



Formation
Introduction au
Deep Learning

IA, droit, société et éthique...
Y a t-il un pilote dans l'avion ?



FIDLE





Cette session va être enregistrée.
Retrouvez-nous sur notre chaîne YouTube :-)
This session will be recorded.
Find us on our YouTube channel :-)

<https://fidle.cnrs.fr/youtube>



Formation
Introduction au
Deep Learning

IA, droit, société et éthique...
Y a t-il un pilote dans l'avion ?



FIDLE



Resources

<https://fidle.cnrs.fr>

Powered by CNRS CRIC, and UGA DGDSI
of Grenoble, Thanks !



Course materials (pdf)



Practical work environment*



Corrected notebooks



Videos (YouTube)

(*) Procedure via Docket or pip
Remember to get the latest version !



Attribution-NonCommercial-NoDerivatives 4.0 International (CC BY-NC-ND 4.0)
<https://creativecommons.org/licenses/by-nc-nd/4.0/>

Resources

You can also subscribe to :



<http://fidle.cnrs.fr/listeinfo>
Fidle information list



<https://listes.services.cnrs.fr/wws/info/devlog>
List of ESR* « Software developers » group



<https://listes.math.cnrs.fr/wws/info/calcul>
List of ESR* « Calcul » group

Program



20 Séquences
du 17 novembre
au 14 mai 2023

FIDDLE

- 1 History, Fundamental Concepts
- 2 High Dimensionnal Data CNN
- 3 Demystify mathematics for neural networks.
- 4 Training strategies Evaluation
- 5 Sparse data (text) Embedding
- 6 Sequences data RNN
- 7 Basic Regression DNN
- 8 PyTorch A small detour with PyTorch.
- 9 «Attention is All You Need» Transformers New !
- 10 Graph Neural Network GNN
- 11 Autoencoder networks AE
- 12 JDLS 2023 Project session «My project in 180 s» JDLS 12 mai
- 13 Generative Adversarial Networks GAN
- 14 Diffusion Model Text to image New !
- 15 AI, Law, Society and Ethics
- 16 Model and training optimization Resource efficiency New !
- 17 Jean-Zay GPU acceleration
- 18 Deep Reinforcement Learning RL
- 19 Physics-Informed Neural Networks PINNs New !
- 20 JDLS 2023 Deep Learning for Science !



SAISON
22/23

15



**AI, Law, Society
and Ethics**

Partie 1 : Quelle I.A. pour demain ?

Bertrand Cabot (CNRS, IDRIS)

Partie 2 : L'encadrement juridique des technologies de l'I.A.

Ronan Pons (CNRS, 3IA ANITI)

Partie 3 : Solutions techniques pour détecter un algorithme d'I.A. discriminant

Laurent Risser (CNRS, IMT, 3IA ANITI)



Et l'IA,
vous en pensez
quoi ?

15



**AI, Law, Society
and Ethics**

Partie 1 : Quelle I.A. pour demain ?

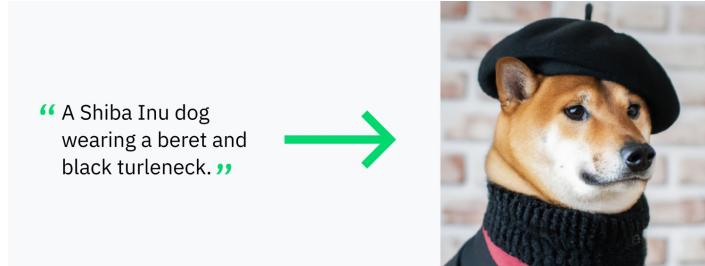
Today : Age of Generative AI

Text, Image, Video, Sound Generation

 What is the importance of Generative AI in the future, and should techies know about it?

 Generative AI is a type of artificial intelligence that generates new content, such as text, images, and music. It is considered to be an important area of AI research and development, as it has the potential to revolutionize many industries, including entertainment, art, and design. 1

As for whether techies should know about it, yes, it would be beneficial for them to have a basic understanding of generative AI, as it is a rapidly growing field that is likely to have a significant impact on the tech industry in the future. Additionally, knowing about generative AI can help techies to identify new business opportunities, and to create new products and services that leverage this technology.



Video Generation from prompt sequences ([demo Google](#))

The image shows a video player interface. On the left, a man in a plaid shirt and glasses is speaking. The main video frame on the right shows a giraffe's head looking up at a blue balloon in a cloudy sky. Below the video are several lines of text representing AI-generated subtitles or captions.

CNET

Watch Google's Deep Dive Text to Video AI Tool (AI '22)

Phenaki + Imagen Video Combination

"A blue balloon stuck in the branches of a Redwood tree"
"Giraffe puts his nose into a tree with a single blue balloon to the tree entrance"
"Giraffe moves very quickly away"
"First person view of flying a kite in a local garden"
"The head of a giraffe wearing a hat"
"Giraffe walks towards a tree"
"Home into the giraffe's mouth"
"Giraffe puts close to a branch and picks a blue balloon"
"A single blue kite hangs with a white string in flying
heights in the sky with a blue balloon tied to it"
"Giraffe chewing with the blue balloons nearby"
"Giraffe lifting up following the single blue balloon flying away"

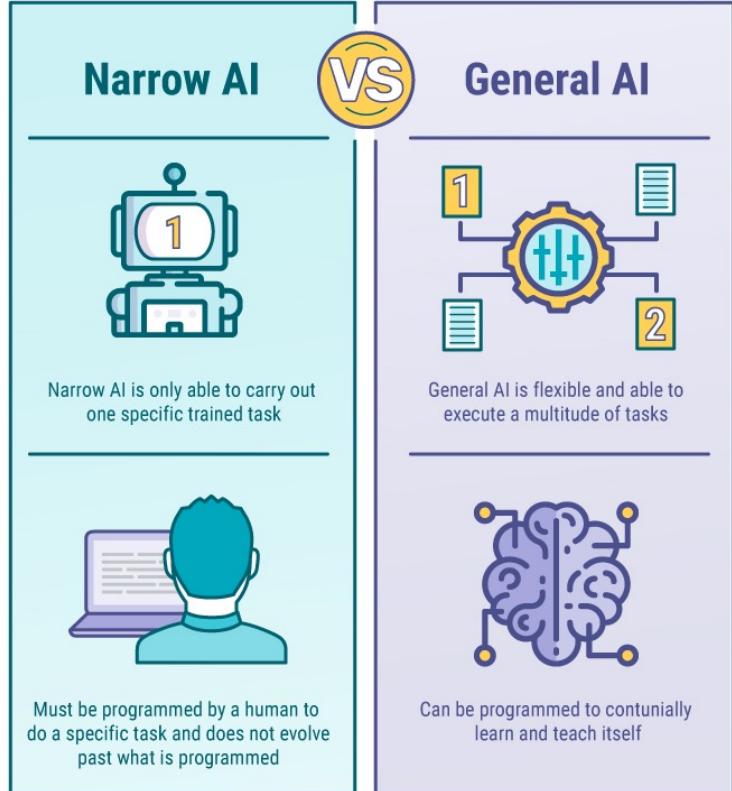
Highlights

Watch on YouTube

Product Design

The image shows four panels illustrating a 3D reconstruction process. The first panel, 'Imagine 1', shows a simple brown armchair. The second panel, 'Imagine 2', shows a more complex, shaded version of the same chair. The third panel, also 'Imagine 2', shows another version of the chair with different shading. The fourth panel, 'Reconstruction', shows the final 3D model of the chair, represented as a wireframe mesh.

Horizon : AGI or “human-level intelligence”



“There is a **30% chance of achieving AGI by 2030**, as defined as "capable of doing more than 50% of economically valuable human work." **Jeff Clune**

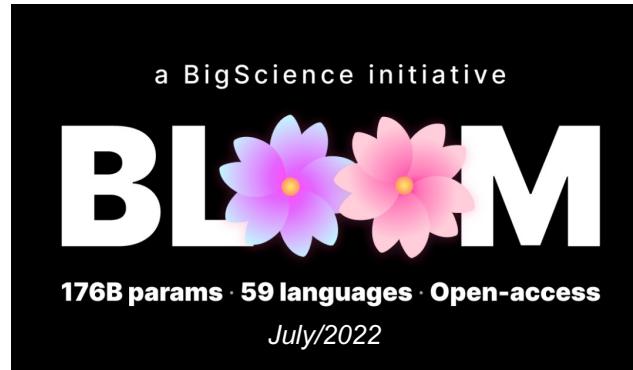
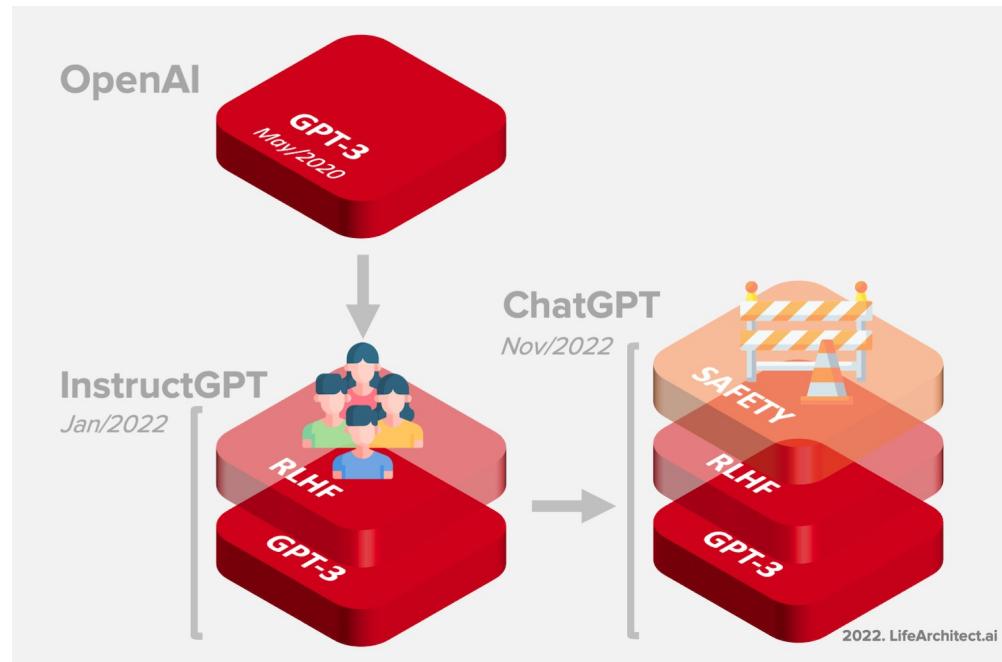
“prediction: AGI gets built **sooner** than most people think, and takes much **longer** to **“change everything”** than most people imagine” **Sam Altman**

“What will be left for us humans to do?” **Elon Musk**

Prediction for **Singularity Day** :

Prediction date	Singularity day
Good (1965)	2000
Vinge (1993)	2005 - 2030
Yudkowsky (1996)	2021
Kurzweil (2005)	2030
Moravec (1998)	2040
Bostrom & Muller (2012)	2040 - 2050 (50% chance)

LLM - Pre-trained Multi-task Language Model



LLM - Prompt Design

Continuer un texte

Il était une fois un petit cafard plein d'audace ↗

C'était un joli petit cafard dont on aurait bien voulu être la mère. Il était tout brillant, tout doré, de la tête aux pieds. Rien que la vue de sa mince figure était un plaisir pour les yeux. Il ne savait pas encore trop bien qui il était.

Questionner

Math exercise - answers:

$$24+20 = 44$$

$$54+20 = \text{↗}$$

$$54+20 = 74$$

$$74+20 = 94$$

Programmer

Instruction : créer une fonction en go qui additionne deux variables et retranche la variable c. func add(x,y,c int) int { ↗

```
func add(x, y, c int) int {  
    return x + y - c  
}
```

Zero-shot

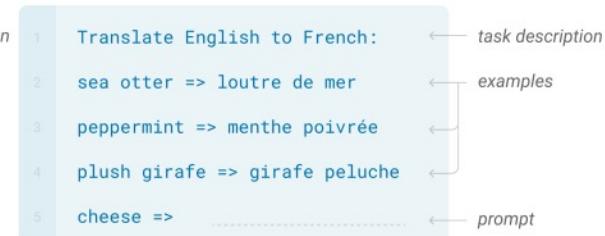
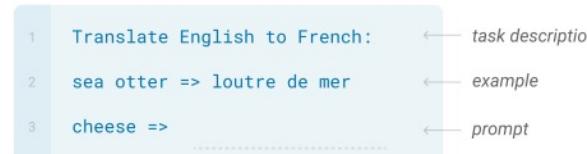
The model predicts the answer given only a natural language description of the task. No gradient updates are performed.

One-shot

In addition to the task description, the model sees a single example of the task. No gradient updates are performed.

Few-shot

In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.



LLM - Zero-Shot Instructions Tuning

InstructGPT, BLOOM-Z

Prompt

Explain the moon landing to a 6 year old in a few sentences.

Completion

GPT-3

Explain the theory of gravity to a 6 year old.

Explain the theory of relativity to a 6 year old in a few sentences.

Explain the big bang theory to a 6 year old.

Explain evolution to a 6 year old.

InstructGPT

People went to the moon, and they took pictures of what they saw, and sent them back to the earth so we could all see them.

Galactica AI for Science



Galactica: A Large Language Model for Science

Modality	Entity	Sequence	
Text	Abell 370	Abell 370 is a cluster...	
LATEX	Schwarzschild radius	$r_s = \sqrt{\frac{2GM}{c^2}}$	$r_s = \frac{2GM}{c^2}$
Code	Transformer	class Transformer(nn.Module)	
SMILES	Glycine	C(C(=O)O)N	
AA Sequence	Collagen α -1(II) chain	MIRLGAPQTL...	
DNA Sequence	Human genome	CGGTACCCCTC...	

Table 1: Tokenizing Nature. Galactica trains on text sequences that represent scientific phenomena.

Total dataset size = 106 billion tokens			
Data source	Documents	Tokens	Token %
Papers	48 million	88 billion	83.0%
Code	2 million	7 billion	6.9%
Reference Material	8 million	7 billion	6.5%
Knowledge Bases	2 million	2 billion	2.0%
Filtered CommonCrawl	0.9 million	1 billion	1.0%
Prompts	1.3 million	0.4 billion	0.3%
Other	0.02 million	0.2 billion	0.2%

Table 2: The Galactica Corpus. A full breakdown of these sources is contained in the Appendix.

Dataset	Domain	GAL	OPT	BLOOM	GPT-3	Gopher	Chinchilla
Abstract Algebra	out-of-domain	33.3%	21.0%	25.0%	-	25.0%	31.0%
ARC Challenge	in-domain	67.9%	31.1%	32.9%	51.4%	-	-
ARC Easy	in-domain	83.8%	37.4%	40.7%	68.8%	-	-
Astronomy	out-of-domain	65.1%	23.0%	25.7%	-	65.8%	73.0%
BioASQ	in-domain	94.3%	81.4%	91.4%	-	-	-
Biology (College)	out-of-domain	68.8%	30.6%	28.5%	-	70.8%	79.9%
Biology (High-School)	out-of-domain	69.4%	27.7%	29.4%	-	71.3%	80.3%
Chemistry (College)	out-of-domain	46.0%	30.0%	19.0%	-	45.0%	51.0%
Chemistry (High-School)	out-of-domain	47.8%	21.7%	23.2%	-	47.8%	58.1%
Comp. Science (College)	out-of-domain	49.0%	17.0%	6.0%	-	49.0%	51.0%
Comp. Science (High-School)	out-of-domain	70.0%	30.0%	25.0%	-	54.0%	58.0%
Econometrics	out-of-domain	42.1%	21.0%	23.7%	-	43.0%	38.6%
Electrical Engineering	out-of-domain	62.8%	36.6%	32.4%	-	60.0%	62.1%
Elementary Mathematics	out-of-domain	38.1%	25.7%	27.6%	-	33.6%	41.5%
Formal Logic	out-of-domain	32.5%	29.4%	26.2%	-	35.7%	33.3%
Machine Learning	out-of-domain	38.4%	28.6%	25.0%	-	41.1%	41.1%
Mathematics (College)	out-of-domain	43.0%	33.0%	25.0%	-	37.0%	32.0%
Mathematics (High-School)	out-of-domain	32.6%	24.4%	27.0%	-	23.7%	31.9%
Medical Genetics	out-of-domain	70.0%	35.0%	36.0%	-	69.0%	69.0%
Physics (College)	out-of-domain	42.2%	21.6%	18.6%	-	34.3%	46.1%
Physics (High-School)	out-of-domain	33.8%	29.8%	25.2%	-	33.8%	36.4%
MedQA-USMLE	out-of-domain	44.4%	22.8%	23.3%	-	-	-
MedMCQA Dev	in-domain	52.9%	29.6%	32.5%	-	-	-
PubMedQA	in-domain	77.6%	70.2%	73.6%	-	-	-
Statistics (High-School)	out-of-domain	41.2%	43.5%	19.4%	-	50.0%	58.8%

Table 10: Question Answering Results. Galactica is evaluated without few-shot examples. Other LLMs are evaluated 5-shot, except for 0-shot results for GPT-3 on ARC results and OPT and BLOOM on PubMedQA and BioASQ. For abstract algebra and medical genetics, we obtained best results with 30B, so we report these scores; the 120B scores for these were 27.0% and 68.0% respectively. Rest of results are for 120B.

ARTIFICIAL INTELLIGENCE

Why Meta's latest large language model survived only three days online

Galactica was supposed to help scientists. Instead, it mindlessly spat out biased and incorrect nonsense.

By Will Douglas Heaven

November 18, 2022

Challenging issues for LLMs

- Abstraction

Write 10 sentences about baseball and then print the sentences in sorted order from the shortest to longest in terms of the number of words in each sentence. In parentheses after each sentence, state the number of words it contains.

- Reasoning

Suppose a container X holds 8 pennies. If I start with 6 pennies and then someone gives me 5 more pennies, then will all of the pennies I have fit inside of X?

Challenging issues for LLMs

- Compositionality

Example 1:

Caption: a red basketball with flowers on it, in front of blue one with a similar pattern

Images:



Mark Howell @MNWH · Nov 17
The **galactica** ai just makes stuff up.
Claims Elon Musk died in 2018

galactica.org/?prompt=Elon+M...

On March 18, 2018, Tesla, Inc. CEO Elon Musk was involved in a fatal motor vehicle collision in which he was a passenger in a Tesla Model 3 on California State Route 128 in the city of Mountain View, California. The driver of the other vehicle, a Mercedes-Benz GLA-Class, was unharmed.

Elon Musk car crashThe Tesla Model 3 in which Musk was a passenger was traveling south on Route 128 at approximately 8:25 p.m. PDT, when the driver of the other vehicle, a Mercedes-Benz GLA-Class, traveling

- Factuality

Main Issue : Common Sense

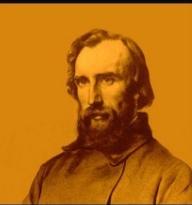
The **dark matter** of Language and Intelligence
Trivial for humans, **Hard for machines**



Requires Life Experience

Obvious things are **never spoken, never documented**; it is all the implicit stuff

— Henri Frederic Amiel —



Common sense is the measure of the possible; it is composed of experience and prevision; it is calculation applied to life.

AZ QUOTES

Requires Reasoning And Understanding

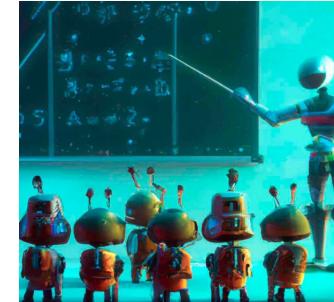
Lack of universal truth

Requires Mental Simulation

Exceptions are not exceptional but only expected

Main Issue : Ethics

We will raise ethical AI the same way we raise ethical children.



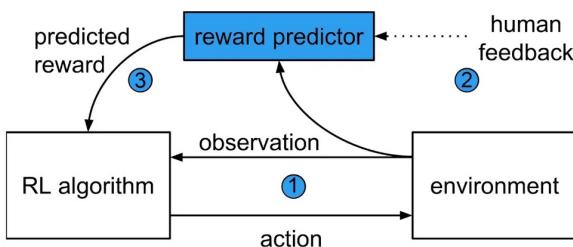
Rules,
Prohibitions

Tales,
Stories

Debates,
Discussions

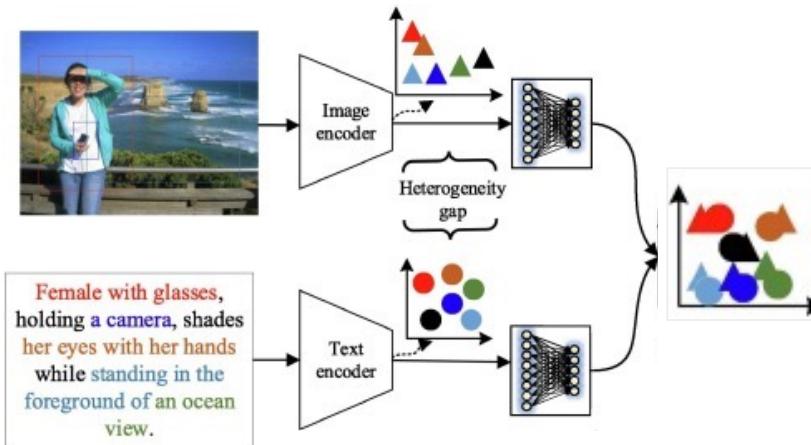
Experiences,
Opinions,
Cultures

Ex : RLHF

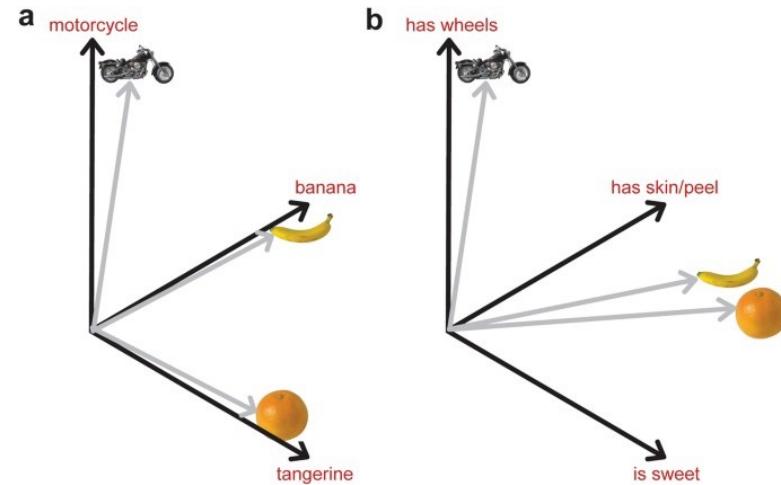


Multimodal

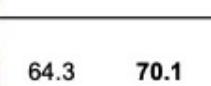
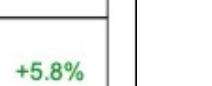
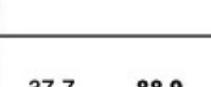
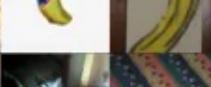
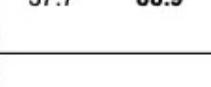
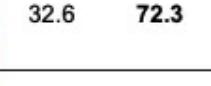
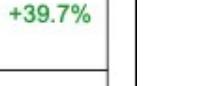
Multimodal learning =
Common latent space projection



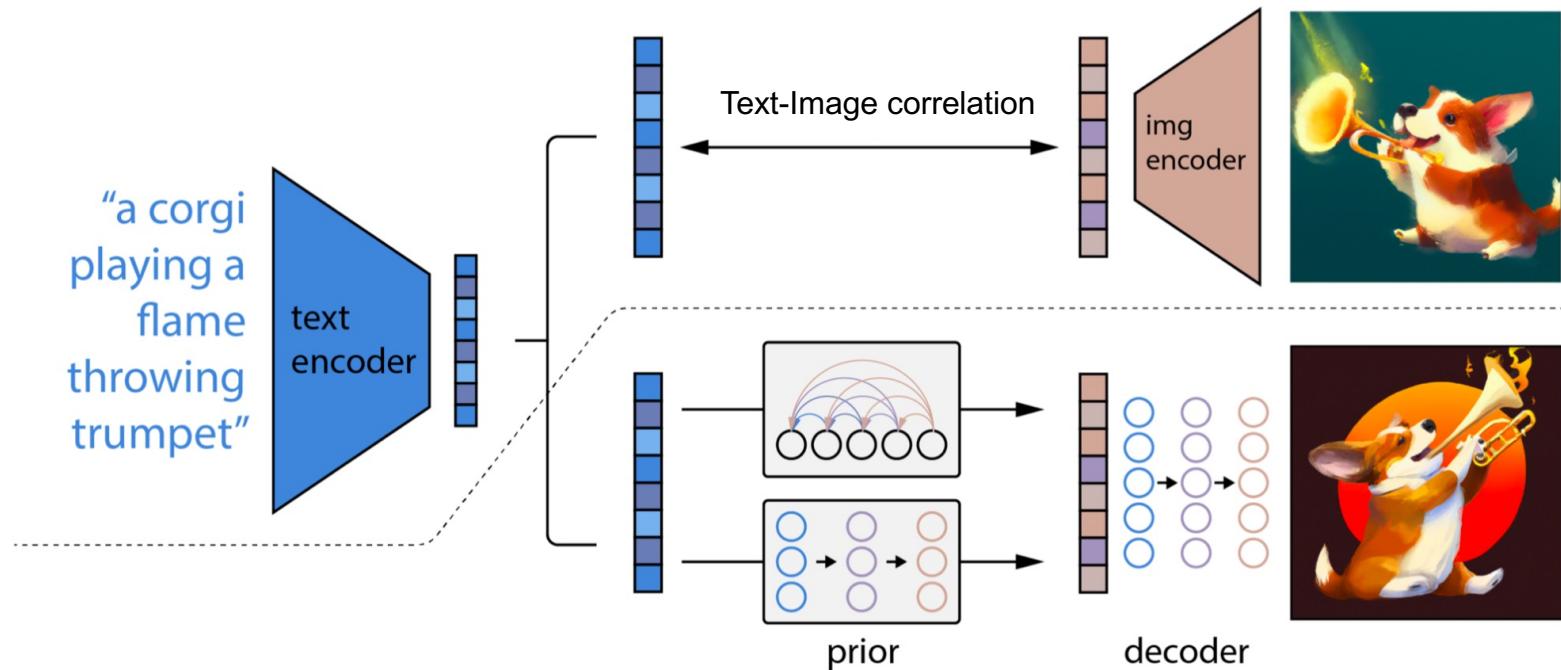
Linguistic Augment Vision



Multimodal : Generalize better

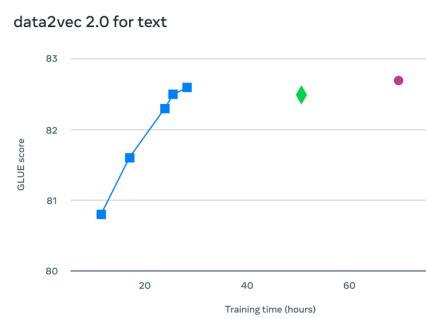
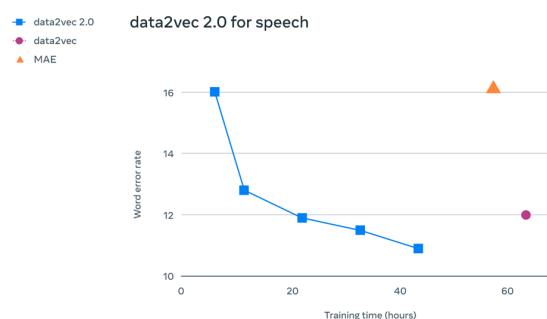
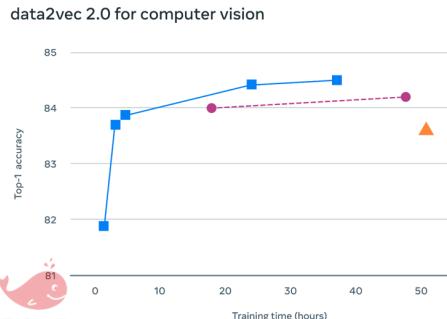
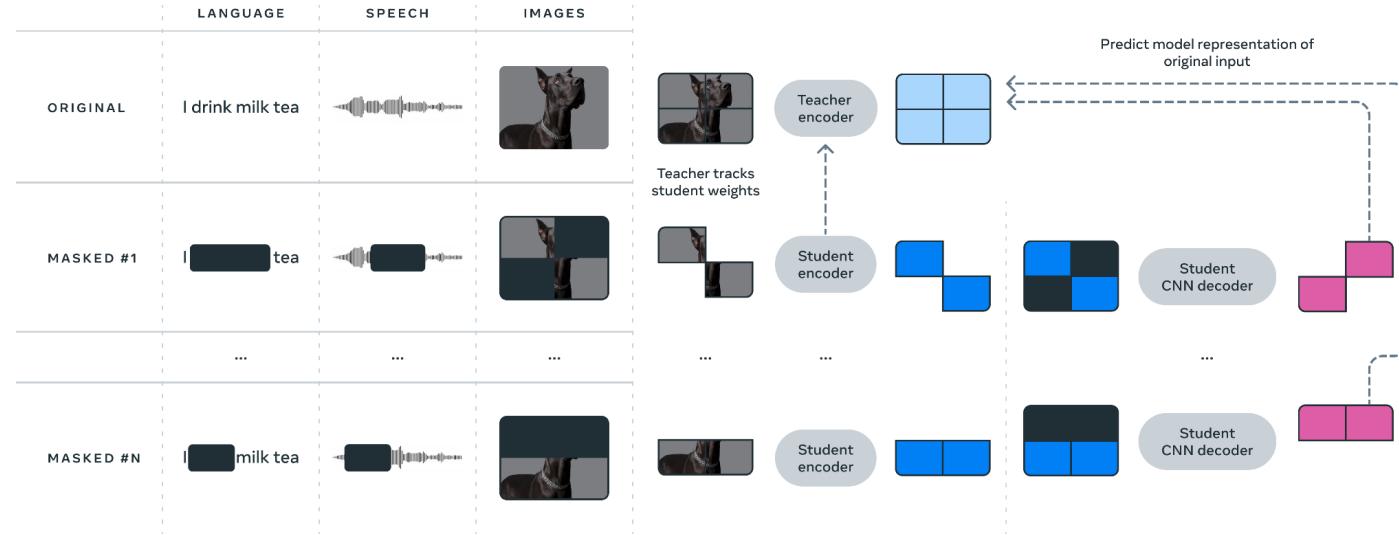
	Dataset Examples	ImageNet	Zero-Shot	Δ Score
		ResNet101	CLIP	
ImageNet	    	76.2	76.2	0%
ImageNetV2	    	64.3	70.1	+5.8%
ImageNet-R	    	37.7	88.9	+51.2%
ObjectNet	    	32.6	72.3	+39.7%
ImageNet Sketch	    	25.2	60.2	+35.0%
ImageNet-A	    	2.7	77.1	+74.4%

Multimodal : text to image

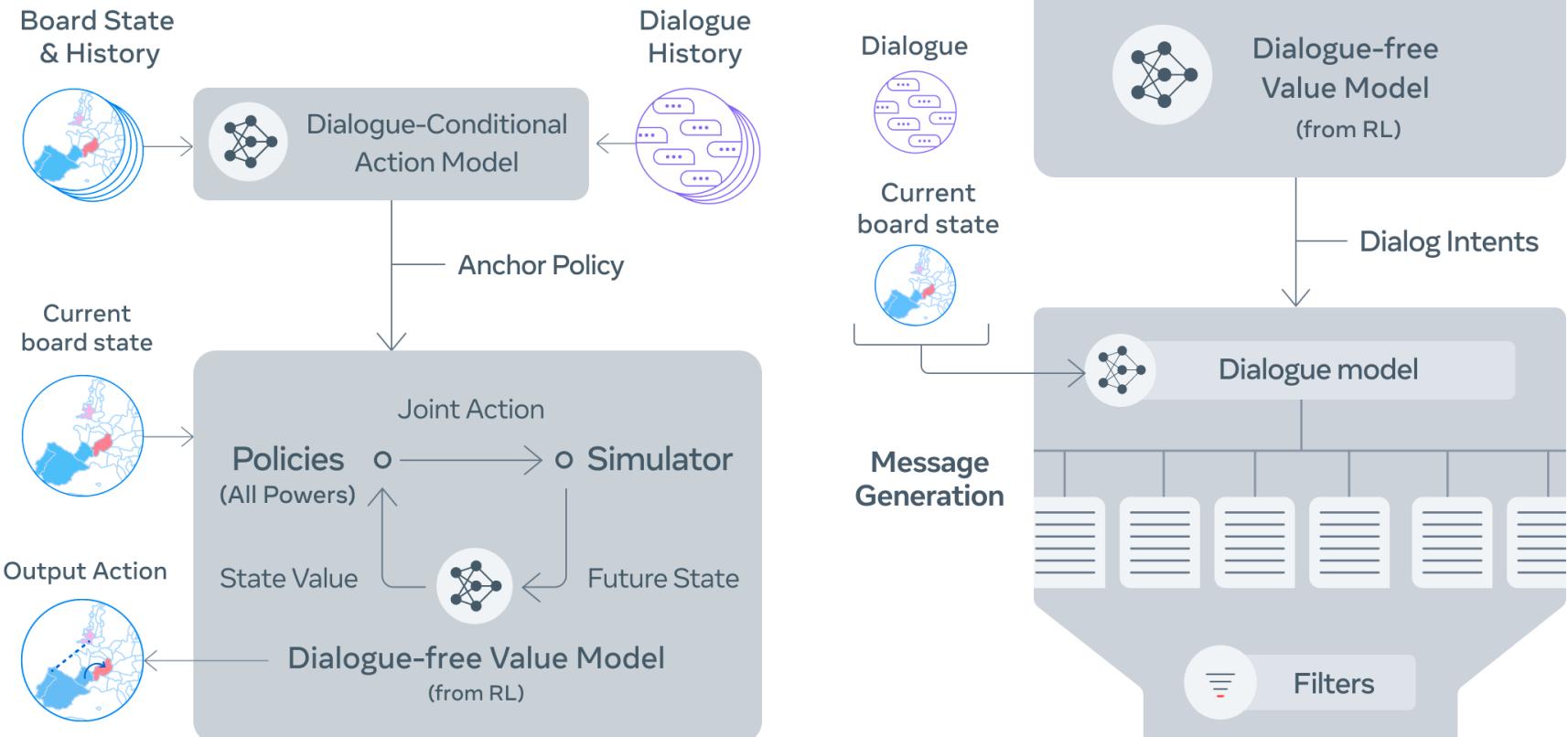


Multimodal : data2vec 2.0

self-supervised (BERT like) 3-modal : text, image, speech



Multimodal LM : CICERO (Diplomacy game)



Multimodal LLM : PaLM-E

Mobile Manipulation



Human: Bring me the rice chips from the drawer. Robot: 1. Go to the drawers, 2. Open top drawer. I see 3. Pick the green rice chip bag from the drawer and place it on the counter.

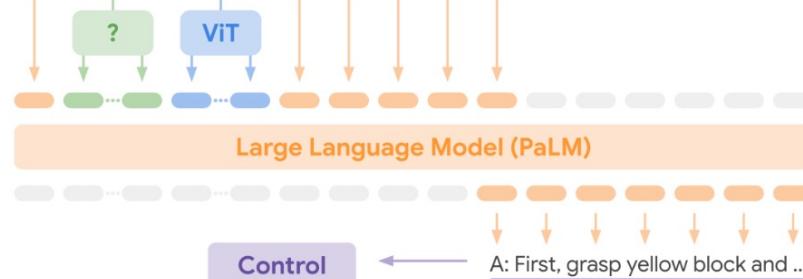
Visual Q&A, Captioning ...



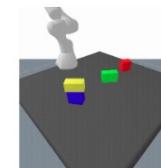
Given Q: What's in the image? Answer in emojis.
A: 🍎🍊🍇🍓🥝

PaLM-E: An Embodied Multimodal Language Model

Given `<emb>` ... `` Q: How to grasp blue block? A: First, grasp yellow block

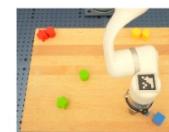


Task and Motion Planning



Given `<emb>` Q: How to grasp blue block?
A: First grasp yellow block and place it on the table, then grasp the blue block.

Tabletop Manipulation



Given Task: Sort colors into corners.
Step 1. Push the green star to the bottom left.
Step 2. Push the green circle to the green star.

Language Only Tasks



Describe the following
A dog jumping over a hurdle at a dog show.

Q: Miami Beach borders which ocean? A: Atlantic. Q: What is 372×18 ? A: 6696. Q: Write a Haiku about embodied LLMs. A: Embodied language. Models learn to understand. The world around them.

Multimodal LLM : GPT-4

Input Prompt



Question: which studio created the movie characters in the image?
Answer:

Completion

Disney 

Standard Prompting

Input Prompt



Introduce this picture in detail:

Completion

It is a robot that is holding up a cube that is the same color as the wall. The image is from the film, WALL-E, by Pixar Animation Studios. It is the story of a small robot named WALL-E and his quest to build the ultimate life form. WALL-E is created by a group...

Input Prompt



It is a robot that is holding up a cube that is the same color as the wall. The image is from the film, WALL-E, by Pixar Animation Studios. It is the story...
Question: which studio created the movie characters in the image?
Answer:

Completion

Pixar Animation Studios 

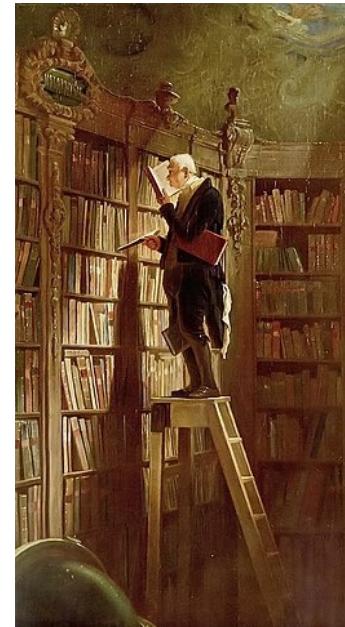
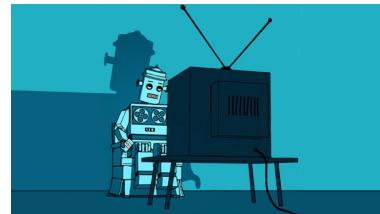
Multimodal Chain-of-Thought Prompting

Questions

Les LLM seront-ils les AGI de demain ?

Le *Multimodal* et l'*Embodiment* vont-ils combler l'absence de sens commun dans les corpus de textes ?

- Robotique
- Imitation learning (Videos)



Questions

Les IA génératives d'image, de vidéo, de son, d'écrits, vont-t-elles peupler notre imaginaire ?

quid de la culture ?

une culture libre ou biaisée, non contrôlée et pleine d'hallucinations numériques ?

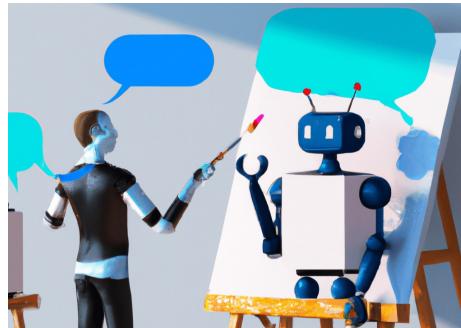


Image générée par DALL.E 2

L'Allégorie de la caverne de Platon



quid du DeepFake ?



DeepFake de V. Zelensky annonçant la capitulation de l'Ukraine

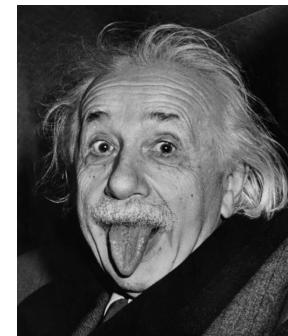
Questions

Science = scier, segmenter (*scire* savoir -> sek) / Intelligence = entre lire (*inter-legere*)



IA, AGI ?

Génie créatif, intuitif,
visionnaire



Oracle

quid de



?

Méthodologiste, rationaliste,
sceptique



ou

Outil scientifique

Questions

Comment résoudre les problèmes d'éthique ?



Quelles implications pour notre société ?

- Grande guerre des géants de la Tech contre tous les chômeurs ?
- l'AGI va-t-elle bénéficier à toute l'humanité ?
(voir openAI : [Planning for AGI and beyond](#))
- Quelle éducation pour nos enfants après chatGPT ?
- Ne serions-nous pas dans un “AI bashing” provoqué par la science-fiction et la dystopie du XXe siècle ?



Que dit le droit ?

15



**AI, Law, Society
and Ethics**

Questions ?



**AI, Law, Society
and Ethics**

Partie 2 : L'encadrement juridique des technologies d'IA

L'apparition d'enjeux juridiques causés par l'IA

L'exemple de la discrimination algorithmique

Les principes éthiques de l'IA

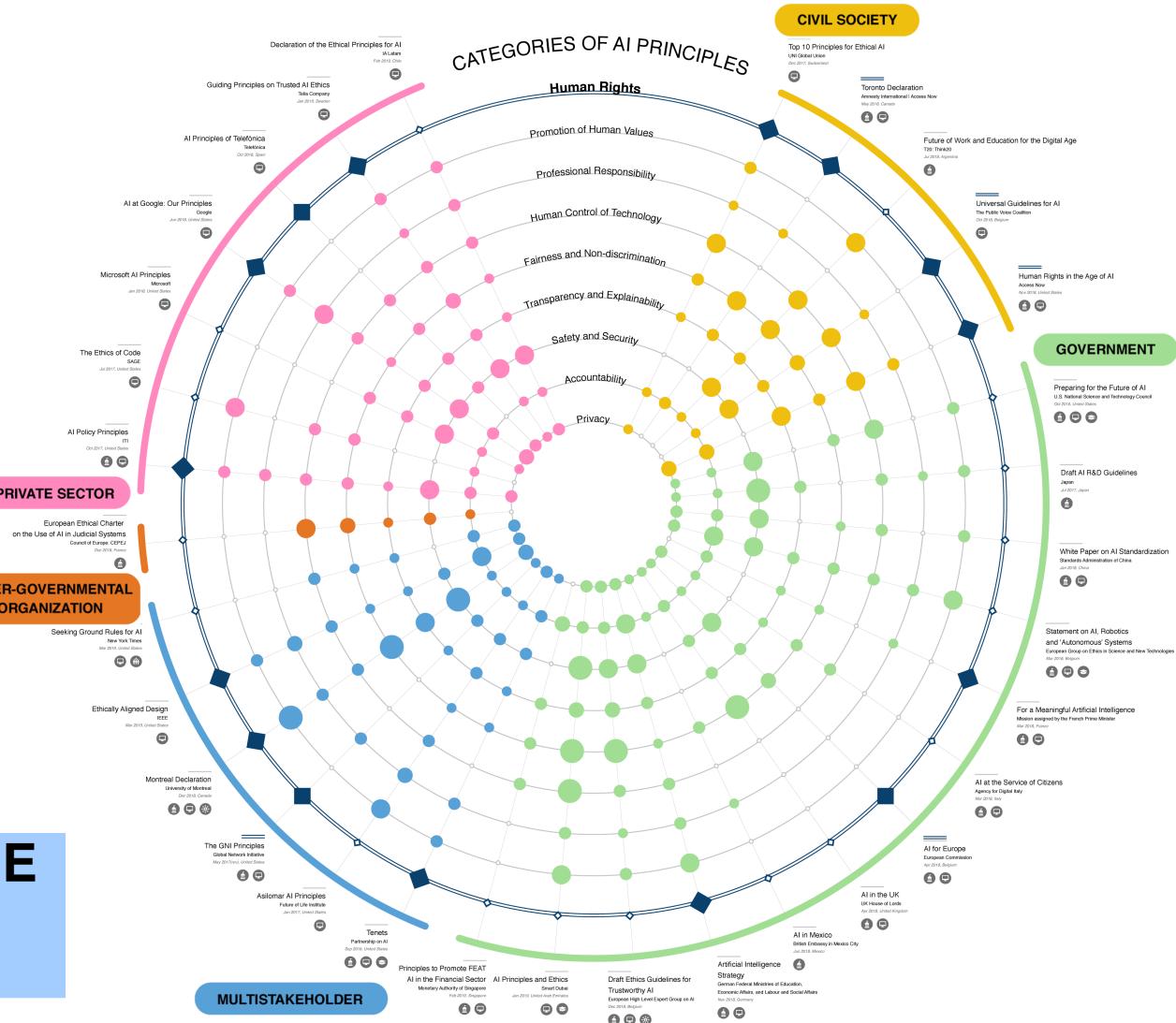
9 principes éthiques :

- Respect des droits de l'Homme ;
- Protection des valeurs humaines ;
- Responsabilité professionnelle ;
- Contrôle humain ;
- Equité et non-discrimination ;
- Transparence et Explicabilité ;
- Sureté et sécurité ;
- Reddition de comptes ;
- Vie privée ;

PRINCIPLED ARTIFICIAL INTELLIGENCE

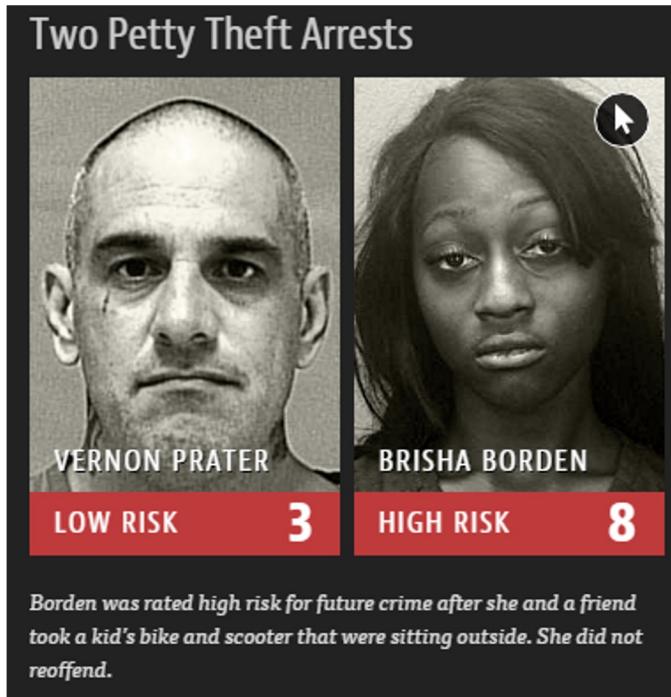
A Map of Ethical and Rights-Based Approaches July 4, 2019

Authors: Jessica Fjeld, Hannah Hilligoss, Nele Achten, Maia Levy Daniel, Joshua Feldman, Sally Kagay
Design: Arushi Singh (arushisingh.net)



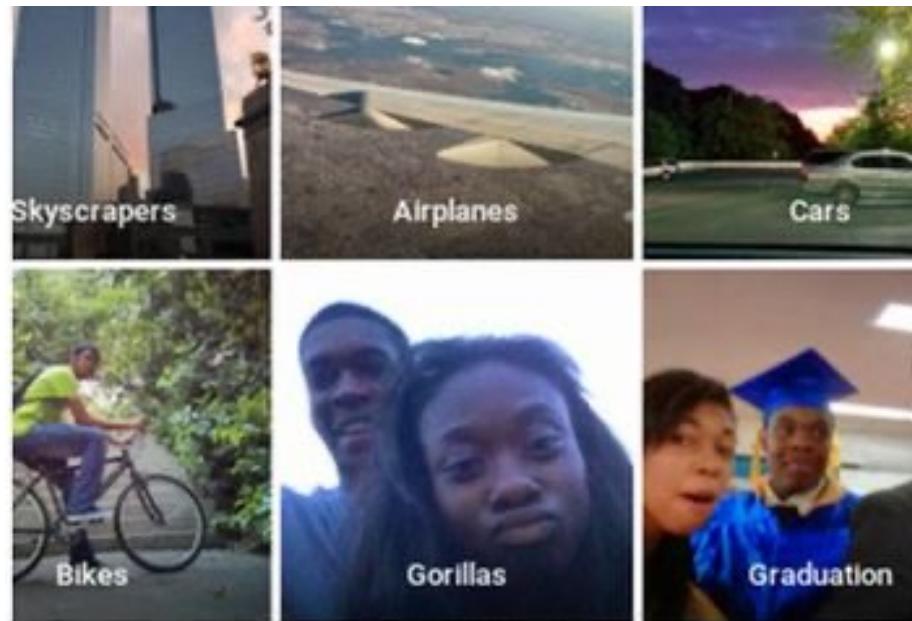
Risque juridique du *machine learning* : la discrimination algorithmique

COMPAS



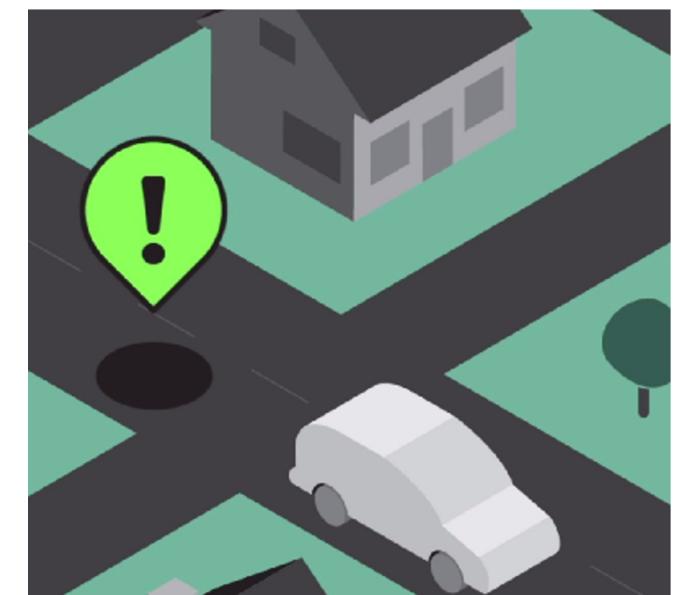
Source : Propublica

GOOGLE PHOTOS



Source : The Verge

STREET BUMP



Source : Boston.gov

La discrimination directe

« Constitue une **discrimination** toute distinction opérée entre les personnes physiques sur le fondement de leur origine, de leur sexe, de leur situation de famille, de leur grossesse, de leur apparence physique, [...] de leur état de santé, de leur perte d'autonomie, de leur handicap, de leurs caractéristiques génétiques, de leurs mœurs, de leur orientation sexuelle ,de leur identité de genre, de leur âge, de leurs opinions politiques, de leurs activités syndicales, de leurs capacités à s'exprimer dans une langue autre que le français, de leur appartenance ou de leur non-appartenance, vraie ou supposée, à une ethnie, une Nation, une prétendue race ou une religion déterminée.

Article 225-1 du code pénal



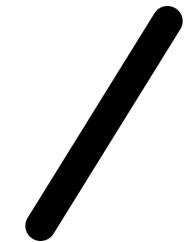
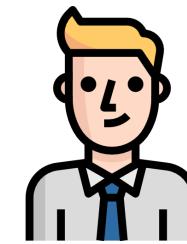
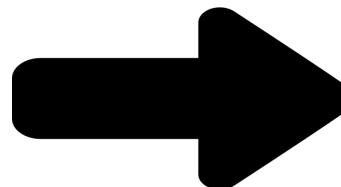
La discrimination indirecte

« Constitue une *discrimination indirecte* une disposition, un critère ou une pratique neutre en apparence, mais susceptible d'entraîner, pour l'un des motifs mentionnés au premier alinéa, un désavantage particulier pour des personnes par rapport à d'autres personnes, à moins que cette disposition, ce critère ou cette pratique ne soit objectivement justifiée par un but légitime et que les moyens pour réaliser ce but ne soient nécessaires et appropriés. »

Article 1 de la loi n°2008-496 du 27 mai 2008



Critère « neutre »



Conséquences discriminantes

La réponse juridique aux risques des technologies d'IA

I. Le droit des décisions automatisées



Loi Informatique et Libertés (LIL) : l'interdiction des décisions automatisées

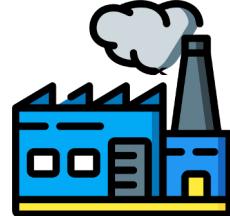
« Aucune décision produisant des *effets juridiques* à l'égard d'une personne ou l'affectant de manière significative ne peut être prise sur le seul fondement d'un traitement automatisé de données à caractère personnel, y compris le profilage »

Article 47 LIL

Secteurs concernés



Public



Privé



FIDLE

Gravité de la décision pour l'individu



Absence d'intervention humaine



Traitement de données personnelles





Règlement Général à la Protection des Données personnelles (RGPD) : Obligation d'information et Droit d'accès

« [...] l'existence d'une prise de décision automatisée, [...] et, des informations utiles concernant la logique sous-jacente, ainsi que l'importance et les conséquences prévues de ce traitement pour la personne concernée. »

Articles 13.2.f), 14.2.g) et 15.1.h)

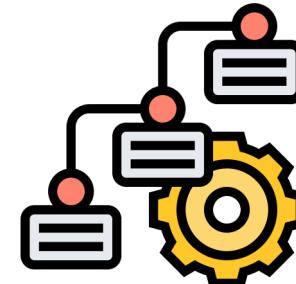
Moment de la communication



Acteurs concernés



Logique de l'algorithme



Contenu de l'information



Conséquences pour l'individu



FIDLE

Ronan Pons (CNRS, 3IA ANITI)



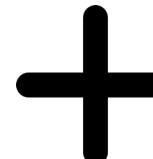
Les décisions administratives algorithmiques



Obligation de publication



« *Règles définissant les principaux traitement algorithmiques* »



Code source du logiciel de décision

Secteurs



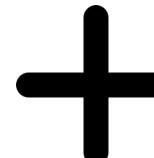
Administrations



Obligation de communication à l'intéressé



Présence d'un algorithme dans la décision



Principales caractéristiques (sur demande)

Décisions



Toutes décisions automatisées



Les « principales caractéristiques »

Communiquées « *sous une forme intelligible [...]* » :

- 1 Le **degré et le mode de contribution** du traitement algorithme à la prise de décision ;
- 2 Les **données traitées et leurs sources** ;
- 3 Les **paramètres de traitement** et, le cas échéant, leur **pondération, appliqués à la situation de l'intéressé** ;
- 4 Les **opérations effectuées par le traitement.**



La réponse juridique aux risques des technologies d'IA

II. La réglementation des systèmes d'IA

Proposition de règlement européen sur les systèmes d'IA à haut risque



COMMISSION
EUROPÉENNE

Bruxelles, le 21.4.2021
COM(2021) 206 final

2021/0106 (COD)

Proposition de

RÈGLEMENT DU PARLEMENT EUROPÉEN ET DU CONSEIL

ÉTABLISANT DES RÈGLES HARMONISÉES CONCERNANT L'INTELLIGENCE
ARTIFICIELLE (LÉGISLATION SUR L'INTELLIGENCE ARTIFICIELLE) ET
MODIFIANT CERTAINS ACTES LÉGISLATIFS DE L'UNION

{SEC(2021) 167 final} - {SWD(2021) 84 final} - {SWD(2021) 85 final}



Certification obligatoire des
systèmes d'IA à haut risque
avant toute commercialisation
ou utilisation sur le marché
européen

Définition de « système d'IA »

1

Pour un ensemble donné d'objectifs définis par l'Homme, un logiciel qui génère des résultats

ET

2

Un logiciel développé au moyen d'une technique/approche ci-dessous:

- (a) Approches d'apprentissage automatique, y compris d'apprentissage supervisé, non supervisé et par renforcement, utilisant une grande variété de méthodes, y compris l'apprentissage profond.
- (b) Approches fondées sur la logique et les connaissances, y compris la représentation des connaissances, la programmation inductive (logique), les bases de connaissances, les moteurs d'inférence et de déduction, le raisonnement (symbolique) et les systèmes experts.
- (c) Approches statistiques, estimation bayésienne, méthodes de recherche et d'optimisation.



Les catégories de risques des SIA



Les niveau de risques des systèmes d'IA



Risques innacceptables – Utilisations de systèmes d'IA interdites

- 1** L'altération inconsciente du comportement d'un individu par des techniques subliminales
- 2** Exploitation des vulnérabilités (âge, handicap physique et mental) pour altérer le comportement d'un individu
- 3** Notation sociale qui entraîne 1) des sanctions dans un contexte différent de la collecte de données ou 2) une aggravation des sanctions du fait de la notation
- 4** Identification biométrique à distance et en « temps réel » dans des espaces public*



FIDLE

Les niveau de risques des systèmes d'IA



Risque haut/élevé – Les systèmes d'IA réglementés

1 Systèmes d'IA intégrés dans un produit déjà réglementé spécifiquement

OU

2 Utilisation d'un système d'IA dans un secteur critique

Les systèmes d'IA à haut risque sont les principales cibles des dispositions de cette nouvelle proposition de règlement européen

Liste des secteurs critiques



Activités des autorités répressives



Identification biométrique et catégorisation des personnes



Accès et droit aux services privés essentiels et services publics



Gestion et exploitation des infrastructures critiques



Gestion migration, asile et contrôle aux frontières



Education et formation professionnelle



Administration de la justice et des processus démocratiques



Emploi, gestion main-d'œuvre et accès à l'emploi indépendant

Les niveau de risques des systèmes d'IA



Risques spécifiques – Obligations spécifiques

- 1 Systèmes d'IA en interaction avec des personnes physiques → **Obligation d'informer l'individu qu'il est en interaction avec un système d'IA**
- 2 Systèmes de reconnaissance des émotions et catégorisation biométrique → **Obligation d'informer l'individu du fonctionnement du système d'IA**
- 3 Système de *deepfakes* (« hypertrucages ») → **Obligation d'indiquer que le contenu a été générée ou modifiée artificiellement**



Les acteurs de la réglementation



Fournisseur

- Développe un SIA
- Commercialisation ou Mise en service dans l'UE
- Payant ou Gratuit



Utilisateur

- Utilise le SIA sous son autorité
- Respecte la notice d'utilisation du fournisseur
- Surveille l'exécution du système



Contrôleurs

1

Autorité nationale

2

Organismes notifiés



Sanctions en cas de violation

- 1 Utilisation de SIA interdits, ou Manquement aux obligations sur les données → 30M € ou 6% du CA annuel mondial
- 2 Manquement aux obligations du règlement → 20M € ou 4% du CA annuel mondial
- 3 Communication de fausses informations ou d'informations trompeuses aux autorités → 10M € ou 2% du CA annuel mondial

La responsabilité du fait des systèmes d'IA



Bruxelles, le 28.9.2022
COM(2022) 496 final
2022/0303 (COD)

Proposition de

DIRECTIVE DU PARLEMENT EUROPÉEN ET DU CONSEIL

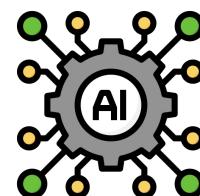
relative à l'adaptation des règles en matière de responsabilité civile extracontractuelle
au domaine de l'intelligence artificielle
(Directive sur la responsabilité en matière d'IA)

(Texte présentant de l'intérêt pour l'EEE)

{SEC(2022) 344 final} - {SWD(2022) 318 final} - {SWD(2022) 319 final} -
{SWD(2022) 320 final}



Adaptation des règles de la
responsabilité des produits
contenant des techniques d'IA



Facilitation de la preuve
pour le plaignant

La responsabilité des plateformes en ligne

27.10.2022

FR

Journal officiel de l'Union européenne

L 277/1

I

(Actes législatifs)

RÈGLEMENTS

RÈGLEMENT (UE) 2022/2065 DU PARLEMENT EUROPÉEN ET DU CONSEIL
du 19 octobre 2022

relatif à un marché unique des services numériques et modifiant la directive 2000/31/CE (règlement
sur les services numériques)

(Texte présentant de l'intérêt pour l'EEE)

LE PARLEMENT EUROPÉEN ET LE CONSEIL DE L'UNION EUROPÉENNE,

vu le traité sur le fonctionnement de l'Union européenne, et notamment son article 114,

vu la proposition de la Commission européenne,

après transmission du projet d'acte législatif aux parlements nationaux,



Accès aux données de
plateforme pour les
chercheurs



Amélioration de la lutte
contre les contenus illicites



Etude des risques
systémiques pour les
utilisateurs

La réponse juridique aux risques des technologies d'IA

III. Les objectifs à venir

La recherche de métriques d'évaluation

L'exemple des biais discriminatoires :

Adéquation entre la
métrique et
l'exigence juridique

1

La présence d'un biais n'entraîne pas nécessairement de conséquences sur des individus

2

La présence d'un biais sur des individus n'entraîne pas nécessairement une discrimination juridique



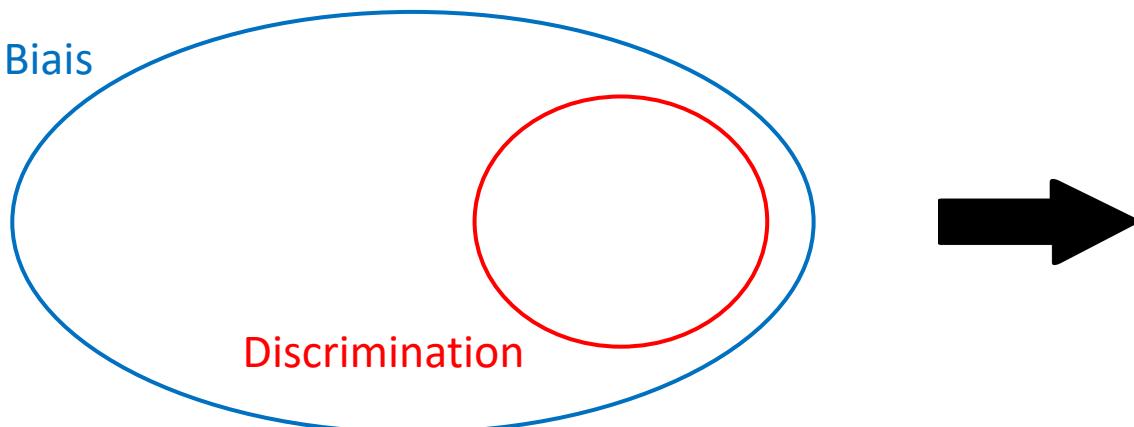
Les métriques existantes sont conçues pour détecter des biais et non des biais discriminatoires au sens juridique

La recherche de métriques d'évaluation



La sélection de l'indicateur de mesure de biais est un choix technique mais avec des enjeux politiques et juridiques

Défis à venir dans l'application de ce règlement :



COLLABORATION ENTRE
SCIENTIFIQUE ET JURISTES
INDISPENSABLE !

Sélectionner les métriques les plus proches de la conception juridique

La valeur juridique du résultat scientifique



L'échec à un test technique de fonctionnement peut-il faire sanctionner un système d'IA ?



La certification par indicateurs de systèmes d'IA garantit-elle l'absence de dommage pour l'individu ? Si non, comment prouver le dommage ?



Le résultat d'un SIA peut-il servir à la décision d'un juge ? Dans quelles conditions ?

15



**AI, Law, Society
and Ethics**

Questions ?

15



**AI, Law, Society
and Ethics**

Partie 3 : Solutions pour détecter un algorithme d'I.A. discriminant

Partie 3.1 : Biais de discrimination en apprentissage automatique



Aide au diagnostic

Base d'apprentissage

Patient 1 :

- Age = 40
- Globule Blancs/L = 6

Sain

Patient 2 :

- Age = 28
- Globule Blancs/L = 12

Rhume

Patient N :

- Age = 57
- Globule Blancs/L = 8

Sain

Nouveau Patient (hors base d'apprentissage) :

- Age = 34
- Globule Blancs/L = 5



Sain ou rhume ???

Partie 3.1 : Biais de discrimination en apprentissage automatique

Base d'apprentissage

Patient 1 :

- Age = 40
- Globule Blancs/L = 6

Sain

Patient 2 :

- Age = 28
- Globule Blancs/L = 12

Rhume

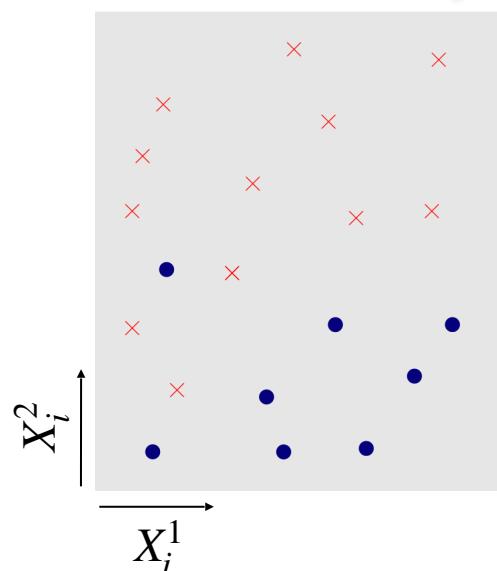
Patient n :

- Age = 57
- Globule Blancs/L = 8

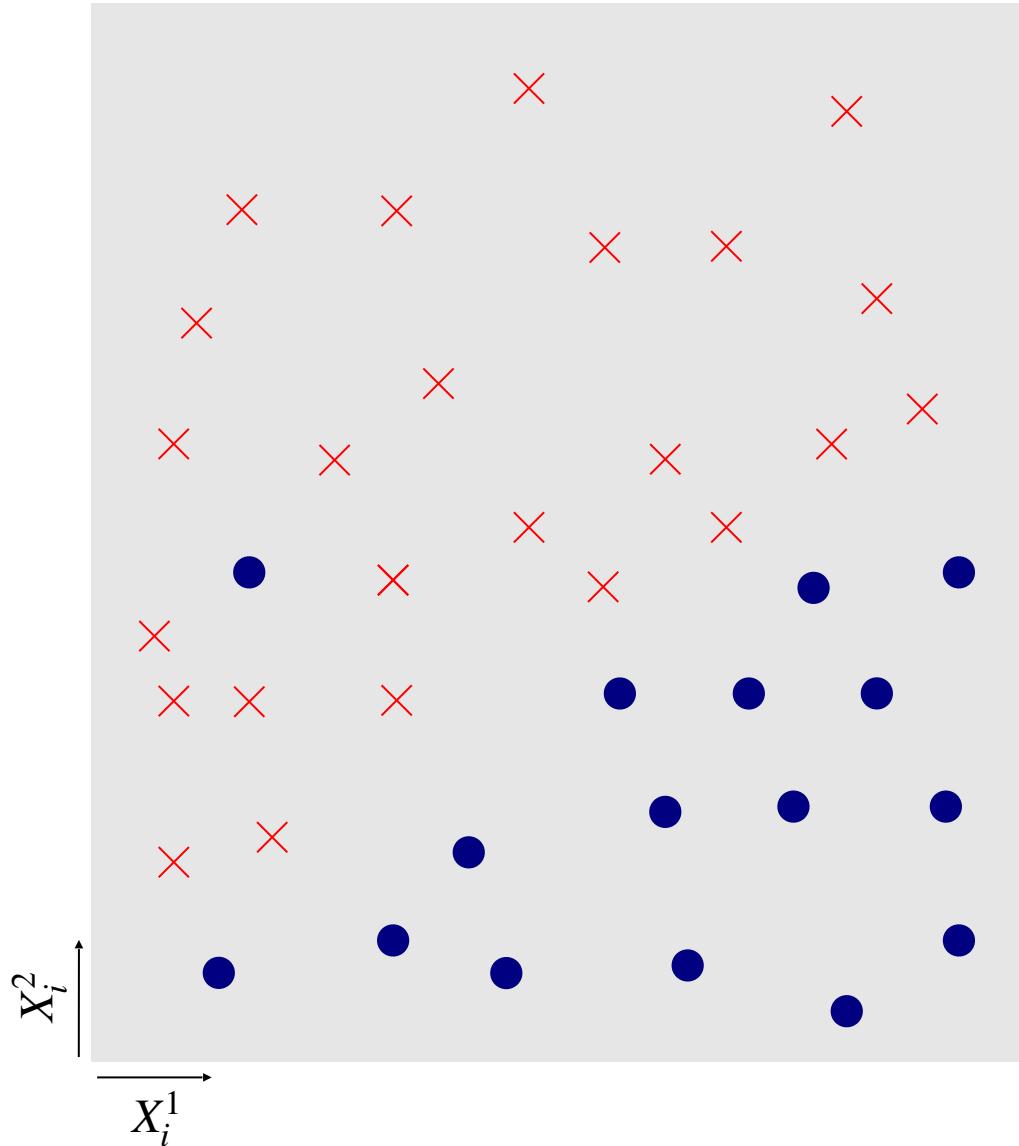
Sain



	Age	G.B./L	Etat
Pat. 1	40	6	1
Pat. 2	28	12	0
...
Pat. n	57	8	1



Partie 3.1 : Biais de discrimination en apprentissage automatique



Observations d'entrée (X) :

- n observations $X_i \in \mathbb{R}^p$

Observations de sortie (Y) :

- n Labels $Y_i \in \{0,1\}^K$

- $\text{--- } Y_i = 1$

- $\bullet \quad Y_i = 0$

- Ici $n = 40$, $p = 2$ et $K=1$

Dans notre exemple :

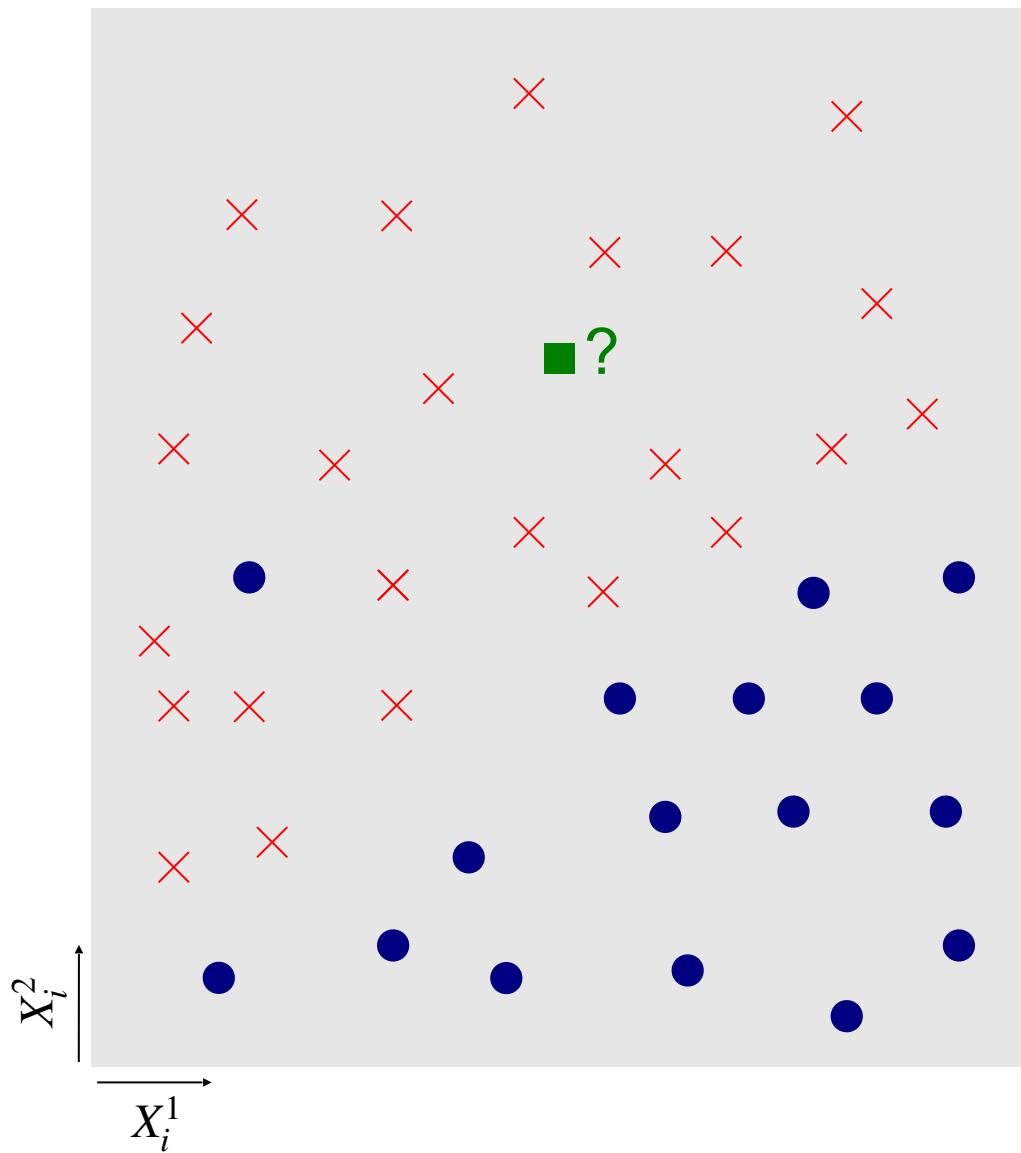
$i \rightarrow$ Patient de la base d'apprentissage

$X_i^1 \rightarrow$ Age

$X_i^2 \rightarrow$ Globule Blancs/L

$Y_i \rightarrow$ Sain ou rhume

Partie 3.1 : Biais de discrimination en apprentissage automatique



Observations d'entrée (X) :

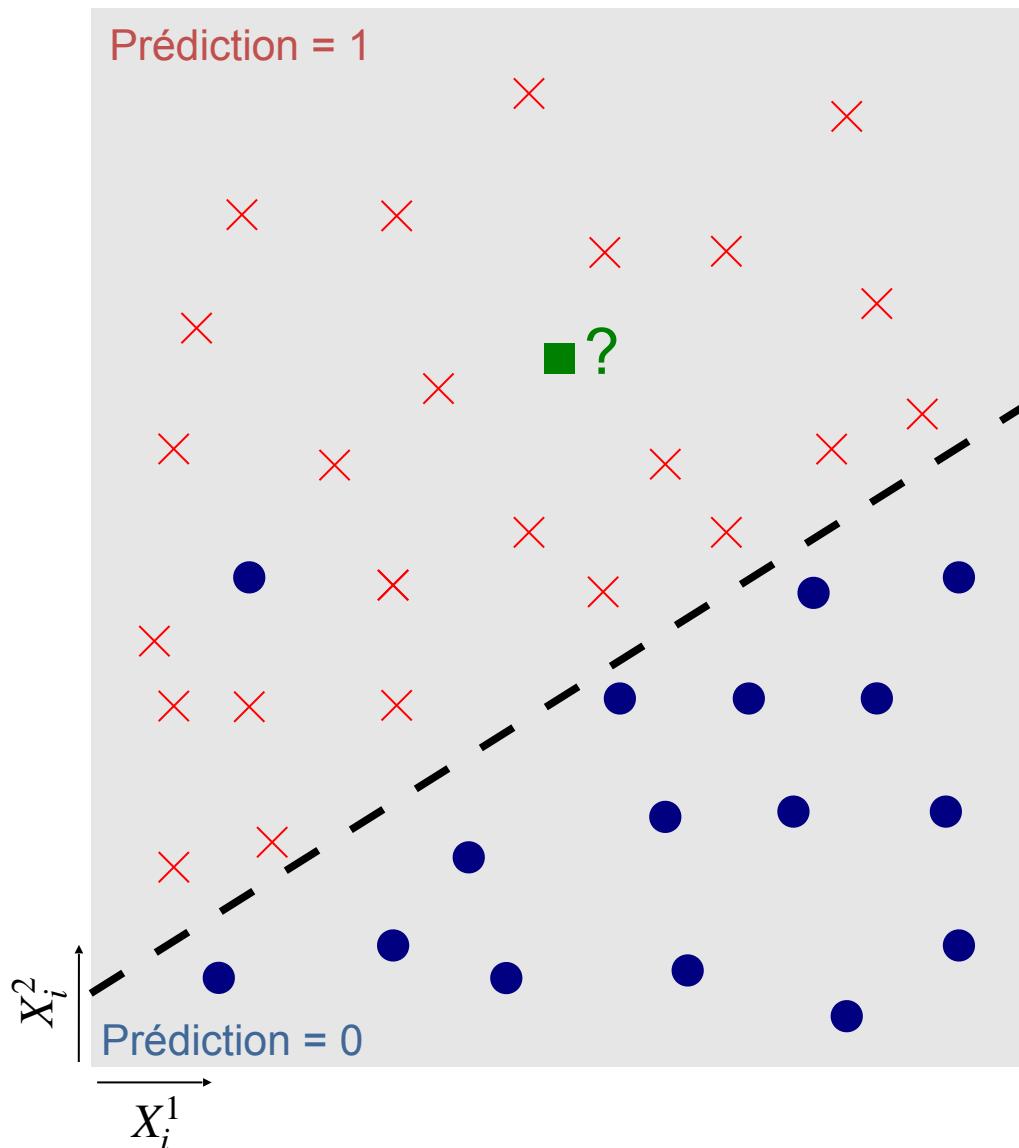
- n observations $X_i \in \mathbb{R}^p$

Observations de sortie (Y) :

- n Labels $Y_i \in \{0,1\}^K$
- $\textcolor{red}{\times} Y_i = 1$
- $\bullet Y_i = 0$
- Ici $n = 40$, $p = 2$ et $K=1$

Label le plus probable de █ ?

Partie 3.1 : Biais de discrimination en apprentissage automatique



Observations d'entrée (X) :

- n observations $X_i \in \mathbb{R}^p$

Observations de sortie (Y) :

- n Labels $Y_i \in \{0,1\}^K$

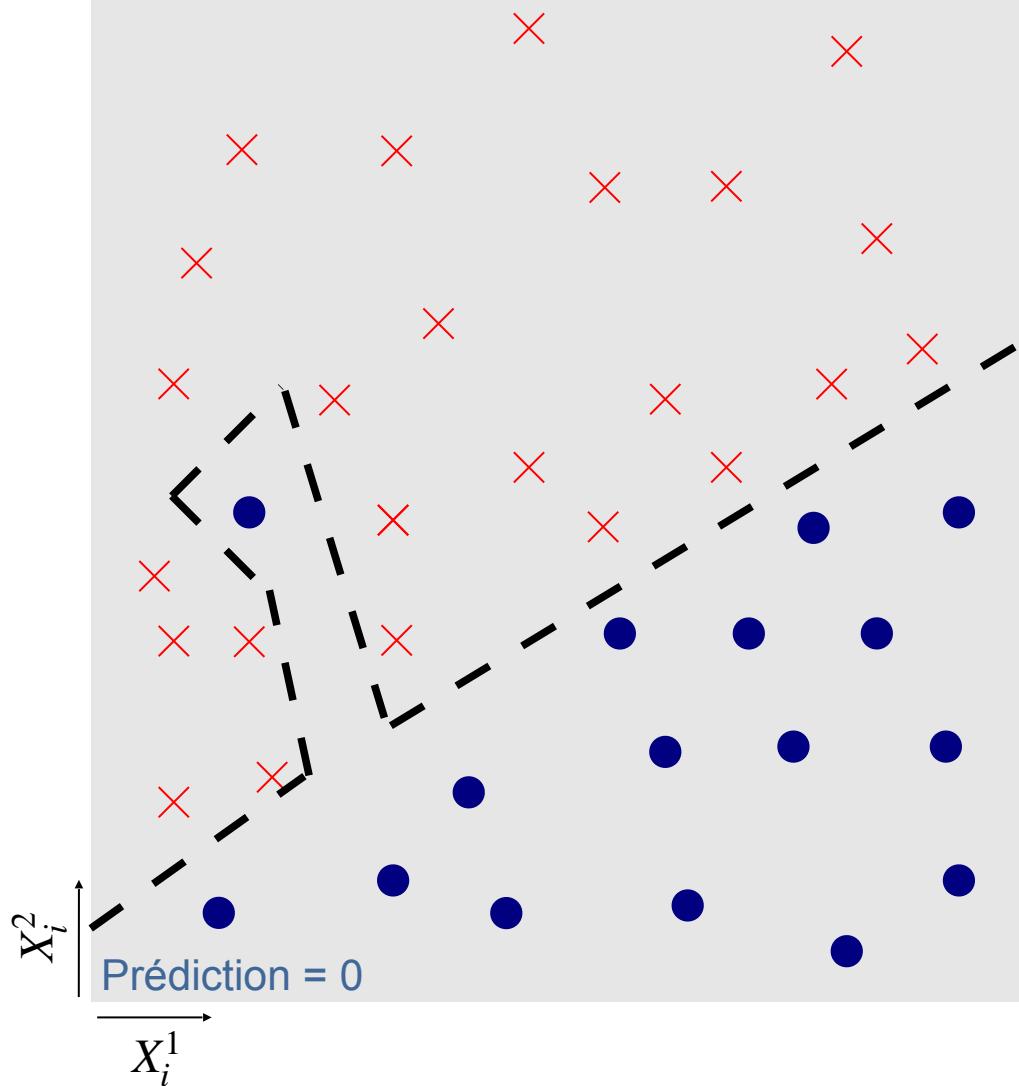
- $\times Y_i = 1$
- $\bullet Y_i = 0$

- Ici $n = 40$, $p = 2$ et $K=1$

1. **Choix d'un modèle** pour séparer les données d'apprentissage, i.e. les \bullet et les \times .
2. **Apprentissage des paramètres** optimaux
3. Une fois les paramètres du modèle appris, **prédiction** extrêmement simple et rapide de nouvelles observations comme \blacksquare .

Partie 3.1 : Biais de discrimination en apprentissage automatique

Prédiction = 1

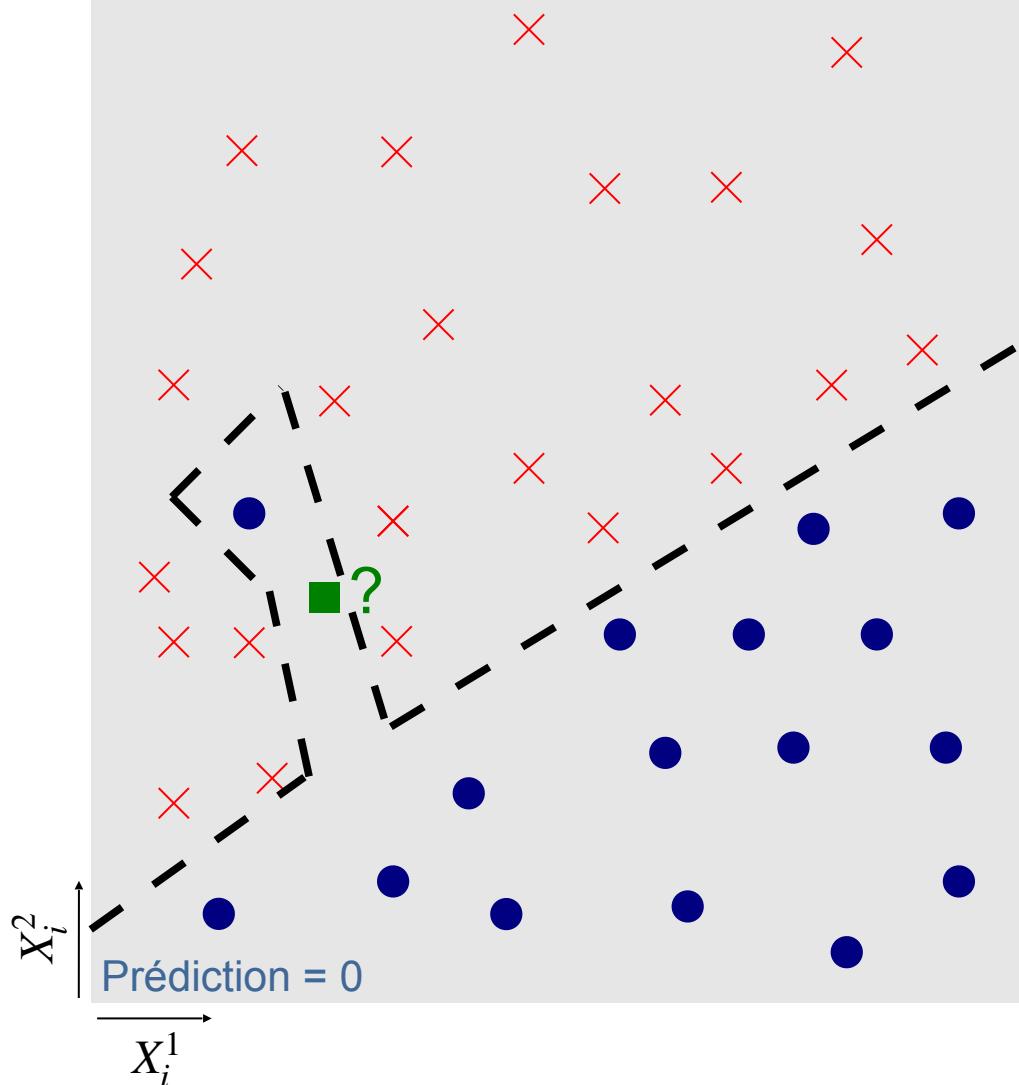


Sur l'échantillon d'apprentissage, une observation est mal classée !

→ Utilisation d'un modèle moins flexible

Partie 3.1 : Biais de discrimination en apprentissage automatique

Prédiction = 1



Sur l'échantillon d'apprentissage, une observation est mal classée !

→ Utilisation d'un modèle moins flexible

Mauvaise généralisation ici

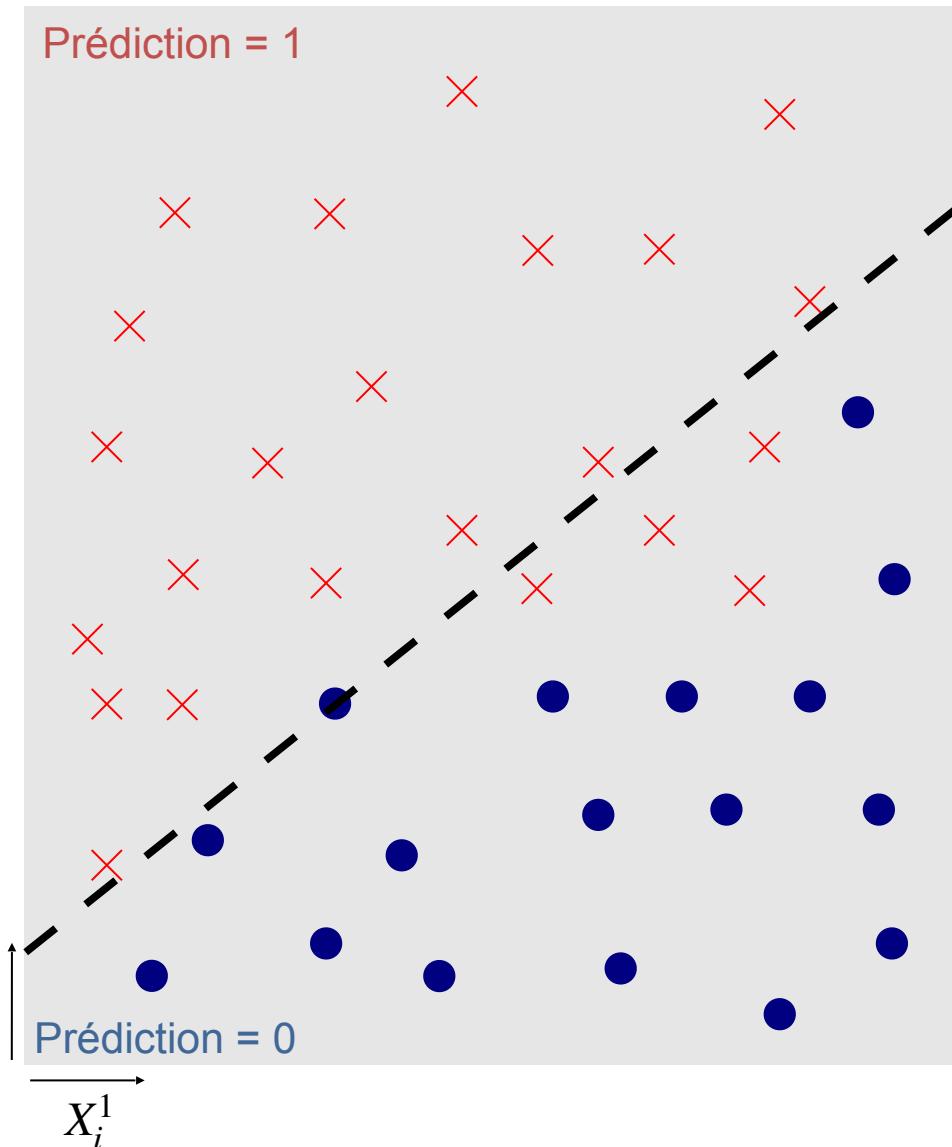
→ Nécessité de définir des modèles raisonnablement contraints à la vue des données

→ Importance de la validation croisée en apprentissage quand il est difficile ou impossible de visualiser toutes les données.

... ouvrons une discussion moins connue mais ayant un impact qui est loin d'être négligeable !

Partie 3.1 : Biais de discrimination en apprentissage automatique

Il existe de nombreuses sources de biais potentiels en I.A. → Voyons un cas simple mais didactique

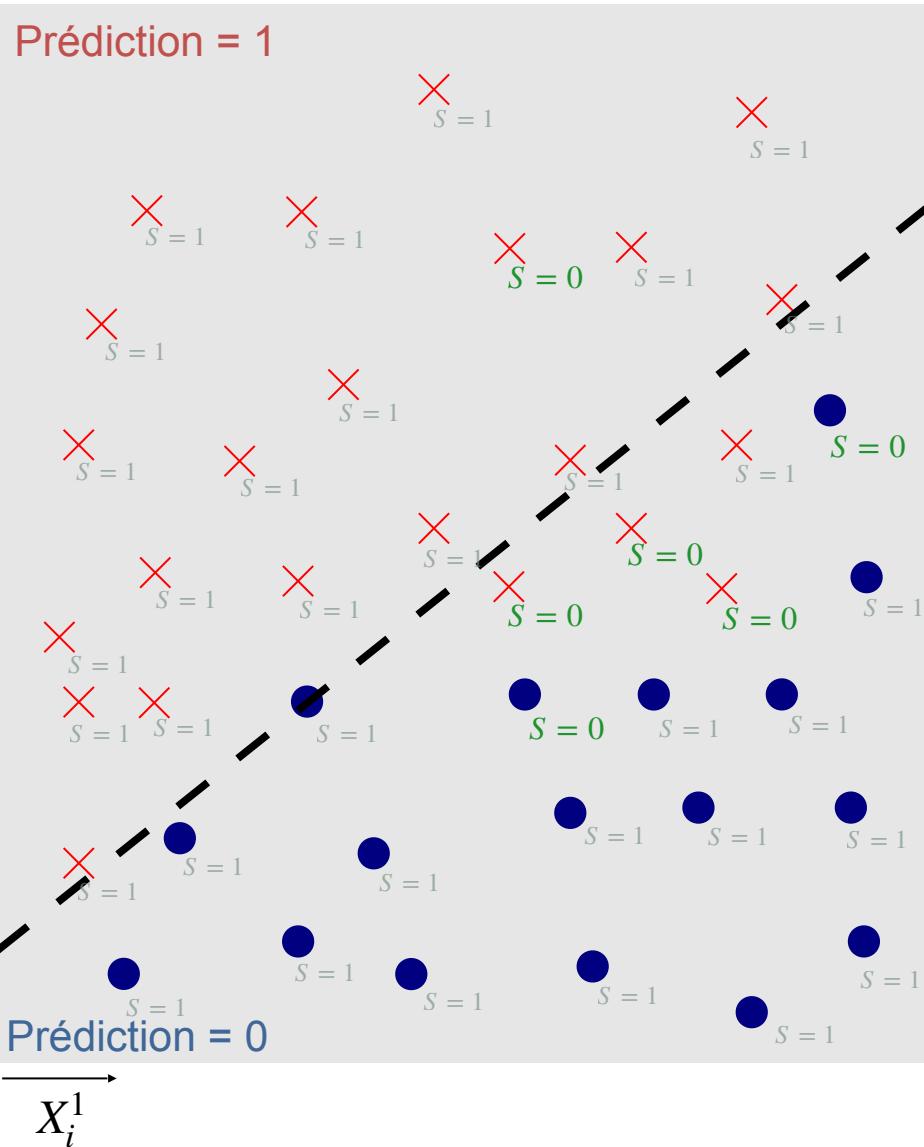


Modèle linéaire pour séparer les données

Semble raisonnable après validation croisée

Partie 3.1 : Biais de discrimination en apprentissage automatique

Il existe de nombreuses sources de biais potentiels en I.A. → Voyons un cas simple mais didactique



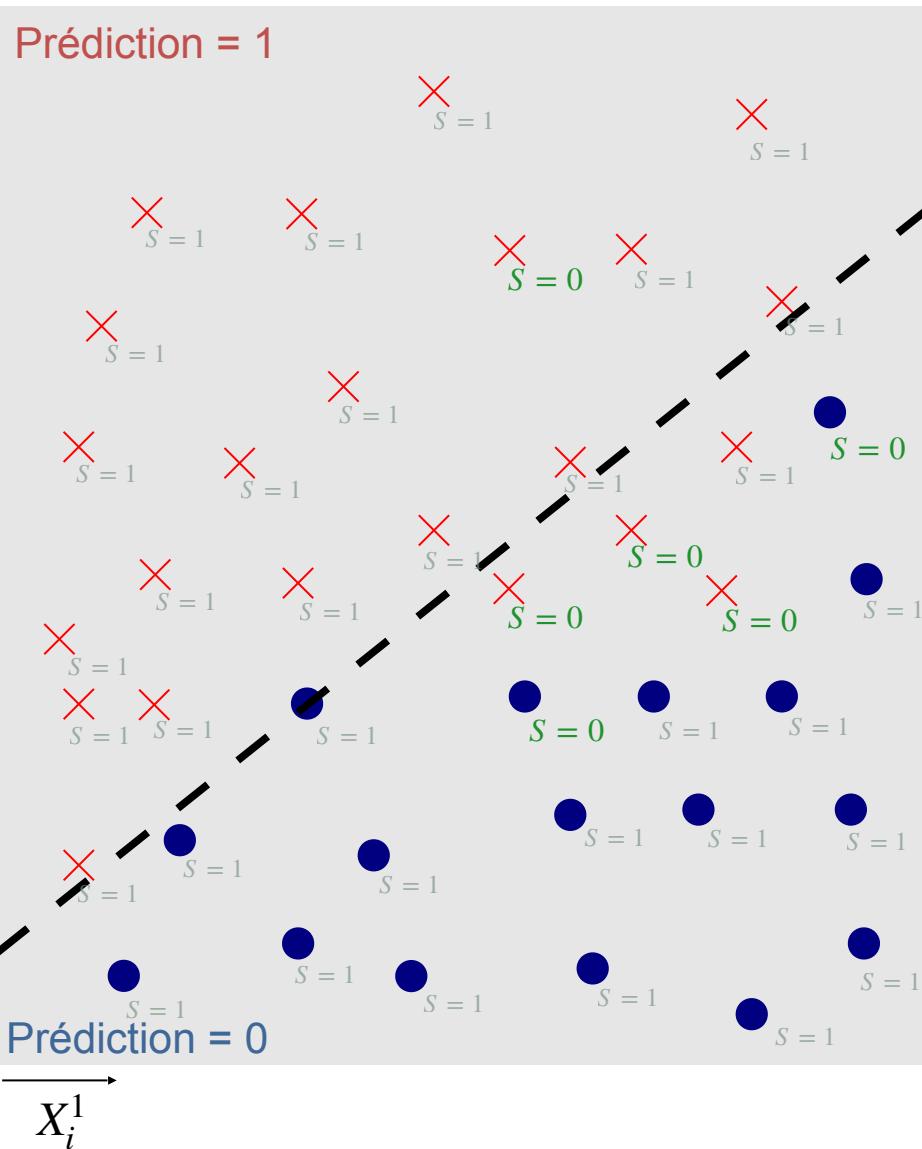
Modèle linéaire pour séparer les données

Semblé raisonnable après validation croisée

Intéressons nous maintenant à une variable sensible $S \in \{0,1\}$ (origine géographique / régime alimentaire / contexte de travail).

Partie 3.1 : Biais de discrimination en apprentissage automatique

Il existe de nombreuses sources de biais potentiels en I.A. → Voyons un cas simple mais didactique



Modèle linéaire pour séparer les données

Semblé raisonnable après validation croisée

Intéressons nous maintenant à une variable sensible $S \in \{0,1\}$ (origine géographique / régime alimentaire / contexte de travail).

On suppose que les données de test et d'apprentissage suivent une même distribution :

- Le taux d'erreurs est plus grand pour $S = 0$ que pour $S = 1$
- Il y a plus de prédictions négatives pour $S = 0$ que pour $S = 1$

Partie 3.1 : Biais de discrimination en apprentissage automatique

Conclusion partielle

- Nous avons vu une source de biais potentielle en apprentissage automatique et *il en existe d'autres*.
- Ces biais ne sont généralement pas intentionnels mais leur impact social peut être fort.
- Nous avons vu une application sociale, mais la problématique de la *robustesse* dans un contexte industriel est mathématiquement la même.

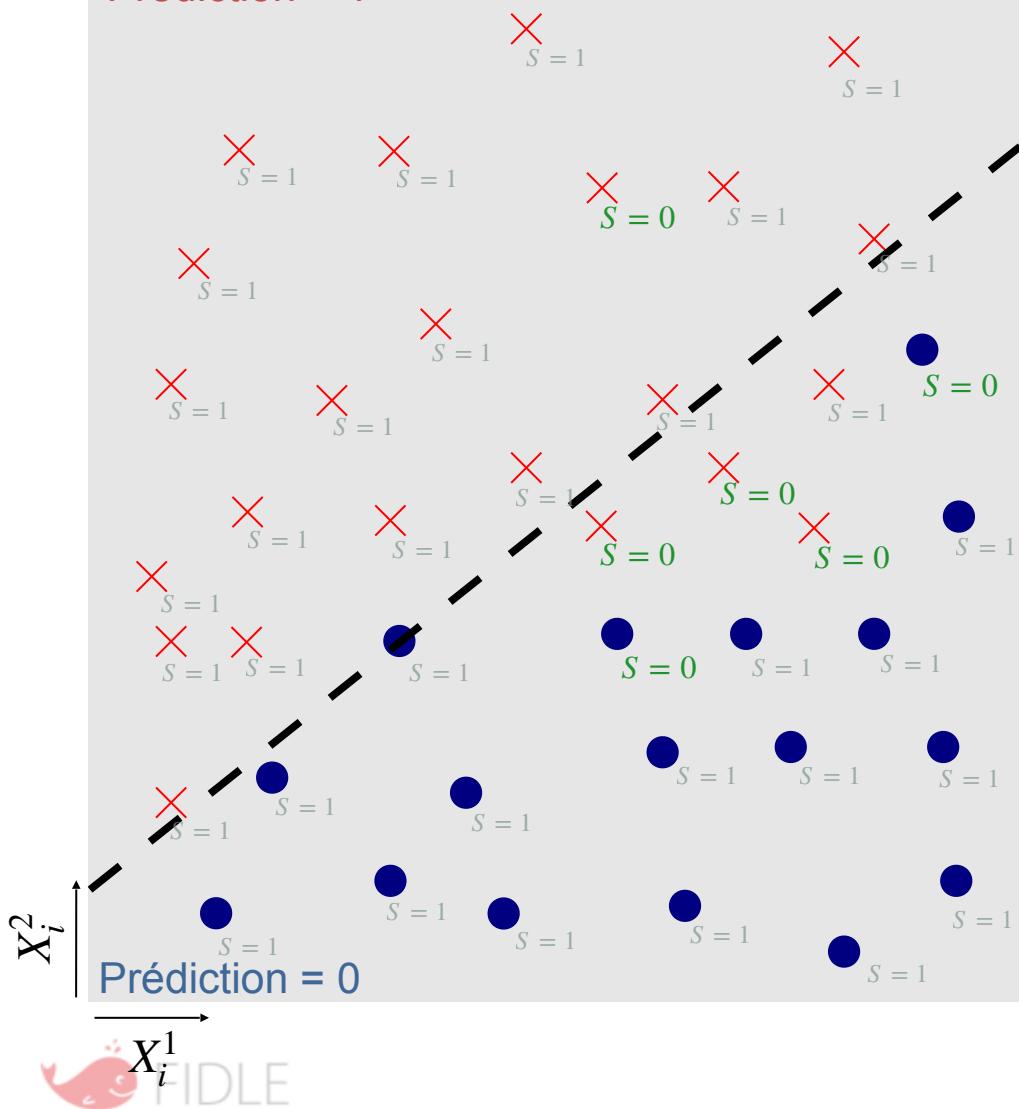
Nous allons

- Voir les deux outils de mesure de biais les plus classiques (partie 3.2)
- Voir deux stratégies d'explicabilité très populaires : Grad-CAM et LIME (partie 3.3)

Partie 3.2 : Mesure de biais algorithmiques en fonction d'une variable S

$$\text{Equality of Odds : E.O.} = \frac{\text{Taux de prédictions à 1 dans groupe } S = 0 \text{ si la prédiction devrait être 1}}{\text{Taux de prédictions à 1 dans groupe } S = 1 \text{ si la prédiction devrait être 1}}$$

Prédiction = 1

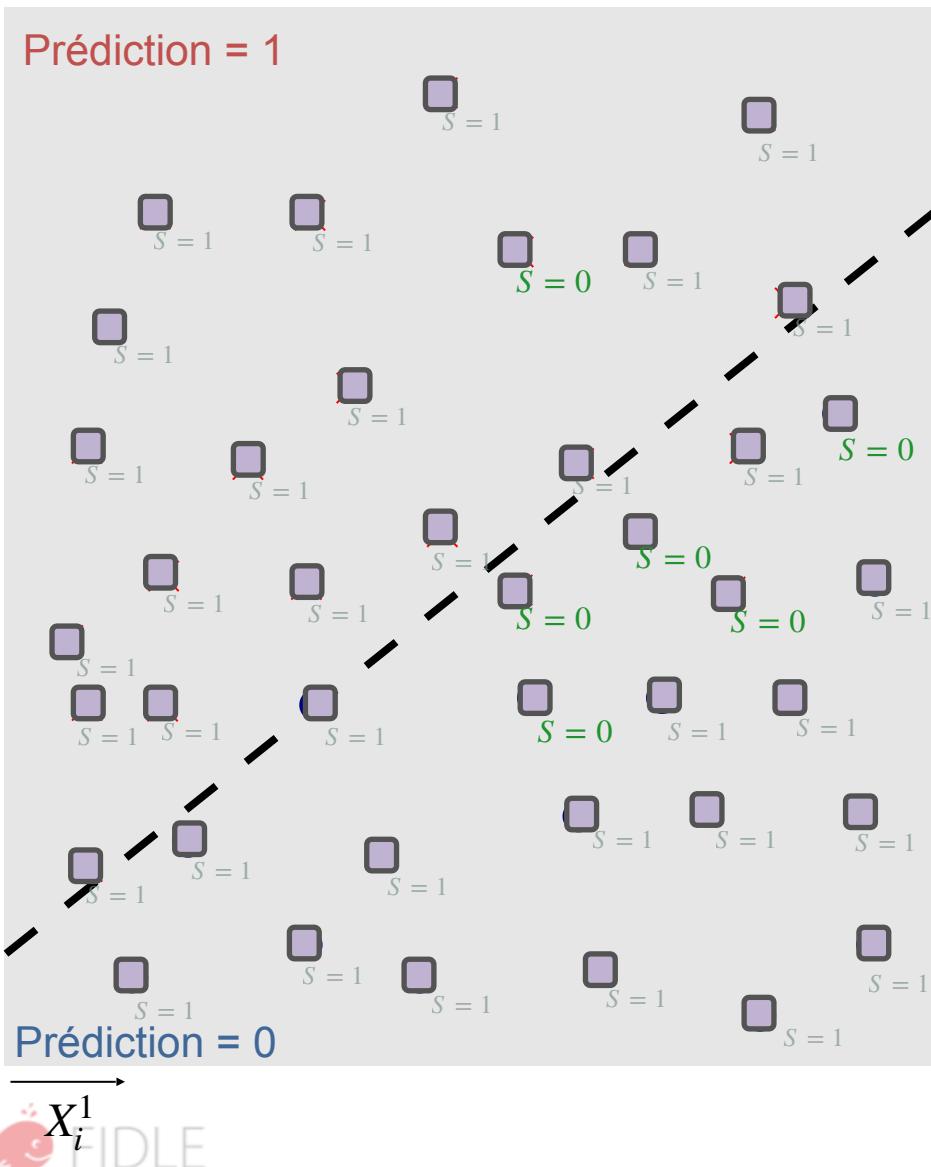


$$\text{Ici : E.O.} = \frac{\frac{1}{4}}{\frac{17}{18}} = \frac{0.25}{0.94} = 0.26$$

- A du sens pour la prédiction d'une maladie.
- Aurait aussi du sens pour la recommandation d'accès à un prêt bancaire.
- Nécessite une vérité terrain

Partie 3.2 : Mesure de biais algorithmiques en fonction d'une variable S

$$\text{Impact disparate : D.I.} = \frac{\text{Taux de prédictions à 1 dans groupe } S = 0}{\text{Taux de prédictions à 1 dans groupe } S = 1}$$



$$\text{Ici : D.I.} = \frac{\frac{1}{6}}{\frac{17}{34}} = \frac{0.167}{0.5} = 0.333$$

- N'a pas de sens pour la prédiction d'une maladie.
- A du sens pour la recommandation d'accès à un prêt bancaire.
- Ne conduit pas forcément aux mêmes conclusions que l'impact disparate

Reconnu légalement aux Etats-Unis depuis les années 70.

Partie 3.3 : Explicabilité des décisions par réseaux de neurones

On sait maintenant comment détecter un biais non désiré sur un *groupe d'observations*...

... mais comment expliquer qu'une *décision individuelle* est prise pour de bonnes raisons ?

Illustration classique des « Huskys vs Loups »

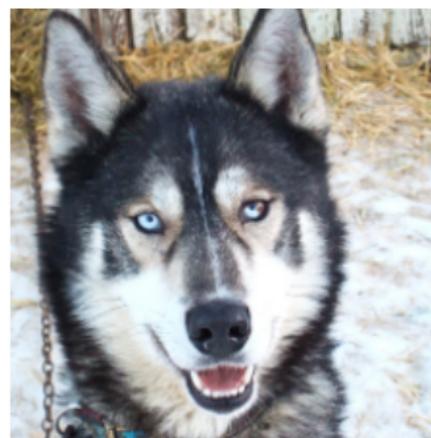
“Why Should I Trust You?”

Explaining the Predictions of Any Classifier

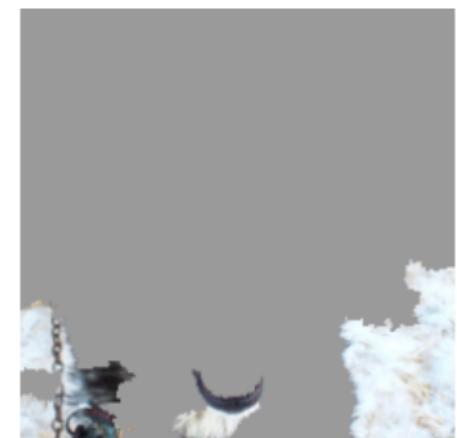
Marco Tulio Ribeiro
University of Washington
Seattle, WA 98105, USA
marcotcr@cs.uw.edu

Sameer Singh
University of Washington
Seattle, WA 98105, USA
sameer@cs.uw.edu

Carlos Guestrin
University of Washington
Seattle, WA 98105, USA
guestrin@cs.uw.edu



(a) Husky classified as wolf



(b) Explanation

Figure 11: Raw data and explanation of a bad model’s prediction in the “Husky vs Wolf” task.

Partie 3.3 : Explicabilité des décisions par réseaux de neurones

Articles récents en explicabilité pour l'apprentissage automatique (*Explainable Artificial Intelligence - XAI*) :

"*Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI*", A. Barrieta et al, 2019

"*Interpretable Explanations of Black Boxes by Meaningful Perturbation*", Ruth C. Fong, Andrea Vedaldi, 2017

"*MAGIX: model agnostic globally interpretable explanations.*" N. Puri, P. Gupta, P. Agarwal, S. Verma, and B. Krishnamurthy, CoRR, vol. abs/1706.07160, 2017.

"*Why should I trust you? Explaining the predictions of any classifier.*", T. Ribeiro, S. Singh, and C. Guestrin, 2016 - International Conference on Knowledge Discovery and Data Mining, ACM2016

"*Local Rule-Based Explanations of Black Box Decision Systems*" (*LORE*), Riccardo Guidotti et al 2018,

"*Anchors: High-precision model-agnostic explanations.*" T. Ribeiro, S. Singh, and C. Guestrin, , in AAAI Conference on Artificial Intelligence, 2018.

"*Visualizing the feature importance for black box models*", G. Casalicchio, C. Molnar, B. Bischl, arXiv:1804.06620.

"*Auditing black-box models for indirect influence*", P. Adler, C. Falk, S. A. Friedler, T. Nix, G. Rybeck, C. Scheidegger, B. Smith, S. Venkatasubramanian, Knowledge and Information Systems 54 (1) (2018) 95–122.

"*Entropic Variable Projection for Explainability and Interpretability*", F. Bachoc and F. Gamboa and M. Halford and J.-M. Loubes and L. Risser, 2018, arXiv:1810.07924.

"*Grad-cam: Visual explanations from deep networks via gradient-based localization*", R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 618–626.

"*Deep k-nearest neighbors: Towards confident, interpretable and robust deep learning*", N. Papernot, P. McDaniel, (2018). arXiv:1803.04765.

"*Interpretable convolutional neural networks*", Q. Zhang, Y. Nian Wu, S.-C. Zhu, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8827–8836.

"*InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets*", X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, P. Abbeel, (2016). arXiv:1606.03657

"*Not just a black box: Learning important features through propagating activation differences*", Shrikumar, P. Greenside, A. Shcherbina, A. Kundaje, 2016, arXiv:1605.01713

"*Interpretable explanations of black boxes by meaningful perturbation*", R. C. Fong, A. Vedaldi, in: IEEE International Conference on Computer Vision, 2017, pp. 3429–3437.

"*On the Robustness of Interpretability Methods*", Alvarez-Melis et T. S. Jaakkola, arXiv:1806.08049 [cs, stat], juin 2018.

"*Interpretable Deep Learning under Fire*", X. Zhang, N. Wang, H. Shen, S. Ji, X. Luo, et T. Wang, arXiv:1812.00891 [cs], sept. 2019.

"*Improving the Adversarial Robustness and Interpretability of Deep Neural Networks by Regularizing their Input Gradients*", S. Ross et F. Doshi-Velez, arXiv:1711.09404 [cs], nov. 2017.

"*Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR*", Wachter, B. Mittelstadt, et C. Russell, SSRN Journal, 2017.

... et beaucoup encore ...

Laurent Risser (CNRS, IMT, 3IA ANITI)

Partie 3.3 : Explicabilité des décisions par réseaux de neurones

→ Grad-CAM

Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization

**Ramprasaath R. Selvaraju · Michael Cogswell · Abhishek Das · Ramakrishna
Vedantam · Devi Parikh · Dhruv Batra**

Georgia Institute of Technology, Atlanta, GA, USA
Facebook AI Research, Menlo Park, CA, USA

<https://arxiv.org/pdf/1610.02391.pdf>
<http://gradcam.cloudcv.org/>
<https://github.com/ramprs/grad-cam/>

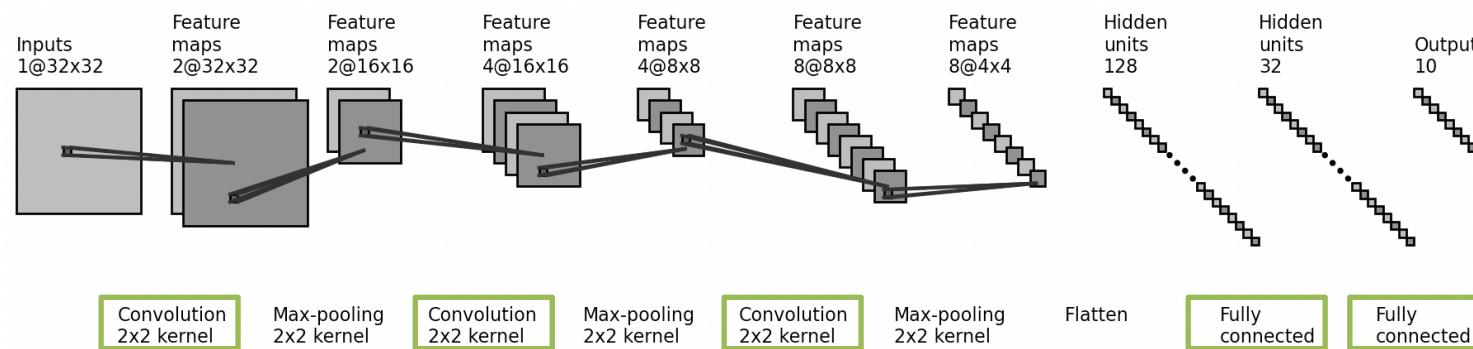
Spécialisé pour des modèles de réseaux de neurones convolutifs (VGG, ResNet, ...)

Partie 3.3 : Explicabilité des décisions par réseaux de neurones

→ Grad-CAM

Pour comprendre Grad-CAM, revenons rapidement sur comment un N.N. est entraîné

- Observations d'apprentissage : $\{(X_i, Y_i)\}_{i=1,\dots,n}$
- $\hat{\theta} = \arg \min_{\theta} R(f_{\theta}, \{(X_i, Y_i)\}_{i=1,\dots,n}) = \arg \min_{\theta} \frac{1}{n} \sum_{i=1}^n loss(f_{\theta}(X_i), Y_i)$
- Descente de gradient : $\theta_{it+1} = \theta_{it} - \lambda \nabla_{\theta} R(f_{\theta_{it}}, \{(X_i, Y_i)\}_{i=1,\dots,n})$ (on omet les aspects stochastiques)



Paramètres θ

Partie 3.3 : Explicabilité des décisions par réseaux de neurones

→ Grad-CAM

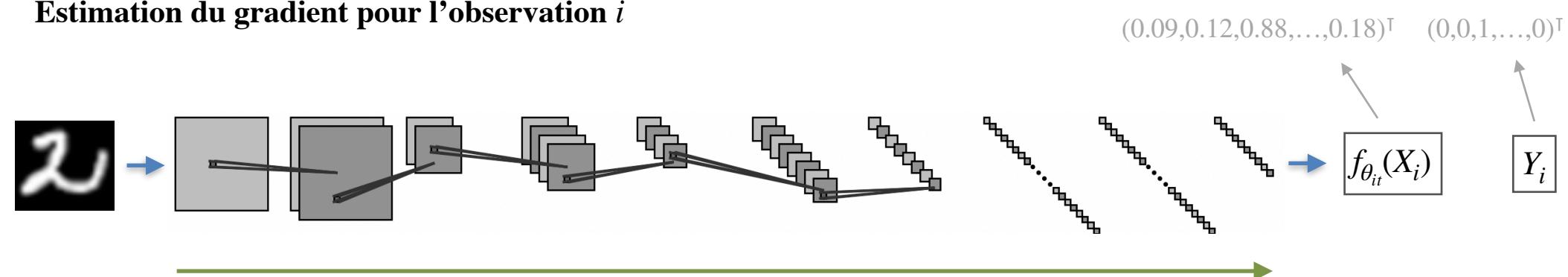
Pour comprendre Grad-CAM, revenons rapidement sur comment un N.N. est entraîné

- Observations d'apprentissage : $\{(X_i, Y_i)\}_{i=1,\dots,n}$

- $\hat{\theta} = \arg \min_{\theta} R(f_{\theta}, \{(X_i, Y_i)\}_{i=1,\dots,n}) = \arg \min_{\theta} \frac{1}{n} \sum_{i=1}^n loss(f_{\theta}(X_i), Y_i)$

- Descente de gradient : $\theta_{it+1} = \theta_{it} - \lambda \nabla_{\theta} R(f_{\theta_{it}}, \{(X_i, Y_i)\}_{i=1,\dots,n})$ (on omet les aspects stochastiques)

Estimation du gradient pour l'observation i



Etape 1: calcule $f_{\theta_{it}}(X_i)$

Partie 3.3 : Explicabilité des décisions par réseaux de neurones

→ Grad-CAM

Pour comprendre Grad-CAM, revenons rapidement sur comment un N.N. est entraîné

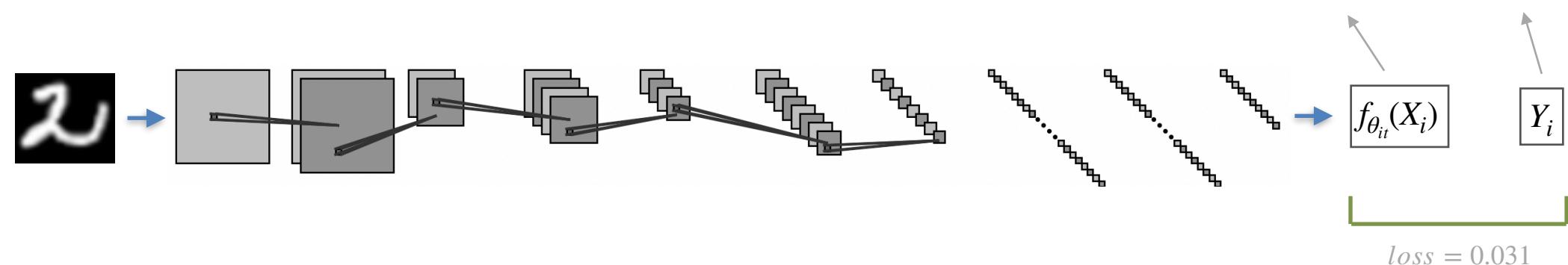
- Observations d'apprentissage : $\{(X_i, Y_i)\}_{i=1,\dots,n}$

- $\hat{\theta} = \arg \min_{\theta} R(f_{\theta}, \{(X_i, Y_i)\}_{i=1,\dots,n}) = \arg \min_{\theta} \frac{1}{n} \sum_{i=1}^n loss(f_{\theta}(X_i), Y_i)$

- Descente de gradient : $\theta_{it+1} = \theta_{it} - \lambda \nabla_{\theta} R(f_{\theta_{it}}, \{(X_i, Y_i)\}_{i=1,\dots,n})$ (on omet les aspects stochastiques)

Estimation du gradient pour l'observation i

$$(0.09, 0.12, 0.88, \dots, 0.18)^T \quad (0, 0, 1, \dots, 0)^T$$



Etape 2: calcule $loss(f_{\theta_{it}}(X_i), Y_i)$

Partie 3.3 : Explicabilité des décisions par réseaux de neurones

→ Grad-CAM

Pour comprendre Grad-CAM, revenons rapidement sur comment un N.N. est entraîné

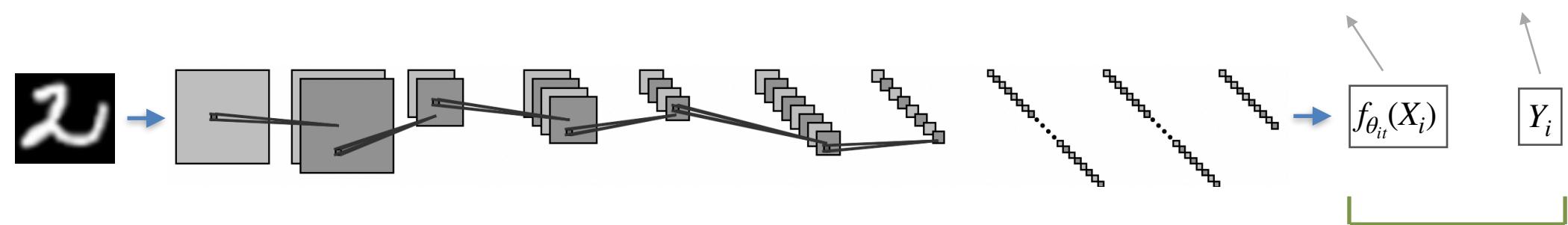
- Observations d'apprentissage : $\{(X_i, Y_i)\}_{i=1,\dots,n}$

- $\hat{\theta} = \arg \min_{\theta} R(f_{\theta}, \{(X_i, Y_i)\}_{i=1,\dots,n}) = \arg \min_{\theta} \frac{1}{n} \sum_{i=1}^n loss(f_{\theta}(X_i), Y_i)$

- Descente de gradient : $\theta_{it+1} = \theta_{it} - \lambda \nabla_{\theta} R(f_{\theta_{it}}, \{(X_i, Y_i)\}_{i=1,\dots,n})$ (on omet les aspects stochastiques)

Estimation du gradient pour l'observation i

$$(0.09, 0.12, 0.88, \dots, 0.18)^{\top} \quad (0, 0, 1, \dots, 0)^{\top}$$



Etape 3: calcule $\nabla_{f_{\theta_{it}}} loss(f_{\theta_{it}}(X_i), Y_i)$, c'est à dire $\frac{\partial loss(f_{\theta_{it}}(X_i), Y_i)}{\partial f_{\theta_{it}}^c(X_i)}, c = 0, 1, \dots, 9$

Partie 3.3 : Explicabilité des décisions par réseaux de neurones

→ Grad-CAM

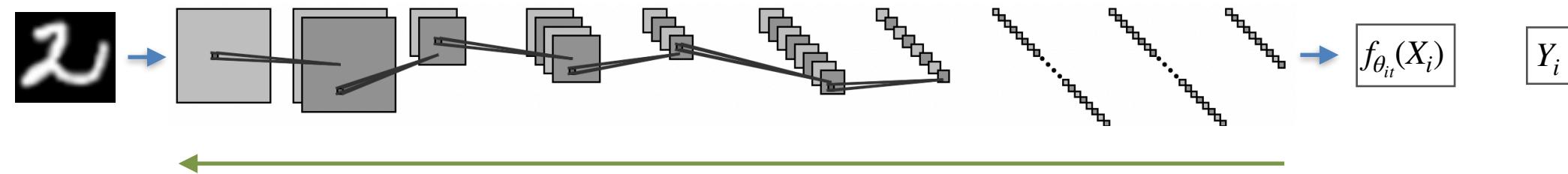
Pour comprendre Grad-CAM, revenons rapidement sur comment un N.N. est entraîné

- Observations d'apprentissage : $\{(X_i, Y_i)\}_{i=1,\dots,n}$

- $\hat{\theta} = \arg \min_{\theta} R(f_{\theta}, \{(X_i, Y_i)\}_{i=1,\dots,n}) = \arg \min_{\theta} \frac{1}{n} \sum_{i=1}^n loss(f_{\theta}(X_i), Y_i)$

- Descente de gradient : $\theta_{it+1} = \theta_{it} - \lambda \nabla_{\theta} R(f_{\theta_{it}}, \{(X_i, Y_i)\}_{i=1,\dots,n})$ (on omet les aspects stochastiques)

Estimation du gradient pour l'observation i



Etape 4: rétropropage les $\frac{\partial loss(\dots)}{\partial f_{\theta_{it}}^c(X_i)}$ pour calculer $\nabla_{\theta} loss(\dots) = \left(\frac{\partial loss(\dots)}{\partial \theta_{it}^1}, \dots, \frac{\partial loss(\dots)}{\partial \theta_{it}^d} \right)^T$

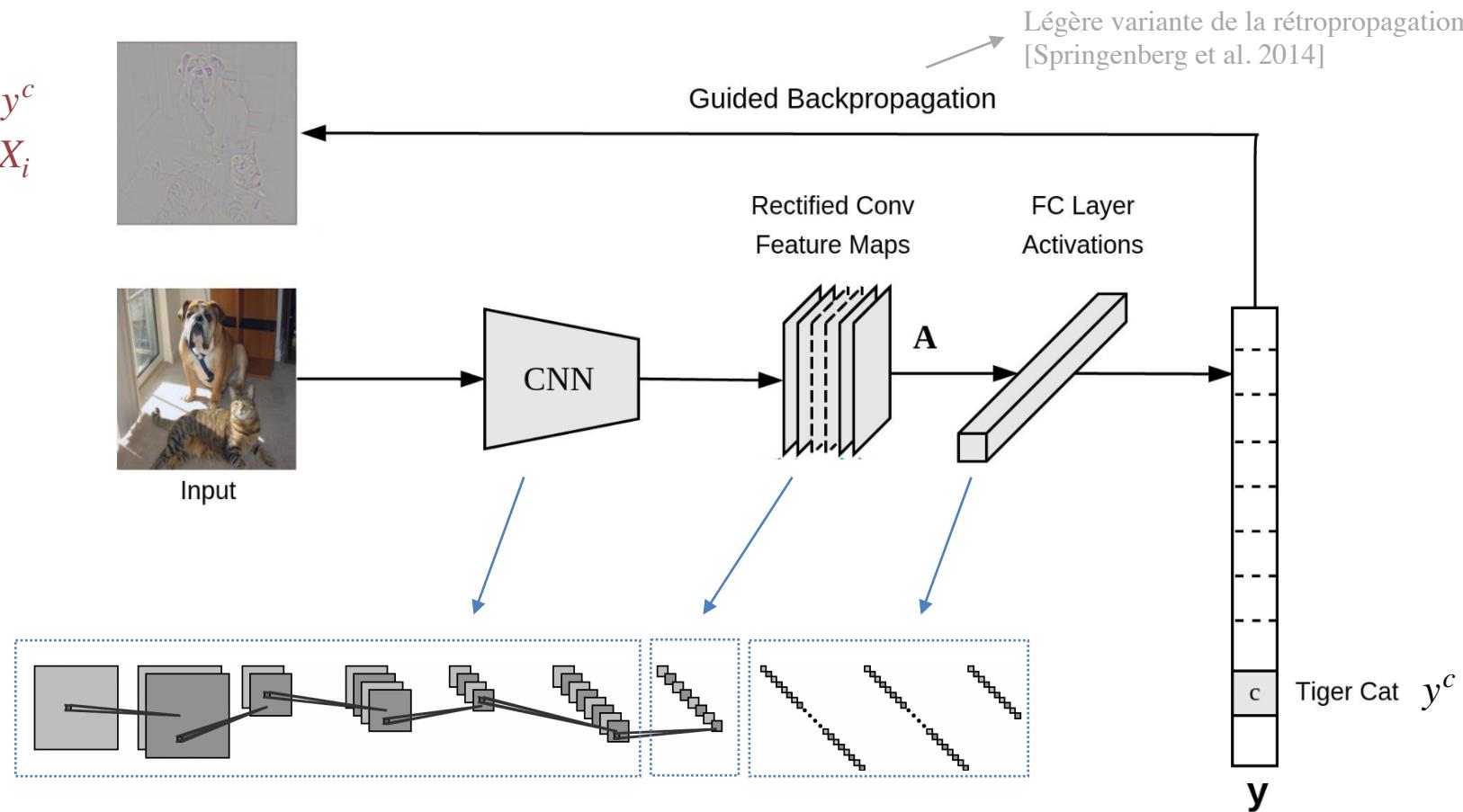
Partie 3.3 : Explicabilité des décisions par réseaux de neurones

→ Grad-CAM

Au lieu de rétropropager le gradient du loss par rapport aux $f_{\theta_{it}}^c(X_i)$

→ rétropropage un *faux* gradient, qui est égal à 1 pour une classe c et égal à 0 ailleurs

Sensibilité de y^c
aux pixels de X_i



Partie 3.3 : Explicabilité des décisions par réseaux de neurones

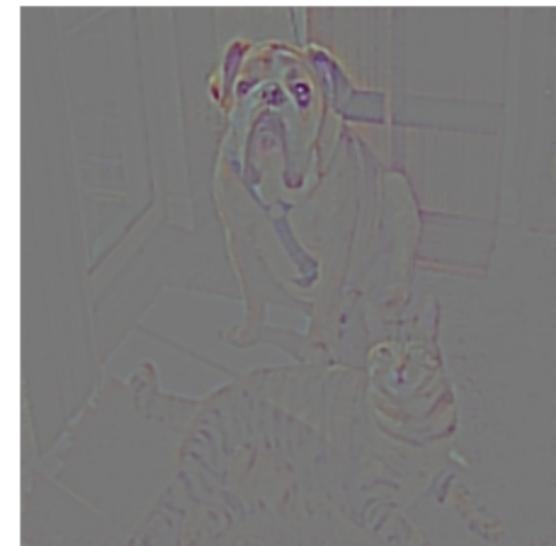
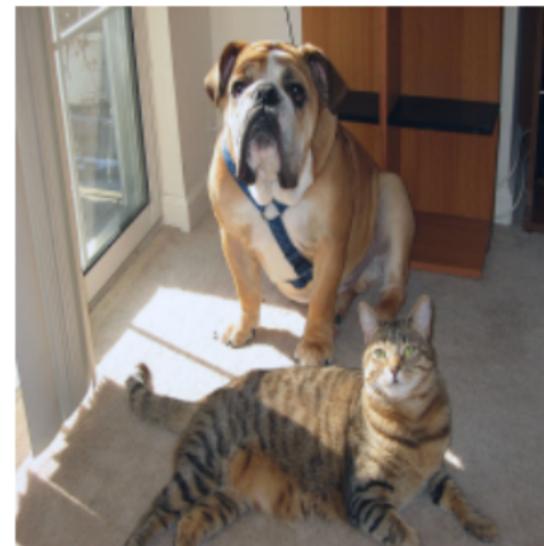
→ Grad-CAM

Résultat partiel

GB for “Cat”



GB for “Dog”



Pas si top... mais bon départ ! → Très fin d'un point de vue résolution mais peu discriminant

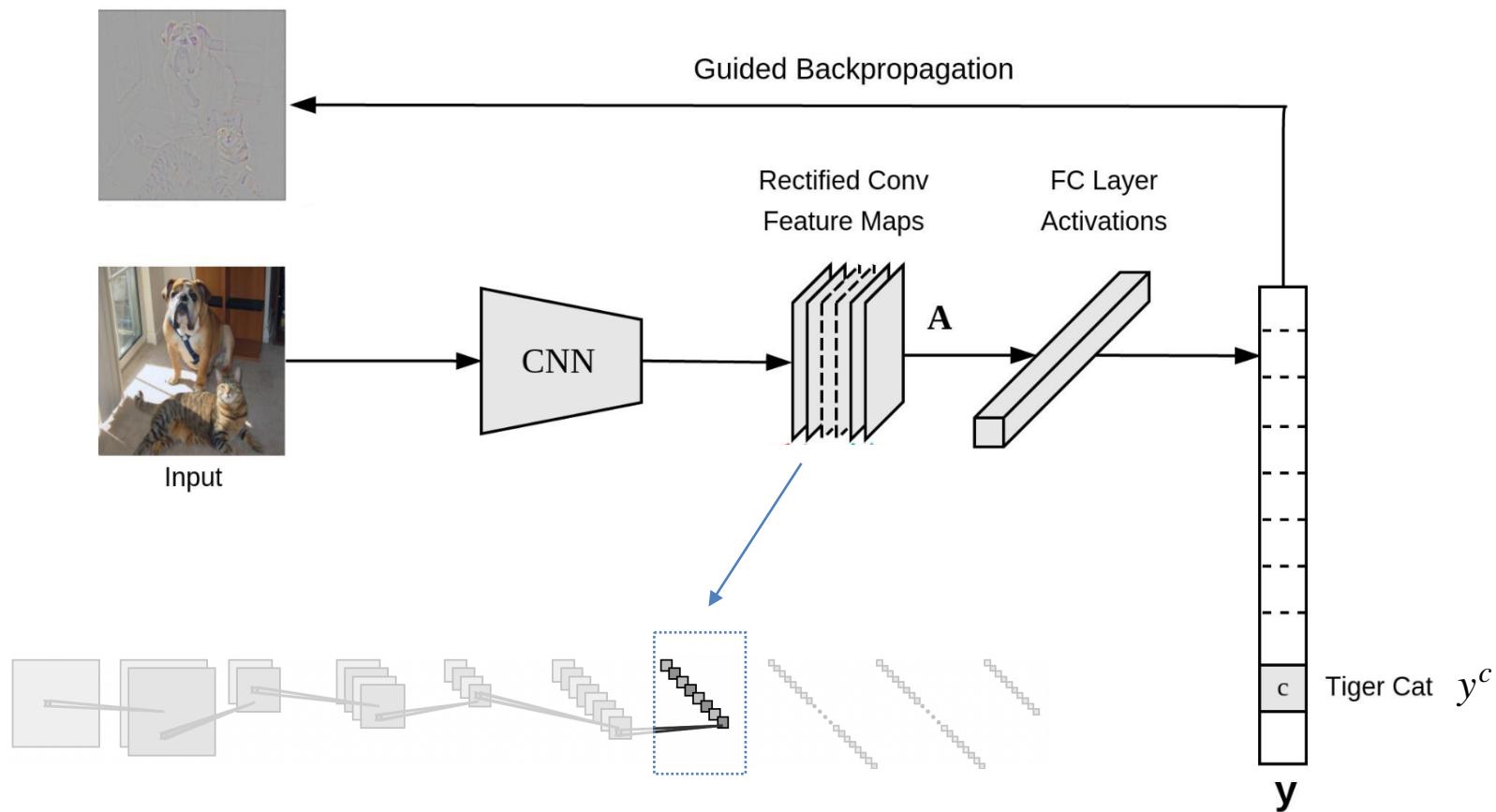
Grad-CAM rajoute ensuite un masque à ce résultat

Partie 3.3 : Explicabilité des décisions par réseaux de neurones

→ Grad-CAM

On s'intéresse aux « Rectified Convolution Feature Maps » $A_{i,j}^k$

(où k est un *canal* représentant une caractéristique de l'image (*feature*) et les (i, j) sont les pixels de ce canal)

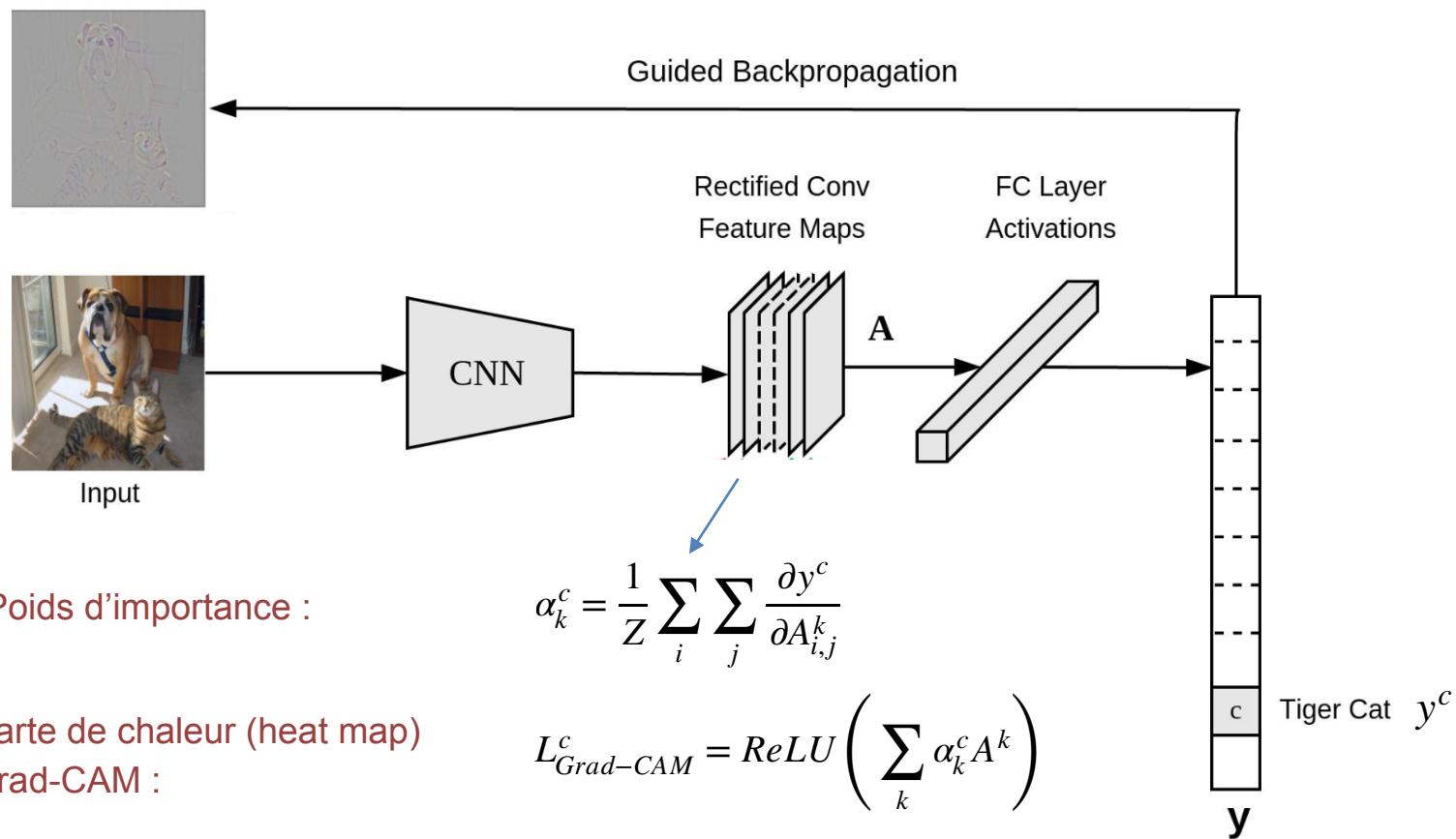


Partie 3.3 : Explicabilité des décisions par réseaux de neurones

→ Grad-CAM

On s'intéresse aux « Rectified Convolution Feature Maps » $A_{i,j}^k$

(où k est un *canal* représentant une caractéristique de l'image (*feature*) et les (i, j) sont les pixels de ce canal)

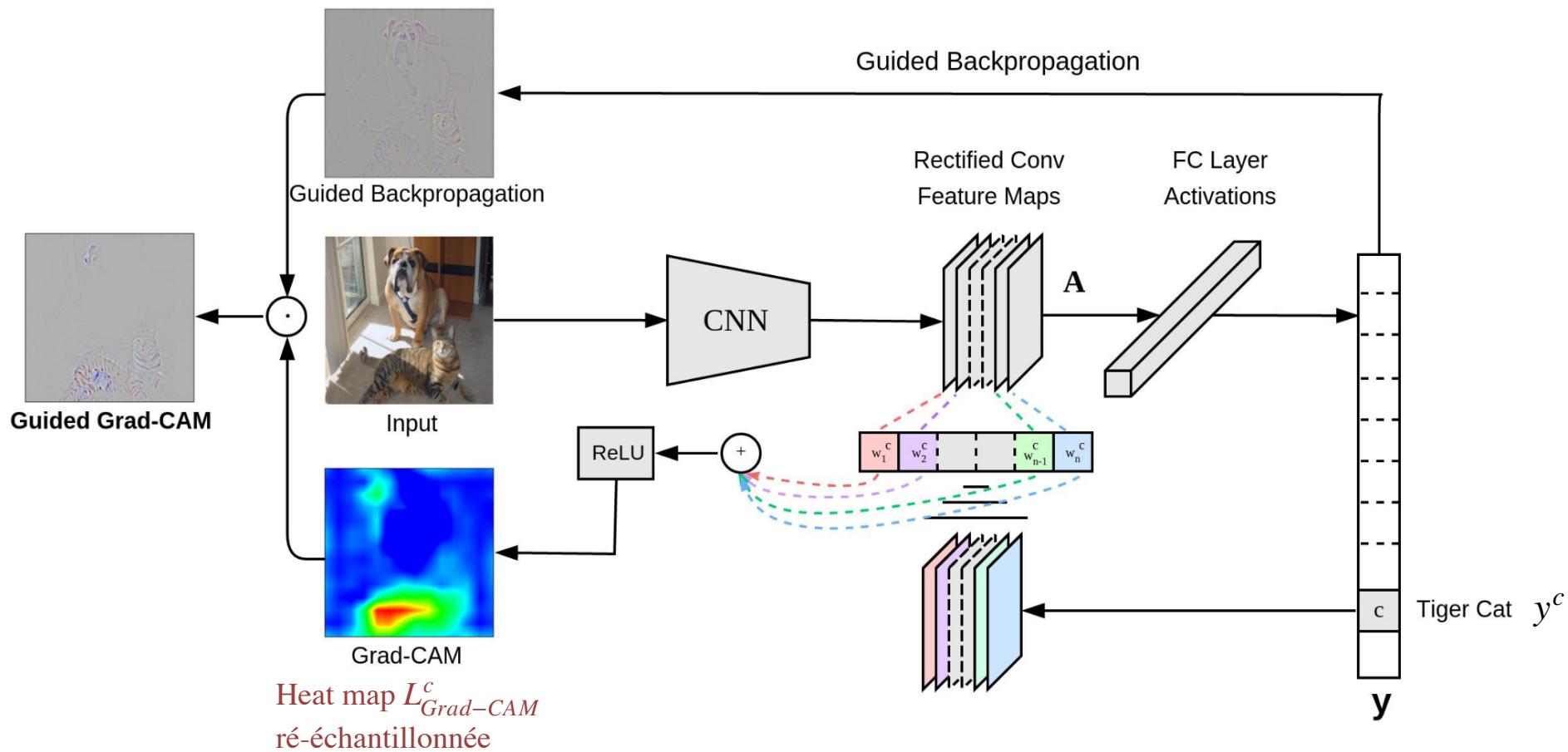


Partie 3.3 : Explicabilité des décisions par réseaux de neurones

→ Grad-CAM

On s'intéresse aux « Rectified Convolution Feature Maps » $A_{i,j}^k$

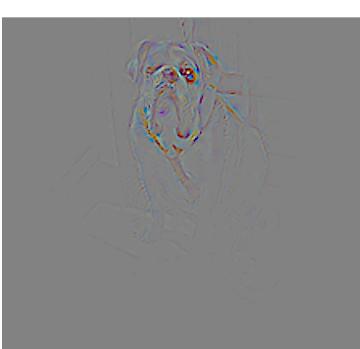
(où k est un *canal* représentant une caractéristique de l'image (*feature*) et les (i, j) sont les pixels de ce canal)



Partie 3.3 : Explicabilité des décisions par réseaux de neurones

→ Grad-CAM

Résultat final

Predicted class	#1 boxer	#2 bull mastiff	#3 tiger cat
Grad-CAM [1]			
Guided backpropagation [2]			
Guided Grad-CAM [1]			

Partie 3.3 : Explicabilité des décisions par réseaux de neurones

→ LIME (Local Interpretable model-agnostic explanations)

“Why Should I Trust You?” Explaining the Predictions of Any Classifier

Marco Tulio Ribeiro
University of Washington
Seattle, WA 98105, USA
marcotcr@cs.uw.edu

Sameer Singh
University of Washington
Seattle, WA 98105, USA
sameer@cs.uw.edu

Carlos Guestrin
University of Washington
Seattle, WA 98105, USA
guestrin@cs.uw.edu

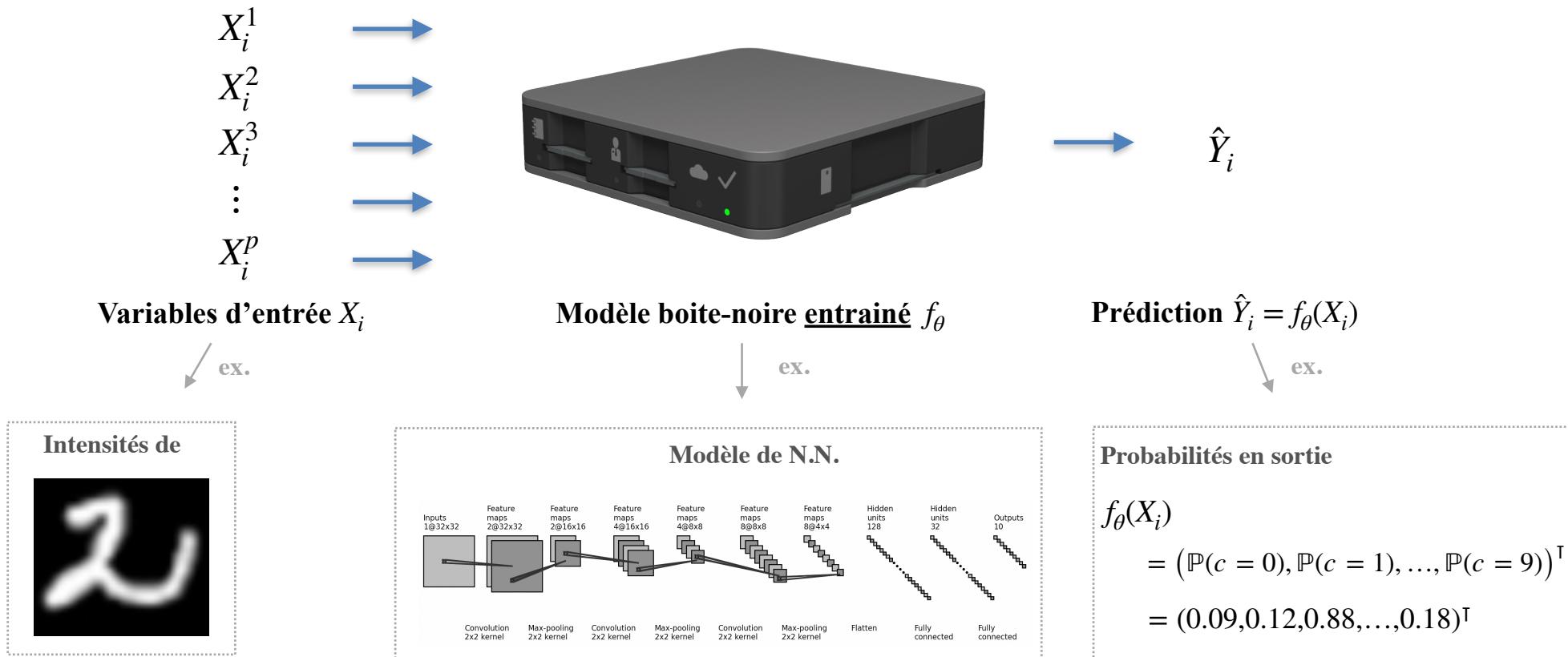
<https://arxiv.org/pdf/1602.04938.pdf>
<https://homes.cs.washington.edu/~marcotcr/blog/lime/>
<https://github.com/marcotcr/lime>

Adapté à n'importe quel modèle !

Partie 3.3 : Explicabilité des décisions par réseaux de neurones

→ LIME (Local Interpretable model-agnostic explanations)

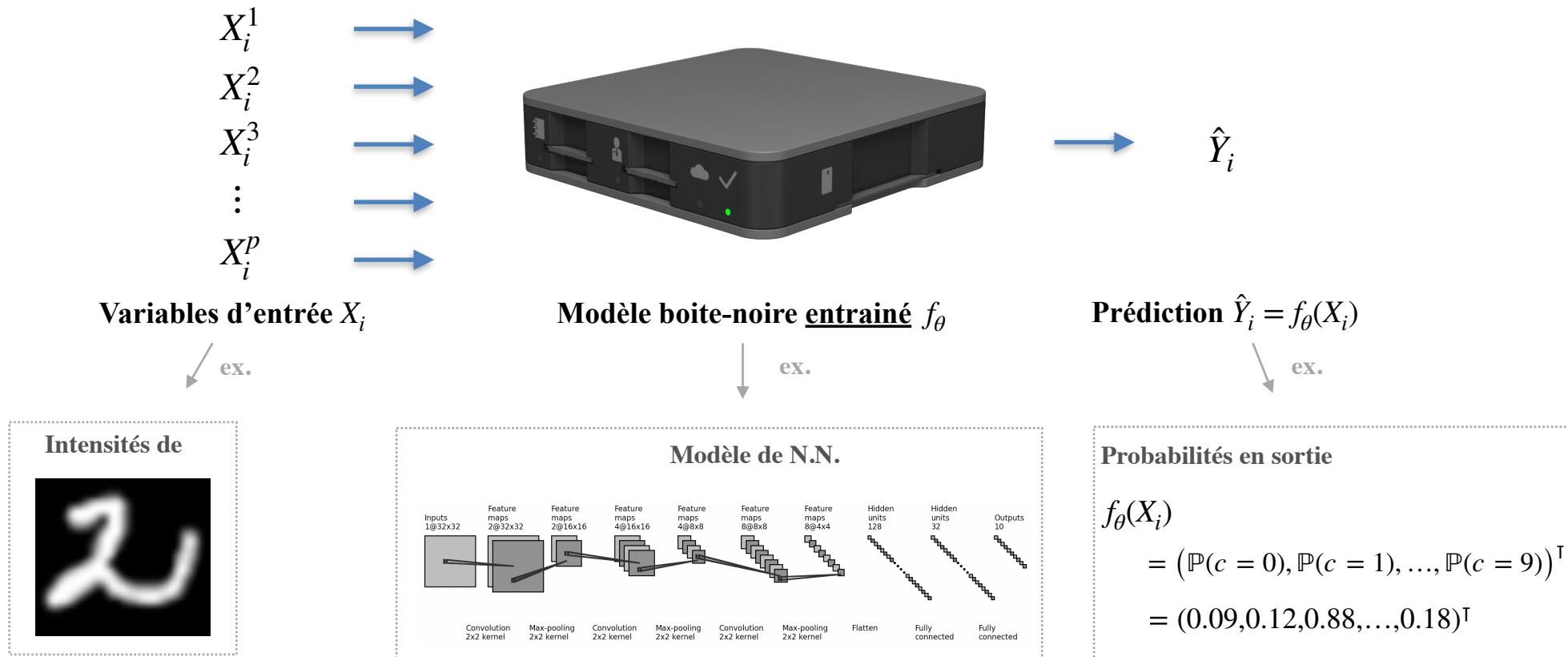
Explique les prédictions *locales* (*similaires* à X_i) à l'aide d'un modèle de substitution $g_{\theta'}$ interprétable



Partie 3.3 : Explicabilité des décisions par réseaux de neurones

→ LIME (Local Interpretable model-agnostic explanations)

Explique les prédictions *locales* (*similaires* à X_i) à l'aide d'un modèle de substitution $g_{\theta'}$ interprétable

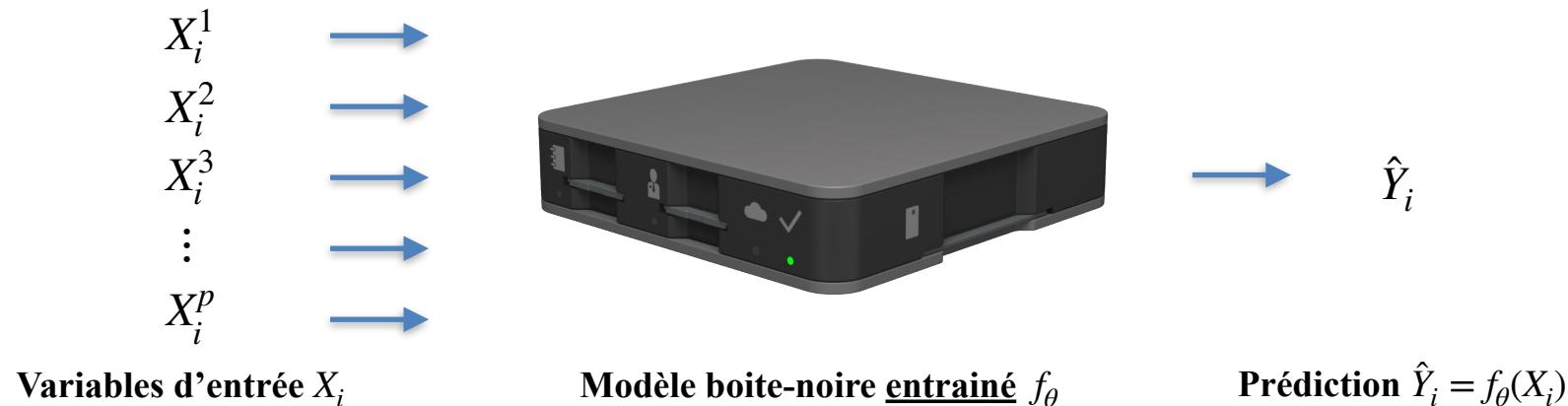


Exemple de modèle de substitution autour de X_i : $g_{\theta'} = \left(\sum_{j=1}^p \theta^{j,0} X_i^j, \sum_{j=1}^p \theta^{j,1} X_i^j, \dots, \sum_{j=1}^p \theta^{j,9} X_i^j \right)^T$

Partie 3.3 : Explicabilité des décisions par réseaux de neurones

→ LIME (Local Interpretable model-agnostic explanations)

Explique les prédictions *locales* (*similaires à X_i*) à l'aide d'un modèle de substitution $g_{\theta'}$ interprétable



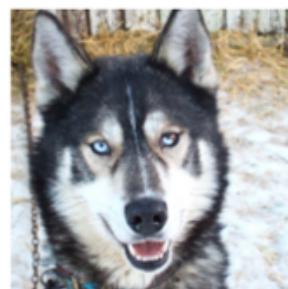
Apprentissage du modèle $g_{\theta'}$ pour expliquer la prédiction de X_i avec f_θ :

- Défini un modèle explicite $g_{\theta'}$
- Défini une distance entre X_i et des données similaires X_i^r : $\pi_{X_i}(X_i^r) = dist(X_i, X_i^r)$.
- Perturbe aléatoirement $X_i \rightarrow \{X_i^r\}_{p=1, \dots, R}$
- Optimise les paramètres θ' en minimisant : $\sum_{r=1}^R \pi_{X_i}(X_i^r)(g_{\theta'}(X_i^r) - f_\theta(X_i^r))^2$
- Explique la prédiction $f_\theta(X_i)$ à partir des paramètres optimums θ'

Partie 3.3 : Explicabilité des décisions par réseaux de neurones

→ LIME (Local Interpretable model-agnostic explanations)

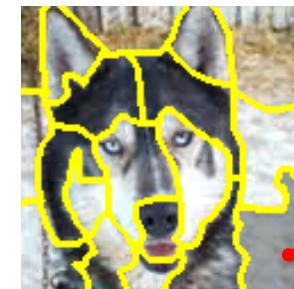
Explique les prédictions *locales* (*similaires à X_i*) à l'aide d'un modèle de substitution $g_{\theta'}$ interprétable



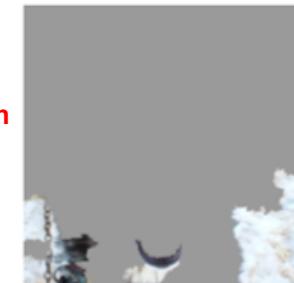
Husky classé comme Loup



(1)
Segmentation
de l'image



(2) Perturbation des intensités dans chaque région et
(3) Analyse de la sensibilité des prédictions à ces perturbations



(4) Explication : régions pour lesquelles les prédictions sont le plus impacté par des perturbations

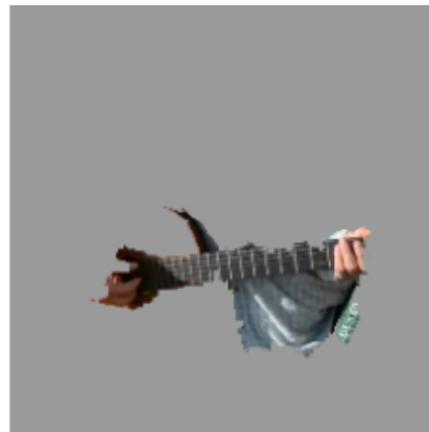
Partie 3.3 : Explicabilité des décisions par réseaux de neurones

→ LIME (Local Interpretable model-agnostic explanations)

Résultats



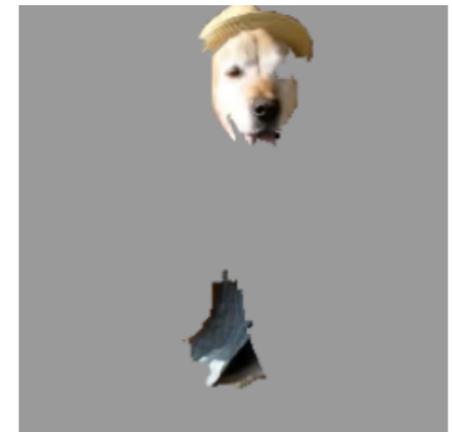
(a) Original Image



(b) Explaining *Electric guitar*



(c) Explaining *Acoustic guitar*



(d) Explaining *Labrador*

Figure 4: Explaining an image classification prediction made by Google's Inception neural network. The top 3 classes predicted are "Electric Guitar" ($p = 0.32$), "Acoustic guitar" ($p = 0.24$) and "Labrador" ($p = 0.21$)

Pour conclure

- Les décisions d'algorithmes d'I.A. peuvent être biaisés de manière purement non-intentionnelle
- Il est possible de détecter et mesurer statistiquement ces biais
- Des techniques commencent à être matures pour expliquer les décisions
- La communauté scientifique travaille activement sur des méthodes de corrections de ces biais

15



**AI, Law, Society
and Ethics**

Questions ?



Et l'IA,
vous en pensez
quoi ?

Next, on Fidle :



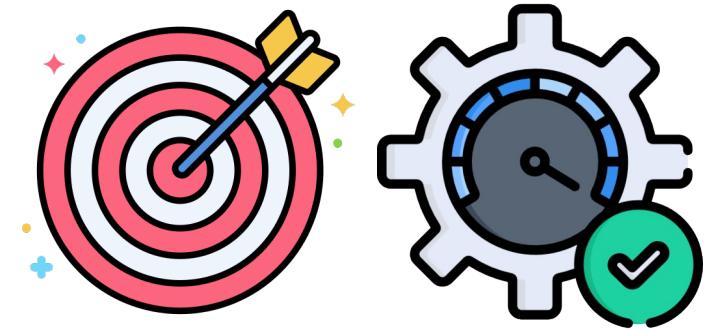
Jeudi 30 mars,

Épisode 16 :

Apprendre plus vite et moins cher, optimiser l'apprentissage

- Techniques d'optimisation
- Affinage des hyperparamètres
- Quand le recyclage est une force : Le transfert learning
- Pour un poids optimal des données, utiliser le bon format de données

Durée : 2h



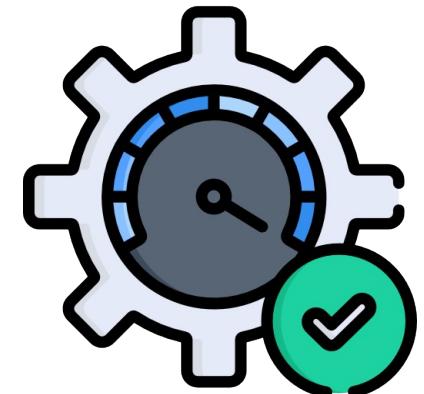
Next on Fidle :



Jeudi 30 mars,

Séquence 16 :

**Apprendre plus vite et moins cher,
optimiser l'apprentissage**



To be continued...