OUTCOMEREA

Comment déterminer des profils cliniques homogènes

Journées OUTCOMEREA

4 décembre 2015

Sébastien BAILLY

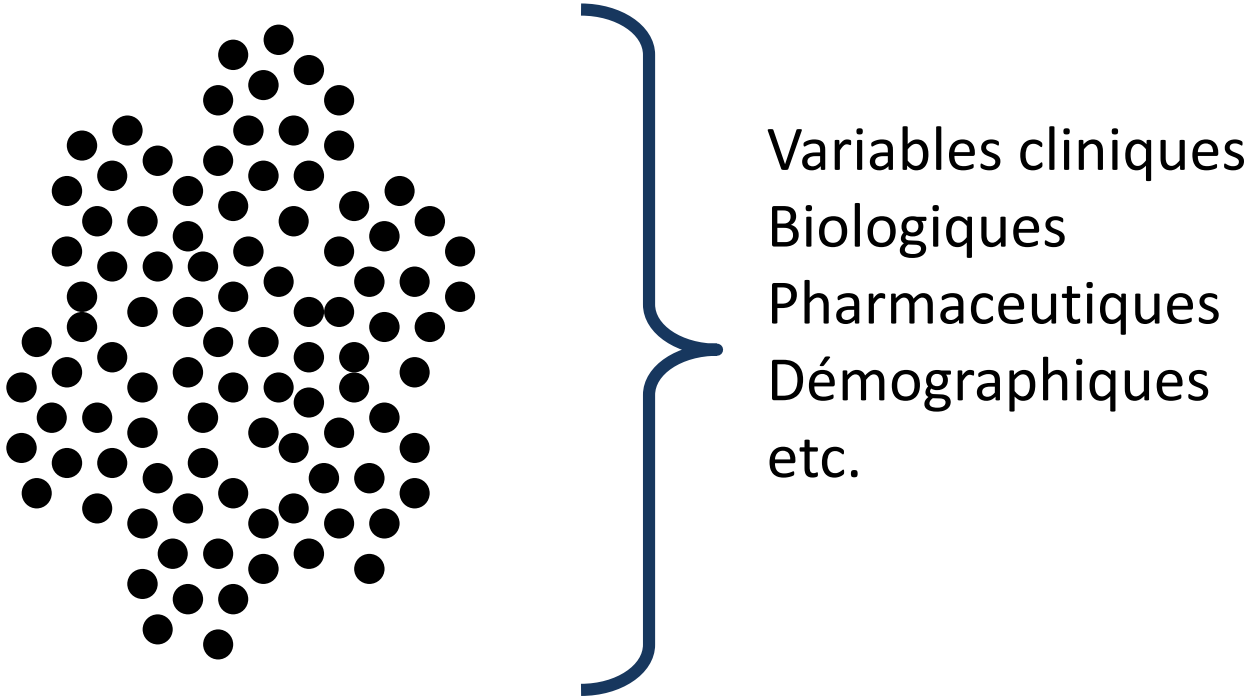
Introduction

Apprentissage non supervisé

Apprentissage supervisé

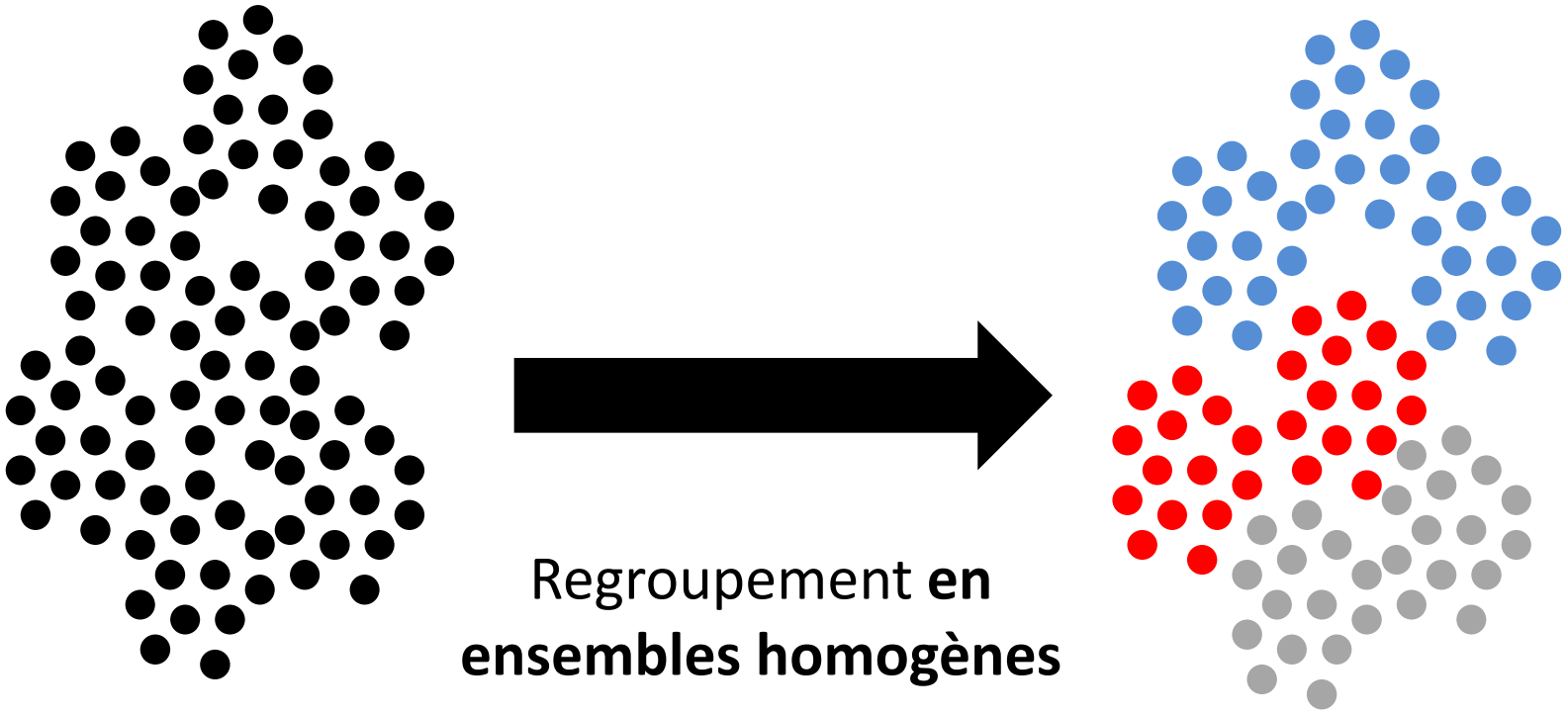
Données longitudinales

Conclusion



X patients

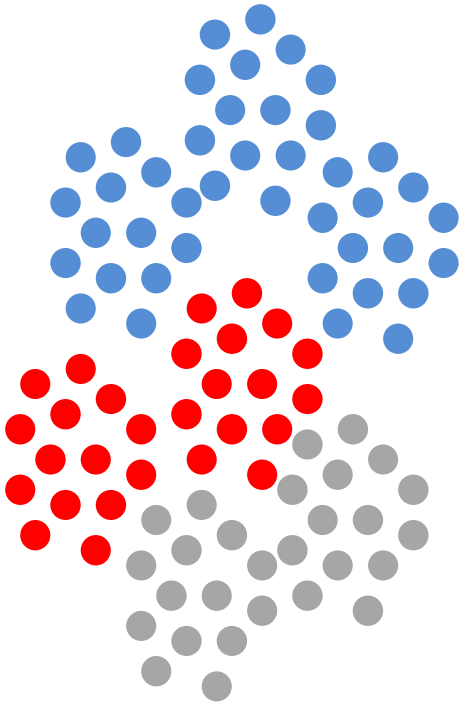
Variables cliniques
Biologiques
Pharmaceutiques
Démographiques
etc.



Regroupement **en ensembles homogènes**

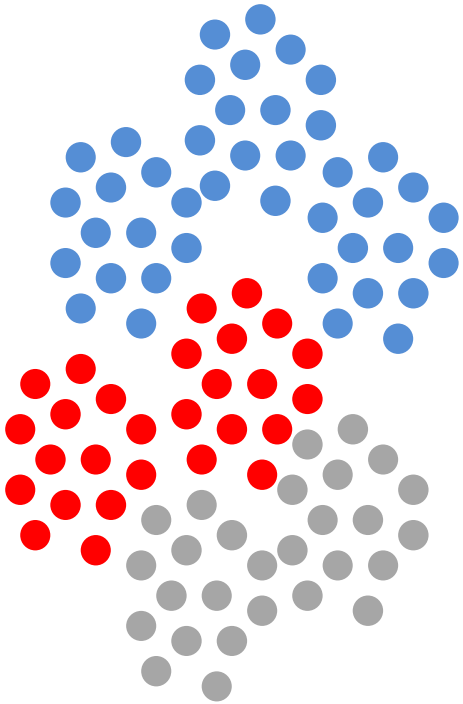
Classer les patients dans des classes ou **clusters**

Cluster analysis



Minimiser la **distance** entre deux individus semblables

Maximiser la **distance** entre deux individus différents

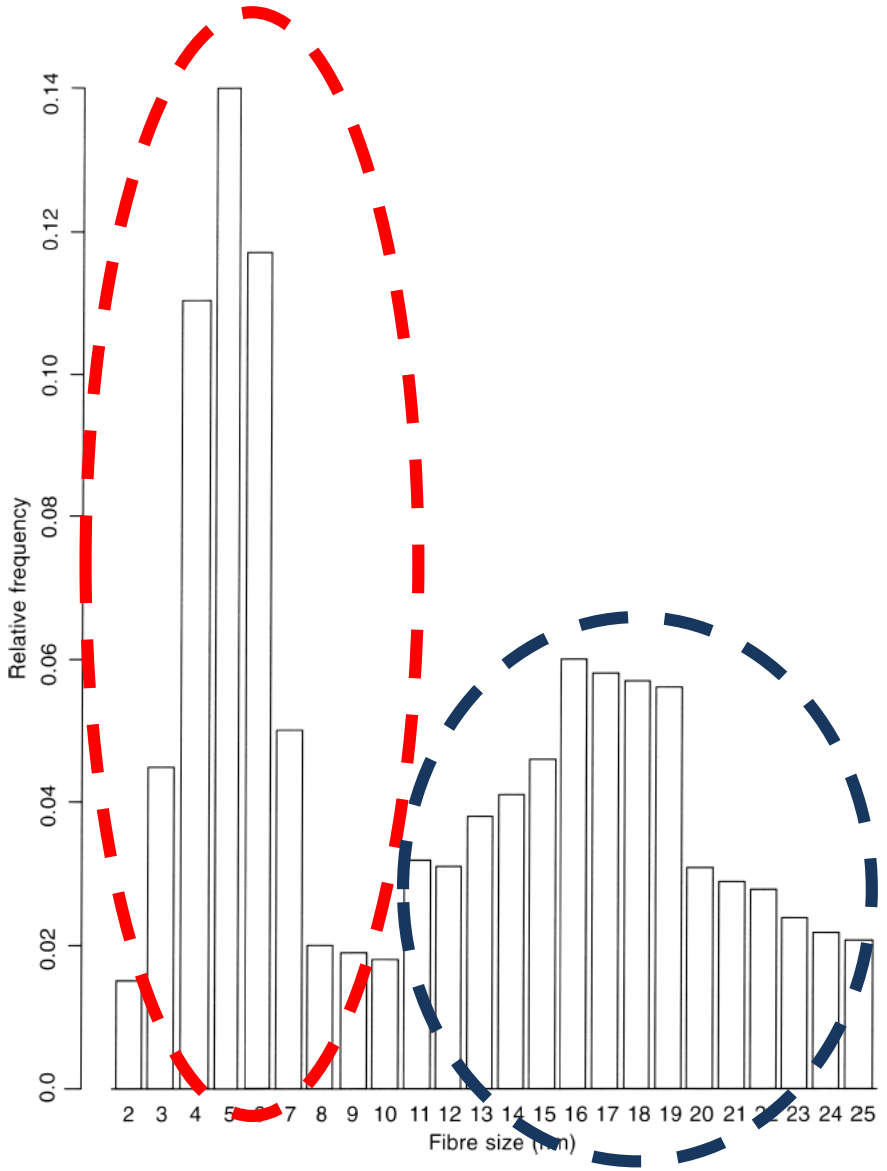


Minimiser la **distance** entre deux individus semblables

Maximiser la **distance** entre deux individus différents

Graphiques à
deux dimensions

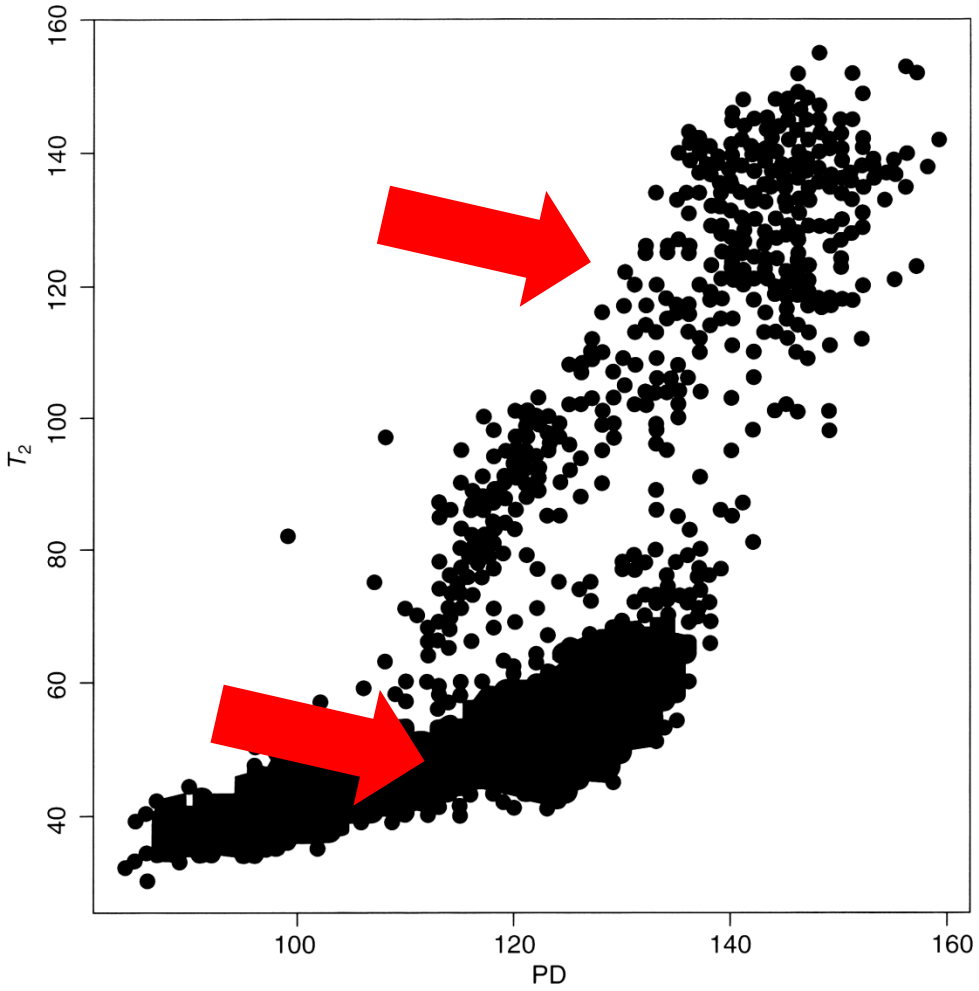
Histogrammes



Everitt et al. Cluster Analysis 5th Edition 2011

Graphiques à
deux dimensions

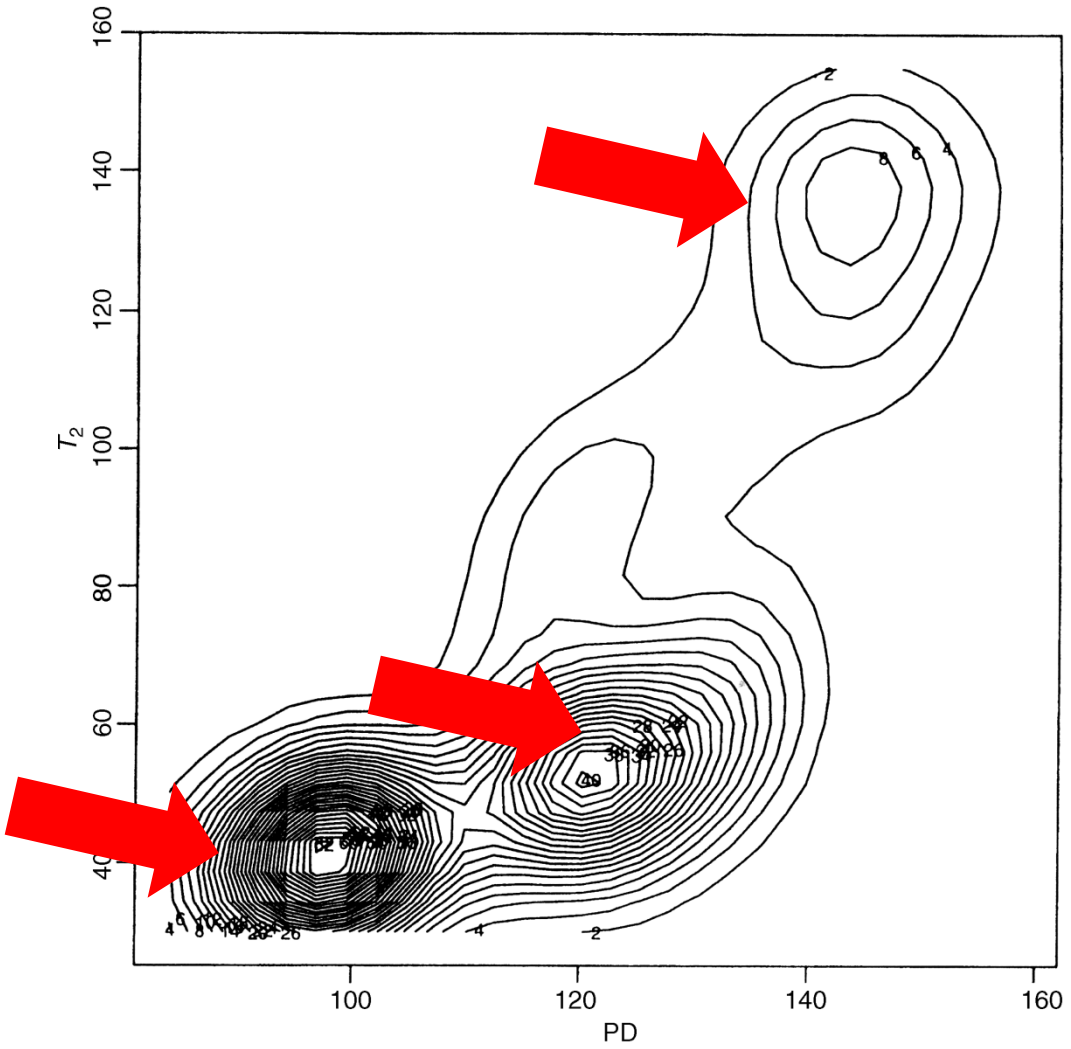
Nuages de points



Everitt et al. Cluster Analysis 5th Edition 2011

Graphiques à
deux dimensions

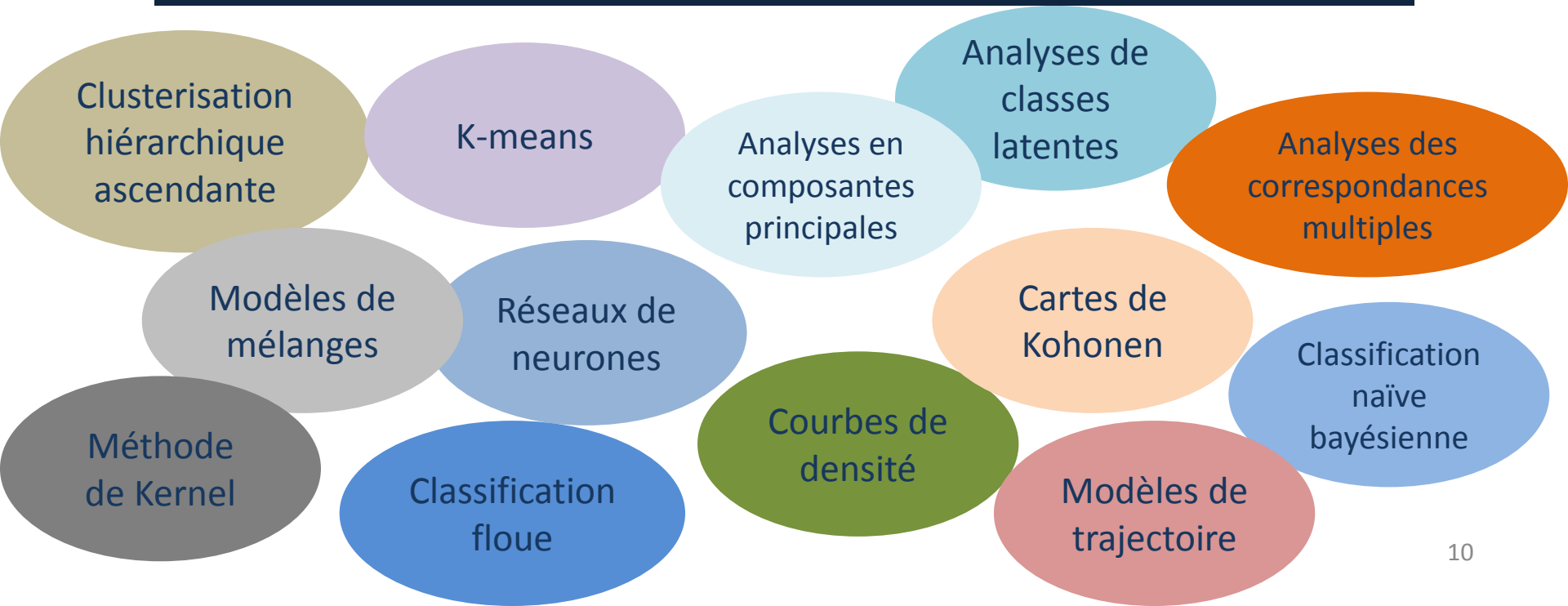
Courbes de densités



Everitt et al. Cluster Analysis 5th Edition 2011

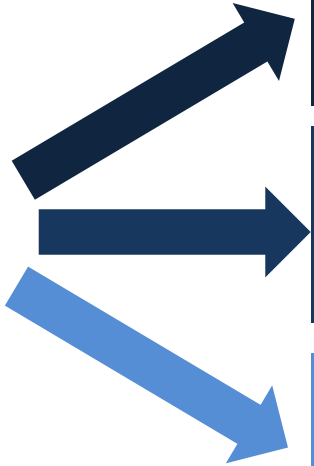
Augmentation du **nombre de variables** et du nombre **de patients**

Approches plus complexes



Augmentation du **nombre de variables** et du nombre **de patients**

Approches plus complexes



Apprentissage non supervisé :
Classification ascendante hiérarchique

Apprentissage supervisé :
Forêts aléatoires

Données longitudinales :
Modèles de trajectoire

Introduction

Apprentissage non supervisé

Apprentissage supervisé

Données longitudinales

Conclusion

Apprentissage non supervisé : classification ascendante hiérarchique

Recherche de classes
absence de données à expliquer

Apprentissage non supervisé : classification ascendante hiérarchique

Recherche de classes
absence de données à expliquer

Apprentissage non supervisé : classification ascendante hiérarchique

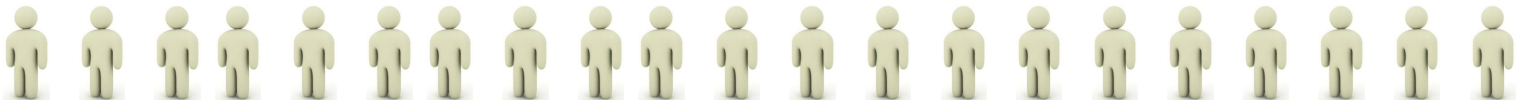
Recherche de classes
absence de données à expliquer

Structuration des données après regroupement

Apprentissage non supervisé : classification ascendante hiérarchique

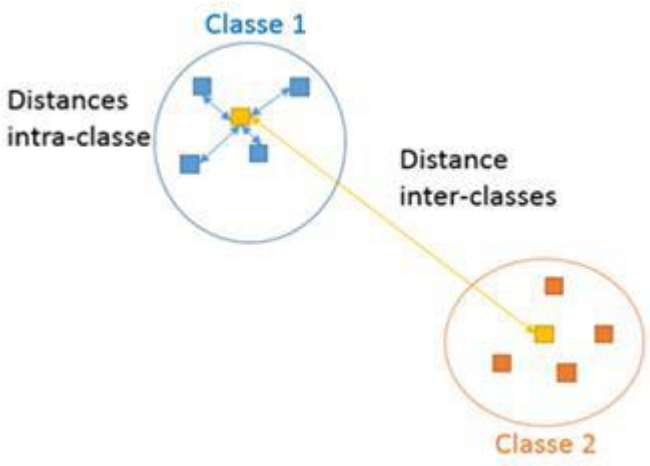
1ère étape : chaque individu est son propre cluster

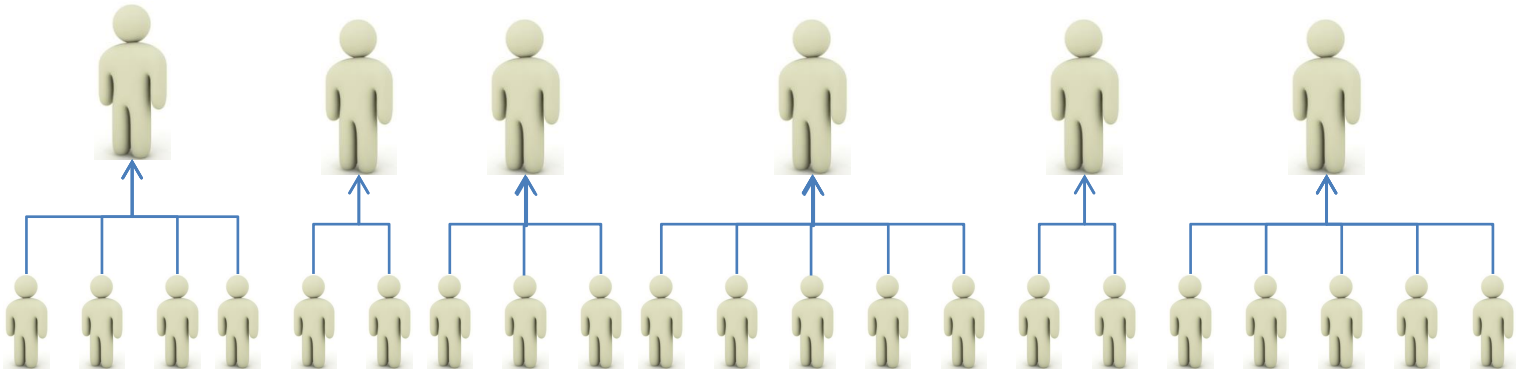
↳ Calcul de la matrice des distance des individus deux à deux



2ème étape : Regroupement des éléments les plus proches

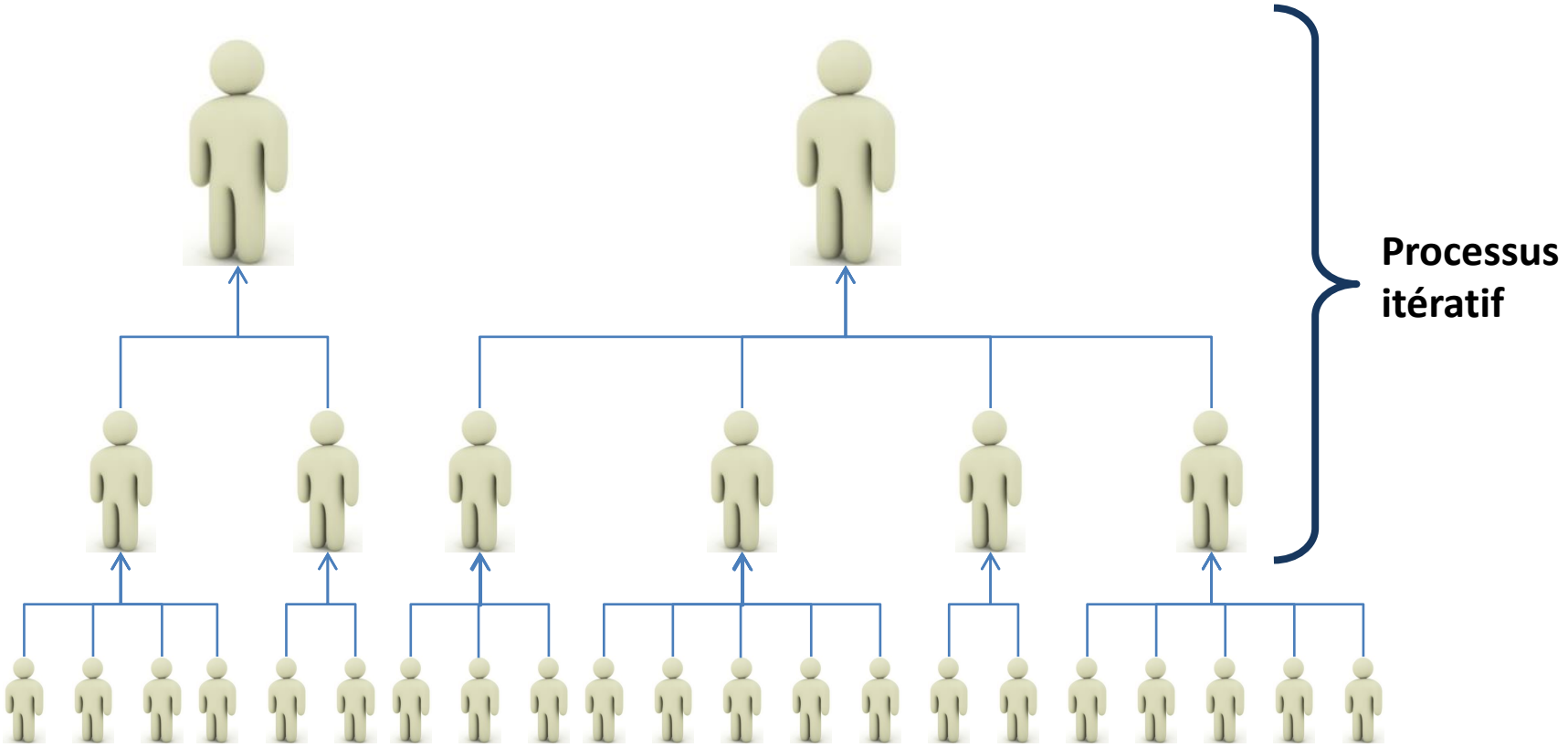
↳ Mise à jour de la matrice des distances en fonction des nouvelles classes



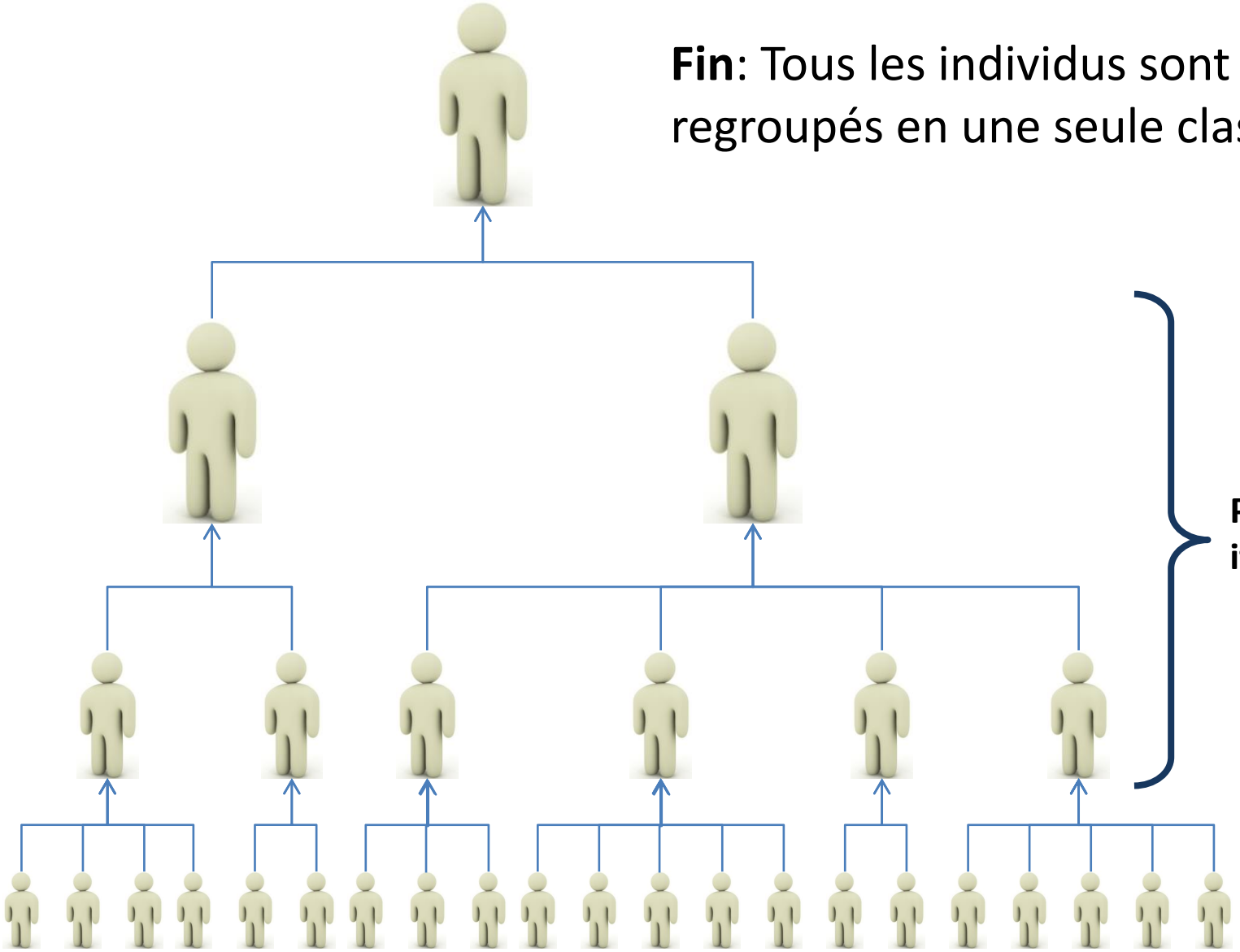


2ème étape : Regroupement des éléments les plus proches

↳ Mise à jour de la matrice des distances en fonction des nouvelles classes



Apprentissage non supervisé : classification ascendante hiérarchique



Fin: Tous les individus sont regroupés en une seule classe

Processus itératif

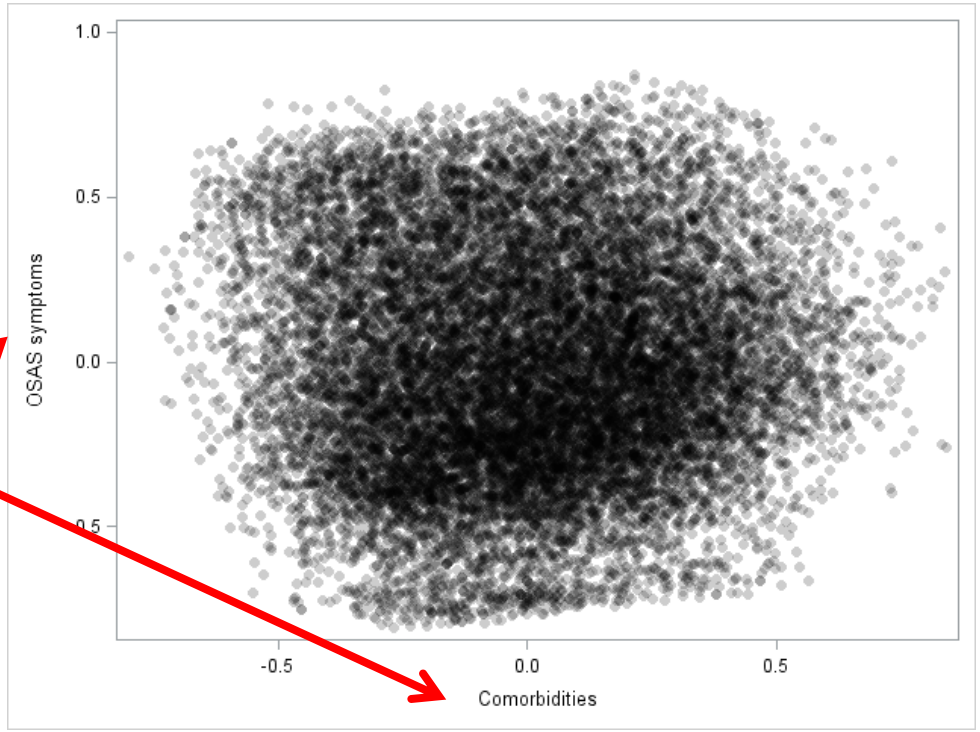
Apprentissage non supervisé : classification ascendante hiérarchique

18 263

Patients ayant un syndrôme d'apnée du sommeil

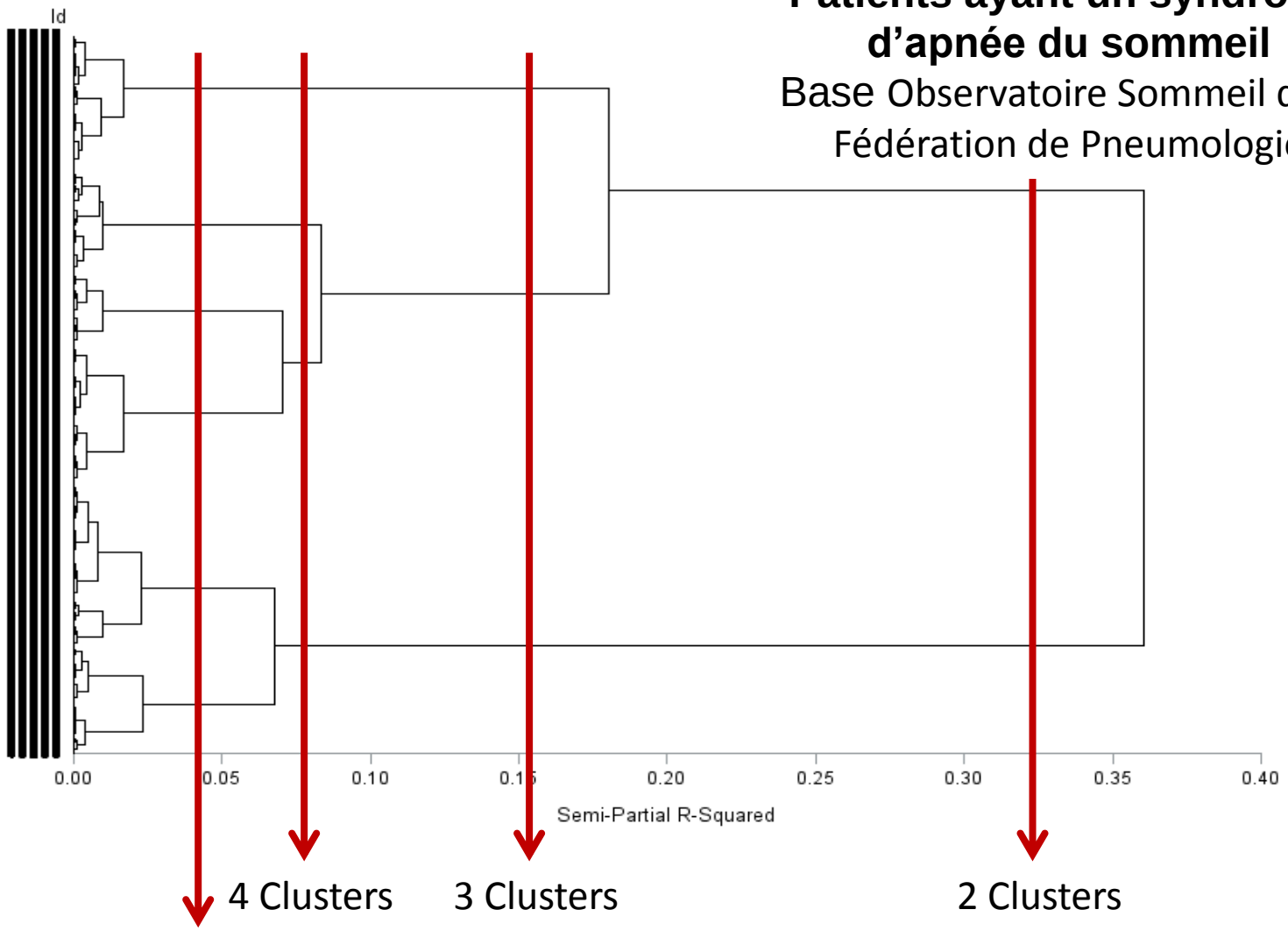
Base Observatoire Sommeil de la Fédération de Pneumologie

Réduction de l'information



Apprentissage non supervisé : classification ascendante hiérarchique

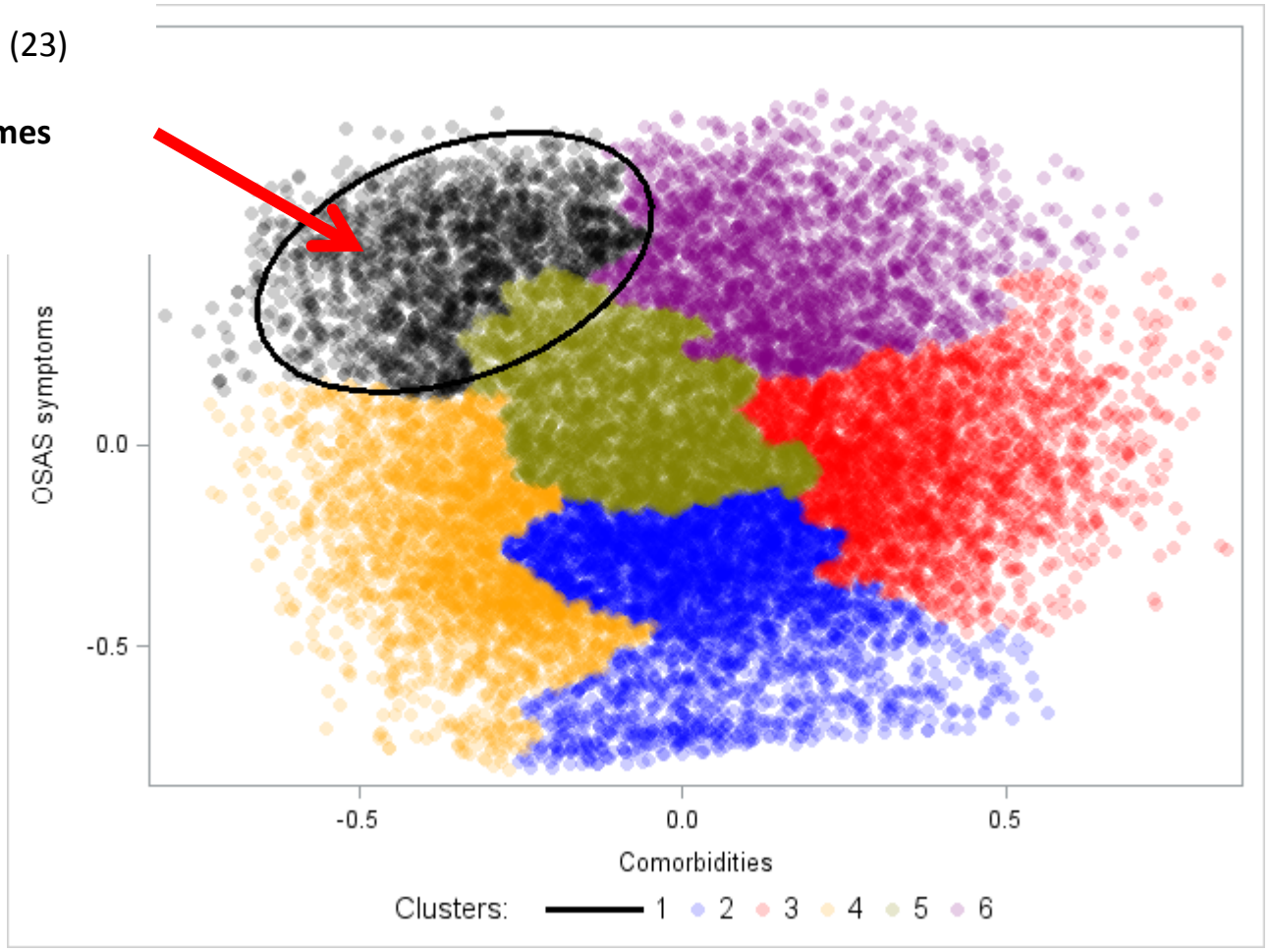
18 263
Patients ayant un syndrôme d'apnée du sommeil
Base Observatoire Sommeil de la Fédération de Pneumologie



Dendogramme

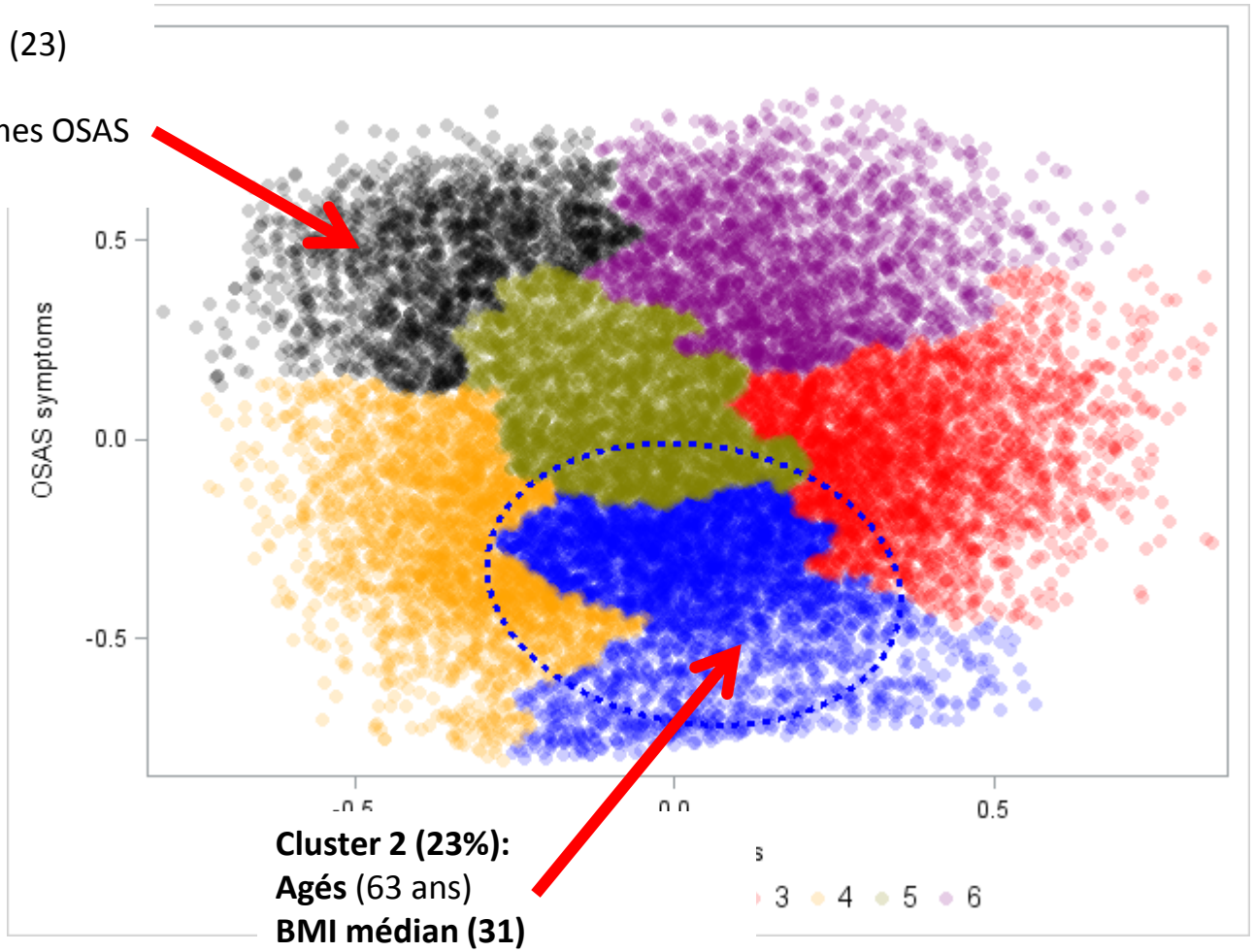
Apprentissage non supervisé : classification ascendante hiérarchique

Cluster 1 (10%):
Jeunes (48a)
BMI bas (29)
AHI bas (32) – ODI bas (23)
Peu de comorbidités
Beaucoup de symptômes
OSAS



Apprentissage non supervisé : classification ascendante hiérarchique

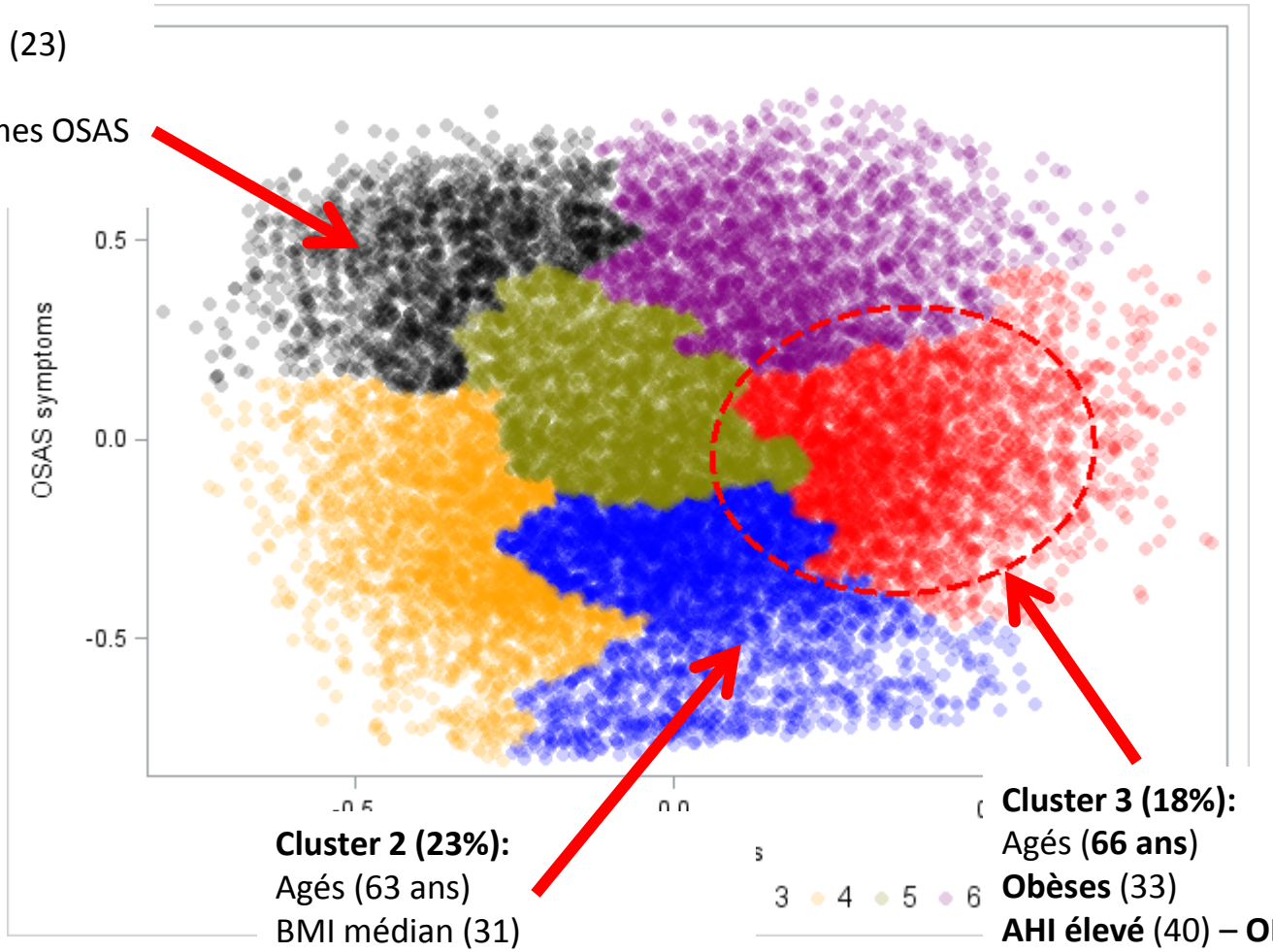
Cluster 1 (10%):
Jeunes (48a)
BMI bas (29)
AHI bas (32) – ODI bas (23)
Peu de comorbidités
Beaucoup de symptômes OSAS



Cluster 2 (23%):
Agés (63 ans)
BMI médian (31)
AHI médian (34) – ODI bas (26)
Peu de comorbidités
Peu de symptômes OSAS

Apprentissage non supervisé : classification ascendante hiérarchique

Cluster 1 (10%):
Jeunes (48a)
BMI bas (29)
AHI bas (32) – ODI bas (23)
Peu de comorbidités
Beaucoup de symptômes OSAS

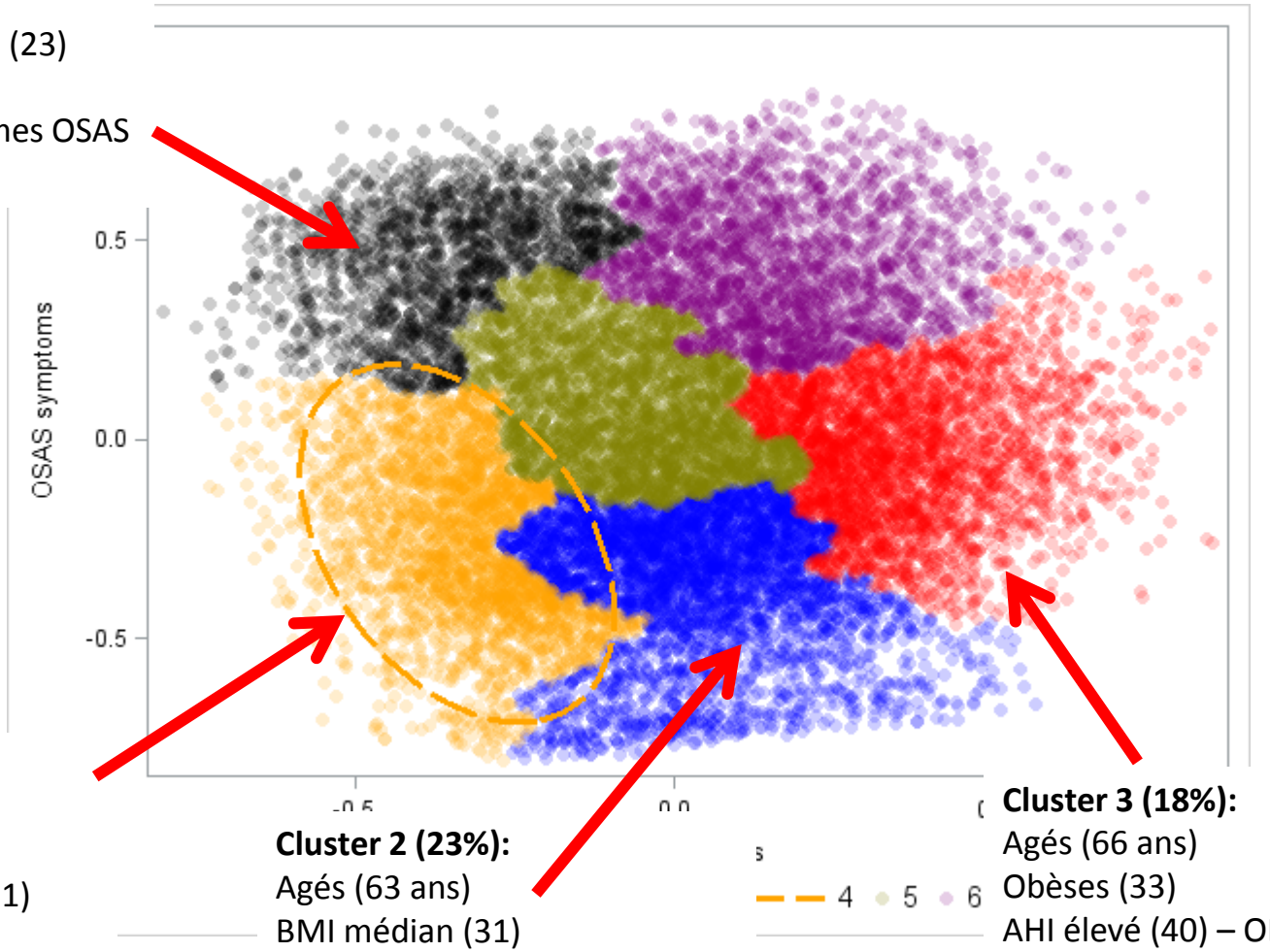


Cluster 2 (23%):
Agés (63 ans)
BMI médian (31)
AHI médian (34) – ODI bas (26)
Peu de comorbidités
Peu de symptômes OSAS

Cluster 3 (18%):
Agés (**66 ans**)
Obèses (33)
AHI élevé (40) – **ODI élevé** (33)
Nombreuses comorbidités
Peu de symptômes OSAS

Apprentissage non supervisé : classification ascendante hiérarchique

Cluster 1 (10%):
Jeunes (48a)
BMI bas (29)
AHI bas (32) – ODI bas (23)
Peu de comorbidités
Beaucoup de symptômes OSAS

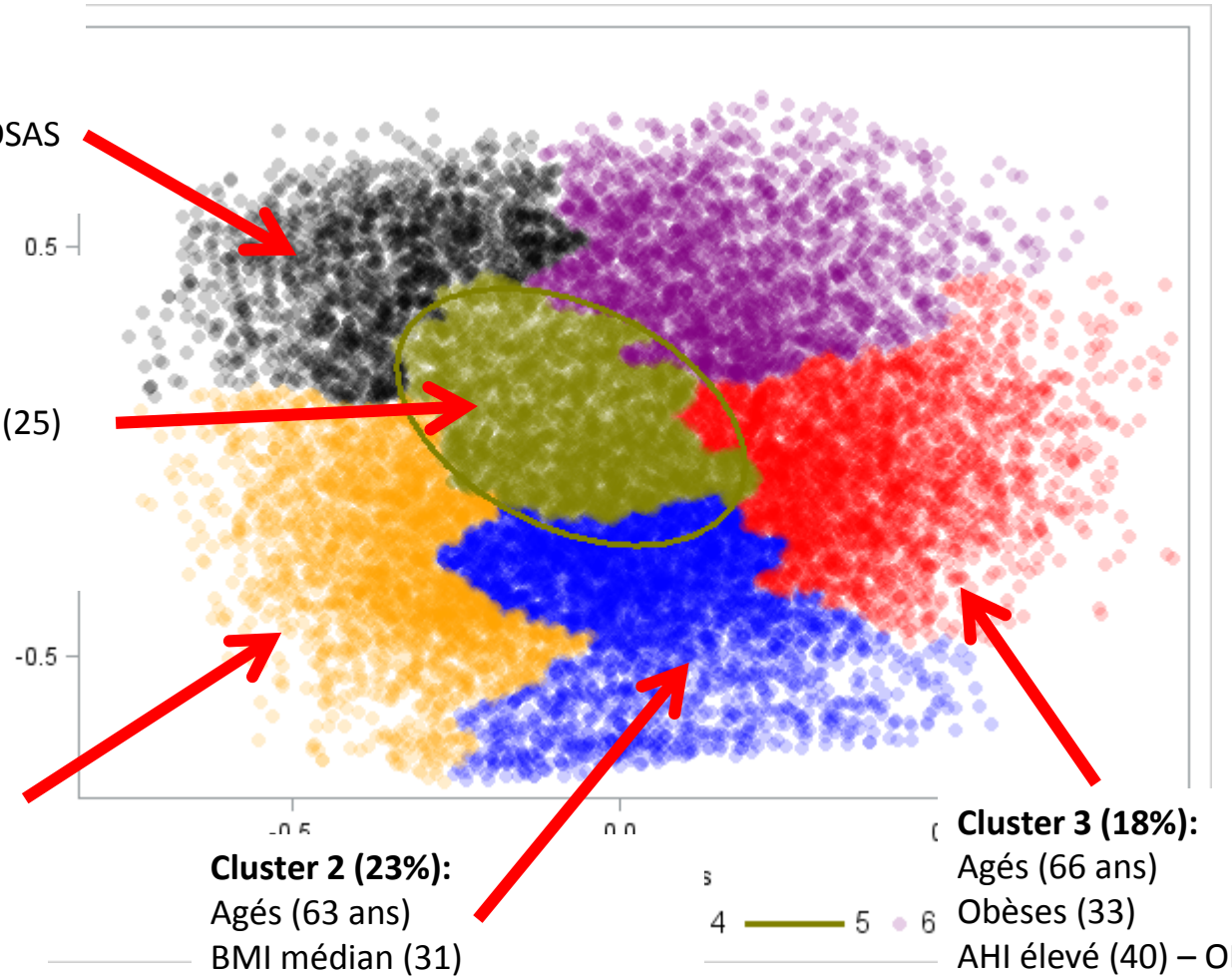


Cluster 4 (15%):
Jeunes (49 ans)
BMI bas (28)
AHI bas (31) – ODI bas (21)
Peu de comorbidités
Peu de symptômes OSAS

Cluster 2 (23%):
Agés (63 ans)
BMI médian (31)
AHI médian (34) – ODI bas (26)
Peu de comorbidités
Peu de symptômes OSAS

Cluster 3 (18%):
Agés (66 ans)
Obèses (33)
AHI élevé (40) – ODI élevé (33)
Nombreuses comorbidités
Peu de symptômes OSAS

Cluster 1 (10%):
Jeunes (48a)
BMI bas (29)
AHI bas (32) – ODI bas (23)
Peu de comorbidités
Beaucoup de symptômes OSAS



Cluster 5 (19%):
Age moyen (56 ans)
BMI moyen (31)
AHI moyen (34) – ODI moyen (25)
Peu de comorbidités
Peu de symptômes OSAS

Cluster 4 (15%):
Jeunes (49 ans)
BMI bas (28)
AHI bas (31) – ODI bas (21)
Peu de comorbidités
Peu de symptômes OSAS

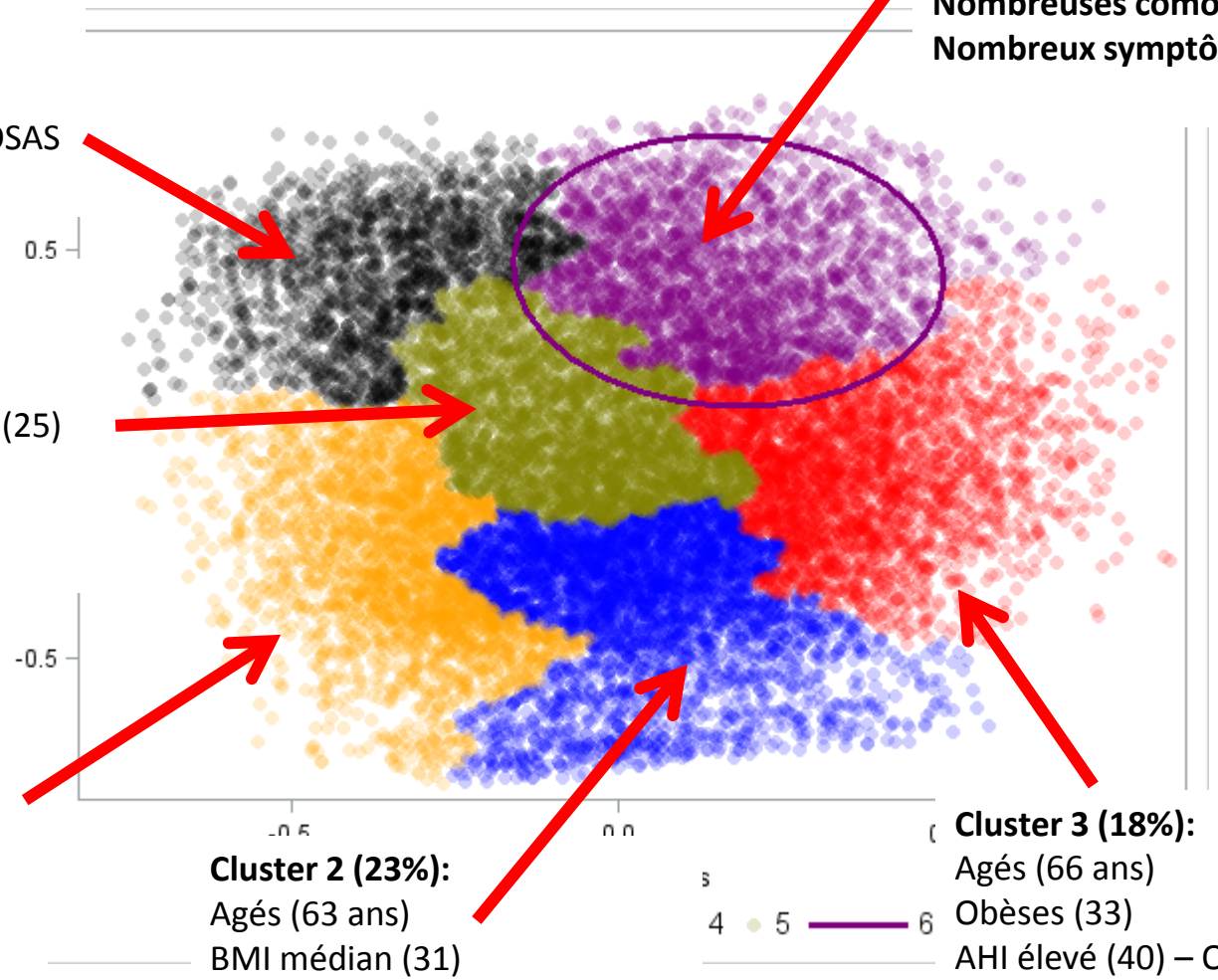
Cluster 2 (23%):
Agés (63 ans)
BMI médian (31)
AHI médian (34) – ODI bas (26)
Peu de comorbidités
Peu de symptômes OSAS

Cluster 3 (18%):
Agés (66 ans)
Obèses (33)
AHI élevé (40) – ODI élevé (33)
Nombreuses comorbidités
Peu de symptômes OSAS

Apprentissage non supervisé : classification ascendante hiérarchique

Cluster 1 (10%):
Jeunes (48a)
BMI bas (29)
AHI bas (32) – ODI bas (23)
Peu de comorbidités
Beaucoup de symptômes OSAS

Cluster 6 (14%):
Age moyen (60 ans)
Obèses (33)
AHI élevé (39) – ODI élevé (31)
Nombreuses comorbidités
Nombreux symptômes OSAS



Cluster 5 (19%):
Age moyen (56 ans)
BMI moyen (31)
AHI moyen (34) – ODI moyen (25)
Peu de comorbidités
Peu de symptômes OSAS

Cluster 4 (15%):
Jeunes (49 ans)
BMI bas (28)
AHI bas (31) – ODI bas (21)
Peu de comorbidités
Peu de symptômes OSAS

Cluster 2 (23%):
Agés (63 ans)
BMI médian (31)
AHI médian (34) – ODI bas (26)
Peu de comorbidités
Peu de symptômes OSAS

Cluster 3 (18%):
Agés (66 ans)
Obèses (33)
AHI élevé (40) – ODI élevé (33)
Nombreuses comorbidités
Peu de symptômes OSAS

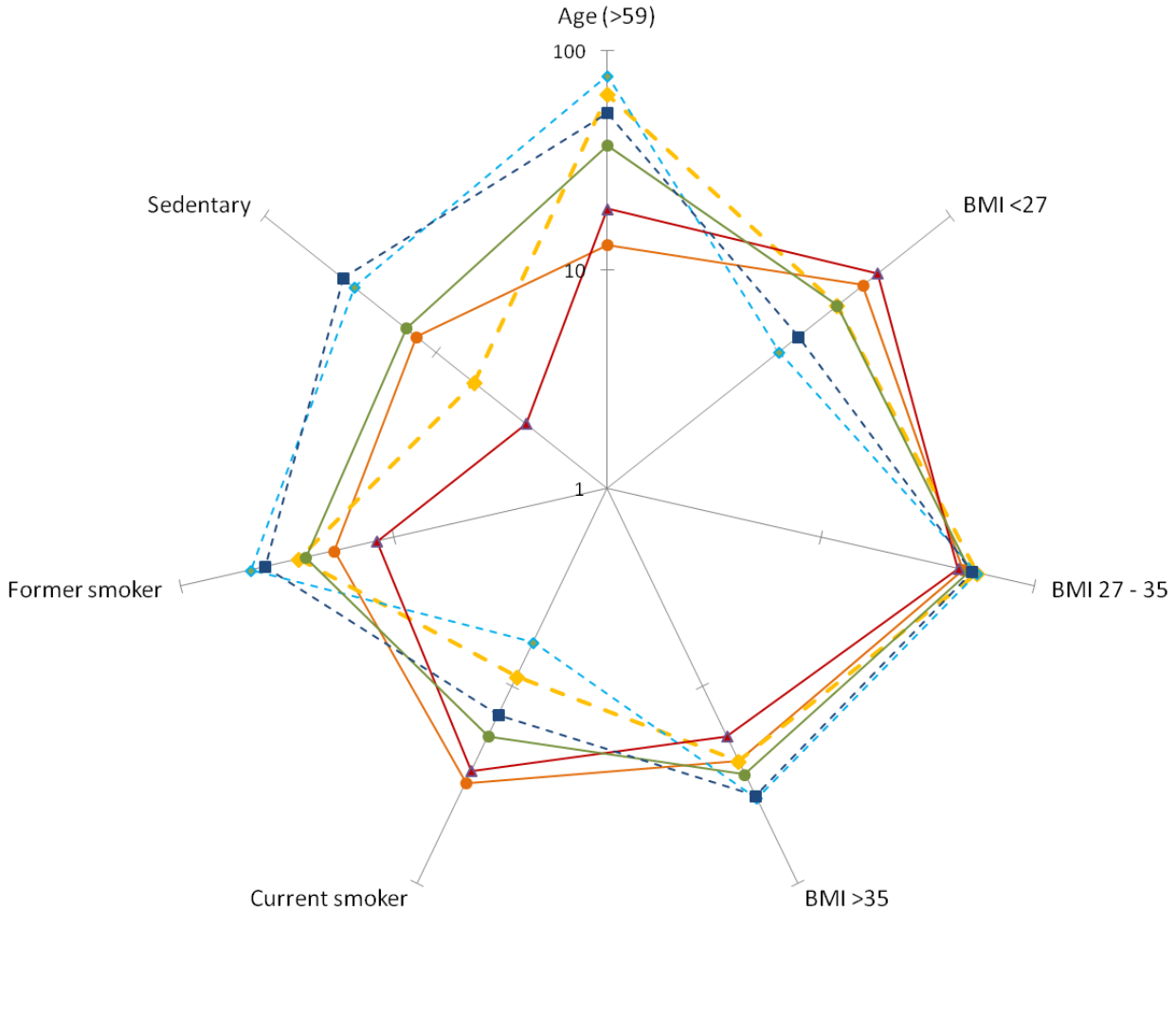
Apprentissage non supervisé : classification ascendante hiérarchique

	Age (years)	BMI (kg/m^2)	AHI (/h)	ODI (/h)	Epworth scale	Co-morbitidies	OSAS symptoms
Cluster 1	Young (48)	Low (29)	Low (31.6)	Low (23)	**High** (12)	Few or no	**Many**
Cluster 2	**Oldest** (63)	Median (31)	Median (34)	Median (26)	Low (8)	Few or no	Few or no
Cluster 3	**Oldest** (66)	**Obese** (33)	**High** (40)	**High** (33)	Low (9)	**Many**	Few or no
Cluster 4	Young (49)	Low (28)	Low (31)	Low (21)	Median (10)	Few or no	Few or no
Cluster 5	Middle age (56)	Median (31)	Median (34)	Median (25)	**High** (11)	Few or no	Few or no
Cluster 6	Middle age (60)	**Obese** (33)	**High** (39)	**High** (31)	**High** (11)	**Many**	**Many**

Bailly S, Destors M, Grillet Y, Richard P, Stach B, Vivodtzev I, Timsit JF, Levy P, Tamisier R, Pepin JL

Apprentissage non supervisé : classification ascendante hiérarchique

Facteurs environnementaux



Apprentissage non supervisé : classification ascendante hiérarchique

- Méthode basée sur les distance entre les individus
- Qualification et interprétation simple des classes créées
- Méthode descriptive
- Dépend du jeu de données : nécessité de faire une validation externe

Introduction

Apprentissage non supervisé

Apprentissage supervisé

Données longitudinales

Conclusion

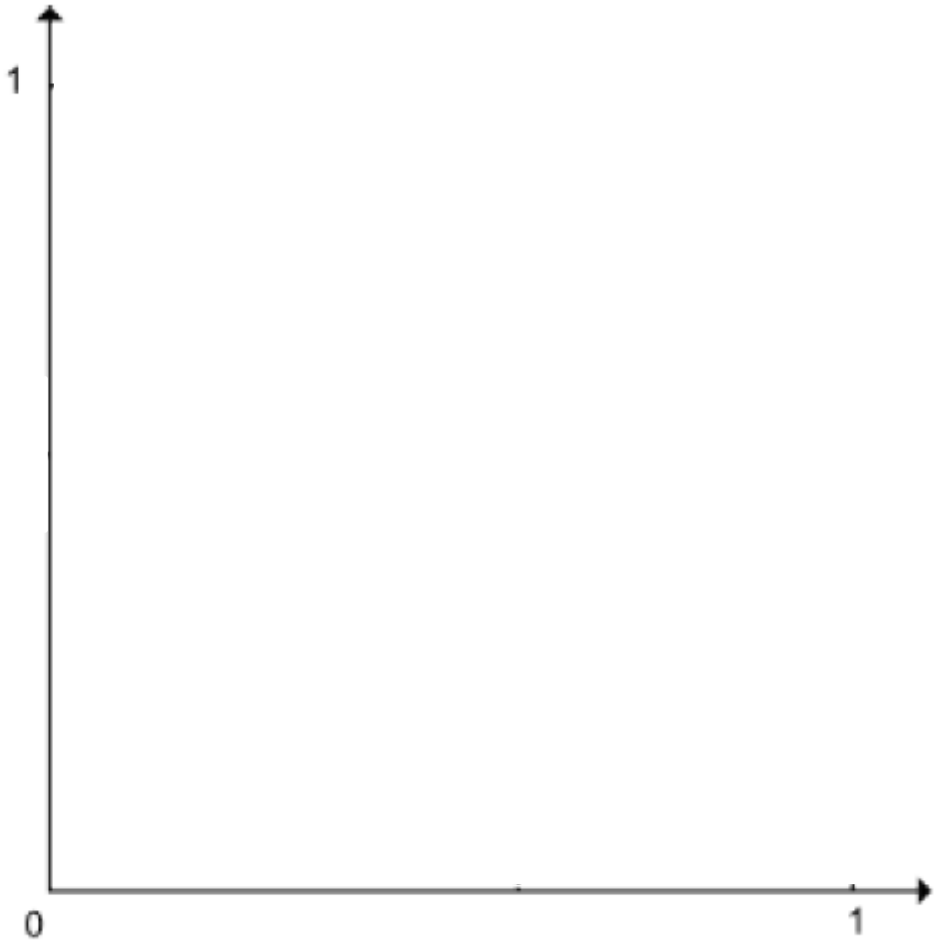
Apprentissage supervisé : Forêts aléatoires – Random forest

Arbres de décisions
Classification and regression tree (CART)
Breiman et al. 1984

→ Prédire la survenue d'une maladie en fonction des données

→ Retracent les questions à se poser pour effectuer le classement

Apprentissage supervisé : Forêts aléatoires – Random forest



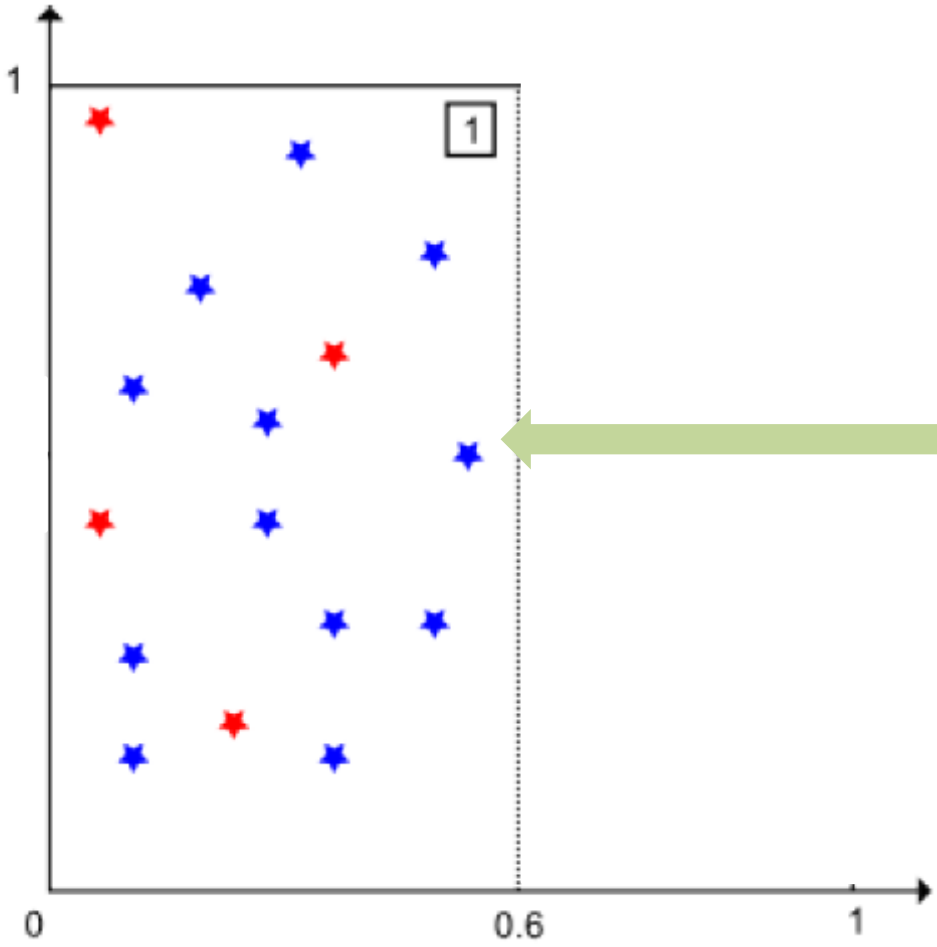
[0 ; 1]*
[0 ; 1]

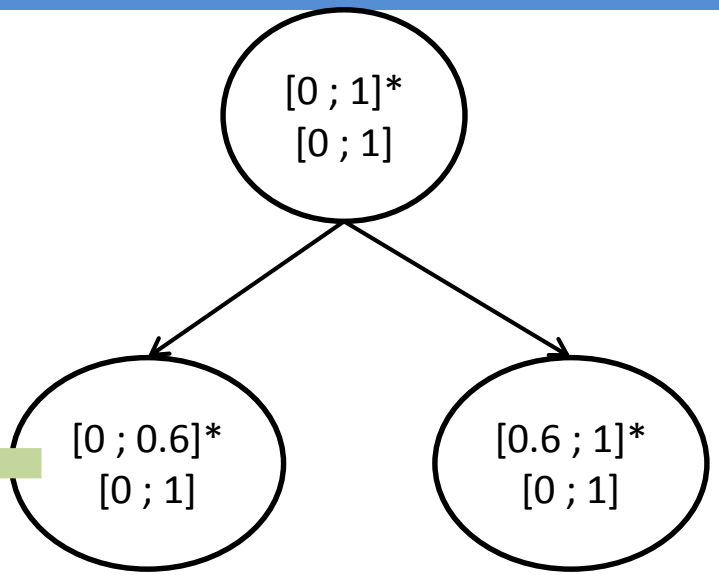
On cherche à prédire une variable binaire

★
★

En fonction de deux variables quantitatives
comprises entre 0 et 1

Apprentissage supervisé : Forêts aléatoires – Random forest





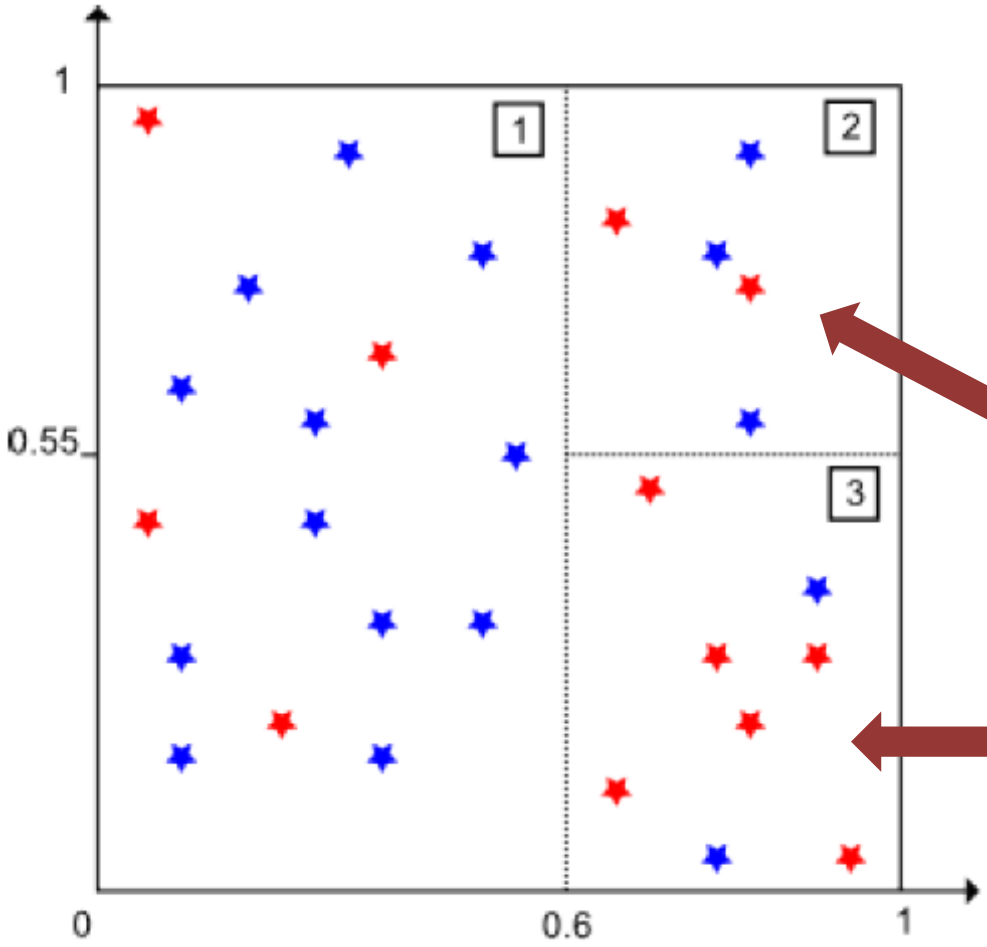
On cherche à prédire une variable binaire

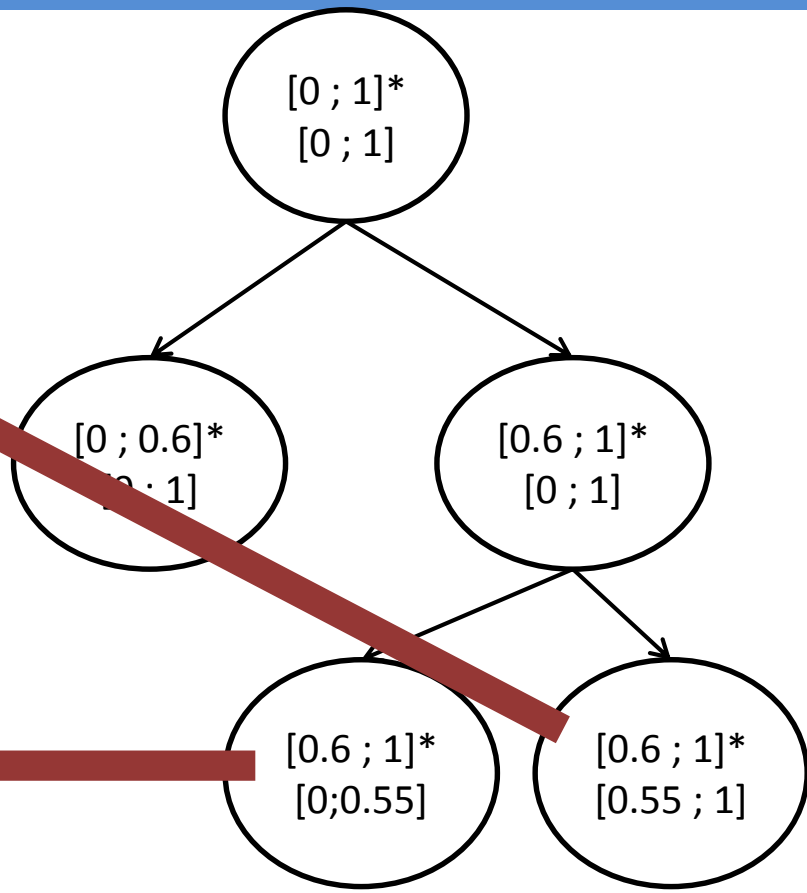
★

★

En fonction de deux variables quantitatives
comprises entre 0 et 1

Apprentissage supervisé : Forêts aléatoires – Random forest



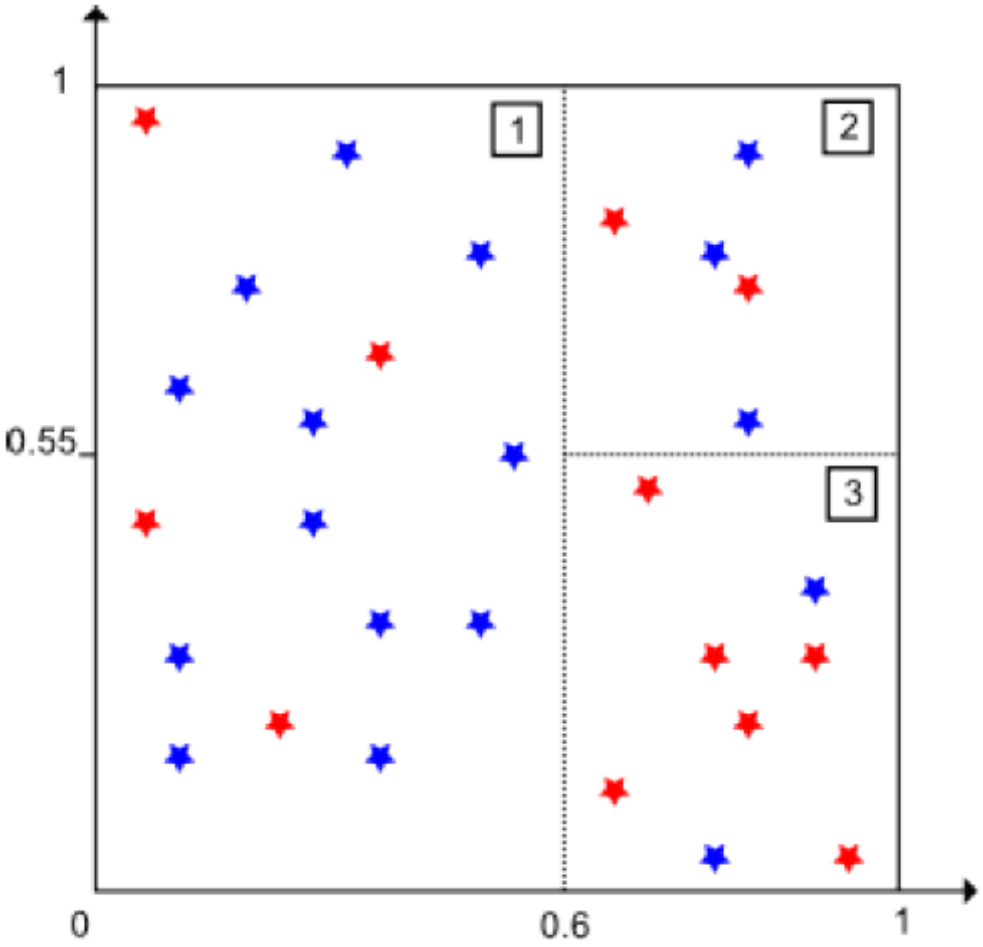


On cherche à prédire une variable binaire

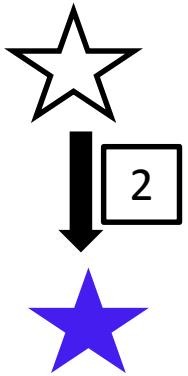
★ ★

En fonction de deux variables quantitatives
comprises entre 0 et 1

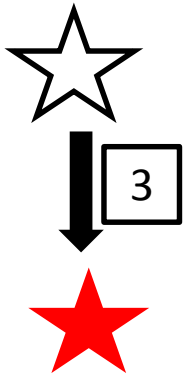
Apprentissage supervisé : Forêts aléatoires – Random forest



0.72*0.75



0.70*0.20



Apprentissage supervisé : Forêts aléatoires – Random forest

- \+ L'arbre peut être directement traduit en règles claires et interprétables
- \+ Sélection automatique des variables pertinentes
  Robuste par rapport aux variables redondantes
- \+ Robuste par rapport aux variables aberantes
  Possibilité de prendre en compte les données manquantes
- \+ Traite rapidement de très grandes bases de données
- \+ Possibilité d'intervenir dans la construction de l'arbre
- \- Sont fortement dépendant de l'échantillon initial

Random forest : Breiman 2001

Introduction d'une partie aléatoire dans la construction des arbres de décision

Apprentissage supervisé : Forêts aléatoires – Random forest

Classification and regression Tree

↓

Un seul arbre vote pour classer un nouveau patient malade ou non malade

Apprentissage supervisé : Forêts aléatoires – Random forest

Random forest

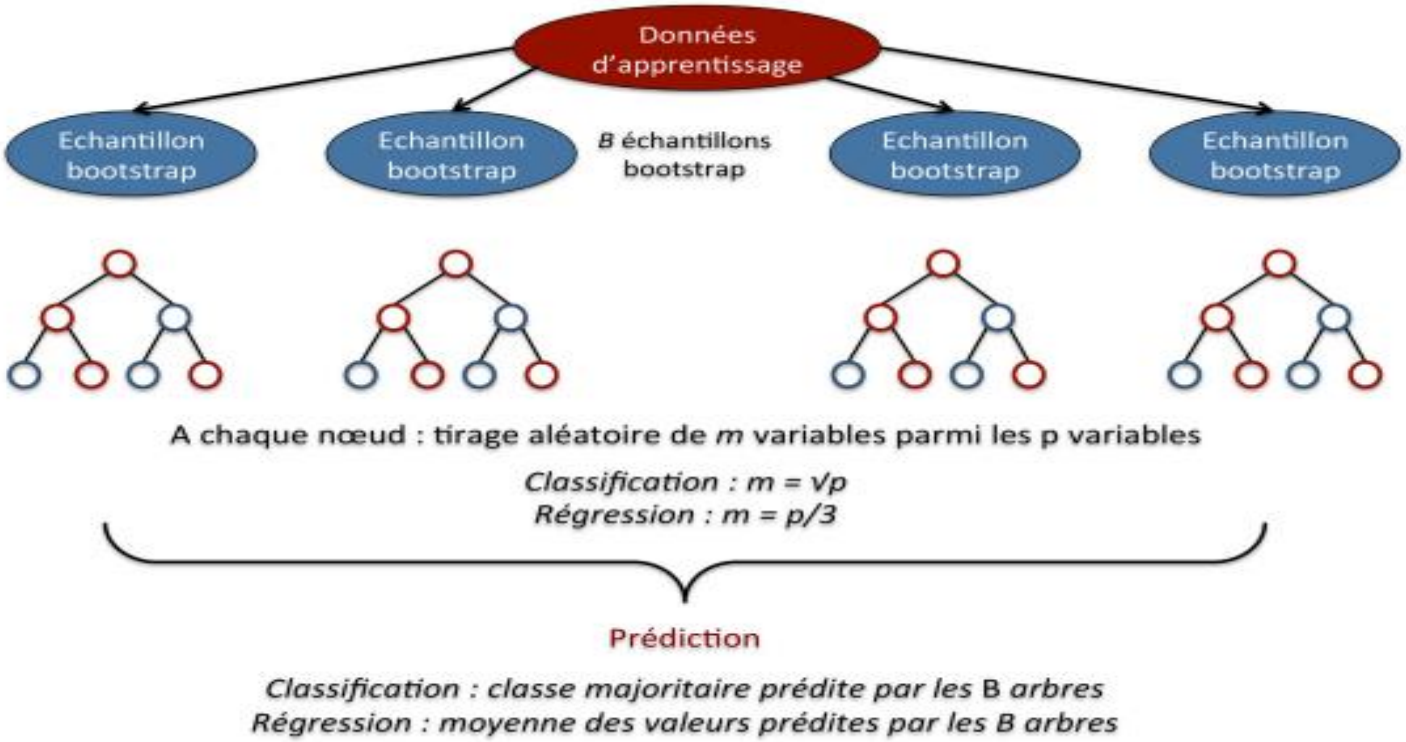
↓

La forêt aléatoire vote pour classer un nouveau patient malade ou non malade

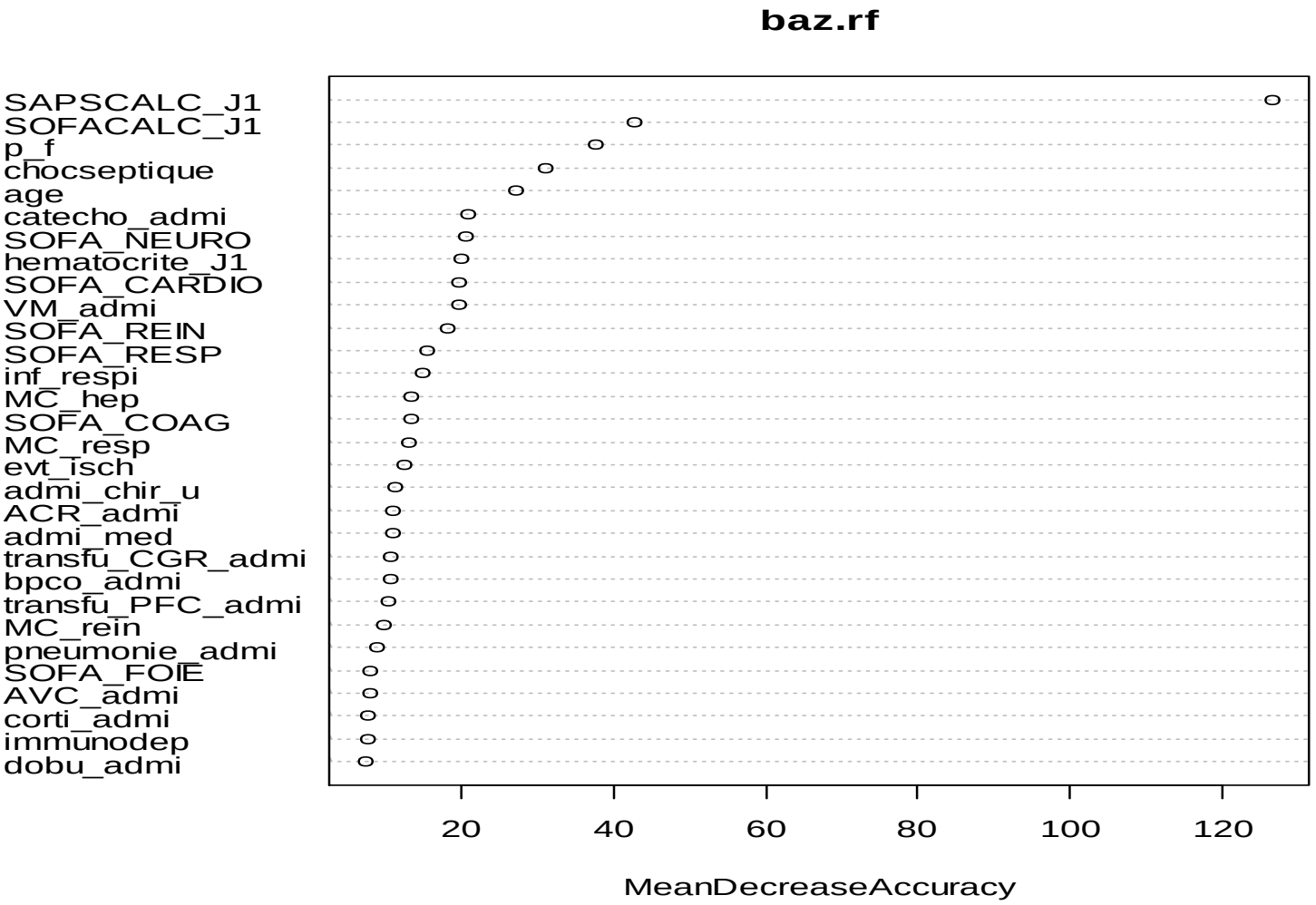
Apprentissage supervisé : Forêts aléatoires – Random forest

Random forest

La forêt aléatoire vote pour classer un nouveau patient malade ou non malade



Apprentissage supervisé : Forêts aléatoires – Random forest



Décès à J30 chez les patients admis en réanimation en choc septique ou sepsis sévère en fonction des variables à l'admission – Base OUTCOMEREA – Dr Claire Dupuis

Apprentissage supervisé : Forêts aléatoires – Random forest

An Empirical Comparison of Supervised Learning Algorithms

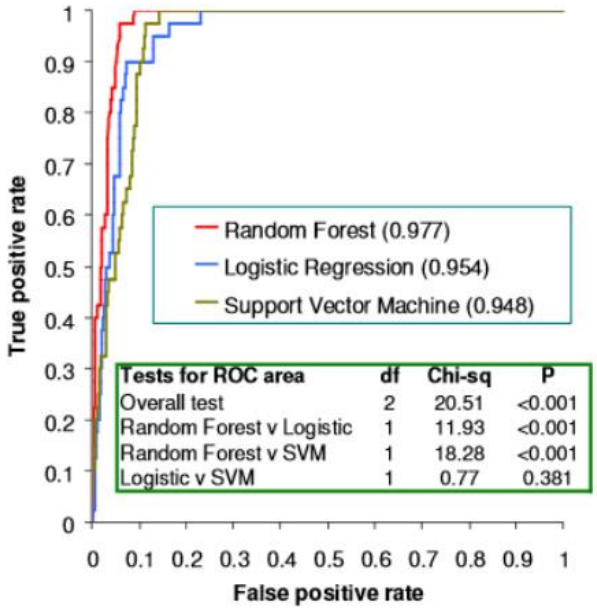
Rich Caruana CARUANA@CS.CORNELL.EDU
Alexandru Niculescu-Mizil ALEXN@CS.CORNELL.EDU
Department of Computer Science, Cornell University, Ithaca, NY 14853 USA

Published in final edited form as:
Int J Appl Sci Technol. 2012 August ; 2(7): 268–.

A Comparison of Logistic Regression, Logic Regression, Classification Tree, and Random Forests to Identify Effective Gene-Gene and Gene-Environmental Interactions

Wonsuk Yoo,

En faveur des forêts aléatoires



Vickers et al. Advanced Research Techniques Forum of the American Marketing Association - 2007

En faveur des forêts aléatoires

MAIS

Problème en **cas de nombre de variables pertinentes très faibles**, dans l'absolu et relativement au nombre total de variables

Manque de lisibilité du méta-modèle

Complexité de la mise en oeuvre

Introduction

Apprentissage non supervisé

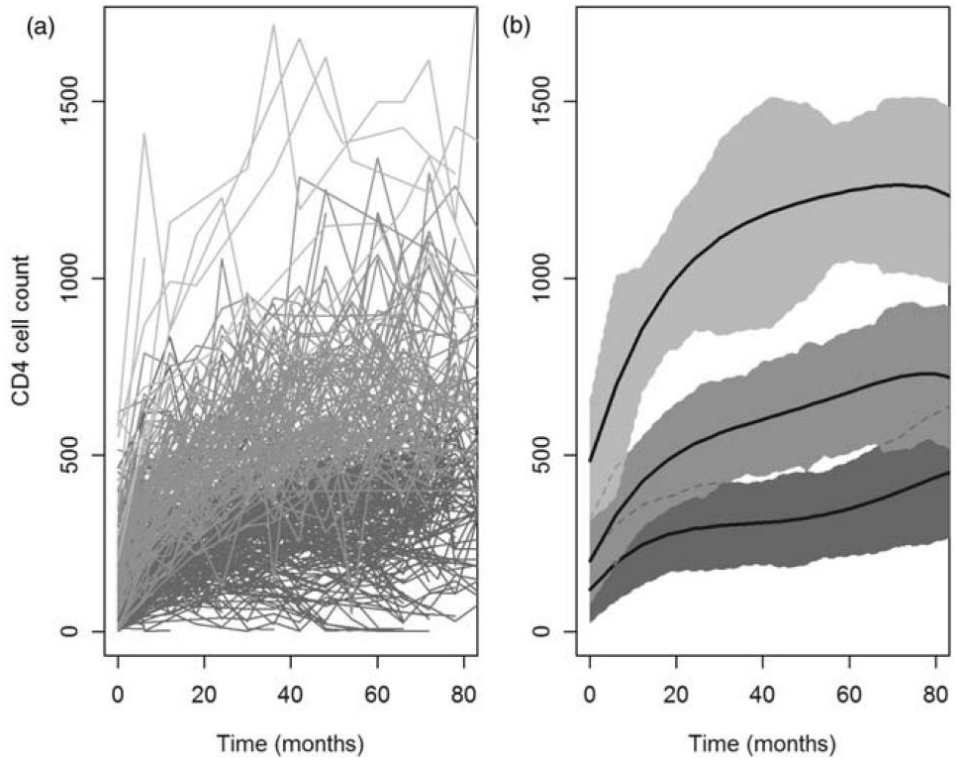
Apprentissage supervisé

Données longitudinales

Conclusion

Données longitudinales : Modèles de trajectoire

Group-based trajectory modeling / Latent Growth Modeling
les patients sont regroupés selon leur probabilité d'appartenir à une trajectoire moyenne

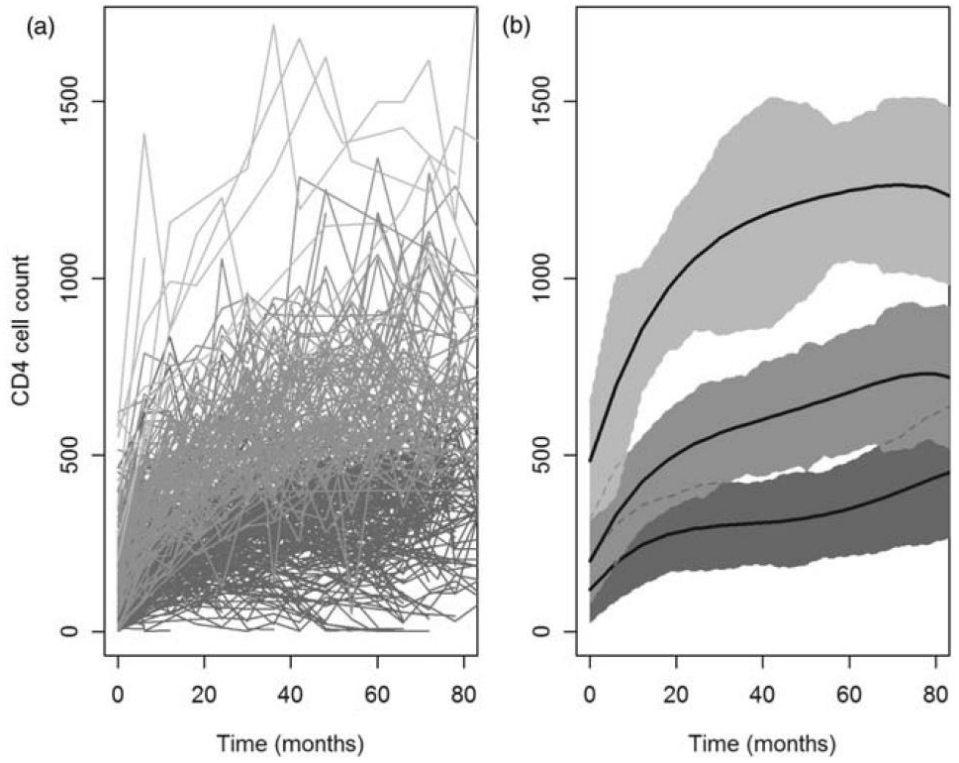


Comptage des lymphocytes CD4 chez des patients VIH sous traitement antirétroviral

Elsensohn et al., Stat Methods Med Res 2013

Données longitudinales : Modèles de trajectoire

Group-based trajectory modeling / Latent Growth Modeling
les patients sont regroupés selon leur probabilité d'appartenir à une trajectoire moyenne



Exploitation des **données longitudinales**

Comptage des lymphocytes CD4 chez des patients VIH sous traitement antirétroviral

Elsensohn et al., Stat Methods Med Res 2013

Introduction

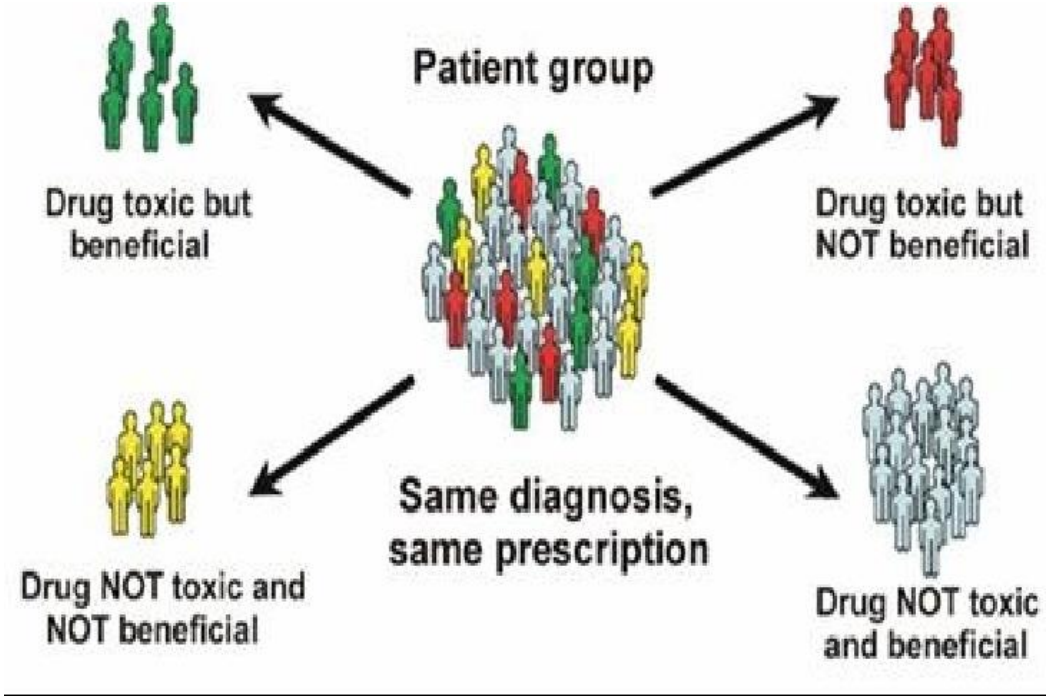
Apprentissage non supervisé

Apprentissage supervisé

Données longitudinales

Conclusion

Détermination de sous-groupes homogènes
Identification de phénotypes de patients



OUTCOMEREA

lame UMR 1137

Merci pour votre attention