XAI
eXplainable Artificial Intelligence
Cours 8 - mardi 21 novembre 2023

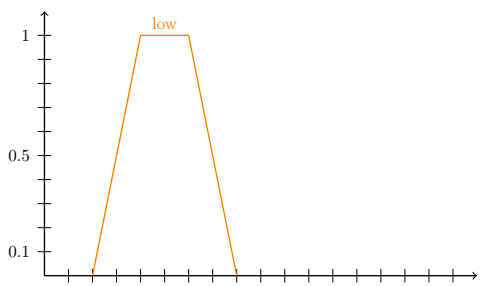
Marie-Jeanne Lesot
Christophe Marsala
Jean-Noël Vittaut
Gauvain Bourgne

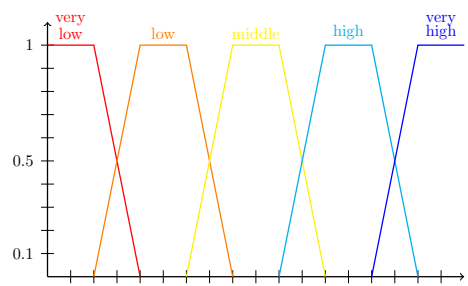
LIP6, Sorbonne Université

Motivations

- Interprétabilité : explicabilité par modèles transparents ?
- Cas des k-ppv
  - compréhension de la distance/similarité ?
- Cas des arbres de décision
  - dépend de leur profondeur
  - les attributs doivent être compréhensibles
    si l'indicateur A843b9 < 12.5
  - problème de compréhension des seuils : trop arbitraires ?
    si la largeur des pétales < 6.327

$\Longrightarrow$ utilisation de sous-ensembles flous



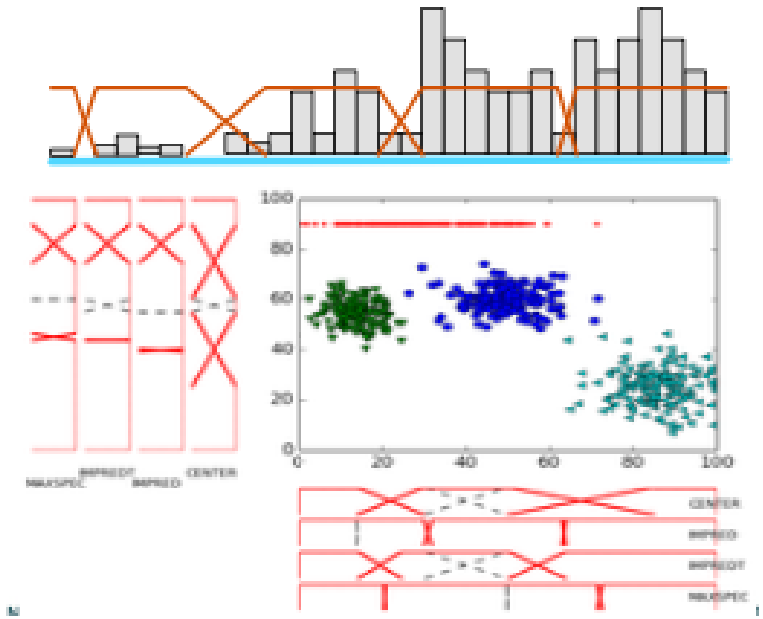


Au programme du jour

- 1. Construction de sous-ensembles flous
  - à partir des données
  - par l'utilisateur
  - construction coopérative
  - approche cognitive
- 2. Mise en œuvre : interrogation flexible de bases de données
- 3. Explications d'exceptions

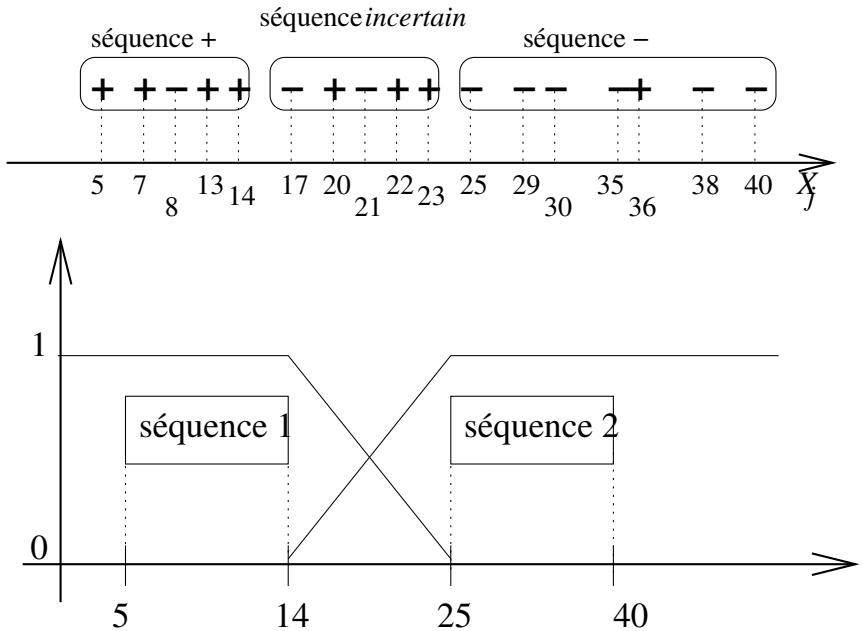
A partir des données

- Approche non supervisée : selon leur distribution



A partir des données

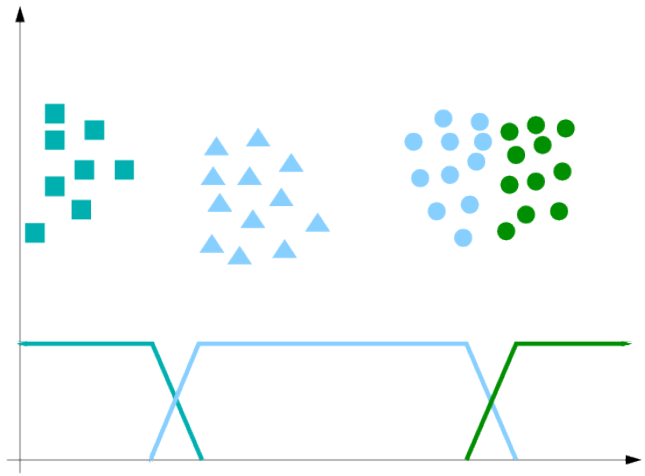
- Approche supervisée : prise en compte des classes
  - outils de morphologie mathématique



- **Risque** : non intuitif pour l'utilisateur
  - en mode supervisé ou non

Définition par l'utilisateur

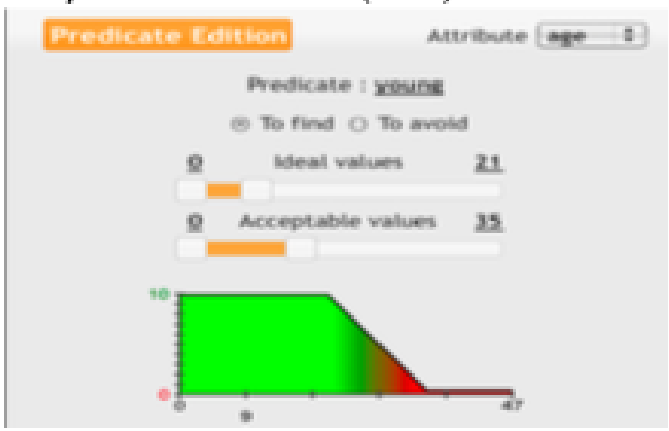
- **Risque** : inadéquation aux données



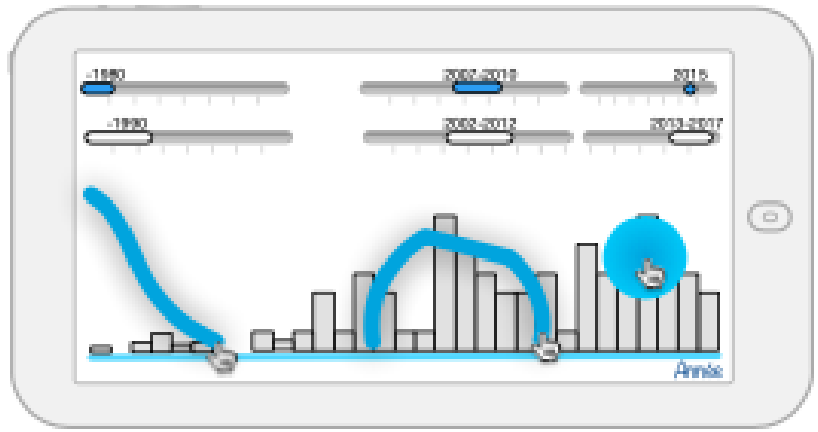
- Besoin d'une double adéquation
  - même description numérique $\Longrightarrow$ même description linguistique
  - même description linguistique $\Longrightarrow$ mêmes sous-groupes numériques
- Méthode de révision de vocabulaire
  - modifications locales par **décomposition de modalités**
  - pour préserver l'interprétabilité par l'utilisateur

Interfaces de saisie

ReqFlex Smits et al. (2013)



iPadFS Yager et al. (2019)

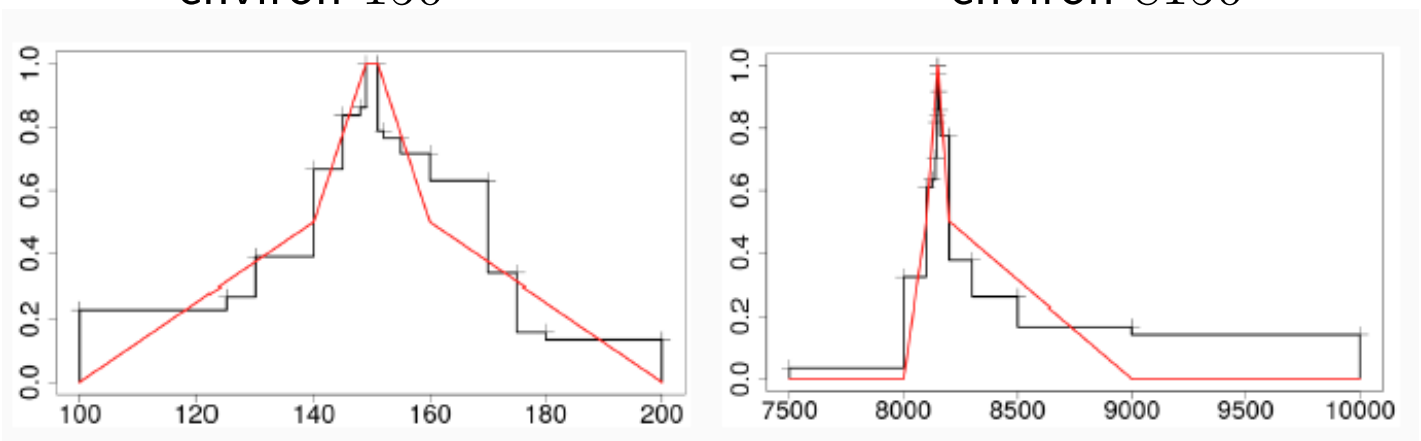


Approche cognitive

- Modélisation des expressions numériques approximatives : **environ** x
  - **How much is about?** (Lefort et al. 2016)
- Collecte de données expérimentales réelles : "selon vous, entre quelles valeurs minimale et maximale se trouve environ x ?"

  $\Longrightarrow$ intervalle symétrique dans 79% des cas étudiés seulement

environ 150 environ 8150



Modèle computationnel

- En fonction de dimensions arithmétiques et cognitives de x

$$x = \sum_{i=0}^{q} c_i \cdot 10^i \text{ avec } c_i \in [\![0, 9]\!]$$

dimension	définition	exemple
- magnitude	x	8150
- dernier chiffre significatif	c_{i^*} avec $i^* = \min\{i / c_i \neq 0\}$	5
- granularité	10^{i^*}	10
- nombre de chiffres significatifs	$q - i^* + 1$	3
- complexité cognitive	NCS(x) - B(x)	2.5

$$\text{avec B}(x) = \begin{cases} 0.5 & \text{si DCS}(x) = 5 \text{ et NCS}(x) > 1 \\ 0.25 & \text{si DCS}(x) \in \{2, 8\} \text{ et NCS}(x) > 1 \\ 0 & \text{sinon} \end{cases}$$

C(400) = 1 C(445) = 2.5 C(446) = 3

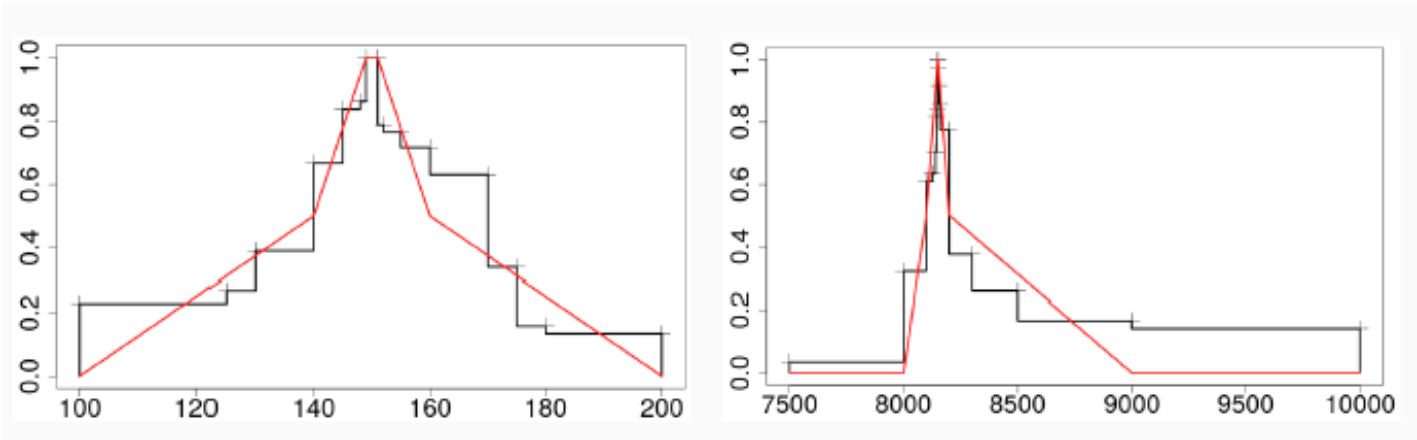
Modèle computationnel

- Validation expérimentale

  $\Longrightarrow$ effet séparés de la magnitude, la granularité et DCS

  $\Longrightarrow$ performance du modèle

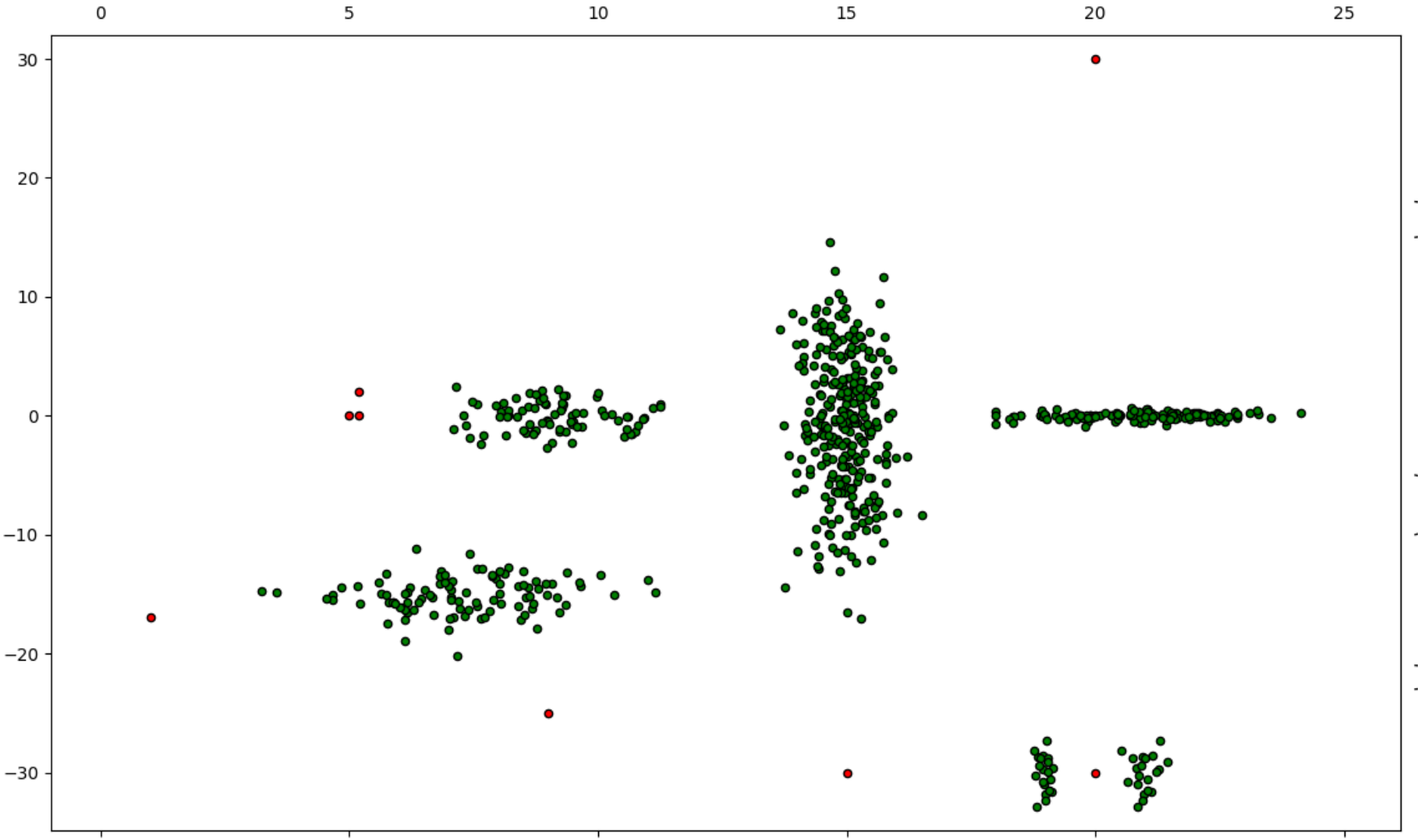
environ 150 environ 8150



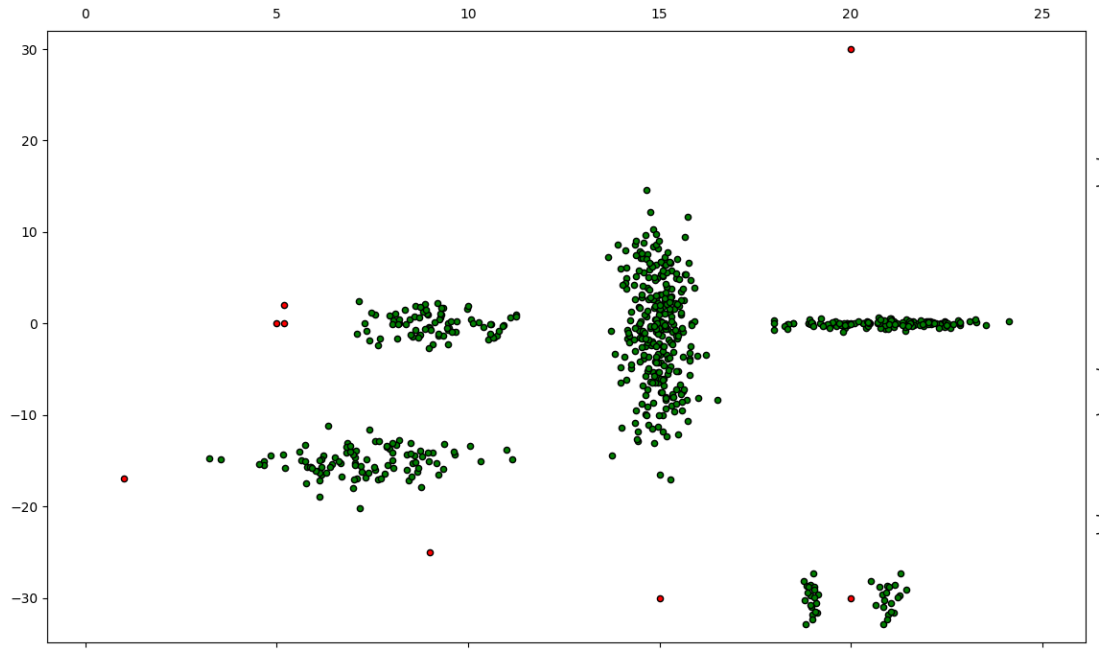
Au programme du jour

- 1. Construction de sous-ensembles flous
- 2. Mise en œuvre : interrogation flexible de bases de données
- **3. Explications d'exceptions** = anomalie, outlier
  - détection d'exceptions
  - génération d'explications

Exceptions : illustration et définition



Exceptions : illustration et définition



- Définition : *"an observation that deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism"*

(Hawkins D., Identification of outliers.
Monographs on applied probabilities and statistics, 1980)

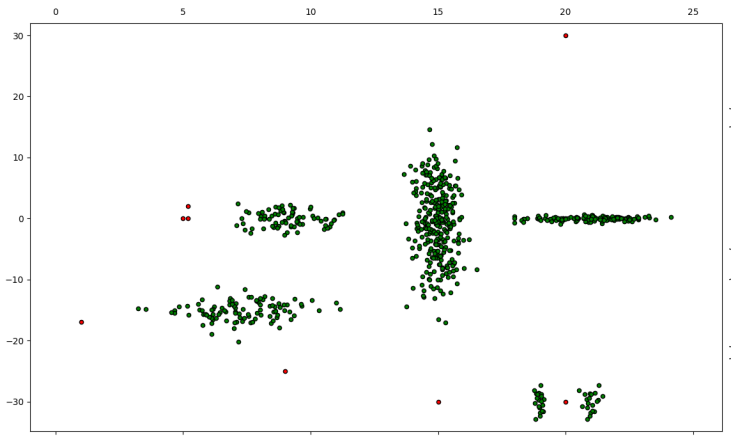
Détection d'exceptions

- Multiplicité de méthodes et de taxonomies

(Goldstein & Ushida, 16; Liu et al, 12; Tchaghe et al, 21)

- Quelques exemples de méthodes :
  - LOF, One-class SVM, auto-encodeurs, forêt d'isolation

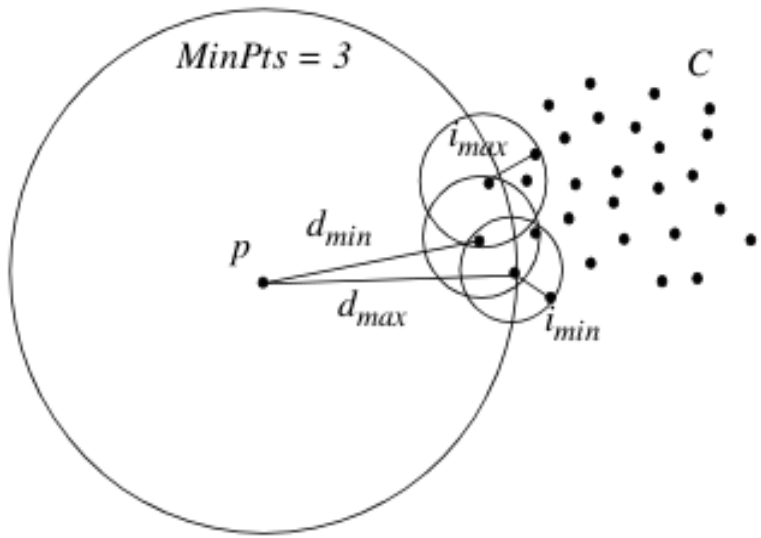
(Breunig et al, 00; Amer et al, 13; Chen et al, 17; Liu et al, 12)



LOF: Local Outlier Factor

(Breunig et al, 00)

- Idée de base : **comparaison de la densité locale autour d'un point à la densité locale de ses voisins**



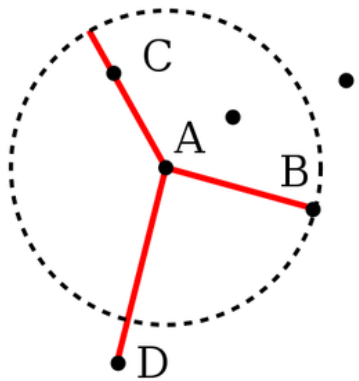
$d_{min} = 4 * i_{max}$
$\Rightarrow LOF_{MinPts}(p) \geq 4$

$d_{max} = 6 * i_{min}$
$\Rightarrow LOF_{MinPts}(p) \leq 6$

LOF: Local Outlier Factor

(Breunig et al, 00)

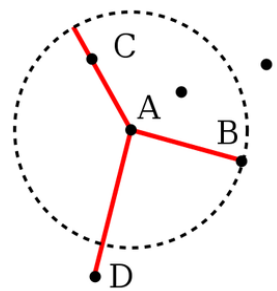
- Idée de base : **comparaison de la densité locale autour d'un point à la densité locale de ses voisins**
- Formellement
  - $kDist(x) = \text{sort}([dist(x,z) \text{ for } z \in \mathcal{X}])[k]$
  - k plus proches voisins $\mathcal{N}_k(x) = \{z \in \mathcal{X} | d(x,z) \leq kDist(x)\}$
  - reachability-distance : $rDist(x,y) = \max(kDist(y), d(x,y))$



LOF: Local Outlier Factor

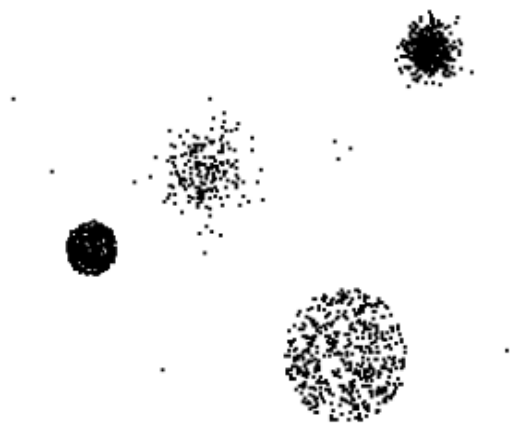
(Breunig et al, 00)

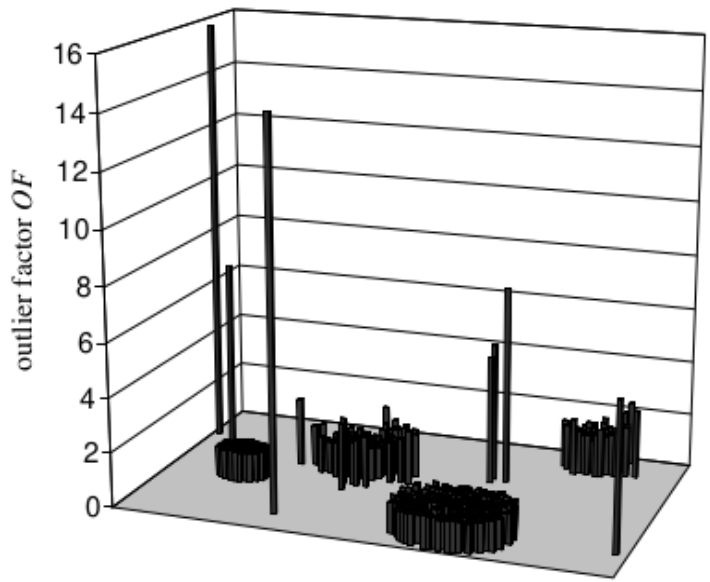
- Idée de base : **comparaison de la densité locale autour d'un point à la densité locale de ses voisins**
- Formellement
  - $kDist(x) = \text{sort}([dist(x,z) \text{ for } z \in \mathcal{X}])[k]$
  - k plus proches voisins $\mathcal{N}_k(x) = \{z \in \mathcal{X} | d(x,z) \leq kDist(x)\}$
  - reachability-distance : $rDist(x,y) = \max(kDist(y), d(x,y))$
  - local reachability distance :
    inverse de la $rDist$ de x **depuis** ses voisins
    $$lrd_k(x) = \left(\frac{\sum_{z \in \mathcal{N}_k(x)} rDist(x,z)}{|\mathcal{N}_k(x)|} \right)^{-1}$$
  - score d'anomalie
    $$LOF_k(x) = \frac{1}{|\mathcal{N}_k(x)|} \sum_{z \in \mathcal{N}_k(x)} \frac{lrd_k(z)}{lrd_k(x)}$$



LOF: Local Outlier Factor

(Breunig et al, 00)

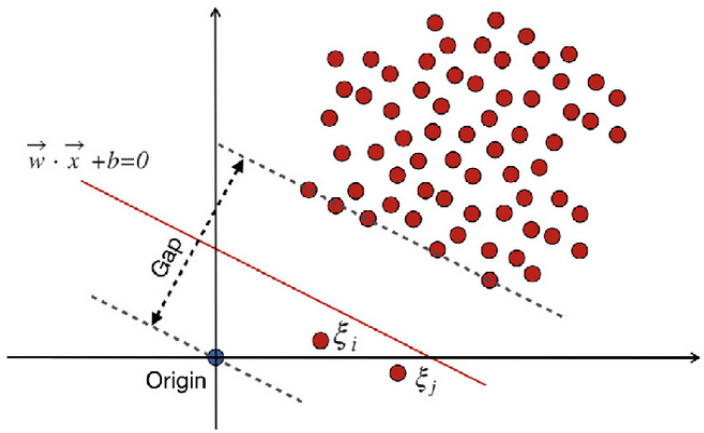


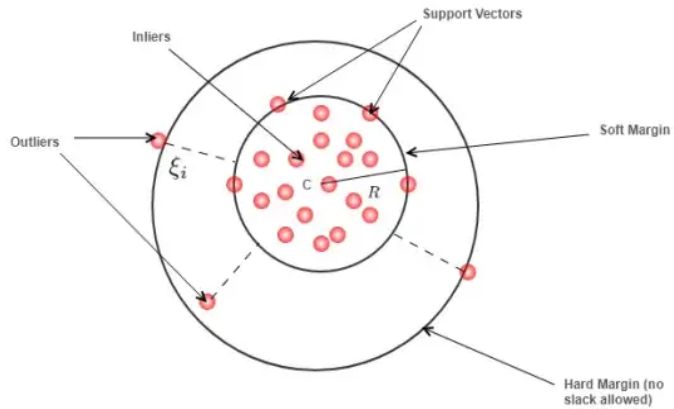


One-Class SVM

(Schölkopf et al, 01)

- Trouver un hyper-plan qui sépare les données de l'origine, en maximisant la marge





One-Class SVM

(Schölkopf et al, 01)

- Trouver un hyper-plan qui sépare les données de l'origine, en maximisant la marge

$$\min \frac{1}{2}\|w\|^2 + \frac{1}{\nu}\sum_i \xi_i - \rho$$

sous contrainte $\forall i, (w \cdot \varphi(x_i)) \geq \rho - \xi_i$ et $\xi_i \geq 0$

- De façon équivalente

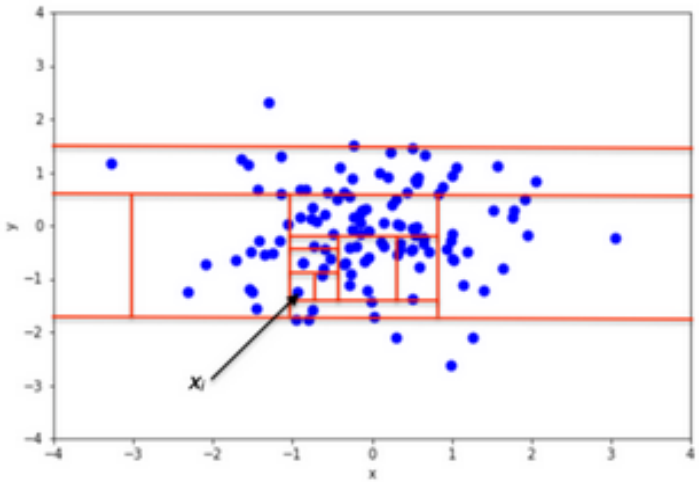
$$\min_\alpha \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j k(x_i, x_j) \text{ sous contrainte } \forall i, 0 \leq \alpha_i \leq \frac{1}{\nu}, \sum_i \alpha_i = 1$$

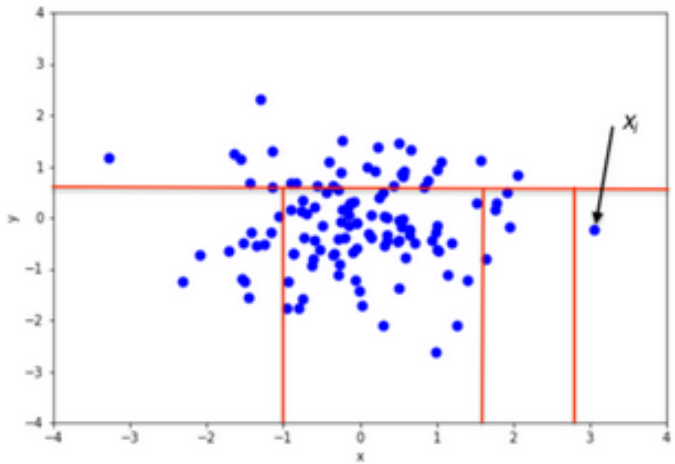
  - souvent utilisé avec un noyau gaussien $k(x, y) = (\varphi(xi) \cdot \varphi(y)) = exp\left(\frac{\|x-y\|^2}{c}\right)$

Forêts d'isolation

(Liu et al, 08)

- Principe : **les exceptions sont faciles à isoler**
  - décomposer récursivement les données
  - par tirage aléatoire d'un attribut et un seuil
  - jusqu'à ce que chaque donnée soit seule dans la feuille
- Construire une forêt d'arbres d'isolation

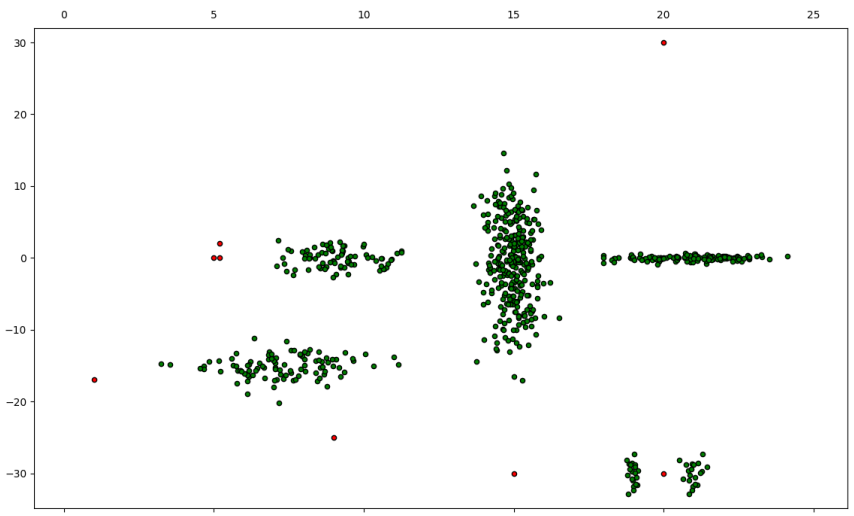


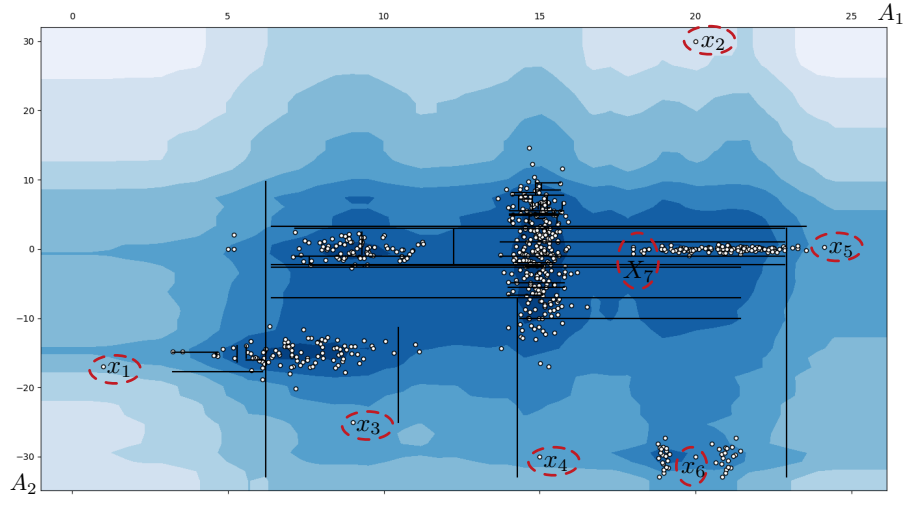


Forêts d'isolation

(Liu et al, 08)

- Exemple





Au programme du jour

- 1. Construction de sous-ensembles flous
- 2. Mise en œuvre : interrogation flexible de bases de données
- 3. Explications d'exceptions = anomalie, outlier
  - détection d'exceptions
  - **génération d'explications**

Critères de comparaison

- Local vs global
  - expliquer une anomalie ou plusieurs
- *Model-agnostic* vs *model-specific*
- 4 familles (Tchaghe et al. 2021)
  - importance d'attribut
  - valeur d'attribut
  - comparaison de points
  - analyse de la structure des données

Importance d'attribut

- **Identification de sous-espaces** : *Group outlying aspects mining*
  - Subspace Outlying Degree : sous-espace minimal dans lequel x est une exception

    (H.-P. Kriegel, P. Kröger, E. Schubert, A. Zimek, Outlier detection in axis-parallel subspaces of high dimensional data, 2009)
  - Correlation Outlier Probabilities : transformation de l'espace initial

    (H.-P. Kriegel, P. Kröger, E. Schubert, A. Zimek, Outlier detection in arbitrarily oriented subspaces, 2012)
  - LookOut : trouver les paires d'attributs qui discriminent les exceptions, en maximisant le score d'exceptionalité calculé dans le sous-espace

    (N. Gupta, D. Eswaran, N. Shah, L. Akoglu, C. Faloutsos, Beyond outlier detection: Lookout for pictorial explanation, 2018)
  - Après clustering des exceptions

    (H. Liu, F. Ma, Y. Wang, S. He, J. Chen, J. Gao, Lp-explain: Local pictorial explanation for outliers, 2020)

Importance d'attribut

- **Pondération d'attributs**
  - avec KernelSHAP
    - anomalie : point avec coût élevé d'auto-encodage
    - calcul de SHAP pour l'auto-encodeur utilisé pour prédire les attributs d'erreur maximale

(L. Antwarg, B. Shapira, L. Rokach,
Explaining anomalies detected by autoencoders using shap, 2019)

  - DIFFI : agrégation de la profondeur d'utilisation des attributs dans une forêt aléatoire

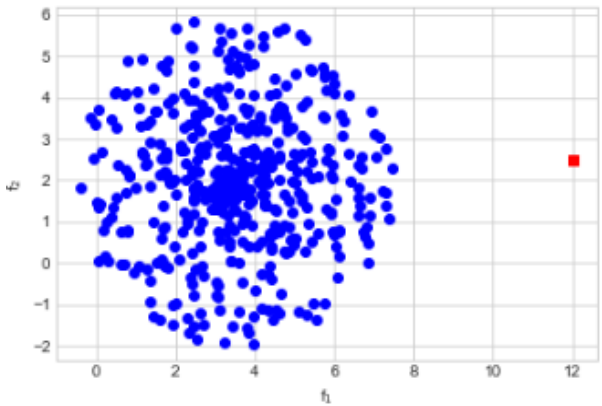
(M. Carletti, M. Terzi, G. A. Susto,
Interpretable anomaly detection with diffi:
Depth-based feature importance for the isolation forest, 2020)

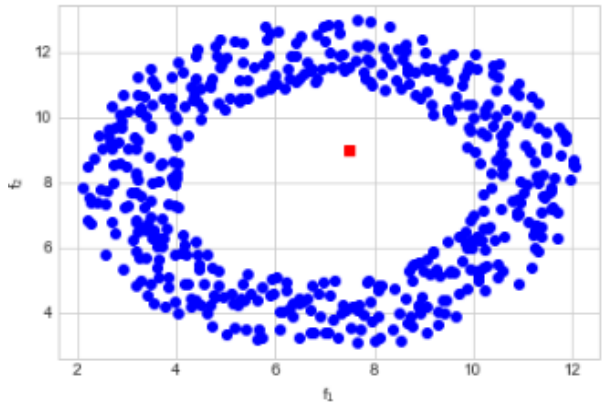
  - gradient des attributs d'un auto-encodeur variationnel : si une variation d'un attribut provoque une importante variation du score d'anomalie

(Q. P. Nguyen, K. W. Lim, D. M. Divakaran, K. H. Low, M. C. Chan,
Gee: A gradient-based explainable variational autoencoder for network anomaly detection, 2019)

Valeur d'attribut

- Il n'y a pas toujours d'attribut responsable





Valeur d'attribut

- Explications = formules logiques sous forme normale disjonctive

$$(\bigwedge_i X_i *_i a_i) \vee \ldots \qquad \text{où } *_i \in \{<, \leq, =, >, \geq\}$$

  - forêt aléatoire et agrégation des branches pointant vers la classe exception

(E. Baseman, S. Blanchard, N. DeBardeleben, A. Bonnie, A. Morrow, Interpretable anomaly detection for monitoring of high performance computing systems, 2016)

  - décomposition récursive des régions sans exceptions en hypercubes, dont les frontières définissent des règles logiques

(A. Barbado, Ó. Corcho, R. Benjamins, Rule extraction in unsupervised anomaly detection for model explainability, 2019)

Comparaison de points

- Comparaison avec la moyenne des données non exceptionnelles
- Approche contrefactuelle : valeurs à changer pour ne plus être une exception

(S. Haldar, P. G. John, D. Saha,
Reliable counterfactual explanations for autoencoder based anomalies, 2021)

- ABOD: Comparaison angulaire avec le point le plus proche appartenant à un cluster

(H.-P. Kriegel, M. Schubert, A. Zimek,
Angle-based outlier detection in high-dimensional data, 2008)

Analyse de la structure des données

- Traitement de l'intégralité des données
  - pas seulement le voisinage de l'exception
  - clustering robuste, capable de détecter les exceptions
- Position relative aux données non anormales
  - capture d'une notion de contexte
- x-PACKS: subspace clustering + approximation par hyper-rectangle des régions avec exceptions + extraction de règles

(M. Macha, L. Akoglu,
Explaining anomalies in groups with characterizing subspace rules, 2018)

PANDA : Personalised ANomaly Detection and Analysis

(Smits et al, 2022)

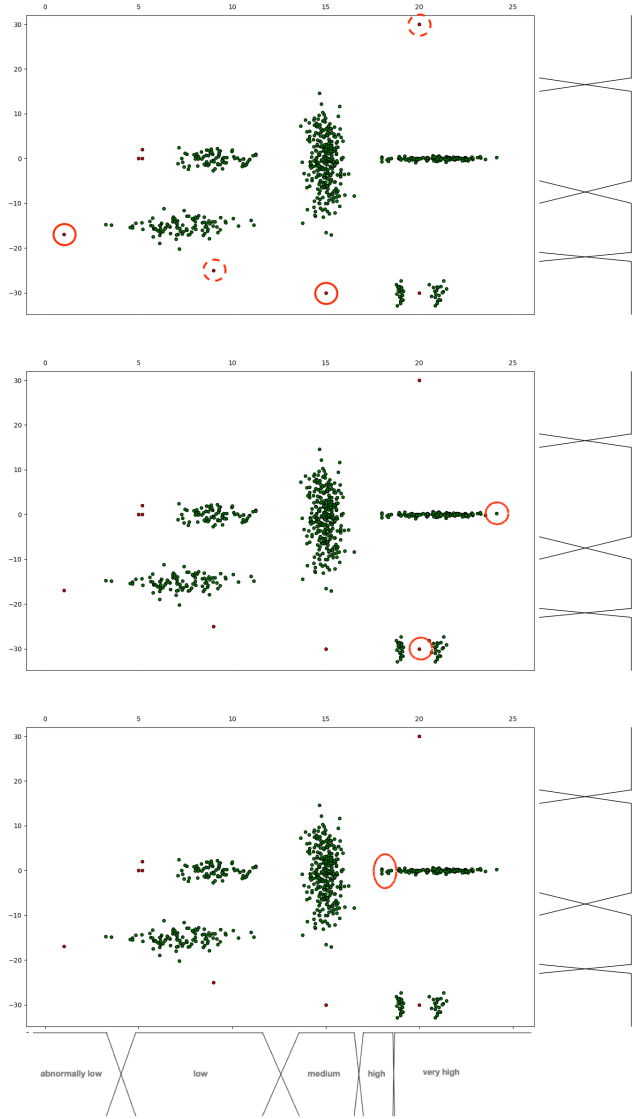
- **1. Double représentation complémentaire**
  - $\mathcal{D}$: description initiale $x = \langle x_1, \ldots, x_m \rangle$
  - $\mathcal{D}^{\mathcal{V}}$ basée sur le vocabulaire défini par l'utilisateur : réécriture
- **2. Double détection d'anomalie** : appliquée à $\mathcal{D}$ et $\mathcal{D}^{\mathcal{V}}$

$$\Rightarrow \begin{cases} \mathcal{A} & = \text{ anomalies guidées par les données} \\ \mathcal{A}^{\mathcal{V}} & = \text{ anomalies guidées par les connaissances} \end{cases}$$

- **3. Analyse croisée** des anomalies détectées
  - $\mathcal{A} \cap \mathcal{A}^{\mathcal{V}}$: anomalies communes $\Longrightarrow$ anomalies confirmées et décrites
  - anomalies propres à chaque représentation :
  - $\mathcal{A} \setminus \mathcal{A}^{\mathcal{V}}$, propres aux données $\Longrightarrow$ anomalies inattendues
  - $\mathcal{A}^{\mathcal{V}} \setminus \mathcal{A}$, propres au vocabulaire $\Longrightarrow$ vocabulaire subtil

PANDA

- **Anomalies confirmées** : $\mathcal{A} \cap \mathcal{A}^{\mathcal{V}}$
  - avec description linguistique
  - NB : description "totale", $\neq$ cause
- **Anomalies inattendues** $\mathcal{A} \setminus \mathcal{A}^{\mathcal{V}}$
  - isolées en terme de densité, mais couvertes par des termes qui les rendent régulières
  - diagnostics : vocabulaire inadéquat ou données incomplètes
- **Points d'intérêt** $\mathcal{A}^{\mathcal{V}} \setminus \mathcal{A}$
  - distinctions non supportées par la distribution des données
  - diagnostics : vocabulaire inadéquat ou attention attirée sur des cas particuliers



TME

- Données
  - base artificielle similaire à celles du transparent 12
  - base de référence
- Détection d'exceptions avec scikit-learn
  - `https://scikit-learn.org/stable/modules/generated/sklearn.svm.OneClassSVM.html`
  - `https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.LocalOutlierFactor.html`
  - `https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.IsolationForest.html`
- Explications naïves
  - que donne une approche contre-factuelle pour les forêts d'isolation ?
  - que donnent les approches par vecteur d'importance d'attributs ?
- Explications moins naïves
  - choisir, implémenter et tester une des approches présentées

Au programme du jour

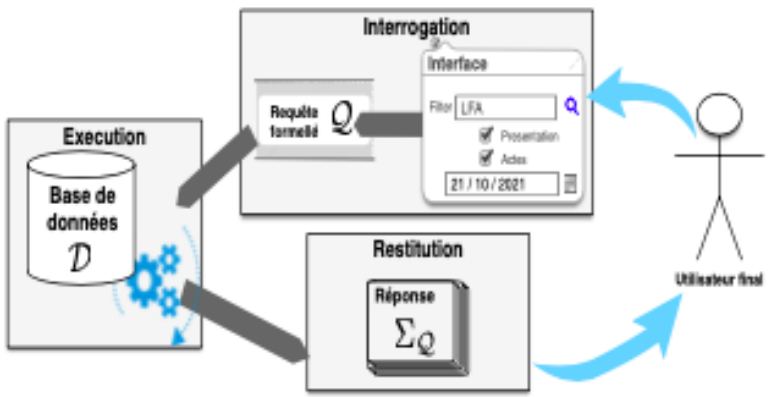
- 1. Construction de sous-ensembles flous
- 2. Mise en œuvre : interrogation flexible de bases de données

Merci à Grégory Smits, IMT Brest

  - interrogation par requête
  - exécution de la requête
  - restitution des réponses
- 3. Explications d'exceptions

Processus d'interrogation de données structurées

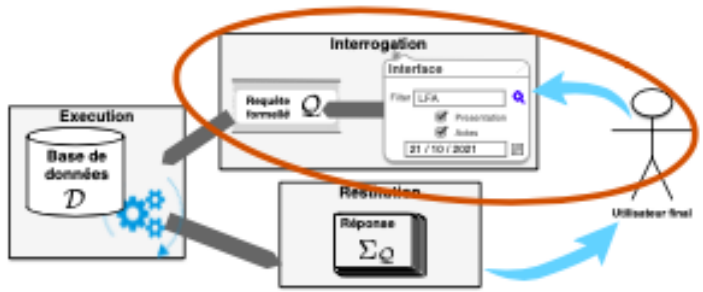
- Boucle d'interaction pour accélérer la transformation des données en connaissances utiles



1. expression d'un besoin d'information
2. exécution de la requête
3. restitution des réponses

- Solution technique, mais
  - expressivité limitée des interfaces et langages d'interrogation
  - systèmes de stockage et exécution dédiés (modèle et langage)
  - résultat difficilement exploitable

Etape d'interrogation



- Langage formel d'interrogation comme canal de communication
  - expressivité limitée
  - formalisme potentiellement abscons pour l'utilisateur final
  - méconnaissance des données et de leur structure

$\Longrightarrow$ simplifier la construction de requêtes

Interrogation par préférence

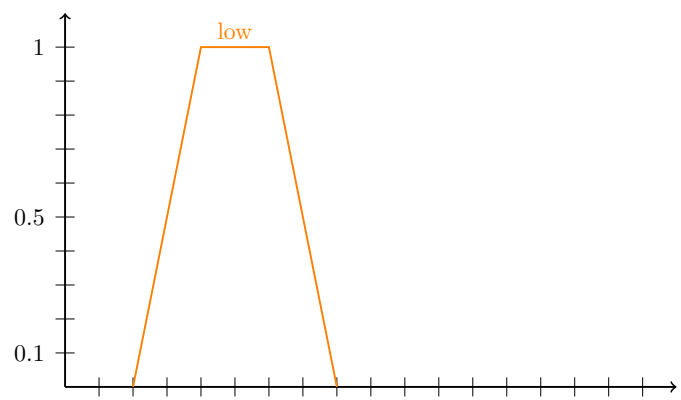
- Prédicat booléen obligatoire $\Longrightarrow$ préférence utilisateur désirée

  (Chang, 76; Lacroix et Lavency , 87; Motro 88)

- Approches qualitatives : relation binaire de préférence
  - requête skyline (Borzsony et al., 01), ordre partiel strict (Kiessling, 02), représentation logique (Chomicki, 03), sémantique ceteris paribus (Brafman, Domshlak, 04; Hadjali et al. 11)

$$\omega_Q(R) = \{r \in R | \neg \exists t' \in R, t' \prec t\}$$

- Approche quantitatives : fonction de score (Agrawal, Wimmers 00), top-k (Chaudhuri, Gravano 99, Hristidis et al. 01), **requête floue** (Tahani 77, Kacprzyk, Ziolkowski 86, Bosc, Pivert 95)
  - ex : voitures *très récentes* de kilométrage *faible*



Implémentations de SGBD flous

- De nombreux POC :
  - données relationnelles MS ACCESS (Kacprzyk, Zadrozny, 95), PostgreSQL (Smits et al., 13),
  - données graphes : NOSQL (Castelltort, Laurent, 14; Pivert et al., 16)

```
SELECT *,  get_mu() as mu
FROM cars
WHERE most(year ~= 'very recent', km ~= 'low, brand = 'VW')
ORBER BY mu LIMIT 10;
```

```
MATCH (x:author)-[authorof|ST IS strong]->(p:paper),
      (p:paper)-[:published]->(j:journal)-[:impactfactor]->(i:impactfactor),
      (j:journal)-[:domain]->(d) WHERE p.year IS recent
WITH x HAVING most(p) ARE (i.value IS high AND d.name="database")
RETURN x
```