

Majeure Machine Learning

Biais et
explicabilité

Contenu



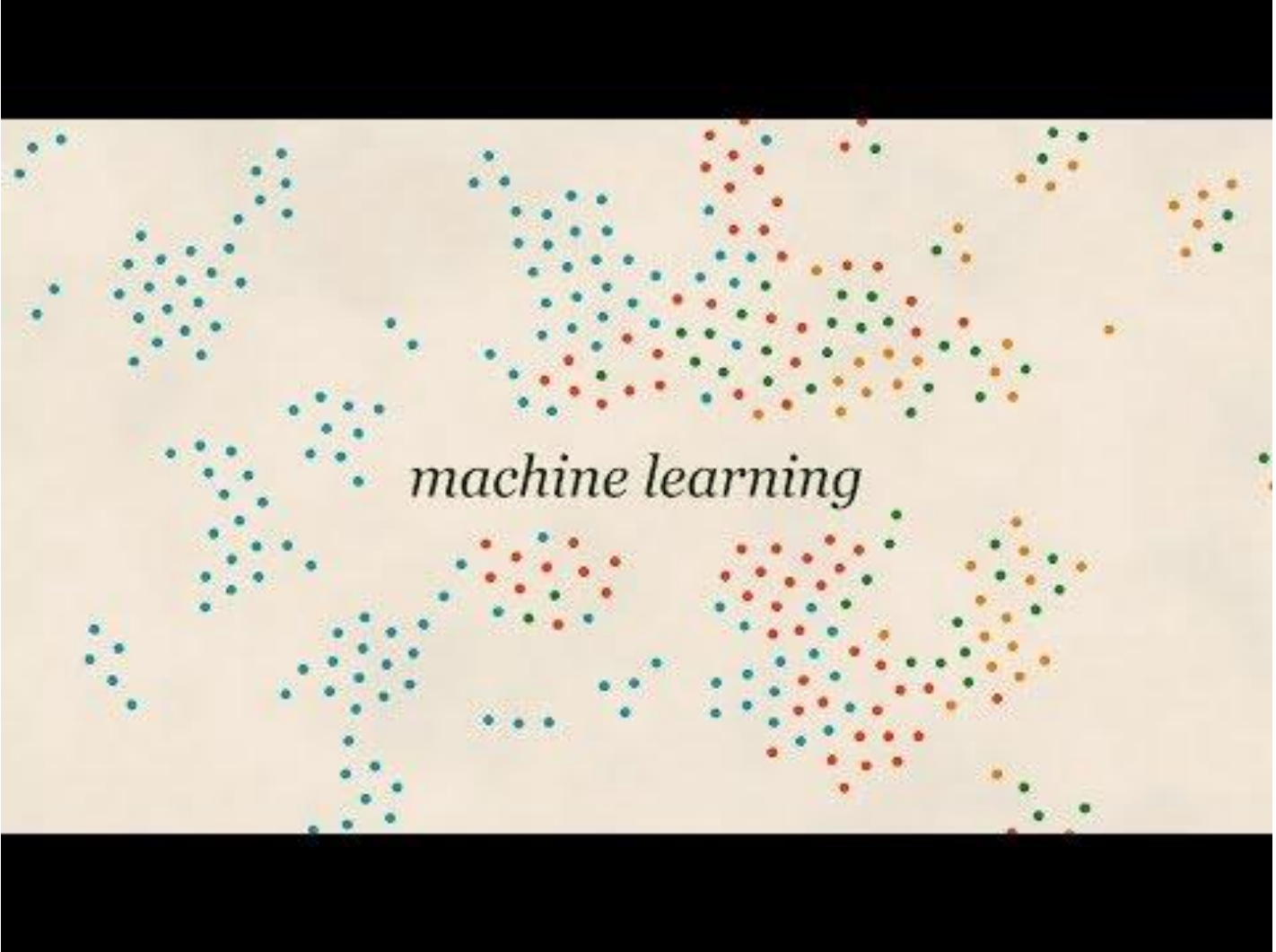
- Présentation des biais
- Définition de l'interprétabilité et de l'explicabilité
- Présentation de LIME et SHAP

Ce que vous devrez savoir faire



- Etre vigilant sur les biais contenus dans vos données
- Etre vigilant sur l'explicabilité demandée par votre projet
- Différencier l'interprétabilité de l'explicabilité
- Connaître les différences entre les modèles “white box” et “black box”

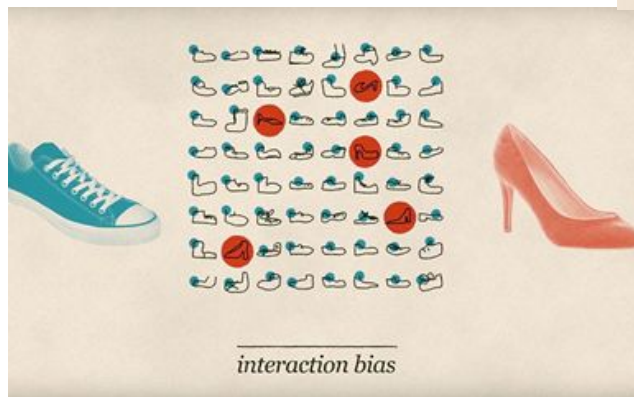
Biais

A scatter plot on a light beige background, framed by a black border at the top and bottom. The plot contains four distinct clusters of small, semi-transparent dots. The clusters are colored blue, red, green, and orange. The blue dots are primarily in the upper-left and lower-left areas. The red dots are concentrated in the upper-middle and lower-middle sections. The green dots are scattered in the upper-right and lower-right regions. The orange dots are mostly in the upper-right and lower-right areas, often overlapping with the green dots. The text "machine learning" is written in a black, serif font, centered horizontally and slightly below the vertical center of the plot.

machine learning

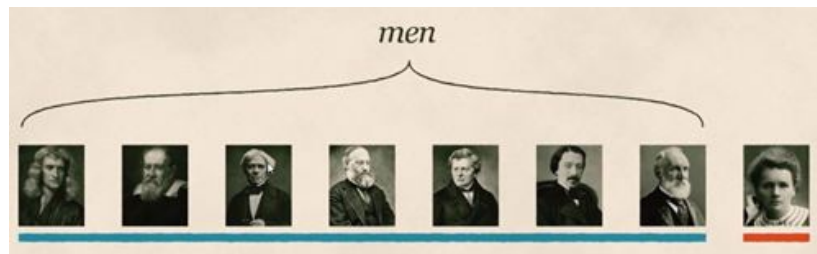
Les 3 biais

Interaction bias



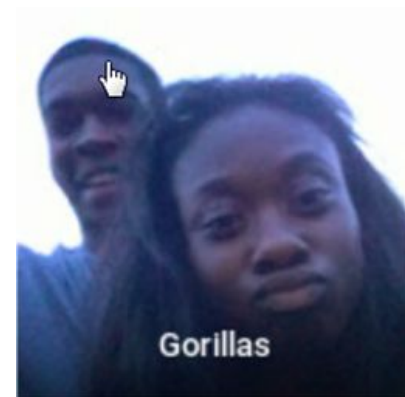
« notre jeu de données composé
des remarques des utilisateurs
est-il représentatif de la réalité »

Latent bias



« notre jeu de données est-il
toujours représentatif de la
réalité »

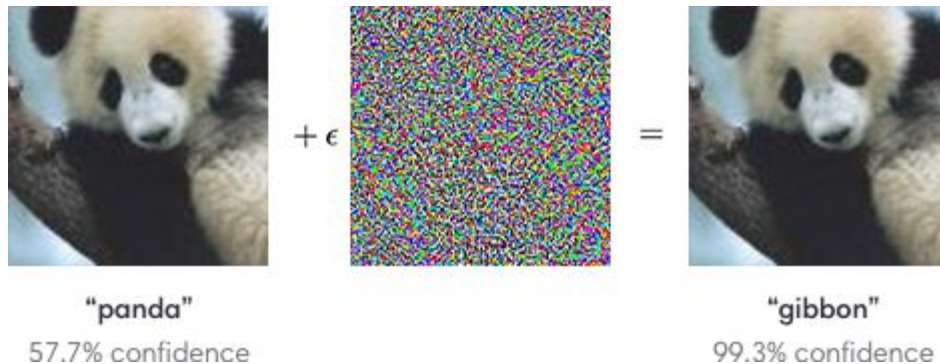
Selection bias



« notre jeu de données
représente-t-il tout le monde
uniformément ? »

Adversarial Examples

Intuition : Les réseaux de neurones peuvent être sensibles au bruit et la frontière de classification peut être très fine



=> Ces failles peuvent être exploitées sous forme d'attaques !

Un algorithme d'optimisation peut être utilisé pour trouver quel bruit produit une classe précise prédite par le réseau (Targeted Attacks)

Exemple dangereux : Voiture autonomes !



Interprétabilité et explicabilité

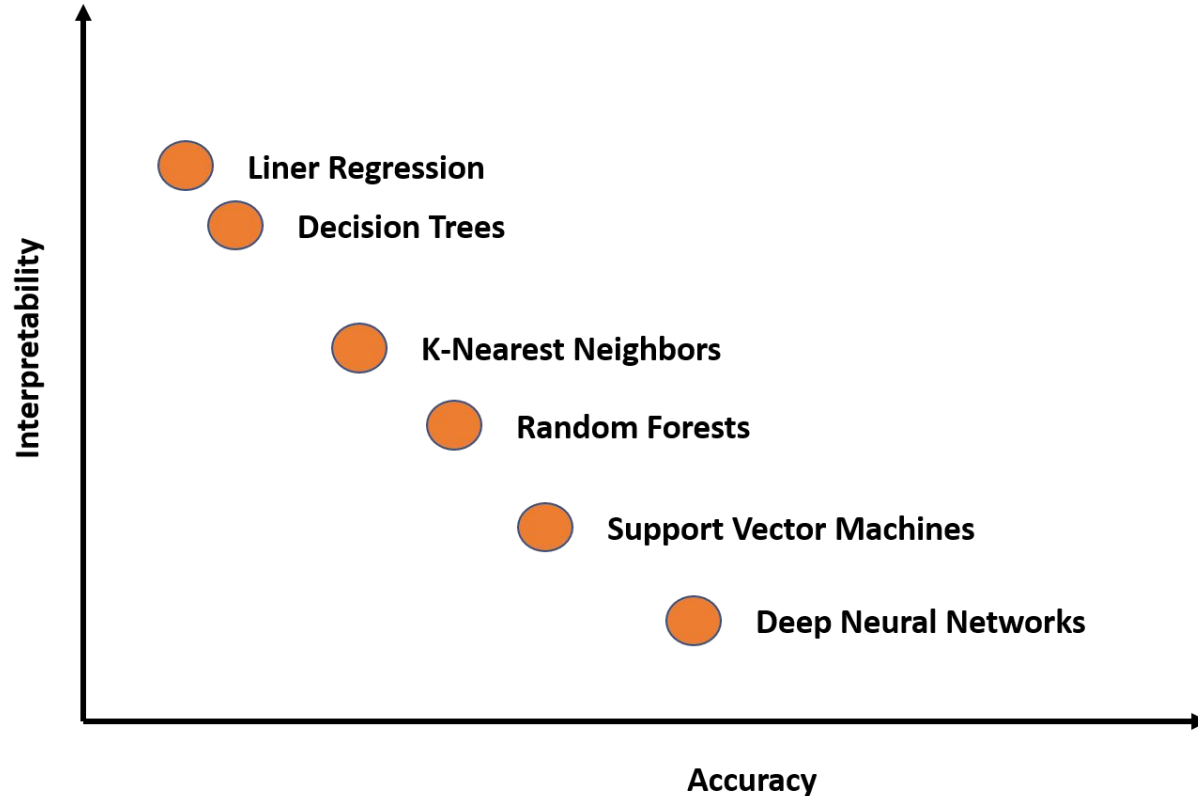
Explicabilité et interprétabilité

Un algo est « interprétable » lorsqu'on comprend précisément son fonctionnement, à l'instar d'un [arbre de décision](#) (pour conseiller un médecin dans l'opération ou non d'une tumeur selon sa taille, l'âge du patient, son poids, etc.).

Notion plus « faible », « l'explicabilité » d'un algo suppose seulement de comprendre quels sont les éléments qui motivent la décision sans forcément comprendre tout le mécanisme de sa construction. Par exemple, d'identifier les pixels qui motivent le résultat pour un réseau de neurones dans un problème de reconnaissance d'images.

[source](#)

Interprétabilité des modèles



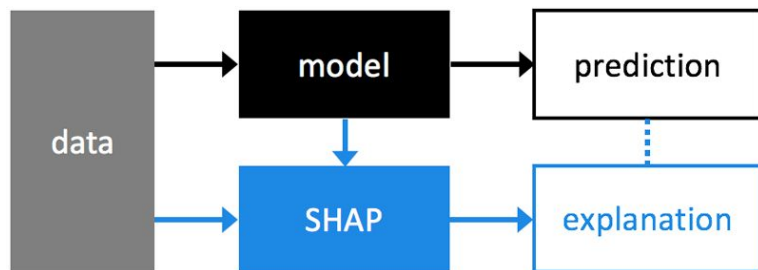
LIME - Pourquoi doit-on croire le modèle ?

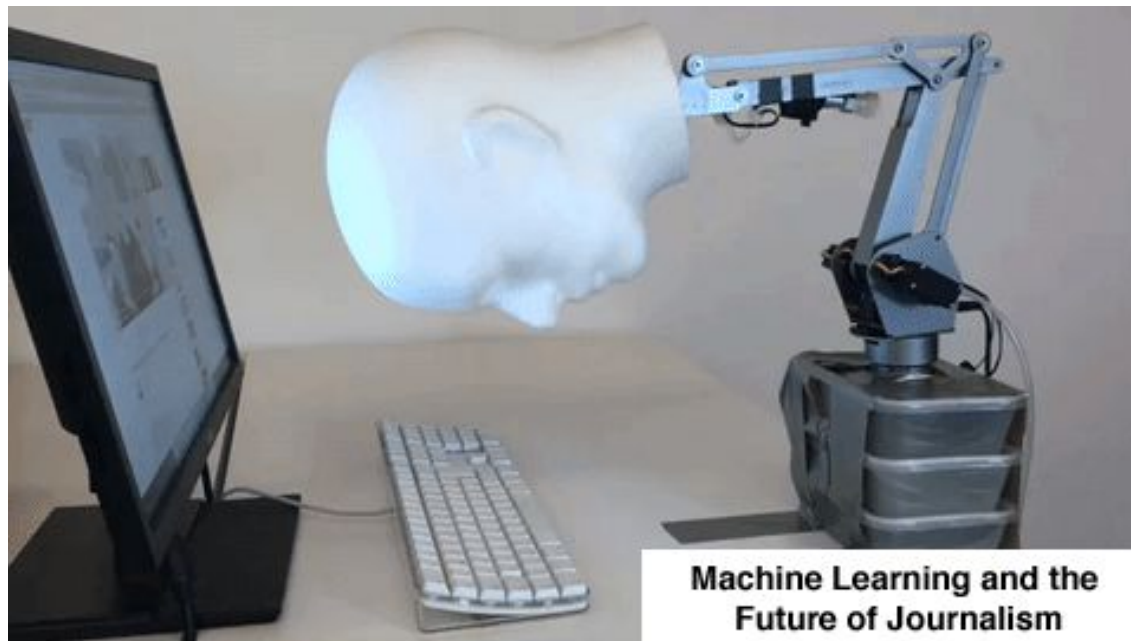


Sometimes you don't know if you can trust a machine learning prediction...



Shap





**Machine Learning and the
Future of Journalism**

Fin du chapitre 6