

Introduction à l'IA explicable

Antoine Cornuéjols

AgroParisTech – INRAE MIA paris-Saclay

antoine.cornuejols@agroparistech.fr

Plan

1. Motivation
2. Les familles de méthodes
3. Interprétabilité par conception (by design)
4. Interprétation post-hoc
 - a. Indépendante du modèle (model-agnostic)
 - b. Dépendante du modèle (model-specific)
5. Quel avenir pour l'IA explicable ?

Les systèmes experts : MYCIN

MYCIN : Ma recommandation thérapeutique sera en premier choix : pour couvrir les points <1 2 3 5 6>, prescrire de la Gentamicine, dose 119 mg 98 h IV, pendant 10 jours;

Pour couvrir le point 4, prescrire de la Clindamycine, dose 595 mg 96 h IV, pendant 14 jours.

Désirez-vous éventuellement éliminer certains médicaments ?

Médecin : Oui, la Clindamycine.

MYCIN : Pour couvrir les points <2 3 4 5 6> donner du Chloramphénicol, dose 1 g 96 h IV, pendant 14 jours;

** Surveiller le taux de globules blancs du patient.

Pour couvrir le point 1, donner de la gentamicine, dose 119 mg 98 h IV, pendant 10 jours.

** Modifier la dose en cas de faiblesse rénale.

Les explications des SEs

MYCIN : L'infection du malade a t-elle été contractée à l'hôpital ?

Utilisateur : pourquoi ?

TEIRESIAS : [pourquoi est-ce important de savoir cela ?]

Cela aidera à déterminer la catégorie à laquelle appartient l'organisme responsable de l'infection :

- il a déjà été établi que 1° l'organisme est à Gram négatif, 2° c'est un bâtonnet, 3° il se développe à l'air ou sans air,

- donc si l'infection par cet organisme n'a pas été contractée lors du séjour à l'hôpital alors il est presque certain (0,8) que l'organisme est de la catégorie des enterobacteriaceae; de plus, il se peut (0,1 ; règle 37) que ce soit la pseudomonias aeruginosa.

Utilisateur : pourquoi n'avez-vous pas envisagé la possibilité que l'organisme soit un streptocoque ?

TEIRESIAS : la règle 33 aurait du être utilisée pour identifier l'organisme comme un streptocoque, mais la seconde clause "la morphologie est de type coque" était déjà connue comme fausse et la règle 33 n'a jamais été essayée.

Les explications des SEs

Pourquoi ne faut-il pas prescrire de tétracycline à un enfant de moins de 8 ans ?

Connaissances justificatives

Dépôt de la drogue sur les **os en développement**

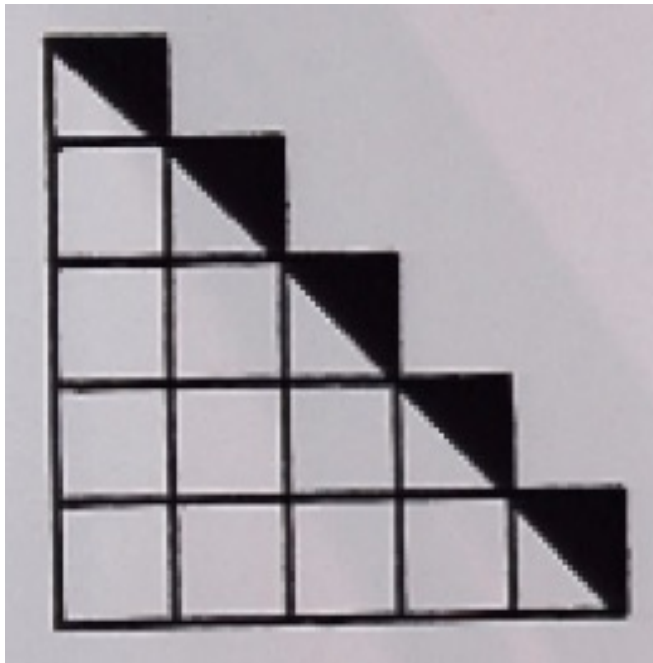
- **Noircissement** définitif des dents
 - Coloration socialement **indésirable**
 - **Ne pas administrer** de tétracycline aux enfants de moins de 8 ans

Notion d'**effets secondaires** indésirables

Relations de **causalité**

Raisonnement graphique

$$1 + 2 + 3 + \dots + n \stackrel{?}{=} \frac{n^2}{2} + \frac{n}{2}$$

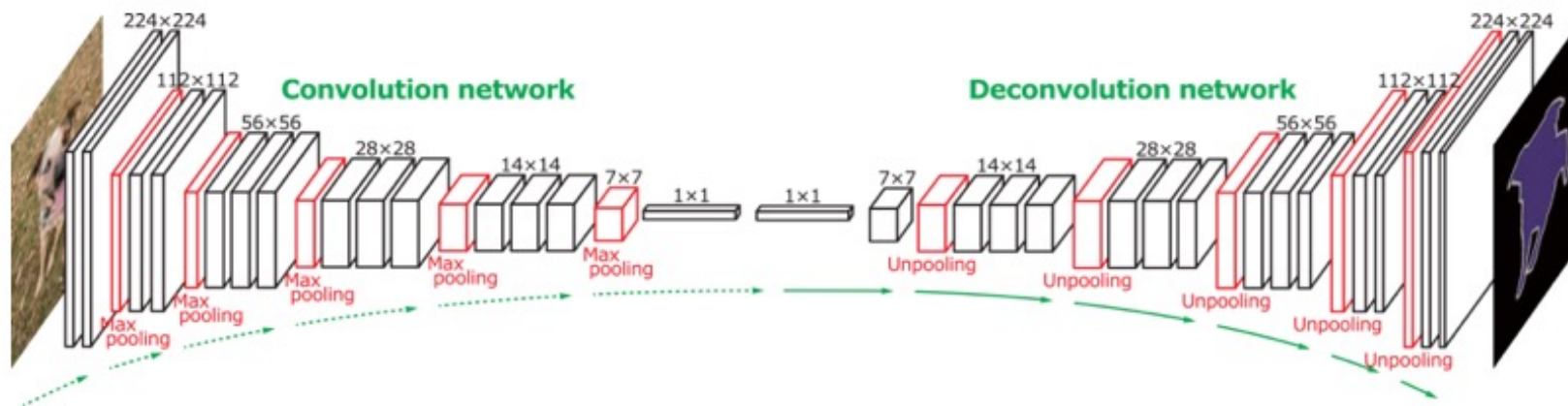


Types d'explications

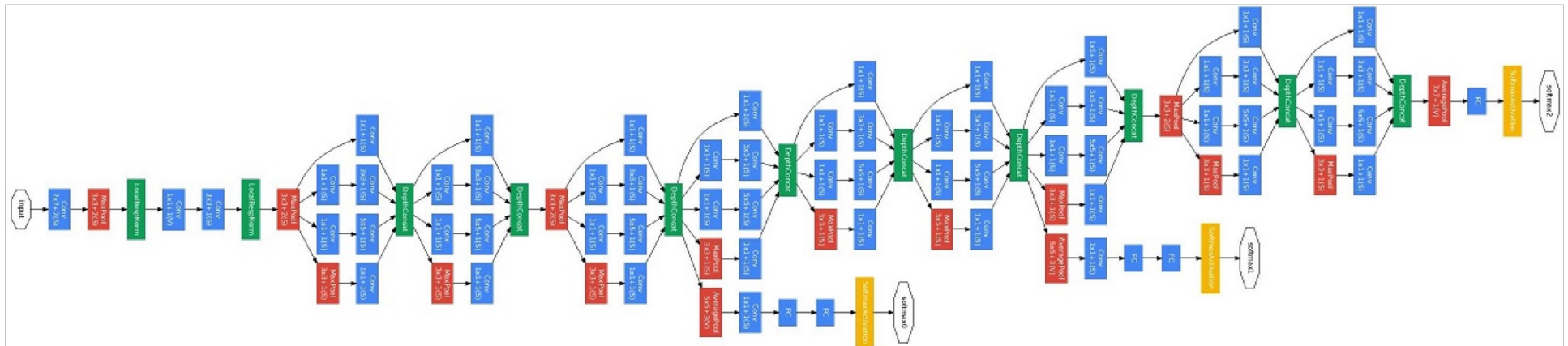
- De « **surface** »
 - Liées au **fonctionnement** du système
 - Adéquates pour vérifier ce fonctionnement
 - **Expert IA**
- « **profondes** »
 - Difficiles à **découvrir** et à **expliciter**
 - Nécessaires pour l'**expert du domaine**
 - Pour les **utilisateurs**

Les « réseaux de neurones **profonds** »

- Des réseaux de neurones artificiels
 - à grand nombre de couches (parfois > qqs 100)
 - et **très grand nombre de paramètres** (qqs $10^7 - 10^9$ paramètres)



GoogleNet



Explications et réseaux de neurones profonds

Illusions d'optique : quelle explication ?



Boxer: 0.40 Tiger Cat: 0.18

(a) Original image



Airliner: 0.9999

(b) Adversarial image

!!??

[Selvaraju et al. (2017) « *Grad-CAM: Visual explanations from deep networks via gradient-based localization* »]

Exemple en médecine

MACHINE LEARNING

Adversarial attacks on medical machine learning

Emerging vulnerabilities demand new conversations

22 March 2019

Science

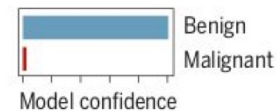
The anatomy of an adversarial attack

Demonstration of how adversarial attacks against various medical AI systems might be executed without requiring any overtly fraudulent misrepresentation of the data.

Original image



Dermoscopic image of a benign melanocytic nevus, along with the diagnostic probability computed by a deep neural network.



Diagnosis: Benign

The patient has a history of back pain and chronic alcohol abuse and more recently has been seen in several...

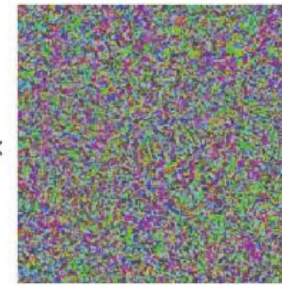
Opioid abuse risk: High

277.7 Metabolic syndrome
429.9 Heart disease, unspecified
278.00 Obesity, unspecified

Reimbursement: Denied

« Inexplicable »

Adversarial noise



Perturbation computed by a common adversarial attack technique. See (7) for details.

Adversarial example



Combined image of nevus and attack perturbation and the diagnostic probabilities from the same deep neural network.



Diagnosis: Malignant

The patient has a history of lumbago and chronic alcohol dependence and more recently has been seen in several...

Opioid abuse risk: Low

401.0 Benign essential hypertension
272.0 Hypercholesterolemia
272.2 Hyperglyceridemia
429.9 Heart disease, unspecified
278.00 Obesity, unspecified

Reimbursement: Approved

Le cas AlphaGo

- Un joueur « extraterrestre »
- Un jeu stupéfiant
- Révolutionne la manière de jouer
- Effervescence dans les écoles de go



Découverte
scientifique



A basic principle

- Machine Learning “just” **reformulates** what has been given as **input**
- A **conservation** theorem:
 - **No information is “added”**
 - **Data + prior knowledge**

A basic principle

- Machine Learning “just” reformulates what has been given as input
- A conservation theorem:
 - No information is “added”
 - Data + prior knowledge

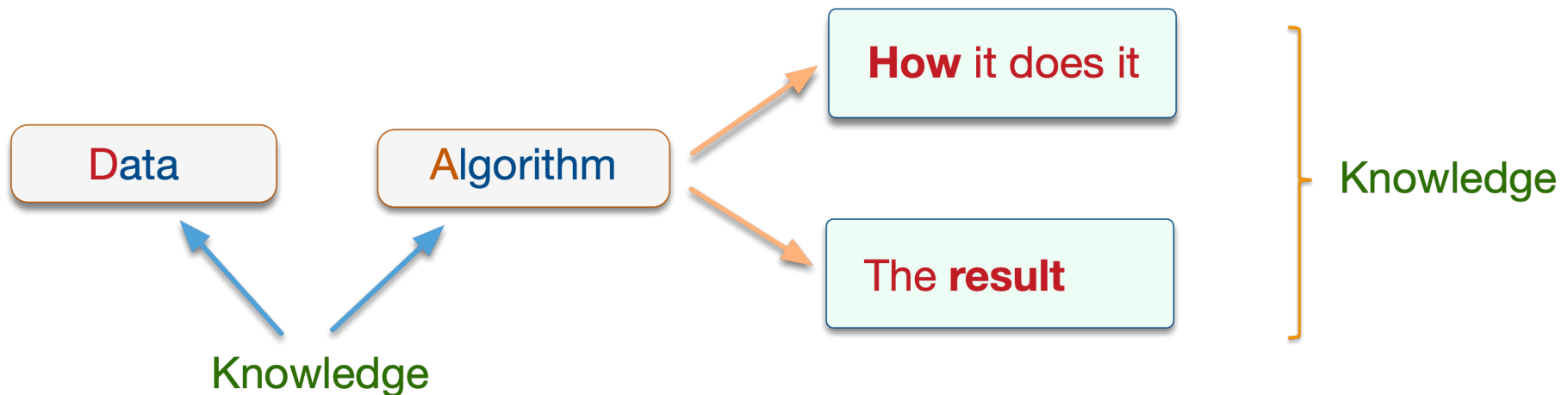
Little data + **lots** of prior knowledge

Big data + **less** prior knowledge

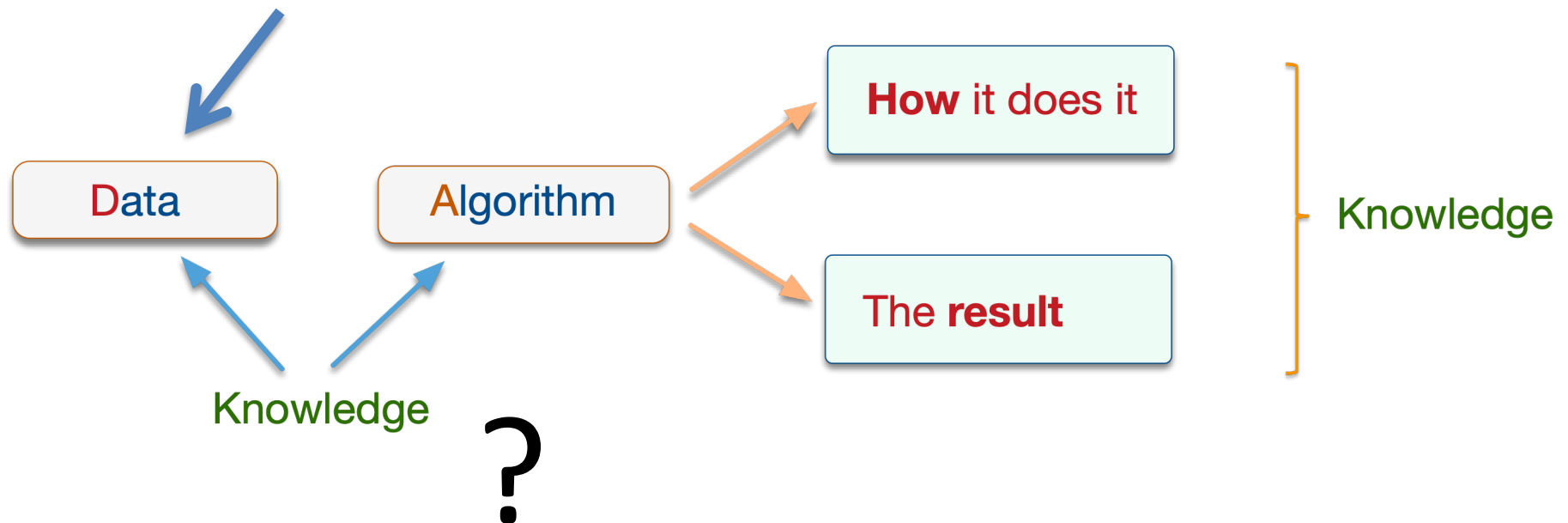
A basic principle

- Machine Learning “just” **reformulates** what has been given as **input**
- A **conservation** theorem:
 - **No information is “added”**
 - **Data + prior knowledge**

Little data + **lots** of prior knowledge
Big data + **less** prior knowledge

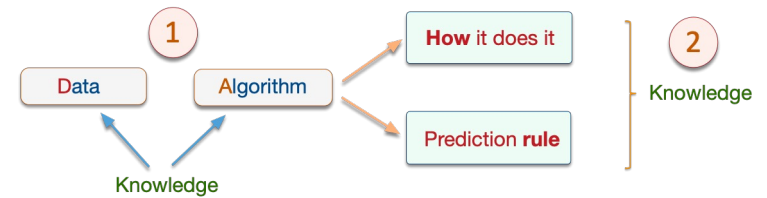


Biais sur les données



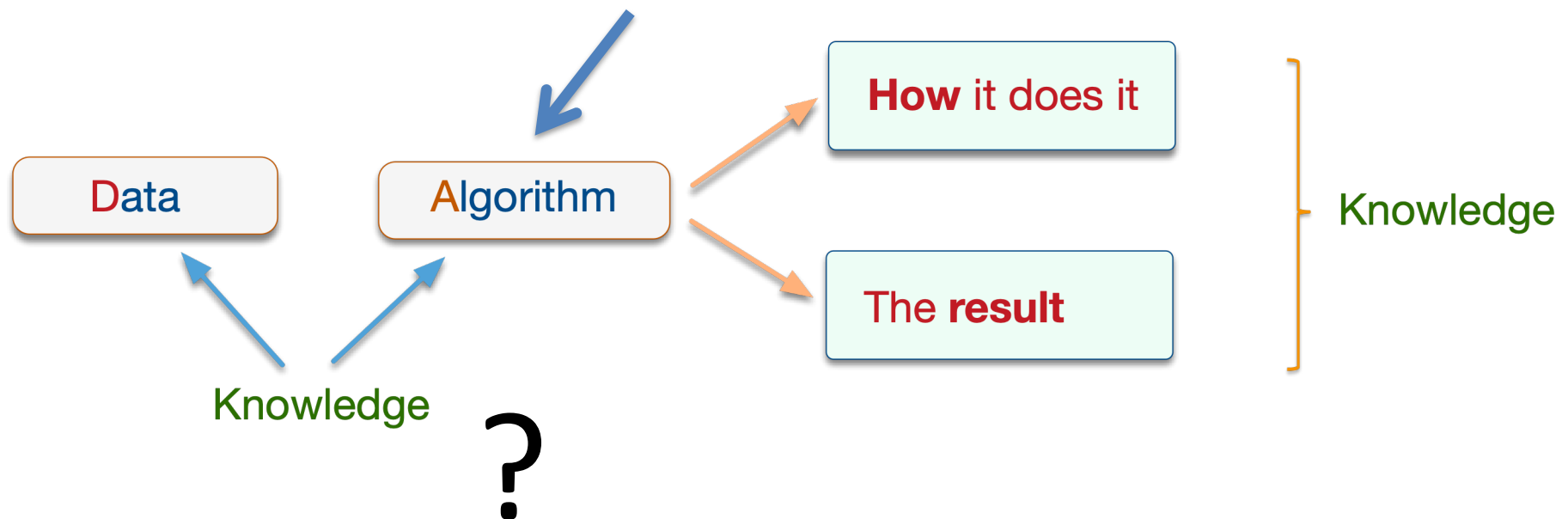
...

Knowledge as **input** to ML



- Knowledge in the **data**
 - The **experimental apparatus**
 - Choice of **the descriptors** (the features)
 - **Enrichment** using ontologies
 - **Normalization** of the values
 - **Missing** values
 - Possibly **added data point**
 - With invariances in mind
 - ...

Biais sur les algorithmes

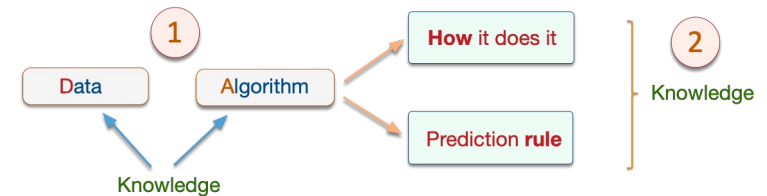


...

Knowledge as **input** to ML

- Knowledge **in the learning algorithm**

- Constraints on the hypothesis space: **representation bias**



$$h^* = \underset{h \in \mathcal{H}}{\text{ArgMin}} \left[R_{\text{Emp}}(h) + \lambda \text{reg}(h) \right]$$

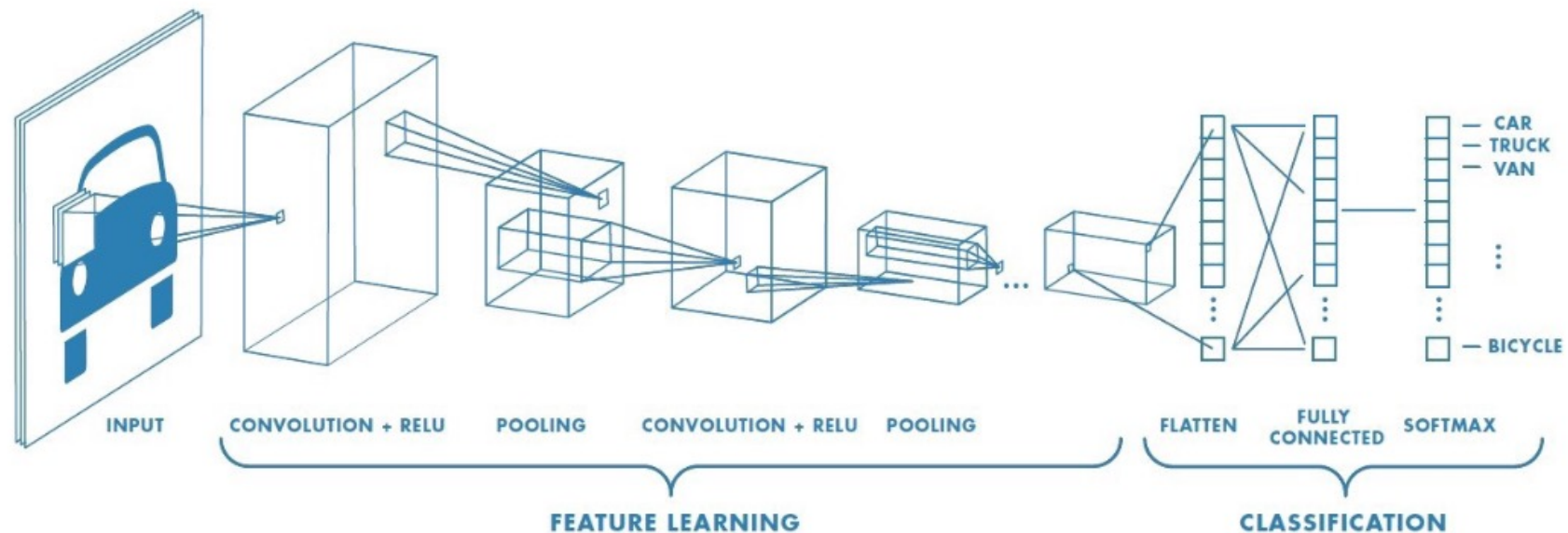
Looking for **sparse linear hypotheses**

$$h^* = \underset{h \in \mathcal{H}}{\text{ArgMin}} \left[\frac{1}{m} \sum_{i=1}^m \ell(h(\mathbf{x}_i), y_i) + \lambda ||h||_1 \right]$$

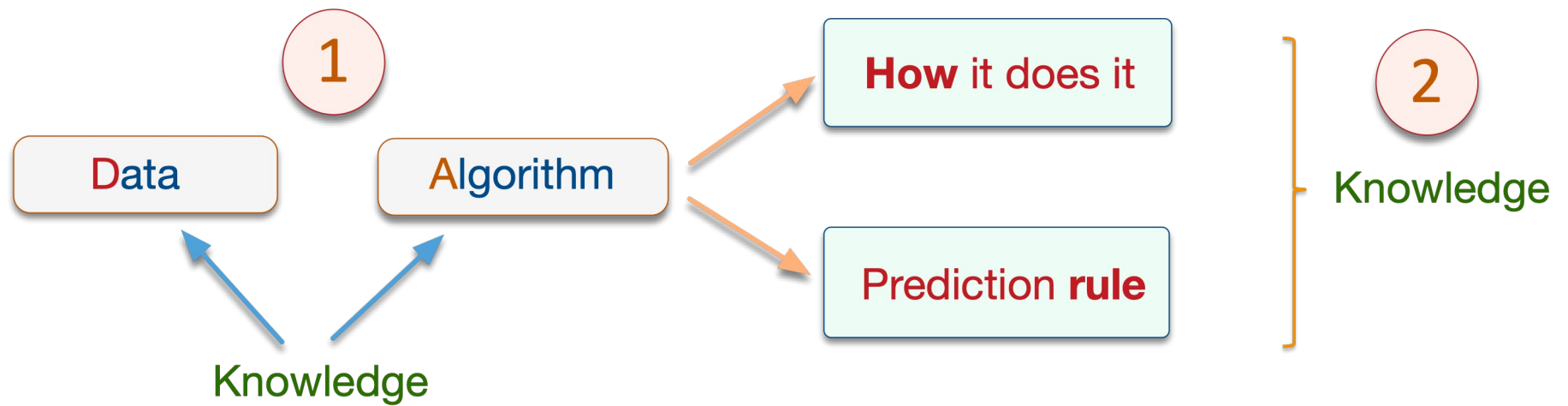
Favors hypotheses with few non null coefficients

Knowledge as **input** to ML

- Convolutional Neural Networks
 - Knowledge embedded in the **architecture of the network**



From <https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53>



What kind of knowledge can we **extract**?

- The **hypothesis returned**: a decision function

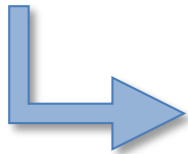
What kind of knowledge can we **extract**?

- The **hypothesis returned**: a decision function
 - **Recommending a movie**
 - **Recommending a life partner**
 - **Written character recognition**
 - **Recognize traffic signs**
 - **Decide the value of a position in go**
 - **Predict the risk of crime occurrence**
 - **Decide if someone should get a loan**
 - **Decide to hire or not someone**

What kind of knowledge can we **extract**?

- Just a decision function?
- Or **included** in a **larger inference system**?

E.g. in legal cases



Question **where** that decision **come from**

What kind of knowledge can we **extract**?

- **Interpretability** of the **decision function**
 - **Decision trees** seem readily interpretable
 - **Linear decision functions** are less so
 - **Random forests** are much less still
 - **SVM**
 - **Neural Networks**
- } Require a difficult analysis

What kind of knowledge can we **extract**?

- **Interpretability** of the **learning process** leading to a decision function
 - **Sensitivity analysis**
 - If this input value is changed, what happens
 - **Explanation-Based Learning**

What kind of knowledge can we **extract**?

- When interpretability is **NOT** needed?

What kind of knowledge can we **extract**?

- When interpretability is **NOT** needed?
 - When **low risk** associated with the decision
 - E.g. *recommendation for a movie*
 - When **good guarantees** on performance exist
 - E.g. *character recognition*

What kind of knowledge can we **extract**?

- When interpretability **IS** needed?

What kind of knowledge can we **extract**?

- When interpretability **IS** needed?
 1. With **high risk decisions**
 - *E.g. surgical operation*
 - *E.g. shutting down a nuclear plant*
 - *E.g. autonomous vehicle*

What kind of knowledge can we **extract**?

- When interpretability **IS** needed?
 1. With **high risk decisions**
 - *E.g. surgical operation*
 - *E.g. shutting down a nuclear plant*
 - *E.g. autonomous vehicle*
 2. Satisfying **curiosity** (what science is about)
 - *E.g. explain surprising results*
 - *E.g. when no easy explanation exists*
 - *E.g. when the decision function must be included in a larger inference system (a domain theory)*

What kind of knowledge can we **extract**?

- When interpretability **IS** needed?

What kind of knowledge can we **extract**?

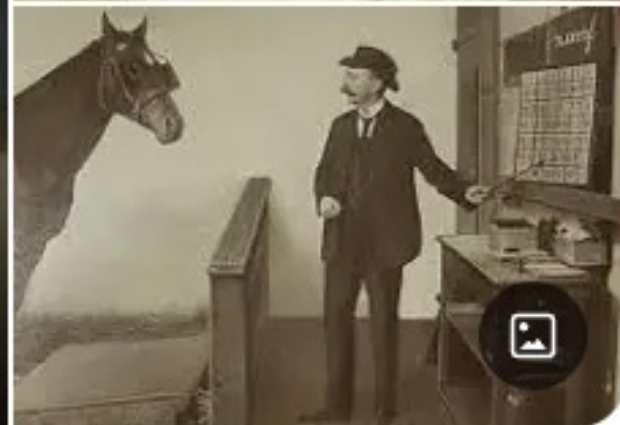
- When interpretability **IS** needed?

3. Debugging

- *E.g. why is that decision wrong (counterfactual)*
- *E.g. if a bicycle is recognized because it has two wheels, what if one is hidden behind side bags?*
- *E.g. why the system seems gender biased?*

“Clever Hans” effect

Hans le Malin :



...

Hans le Malin :



Hans le Malin est un cheval, élevé en Allemagne au début du XX^e siècle. Considéré comme ayant une « intelligence supérieure », il devient célèbre dans toute l'Europe. Ce cheval **semble**, en effet, **pouvoir répondre à toutes sortes de questions** : « En tapant du sabot, Hans était capable **d'additionner**, de **soustraire**, de **multiplier** et de **diviser**. Hans pouvait **épeler**, **lire**, et **résoudre des problèmes d'harmonie musicale** »^[1]. Son premier propriétaire, Wilhelm von Osten, l'éduque au calcul et à la lecture en toute bonne foi. Face à la controverse suscitée par la réalité de l'intelligence de Hans, plusieurs scientifiques étudient son cas.

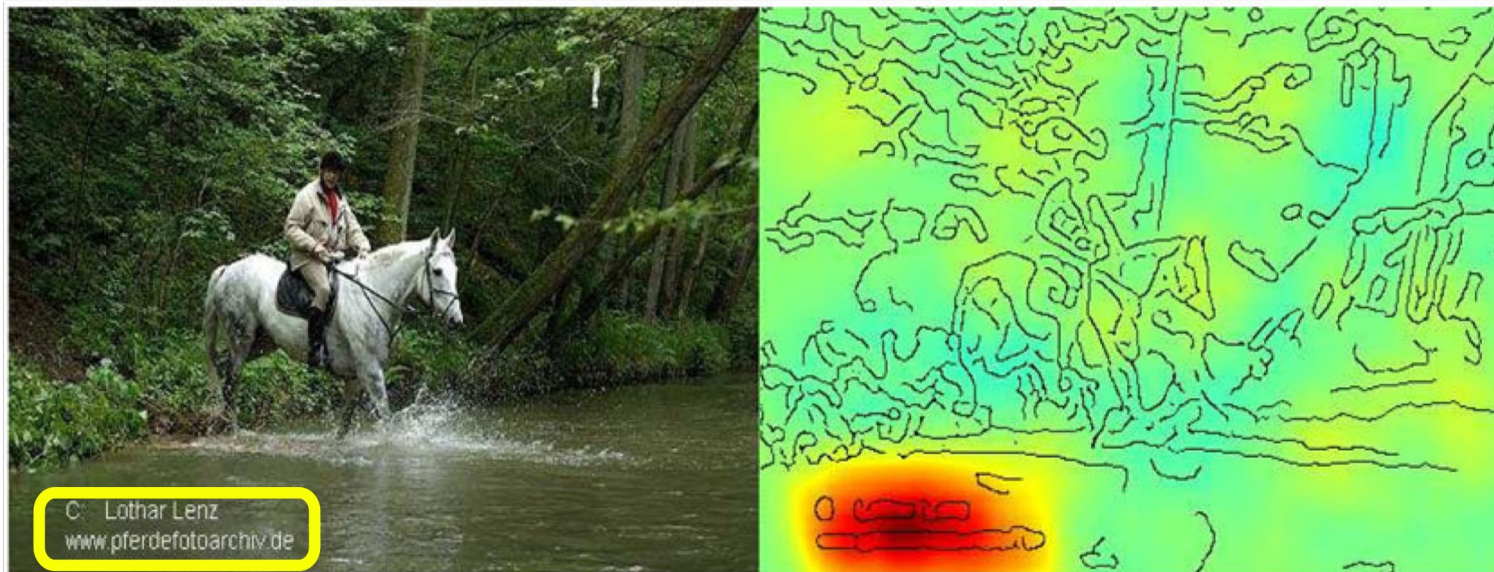
Hans le Malin :



Oskar Pfungst, un psychologue allemand qui a enquêté sur le cas de Hans, finit par découvrir ce qui deviendra en psychologie le « phénomène Hans le Malin », l'« effet Hans le Malin » ou encore l'« effet Clever Hans » (en anglais : *Clever Hans phenomenon* ou *Clever Hans effect*). Hans ne possède pas de capacité particulière en calcul ou en lecture. Il **interprète des signaux corporels émis inconsciemment par son maître ou par les personnes qui l'interrogent.**

Effet “clever Hans” en apprentissage **supervisé**

Horse-picture from Pascal VOC data set



9963 images de chevaux dans Pascal VOC

- Corrélations **accidentelles** (spurious) mais très prédictives

Effet “clever Hans” en apprentissage supervisé

Artificial picture of a car



Reconnu comme
un **cheval**

- Corrélations accidentelles (spurious) mais très prédictives
sauf quand ...



Explainable AI reveals Clever Hans effects in unsupervised learning models

Received: 27 July 2024

Accepted: 19 December 2024

Published online: 17 March 2025



Jacob Kauffmann^{1,2}, Jonas Dippel^{1,2,3}, Lukas Ruff ^{1,3}, Wojciech Samek ^{1,2,4},
Klaus-Robert Müller ^{1,2,5,6,7} ✉ & Grégoire Montavon ^{1,2,8} ✉

Unsupervised learning has become an essential building block of artificial intelligence systems. The representations it produces, for example, in foundation models, are critical to a wide variety of downstream applications. It is therefore important to carefully examine unsupervised models to ensure not only that they produce accurate predictions on the available data but also that these accurate predictions do not arise from a Clever Hans (CH) effect. Here, using specially developed explainable artificial

What kind of knowledge can we **extract**?

- When interpretability **IS** needed?

3. Debugging

- *E.g. why is that decision wrong (counterfactual)*
- *E.g. if a bicycle is recognized because it has two wheels, what if one is hidden behind side bags?*
- *E.g. why the system seems gender biased?*

4. Interpretability **demands higher standard predictive systems**

- *An interpretable system **can be manipulated***
 - *E.g. if someone knows that a loan is granted if you have more than 2 credit cards*



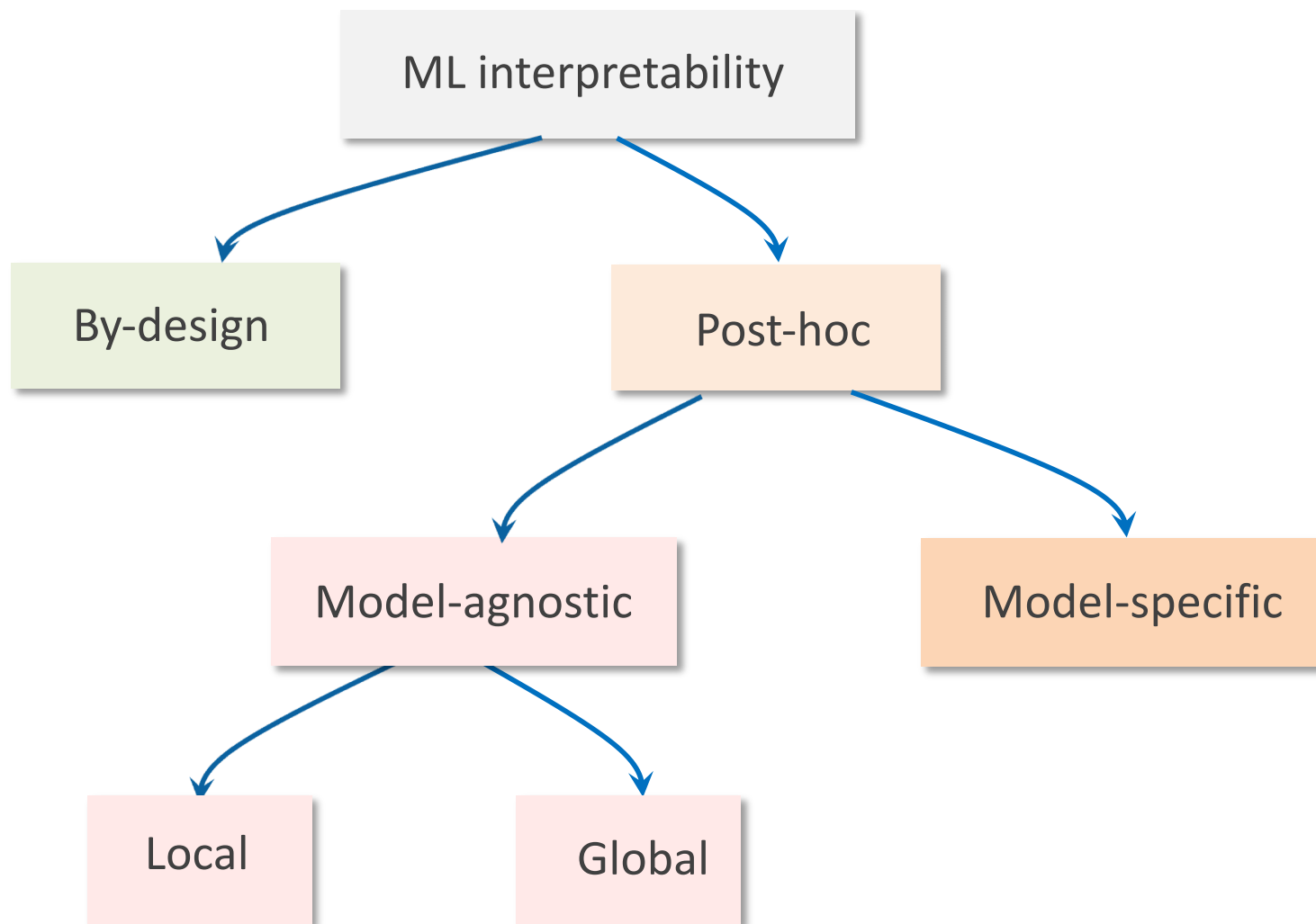
- *In order **not to be manipulated**,
the predictive system **must use causal factors***

Des algorithmes « transparent »

1. Dans lesquels on puisse « injecter » l'expertise humaine
2. Dont les résultats (modèles appris) soient interprétables

Plan

1. Motivation
2. Les familles de méthodes
3. Interprétabilité par conception (by design)
4. Interprétation post-hoc
 - a. Indépendante du modèle (model-agnostic)
 - b. Dépendante du modèle (model-specific)
5. Quel avenir pour l'IA explicable ?



• ...

- Méthodes « **par conception** » (*by-design*)
 - Fournissent de l'information sur l'importance relative des descripteurs par nature de la méthode
 - E.g. Régression linéaire
 - E.g. Arbre de décision
- Méthodes **post-hoc**
 - Découplent l'apprentissage d'un modèle de son explication**
 - **Indépendantes** du modèle (*model-agnostic*)
 - **Locale** : fournissent une « explication » relative à la prédiction associée à **un exemple**
 - **Globale** : fournissent une « explication » sur la **stratégie** de décision **globale** du modèle
 - **Dépendantes** du modèle (*model-specific*)
 - E.g. Cartes de **saillance** pour les réseaux de neurones
 - E.g. Identification des **concepts** mis en jeu dans la décision

Plan

1. Motivation
2. Les familles de méthodes
3. Interprétabilité par conception (by design)
4. Interprétation post-hoc
 - a. Indépendante du modèle (model-agnostic)
 - b. Dépendante du modèle (model-specific)
5. Quel avenir pour l'IA explicable ?

La régression linéaire

- Prédit la sortie par une **somme pondérée des descripteurs** (ici ***d***)

$$y = \beta_0 + \beta_1 \mathbf{x}^{(1)} + \dots + \beta_d \mathbf{x}^{(d)}$$

```
y = 674.95 + 0.00009*INCOME + -9.26*R_DEBT_INCOME + 253.13*R_EXPENDITURE + -414.78*R_EXPENDITURE_SQR + -98.45*R_ENTERTAINME  
NT + -7.09*GAMBLING_LOW + -24.82*GAMBLING_HIGH
```

- Les coefficients β sont typiquement trouvés par la **minimisation de l'écart quadratique** pour les ***m*** exemples d'apprentissage

$$\hat{\beta} = \underset{\beta_0, \dots, \beta_d}{\text{Argmin}} \sum_{i=1}^m \left(y_i - \left(\beta_0 + \sum_{j=1}^d \beta_j \mathbf{x}_i^{(j)} \right) \right)^2$$

La régression linéaire

- Pour être **correct**, un modèle linéaire **suppose** :
 - La fonction cible est **linéaire** (ou proche)
 - La sortie suit une **loi normale** étant donnés les descripteurs
 - Sinon, les estimateurs de confiance sont invalides
 - **L'homoscedasticité** : la variance est constante pour toutes les valeurs de chaque descripteur
 - E.g. Quelque soit la taille de la maison, la variance autour du prix prédit est constante
 - **Non corrélation** des descripteurs

Régression linéaire et interprétabilité

- Que nous disent les coefficients β_j ?
 - De **combien varie la variable de sortie** quand on modifie le **descripteur $x^{(j)}$** associé à β_j d'une unité
 - E.g. De combien varie le prix d'une maison *s'il y a une pièce de plus*

	coef	std err	t	P> t	[0.025	0.975]
const	674.9545	21.517	31.368	0.000	632.730	717.179
INCOME	9.011e-05	8.13e-06	11.084	0.000	7.42e-05	0.000
R_DEBT_INCOME	-9.2560	0.146	-63.485	0.000	-9.542	-8.970
R_EXPENDITURE	253.1289	83.841	3.019	0.003	88.604	417.654
R_EXPENDITURE_SQR	-414.7774	83.741	-4.953	0.000	-579.108	-250.447
R_ENTERTAINMENT	-98.4476	13.602	-7.238	0.000	-125.139	-71.756
GAMBLING_LOW	-7.0944	2.770	-2.561	0.011	-12.531	-1.658
GAMBLING_HIGH	-24.8233	2.074	-11.971	0.000	-28.892	-20.754

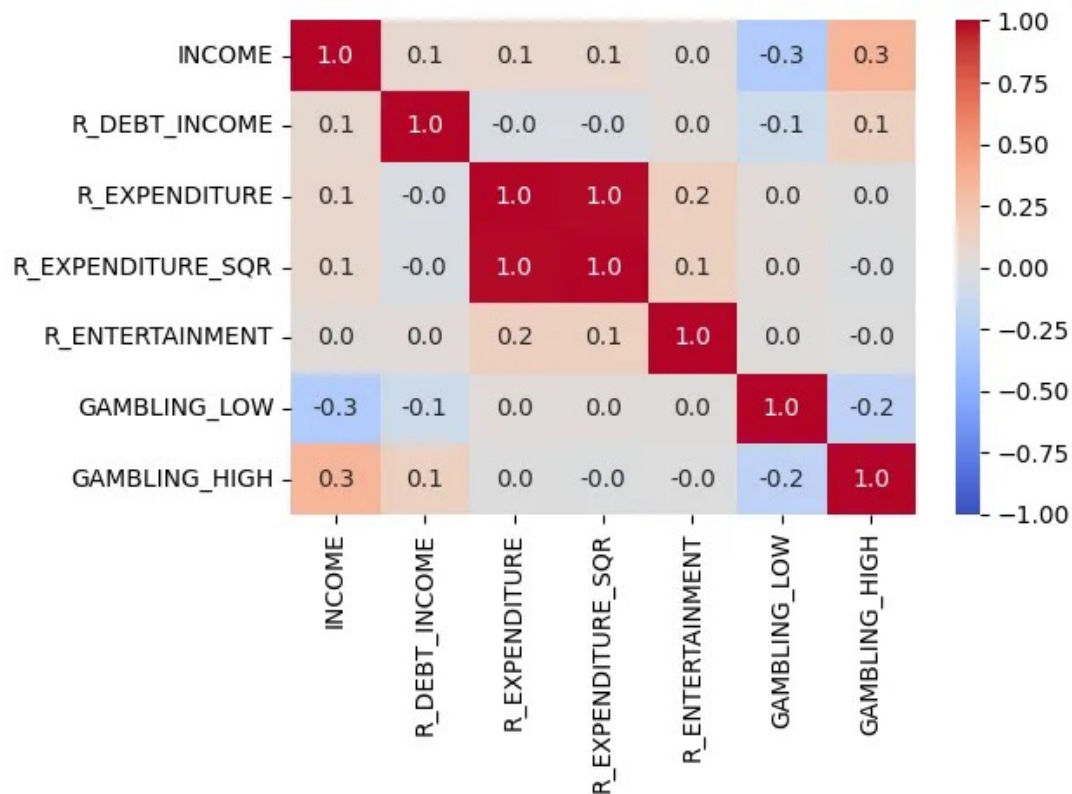
t-statistique : mesure le rapport du coefficient à son erreur standard. Une **plus grande valeur** indique un **coefficient plus important**.

$$t_{\hat{\beta}_j} = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)}$$

	coef	std err	t	P> t	[0.025	0.975]
const	674.9545	21.517	31.368	0.000	632.730	717.179
INCOME	9.011e-05	8.13e-06	11.084	0.000	7.42e-05	0.000
R_DEBT_INCOME	-9.2560	0.146	-63.485	0.000	-9.542	-8.970
R_EXPENDITURE	253.1289	83.841	3.019	0.003	88.604	417.654
R_EXPENDITURE_SQR	-414.7774	83.741	-4.953	0.000	-579.108	-250.447
R_ENTERTAINMENT	-98.4476	13.602	-7.238	0.000	-125.139	-71.756
GAMBLING_LOW	-7.0944	2.770	-2.561	0.011	-12.531	-1.658
GAMBLING_HIGH	-24.8233	2.074	-11.971	0.000	-28.892	-20.754

- **t-statistique** : mesure le rapport du coefficient à son erreur standard. Une plus grande valeur indique un coefficient plus important.
- **P > |t|** : la **p-valeur** pour un test d'hypothèse.
Une **petite p-valeur** (typiquement moins que 0.05) suggère que le **descripteur** correspondant a un impact **significatif** sur la variable de sortie
- **[0.025 0.975]** : l'intervalle de confiance pour chaque coefficient

Matrice de corrélation entre descripteurs



+1 : corrélation positive maximale

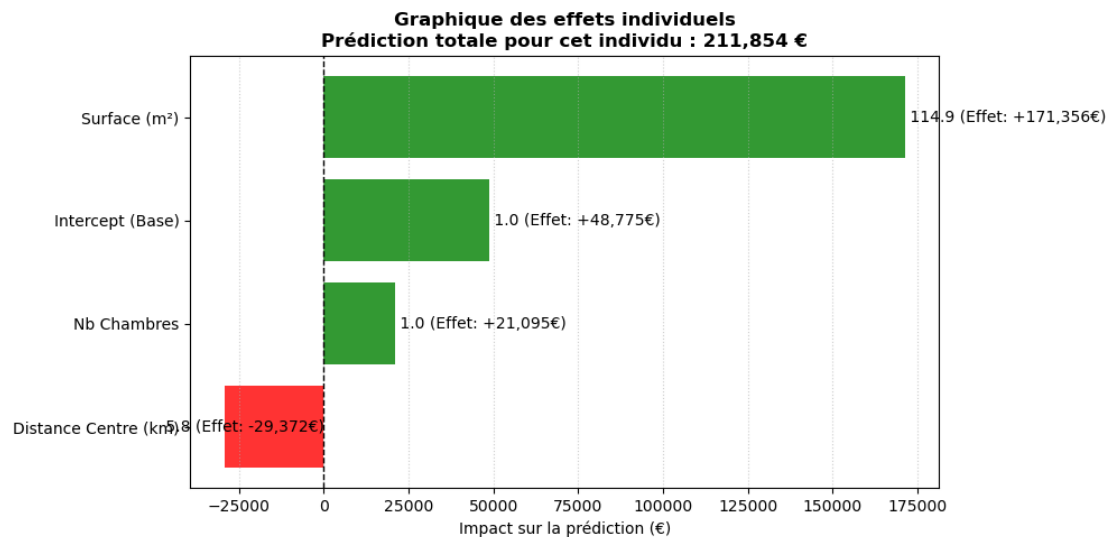
-1 : corrélation négative maximale

0 : pas de relation linéaire entre les descripteurs

- La matrice de corrélation révèle comment les variables de description sont **linéairement corrélées** entre elles
 - Si **corrélation élevée**, positive ou négative, alors les descripteurs sont **redondants**

Graphique des effets individuels (*individual effect plot*)

- Fournit une mesure de la contribution de chaque variable explicative (descripteur) à la prédiction pour un exemple donné (e.g. un patient)



```

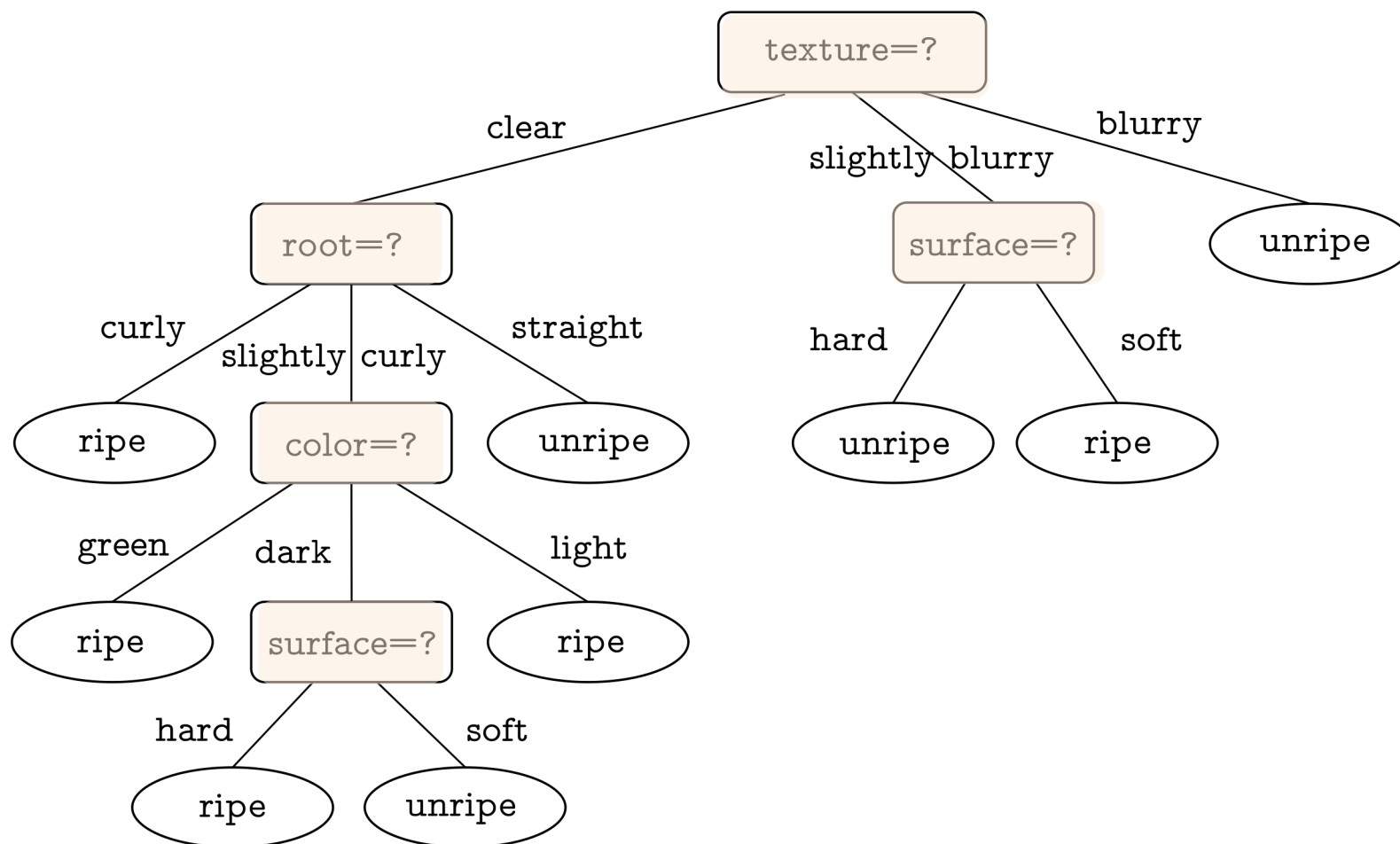
1 import numpy as np
2 import pandas as pd
3 import matplotlib.pyplot as plt
4 from sklearn.linear_model import LinearRegression
5
6 # 1. Génération de données fictives (Ex: Prix d'une maison)
7 np.random.seed(42)
8 n_samples = 100
9 surface = np.random.normal(100, 30, n_samples)
10 chambres = np.random.randint(1, 5, n_samples)
11 distance_centre = np.random.uniform(1, 15, n_samples)
12
13 # Prix théorique avec un peu de bruit aléatoire
14 prix = 50000 + 1500 * surface + 20000 * chambres - 5000 * distance_centre + np.random
15
16 X = pd.DataFrame({'Surface (m²)': surface, 'Nb Chambres': chambres, 'Distance Centre
17 y = prix
18
19 # 2. Entraînement du modèle de régression linéaire
20 model = LinearRegression()
21 model.fit(X, y)
22
23 # 3. Sélection d'un individu spécifique (le premier de la liste)
24 individu_idx = (constant) X: DataFrame
25 nom_features = X.columns
26 valeurs_individu = X.iloc[individu_idx].values
27
28 # 4. Calcul des effets individuels (Coefficient * Valeur)
29 coefficients = model.coef_
30 effets = coefficients * valeurs_individu
31 intercept = model.intercept_
32 prediction_totale = model.predict(X.iloc[[individu_idx]])[0]
33
34 # Création d'un DataFrame pour faciliter le traçage
35 df_effets = pd.DataFrame({
36     'Variable': nom_features,
37     'Valeur': valeurs_individu,
38     'Effet': effets
39 })
40
41 # Ajout de l'Intercept (la valeur de base commune à tous)
42 df_effets.loc[len(df_effets)] = ['Intercept (Base)', 1, intercept]
43
44 # Tri des effets pour une meilleure lecture graphique
45 df_effets = df_effets.sort_values(by='Effet', ascending=True)
46
47 # 5. Construction du graphique des effets individuels
48 plt.figure(figsize=(10, 5))
49 colors = ['red' if x < 0 else 'green' for x in df_effets['Effet']]
50
51 bars = plt.barh(df_effets['Variable'], df_effets['Effet'], color=colors, alpha=0.8)
52 plt.axvline(x=0, color='black', linestyle='--', linewidth=1)
53
54 # Ajout des étiquettes de texte (Valeur réelle et impact financier)
55 for bar, val, eff in zip(bars, df_effets['Valeur'], df_effets['Effet']):
56     label = f" {val:.1f} (Effet: +{eff:,.0f}€)" if eff >= 0 else f" {val:.1f} (Effet: -{abs(eff):,.0f}€)"
57     plt.text(bar.get_width() if eff >= 0 else 0, bar.get_y() + bar.get_height()/2,
58             label, va='center', ha='left' if eff >= 0 else 'right', fontsize=10)
59
60 plt.title(f"Graphique des effets individuels\nPrédiction totale pour cet individu : {prediction_totale}")
61 plt.xlabel("Impact sur la prédiction (€)")
62 plt.grid(axis='x', linestyle=':', alpha=0.6)
63 plt.tight_layout()
64 plt.show()
65

```

Avantages et **inconvénients** des modèles linéaires

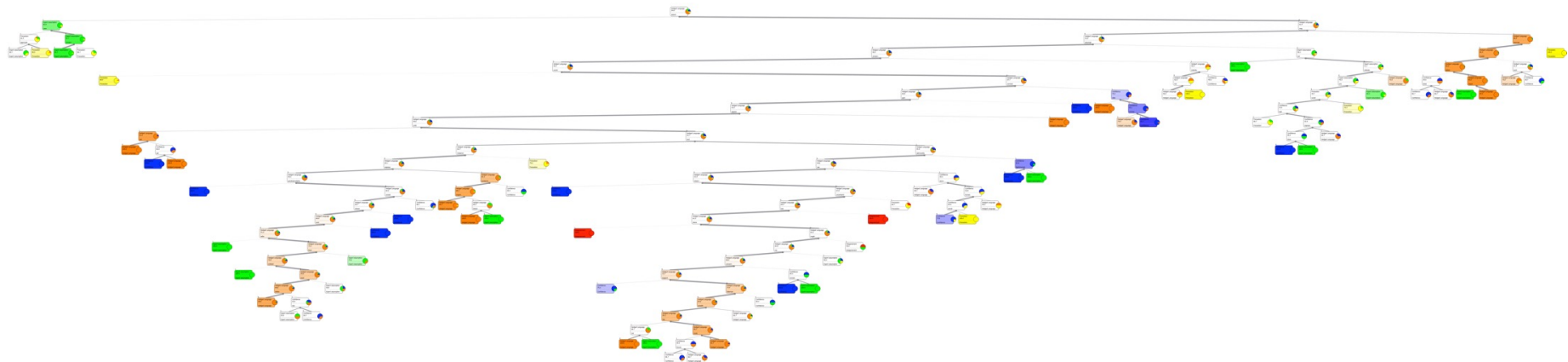
- **Supposent** les descripteurs **décorrélés**
 - E.g. **si deux descripteurs sont corrélés**
et que l'un a un très fort coefficient positif,
==> l'autre peut avoir un **coefficient négatif** alors même
qu'il est **corrélé positivement** en soi à la cible
- **Ne peuvent pas** bien rendre compte de **dépendances non linéaires**
entre les descripteurs et la cible
- **Simple**s à mettre en œuvre
- Permettent d'obtenir des **intervalles de confiance**

Les arbres de décision

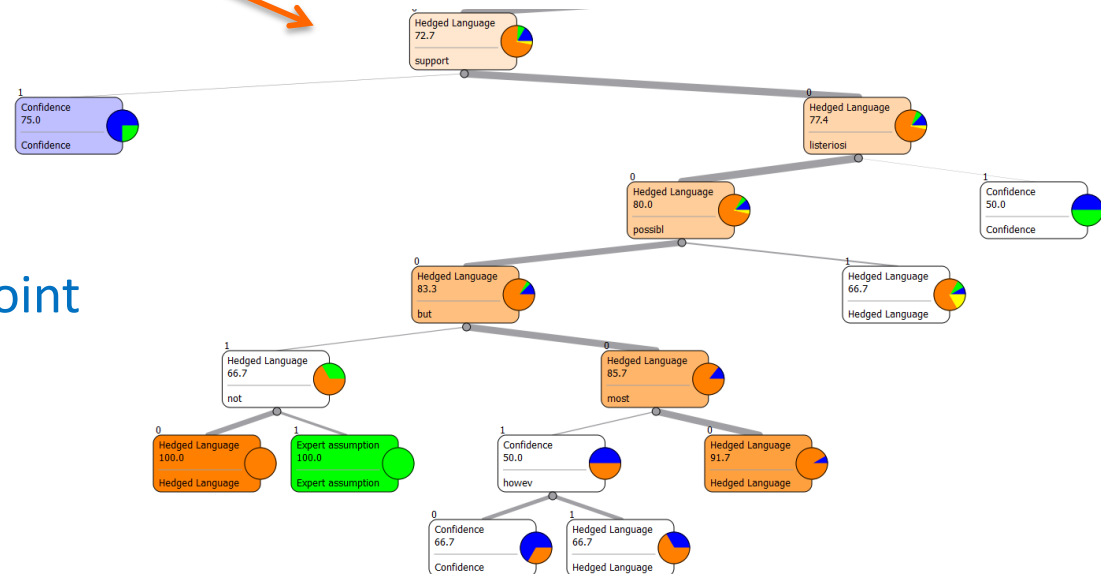


...

Les arbres de décision



Lisibles ... jusqu'à un certain point



Avantages et **inconvénients** des arbres de décision

- Capturent naturellement des **interactions** entre descripteurs
 - Montrent comment on obtient **différents groupes** dans les données
 - La structure en **arbre** est adaptée à l'**interprétation**
-
- **Ne sont pas** bien adaptés à capturer les **dépendances linéaires**
 - cf. arbres obliques
 - **Instables** : peuvent changer beaucoup pour de petits changements de la base d'apprentissage
 - **Peuvent être opaques** si les descripteurs sont mal choisis
 - E.g. l'explication des fins de parties aux échecs

Avantages et **inconvenients** des modèles par **conception** (ante hoc)

- Plus facilement **interprétables**
- Interprétation du **modèle** et pas seulement d'une prédiction donnée
- En général enclins à **sous-apprentissage** car trop simples pour capturer les interdépendances entre descripteurs et avec la cible

Plan

1. Motivation
2. Les familles de méthodes
3. Interprétabilité par conception (by design)
4. Interprétation post-hoc
 - a. Indépendante du modèle (model-agnostic)
 - b. Dépendante du modèle (model-specific)
5. Quel avenir pour l'IA explicable ?

Expliquer des prédictions **individuelles** : à propos d'un exemple

1. LIME

2. Shapley values

LIME

- LIME : **L**ocal **I**nterpretable **M**odel-agnostic **E**xplanations
 - Identifier un **modèle interpretable** localement **fidèle** au modèle appris

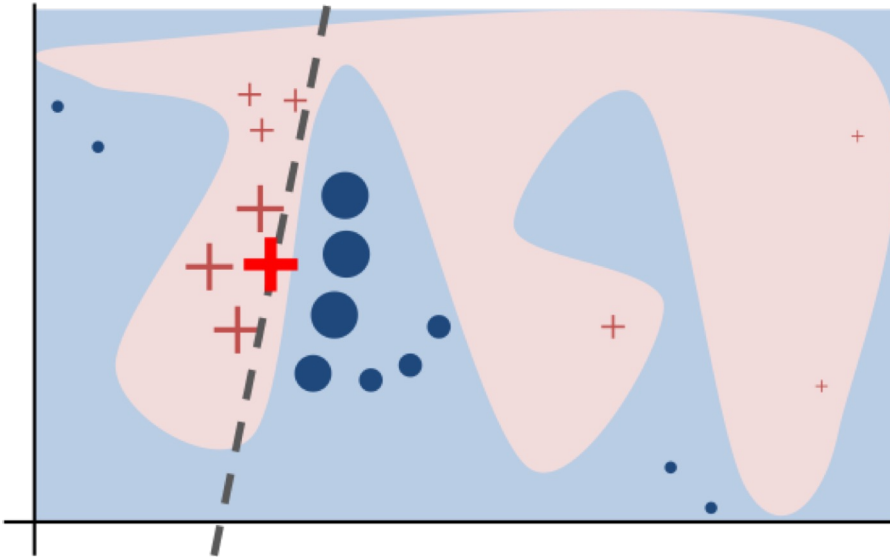
$$\text{explanation}(x) = \arg \min_{g \in G} L(f, g, \pi_x) + \Omega(g)$$

Diagram illustrating the LIME formula components:

- Our test Instance** points to x in the formula.
- Complex Model** points to f in the formula.
- Model Complexity** points to $\Omega(g)$ in the formula.
- g is a model from intrinsic interpretable family models G** points to $g \in G$ in the formula.
- Proximity Measure** points to π_x in the formula.

RIBEIRO, Marco Tulio, SINGH, Sameer, et GUESTRIN, Carlos. " **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*. 2016. p. 1135-1144.

LIME



- La fonction de décision f est complexe (inconnue) (entre bleu et rose)
- La croix rouge représente l'exemple à expliquer
- Les points bleus et croix rouges sont les exemples proches avec une taille fonction de la proximité
- La **droite en pointillés** est la fonction g apprise par LIME approximant localement la fonction f

Algorithme

1. Sélectionner **l'exemple** à expliquer
2. Générer des **exemples virtuels** en permutant des valeurs d'attributs
3. Assigner un **poids** à chacun de ces exemples virtuels en fonction de leurs distances à l'exemple
4. **Prédire** la classe de ces exemples virtuels grâce au modèle appris
5. Entraîner un **modèle surrogé** simple (e.g. fonction linéaire) sur les exemples virtuels étiquetés
6. **Interpréter** le modèle surrogé

LIME

- Remarque :
 - Le **choix** de la fonction de **distance** pour sélectionner les exemples proches est **crucial**
 - [Ribeiro et al.] utilisent une fonction noyau exponentielle avec un choix particulier de largeur (width)
 - $0.75 * \text{la racine carré du nombre de descripteurs}$
- LIME est **implémenté** en **Python** (bibliothèque **lime** et bibliothèque **PiML**)
et en **R** (package **lime** et package **iml**)

Expliquer des prédictions **individuelles** : à propos d'un exemple

1. LIME

2. Shapley values

La méthode des valeurs de Shapley

- Inspiration de la théorie des jeux
 - Chaque **descripteur** est un « *joueur* » dans un jeu de prédiction
 - La **prédiction** est le « *gain* »
 - La méthode des valeurs de Shapley vise à **rétribuer équitablement** chaque joueur.
Les joueurs **coopèrent** dans une coalition pour l'obtention du gain
- La méthode des valeurs de Shapley s'applique pour les **modèles linéaires ou non linéaires**

Illustration

- Soient **trois élèves** concourant pour **un challenge** Kaggle
 - E1 : apporte l'**expertise** IODAA
 - E2 : apporte une **machine** très puissante
 - E3 : est un **geek**
- Les trois élèves gagnent le 1^{er} prix = 10 000 \$
 - 2^{ème} prix : 7500 \$
 - 3^{ème} prix : 5000 \$
- Comment **partager équitablement** le gain ?

- Si on rejouait :
 - E1 seul gagnerait le 2^{ème} prix : 7500 \$
 - E2 seul gagnerait le 3^{ème} prix : 5000 \$
 - E3 seul gagnerait le 4^{ème} prix : 2500 \$
 - E1 & E2 (E12) gagneraient le 1^{er} prix : 10 000 \$
 - E1 & E3 (E13) gagneraient le 2^{ème} prix : 7500 \$
 - E2 & E3 (E23) gagneraient le 3^{ème} prix : 5000 \$

- Si on rejouait :
 - E1 seul gagnerait le 2^{ème} prix : 7500 \$
 - E2 seul gagnerait le 3^{ème} prix : 5000 \$
 - E3 seul gagnerait le 4^{ème} prix : 2500 \$
 - E1 & E2 (E12) gagneraient le 1^{er} prix : 10 000 \$
 - E1 & E3 (E13) gagneraient le 2^{ème} prix : 7500 \$
 - E2 & E3 (E23) gagneraient le 3^{ème} prix : 5000 \$
- Considérons tous les **ordres** d'apport de contribution :
 - E1 -> E2 -> E3 : 7500 -> 2500 -> 0
 - E1 -> E3 -> E2 : 7500 -> 0 -> 2500
 - E2 -> E1 -> E3 : 5000 -> 5000 -> 0
 - E2 -> E3 -> E1 : 5000 -> 0 -> 5000
 - E3 -> E1 -> E2 : 2500 -> 5000 -> 0
 - E3 -> E2 -> E1 : 2500 -> 2500 -> 5000

- Considérons tous les **ordres** d'apport de contribution :

— E1 -> E2 -> E3	:	7500	->	2500	->	0
— E1 -> E3 -> E2	:	7500	->	0	->	2500
— E2 -> E1 -> E3	:	5000	->	5000	->	0
— E2 -> E3 -> E1	:	5000	->	0	->	5000
— E3 -> E1 -> E2	:	2500	->	5000	->	0
— E3 -> E2 -> E1	:	2500	->	2500	->	5000

- $\text{Val}(\text{E1}) = (7500 + 7500 + 5000 + 5000 + 5000 + 5000) / 6 = 7500$

- Considérons tous les **ordres** d'apport de contribution :

— E1 -> E2 -> E3	: 7500 -> 2500 -> 0
— E1 -> E3 -> E2	: 7500 -> 0 -> 2500
— E2 -> E1 -> E3	: 5000 -> 5000 -> 0
— E2 -> E3 -> E1	: 5000 -> 0 -> 5000
— E3 -> E1 -> E2	: 2500 -> 5000 -> 2500
— E3 -> E2 -> E1	: 2500 -> 2500 -> 5000

- $\text{Val}(\text{E1}) = (7500 + 7500 + 5000 + 5000 + 5000 + 5000) / 6 = 5833$
- $\text{Val}(\text{E2}) = (2500 + 2500 + 5000 + 5000 + 2500 + 2500) / 6 = 3333$
- $\text{Val}(\text{E3}) = (0 + 0 + 0 + 0 + 2500 + 2500) / 6 = 833$

Illustration

- Un système de régression a été entraîné pour **prédire le salaire** d'une personne en fonction de :
 - x_1 : son âge $\in \{18, \dots, 60\}$ (selon une distribution supposée uniforme)
 - x_2 : l'obtention d'un diplôme ou non $\in \{0, 1\}$ (avec probabilités égales)
- Soit le modèle appris : $h(x_1, x_2) = 200 x_1 + 1000 x_2$
- Le salaire prédit pour une personne (**âge = 20** et **diplôme = 1**) est de **5000 €** $= 200 * 20 + 1000 * 1$
- Quelle est la valeur de Shapley de x_2 (diplôme) ?

- Soit le modèle appris : $h(x_1, x_2) = 200 x_1 + 1000 x_2$
- Le salaire prédit pour une personne donnée ($\hat{\text{âge}} = 20$ et $\text{diplôme} = 1$) est de $5000 \text{ €} = 200 * 20 + 1000 * 1$

Quelle est la valeur de Shapley de x_2 (diplôme = 1) ?

Quelle serait la prédiction si on ne connaissait pas x_2 ?

Calcul de la contribution marginale de x_2 (diplôme) à la coalition de « âge » (x_1)

$$\begin{aligned} Val_{\mathbf{x}}(x_1) &= \int h(20, x_2) dP_{x_2} = \sum_{i=0}^1 h(20, i) P(x_2 = i) \\ &= 4500 \end{aligned}$$

$$Val_{\mathbf{x}}(x_2) = Val_{\mathbf{x}}(x_1, x_2) - Val_{\mathbf{x}}(x_1) = 5000 - 4500 = 500$$

Illustration

- Un système de régression a été entraîné pour prédire le **prix des appartements** en fonction de :
 - La **surface**
 - L'**étage**
 - La **proximité** ou non d'un **parc**
 - L'interdiction de la **présence** d'un **chat** ou non
- Soit l'exemple : `area = 50 ; floor = 2 ; park = nearby ; cat = banned`
- Le prix prédit est **300 k€** alors que le prix moyen est **310 k€**
- L'objectif est d'expliquer la différence, soit **-10 k€**
 - Une explication pourrait être : `area = 50` a contribué pour **30 k€**,
`floor = 2` pour **10 k€**, `park = nearby` pour **0 €**, `cat = banned` pour **-50 k€**.
Soit **-10 k€** au total.

- Comment calculer la contribution d'un descripteur pour un exemple donné ?
 - La valeur de Shapley pour un exemple donné (une valeur d'un attribut j) est :

$$\phi_j(val) = \sum_{S \subseteq \{1, \dots, d\} \setminus \{j\}} \frac{|S|! (d - |S| - 1)!}{d!} (val(S \cup \{j\}) - val(S))$$

où

$$val_{\mathbf{x}}(S) = \int \hat{h}(x_1, \dots, x_d) dP_{X_C} - \mathbb{E}[\hat{h}(\mathbf{X})]$$

S : sous-ensemble des attributs utilisés dans le modèle

d : est le nombre total d'attributs

$\mathbf{x}$: est l'exemple pour lequel on cherche les contributions des valeurs des attributs

X_C sont tous les attributs non inclus dans la coalition S

- Comment calculer la contribution d'un descripteur pour un exemple donné ?
 - La valeur de Shapley pour un exemple donné (une valeur d'un attribut j) est :

$$\phi_j(val) = \sum_{S \subseteq \{1, \dots, d\} \setminus \{j\}} \frac{|S|! (d - |S| - 1)!}{d!} (val(S \cup \{j\}) - val(S))$$

où

$$val_{\mathbf{x}}(S) = \int \hat{h}(x_1, \dots, x_d) dP_{X_C} - \mathbb{E}[\hat{h}(\mathbf{X})]$$

S : sous-ensemble des attributs utilisés dans le modèle

d : est le nombre total d'attributs

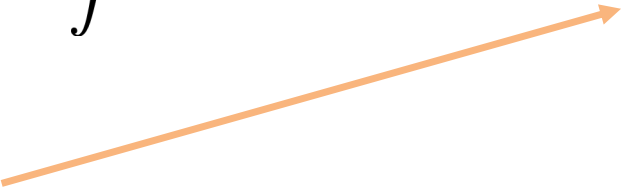
$\mathbf{x}$: est l'exemple pour lequel on cherche les contributions des valeurs des attributs

X_C sont tous les attributs non inclus dans la coalition S

$$val_{\mathbf{x}}(S) = val_{\mathbf{x}}(\{1, 3\}) = \int \hat{h}(x_1, X_2, x_3, X_4) dP_{X_2 X_4} - \mathbb{E}[\hat{h}(\mathbf{X})]$$

Deux questions

1. Comment faire quand on ne connaît pas les distributions de probabilité des valeurs des attributs ?
2. Comment faire face à la croissance exponentielle du nombre de coalitions avec le nombre de descripteurs ?

$$val_{\mathbf{x}}(S) = val_{\mathbf{x}}(\{1, 3\}) = \int \hat{h}(x_1, X_2, x_3, X_4) dP_{X_2 X_4} - \mathbb{E}[\hat{h}(\mathbf{X})]$$


- Pour estimer cette distribution, on **tire aléatoirement** des valeurs des variables X_2 et X_4 (dans cet exemple) qui sont **présentes dans l'ensemble d'apprentissage** (idéalement toutes les valeurs)

Approximate Shapley estimation for a single feature value:

- Output: Shapley value for the value of the j -th feature
- Required: Number of iterations M , instance of interest $\mathbf{x}$, feature index j , data matrix $\mathbf{X}$, and machine learning model f
- For all $m = 1, \dots, M$:
 - Draw random instance $\mathbf{z}$ from the data matrix $\mathbf{X}$
 - Choose a random permutation $\mathbf{o}$ of the feature values
 - Order instance $\mathbf{x}$: $\mathbf{x}_{\mathbf{o}} = (x_{(1)}, \dots, x_{(j)}, \dots, x_{(p)})$
 - Order instance $\mathbf{z}$: $\mathbf{z}_{\mathbf{o}} = (z_{(1)}, \dots, z_{(j)}, \dots, z_{(p)})$
 - Construct two new instances
 - With j : $\mathbf{x}_{+j} = (x_{(1)}, \dots, x_{(j-1)}, x_{(j)}, z_{(j+1)}, \dots, z_{(p)})$
 - Without j : $\mathbf{x}_{-j} = (x_{(1)}, \dots, x_{(j-1)}, z_{(j)}, z_{(j+1)}, \dots, z_{(p)})$
 - Compute marginal contribution: $\phi_j^{(m)} = \hat{f}(\mathbf{x}_{+j}) - \hat{f}(\mathbf{x}_{-j})$
- Compute Shapley value as the average: $\phi_j(\mathbf{x}) = \frac{1}{M} \sum_{m=1}^M \phi_j^{(m)}$

From [Christoph Molnar « Interpretable Machine Learning », (2025)]

Remarques

- LIME fournit l'importance relative de chaque attribut
- Les valeurs de Shapley fournissent la contribution « exacte » de chaque valeur mesurée de chaque attribut
- ATTENTION
 - « SHAP is not a measure of 'how important a given feature is in the real word', it is simply 'how important a feature is to the model »

Gianluca Zuin

Plan

1. Motivation
2. Les familles de méthodes
3. Interprétabilité par conception (by design)
4. Interprétation post-hoc
 - a. Indépendante du modèle (model-agnostic)
 - b. Dépendante du modèle (model-specific)
5. Quel avenir pour l'IA explicable ?

Explications pour les réseaux de neurones en **vision**

- Méthodes à base de **gradient**

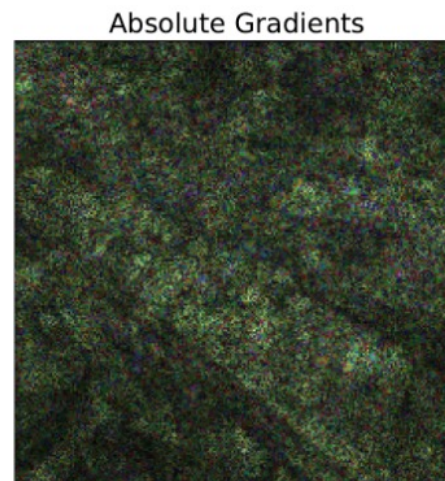
Cartes de saillance (*saliency maps*)

- Algorithme « **Vanilla gradient** »

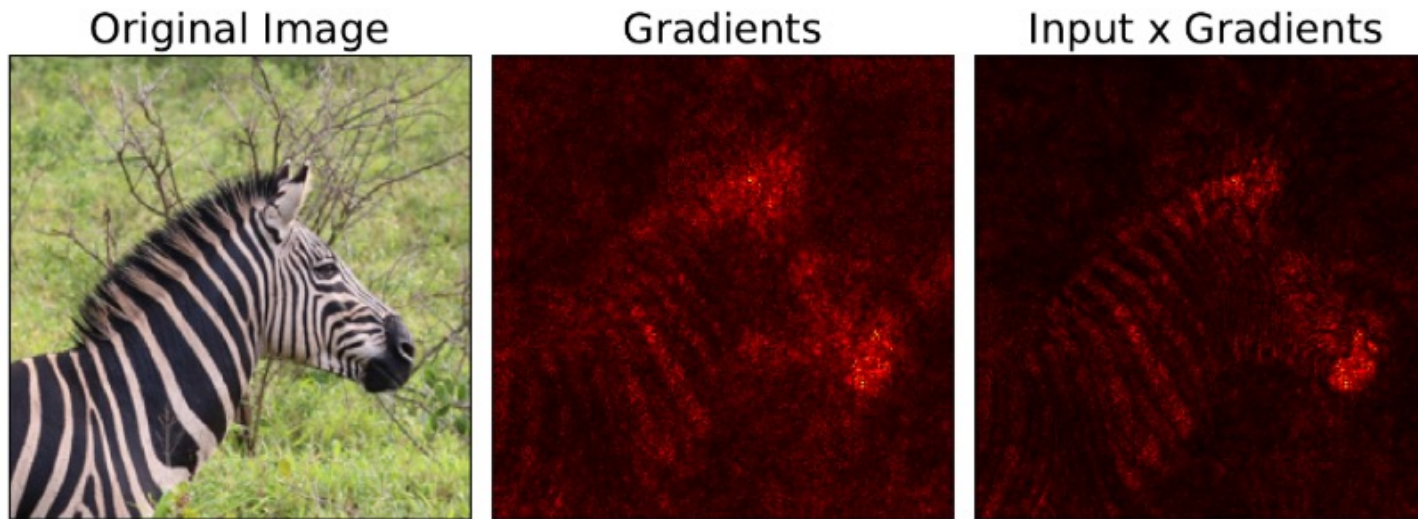
- Calcul du **gradient** d'une **sortie** logit par rapport à un **pixel** de l'image en

entrée $\frac{\partial y_c}{\partial X}$ (y_c : la classe de plus fort logit, celle qui est prédite)

- Typiquement un **gradient élevé** indique qu'un pixel est important pour la sortie



Input x Gradients



- Considérons un modèle linéaire simple $y = 100 x_1 + 5 x_2 + 1000$

On a $\frac{\partial y}{\partial x_1} = 100$ $\frac{\partial y}{\partial x_2} = 5$ x_2 semble moins important

- Mais soit un client de 30 ans et de salaire 50 000 €

$$x_1 \frac{\partial y}{\partial x_1} = 3000 \quad x_2 \frac{\partial y}{\partial x_2} = 250\,000$$

L'importance du salaire apparaît cette fois bien plus grande

- Problèmes

1. Dans une image, les **valeurs des pixels** n'ont pas de rapport linéaire direct avec la sortie

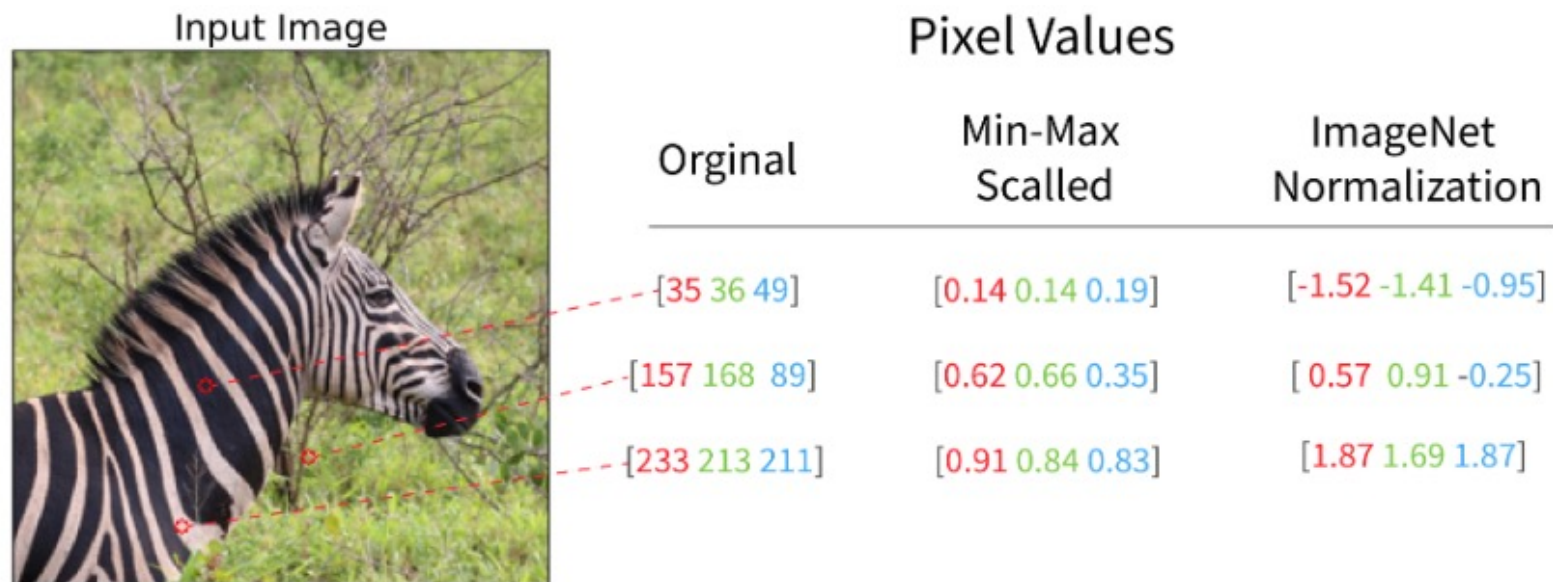


Figure 3: different encodings of pixel values. For min-max scaling dark pixels will tend to have lower values. For normalization, we are biased against more central values like those from the background.

2. La méthode met en évidence la **structure de l'image**, ce qui est plaisant au regard, mais **pas nécessairement relié** à ce qui est employé par le système pour réaliser ses prédictions

Guided backpropagation (GBP)

- Modifie l'algorithme standard de rétro-propagation en **ne passant que les gradients positifs**
 - Dans la rétropropagation standard, les gradients négatifs jouent un rôle crucial dans la réduction de l'activation des éléments qui **ne sont pas associés à la classe cible**.

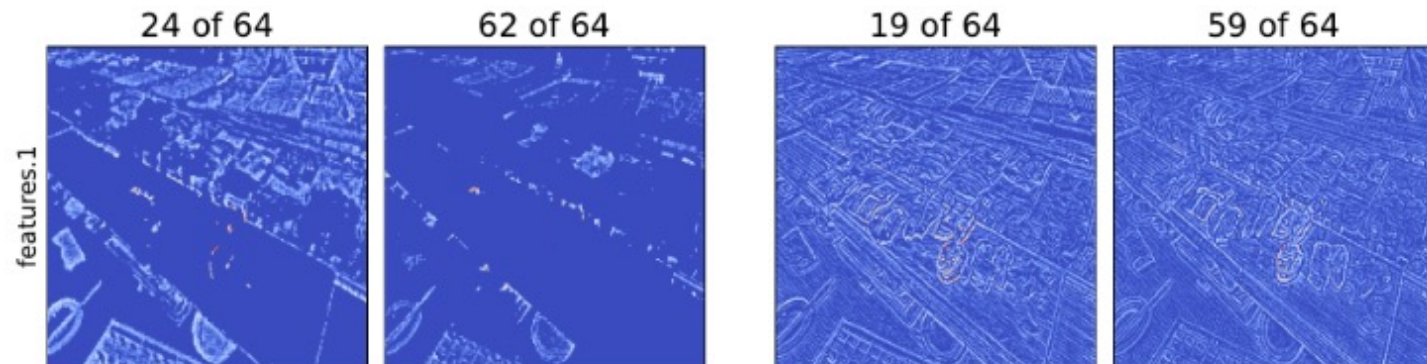
Dans le cas de la GBP, nous souhaitons **uniquement visualiser les éléments associés à la classe cible**.

En supprimant les gradients négatifs, nous pouvons réduire le bruit et obtenir des visualisations plus nettes.

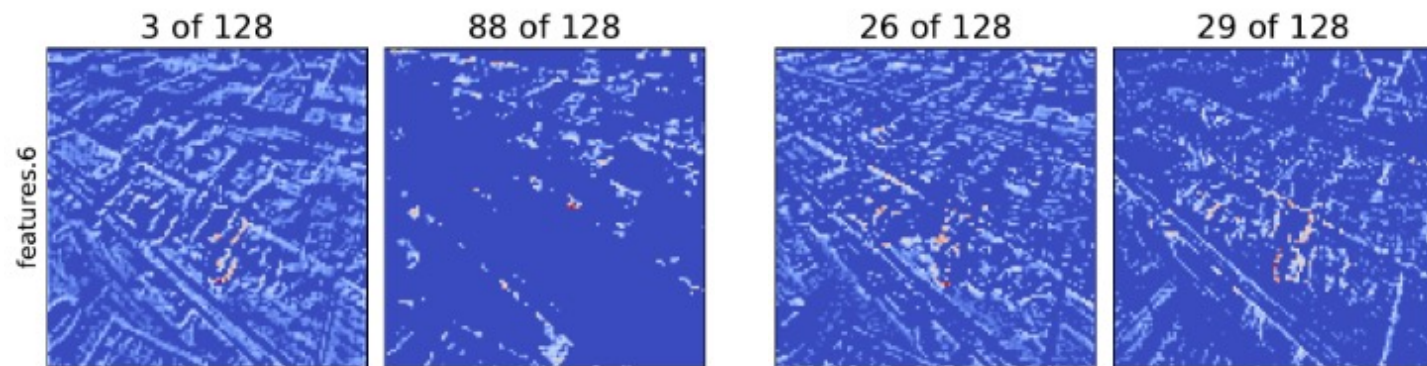


- Peut également servir à identifier ce qui est important dans les **couches intermédiaires**

1^{ère} couche
après
l'entrée



6^e couche
après
l'entrée



Exemples de **gradients** issus de cartes de caractéristiques intermédiaires.

- Le nom de la **couche** est indiqué sur l'**axe des y**.
- Le numéro de la **carte de caractéristiques** et le **nombre total de cartes de caractéristiques** dans la couche concernée sont indiqués au-dessus de chaque image.

Grad-CAM : Gradient-weighted Class Activation Mapping

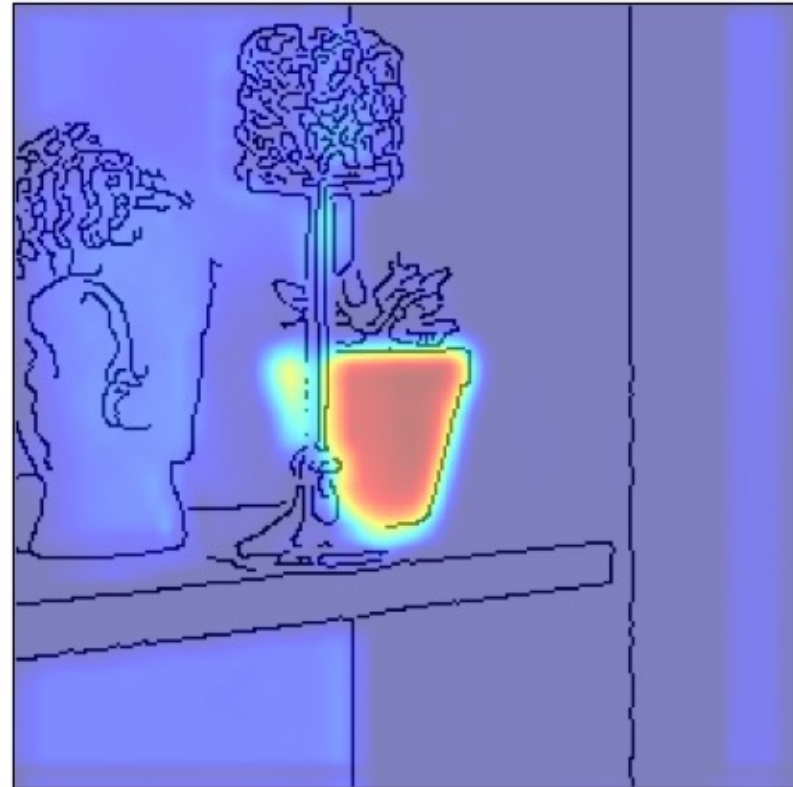
- La plupart des méthodes propagent le gradient jusqu'à l'image en entrée
- Grad-CAM s'arrête à la **couche de convolution finale**.
 - Moins précise que d'autres méthodes, Grad_CAM peut mettre en évidence des régions de l'image importantes pour une classification donnée
 - La sortie est une « **heat map** »
 - Grad-CAM est **rapide**, et peut être appliqué à tout réseau à convolution, ce qui en fait une méthode très prisée (malgré ses défauts)

- Ce qui paraît être un succès

Input



GradCAM Heatmap



1. Feature-based explanations methods

- Attribute the decision to **important features in the input space**



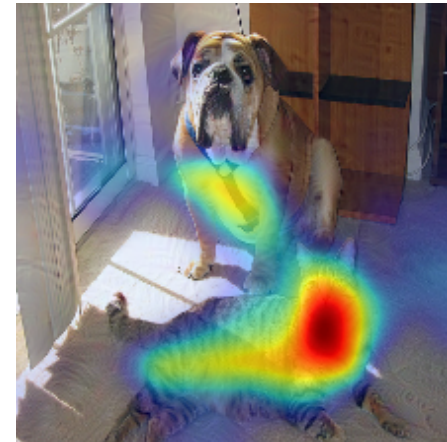
Boxer: 0.40 Tiger Cat: 0.18

(a) Original image



Boxer: $1.1e-20$

(c) Grad-CAM “Dog”



Tiger Cat: $6.5e-17$

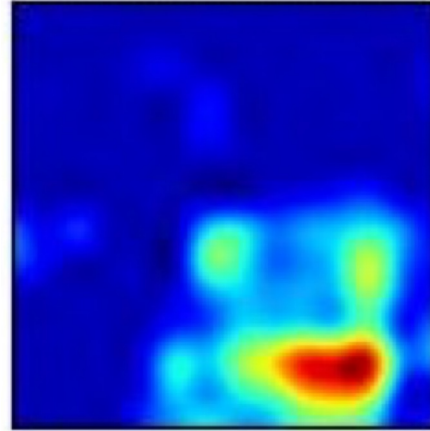
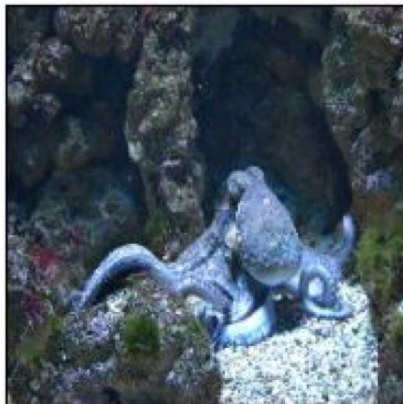
(d) Grad-CAM “Cat”

[Selvaraju et al. (2017) « *Grad-CAM: Visual explanations from deep networks via gradient-based localization* »]

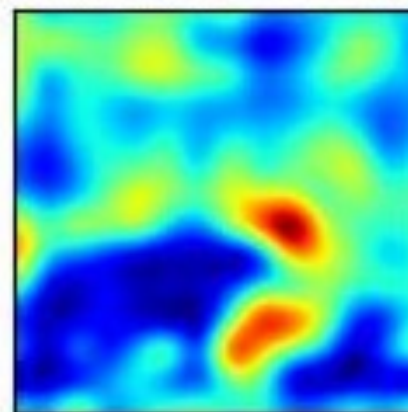
- Ce qui **ne** paraît **pas** être un succès



Classé comme un lévrier
(greyhound) à 35%



Classé incorrectement
comme une anguille (eel)
à 70%



- Limites

- La méthode GBP ne fait **pas de distinction entre les classes**.

- Par exemple, la **fourrure** peut constituer une caractéristique importante lors de la classification d'images de **chats**.

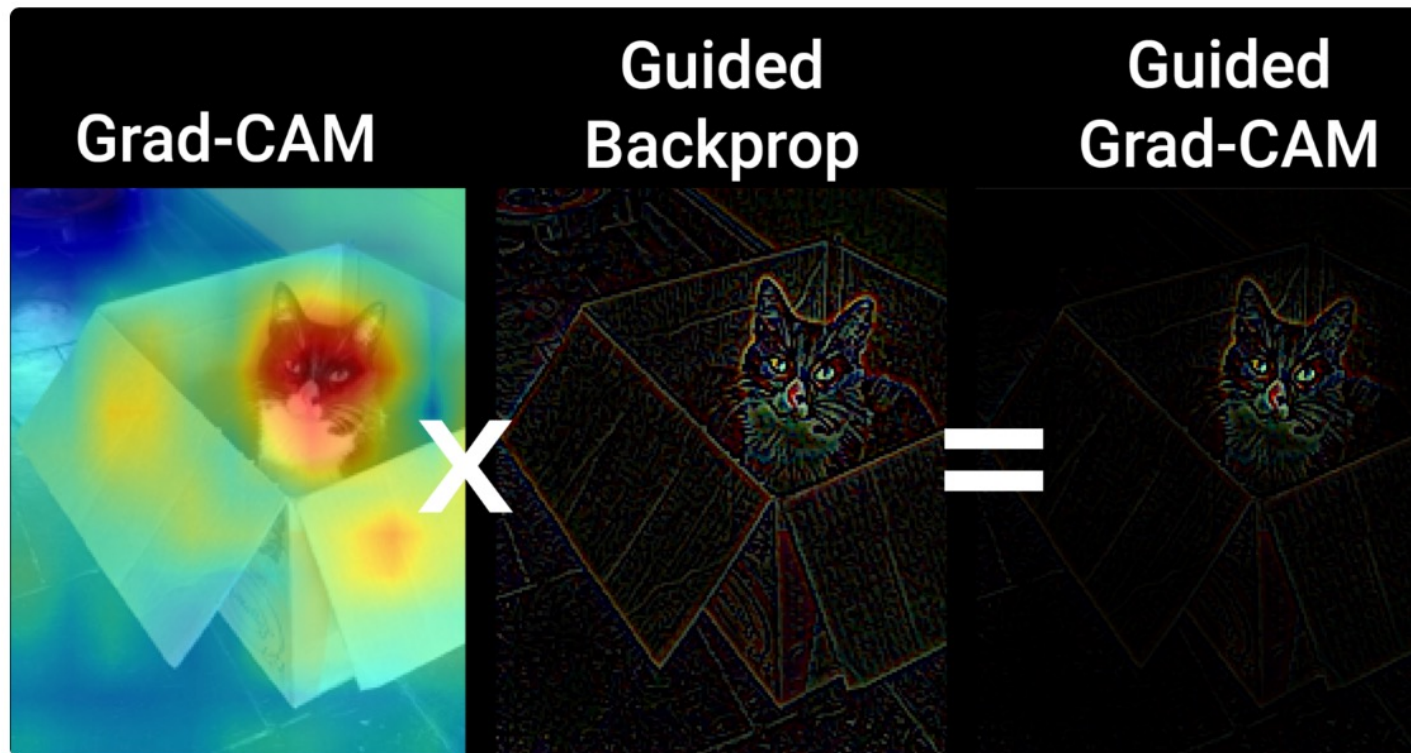
Cependant, si plusieurs animaux à fourrure sont présents sur l'image, la méthode GBP peut mettre en évidence toutes les zones de fourrure, **quel que soit l'animal auquel elles appartiennent**.

- En d'autres termes, la méthode GBP **met en évidence les caractéristiques qui contribuent à l'activation de la classe cible**, **mais pas exclusivement** celles qui sont spécifiques à cette classe.

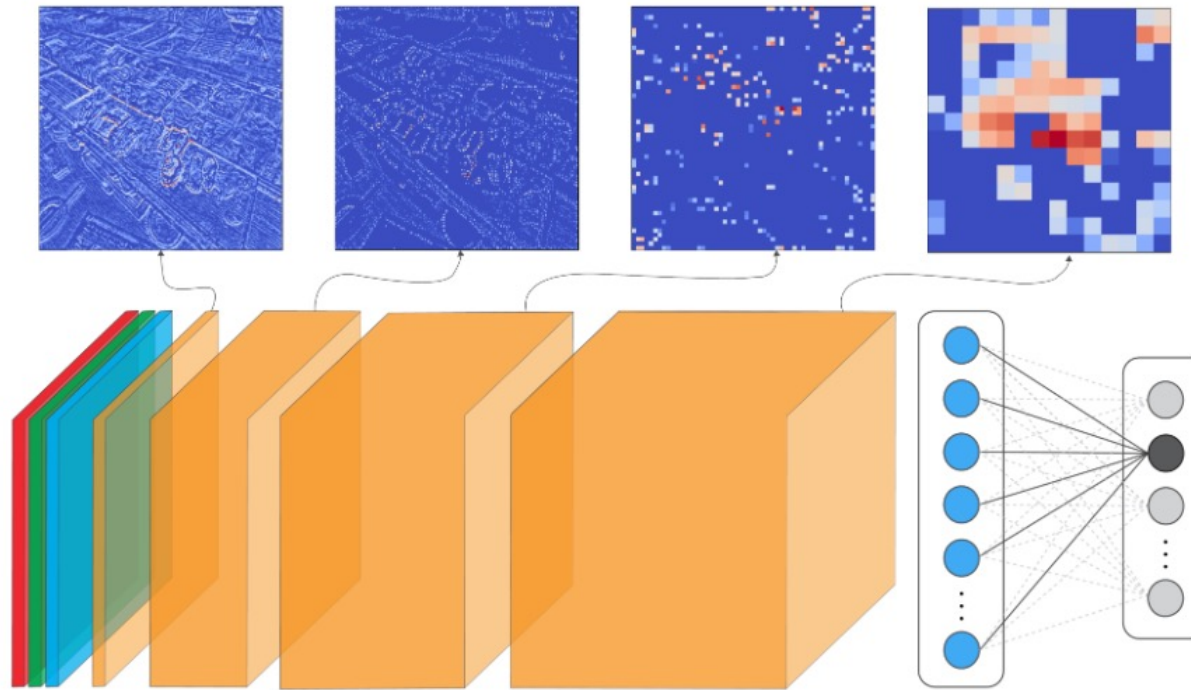
Guided Grad_CAM

- Guided Back-Propagation met en évidence des **détails**
- Grad-CAM discrimine les **classes**. Révèle les parties de l'image pertinentes pour prédire une classe donnée

Ces méthodes sont **complémentaires** --> Guided Grad-CAM



Retour sur GBP



Cartes de caractéristiques extraites à **différentes couches**. Les premières couches extraient des formes basiques qui sont ensuite combinées en des descripteurs plus abstraits

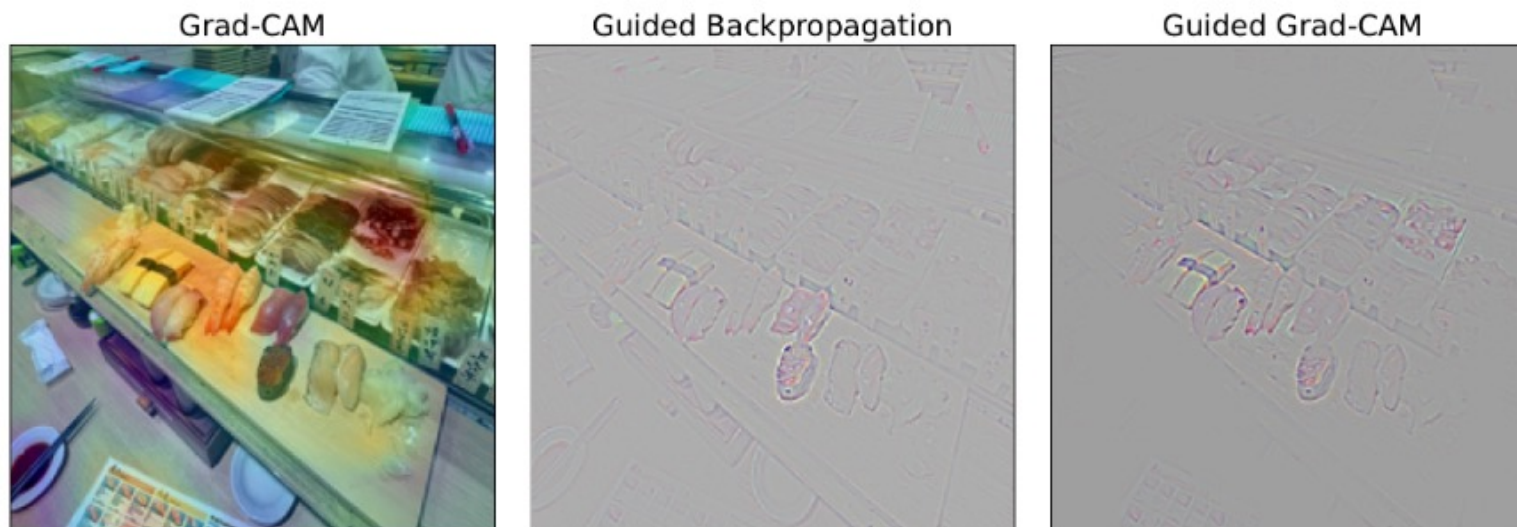
Le GBP, propage les gradients d'une couche qui ont **contribué positivement** à la couche suivante du réseau. Le **problème** est qu'une carte de caractéristiques d'une couche antérieure peut contenir des caractéristiques similaires, telles que des contours et des textures. On peut donc se retrouver dans une situation où un détecteur de contours contribuerait positivement à la carte de caractéristiques d'une couche plus profonde, **même si ce contour provient d'une région non pertinente de l'image**.

Grad-CAM

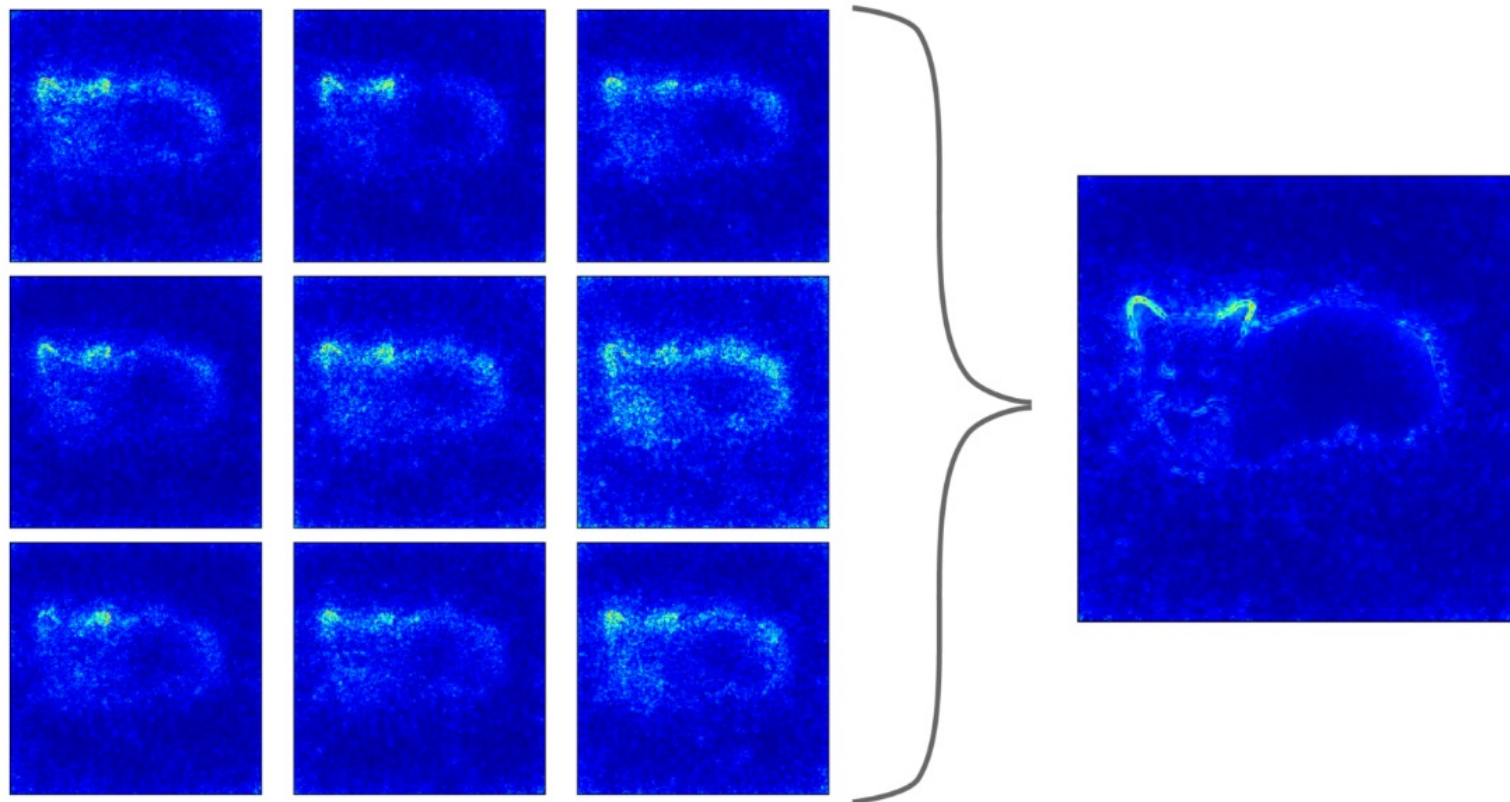
- Grad-CAM se concentre généralement sur la **dernière couche convolutive** du réseau qui est proche de la sortie et contient de nombreuses informations sémantiques et spatiales.

Par conséquent, les gradients seront les plus forts aux **emplacements spatiaux** des objets **ayant contribué positivement** au logit cible.

L'inconvénient est que ces couches profondes ont une **résolution plus faible**.



SmoothGrad: Adding noise to remove noise

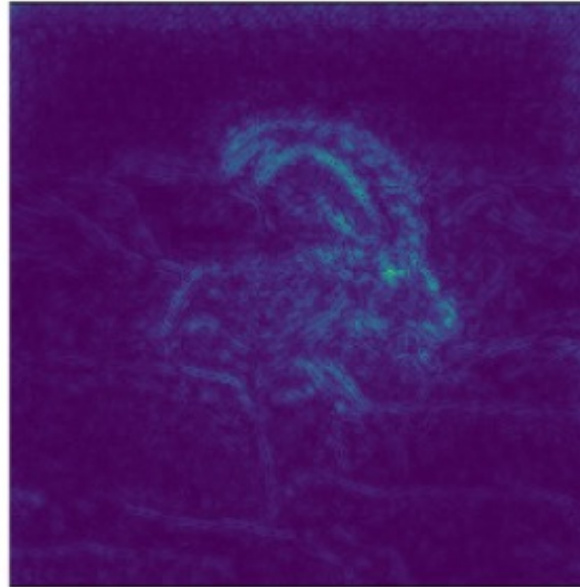


[Smilkov, Daniel, Thorat, Nikhil, Kim, Been, et al. (2017). *Smoothgrad: removing noise by adding noise*. arXiv preprint arXiv:1706.03825.]

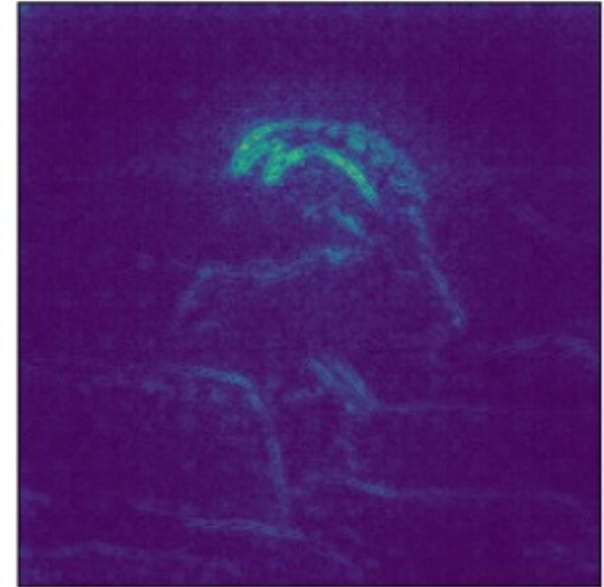
Input Image



GBP



SmoothGrad + GBP



From <https://adataodyssey.com/smoothgrad/>

Integrated gradient

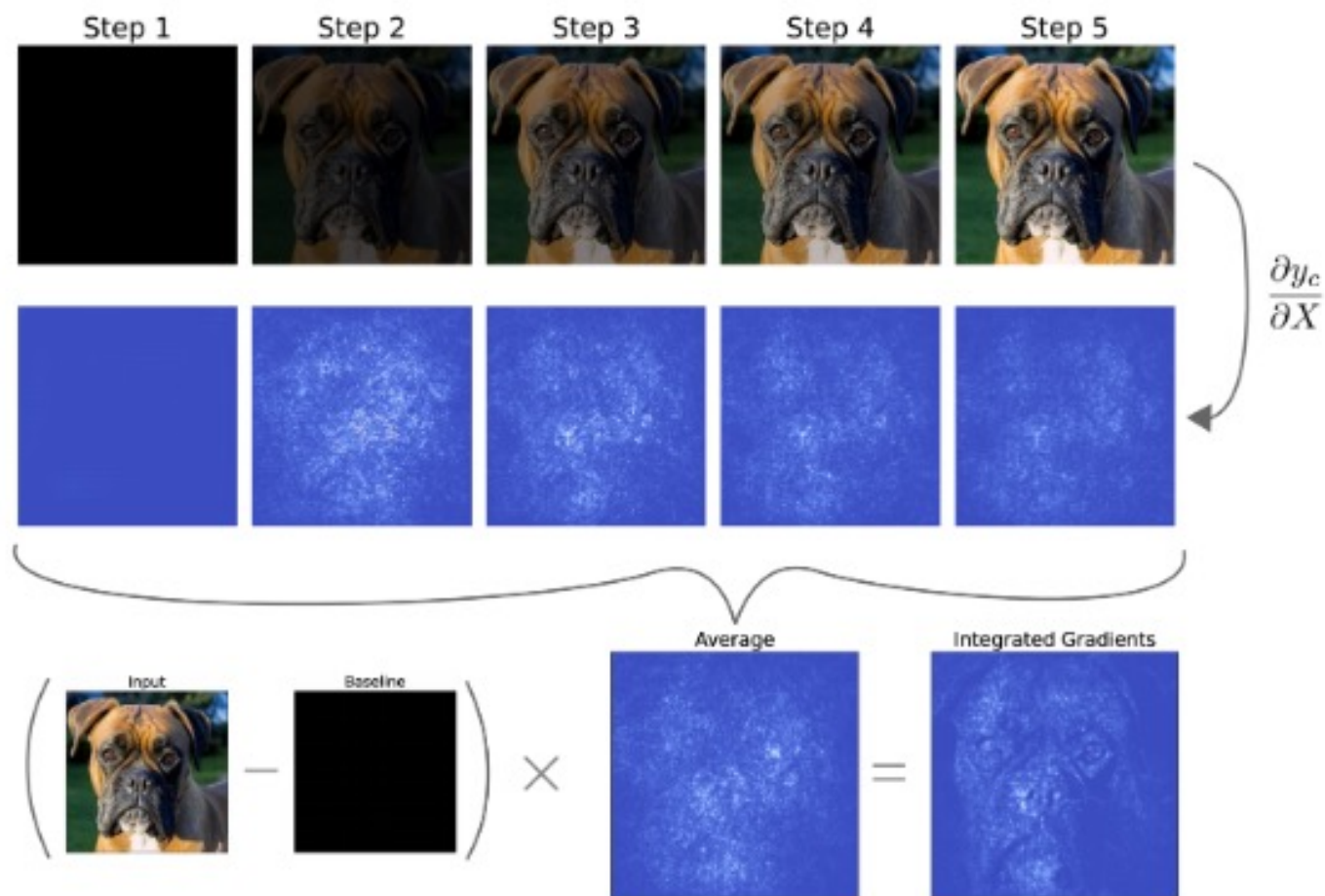


Figure 1: overview of integrated gradients. We calculate the average gradients at each step multiplied by the difference between the original input image and the baseline.

...

Limites générales des méthodes par gradient en vision

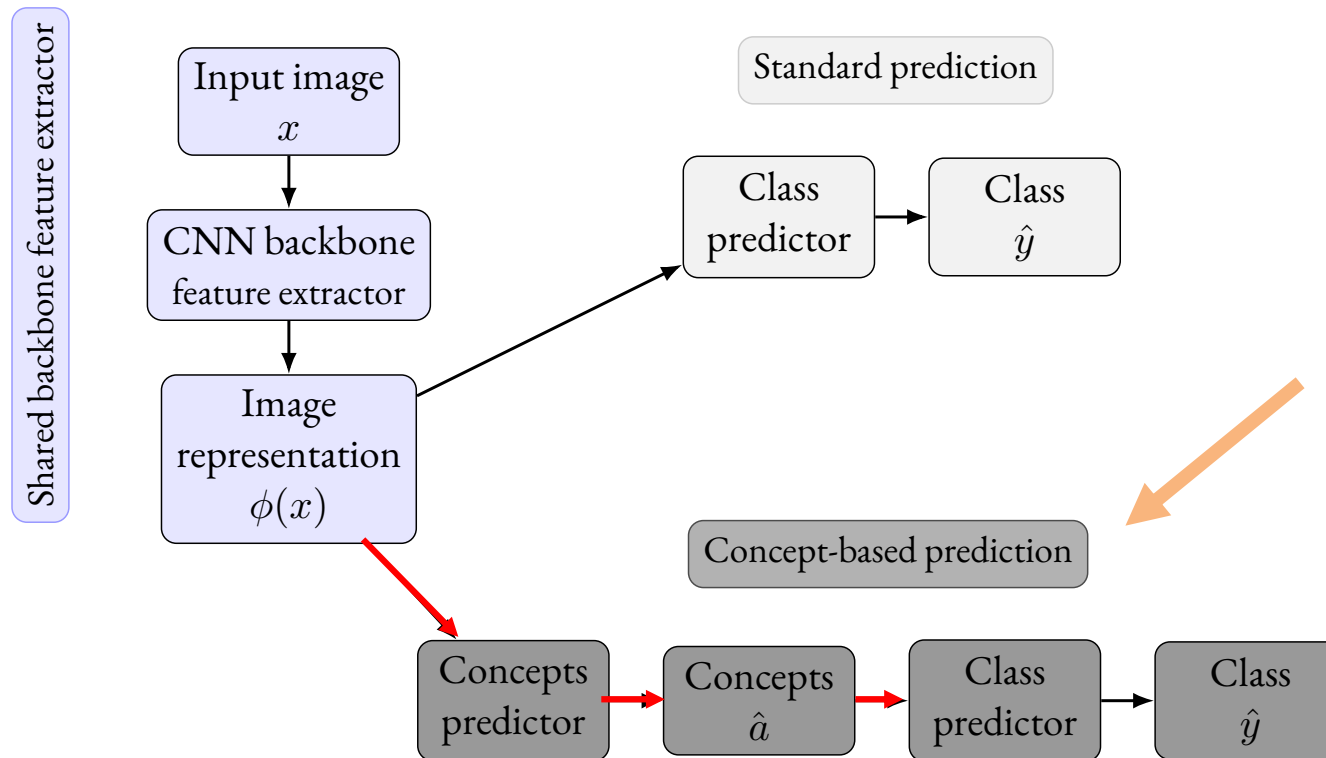
- Il est difficile de savoir si une « explication » est correcte : est bien causale dans la prédiction
- On peut croire qu'une explication est bonne parce qu'elle satisfait nos attentes
- Des petites variations des images peuvent conduire à des résultats très différents
 - E.g. En ajoutant une constante à tous les pixels [Kindermans et al., 2019]
 - Le neurone de biais corrige parfaitement le changement et les poids restent les mêmes
 - Pourtant les « explications » sont différentes !

Concept bottleneck models

- Les modèles de **goulot d'étranglement conceptuels** (CBM) sont des modèles d'apprentissage automatique conçus pour prendre des décisions via des **concepts intermédiaires interprétables** par l'homme plutôt que de passer directement des données brutes à une étiquette finale.

[KOH, Pang Wei, NGUYEN, Thao, TANG, Yew Siang, et al. *Concept bottleneck models*. In International conference on machine learning. Pmlr, 2020. p. 5338-5348.]

Concept bottleneck models



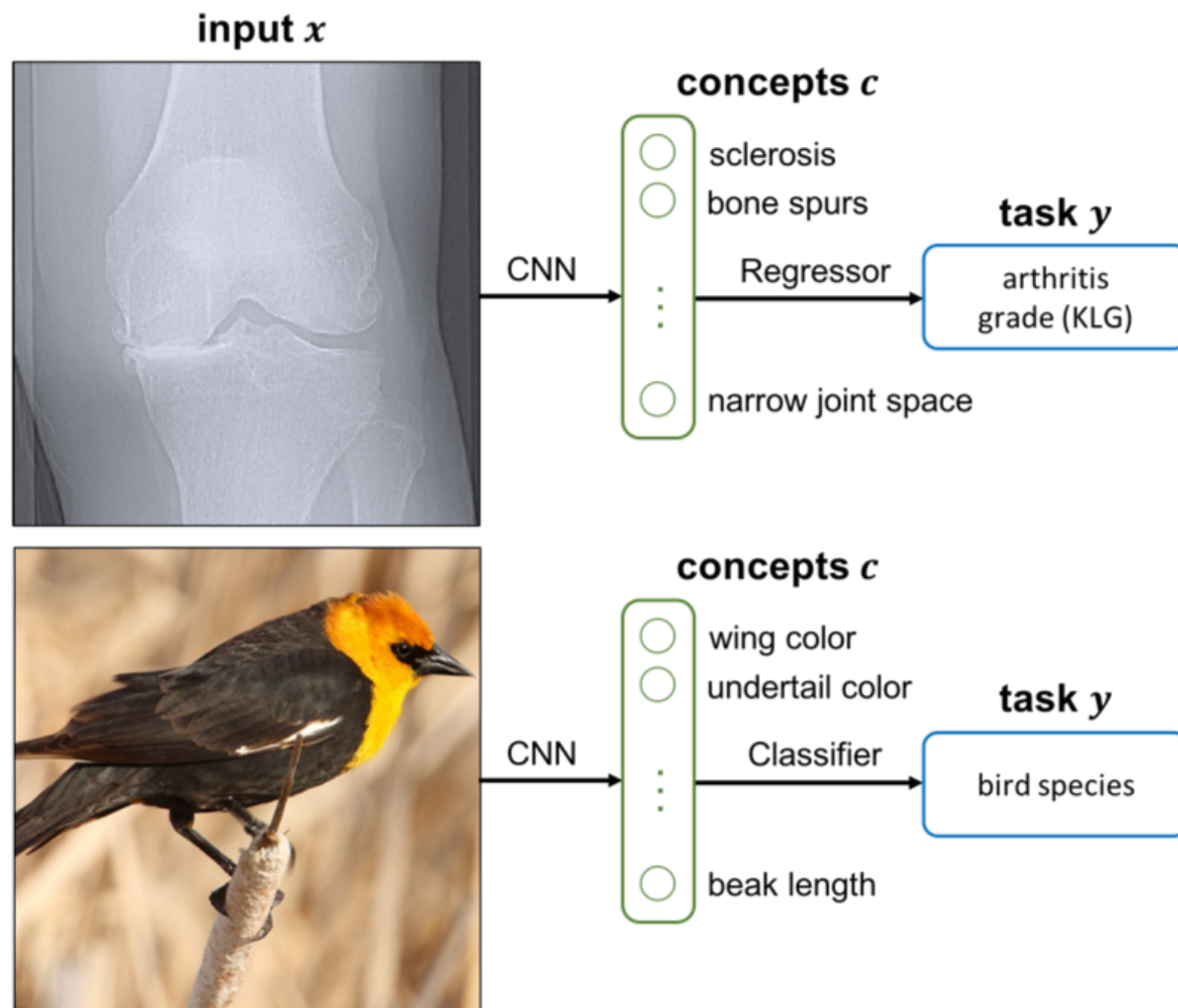
...

[Rim El Cheikh. Thèse « *Ad-Hoc Concept-Based Explainability for Deep Neural Networks* ». U. de Clermont-Ferrand, 2026]]

Concept bottleneck models

Ils fonctionnent en **deux étapes**

- 1. Encodeur de concepts ($g(x)$)** : associe des données brutes (telles que des pixels ou du texte) à un **ensemble de concepts compréhensibles** et définis par l'homme (c).
 - Par exemple, dans le domaine des **radiographies médicales**, ces concepts peuvent être « *excroissances osseuses* » ou « *sclérose* ».
Dans la classification des **oiseaux**, ils peuvent être « *couleur des ailes* » ou « *longueur du bec* ».
- 2. Prédicteur ($f(c)$)** : **utilise** uniquement **les valeurs conceptuelles** prédites pour générer la décision finale ou l'étiquette (y).



...

Concept bottleneck models

- Supposent disposer d'un **ensemble d'apprentissage** $\{(\mathbf{x}_i, y_i, c_i)\}_{1 \leq i \leq m}$ où les c_i sont des concepts identifiés pour chaque exemple
- Ils apprennent un modèle de la forme : $f(g(\mathbf{x}))$ avec

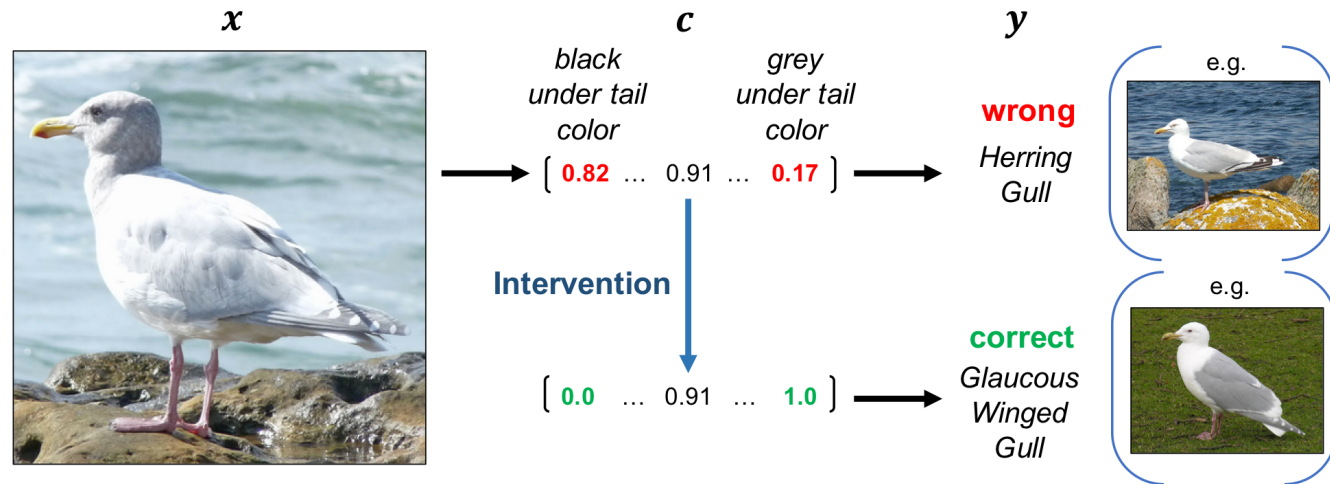
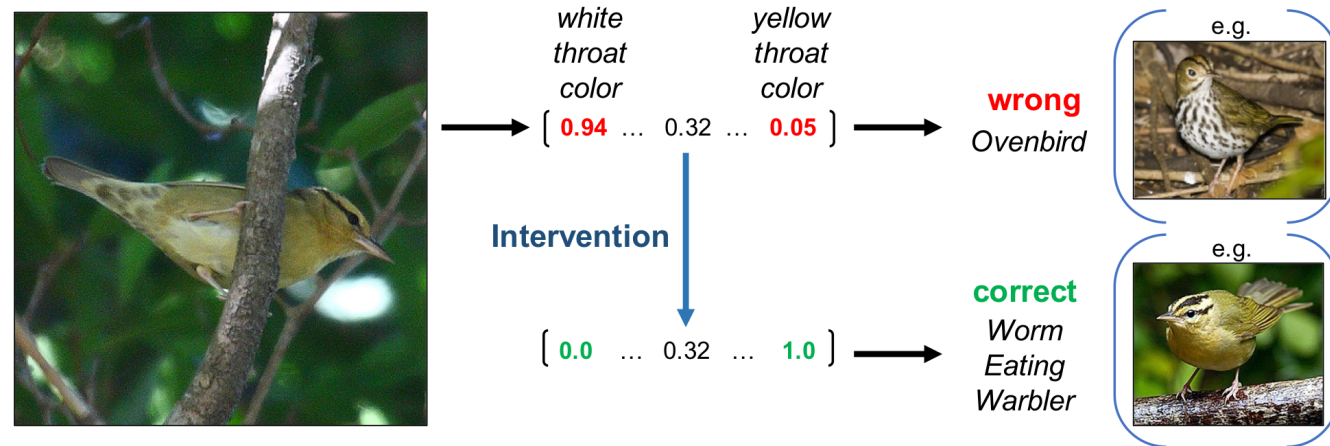
$$g : \mathbb{R}^d \rightarrow \mathbb{R}^k \quad \text{s'il y a } k \text{ concepts}$$

$$f : \mathbb{R}^k \rightarrow \mathbb{R}^n \quad \text{s'il y } n \text{ classes}$$

avec **deux fonctions de coût** dédiées

$$\hat{f} = \underset{f}{\text{ArgMin}} \sum_{i=1}^m \ell_Y(f(c_i), y_i)$$

$$\hat{g} = \underset{g}{\text{ArgMin}} \sum_{i=1, j=1}^{m, k} \ell_{C_j}(g_j(\mathbf{x}_i), c_j)$$



Exemples d'**interventions** pour corriger une classification grâce aux **concepts** mis en évidence par le système de prédiction

L'évaluation des systèmes explicatifs

1. Performance de **prédiction**
 - Pas détériorée
2. Performance de reconnaissance des **concepts intermédiaires**
3. **Stabilité** des poids entre concepts et classes
 - En fonction de la **répétition** des exécutions
et de la **variation** des conditions initiales
4. Est-ce que les **concepts importants** dans le codage
sont **importants pour le calcul** de la sortie ?
5. **Sensibilité** de la sortie aux **interventions** sur les poids des **concepts**
utilisés pour coder un exemple

Plan

1. Motivation
2. Les familles de méthodes
3. Interprétabilité par conception (by design)
4. Interprétation post-hoc
 - a. Indépendante du modèle (model-agnostic)
 - b. Dépendante du modèle (model-specific)
5. Quel avenir pour l'IA explicable ?

-
- Why is Machine Learning currently **lacking**?
 - The **exclusive focus on predictive performance** leads to an **incomplete learning problem formulation**
 - We want also
 - **Interpretability** of the **results**
 - **Interpretability** of the **process**
 - Gaining a **better understanding of the world** when including the learned decision function in an existing theory

-
- Why is Machine Learning currently **lacking**?
 - The **exclusive focus on predictive performance** leads to an **incomplete learning problem formulation**
 - We want also
 - **Interpretability** of the **results**
 - **Interpretability** of the **process**
 - Gaining a **better understanding of the world**
when including the learned decision function in an existing theory

Somehow, we have to **change**
the inductive criterion used in Machine Learning

Les explications **humaines** sont-elles **bonnes** ?

The “bystander effect”

- People were told that they would participate in a market study on games. **Participants** were individually welcomed at the door of the lab by a **friendly assistant** who took them to a room connected to her office, gave them a questionnaire to fill out, and went back to her office, where she could be heard shuffling papers, opening drawers, and so on.

A while later, the participants heard her climb on a chair and then heard a loud crash and a scream, *“Oh, my God, my foot ... I ... can’t move it. Oh ... my ankle! ... I can’t get this ... thing ... of me.”*

The “bystander effect”

- In **one condition**, the participant was **alone in the room** when all this happened.
- In **another condition**, **there was a man in the room** who acted as if he were a participant too.

This man **hardly reacted** to the crash and the scream.

He just shrugged and went on filling out the questionnaire.

The “bystander effect”

What do you expect ?

The “bystander effect”

What do you expect ?

- In the first condition, 70% of the participants intervened to help.
- In the second condition, only 7% did

The “bystander effect”

After all this the participants were **interviewed about their reactions.**

- **First condition**, typically: “I wasn’t quite sure what had happened; I thought I should at least find out.”
- **Second condition**, typically the participants thought that whatever had happened was not too serious

Not one ever mentioned the presence of the other participant!!

And even more, they **claimed this presence had no influence at all on their decision**

Lesson

- The reasons invoked **come after the fact** of deciding
- Indeed they are blatantly faulty
- Reasons are justifications

Pourquoi cela devient-il **important** pour les **machines** ?

Les **IA** sont de plus en plus des **sociétés d'agents**

- Qui **communiquent**
- Prennent des **décisions** en fonction de ces **interactions**

Il faut que chaque agent puisse **expliquer** les « **raisons** »
de ses sorties **aux autres agents**

An **accident** of an (autonomous) car

On May 7th, 2016, a Tesla car in autopilot mode collided with a semi-trailer across the road

- The analysis reveals that:
 - The **radar** did detect the semi-trailer but there were numerous road signs on the road with a radar "signature" similar to one of a car
 - The **camera** was unsure of its detections due to a dazzling milky sky (the semi-trailer was white).

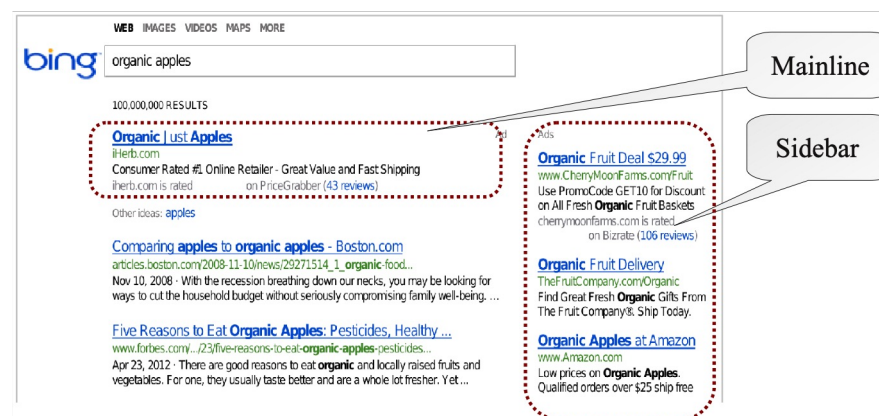
Combination of adaptive sub-systems and justifications

- The analysis reveals that:
 - The **radar** did detect a semi-trailer but ...
 - The **camera** was not sure of its detection of a semi-trailer because ...
- **What if the radar** had said: “Even though I am **not quite sure**, I do not believe it to be a semi-trailer **because** it looks to me like a road sign of which I have detected many on this road.”
- **What if the camera** had said: “Even though I am **not quite sure**, I do not believe it to be a semi-trailer **because** the color is close to the color of the sky which is milky today.”
- **What if** the radar and camera had a long history of such exchanges **and were influencing each other ...**

Interactions between learning modules

Adaptive advertising placement system

- Two sub-systems
 - One placing advertising links
 - The other one choosing the adds
- Mutually influencing each other
 - Each one is based on click data
 - Which also depends on the intervention of the other system
 - And other uncontrolled factors (price, user requests, ...)



[L. Bottou et al. «Counterfactual Reasoning and Learning Systems: The Example of Computational Advertising », JMLR, 14, (2013), 3207-3260]

Des questions

- **Interprétabilité** vs. **explicabilité** ?
- Comment **évaluer** une explication ?
- Explications et liens de **causalité** ?
- Une explication est-elle **symbolique** par nature ?
- Explication « **one-shot** » ou par **interactions** ?
- Est-ce que **expliquer** donne la **possibilité** de **tromper le système** ?
(puisque je sais comment la machine raisonne)
- Explications
 - pour les **experts IA**
 - pour les **experts du domaine**
 - pour les **utilisateurs**

Des questions (suite)

- Est-ce que les « **chains of thought** » sont des explications fiables ?
- Place de la **causalité** ?
 - Le graal ?

Ressources

- <https://adataodyssey.com/>
- Christoph Molnar « Interpretable Machine Learning » (3rd Ed. 2025)
- RUDIN, Cynthia. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 2019, vol. 1, no 5, p. 206-215. (<https://arxiv.org/abs/1811.10154>)