

EXPLIQUABILITÉ ET INTERPRÉTABILITÉ





01

QU'EST CE QUE C'EST ?

Les modèles utilisés en ML et DL peuvent être **complexes et dur à comprendre**.

Pourtant, il y a plein de **bénéfices** à comprendre les décisions des modèles:

- **Pour le datascientiste**
 - Pour **évaluer** que les modèles fonctionnent correctement
 - Qu'ils n'ont pas appris par cœur, ou appris à détecter la présence de neige pour faire la différence entre chien et loups
 - Pour déboguer
 - Pour identifier les limites des modèles et les améliorer
- **Pour l'utilisateur**
 - Pour **accepter** d'utiliser un modèle
 - Effet « boîte noire »
 - Pour savoir quelle **confiance** donner à une décision du modèle

Deux termes sont souvent utilisés dans le domaine : **interprétabilité et explicabilité**

Il n'y a **pas de définition précise, commune** à la communauté IA de ces deux termes.

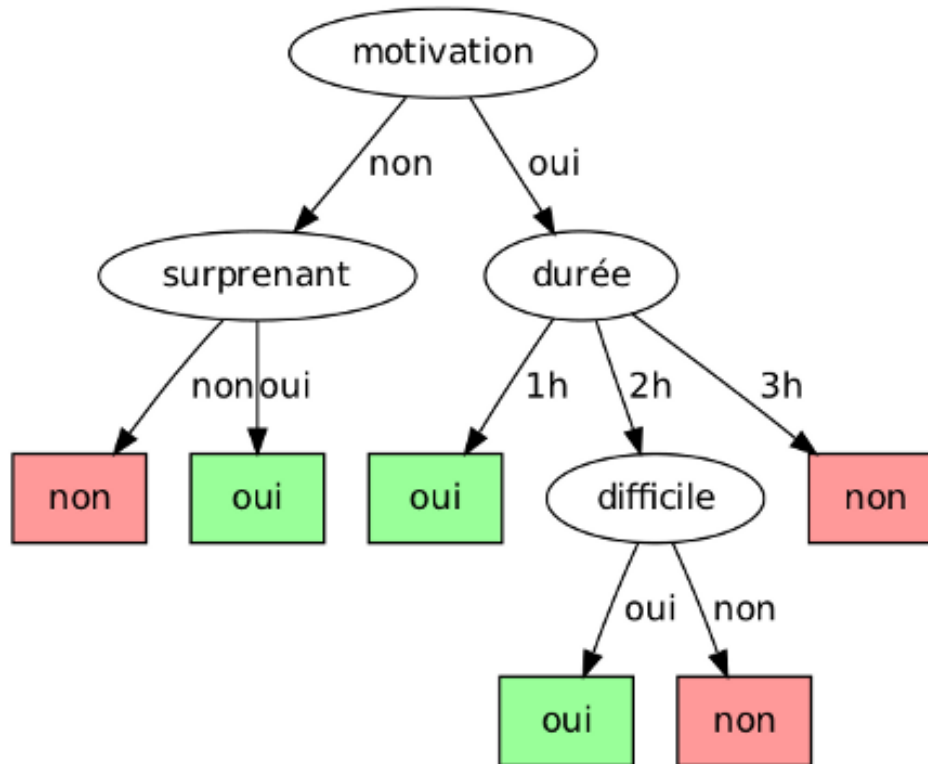
Définition de Rudin, 2019 (C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, 2019) :

Interprétabilité : Caractéristique **inhérente** aux modèles, qui permettent d'eux même d'interpréter leur décision.

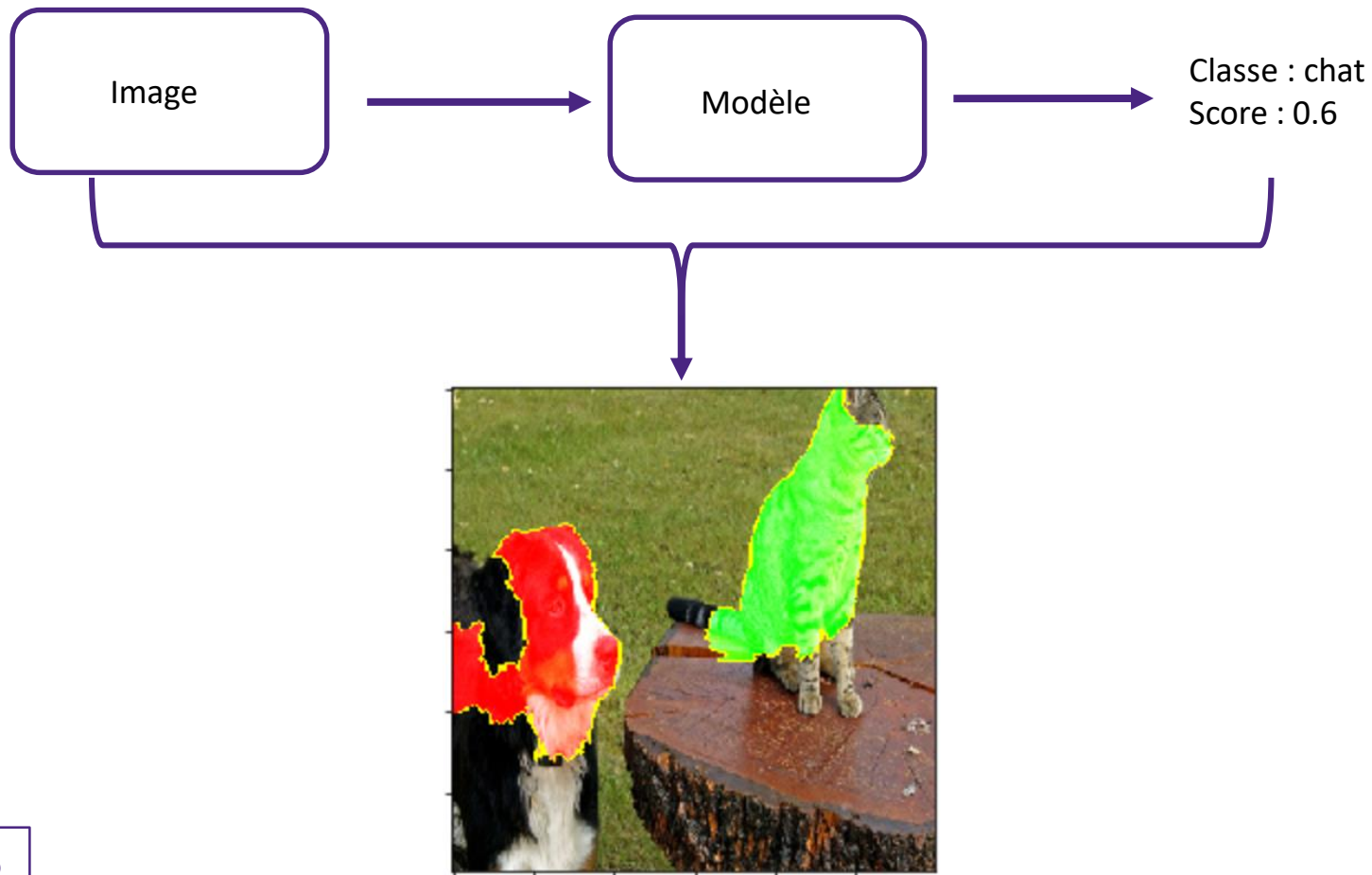
- *Propriété « passive » des modèles, « a priori »*
- *Un arbre de décision est **interprétable***
- *Un modèle est **interprétable** si en le regardant on peut comprendre **comment il fonctionne**, comment il a pris sa décision.*
- *Peut demander **de connaître le fonctionnement du modèle, d'avoir des connaissances en IA.***

Explicabilité : Explication à posteriori de la décision d'un modèle qui peut être une boîte noire.

- *Propriété « active » : un traitement particulier est effectué pour expliquer la décision*
- *« a posteriori »*
- *Quelles informations je peux obtenir de ce modèle, de cette décision ?*
- *Ne demande pas forcément de connaissance du modèle voir de l'IA.*



Un arbre de décision



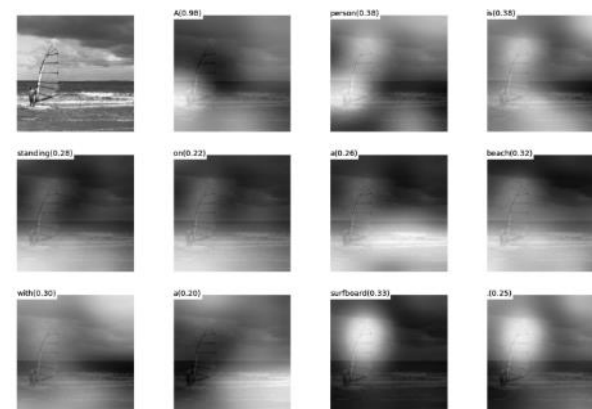
EXPLICABILITÉ OU INTERPRÉTABILITÉ ?

by ent423 ,ent261 correspondent updated 9:49 pm et ,thu march 19,2015 (ent261) a ent114 was killed in a parachute accident in ent45 ,ent85 ,near ent312 ,a ent119 official told ent261 on wednesday .he was identified thursday as special warfare operator 3rd class ent23 ,29 ,of ent187 , ent265 .`` ent23 distinguished himself consistently throughout his career .he was the epitome of the quiet professional in all facets of his life ,and he leaves an inspiring legacy of natural tenacity and focused . . .

by ent270 ,ent223 updated 9:35 am et ,mon march 2 ,2015 (ent223) ent63 went famillial for fall at its fashion show in ent231 on sunday ,dedicating its collection to `` mamma " with nary a pair of `` mom jeans " in sight .ent164 and ent21 , who are behind the ent196 brand ,sent models down the runway in decidedly feminine dresses and skirts adorned with roses ,lace and even embroidered doodles by the designers 'own nieces and nephews .many of the looks featured saccharine needlework phrases like `` i love you , . . .

ent119 identifies deceased sailor as X ,who leaves behind a wife

X dedicated their fall fashion show to moms



(b) A person is standing on a beach with a surfboard.

EXPLICABILITÉ OU INTERPRÉTABILITÉ ?

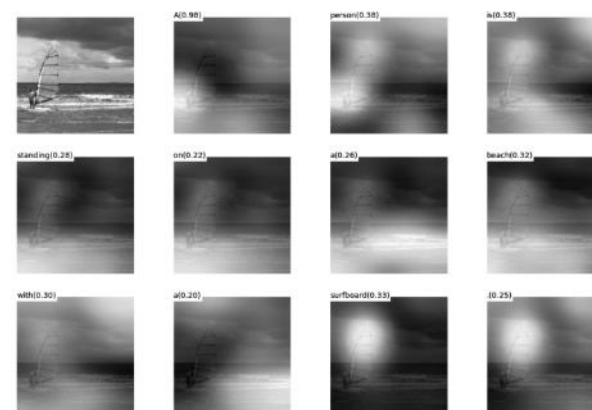
by ent423 ,ent261 correspondent updated 9:49 pm et ,thu march 19,2015 (ent261) a ent114 was killed in a parachute accident in ent45 ,ent85 ,near ent312 ,a ent119 official told ent261 on wednesday .he was identified thursday as special warfare operator 3rd class ent23 ,29 ,of ent187 , ent265 .`` ent23 distinguished himself consistently throughout his career .he was the epitome of the quiet professional in all facets of his life ,and he leaves an inspiring legacy of natural tenacity and focused . . .

ent119 identifies deceased sailor as X ,who leaves behind a wife

by ent270 ,ent223 updated 9:35 am et ,mon march 2 ,2015 (ent223) ent63 went famillial for fall at its fashion show in ent231 on sunday ,dedicating its collection to `` mamma " with nary a pair of `` mom jeans " in sight .ent164 and ent21 , who are behind the ent196 brand ,sent models down the runway in decidedly feminine dresses and skirts adorned with roses ,lace and even embroidered doodles by the designers 'own nieces and nephews .many of the looks featured saccharine needlework phrases like `` i love you , . . .

X dedicated their fall fashion show to moms

Explicabilité



(b) A person is standing on a beach with a surfboard.

EXPLICABILITÉ OU INTERPRÉTABILITÉ ?

by ent423 ,ent261 correspondent updated 9:49 pm et ,thu march 19,2015 (ent261) a ent114 was killed in a parachute accident in ent45 ,ent85 ,near ent312 ,a ent119 official told ent261 on wednesday .he was identified thursday as special warfare operator 3rd class ent23 ,29 ,of ent187 , ent265 .`` ent23 distinguished himself consistently throughout his career .he was the epitome of the quiet professional in all facets of his life ,and he leaves an inspiring legacy of natural tenacity and focused . . .

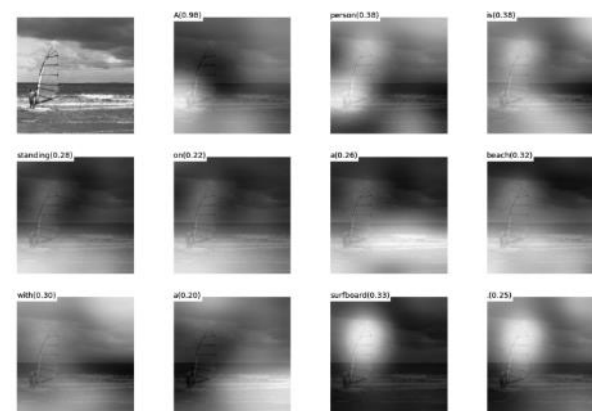
by ent270 ,ent223 updated 9:35 am et ,mon march 2 ,2015 (ent223) ent63 went familial for fall at its fashion show in ent231 on sunday ,dedicating its collection to `` mamma " with nary a pair of `` mom jeans " in sight .ent164 and ent21 , who are behind the ent196 brand ,sent models down the runway in decidedly feminine dresses and skirts adorned with roses ,lace and even embroidered doodles by the designers 'own nieces and nephews .many of the looks featured saccharine needlework phrases like `` i love you , . . .

ent119 identifies deceased sailor as X ,who leaves behind a wife

X dedicated their fall fashion show to moms

Explicabilité

- Il y a un **traitement supplémentaire** effectué
- Je **n'interprète pas le fonctionnement interne**, ni comment la décision est prise
- J'explique uniquement sur quelle partie de l'entrée le modèle a travaillé



(b) A person is standing on a beach with a surfboard.

Explicabilité

En NLP, en traitement d'image : les modèles à l'état de l'art sont peu interprétables.

Le deep learning est généralement peu interprétable.

-> On va se tourner sur **l'explicabilité**.

- Modèle agnostique
 - Frameworks tels que **Lime, Shapely, ...**
 - Modèle agnostique mais pas data agnostique.
- Dépendants du modèle
 - L'attention peut être utilisée pour visualiser si le modèle en a
 - Le traitement exact dépend de ce que reçoit l'attention et sa place dans l'architecture du modèle.

DEUX APPROCHES: LOCAL VS GLOBAL

Global:

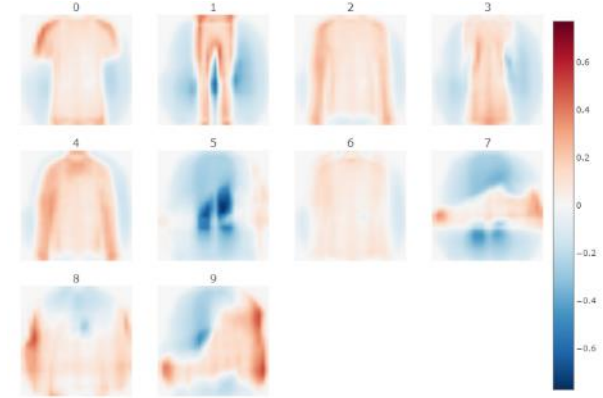
Fournit une explication sur l'ensemble du réseau et sur l'ensemble du jeu de données .

Pour comprendre si mon modèle a appris les bonnes features

Local:

Pour une seule prédiction, pourquoi cette prédiction est sortie

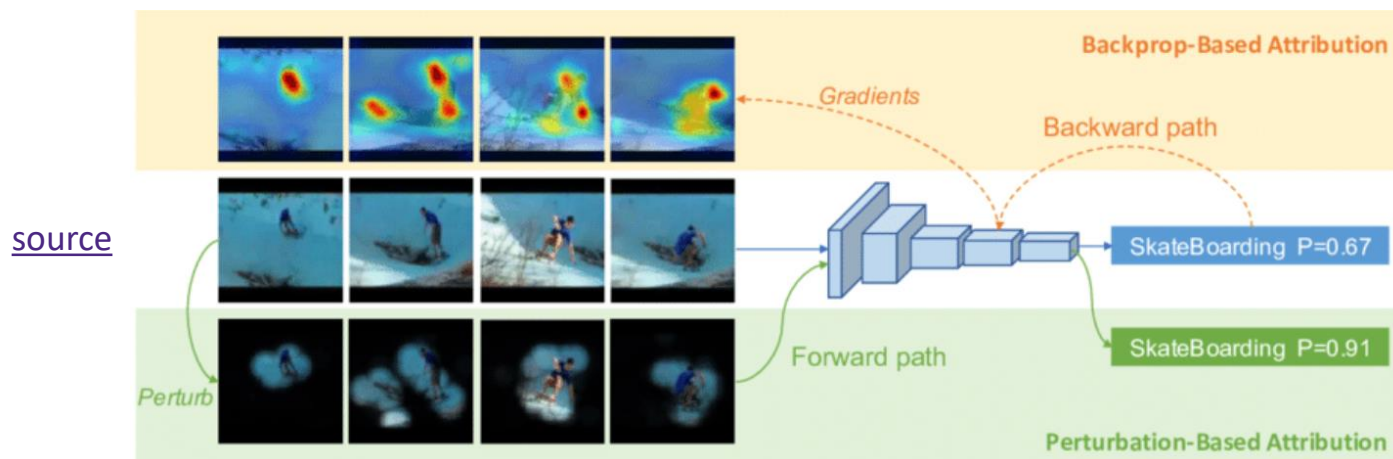
Pour un patient, client, ou pour la défense



DEUX APPROCHES DE CALCUL: PERTURBATIONS ET GRADIENTS

Perturbations :

Consiste à modifier les données envoyées au modèle afin de conclure quelle partie de la donnée permet d'obtenir une certaine prédiction



Gradients:

On part de la sortie du réseau et on attribue un score d'importance à chaque valeur intermédiaire calculée lors du *forward*.

FAIRE UNE EXPLCIATION PARFAITE

Idéalement, nous voudrions que nos explications vérifient certaines propriétés, comme par exemple :

- **La fidélité** : une explication est fidèle si elle reflète correctement le comportement du modèle.
- **La stabilité** : une explication est stable si elle ne change pas lorsque l'entrée subit des transformations sémantiques.
- **La consistance** : une explication est consistante si, pour différents modèles entraînés sur la même tâche, les explications sont similaires.
- **La représentativité** : une explication est représentative si elle couvre bien les différents phénomènes qui peuvent avoir lieu dans une scène.

L'explicabilité est un des domaines de recherche les plus actifs. Il existe actuellement des méthodes pour expliquer mais elles présentent un certain nombre de limitations (long à calculer, pas assez précis, marche moins bien sur des modèles complexes, etc.)

Pour ces raisons, l'explicabilité en industrie n'est pas encore déployé et utilisé pour les modèles de Deep Learning.

Ce facteur peut être limitant est impliqué que seul les modèles à bases d'arbres peuvent être utilisés dans certains domaines spécifiques (pas forcément d'un point de vue légal, mais plus d'acceptation par le métier)



02

LIME

- Framework python, activement maintenu
- Opensource. <https://github.com/marcotcr/lime> : 8000+ stars, 1000+ forks
- Modèle agnostique
- Condition sur les données : savoir les découper en **morceaux pertinents**
 - **Images, textes**
 - Series temporelles ?
 - Ne m'a **pas convaincu sur des images Sar**
 - Bruits, couleurs rendent la segmentation compliquée
- Heuristique
 - Quand l'explication n'est pas convaincante, cela :
 - Peut venir du modèle qui ne fonctionne pas aussi bien que prévu
 - Peut venir de Lime qui ne réussit pas à expliquer
- Fonctionnement (repris dans les slides suivantes) :
 - <https://www.oreilly.com/content/introduction-to-local-interpretable-model-agnostic-explanations-lime/>

Image Initiale



Modèle entraîné

Classe : crapaud
Score : 0.62

Entrée : échantillon et modèle entraîné

LIME

Sortie : Éléments de l'échantillon ayant le plus influencés la décision du modèle



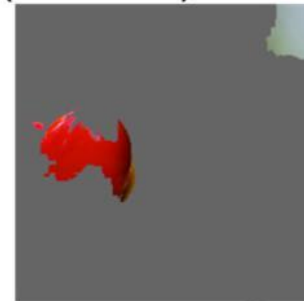
$P(\text{ }) = 0.54$



$P(\text{ }) = 0.07$



$P(\text{ }) = 0.05$



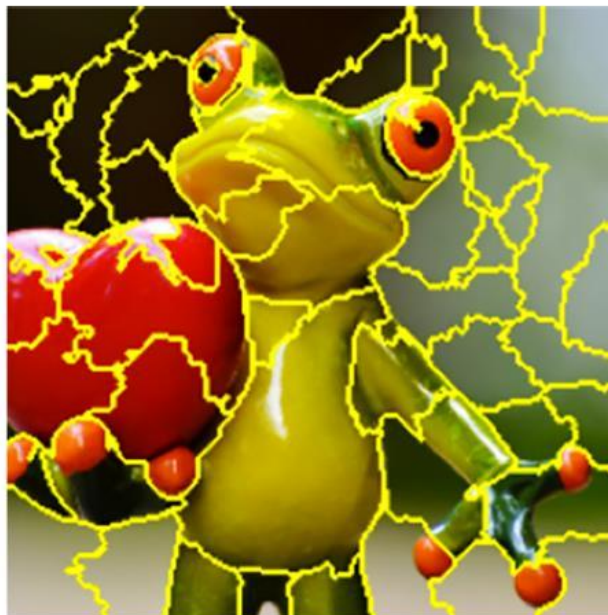
COMMENT MARCHÉ LIME ?

Image Initiale



Modèle entraîné

Classe : crapaud
Score : 0.62



Etape 1 : séparer l'entrée en
partie en « **composants
interprétables** »

- **Suffisamment gros** pour que le modèle arrive à prédire dessus
- **Suffisamment petits** pour que leur nombre donne une explication de bonne qualité

COMMENT MARCHÉ LIME ?







Image Initiale



Modèle entraîné

Réponse :
crapaud 0.85



	 0.85
	 0.00001
	 0.52

Etape 2 : Envoyer des **combinaisons de « composants interprétables »** au modèle entraîné

- Le modèle donne sa prédiction et le score associé

COMMENT MARCHÉ LIME ?

Image Initiale



Modèle entraîné

Réponse :
crapaud 0.85

	 0.85
	 0.00001
	 0.52

Données

Labels

Dataset de régression

Etape 2 : Envoyer des combinaisons de « **composants interprétables** » au modèle entraîné

- Le modèle donne sa prédiction et le score associé
- L'ensemble (combinaisons, scores) forme un **dataset de régression**

COMMENT MARCHE LIME ?

Image Initiale



Modèle entraîné

Réponse :
crapaud 0.85



	 0.85
	 0.00001
	 0.52

Données Labels

Dataset de régression

Etape 2 : Envoyer des combinaisons de « **composants interprétables** » au modèle entraîné

- Le modèle donne sa prédiction et le score associé
- L'ensemble (combinaisons, scores) forme un **dataset de régression**
 - **Nouvelle tâche** : prédire le score donné par le modèle entraîné sur une combinaison de composants interprétable

COMMENT MARCHÉ LIME ?

Image Initiale



Modèle entraîné

Réponse :
crapaud 0.85



Données Labels

Dataset de régression

Etape 2 : Envoyer des combinaisons de « **composants interprétables** » au modèle entraîné

- Le modèle donne sa prédiction et le score associé
- L'ensemble (combinaisons, scores) forme un **dataset de régression**
 - **Nouvelle tâche** : prédire le score donné par le modèle entraîné sur une combinaison de composants interprétable

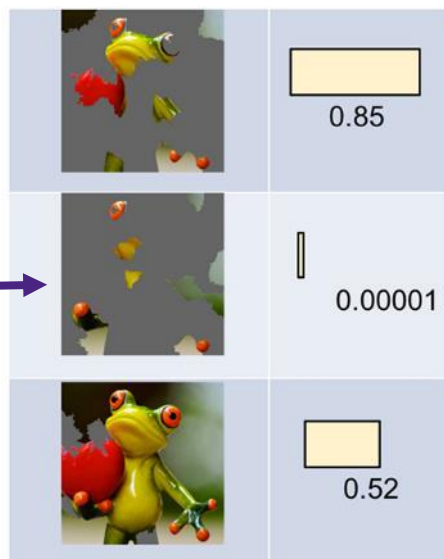
COMMENT MARCHE LIME ?

Image Initiale

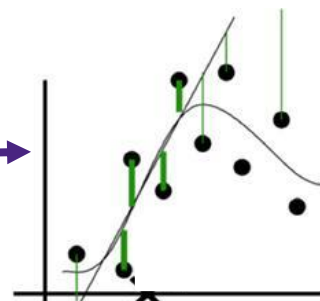


Etape 3 : Entrainer un modèle de **régression interprétable** sur ce jeu de donnée de regression.

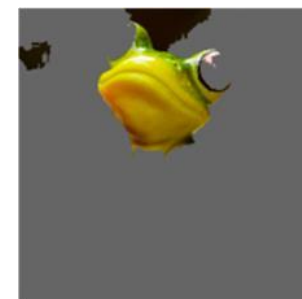
- LIME utilise une **régression linéaire**
- Les features les plus utilisées pour la régression linéaire sont **l'explication** : les features décidant la décision du modèle entraîné sur l'image initiale



Dataset de regression



Régression linéaire
interprétable



Meilleures features

Lime, pour résumer :

- Entraîne un second modèle interprétable à prédire la décision du modèle à expliquer
 - **Très sensible au découpage en morceau utilisé pour l'entraînement**
 - Comment choisir ces morceaux ?
 - Comment compléter le reste de l'image/du texte ? (« valeur zero » ?)
- Les morceaux **les plus utiles au modèle interprétable** permettent **d'expliquer la décision modèle initial**
 - **Nécessite que le second modèle réussisse correctement à prédire le score du premier.**
 - Ce qui n'est pas forcément un problème facile !

Pros :

- Modèle agnostique
- Facilement interprétable même pour quelqu'un ne connaissant pas l'IA

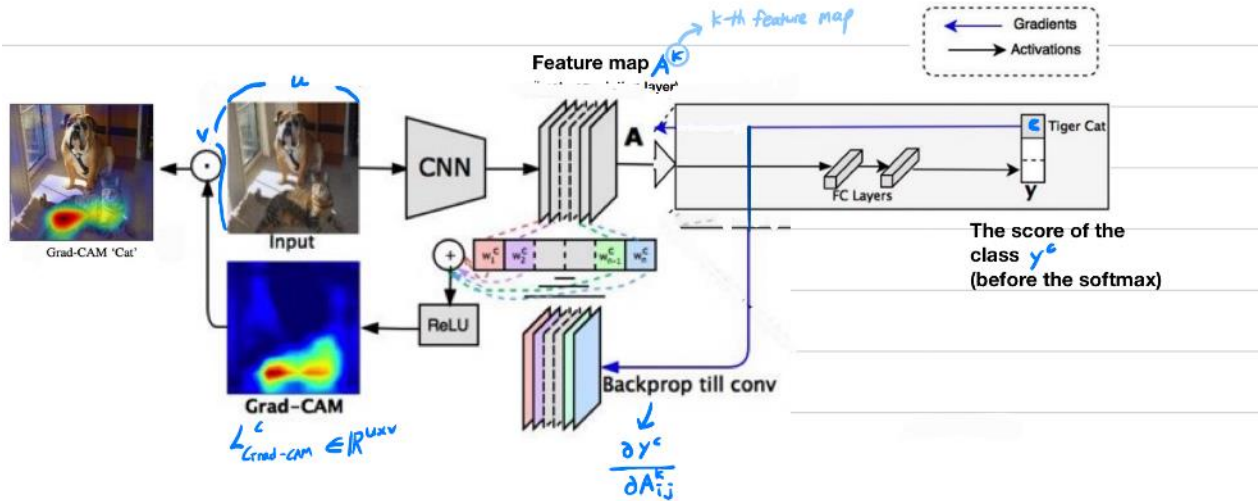
Cons :

- Heuristique, peut ne pas bien fonctionner
- **Long**
- **Explicabilité locale (i.e sur un échantillon, par opposition à globale, sur l'ensemble des données)**



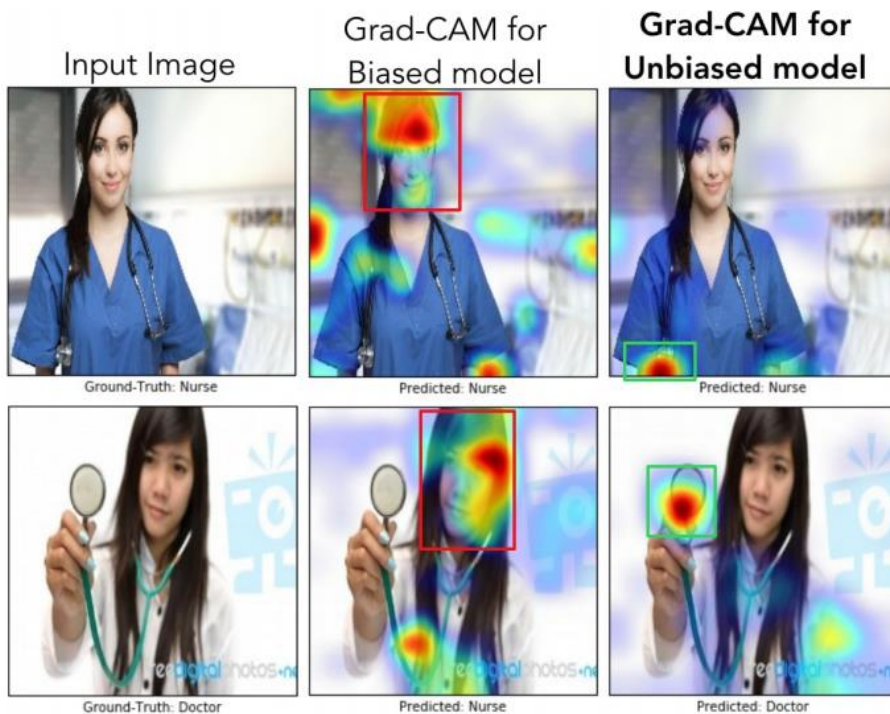
3

GRADIENTS



- Travail à partir de la dernière couche de convolution
- Evalue les features maps qui ont le plus d'importance et donc les parties de l'image
- Ne Nécessite pas de réentraînement
- Agit comme une couche d'attention post-entraînement

[source](#)



[source](#)

Explainable AI

Cheat sheet ex.pegg.io v.0.2

