# Exploration des données

## CM1-2 : Détection d'anomalies

Mickaël Martin Nevot

V1.0.0

# Exploration des données

# Rappel : nettoyage des données

- Deux étapes de traitement :
  - **Valeurs manquantes**, doublons, violation de contraintes, valeurs incorrectes
  - **Valeur aberrantes** *(outliers)*

# Rappel : données manquantes

- Problèmes des données manquantes :
  - Diminution des performances
  - Difficultés dans le traitement des données
  - Biais dans les estimations
- Types :
  - **MCAR** *(missing completely at random)* : par hasard
  - **MAR** *(missing at random)* : ne dépend que des autres var.
  - **NMAR** *(missing not at random)* : dépend de la var. manquante

# Gestion des données manquantes

- Méthode de **suppression** :
  - Analyse des cas complets (ACC)
  - Méthode des cas disponibles (MCD) / par paire
- Méthode **d'imputation** : ← Données manquantes < 20 %
  - Simple (moyenne, médiane, mode) ← Données manquantes < 5 %
  - Multiple
  - Autres méthodes : hot deck, last observation carried forward (LOCF), next observation carried backward (NOCB)
- Méthode basée sur un modèle :
  - KNN, RMSE, MAE, etc. ← Données manquantes < 20 %

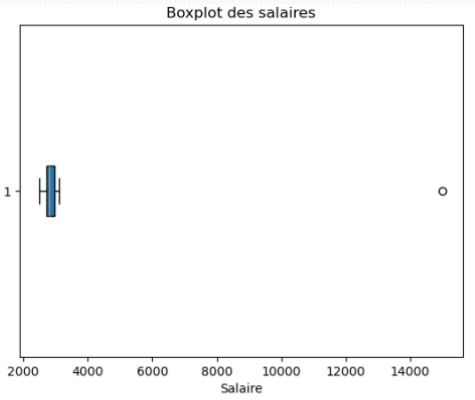
# Valeurs aberrantes *(outliers)*

- **Valeur qui s'écarte considérablement des autres valeurs**
- Peut être causée par erreurs de mesure / d'exécution
- Impact :
  - Influence sur les statistiques descriptives
  - Interprétation erronée
- Nettoyage : conserver, modifier, supprimer ?

Traitées comme valeurs manquantes ?

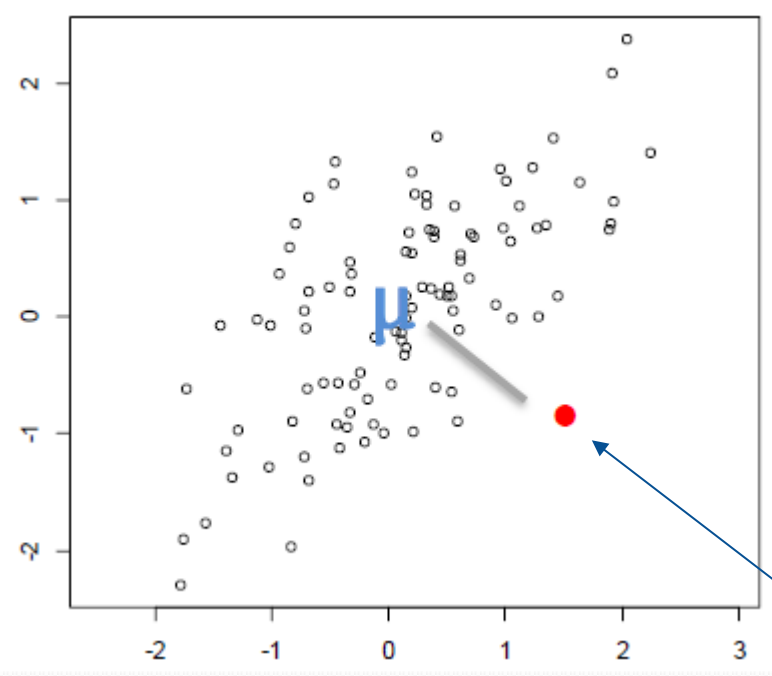
Les valeurs aberrantes peuvent contenir des informations importantes !

| Employé | Salaire |
|---|---|
| 1 | 2,500 |
| 2 | 2,700 |
| 3 | 2,800 |
| 4 | 3,000 |
| 5 | 2,600 |
| 6 | 2,900 |
| 7 | 3,100 |
| 8 | 15,000 |
| 9 | 2,800 |
| 10 | 2,750 |



# *Outliers* : par naïveté

- Distance de Mahalanobis :
  - Calcul de la distance de chaque point par rapport au barycentre (μ) en considérant le nuage de points *via* la covariance (Σ)

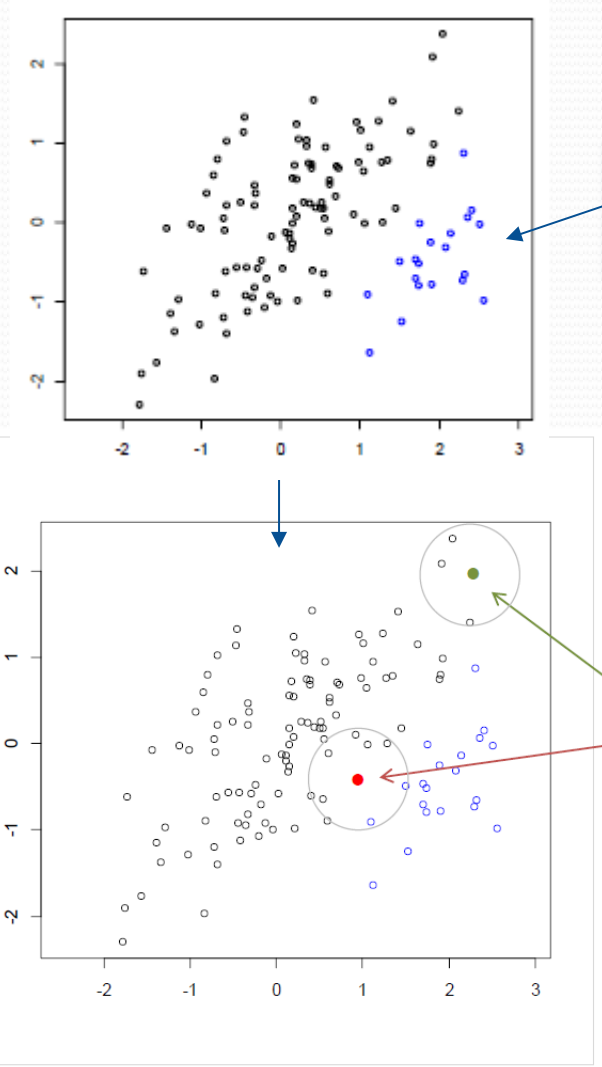



$$d^2_{(\mu,\Sigma)}(x_i) = (x_i - \mu)'\Sigma^{-1}(x_i - \mu)$$

Pas atypique sur chaque axe individuellement, mais l'est par rapport au nuage de points

# *Outliers* : par naïveté

Limitations




En cas de données non-gaussiennes, ou *clustérisées*, le barycentre global est non significatif

Dans une zone à forte densité, un point qui s'écarte de ses voisins devrait plus interroger qu'un se situant dans une zone moins dense

# *Outliers* : par statistiques

- Score Z :

$$Z = \frac{x - \mu}{\sigma}$$

- X : valeur de la donnée
- $\mu$ : moyenne de l'ensemble des données
- $\sigma$ : écart-type de l'ensemble des données

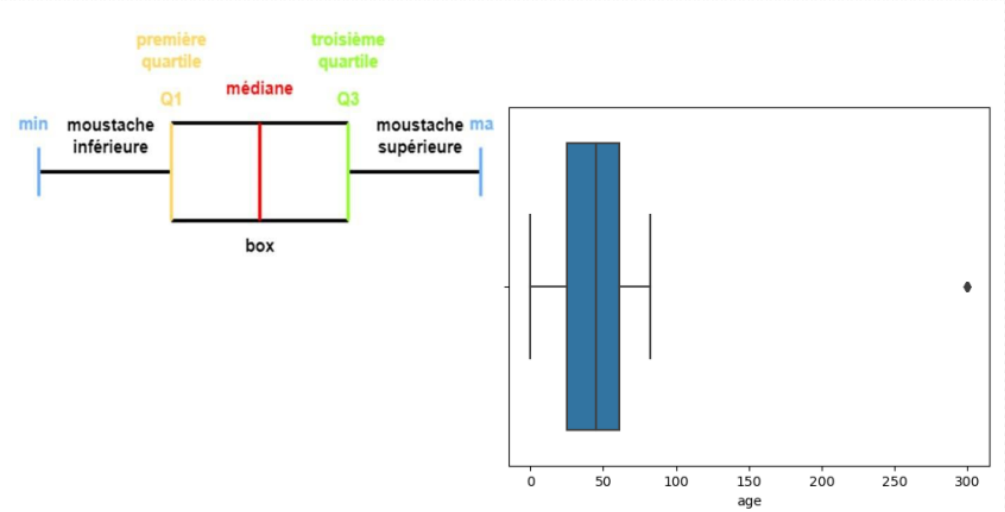
- Approche IQR *(inter quartile range)*

**IQR = Quartile3 – Quartile1**

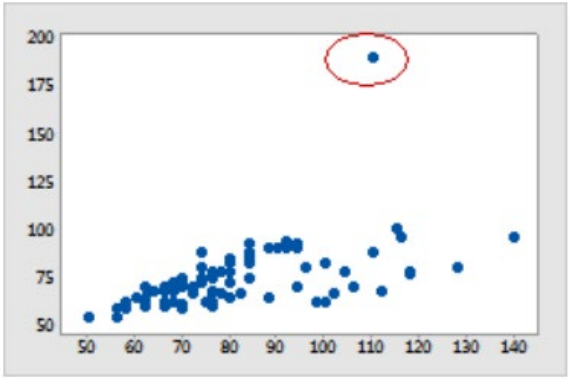
> IQR * 1,5 au-dessus du Quartile3 ou en dessous du Quartile1

# *Outliers* : par visualisation

- Boite à moustache :




- Nuage de points :

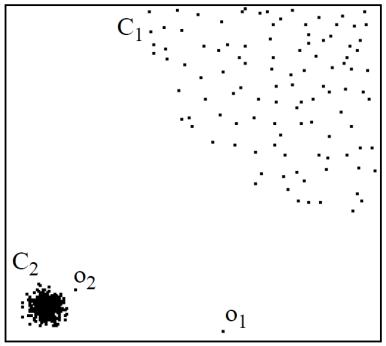


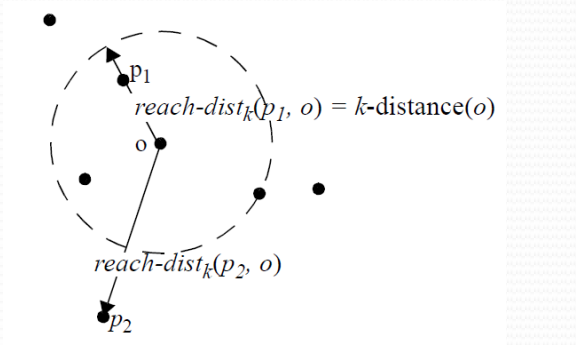

+ histogrammes

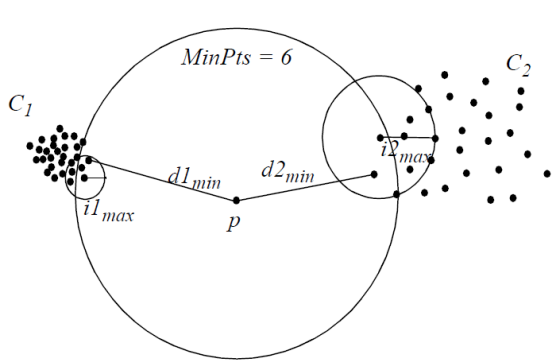
# *Outliers* : par algo. de densité

- **Données binaires et vision globale** :
  - Partitionnement : sensible aux densités hétérogènes
  - Enveloppes convexes : **DBSCAN**
- **Degré d'aberration et vision locale** :
  - Distance des voisins
  - Densité du voisinage : **LOF**

Tester plusieurs approches pour voir comment elles influencent les performances des modèles



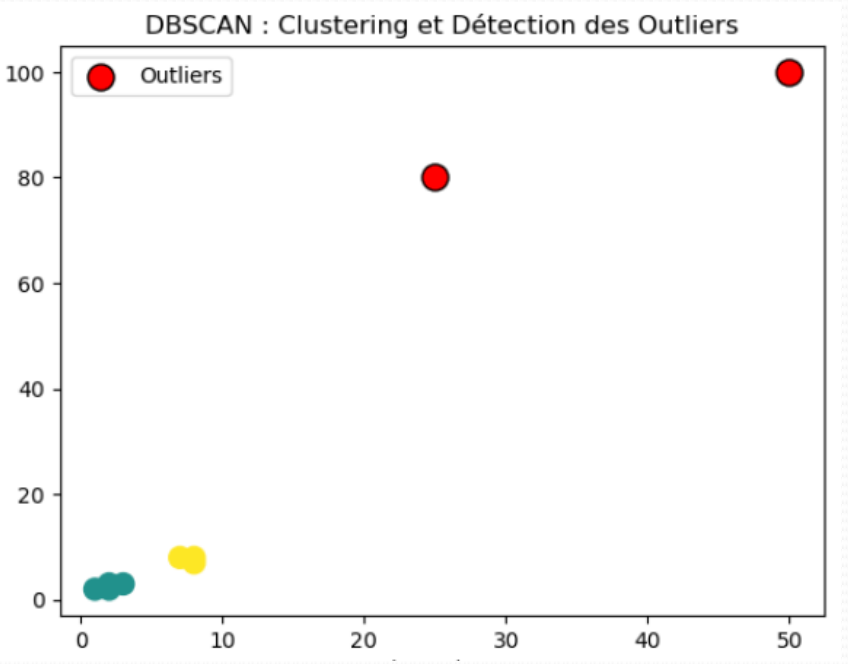




Algorithmes non supervisés

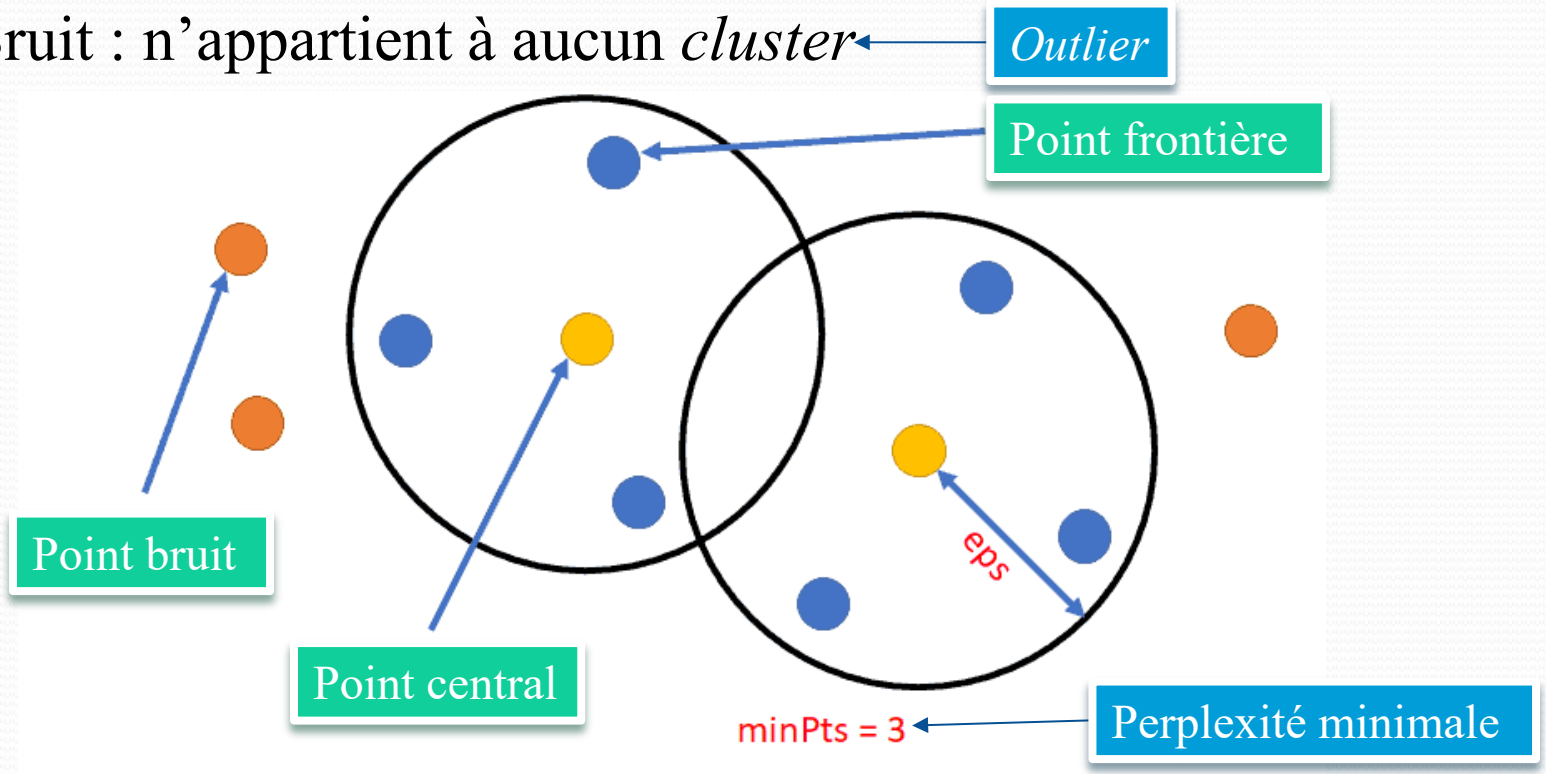
# *Outliers* : par DBSCAN

- *Density-based spatial clustering of applications with noise*
- **Algorithme** très connu en matière de **partitionnement** des données *(clustering)*
- Utilise l'estimation de la **densité locale**
- Les valeurs hors *cluster* denses sont considérées, de manière manichéenne, comme des bruits, et donc des *outliers*

# *Outliers* : par DBSCAN

- Types de points :
  - Central : au cœur d'un *cluster*
  - Frontière : proche point central mais pas assez de voisins
  - Bruit : n'appartient à aucun *cluster*



# *Outliers* : par DBSCAN

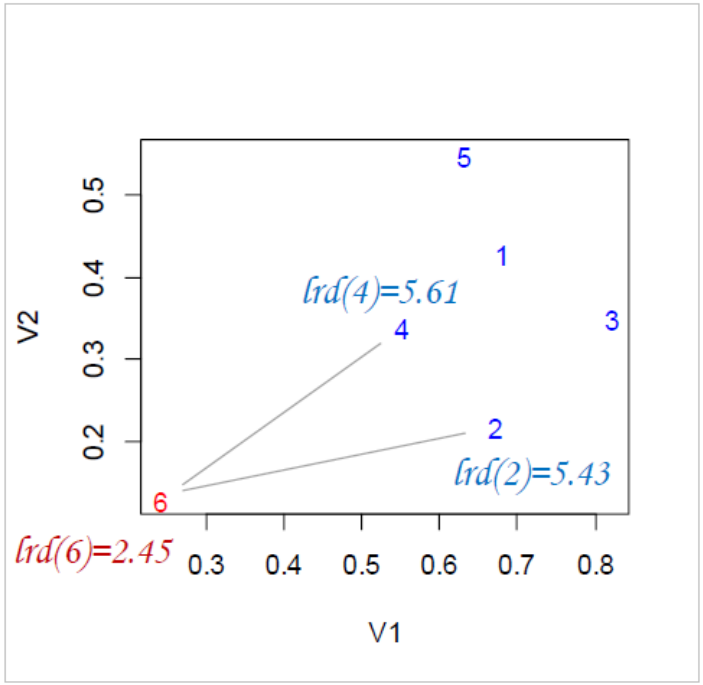
- Paramètres :
  - ε (eps) : distance maximale entre deux points considérés comme voisins :
    - Si trop petit : aucun voisin, que des *outliers*
    - Si trop grand : que des voisins, aucun *outlier*

    Utilisation de la méthode du graphe des k-distances
  - `minPts` : nombre min. de points (moins celui central) requis dans un *cluster*
    - Si trop petit : qu'un *cluster* dense, aucun *outlier*
    - Si trop grand : aucun *cluster* dense, que des *outliers*

    Après choix de ε : définir `minPts` comme nombre moyen de points dans les ε-voisinages

# *Outliers* : par LOF

- *Local outlier factor* :
  - Calcul de densité locale envers k-plus proches voisins :
    - *Local reachability density* (lrd) : mesure de densité locale d'un point




K = 2

$$\begin{cases} lrd(6) \ll lrd(4) \\ lrd(6) \ll lrd(2) \end{cases}$$

Le point n°6 est potentiellement atypique

# *Outliers* : par LOF

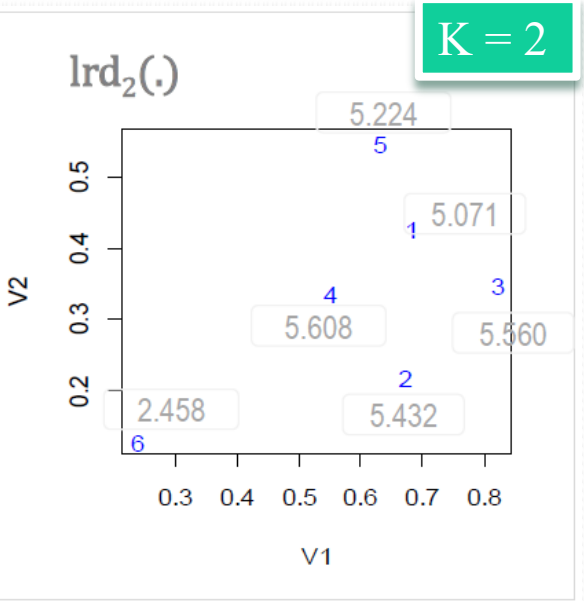
- LOF < 1 : densité > voisins, *inlier*
- LOF ≃ 1 : densité ≃ voisins
- LOF > 1 : densité < voisins, *outlier*

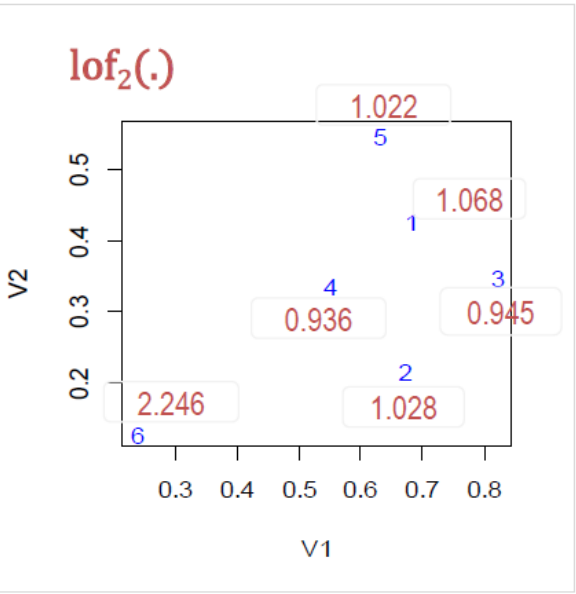
$$lof_k(A) = \frac{\sum_{B \in N_k(A)} \frac{lrd_k(B)}{lrd_k(A)}}{|N_k(A)|}$$

$$lof_2(6) = \frac{\frac{5.608}{2.458} + \frac{5.432}{2.458}}{2} = 2.246$$

Penser à ordonner les données !

K = 2

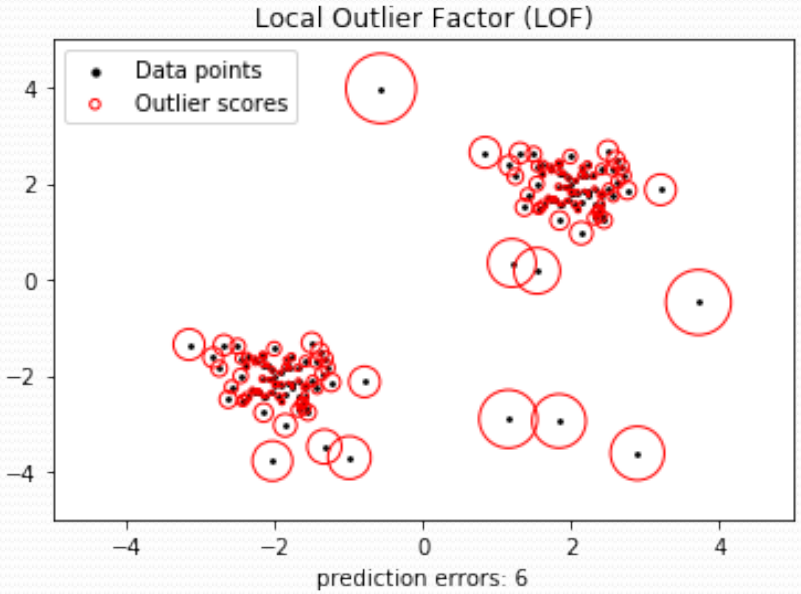


# *Outliers* : par LOF

- Paramètre :
  - k : nombre de voisins à considérer
    - Si k petit : plus précis mais instabilité des résultats
    - Si k grand : plus lissé mais risque de masquer les informations

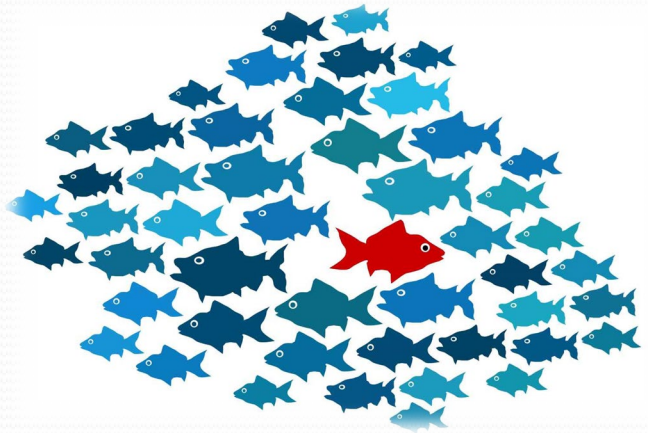
Le choix de k reste un problème ouvert



# Détection d'anomalies

- Applications :
  - Identification **d'observations d'une autre population**
  - **Situations exceptionnelles**
  - **Comportements déviants**, détection des intrusions
- Formes d'anomalies :
  - ***Outlier*** : point « isolé » dans une zone peu dense
  - ***Novelty*** : nouveau point par rapport à un échantillon de référence « propre »

Stratégie : intégration ou élimination ?



# Boîte à outils

- Détection :
  - Great Expectations (bibliothèque Python) : https://greatexpectations.io
  - Pandera : https://github.com/unionai-oss/pandera
  - Pydantic : https://github.com/pydantic/pydantic
- Corrections :
  - OpenRefine : https://openrefine.org
  - HoloClean : http://www.holoclean.io

# Crédits

## Auteur

Mickaël Martin Nevot
mmartin.nevot@gmail.com

Carte de visite électronique

## Relecteurs

Cours en ligne sur : www.mickael-martin-nevot.com