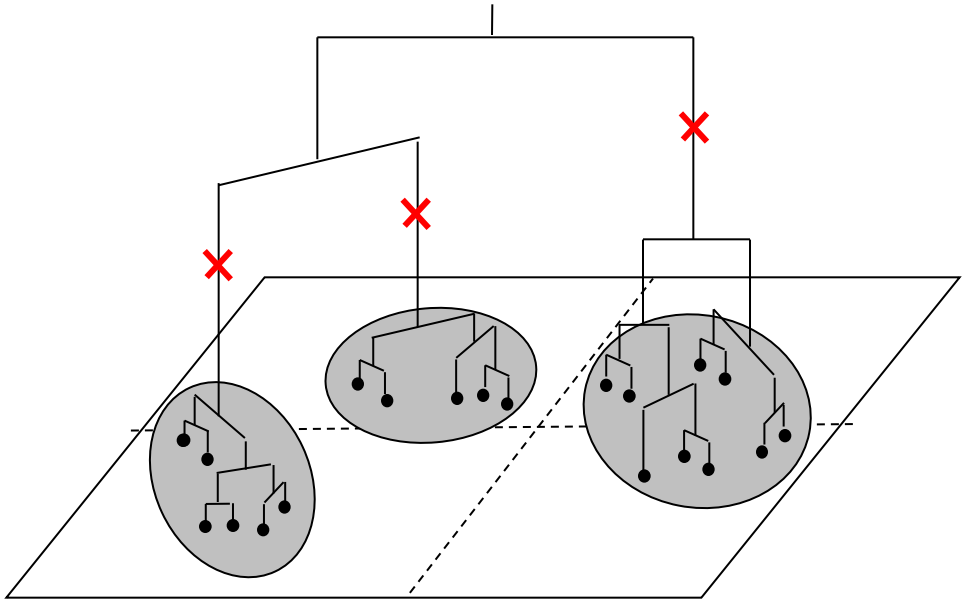
Méthode de la variance minimum de Ward



3 clusters

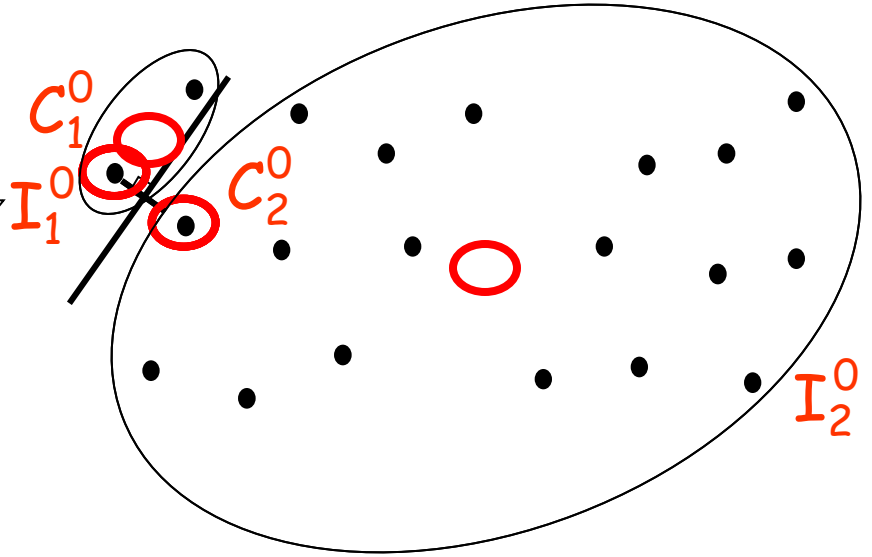
# Méthodes de classification

Méthode de la variance minimum de Ward



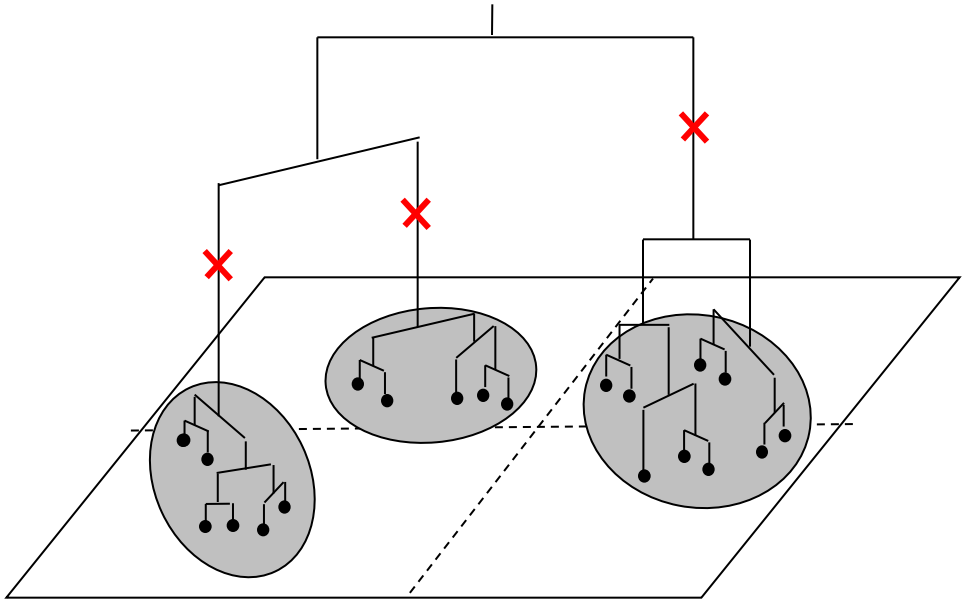
3 clusters

Algorithme K-means de MacQueen



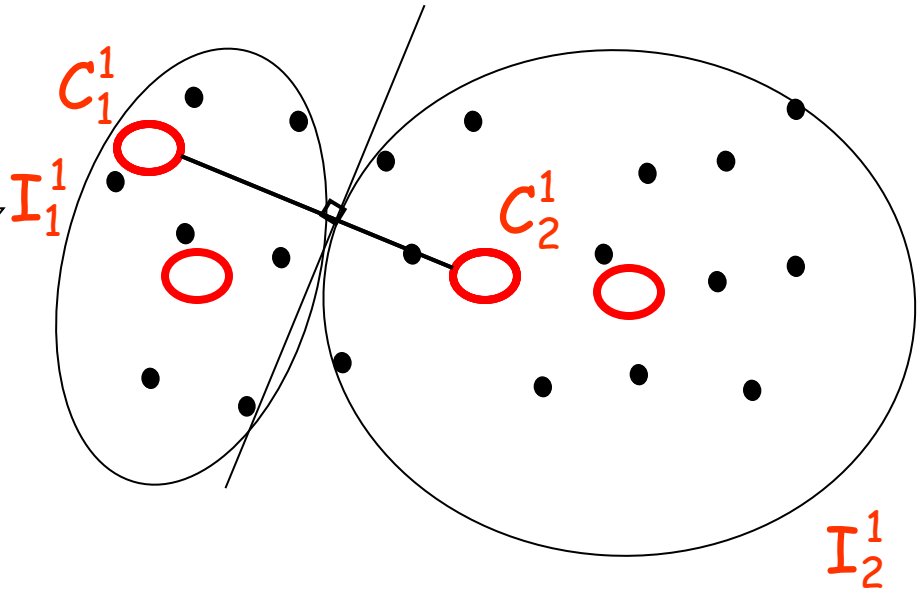
# Méthodes de classification

Méthode de la variance minimum de Ward



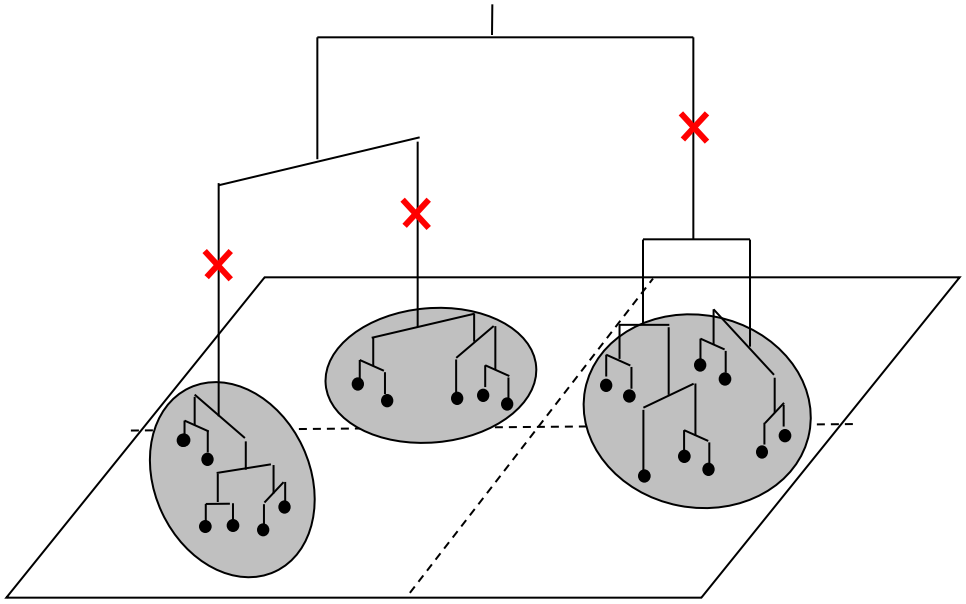
3 clusters

Algorithme K-means de MacQueen



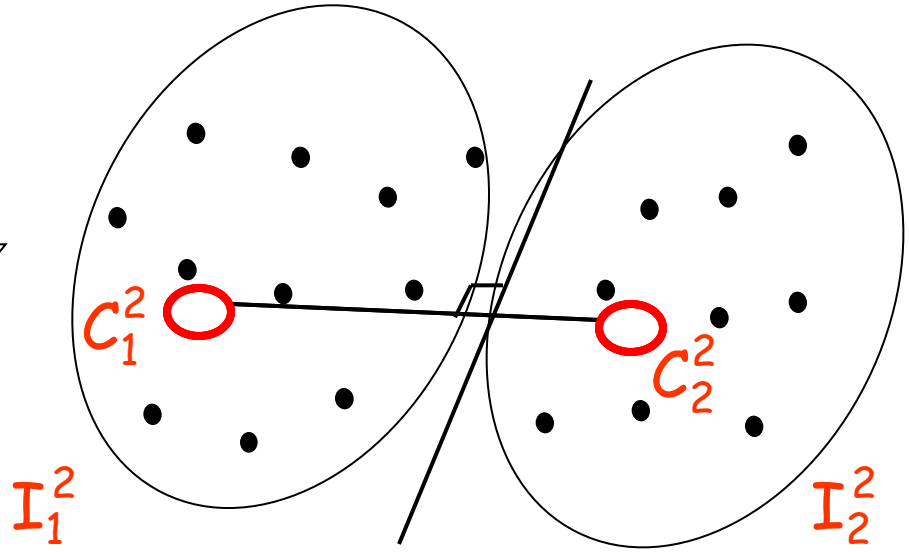
# Méthodes de classification

Méthode de la variance minimum de Ward



3 clusters

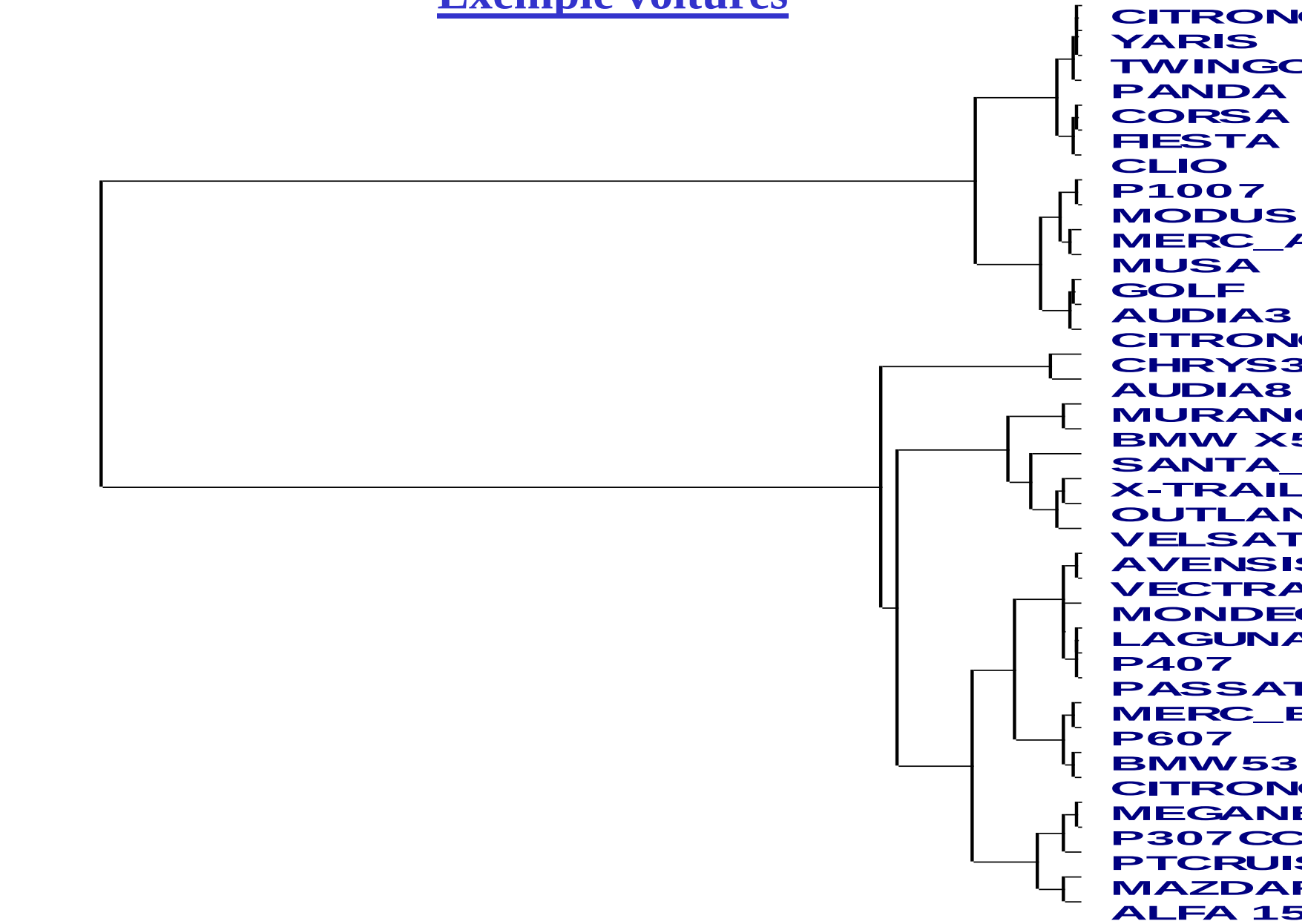
Algorithme K-means de MacQueen




2 clusters

# Classification hiérarchique directe

## Exemple voitures



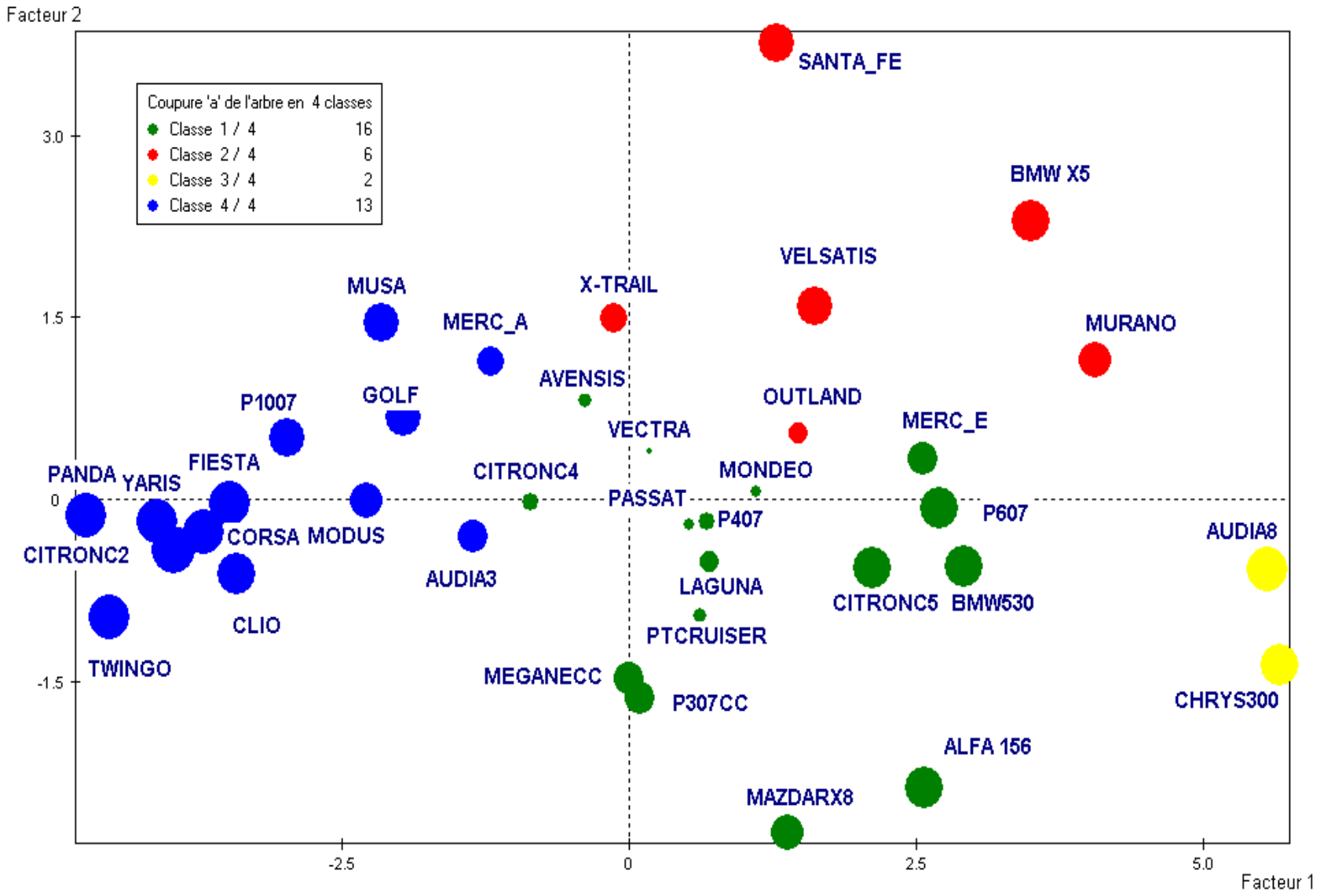
# CLASSIFICATION HIERARCHIQUE (VOISINS RECIPROQUES)

## SUR LES 10 PREMIERS AXES FACTORIELS DESCRIPTION DES NOEUDS

| NUM. | AINE | BENJ | EFF. | POIDS | INDICE | HISTOGRAMME DES INDICES DE NIVEAU |
|---|---|---|---|---|---|---|
| 56 | 16 | 1 | 2 | 2.00 | 0.07846 | ** |
| 57 | 42 | 54 | 6 | 6.00 | 0.08036 | ** |
| 58 | 41 | 31 | 3 | 3.00 | 0.08042 | ** |
| 59 | 22 | 5 | 2 | 2.00 | 0.08235 | ** |
| 60 | 43 | 52 | 4 | 4.00 | 0.09480 | ** |
| 61 | 53 | 35 | 3 | 3.00 | 0.10895 | ** |
| 62 | 50 | 46 | 7 | 7.00 | 0.11433 | ** |
| 63 | 7 | 3 | 2 | 2.00 | 0.14172 | *** |
| 64 | 60 | 51 | 7 | 7.00 | 0.19197 | *** |
| 65 | 58 | 56 | 5 | 5.00 | 0.22214 | **** |
| 66 | 32 | 61 | 4 | 4.00 | 0.24492 | **** |
| 67 | 57 | 55 | 10 | 10.00 | 0.33474 | ****** |
| 68 | 59 | 66 | 6 | 6.00 | 0.36644 | ****** |
| 69 | 62 | 64 | 14 | 14.00 | 0.54735 | ********* |
| 70 | 67 | 65 | 15 | 15.00 | 0.56610 | ********* |
| 71 | 68 | 70 | 21 | 21.00 | 0.96198 | *************** |
| 72 | 63 | 71 | 23 | 23.00 | 1.05146 | ****************** |
| 73 | 69 | 72 | 37 | 37.00 | 5.20370 | **********…..********** |

SOMME DES INDICES DE NIVEAU = 10.99623

# Partition en 4 classes



# EXEMPLE de description d'une classe  Classe 2 / 4

| V.TEST | PROBA | MOYENNES | | ECARTS TYPES | | |
|---|---|---|---|---|---|---|
| | | CLASSE | GENERALE | CLASSE | GENERAL | NUM.LIBELLE |
| Classe 2 / 4 | | ( POIDS = | 6.00 | EFFECTIF = | 6 ) | |
| 4.52 | 0.000 | 168.17 | 150.92 | 5.01 | 10.08 | 6.hauteur |
| 3.21 | 0.001 | 1762.00 | 1398.76 | 186.98 | 298.84 | 7.poids |
| 2.50 | 0.006 | 73.33 | 60.46 | 12.46 | 13.61 | 9.reservoir |
| 2.30 | 0.011 | 491.33 | 369.84 | 157.70 | 139.20 | 8.coffre |
| 1.91 | 0.028 | 183.67 | 177.11 | 4.50 | 9.04 | 5.largeur |
| 1.68 | 0.046 | 463.50 | 434.97 | 14.50 | 44.72 | 4.longueur |
| 1.65 | 0.050 | 222.67 | 190.43 | 37.43 | 51.67 | 11.emission_CO2 |
| 1.55 | 0.061 | 36988.33 | 28476.75 | 8748.07 | 14492.51 | 12.prix |
| 1.28 | 0.100 | 8.78 | 7.68 | 1.86 | 2.27 | 10.consommation |
| 1.00 | 0.159 | 2475.17 | 2138.30 | 569.30 | 891.44 | 2.cylindrée |
| 0.94 | 0.174 | 177.50 | 153.11 | 42.17 | 68.63 | 1.puissance |
| | | | | | | |
| -0.41 | 0.340 | 197.00 | 201.73 | 16.48 | 30.33 | 3.vitesse |

# Le Data Mining ( 1991 )

« Exploration de données »

« Forage de données »

Le data mining traite de la découverte de connaissances cachées, de modèles inattendus et de nouvelles règles issues de grandes bases de données.

Il est considéré comme l'élément clé d'un processus bien plus élaboré appelé découverte de connaissance dans les bases de données.

**( Knowledge Discovery in Databases : KDD)**

# Le Data Mining ( 1991 )

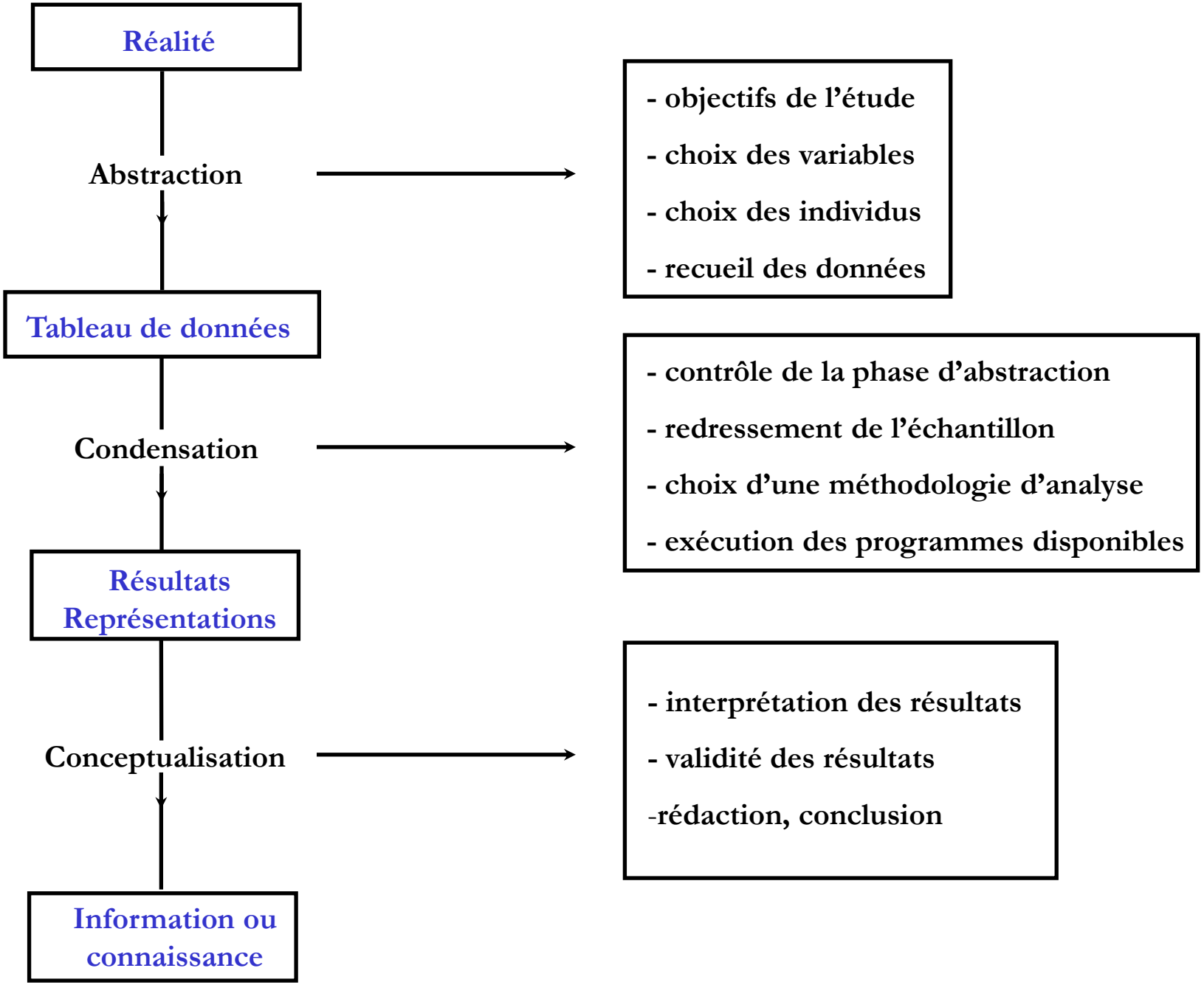
On utilise des méthodes descriptives et explicatives

- **Visualisation** : histogramme, nuage de points

- **Analyse de données** : analyse factorielle, classification

- **Objectif de classement** : analyse discriminante, régression logistique, réseau de neurones

**Explication** : régression, segmentation


Réalité
Abstraction
- objectifs de l'étude
- choix des variables
- choix des individus
- recueil des données
Tableau de données
Condensation
- contrôle de la phase d'abstraction
- redressement de l'échantillon
- choix d'une méthodologie d'analyse
- exécution des programmes disponibles
Résultats
Représentations
Conceptualisation
- interprétation des résultats
- validité des résultats
-rédaction, conclusion
Information ou
connaissance

# METHODES DESCRIPTIVES