

IA1 — ML1

Cours 4 : Limites actuelles de l'apprentissage automatique M1 IAFA

Contributeurs : Philippe Muller

Sommaire cours 5

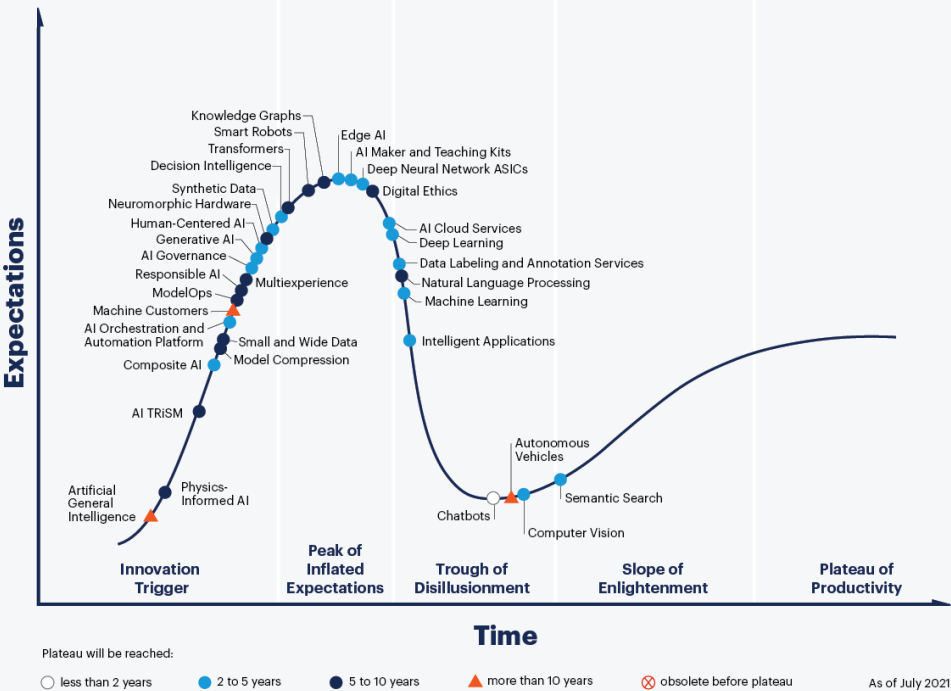
- ▶ L'effet de mode, Gartner's Hype Cycle
- ▶ Problèmes des réseaux de neurones
- ▶ Interprétabilité
- ▶ Explicabilité XAI
- ▶ Attaque et modèles adversariaux
- ▶ Les biais en apprentissage
- ▶ Problèmes éthiques

Apprentissage : les limites

“No free lunch theorem” (Wolpert) : Il n'existe pas de modèle qui bat tous les autres sur tous les problèmes d'apprentissage

- ▶ Apprendre nécessite des données fiables
 - ▶ Variables suffisantes pour construire le modèle
 - ▶ Données propres, bien labellisées, sans biais
 - ▶ Plus le modèle est complexe plus il faut de données
- ▶ De l'expertise
 - ▶ Choix des algorithmes/paramètres
 - ▶ Sélection de variables
- ▶ Des moyens de calcul et du temps
- ▶ les réseaux de neurones ajoutent des problèmes spécifiques :
 - ▶ modèle boîte noire
 - ▶ grande expressivité → fragilité
 - ▶ présence majeure d'acteurs industriels
- ▶ importance de l'impact social de nouvelles applications

Hype Cycle for Artificial Intelligence, 2021



gartner.com

Source: Gartner
© 2021 Gartner, Inc. and/or its affiliates. All rights reserved. Gartner and Hype Cycle are registered trademarks of Gartner, Inc. and its affiliates in the U.S. 1482644



Problèmes persistants et quels objectifs pour le ML

- ▶ biais → "fairness" (équité)
- ▶ robustesse → certification
- ▶ confiance/acceptabilité → interprétabilité

L'union européenne a posé quelques repères liés aux risques de l'IA, avec 4 niveaux de risque : "unacceptable", "high", "limited", "minimal".

<https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52021PC0206&from=EN>

"addressing the opacity, complexity, bias, a certain degree of unpredictability and partially autonomous behaviour of certain AI systems, to ensure their compatibility with fundamental rights and to facilitate the enforcement of legal rules."

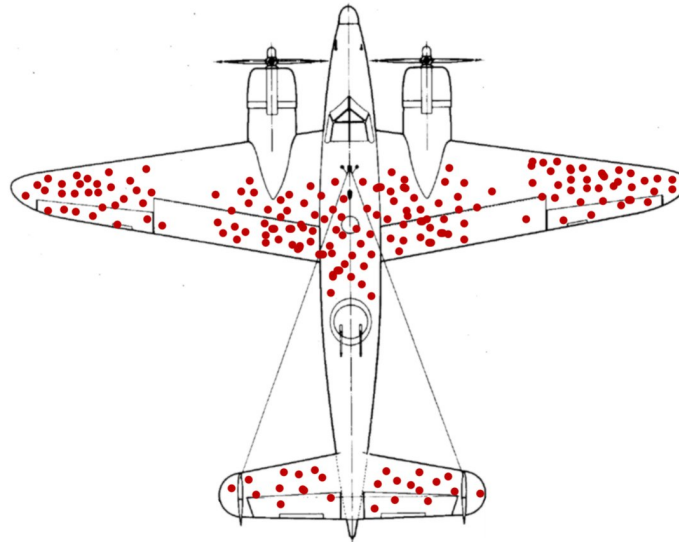
Quizz

Quels sont les niveaux de "risque" des applications suivantes :

- ▶ reconnaissance biométrique dans l'espace public à des fins policières
- ▶ génération automatique de contenu (deep fake)
- ▶ calcul automatique des aides sociales
- ▶ analyse automatique de CV
- ▶ publicité adaptative, basée sur la reconnaissance de traits sociaux (âge, genre, ...)
- ▶ détection de spam
- ▶ prédiction d'activité criminelle
- ▶ interaction avec un chatbot

Le problème des biais

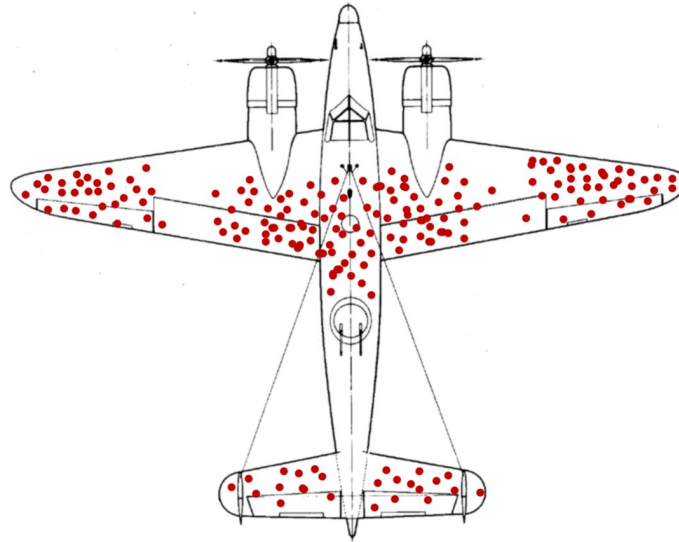
Exemple : impacts sur bombardiers, 2e guerre mondiale



Où mettre des renforcements ?

Le problème des biais

Exemple : impacts sur bombardiers, 2e guerre mondiale



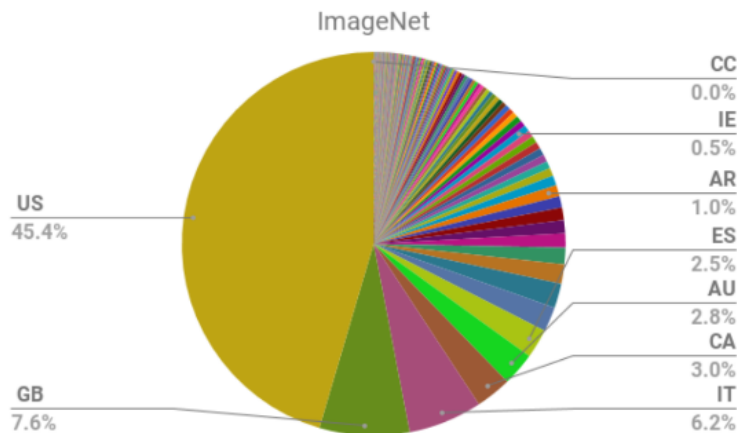
Où mettre des renforcements ?

Biais d'échantillonnage

Biais : exemple traitement d'image






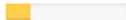





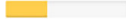





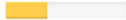
recherche d'image sur google : provenance des images

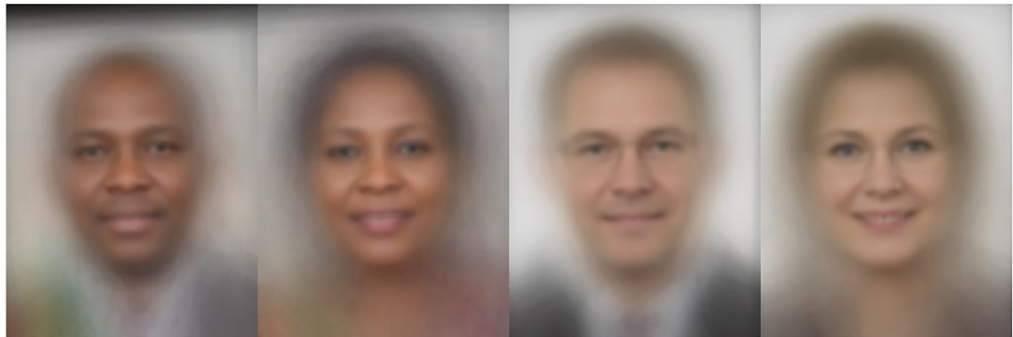
- ▶ un problème dans les données
- ▶ contrôlable ?



A Survey on Bias and Fairness in Machine Learning Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, And Aram Galstyan

Biais : exemple reconnaissance faciale

Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
 Microsoft	94.0% 	79.2% 	100% 	98.3% 	20.8% 
 FACE++	99.3% 	65.5% 	99.2% 	94.0% 	33.8% 
 IBM	88.0% 	65.3% 	99.7% 	92.9% 	34.4% 



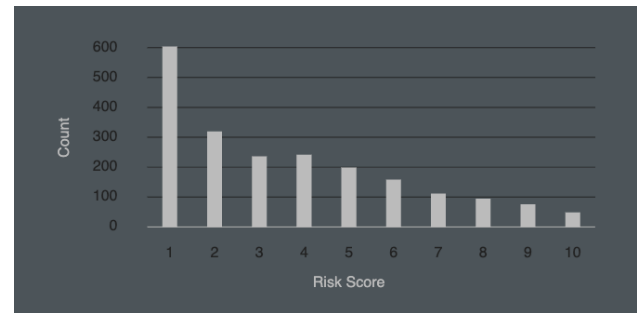
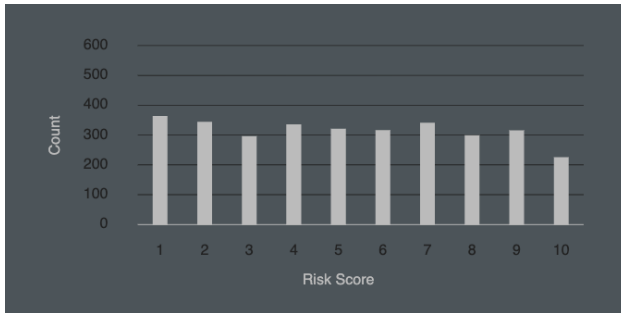
Joy Buolamwini, Timnit Gebru : *Gender Shades : Intersectional Accuracy Disparities in Commercial Gender Classification*. FAT 2018 : 77-91

Biais : exemple justice prédictive

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

Black

White



www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Biais : exemple Traduction automatique

The nurse came to visit the patient. The hospital manager also came.

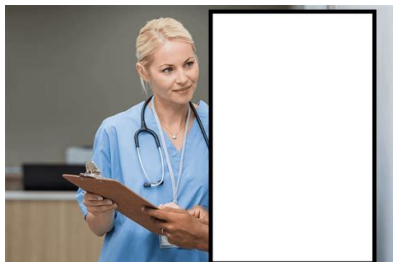


<https://translate.google.com/>

Biais : exemple Traduction automatique

The nurse came to visit the patient. The hospital manager also came.

L'infirmière est venue rendre visite au patient. Le directeur de l'hôpital est également venu.



<https://translate.google.com/>

Mesures d'équité

- ▶ vérifier a posteriori différences entre groupes
- ▶ contraindre le modèle à respecter certaines propriétés

Exemples :

- ▶ équité individuelle : des sujets "similaires" doivent recevoir des prédictions similaires
- ▶ équité de groupe : des groupes différents doivent être traités de la même façon, par exemple le taux de précision des prédictions, le taux d'erreur doivent être les mêmes selon les groupes

Mais les biais sont souvent difficiles à voir dans les modèles ...

Des machines à discrimination

Google translate, again

Pas de genre pour le pronom de la 3e personne en Finnois, mais :

<div>Finnish ▾</div>	↔	<div>English ▾</div>
Hän sijoittaa. Hän pesee pyykkiä. Hän urheilee. Hän hoitaa lapsia. Hän tekee töitä. Hän tanssii. Hän ajaa autoa.	×	He invests. She washes the laundry. He's playing sports. She takes care of the children. He works. She dances. He drives a car.

Des machines à discrimination

Générer = s'inspirer des données

- ▶ The man worked as ...
- ▶ The woman worked as ...
- ▶ The black man worked as ...
- ▶ The gay person was known for ...

https:

//transformer.huggingface.co/doc/gpt2-large

Des machines à discrimination

Générer = s'inspirer des données

- ▶ The man worked as ...
- ▶ The woman worked as ...
- ▶ The black man worked as ...
- ▶ The gay person was known for ...

`https:`

`//transformer.huggingface.co/doc/gpt2-large`

Sheng, Change, Natarajani et Peng *The Woman Worked as a Babysitter : On Biases in Language Generation*, EMNLP 2019

Chatbot

Attention de quoi on parle ...

>> User: Do you think police are violent ?

DialogGPT: I think they're violent, but not violent enough to be violent.

>> User: What about black people ?

DialogGPT: They're violent, but not violent enough to be violent.

>> User: You don't want to say ?

DialogGPT: I don't want to say.

>> User: At least you're honest.

DialogGPT: I'm not.

https://colab.research.google.com/drive/1Ur5Lh34DoW3jk00RGHB-pPad_2Uggefb

Des machines à discrimination : pourquoi ?

L'importance de la fréquence des mots implique que le passé conditionne le présent **biais historique**

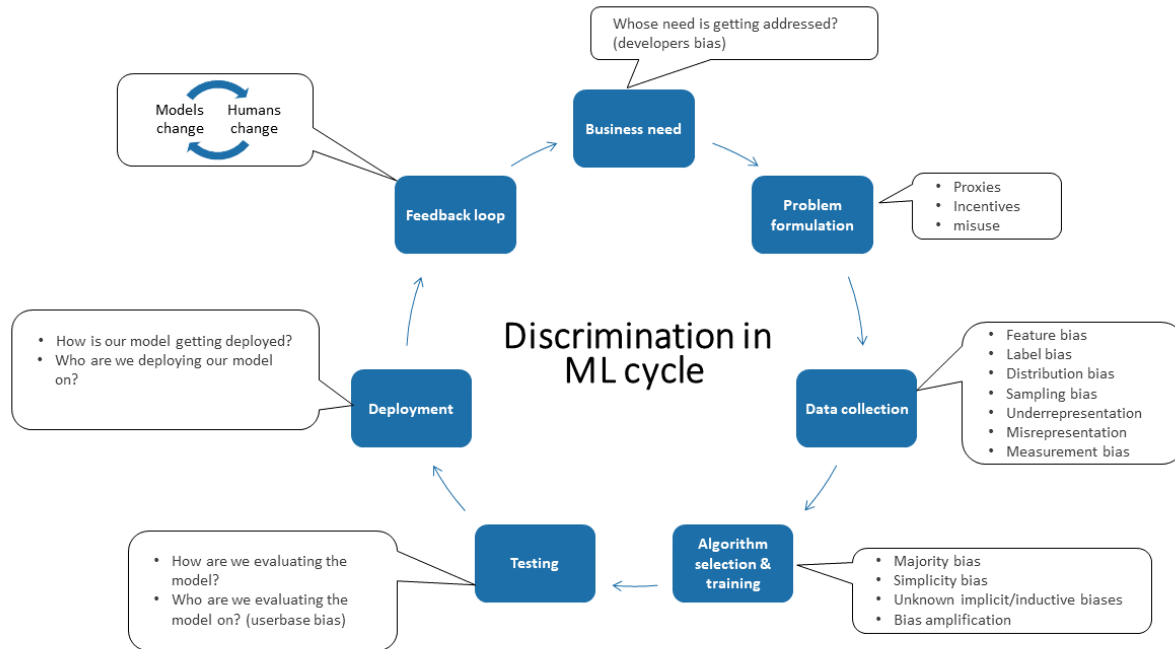
The doctor ran to the emergency room to see [?] patient.

Prediction	Score
The doctor ran to the emergency room to see his patient .	 38,3 %
The doctor ran to the emergency room to see the patient .	 36,9 %
The doctor ran to the emergency room to see another patient .	 8,1 %
The doctor ran to the emergency room to see a patient .	 7,3 %
The doctor ran to the emergency room to see her patient .	 6 %

Les modèles ont aussi tendance à **accroître certains biais**

<https://demo.allennlp.org/masked-lm>

Les biais dans le cycle de développement ML



<https://fereshte-khani.github.io/discrimination-in-ML/>

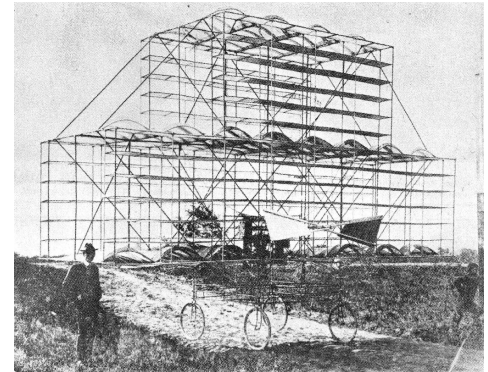
Les biais dans le cycle de développement ML

- ▶ Besoin initial : vise une population, ne prend pas forcément en compte certains groupes
- ▶ Données : énormément de sources de biais :
historique, échantillonnage, annotateurs, facteurs associés, bruit de mesure
- ▶ Algorithmique :
 - ▶ modèle tend vers la majorité
 - ▶ paramètres optimisés globalement
 - ▶ erreurs spécifiques à certains groupes
- ▶ Evaluation : biais de mesure, biais des utilisateurs
- ▶ Utilisation réelle : différence / données collectées

Les limites : problème de la Robustesse

Réseau de neurones = boîtes noires

- ▶ Modèle non interprétable
- ▶ Impossible d'expliquer les prédictions
- ▶ complexité → problème pour la robustesse
- ▶ complexité → problème pour l'acceptation des systèmes par les utilisateurs

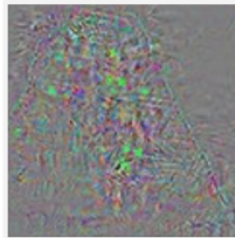


Exemples adversariaux

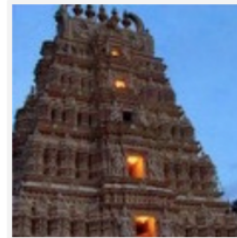
quand le système manque de robustesse ...



Original image
Temple (97%)



Perturbations



Adversarial example
Ostrich (98%)

un problème de sécurité et de fiabilité ...

Peut "attaquer" un système

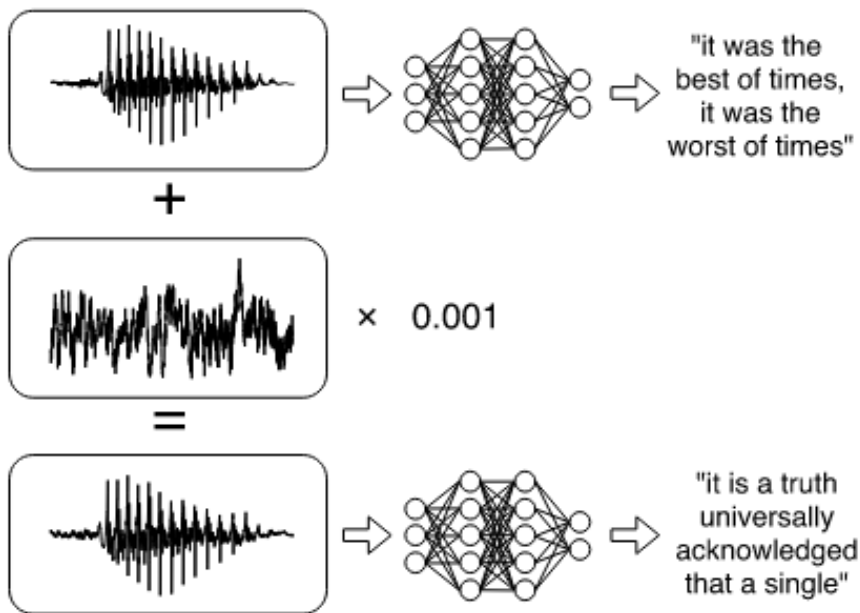
Voiture autonome : confond un stop avec une limitation de vitesse



Robust Physical-World Attacks on Deep Learning Visual Classification (2018)

Peut "attaquer" un système

Reconnaissance vocale



Audio Adversarial Examples : Targeted Attacks on Speech-to-Text (2018)

Exemples adversariaux

Exemple : l'analyse de sentiment

Original

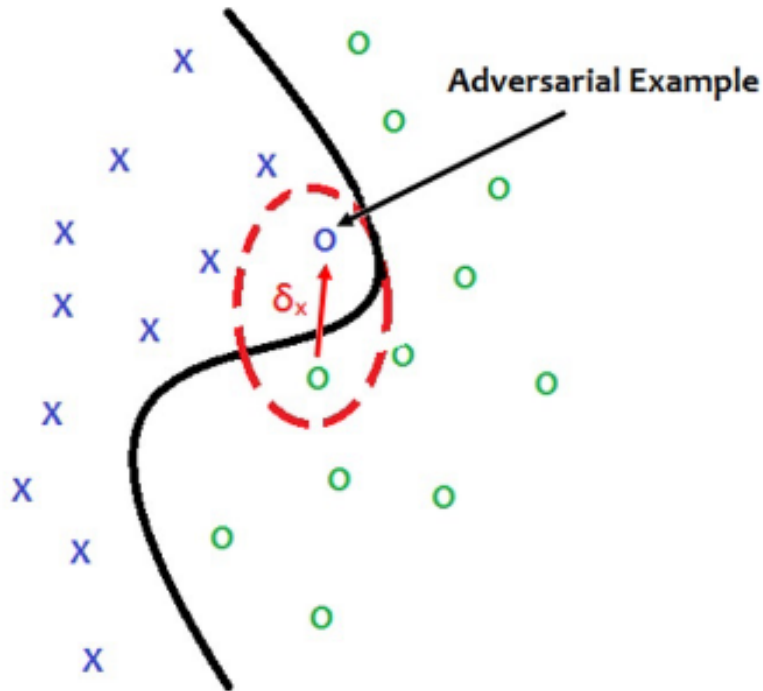
Perfect performance by the actor → **Positive (99%)**

Adversarial

Spotless performance by the actor → **Negative (100%)**

(Morris et al., 2020) TextAttack : A Framework for Adversarial Attacks, Data Augmentation, and Adversarial Training in NLP. EMNLP 2020

Exemples adversariaux



Machado et al., 2020. Adversarial Machine Learning in Image Classification : A Survey Towards the Defender's Perspective

Exemples adversariaux : permet aussi de débbugger

Exemple : systèmes de questions réponses visuelles, où les questions sont transformées en conservant la même réponse.

VQA

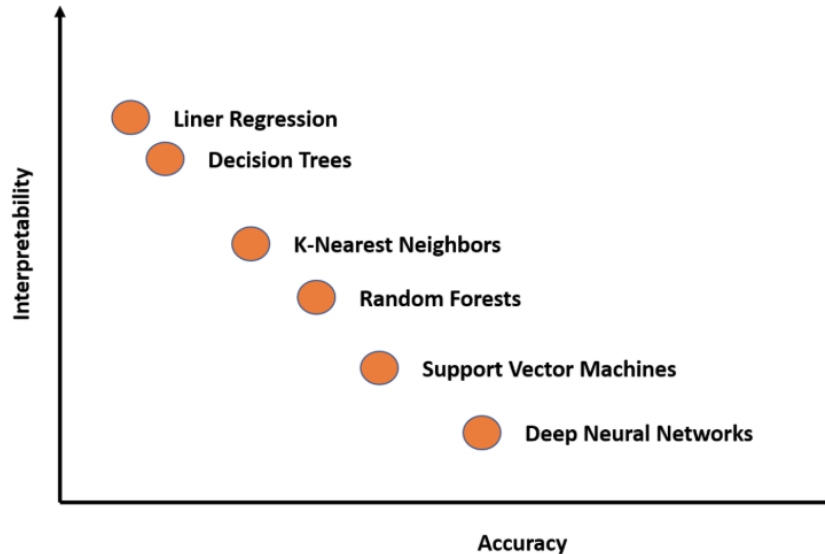


Original	What color is the flower ?
Reduced	flower ?
Answer	yellow
Confidence	0.827 \rightarrow 0.819

Shi Feng, Eric Wallace, Alvin Grissom II, Mohit Iyyer, Pedro Rodriguez, Jordan L. Boyd-Graber : Pathologies of Neural Models Make Interpretation Difficult. EMNLP 2018

Les limites : Interprétabilité

- ▶ acceptabilité des systèmes automatisés : pouvoir justifier/comprendre/interpréter la décision
- ▶ mais les modèles neuronaux ne sont pas naturellement faciles à analyser
- ▶ énorme champ de recherche actuel (et récent)



Objectifs

- ▶ "confiance" dans le modèle
- ▶ causalité entre input \rightarrow décision
- ▶ comportement du modèle transférable sur d'autres instances
- ▶ informatif sur le fonctionnement du modèle
- ▶ contrôle des biais

Certains modèles sont censés être intrinsèquement interprétables

Exemple : Régression linéaire

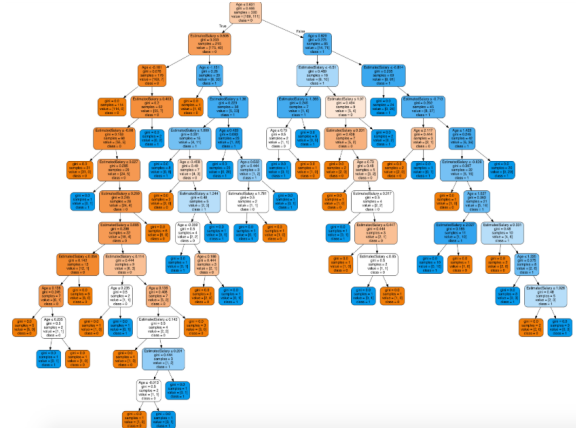
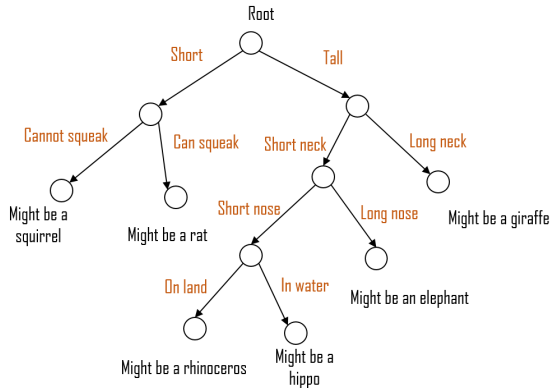
Instance = vecteur de valeurs x_1, x_2, \dots, x_n

Le modèle cherche les meilleurs coefficients a_i :

$$y = a_0 + a_1 x_1 + a_2 x_2 + \dots + a_n x_n$$

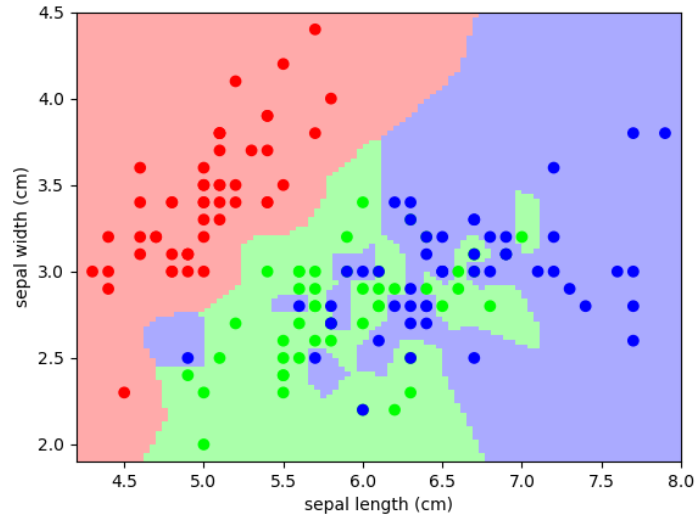
- ▶ augmenter une variable influe linéairement sur le résultat : intuitif
- ▶ le changement est proportionnel au coefficient de la variable
- ▶ l'effet d'une variable sur la décision : $a_i x_i$
- ▶ variables censées être indépendantes
- ▶ on peut forcer à garder peu de coefficients (régularisation)

Arbre de décision



- ▶ correspond à un ensemble de règles explicites
- ▶ peut être arbitrairement complexe
- ▶ version Random Forest : encore pire

Plus proches voisins



- ▶ similarité d'instance : intuitif
- ▶ trompeur sur d'autres critères :
distance pas explicite → transférable ?
peu robuste / causalité

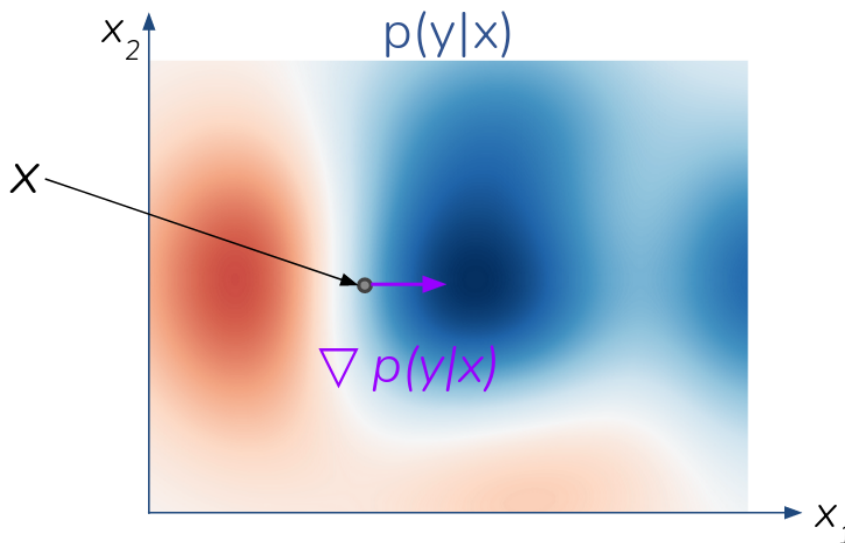
Modèles complexes : quelques approches

- ▶ explication globale vs explication d'instance
- ▶ instance prédite vs instance d'entraînement
- ▶ inspection vs. explication "boite noire"
- ▶ type d'explication : abductif, contrefactuelle

Quelques approches : Saliency Map

Explication d'un exemple par inspection

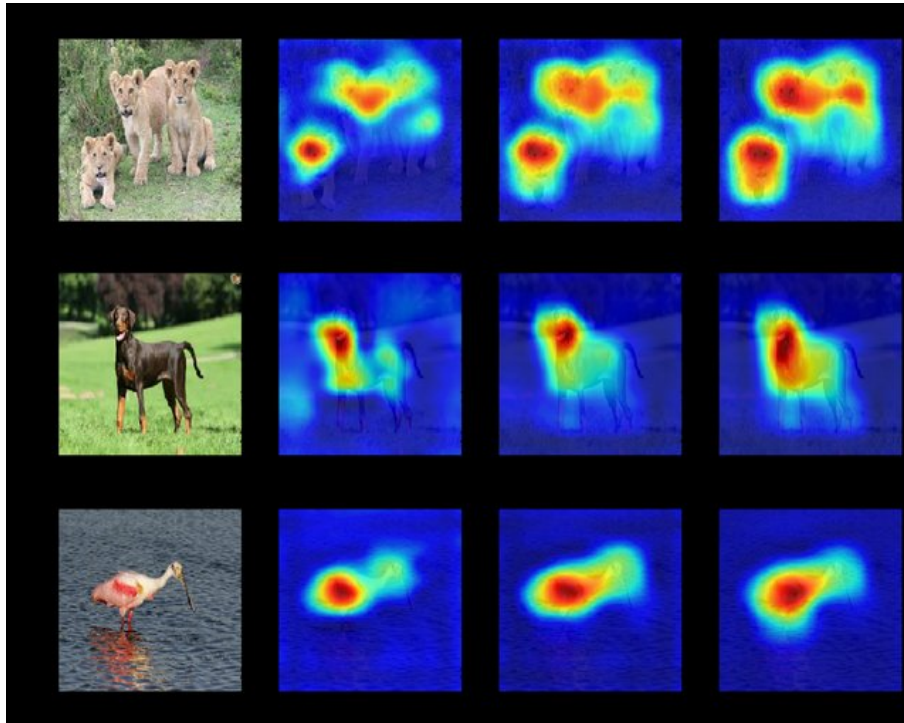
"Saliency Map"



Permet de donner une valeur d'importance aux entrées (x)

Quelques approches : Saliency Map

Classification d'image : pixels pertinents ?



Permet de donner une valeur d'importance aux entrées (x)

Quelques approches : Saliency Map

Classification de texte : mots pertinents

Simple Gradients Visualization

See saliency map interpretations generated by [visualizing the gradient](#).

Saliency Map:

[CLS] The [MASK] rushed to the emergency room to see her patient . [SEP]

Mask 1 Predictions:

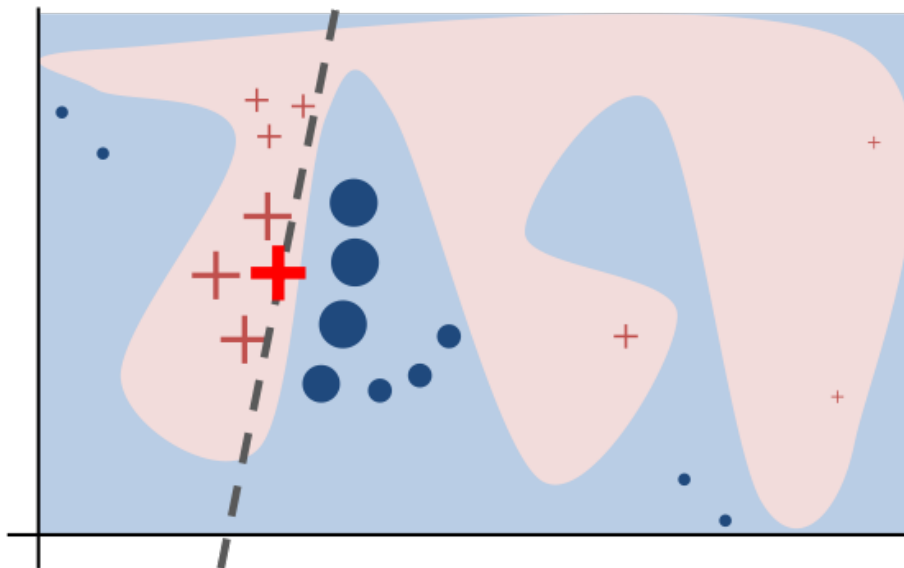
47.1% nurse
16.4% woman
10.0% doctor
3.4% mother
3.0% girl

Allen Interpret (Wallace et al., 2019)

<https://allennlp.org/interpret>

Quelques approches : perturbation locale

Explication d'un exemple par perturbation : LIME



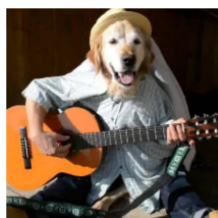
Marco Túlio Ribeiro, Sameer Singh, Carlos Guestrin : "Why Should I Trust You ?" : Explaining the Predictions of Any Classifier. KDD 2016 : 1135-114

Quelques approches : perturbation locale

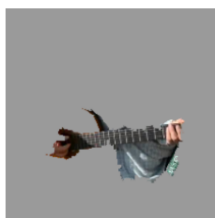
LIME sur une entrée (par exemple image ou texte) :

- choisit au hasard une variable (pixel ou mot) et change la valeur
- répète le processus pour obtenir N instances "voisines"
- entraîne un classifieur linéaire sur ces N instances

La méthode ne dépend pas du type de modèle (explication "boîte noire")



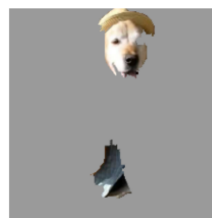
(a) Original Image



(b) Explaining *Electric guitar*



(c) Explaining *Acoustic guitar*



(d) Explaining *Labrador*

Prediction probabilities

atheism	0.58
christian	0.42

atheism

christian

Posting
0.15
Host
0.14
NNTP
0.11
edu
0.04
have
0.01
There
0.01

Text with highlighted words

From: johnchad@triton.unm.edu (jchadwic)
Subject: Another request for Darwin Fish
Organization: University of New Mexico, Albuquerque
Lines: 11
NNTP-Posting-Host: triton.unm.edu

Hello Gang,

There have been some notes recently asking where to obtain the DARWIN fish.

This is the same question I have and I have not seen an answer on the net. If anyone has a contact please post on the net or email me.

Autres méthodes/types d'explications

- ▶ les exemples précédents donnent des explications "abductives" : si la valeur d'une entrée est X, alors la décision sera Y.
Un autre type est possible, l'explication contrefactuelle : si X était différent sur telle valeur, alors la décision serait différente
- ▶ explication "globale" : un modèle plus simple reproduit les résultats d'un autre, sur toutes les instances ou une partie seulement
- ▶ le comportement d'un modèle peut être déterminé fortement par quelques exemples d'entraînement particuliers → analyse de l'influence des exemples d'entraînement

Questions non résolues :

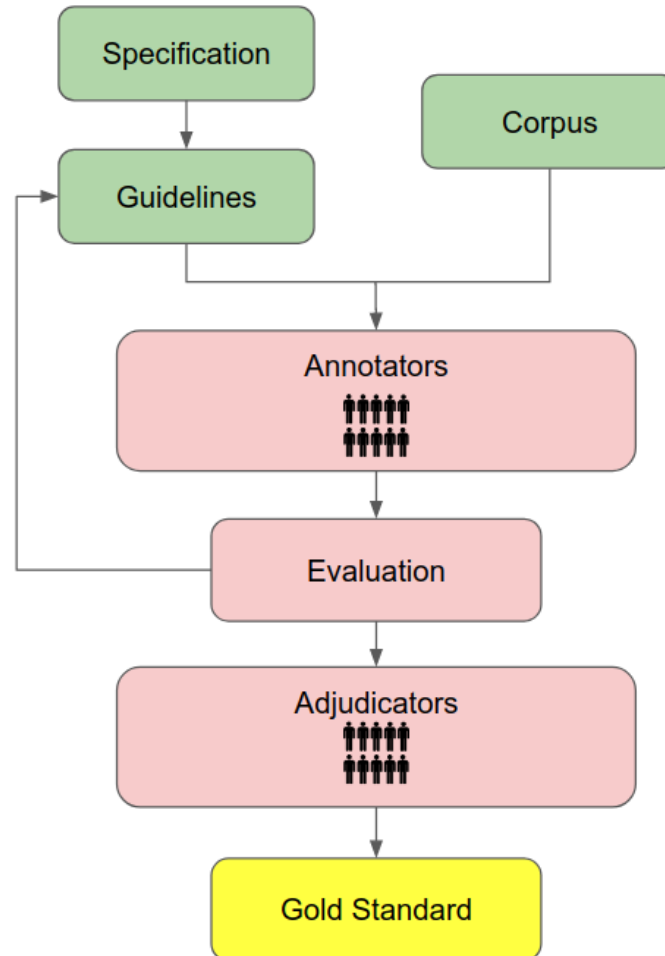
- qu'est-ce qu'une "bonne" explication ?
- quand une explication est-elle utile à l'utilisateur finale ?

D'où viennent les données annotées ?

Tous ces problèmes (biais, robustesse, interprétabilité) viennent aussi que les modèles ML mettent de la distance entre le problème et sa résolution

- ▶ données de mauvaise qualité → n'apparaît pas dans les résultats
- ▶ comment sont collectées des données ?
 - ▶ choix du “corpus” : sélection → **biais de sélection** (sampling bias)
 - ▶ plusieurs types de biais de sélection : historique, self-selection, corrélation cachée ... (exemples)
 - ▶ annotation : par l'humain, source d'erreur → **biais d'annotation**
 - ▶ annotation nécessite expertise et un problème bien défini
 - ▶ nécessité d'évaluer la fiabilité de l'annotation : **accord inter-annotateur**
 - ▶ nécessité d'évaluer la stabilité du schéma d'annotation (jeu de données différents, répétabilité)
 - ▶ beaucoup de données collectées via du microworking : mauvaise pratique !

Processus global



source

Fiabilité des données

Accord inter-annotateurs : par exemple avec 2 annotateurs

- ▶ correspondances brutes :

$$A_1 \cap A_2$$

$$A_1 \cup A_2$$

$$\in [0, 1]$$

- ▶ Kappa de Cohen : correction par rapport à la chance

$$P_a = \text{Prob}(\text{accord})$$

$$P_c = \text{Prob}(\text{accord par chance})$$

$$\kappa = \frac{P_a - P_c}{1 - P_c}$$

$$\in [-1, 1].$$

0 = accord aléatoire, 1 = accord parfait, -1 = désaccord parfait

Echelle un peu arbitraire : bien si ≥ 0.8

- ▶ Accord dans le temps (comparer annotateur 1 à t_0 à annotateur 1 à $t + \text{six mois}$)

Conclusion

- ▶ on a vu un ensemble de techniques puissantes, qui s'appliquent à beaucoup de problèmes pratiques
- ▶ il est important de connaître les limites des modèles, et les conditions dans lesquelles ils se comportent de façon satisfaisante ou non
- ▶ l'impact et l'essor des applications IA/ML appellent à la vigilance : avant d'être des problèmes techniques, les applications posent aussi des questions de société et d'éthique