Science des données III : cours 2

Classification supervisée (partie 2)

Philippe Grosjean & Guyliann Engels

Université de Mons, Belgique
Laboratoire d'Écologie numérique des Milieux aquatiques

http://biodatascience-course.sciviews.org
sdd@sciviews.org

UMONS

Validation croisée et algorithmes

UMONS

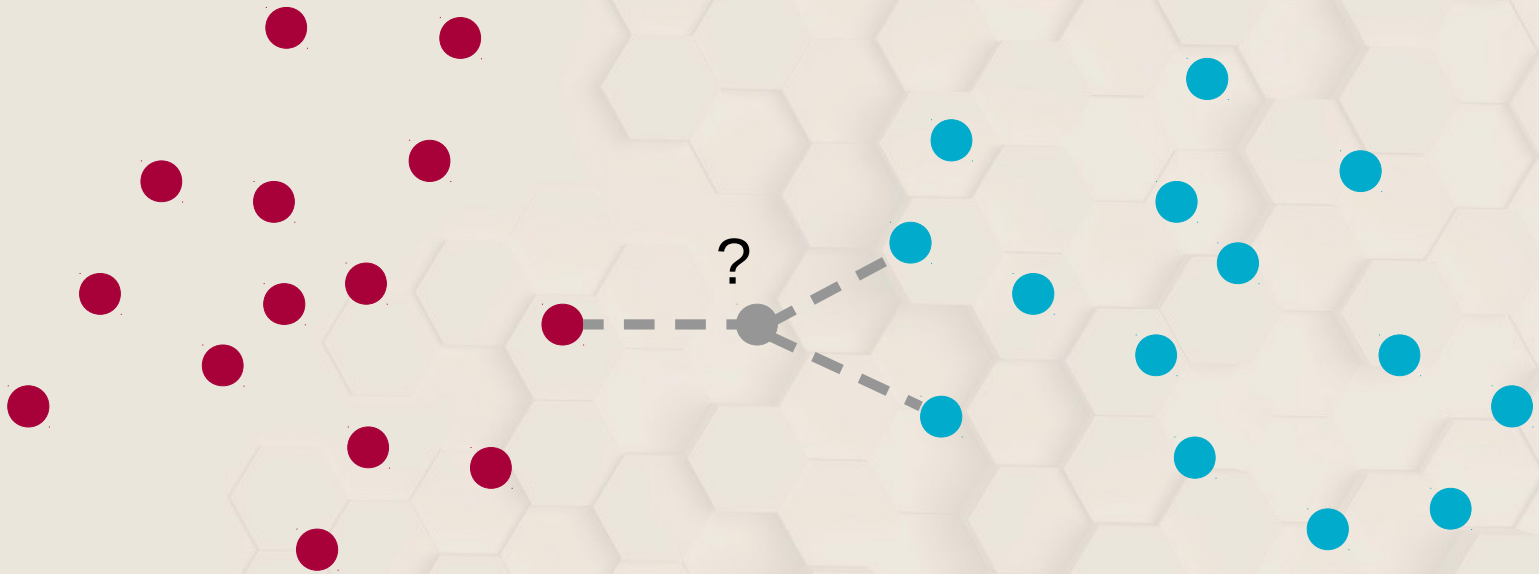
Objectifs du cours

- Utiliser la validation croisée
- Connaitre d'autres algorithmes de classification supervisée : les k plus proches voisins et l'apprentissage par quantification vectorielle

UMONS

Méthode des k plus proches voisins

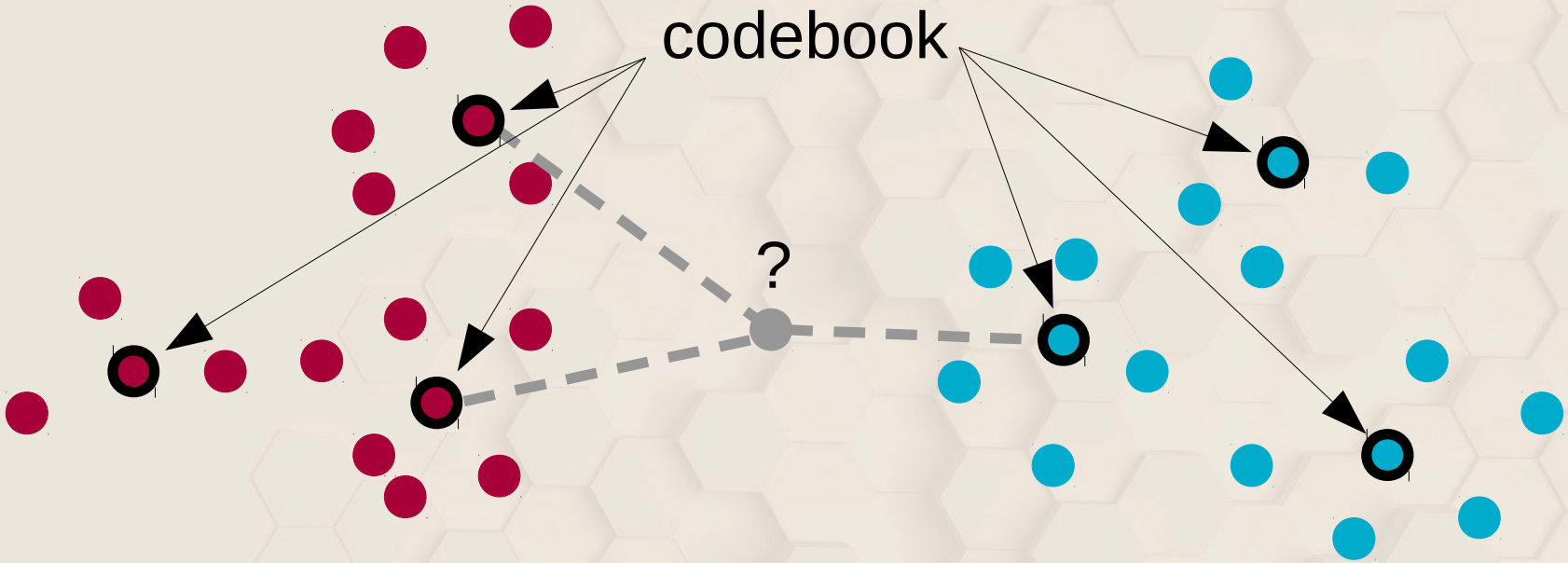
"k-Nearest Neighbours" (k-NN). Rapide à calculer, mais performances moyennes.
Distance de Malahanobis.
(exemple en utilisant k = *3 voisins) :*



UMONS

Apprentissage par quantification vectorielle

k-NN compare à tous les individus du set d'apprentissage. LVQ ("Learning Vector Quantization") crée des "portraits robots" pour chaque classe (= codebook) et compare uniquement à ces derniers.



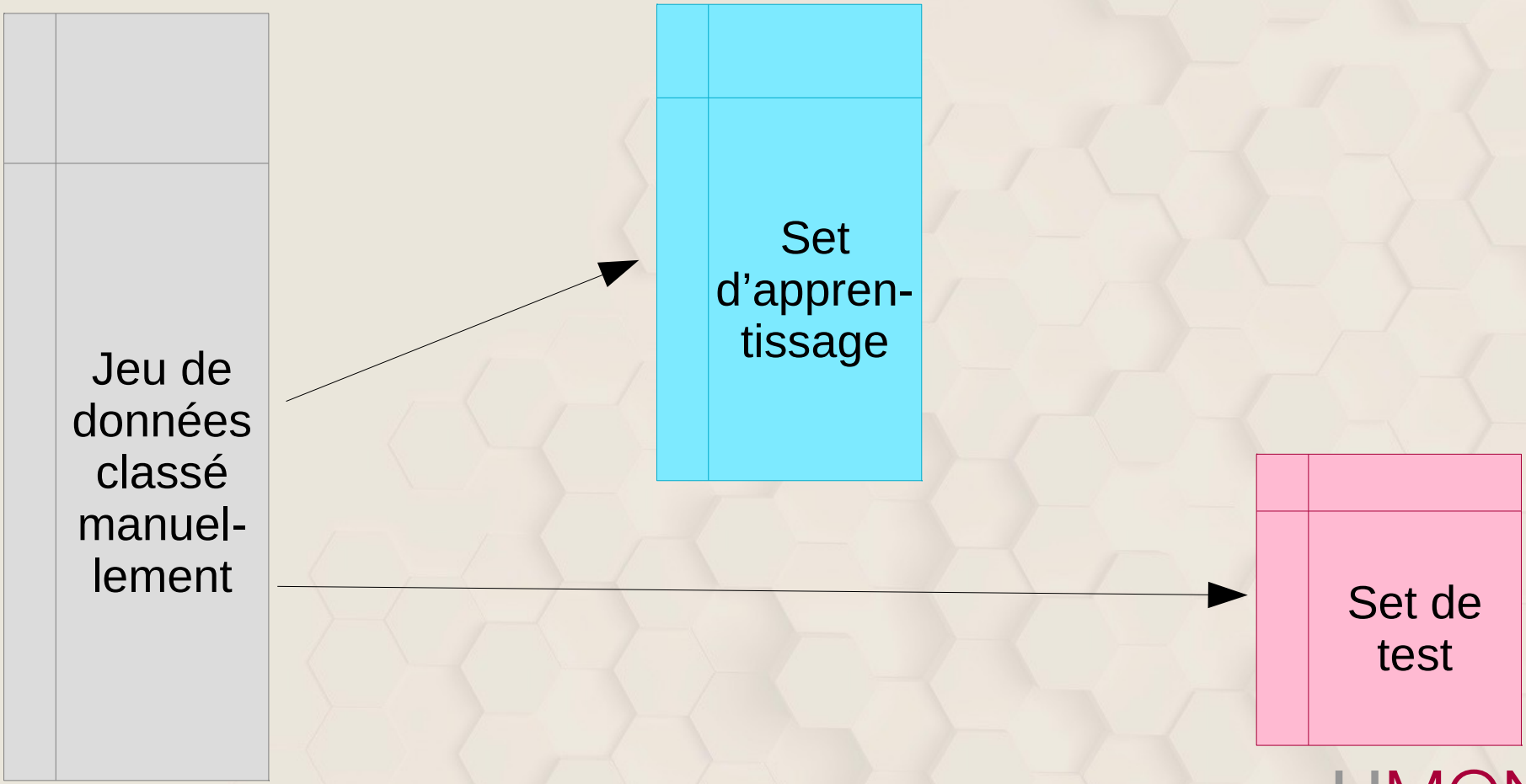
UMONS

Validation croisée

UMONS

La validation croisée

L'obligation de séparer ses précieuses données en set d'apprentissage et set de test est très contraignante.



UMONS

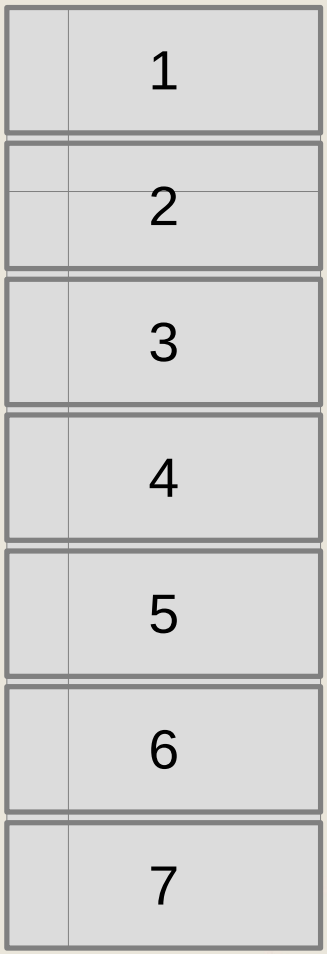
La validation croisée (2)

- La **validation croisée** est une technique générale qui permet d'utiliser toutes les données en apprentissage et en test, mais **pas simultanément**
- L'évaluation des performances en non biaisée

UMONS

La validation croisée - principe

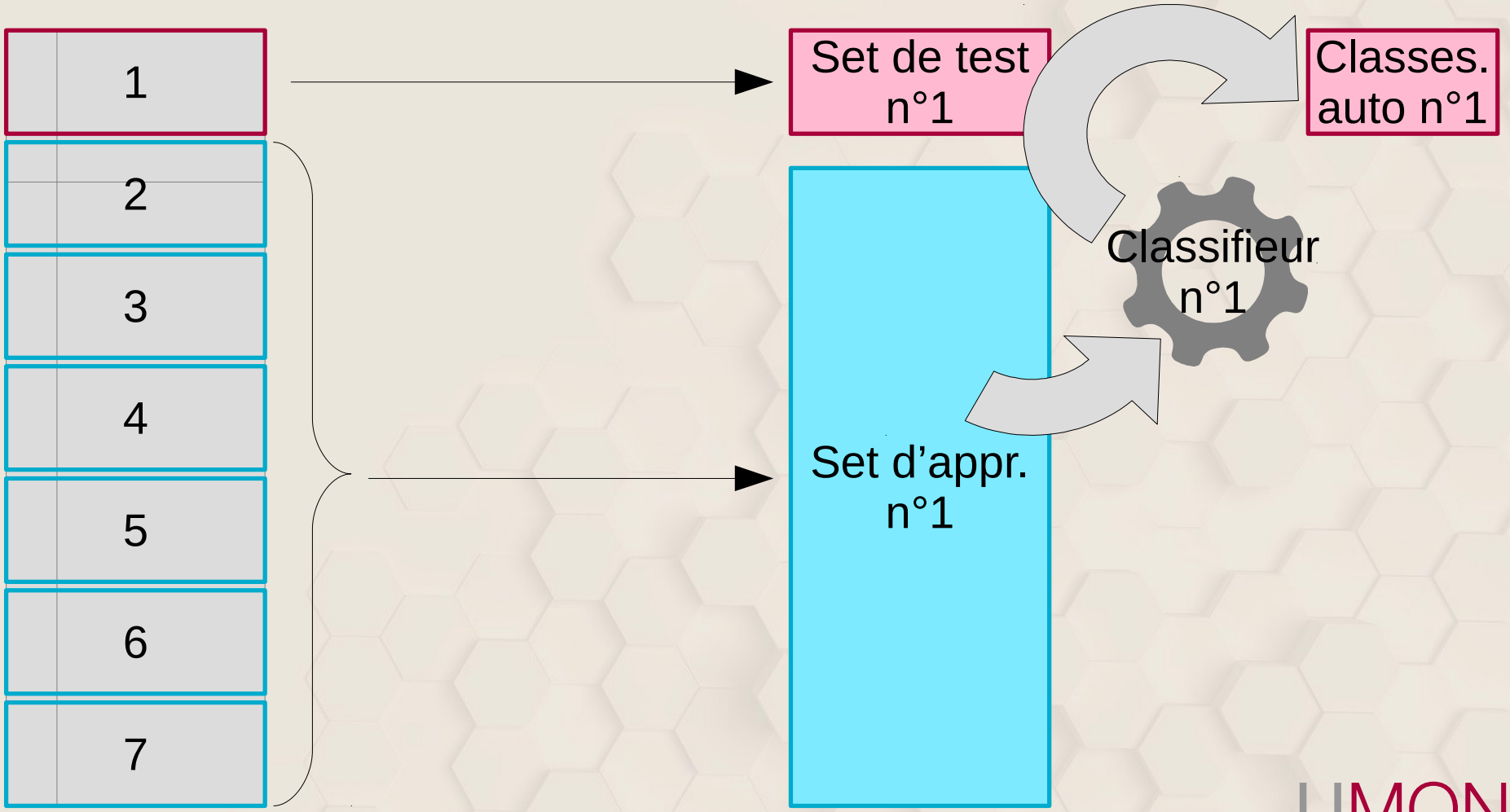
Séparation aléatoire en k sous-ensembles de mêmes effectifs
(exemple : jeud de données classé manuellement séparé en $k = 7$ blocs).



UMONS

La validation croisée - principe

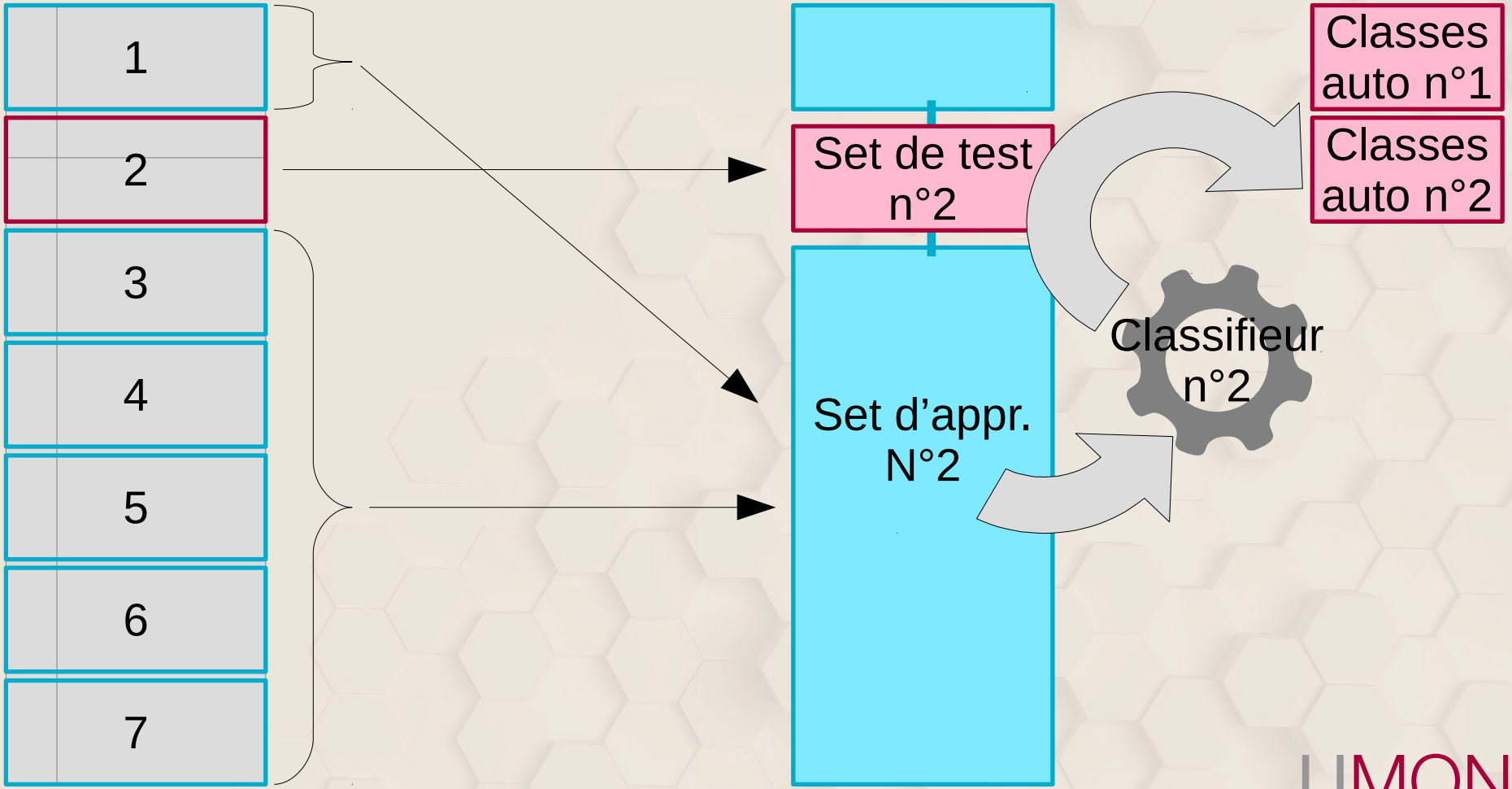
Etape 1 : le premier bloc est set de test, le reste set d'apprentissage. Classification du set de test.



UMONS

La validation croisée - principe

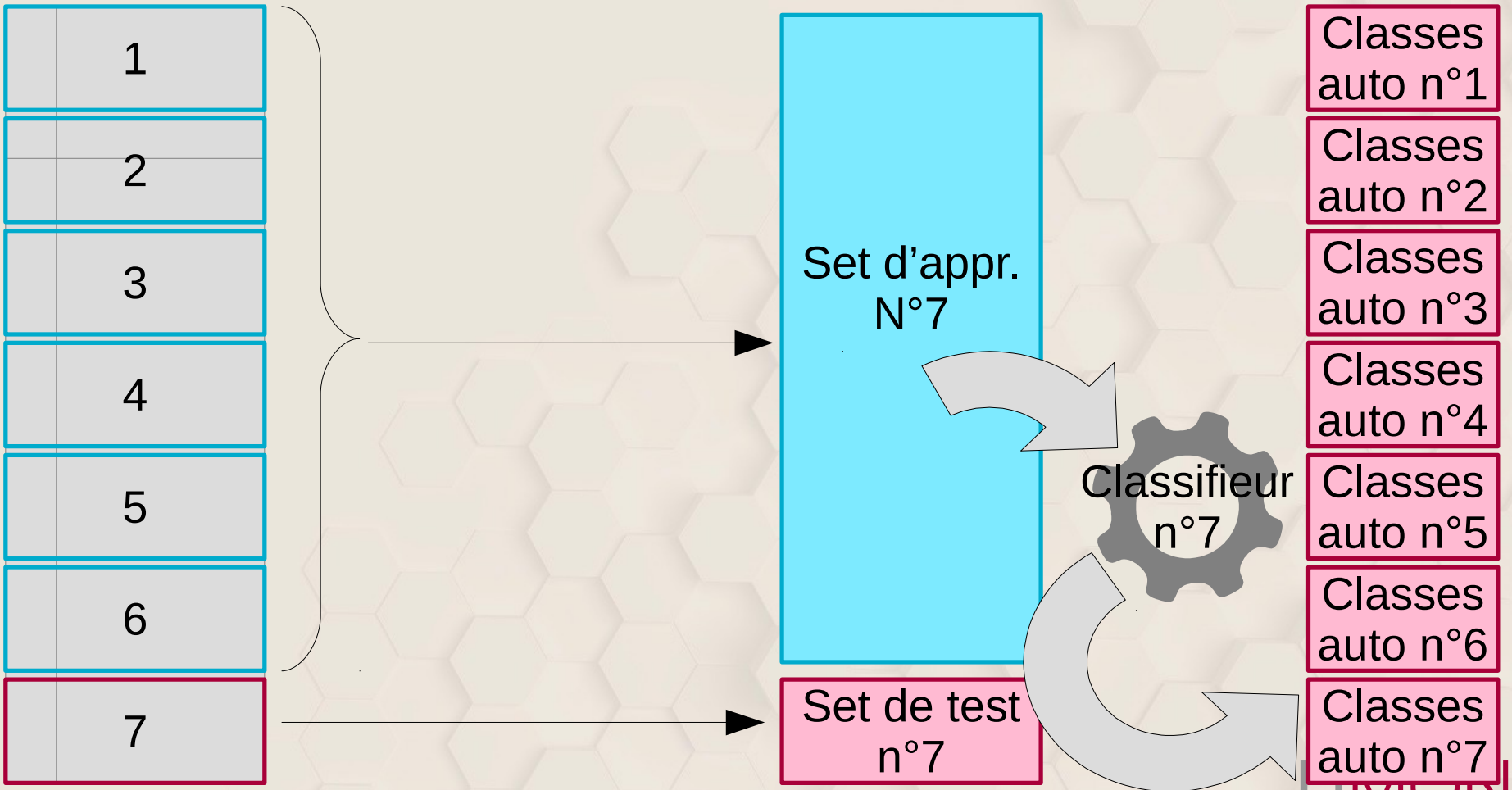
Etape 2 : le second bloc est set de test, le reste set d'apprentissage. Pooler la classification du set de test avec celle issue de l'étape précédente.



UMONS

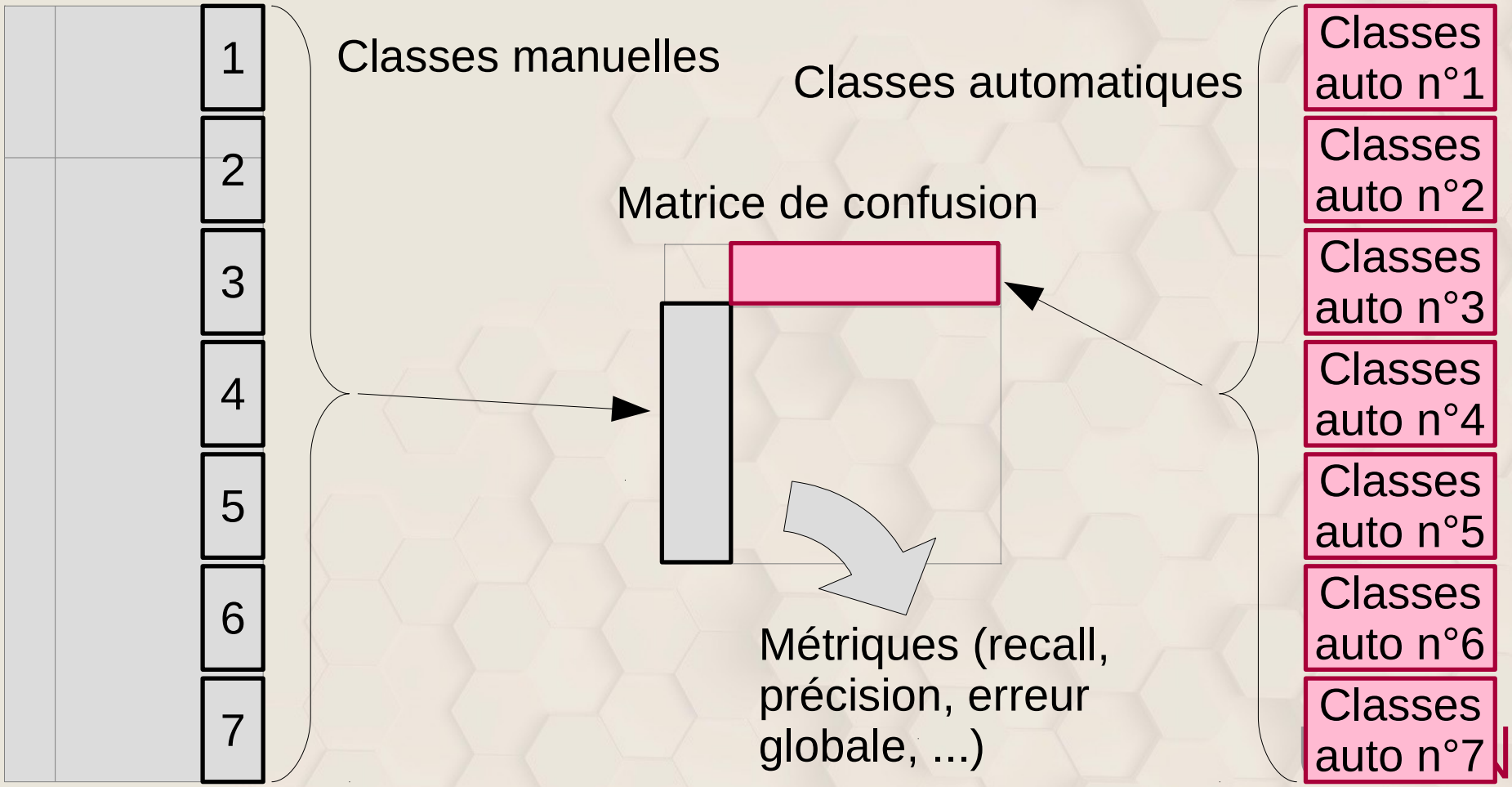
La validation croisée - principe

Etape 7 : à ce stage, k classifieurs ont été utilisés pour classer au final tous les individus du jeu de données initial.



La validation croisée - principe

Etape finale : les classes manuelles et automatiques de l'ensemble sont croisées dans la matrice de confusion, et les métriques sont calculées pour quantifier les performances.



Validation croisée en pratique

Application sur `iris`.

Comparaison de LDA, kNN et LVQ

Utilisation d'une validation croisée avec $k = 10$ pour comparer les performances de classification de 3 algorithmes sur le jeu de données `iris` (démo).

UMONS