Explicabilité des Modèles d'Apprentissage Automatique.

G. Laberge

Polytechnique Montréal

14 octobre 2020

Croissance de l'Apprentissage Automatique

Dûe à la croissance exponentielle des tailles des ensembles de données (*BigData*) et de la puissance des ordinateurs.

1. https://www.20minutes.fr/high-tech/1804887-20160312-jeu-go-ordinateur-google-remporte-3e-duel-face-champion-monde

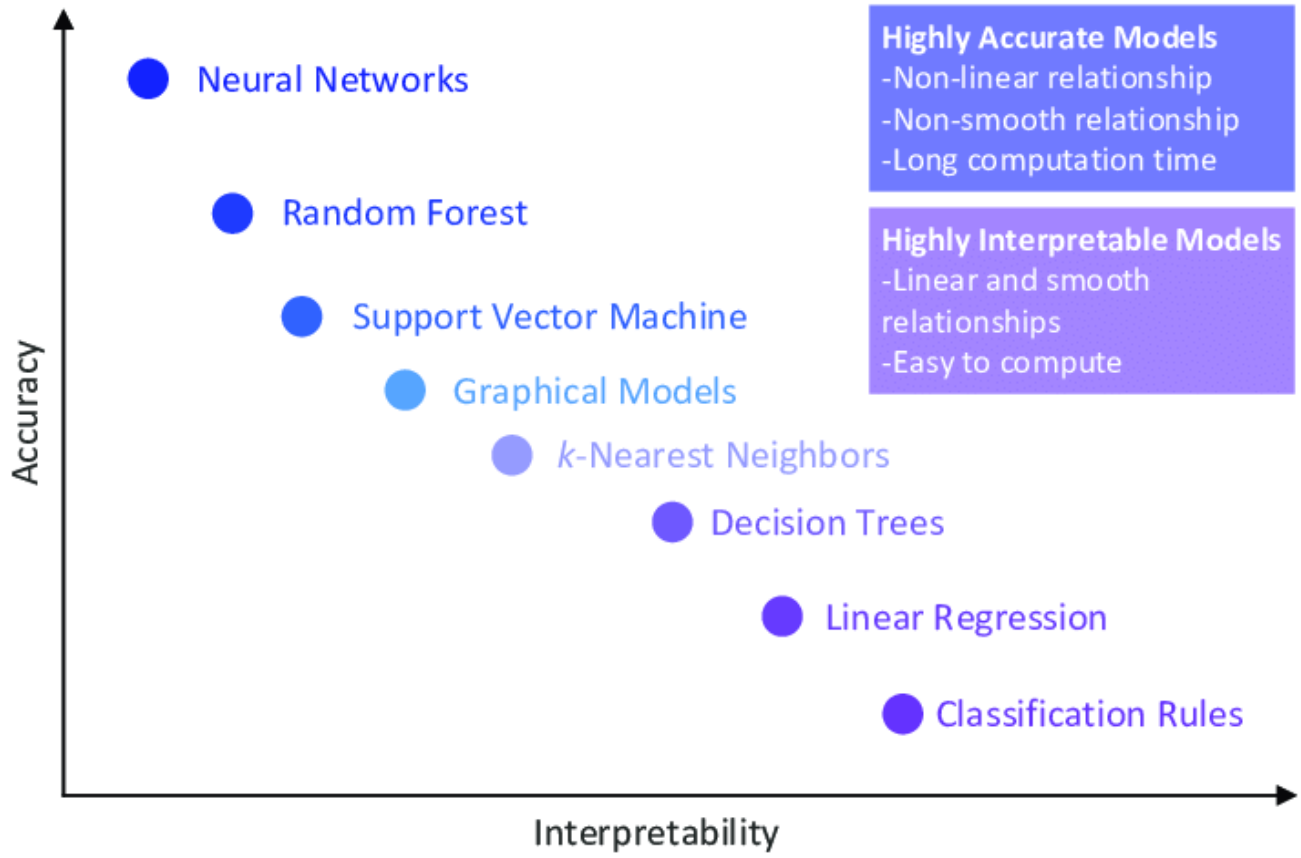
Croissance de l'Apprentissage Automatique

Dûe à la croissance exponentielle des tailles des ensembles de données (*BigData*) et de la puissance des ordinateurs.

1

1. https://www.20minutes.fr/high-tech/1804887-20160312-jeu-go-ordinateur-google-remporte-3e-duel-face-champion-monde

Problématique Actuelle



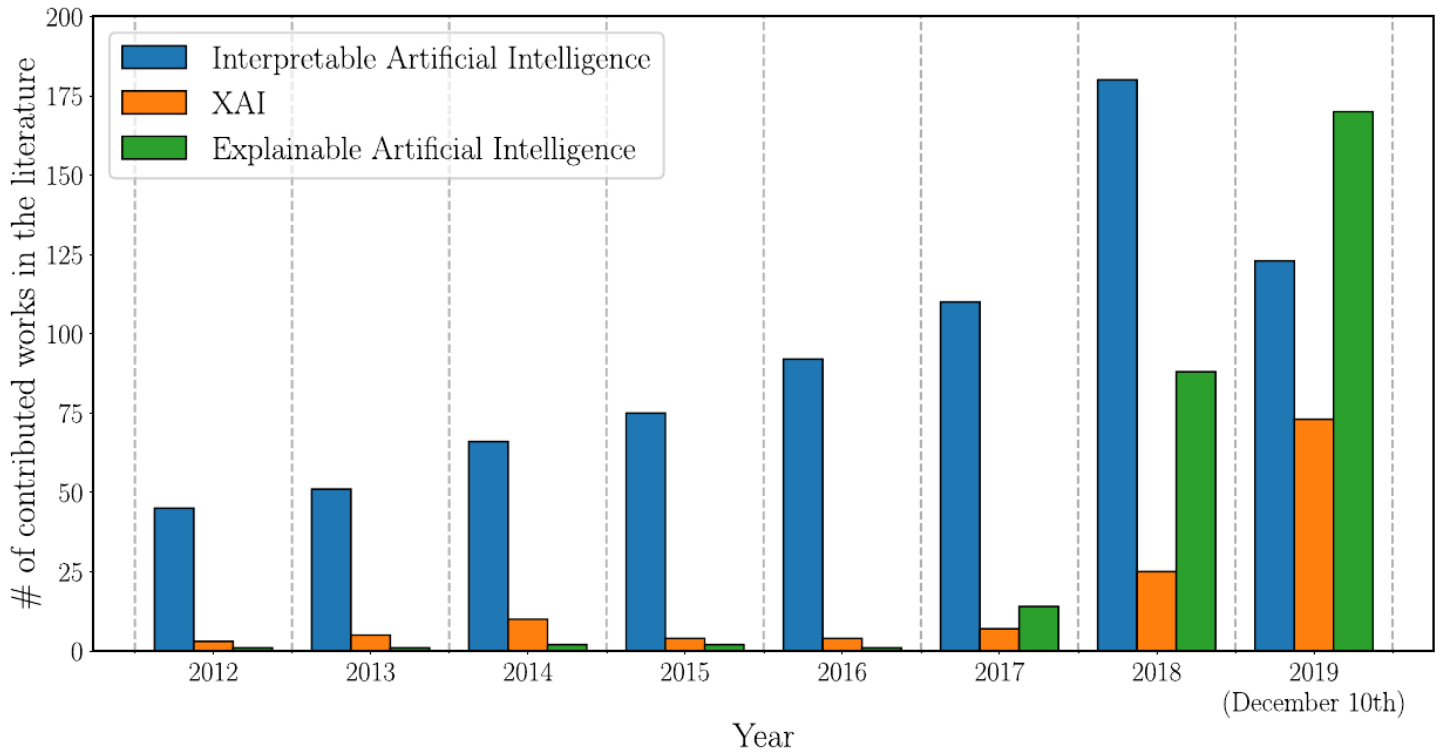
2

2. https://www.researchgate.net/publication/335937022_Machine_Learning_for_5GB5G_Mobile_and_Wireless_Communications_Potential_Limitations_and_Future_Directions

Recherche Actuelle

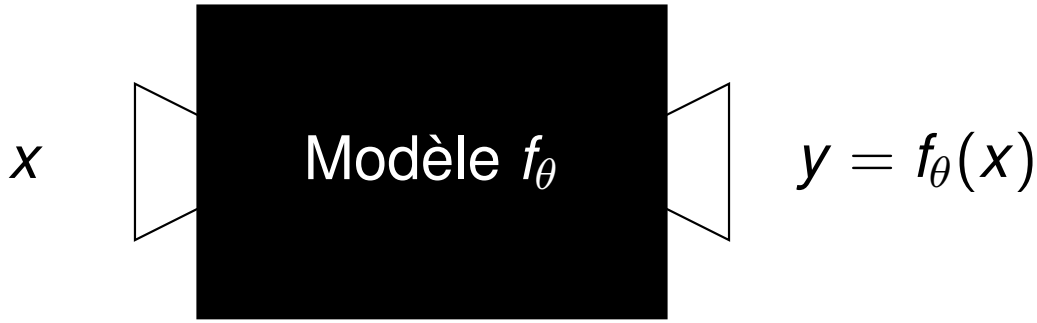
Développement de mécanismes d'explications.

1. DEEL (DEpendable and Explainable Learning)
2. XAI (eXplainable Artificial Intelligence)

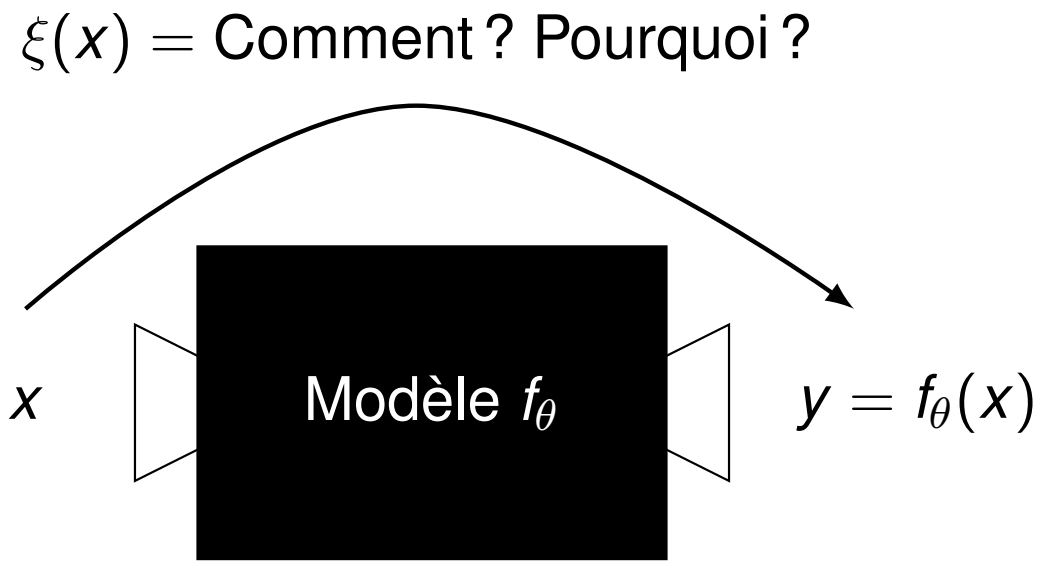


A. Barredo Arrieta, et al. (2020)

Qu'est-ce qu'une explication ?



Qu'est-ce qu'une explication ?



Notes :

1. Le modèle f_θ a déjà été calibré.
2. L'explication ξ dépend de l'entrée x.

Qualité d'une Explication

Une bonne explication est :

1. **Fidèle** : Décrit bien le processus décisionnel du modèle.
2. **Interprétable** : Peut être comprise par un être humain.
3. **Généralisable** : Peut s'appliquer à divers types de modèles.

Classification des Explications

Importance des Caractéristiques

1. \+ Identifier les composante de la variable d'entrée x qui ont eux le plus d'influence sur la décision.
2. \- Résultats contre intuitifs si les composantes de x sont hautement corrélées.

Classification des Explications

Importance des Caractéristiques

1. \+ Identifier les composante de la variable d'entrée x qui ont eux le plus d'influence sur la décision.
2. \- Résultats contre intuitifs si les composantes de x sont hautement corrélées.

Explication Locale

1. \+ Approximer localement f_θ par un modèle simplifié dont on peut expliquer les décisions.
2. \- Notion de localité mal définie en haute dimension.

Classification des Modèles

Spécifique au modèle

1. + Obtenir des explications exploitants la structure interne du modèle.
2. + Implémentation plus efficace.
3. - Moins généralisable.

Classification des Modèles

Spécifique au modèle

1. + Obtenir des explications exploitants la structure interne du modèle.
2. + Implémentation plus efficace.
3. - Moins généralisable.

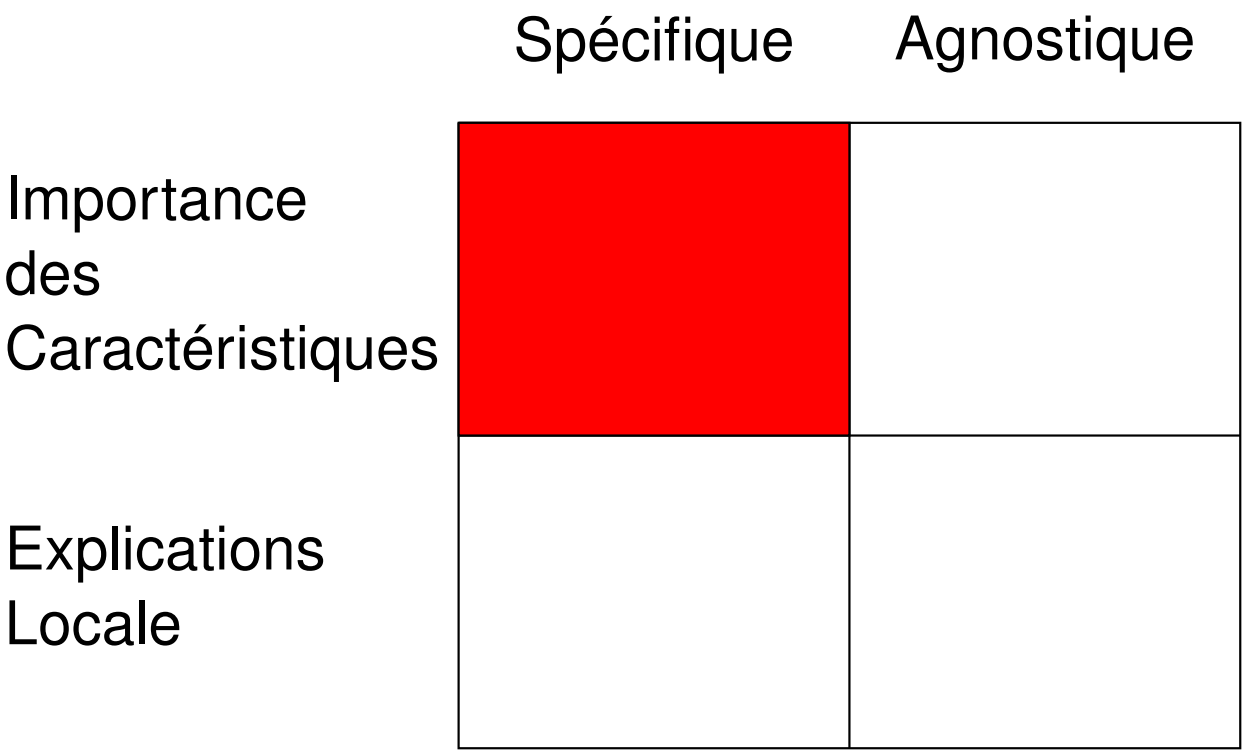
Agnostique au modèle

1. + Expliquer n'importe quel modèle.
2. + Beaucoup de codes en libres accès (LIME, SHAP, anchors)
3. - Peut souffrir de basse fidélité.
4. - Basée sur des perturbations de x. (Susceptible aux attaques adverserielles)

Tableau des Explications

	Spécifique	Agnostique
Importance des Caractéristiques		
Explications Locale		

Tableau des Explications



Explications à l'aide du Gradient

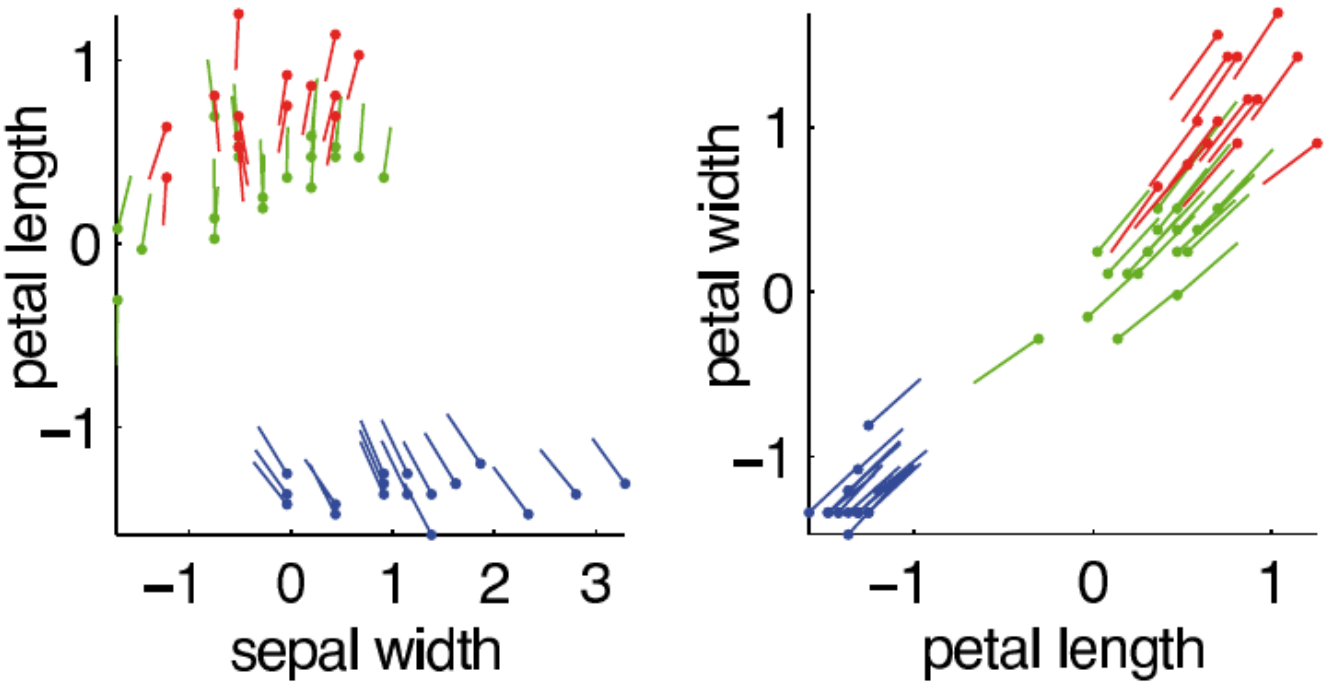
Explication de classificateurs qui modélisent la probabilité $p_c(x)$ associé à une classe c.

$$\xi(x) = \nabla p_c(x) \tag{1}$$

Explications à l'aide du Gradient

Explication de classificateurs qui modélisent la probabilité $p_c(x)$ associé à une classe c.

$$\xi(x) = \nabla p_c(x) \tag{1}$$



D. Baehrens, et al. (2010)

Application aux Réseaux de Neurones

K. Simonyan, et al. (2014)

- Peut donner des explications très bruitées et diffuses.

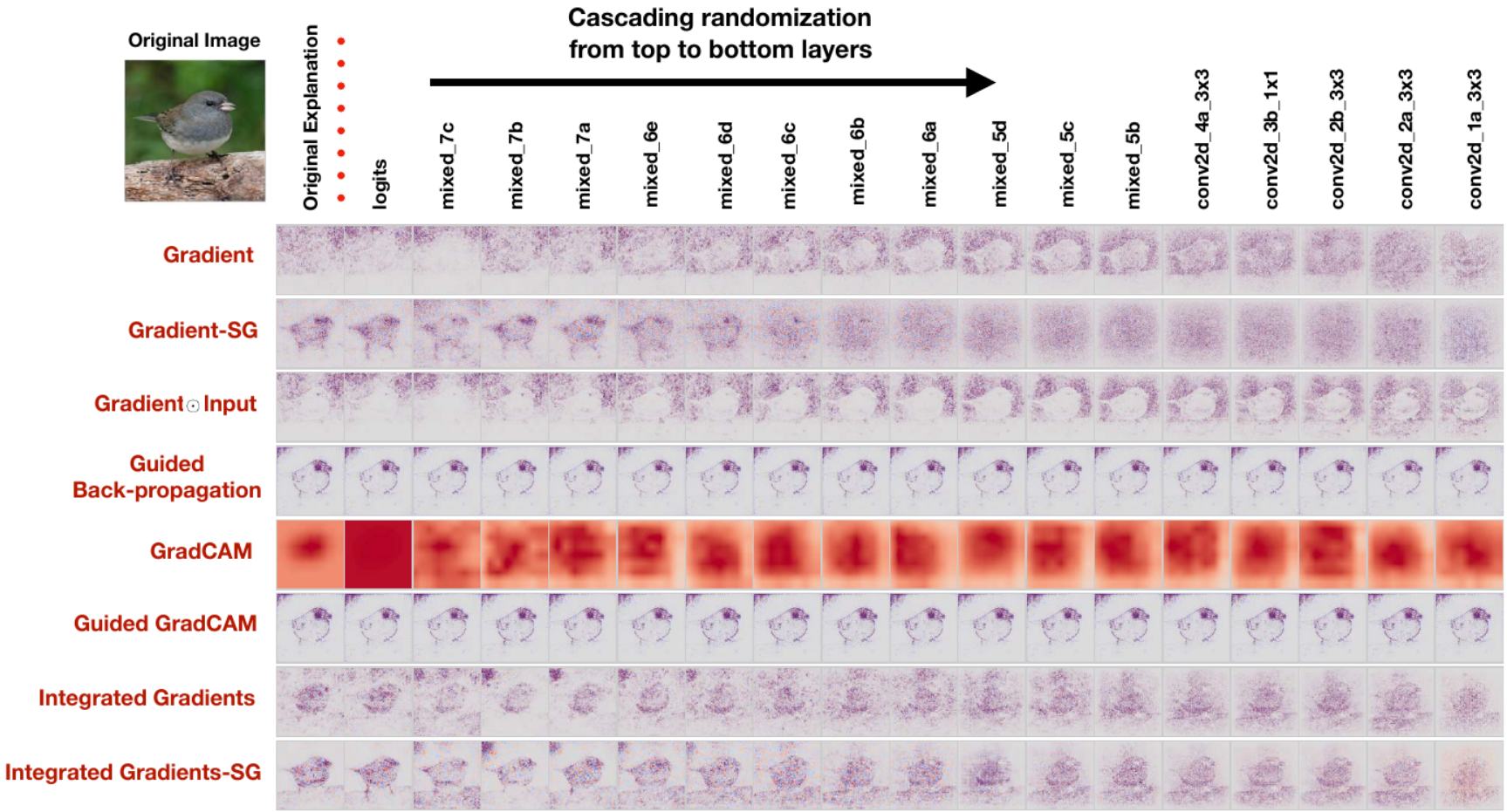
Améliorations Possibles

1. Guided Back-Propagation (J. T. Springenberg et al. , 2015)
2. Integrated Gradient (M. Sundararajan et al. , 2017)
3. Smooth-Grad (D. Smilkov et al. , 2017)
4. GradCAM (R. R. Selvaraju et al. , 2019)

Disponibles en libre accès

1. TensorFlow `https://github.com/PAIR-code/saliency`
2. Pytorch `https://github.com/hs2k/pytorch-smoothgrad`

Manque de Sensibilité aux paramètres du réseau.



J. Adebayo, et al. (2018) `https://github.com/jendawkins/saliencySanity`

Attaques Adverserielles.

A. K. Dombrowski, et al (2019) `https://github.com/pankessel/explanations_can_be_manipulated`

Tableau des Explications

	Spécifique	Agnostique
Importance des Caractéristiques	$\xi(x_0) = \nabla p_c(x)$ - Bruit et diffusion - Manque de sensibilité - Attaques adverserielles	
Explications Locale		

Explications à l'aide de Perturbations

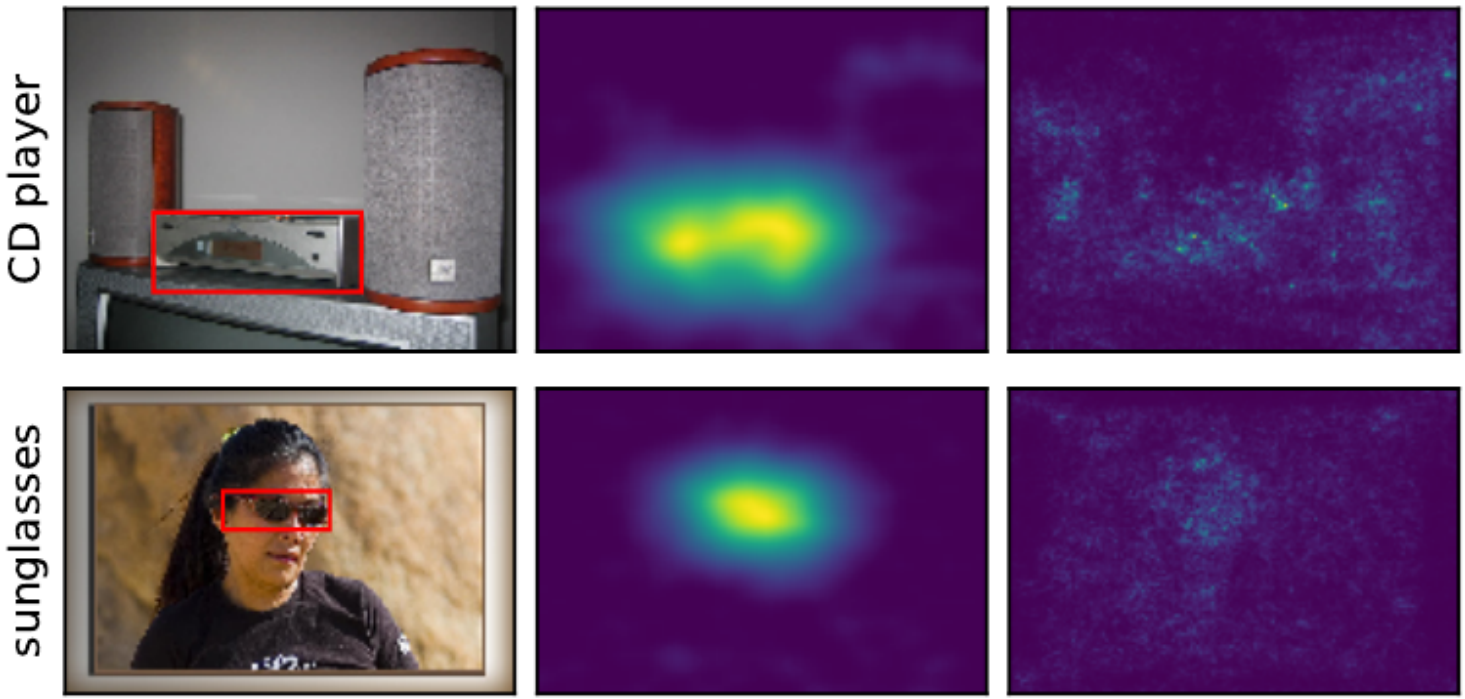
Soit une images avec d pixels. On applique un masque $m \in [0,1]^d$

On choisi le masque m qui affecte le plus la probabilité $p_c(x)$.

Explications à l'aide de Perturbations

Soit une images avec d pixels. On applique un masque $m \in [0,1]^d$

On choisi le masque m qui affecte le plus la probabilité $p_c(x)$.



R. C. Fong, et al. (2017) `https://github.com/dizcza/pytorch-mighty`

SHapley Additive exPlanations (SHAP)

Dans cette méthode, les masques reviennent à garder ou enlever des composants i.e. $m \in \{0,1\}^d$.

La *Shapley Value* de chaque composante est ensuite calculée.
(S. M. Lundberg et al. 2017)

1. + Libre accès `https://github.com/slundberg/shap`.
2. - Couteux.
3. - Suppose l'indépendance des composantes de x.
4. - Basées sur des perturbations de x. (Susceptible aux attaques adverserielles)

Tableau des Explications

	Spécifique	Agnostique
Importance des Caractéristiques	$\xi(x) = \nabla p_c(x)$ - Bruit et diffusion - Manque de sensibilité - Attaques adverserielles	$\xi(x) \in \mathbb{R}^d$ - Couteux - Attaques adverserielles
Explications Locale		

Local Interpretable Model-agnostic Explanations

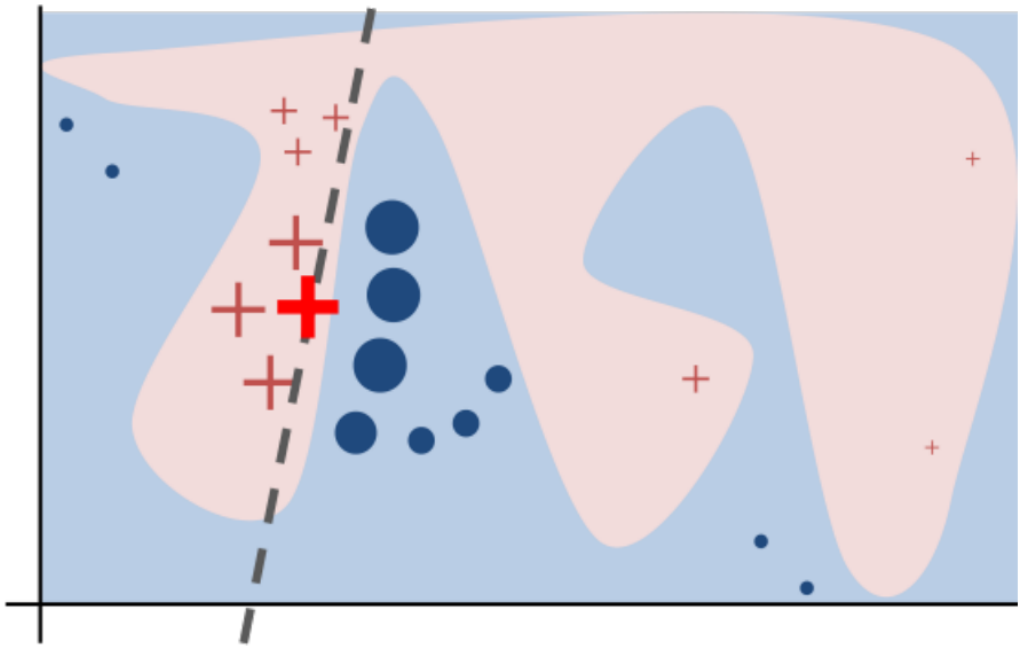
On approximate **localement** le modèle f_θ par une modèle g interprétable.

$$\xi(x) = g \tag{2}$$

Local Interpretable Model-agnostic Explanations

On approximate **localement** le modèle f_θ par une modèle g interprétable.

$$\xi(x) = g \tag{2}$$

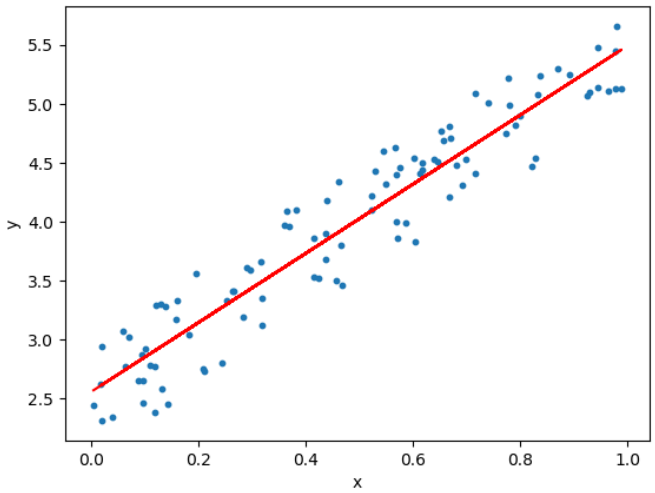


M. T. Ribeiro, et al. (2016)

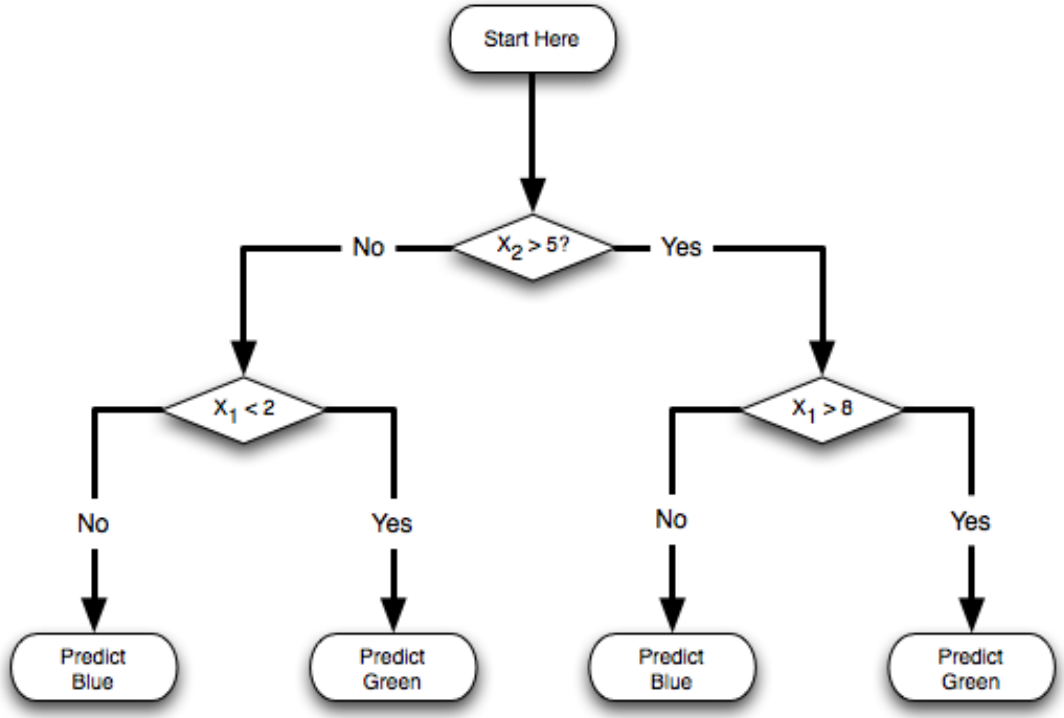
Modèle Interprétables

g : Régression Linéaire

$$y = g(x) = \omega x + b$$



g : Arbre de décision



https://towardsdatascience.com/linear-regression-using-python-b136c91bf0a2
https://alliance.seas.upenn.edu/~cis520/wiki/index.php?n=Lectures.DecisionTrees

Comment Perturber x ?

3 différent modules de LIME.

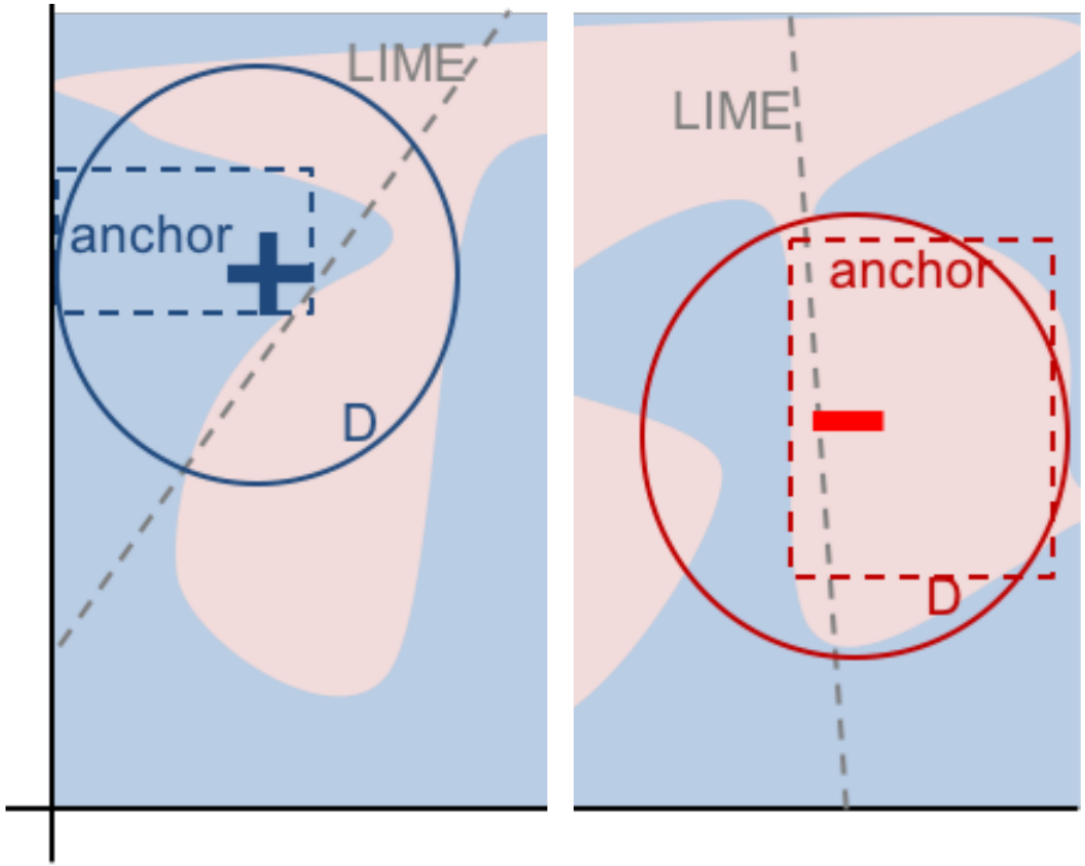
1. **lime_text.py** Perturbe un texte en enlevant certains mots.
2. **lime_image.py** Perturbe une image en enlevant certains super-pixels.
3. **lime_tabular.py** En generale quand $x \in \mathbb{R}^d$, applique une perturbation $x'_i = x_i + \epsilon_i$ avec $\epsilon \sim \mathcal{N}(0, \sigma_i^2)$.

Forces et Faiblesses

1. + Simple à expliquer
2. + Libre accès `https://github.com/marcotcr/lime`.
3. - Localité mal définie en haute dimension.
4. - La région où l'explication est valide n'est pas clairement définie.
5. - Basées sur des perturbations de x. (Susceptible aux attaques adverserielles)

Anchors

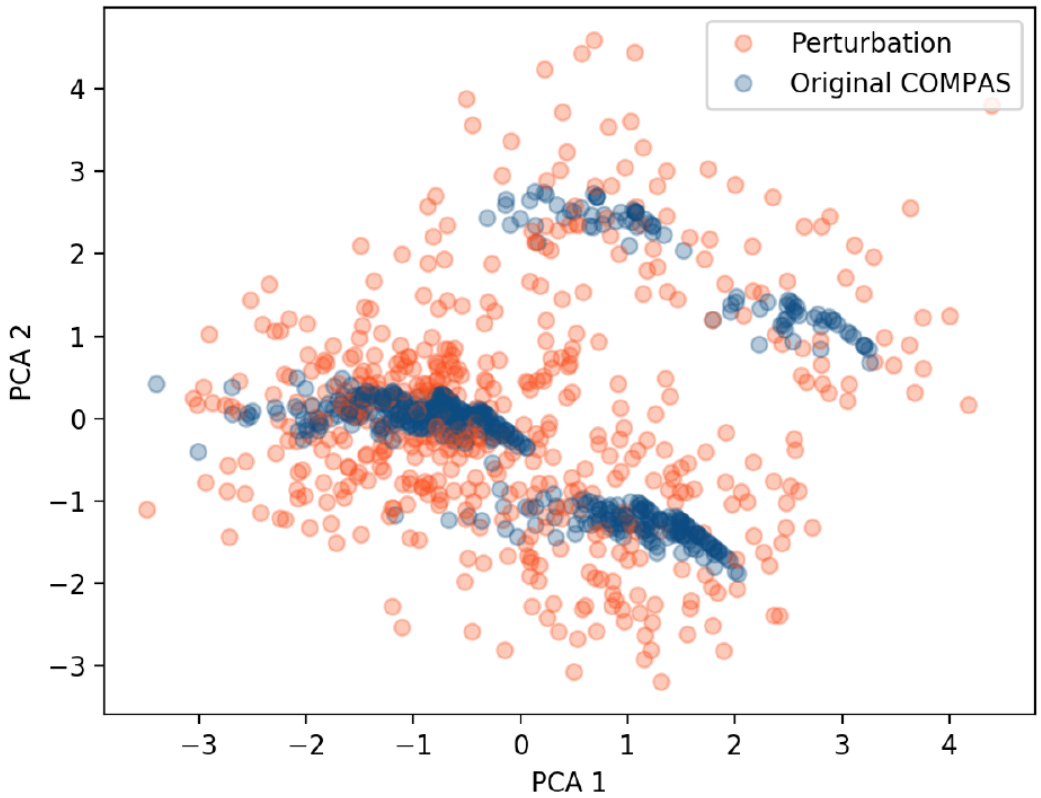
Des explications locales avec un support bien définie. Basé sur des règles **If** ($x_1 < 10$) **and** ($x_2 > 12$) **Then (y = 0)**



M. T. Riberio et al. (2018) `https://github.com/marcotcr/anchor`

Attaques sur LIME et SHAP

Les perturbations de x sont **hors distribution**. On peut ainsi créer un modèle très biaisé, mais que LIME et SHAP ne peuvent diagnostiquer.



D.Slack et al. (2020) `https://github.com/dylan-slack/Fooling-LIME-SHAP`

Tableau des Explications

	Spécifique	Agnostique
Importance des Caractéristiques	$\xi(x) = \nabla p_c(x)$ - Bruit et diffusion - Manque de sensibilité - Attaques adverserielles	$\xi(x) \in \mathbb{R}^d$ - Couteux - Attaques adverserielles
Explications Locale		$\xi(x) = g$ - Localité mal définie. - Attaques adverserielles

Question 1 de recherche.

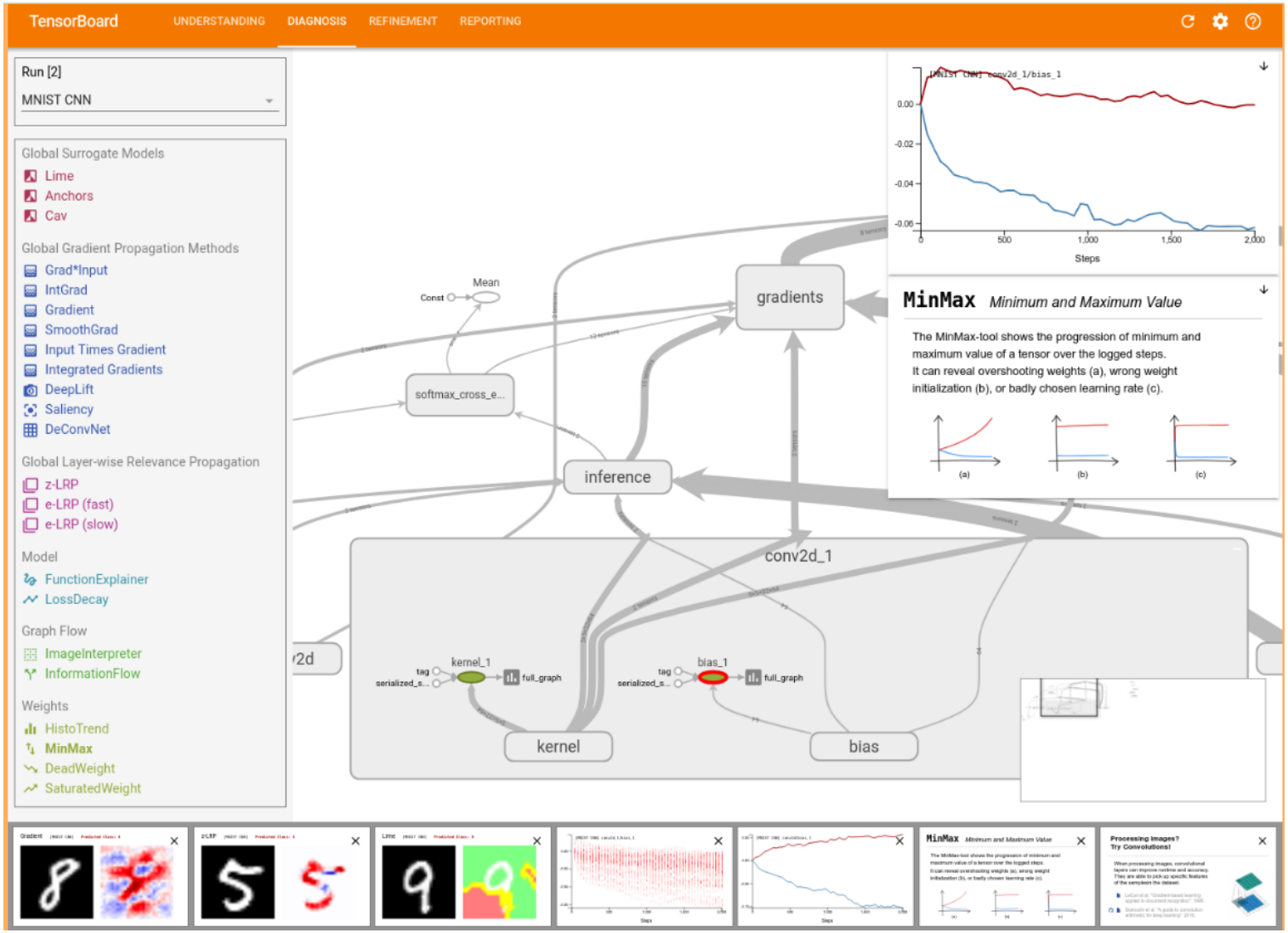
Quel est le gain en information obtenu par l'utilisation d'une explication spécifique au modèle plutôt qu'une explication agnostique au modèle ?

Étude Empirique.

Familiarisation avec les codes en libre accès

1. TensorFlow `https://github.com/PAIR-code/saliency`
2. Pytorch `https://github.com/hs2k/pytorch-smoothgrad`
3. Shap `https://github.com/slundberg/shap`
4. LIME `https://github.com/marcotcr/lime`
5. Anchors `https://github.com/marcotcr/anchor`

ExplAIner, une extension de TensorBoard.

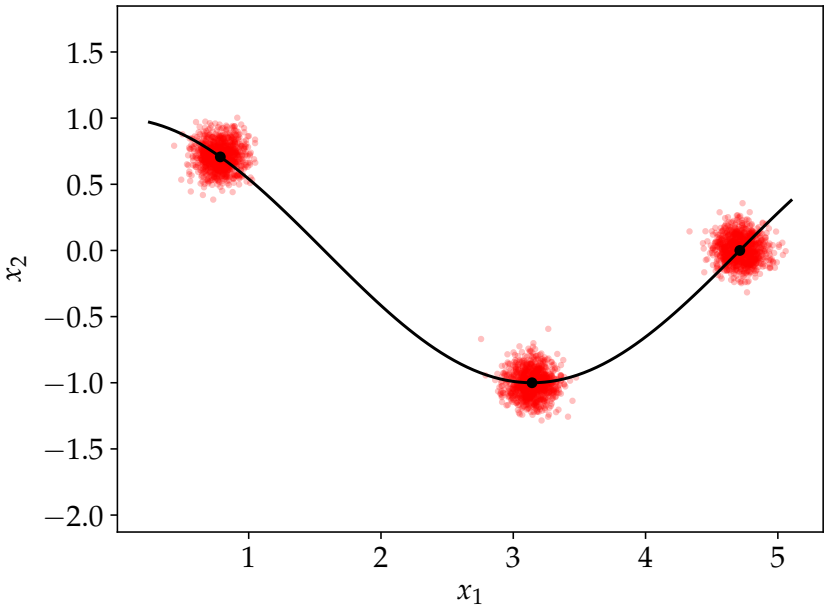


T. Spinner et al. (2020)

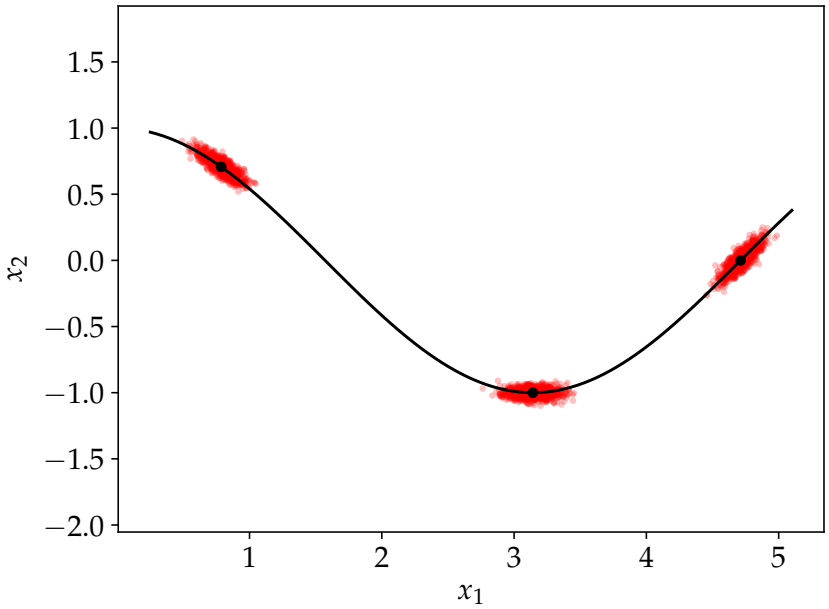
Question 2 de recherche.

Peut-on améliorer LIME dans le cas de données tabulaire en applicant des perturbations qui tiennent compte des structures locales des données ?

Manifold Parzen Window.



(a) Bruit Gaussien.



(b) Manifold Parzen Window.