# La robustesse : un nouveau critère d'évaluation des explications en apprentissage automatique

ATELIER FENDER, 8 JUILLET 2021

**Robin Cugny,** Julien Aligon, Max Chevalier, Geoffrey Roman Jimenez, Chantal Soulé-Dupuy, Olivier Teste

# Sommaire

# L’explicabilité dans l’apprentissage automatique

## Motivations pour l’explicabilité

### Demandes côté utilisateur [1]

- **Traçabilité des données personnelles**
- **La fiabilité**

- **La causalité**
- **La confiance**

- **L’équité**

### Usage côté développeur [2]

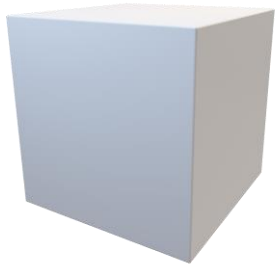
- **Déboguer et améliorer**
- **Fuite de données**
- **Décalage dans les données**


1. Doshi-Velez, F., & Kim, B. (2017).
2. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August).

# L’explicabilité dans l’apprentissage automatique

## Taxonomie

Approches globales



$X \xrightarrow{?} Y$

Approches locales

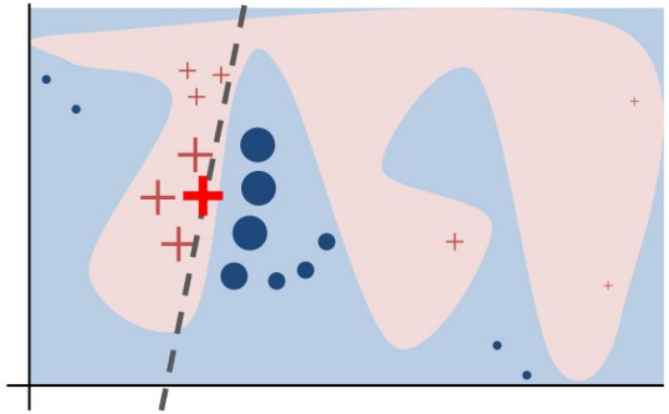
$x_i \xrightarrow{?} y_i$

# L’explicabilité dans l’apprentissage automatique

## Approches locales

Méthodes en bref

LIME [2]

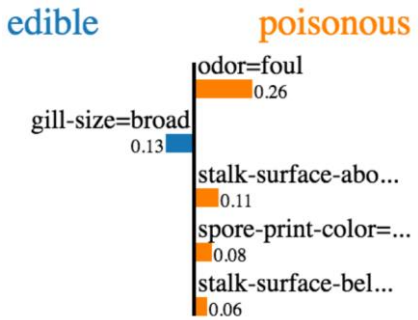


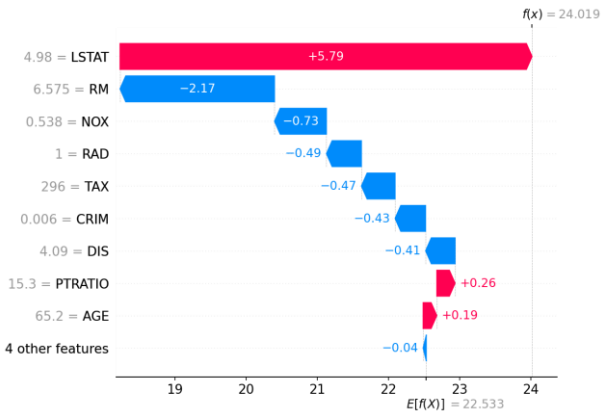
SHAP [3]

Additive Feature Attribution Methods

$$f(x) = g(x') = \phi_0 + \sum_{i=1}^{M} \phi_i x'_i$$

Explications produites




2. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August).
3. Lundberg, S., & Lee, S. I. (2017).

# Motivations pour l'évaluation des explications

## Comment évaluer les explications ?

**Qualitatif et subjectif [2]**

Centré autour de l'humain, tests effectués :

- sélection de modèles (décalage dans les données)
- sélection d'attributs (fuite de données)

**Quantitatif et objectif [4]**

Propriétés des explications individuelles :

- Précision
- Fidélité
- Cohérence
- Représentativité
- Nouveauté
- Certitude
- Stabilité ou **Robustesse**

2. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August).
4. Robnik-Šikonja, M., & Bohanec, M. (2018).

# Travaux sur la robustesse en explicabilité

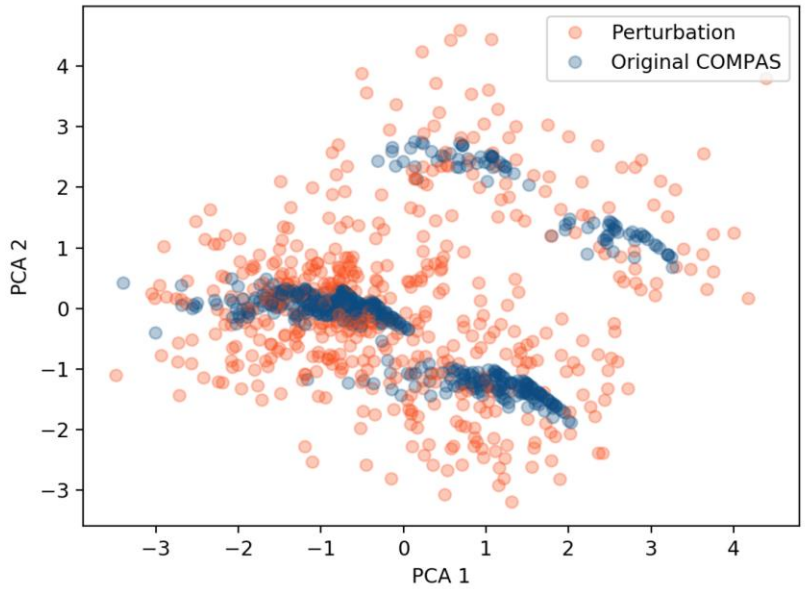
## Approches et définitions

- **La robustesse face à un modèle antagoniste [5]**
- **La robustesse face aux perturbations [6]**
- **Les explications robustes [7]**

5. Slack, D., Hilgard, S., Jia, E., Singh, S., & Lakkaraju, H. (2020, February).
6. Alvarez-Melis, D., & Jaakkola, T. S. (2018).
7. Hancox-Li, L. (2020, January).

# Travaux sur la robustesse en explicabilité

## Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation Methods [5]




Le jeu de données COMPAS projeté en 2 dimensions à l'aide d'une ACP.
Les **données originales** sont en bleu.
Les **perturbations** de LIME et Kernel SHAP sont en rouge.

3 modèles :

- Un modèle biaisé (utilise uniquement un attribut discriminant)
- Un modèle non biaisé (n'utilise pas d'attribut discriminant)
- Un modèle capable de reconnaître la distribution et qui choisit lequel des 2 modèles précédents doit répondre.

Ces approches ne peuvent pas garantir l'équité des modèles étudiés

5. Slack, D., Hilgard, S., Jia, E., Singh, S., & Lakkaraju, H. (2020, February).

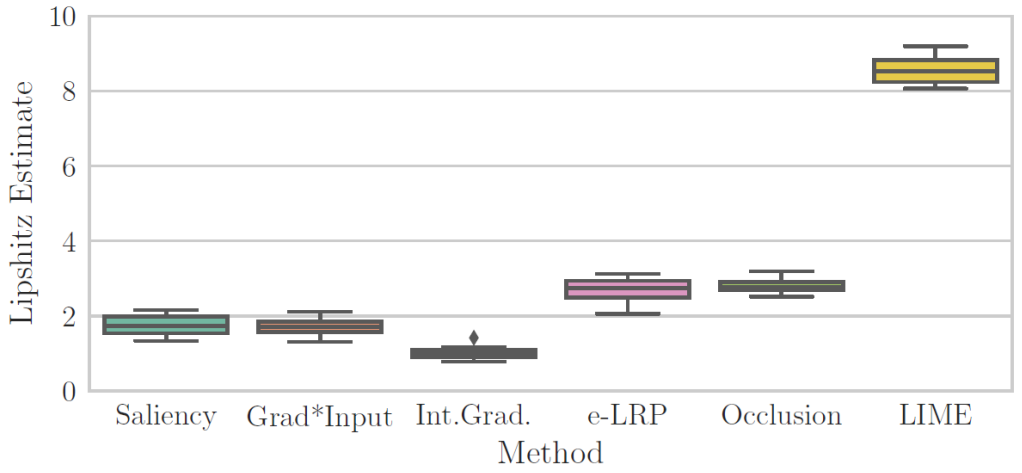
# Travaux sur la robustesse en explicabilité

## On the Robustness of Interpretability Methods [6]

**Mesure de la robustesse** :

$$\hat{L}(x_i) = \underset{x_j \in B_\epsilon(x_i)}{\operatorname{argmax}} \frac{\|f(x_i) - f(x_j)\|_2}{\|x_i - x_j\|_2}$$

Avec $B_\epsilon(x_i)$ une boule de rayon $\epsilon$ et de centre $x_i$.




Robustesse pour différentes méthodes d'explications locales

Modèle instable ?

Dans le cas où l'objectif est de découvrir un phénomène alors il est impératif pour la méthode d'explication d'être robuste.

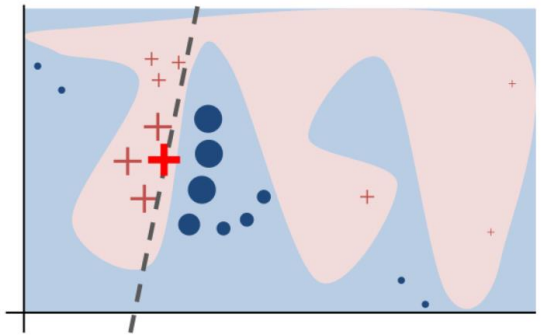
6. Alvarez-Melis, D., & Jaakkola, T. S. (2018).

# Regard critique sur la robustesse

## Robustness in Machine Learning Explanations: Does It Matter? [7]

Le manque de robustesse

Erreurs dans les mesures



Confiance dans le modèle imitateur

Comparaison avec les modèles scientifiques

7. Hancox-Li, L. (2020, January).

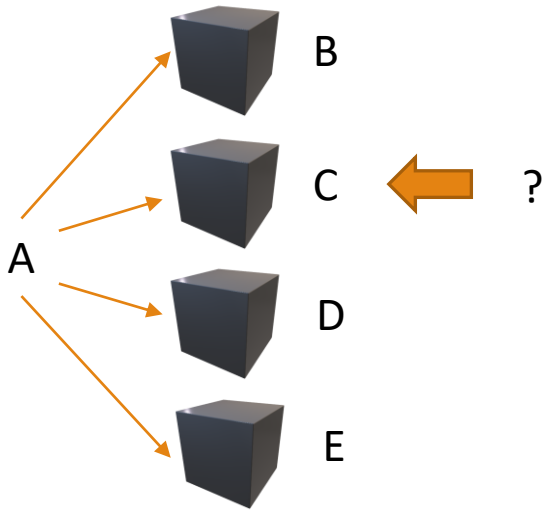
# Regard critique sur la robustesse

## Robustness in Machine Learning Explanations: Does It Matter? [7]

Problématiques éthiques

L'effet Rashomon [8]


Explications non trompeuses




Fairwashing [9]

7. Hancox-Li, L. (2020, January).
8. Breiman, L. (2001).
9. Lakkaraju, H., & Bastani, O. (2020, February).

# En synthèse

Il est important de proposer des méthodes objectives d'évaluation des explications comme **la robustesse**.

La robustesse peut être définie de plusieurs façons : face aux **attaques**, face aux **perturbations**, face à la **multiplicité** des explications.

Les méthodes d'explication locales les plus connues sont sensibles aux perturbations et aux attaques.

La robustesse face aux perturbation est souhaitable lorsqu'on s'intéresse aux **phénomènes réels**.

Avoir plusieurs explications soulève des **problématiques éthiques** à surveiller.

# Merci pour votre attention !

# Références


1. Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.
2. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). " Why should i trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 1135-1144).
3. Lundberg, S., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *arXiv preprint arXiv:1705.07874*.
4. Robnik-Šikonja, M., & Bohanec, M. (2018). Perturbation-based explanations of prediction models. In *Human and machine learning* (pp. 159-175). Springer, Cham.
5. Slack, D., Hilgard, S., Jia, E., Singh, S., & Lakkaraju, H. (2020, February). Fooling lime and shap: Adversarial attacks on post hoc explanation methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (pp. 180-186).
6. Alvarez-Melis, D., & Jaakkola, T. S. (2018). On the robustness of interpretability methods. *arXiv preprint arXiv:1806.08049*.
7. Hancox-Li, L. (2020, January). Robustness in machine learning explanations: does it matter?. In *Proceedings of the 2020 conference on fairness, accountability, and transparency* (pp. 640-647).
8. Breiman, L. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, *16*(3), 199-231.
9. Lakkaraju, H., & Bastani, O. (2020, February). " How do I fool you?" Manipulating User Trust via Misleading Black Box Explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (pp. 79-85).