Explications de données et de classifieurs : quelques méthodes et risques notables

Marie-Jeanne Lesot

Laboratoire d'Informatique de Paris 6

Sorbonne Université

PDIA 2021

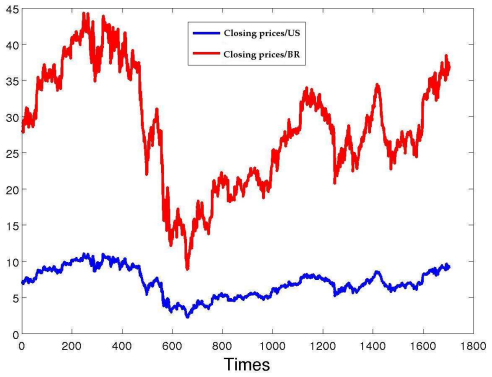
eXplainable Artificial Intelligence

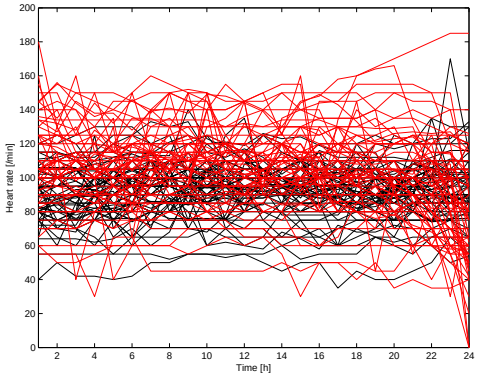
- **Omniprésence de l'IA** et complexité accrue des modèles
  - qualité des résultats obtenus dans de multiples domaines
  - réseaux profonds, XGBoost, forêts aléatoires : modèles boîtes noires
- **Dangers : risque de biais, d'opacité**, de discrimination, manque de confiance, incompréhension
  - point de vue institutionnel : droit à l'explication
  - RGPD : règlement général sur la protection des données, UE 2016
- Appel à projet de la DARPA 2017 : terme **XAI**
  - comme une implémentation du droit à l'explication
  - ajout de l'objectif d'interprétation
- Multiples termes liés : **interprétabilité, explicabilité, accountability**, transparence, équité, ...

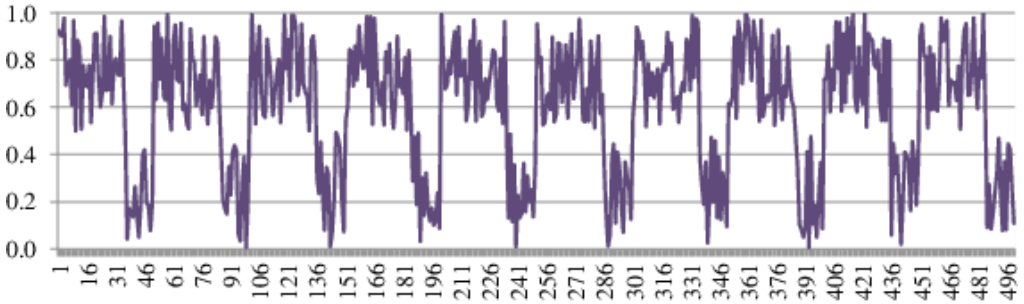
Plan

Résumé linguistique pour l'analyse exploratoire de données

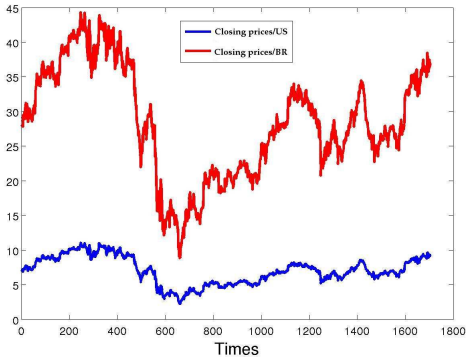
- **Faciliter la compréhension en rendant les données plus lisibles**
  - fournir une vue synthétique
  - exprimer sous forme linguistique
  - cadre **data-to-text**
- Illustration pour des séries temporelles financières ou médicales



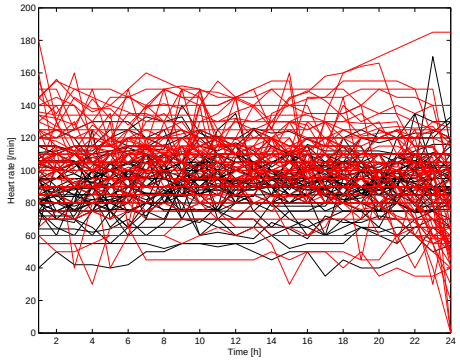




Illustration

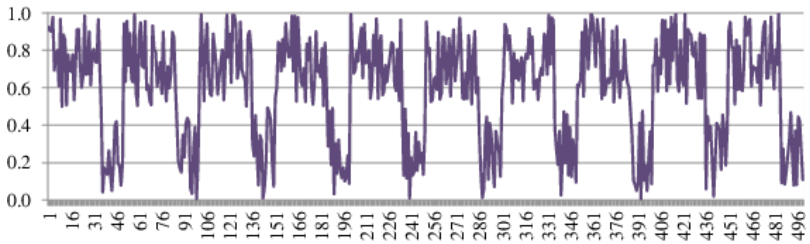


The higher the Brazilian closing price,
the higher the USA closing price,
especially if the USA closing price is in $[8.8; 10.4]$
or in $[17.4; 18.1]$



Few patients have a medium value of heart rate most of time

Almost all deceased patients have a very high value of HR, while recovered patients do not



Approximately every 30 minutes,
the data take high values

Trois familles principales

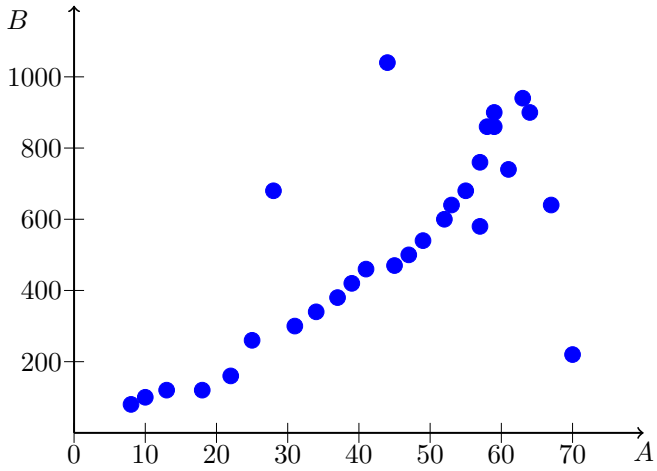
- Approches linguistiques (Reiter & Dale 00, Danlos 02)
- Résumés texte $\rightarrow$ texte (Sustkever 11, Bowman 16)
  - approches statistiques, p. ex. réseaux de neurones récurrents
- **Résumés linguistiques flous**, basés sur des **protoformes**

  p. ex. QRX **sont** P (Yager 82, Zadeh 83, Kacprzyk 02)
  - Q quantificateur, R qualificateur, P résumé : **sous-ensembles flous**
  - associés à un **degré de validité** $\eta \in [0, 1]$
  - exemples: *certains jeunes clients sont grands, 0.66*

    peu de patients ont un rythme cardiaque moyen, 0.9

Un exemple : les motifs graduels

- Expression linguistique
  - **plus A est V, plus B est W**
  - **plus A est élevé, plus B est élevé**
  - exemple : *plus la vitesse du vent est faible, plus la température est élevée*



- Formellement
  - données numérique : $\mathcal{D}$
  - item graduel : $(A, *)$, A un attribut, $* \in \{\leq, \geq\}$
  - motif graduel : $M = \{(A_j, *_j), j = 1..|M|\}$
  - ordre induit : $o \preceq_M o'$ ssi $\forall j = 1..|M|, A_j(o) *_j A_j(o')$

Un exemple : les motifs graduels

- **Piège d'interprétabilité !**
  - semblent intuitifs et faciles à comprendre
  - mais tant d'interprétations possibles !
  - et donc formalisation, critères de qualité et motifs extraits variables
- $\Longrightarrow$ Le destinataire des résumés considère-t-il la même interprétation que celle qui est implémentée ?

Interprétations possibles

the more M_1, the more M_2

plus la vitesse du vent est faible, plus la température est élevée

- **Généralisation floue des règles d'association**

  (Bouchon-Meunier & Desprès 90, Dubois & Prade 92, Hüllermeier 01)

  - la présence (floue) de M_1 implique la présence (floue) de M_2

$$support(M_1 \rightarrow M_2) = \sum_{x \in \mathcal{D}} i(M_1(x), M_2(x))$$

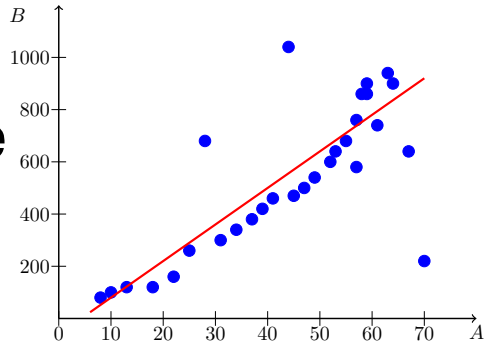
  - chaque donnée x est considérére individuellement

- **Co-variation d'attributs : tendance graduelle globale**

Interprétations possibles

the more M_1*, the more* M_2

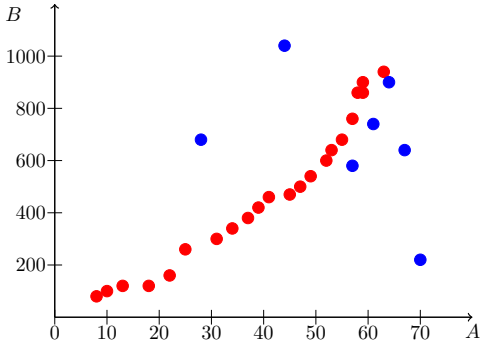
- **Corrélation de valeurs**: régression (Hüllermeier 02)
  - support : qualité de la régression et pente de la droite



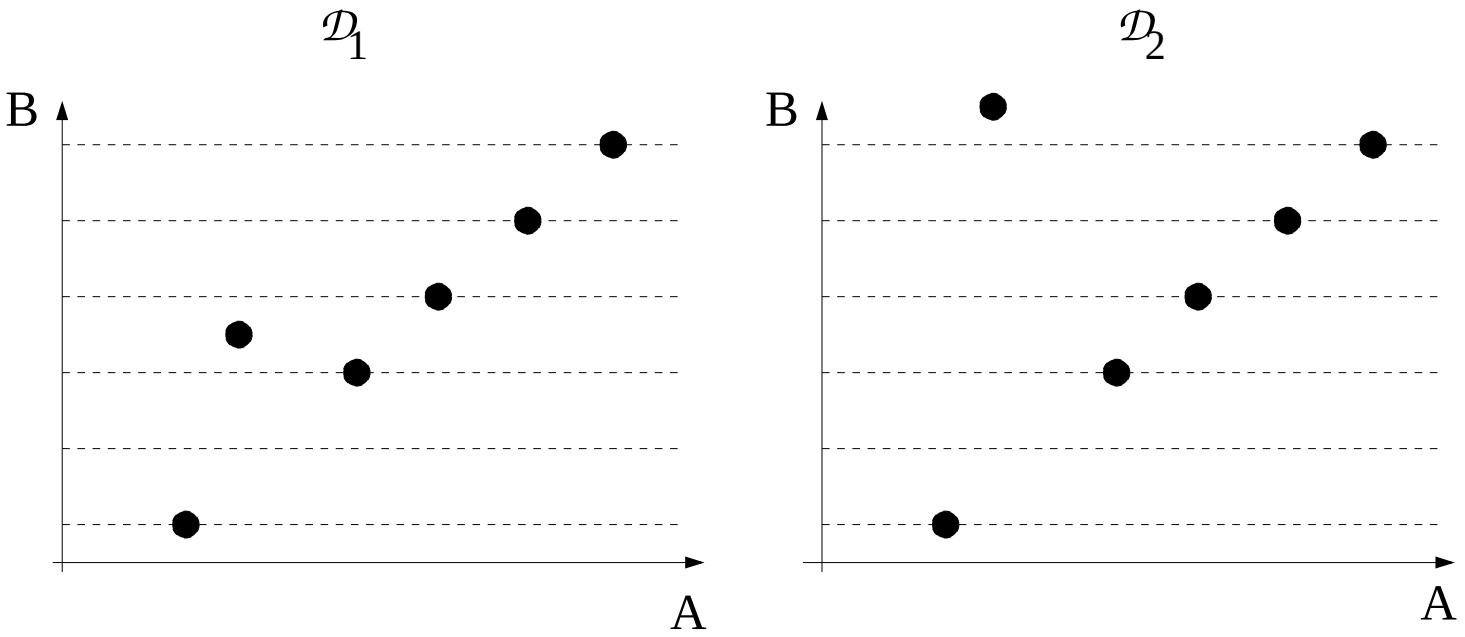
- **Corrélation des ordres induits**
  - support: proportion de couples de points satisfaisant les contraintes d'ordre (Berzal et al. 07)
  - tau de Kendall (GRAANK, FQAS 09)
- **Sous-ensembles de données compatibles**

(Di Jorio et al. 08, 09)

  - support : nombre maximal de données qui peuvent être ordonnées de façon à satisfaire les contraintes d'ordre



Différences d'interprétation



- Rôle de l'amplitude de déviation

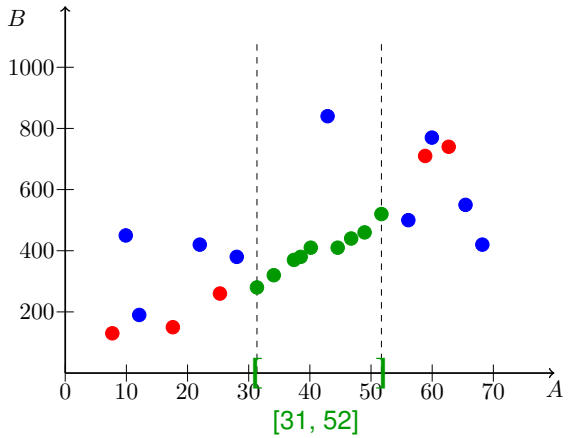
$$support_{\mathcal{D}_1}(A^{\geq}B^{\geq}) \quad \textbf{?} \quad support_{\mathcal{D}_2}(A^{\geq}B^{\geq})$$

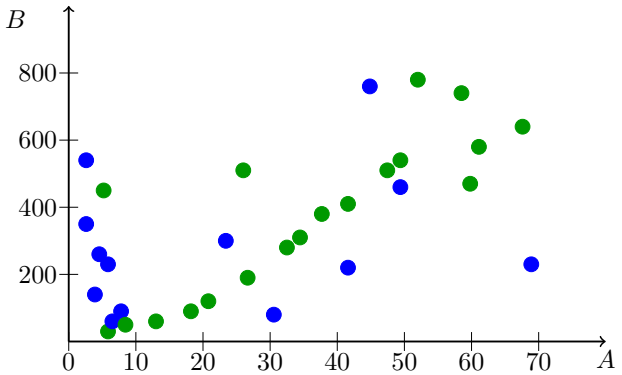
> ou **=**

Motifs graduels enrichis

(WCCI 10, EUSFLAT 13)

- Clauses additionnelles
  - caractérisation : *plus la vitesse du vent est faible, plus la température est élevée, surtout si la vitesse du vent* $\in [1, 10]$
  - renforcement : *plus la vitesse du vent est faible, plus la température est élevée, d'autant plus que l'humidité est faible*
- De même : **multiples choix d'interprétation !**
  - de formalisation, de définition de critères de qualité, de motifs extraits





Plan

Quelques mesures d'interprétabilité pour les résumés linguistiques

- Phrases : degré de vérité, spécialisation, longueur, adéquation

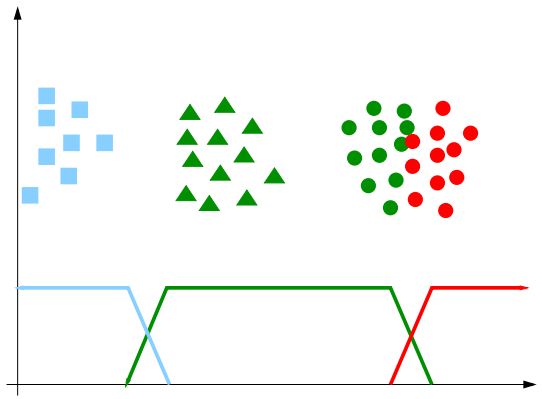
  (Kacprzyk et al 01, 09)

- Résumés : couverture, longueur, naturel (Castilla-Ortega et al. 12)
- Couples de phrases : inclusion, similarité, cohérence (Wilbik et al. 12)
- Vocabulaire : nombre de modalités, distinction entre elles, couverture, type de partition, régularité

  (Zhou et al. 08, Mencar et al. 08, Alonso et al. 09)

Adéquation aux données

(fuzzIEEE 13, IPMU 14, IFSA 17)



- Entrée : données numériques et vocabulaire
- **Double adéquation nécessaire**
  - mêmes descriptions numériques
    $\Rightarrow$ mêmes descriptions linguistiques
  - mêmes descriptions linguistiques
    $\Rightarrow$ membres des mêmes groupes
- Mesures de qualité : basées sur les clusters
  - **comparaison des partitions** induites dans les deux espaces
  - **évaluation croisée** de la qualité de la partition
- Méthode de révision du vocabulaire
  - modifications locales par **décomposition de modalités**
  - pour préserver l'interprétabilité par l'utilisateur

Adéquation linguistique

(SSCI FOCI 13)

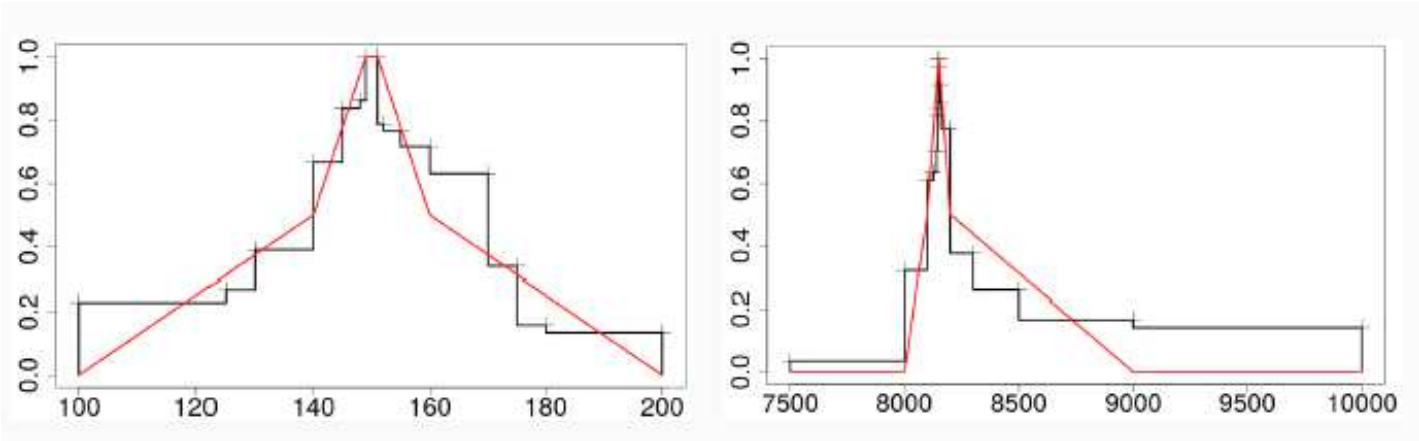
- Exemple de l'expression linguistique d'une période
  p. ex. environ toutes les 45 minutes, les valeurs sont élevées
  - **éviter les nombres trop grands ou trop petits**
    - "toutes les semaines" plutôt que "toutes les 168 heures"
  - **favoriser les entiers plutôt que les nombres décimaux**
    - "toutes les 45 minutes" plutôt que "toutes les 44.2 minutes"
  - **enrichir avec des adverbes** comme "environ"
    - pour exprimer la qualité de l'approximation
- Méthodologie en trois étapes
  - sélection de l'unité
  - approximation de la période
  - sélection de l'adverbe

Adéquation cognitive

(IPMU 16, IUI 17, IJAR 17)

- **Combien vaut "environ x"?**
  - identification de dimensions pertinentes à partir de données réelles
    - magnitude, dernier chiffre significatif, granularité, saillance cognitive
  - **modèle computationel** : compromis entre les plages de valeurs dénotées et la complexité des bornes

environ 150 environ 8150



Plan

- 1. Interprétation de données par résumés linguistiques
  - 1.1 Exemple des motifs graduels
  - 1.2 Questions d'interprétabilité et problème de choix des mots

  $\Longrightarrow$ **l'utilisation d'expressions linguistiques fournit des outils à l'interprétation des données et soulève des questions variées**
- 2. Interprétation de classifieurs par explications contre-factuelles
  - 2.1 Concepts-clés
  - 2.2 Principes
  - 2.3 Exemples de questions d'interprétabilité
- 3. Conclusions et perspectives

Plan

- 1. Interprétation de données par résumés linguistiques
  - 1.1 Exemple des motifs graduels
  - 1.2 Questions d'interprétabilité et problème de choix des mots

  $\Longrightarrow$ l'utilisation d'expressions linguistiques fournit des outils à l'interprétation des données et soulève des questions variées
- **2. Interprétation de classifieurs par explications contre-factuelles**
  - 2.1 Concepts-clés
  - 2.2 Principes
  - 2.3 Exemples de questions d'interprétabilité
- 3. Conclusions et perspectives

Ampleur du domaine

- Absence de consensus (Doshi-Velez et Kim 17, Lipton 17, Mueller et al. 19, Weller 19)
  - sur une définition formelle
  - sur des propriétés souhaitées
- Multiplicité des besoins
  - exemple : peur d'une prédiction erronée ou prédiction comme résultat intermédiaire
  - niveau d'expertise du destinataire
- Multiplicités d'approches, de catégorisations, d'axes de discussion

  (Guidotti et al. 18, Biran et Cotton 19, Artelt et Hammer 19, Carvalho et al. 19, Molnar 19)
- Problématique transdisciplinaire (Miller 19, Wachter et al. 18)
  - informatique, sciences cognitives, philosophie, droit, ...

Quelques axes de discussion

(Doshi-Velez et Kim 17, Lipton 17, Guidotti et al. 18 Carvalho et al. 19)

- Explication simultanée ou postérieure à la classification
  - auto-explication : classifieur générant ses propres explications
  - explication *post hoc*
- Connaissances disponibles : hypothèses d'agnosticité
  - le classifieur, le type de classifieur
  - les données d'apprentissage, leur distribution, d'autres données, des graphes causaux sur les attributs
- Explications globales ou locales
  - sur le classifieur en général, son comportement général
  - sur une prédiction particulière

Multiplicité des formes d'explications

- Représentation graphique
- Conditions suffisantes pour la classification : règles de décision
  - MES : sélection de la meilleure explication candidate par score d'information mutuelle par rapport au modèle à interpréter (Turner, 15)
  - LORE : extraction d'un arbre de décision appris sur le voisinage

    (Guidotti et al., 18)
- Vecteur d'importance d'attribut : *feature importance vector*
  - décroissance de précision quand on permute les valeurs d'un attribut

    (Breiman 01, Fisher et al. 19)
  - LIME : modèle linéaire local (Ribeiro et al. 16)
  - gradient du classifieur (Baehrens et al. 10, Selvaraju et al. 16)
  - SHAP et ses variantes (Stumberlj et al. 09, Lundberg et Lee 17)

Multiplicité des formes d'explications

- Classifieur lui-même, s'il est suffisamment simple
  - arbre de décision peu profond, régression parcimonieuse
  - *self-explaining classifier* (Alvarez Melis and Jaakoloa, 18)
  - classifieur de substitution ou *surrogate models*

(Craven et Shavlik 96, Hara et Hayashi 16)

- Données particulières
  - prototype (Kim et al. 14)
  - données d'apprentissage les plus influentes, identifiées par réapprentissage (Kabra et al. 15, Sharchilev et al. 18)
  - **explication contrefactuelle**

(Martens et Provost 14, Lash et al. 17, Wachter et al. 18)

Principes

- Raisonnement contrefactuel
  - **modifier en imagination l'issue d'un événement en modifiant l'une de ses causes**
  - ex : si James Dean avait pris le train le jour de son accident de voiture, il ne serait pas mort
  - étudié du point de vue de la philosophie, la psychologie, les sciences cognitives
- Cas des explications (Bottou et al 13, Wachter et al. 18, Artelt et Hammer 19)
  - analyser les prédictions en envisageant des modifications susceptibles de changer les conclusions
  - considérer des variantes de la donnée x à expliquer
  - **que faudrait-il changer pour avoir une prédiction différente ?**

Principes

- Variante de la donnée à expliquer : exemple typique (fictif)
  - la demande de crédit a été rejetée
  - pour qu'elle soit acceptée, le candidat aurait dû gagner 500 euros de plus par mois et avoir un accident de moins par an
- Motivations cognitives
  - mode de raisonnement naturel
  - aide à l'apprentissage : expliquer par l'exemple, en comparant
    - exemple : si l'animal avait des oreilles plus longues, ce serait un lièvre et non un lapin
- Motivations pratiques
  - donne des indications à l'utilisateur sur les actions éventuelles à réaliser

Formalisation

- Etant donné un classifieur f et une donnée x
  - construire e tel que $f(e) \neq f(x)$
  - changement à apporter : $e - x$ = explication

Formalisation

- Etant donné un classifieur f et une donnée x
  - construire e tel que $f(e) \neq f(x)$ **en minimisant l'effort**
  - changement **minimal** à apporter : $e - x$ = explication

Formalisation

- Etant donné un classifieur f et une donnée x
  - construire e tel que $f(e) \neq f(x)$ **en minimisant l'effort**
  - changement **minimal** à apporter : $e - x$ = explication
- soit

$$e^* = \arg\min_{e \in \mathcal{X}} c_x(e) \qquad \text{tel que} \quad f(e) \neq f(x)$$

Formalisation

- Etant donné un classifieur f et une donnée x
  - construire e tel que $f(e) \neq f(x)$ **en minimisant l'effort**
  - changement **minimal** à apporter : $e - x$ = explication
- soit

$$e^* = \arg\min_{e \in \mathcal{X}} c_x(e) \qquad \text{tel que} \quad f(e) \neq f(x)$$

- A définir :
  - la fonction de coût c_x
  - l'espace de recherche pour e
  - la méthode d'optimisation

Quelques exemples

$$e^* = \arg\min_{e \in \mathcal{X}} c_x(e) \qquad \text{tel que} \quad f(e) \neq f(x)$$

- Fonction de coût c_x
  - p. ex. distances à x : l_2, l_1, l_0 (Lash et al. 17, Wachter et al. 18, Guidotti et al. 18)
  - coût non uniforme sur tous les attributs
- Espace de recherche : $\mathcal{X}$ ou un sous-ensemble
  - notion d'attribut actionnable
  - $\mathcal{X} = \mathcal{X}_d \cup \mathcal{X}_i \cup \mathcal{X}_u$ (Lash et al. 17)
  - corrélation entre attributs, voire causalité
- Méthode d'optimisation suivant les hypothèses d'agnosticité
  - p. ex. méthodes efficaces pour classifieur linéaire (Ustun et al. 19, Russell et al. 19)
  - échantillonnage aléatoire (Laugel et al. 18)

Growing Spheres

(Laugel et al. 18)

$$e^* = \arg\min_{e \in \mathcal{X}} c(e) \qquad \text{tel que} \quad f(e) \neq f(x)$$

- Fonction de coût $c_x(e) = ||e - x||_2 + ||e - x||_0$
- Optimisation séquentielle
  - minimisation de l_2 par une méthode de Monte Carlo

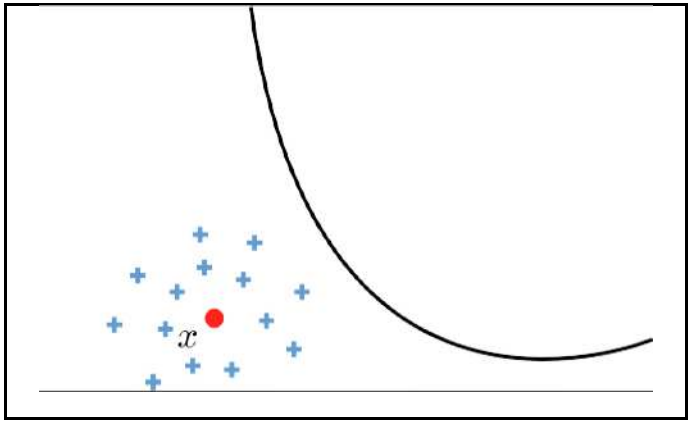
$$\tilde{e} \approx e^* = \arg\min_{z \in \mathcal{X}} \{||x - z||_2 \quad \text{tel que} \quad f(z) \neq f(x)\}$$

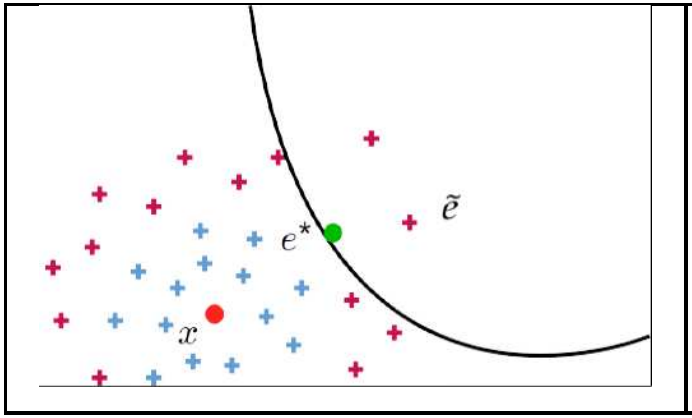
  - minimisation de l_0 : rendre le résultat intermédiaire $\tilde{e}$ parcimonieux

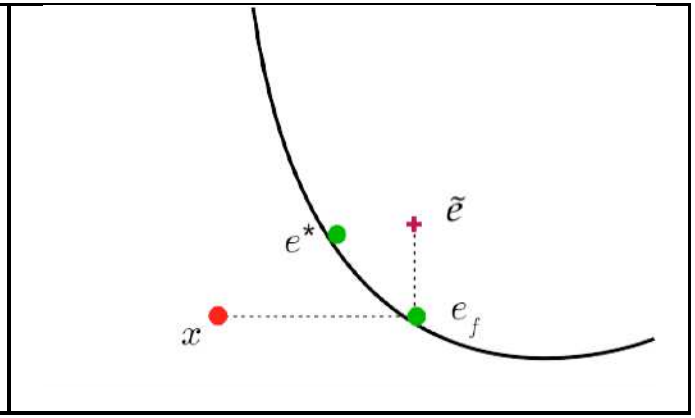
$$e_f = \arg\min_{e \in \mathcal{P}_{\tilde{e}}} ||e - x||_0 \qquad \text{tel que} \quad f(e) \neq f(x)$$

Growing Spheres

(Laugel et al. 18)







LORE: LOcal Rule-based Explanation

(Guidotti et al. 18)

- Composante d'explication par classifieur de substitution
  - génération de données équilibrées autour de x,
    par algorithme génétique
  - apprentissage d'un arbre de décision C4.5 : approximation locale du classifieur f
  - règle : chemin de x dans l'arbre
- Composante d'explication contrefactuelle
  - pour chaque feuille $\mathcal{Q}$ de l'arbre qui prédit $l \neq f(x)$
  - compter le nombre d'attributs à modifier pour affecter x à $\mathcal{Q}$
  - explication : chemins des feuilles de score minimal

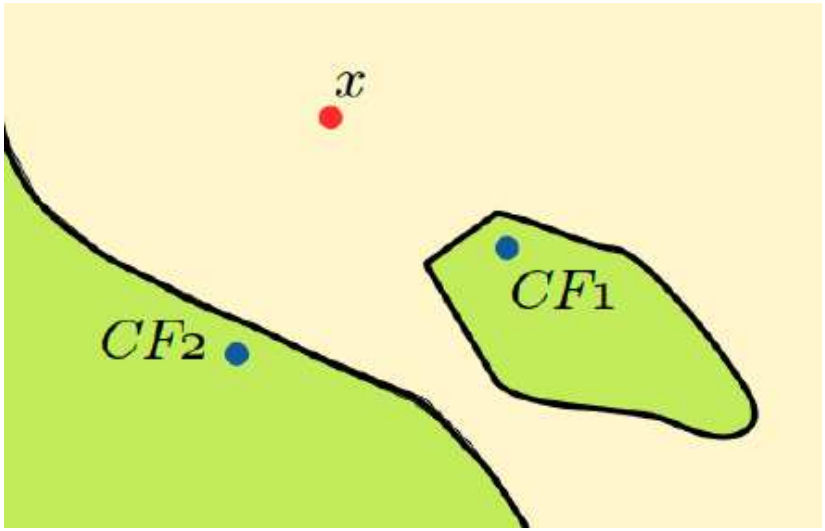
Plan

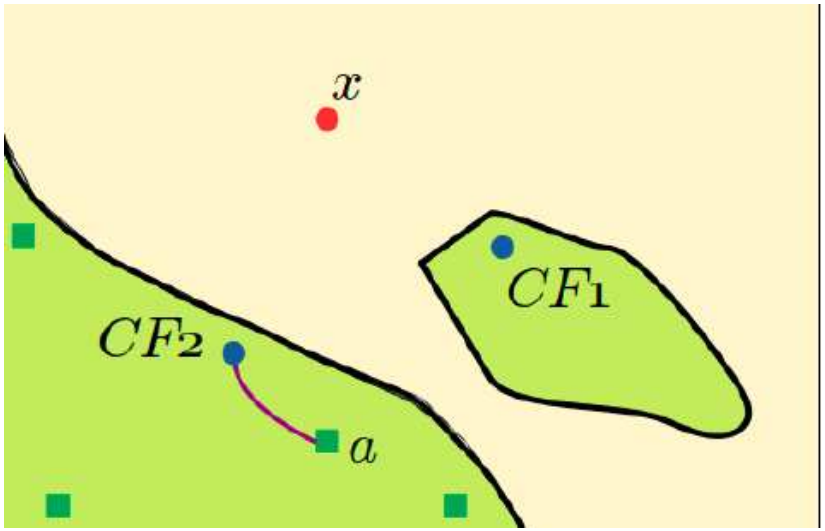
Risque d'explication non justifiée

(Laugel et al. 19)

- Explication contrefactuelle liée à des artefacts du classifieur
- Proposition d'une méthode de diagnostic
  - utilisant les données d'apprentissage X

$\Longrightarrow$ les méthodes existantes sont très sensibles à ce risque

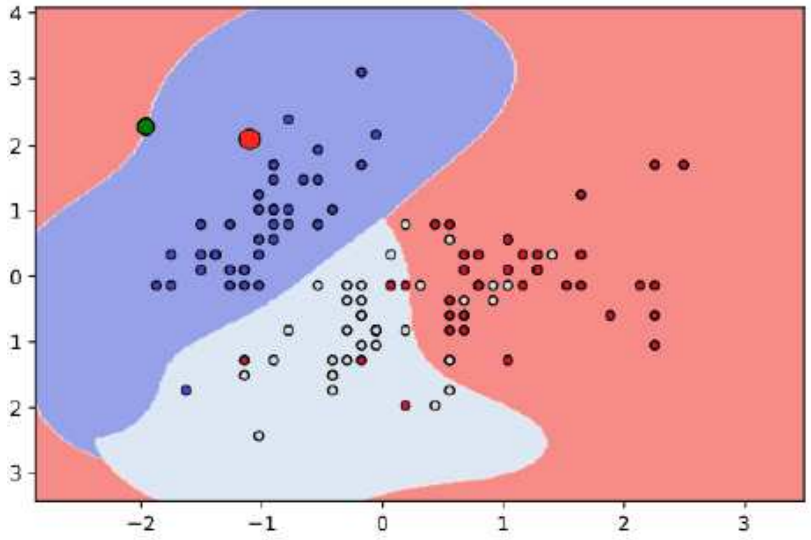




Risque d'explication hors distribution

(Laugel et al. 19)

- Le classifieur peut "improviser" dans les zones où il ne dispose pas de données



- Risque d'explication contrefactuelle non pertinente : par exemple
  - le demandeur devrait être âgé de 135 ans
  - le demandeur devrait avoir le même âge et un niveau d'étude strictement supérieur

Eviter les explications hors distribution

- **Hypothèses d'agnosticité moins radicales**

- Cas où des données sont disponibles
  - FACE : Feasible and Actionable Counterfactual Explanations

    (Poyiadzi et al. 20)

  - trouver un chemin pondéré optimal entre x et e
  - passant par les données disponibles où la densité estimée est supérieure à un seuil
  - avec poids des arcs entre données $w_{ij} = p\left(\frac{x_i+x_j}{2}\right) d(x_i, x_j)$

Eviter les explications hors distribution

- **Hypothèses d'agnosticité moins radicales**

- Cas où la distribution des données est connue (Artelt et Hammer 20)
  - contrainte additionnelle $p_{classe}(e) \geq \delta$
  - e dans région à densité élevée

- Graphe de causalité (Mahajan et al. 19)
  - score causal pénalisant les valeurs d'attributs non compatibles
  - e réaliste

Plan

En guise de conclusion

Merci à : Anne Laurent, Amal Oudni, Maria Rifqi, Bernadette Bouchon-Meunier, Grégory Smits, Olivier Pivert, Gilles Moyse, Sébastien Lefort, Charles Tijus, Elisabetta Zibetti, Thibault Laugel, Christophe Marsala, Marcin Detyniecki, Xavier Renard

- **XAI : un domaine très riche, passionnant**
  - non figé, en pleine expansion
  - absence de consensus, composante subjective
  - multiples approches, multiples questionnements
  - à la croisée de plusieurs disciplines
  - **perspective et défi de l'IA !**
- Multiples sessions spéciales, workshops, conférences dédiées
  - WHI/HiLL@ICML, XAI@IJCAI, FATML@KDD, NeurIPS, IUI, ...
  - AIES, FAT*, CHI, ...

Et pour citer d'autres questions

- Questions de **localité**, de robustesse
- Questions de **présentation des résultats**
  - parfois trop techniques, tournés vers des utilisateurs experts en IA
  - domaine des **interfaces explicatives**
  - possibilité d'expression linguistique

  interprétabilité de "si le demandeur gagnait 317.62 euros de plus par mois" ?
- Questions d'**évaluation** : nécessité de mettre l'utilisateur au cœur
  - a-t-il compris ? quel est l'"objectif pédagogique" ?
  - sait-il quoi faire ? sans le pousser à la fraude
- De telles approches vont-elles rétablir la confiance des utilisateurs ou donner des outils pour les manipuler ?