Explainability and Robustness in Machine Learning

Laurent Risser
CNRS research engineer at *Institut de Mathématiques de Toulouse* and *3IA ANITI*

lrisser@math.univ-toulouse.fr

1. Why neural-networks have become so popular in A.I.?
2. Why should we be cautious with algorithmic bias?
3. How the E.U. starts regulating A.I.?
4. Can we measure, explain or control algorithmic biases?

1. Why neural-networks have become so popular in A.I.?
2. Why should we be cautious with algorithmic bias?
3. How the E.U. starts regulating A.I.?
4. Can we measure, explain or control algorithmic biases?

ALGORITHMIC BIAS IN ARTIFICIAL INTELLIGENCE

1 — Why neural-networks have become so popular in A.I.?

RISE OF NEURAL NETWORKS MODELS FOR AUTOMATIC PREDICTIONS

Affordable access to massive computational ressources

+

Explosion of acquired quantified data

=

High potential for innovative solutions

ALGORITHMIC BIAS IN ARTIFICIAL INTELLIGENCE

1 — Why neural-networks have become so popular in A.I.?

RISE OF NEURAL NETWORKS MODELS FOR AUTOMATIC PREDICTIONS

Affordable access to massive computational ressources

\+

Explosion of acquired quantified data

Machine learning : Makes automatic predictions/decisions by mimicking the behaviours observed in reference data

Neural-Networks : Machine learning models that are particularly suited to treat complex data (images, texts, voice, …)

ALGORITHMIC BIAS IN ARTIFICIAL INTELLIGENCE

1 — Why neural-networks have become so popular in A.I.?

MAIN PRINCIPLES OF MACHINE LEARNING – DIAGNOSTIC AID EXAMPLE

Diagnostic aid

Training base	
Patient 1: • Age = 40 • White globule density = 6	Healthy
Patient 2: • Age = 28 • White globule density = 12	Diseased
⋮	⋮
Patient n: • Age = 57 • White globule density = 8	Healthy

New patient:
- Age = 35
- White globule density = 5

Diagnostic:
Healthy or Diseased ???

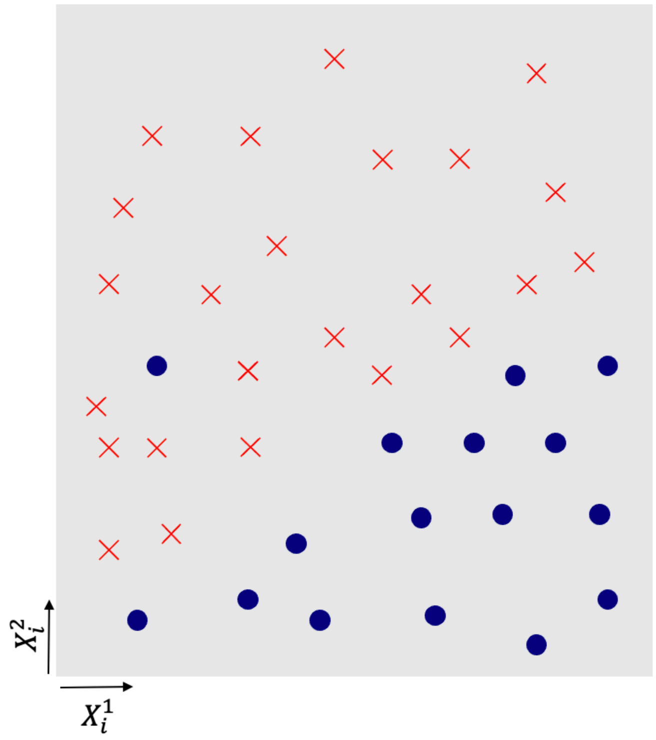
ALGORITHMIC BIAS IN ARTIFICIAL INTELLIGENCE

1 — Why neural-networks have become so popular in A.I.?

MAIN PRINCIPLES OF MACHINE LEARNING – DATA PREPARATION

Training base	
Patient 1: • Age = 40 • White globule density = 6	Healthy
Patient 2: • Age = 28 • White globule density = 12	Diseased
⋮	⋮
Patient n: • Age = 57 • White globule density = 8	Healthy

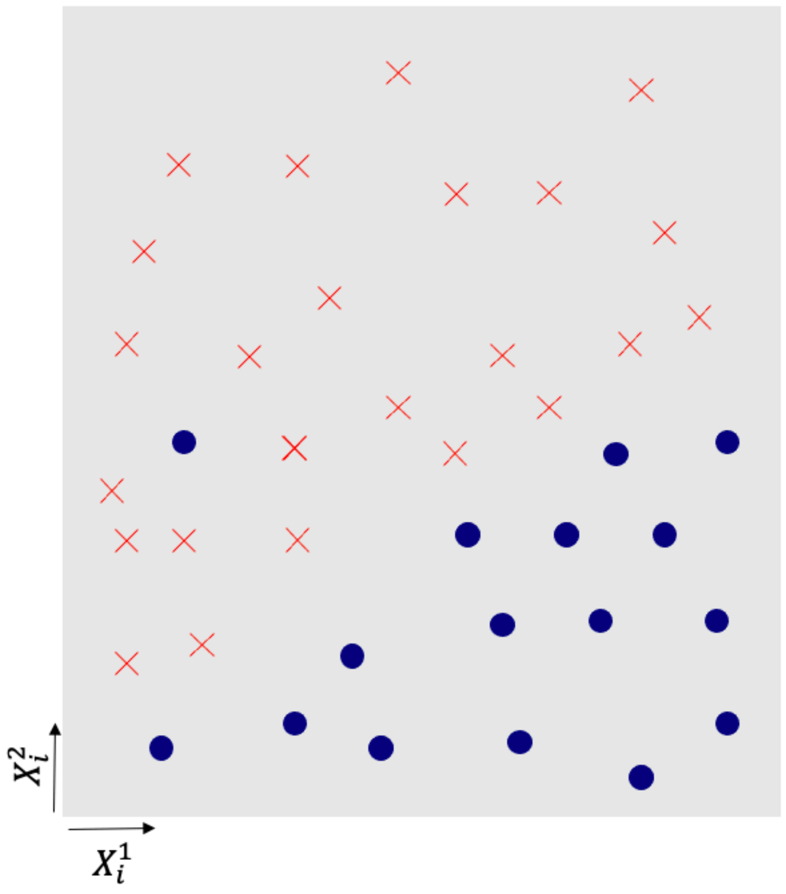
	Age	w.g.d.	State
Pat. 1	40	6	1
Pat. 2	28	12	0
	…	…	…
Pat. n	57	8	1



ALGORITHMIC BIAS IN ARTIFICIAL INTELLIGENCE

1 — Why neural-networks have become so popular in A.I.?

MAIN PRINCIPLES OF MACHINE LEARNING – DATA PREPARATION



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

(here *n=40* and *p=2)*

Output observations Y:
- n Labels $Y_i \in \{0,1\}$
- × $Y_i = 1$
- ● $Y_i = 0$

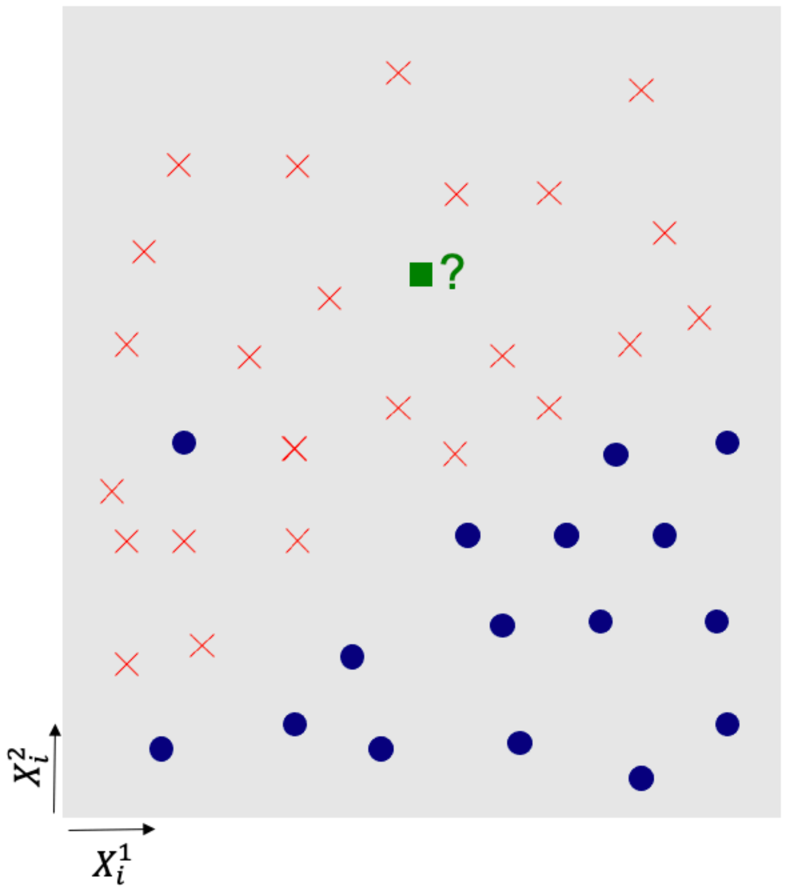
In the *diagnostic aid example*:

$i \longrightarrow$ Patient of the training base
$X_i^1 \longrightarrow$ Age
$X_i^2 \longrightarrow$ White globule density
$Y_i \longrightarrow$ Healthy or Diseased

ALGORITHMIC BIAS IN ARTIFICIAL INTELLIGENCE

1 — Why neural-networks have become so popular in A.I.?

MAIN PRINCIPLES OF MACHINE LEARNING – THE TRAINING/PREDICTION PRINCIPLE



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

(here *n=40* and *p=2)*

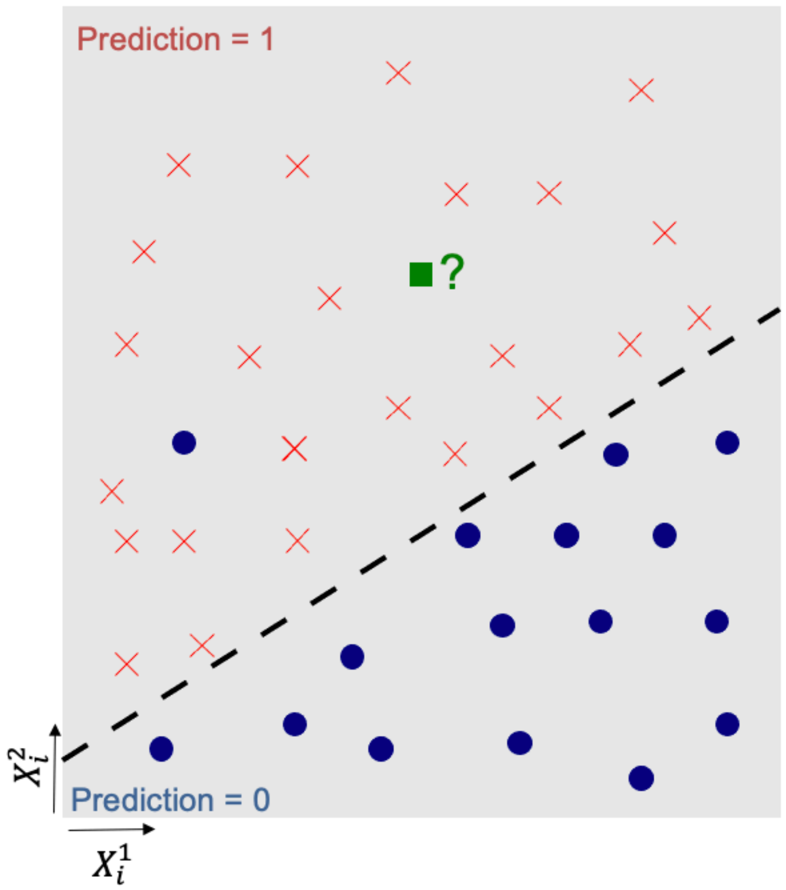
Output observations Y:
- n Labels $Y_i \in \{0,1\}$
- × $Y_i = 1$
- ● $Y_i = 0$

Most likely label for ■ ?

ALGORITHMIC BIAS IN ARTIFICIAL INTELLIGENCE

1 — Why neural-networks have become so popular in A.I.?

MAIN PRINCIPLES OF MACHINE LEARNING – THE TRAINING/PREDICTION PRINCIPLE



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

(here *n=40* and *p=2)*

Output observations Y:
- n Labels $Y_i \in \{0,1\}$
- × $Y_i = 1$
- ● $Y_i = 0$

1. **Choose a prediction model** to split the training data into the ● and the ×.
2. **Train** the optimal **parameters**.
3. Once the **prediction** model parameters trained, predicting the label of new observations like ■ is extremely simple and fast.

ALGORITHMIC BIAS IN ARTIFICIAL INTELLIGENCE

1 — Why neural-networks have become so popular in A.I.?

SUCCESS OF NEURAL-NETWORKS TO TREAT COMPLEX DATA

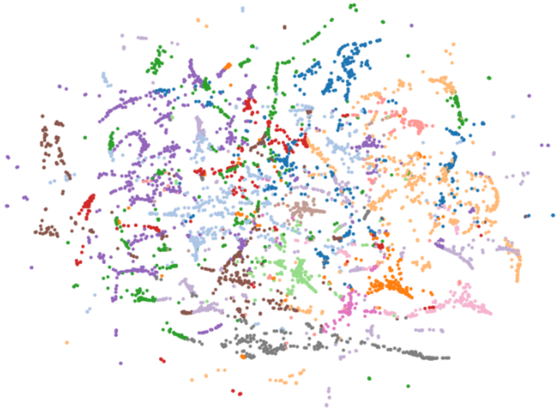
Example of the « Bios dataset », which was made public by linkedin/microsoft, to predict the job occupation using neural networks.

```
 ' Her areas of clinical expertise include arthritis, ba
ck injuries and shoulder disorders, among many others.D
r. Pichard-Encina obtained her undergraduate degree from
the University of Maryland in College Park. She complete
d her medical degree and orthopaedic surgery residency a
t Johns Hopkins. During her residency she was elected to
the American Orthopaedic Association resident leadership
forum.Her research interests include musculoskeletal edu
cation to non-orthopaedic surgery colleagues, as well as
conditions affecting the hand.Dr. Pichard-Encina was hon
ored to appear in the American Academy of Orthopaedic Su
rgery "Heroes" Public Service Announcement Campaign. She
is a member of several professional organizations, inclu
ding the American Academy of Orthopaedic Surgeons, the A
merican Orthopaedic Association and the Ruth Jackson Ort
hopaedic Society.']
```

Input data

(A biography on linkedin)

Data preparation (generally by using a generic pre-trained neural-network)



Optimal data representation

(embedding)

Prediction (using a specific neural-network)

"Surgeon"

Job recommendation

(out of a list of known jobs)

ALGORITHMIC BIAS IN ARTIFICIAL INTELLIGENCE

1 — Why neural-networks have become so popular in A.I.?

SUCCESS OF NEURAL-NETWORKS TO TREAT COMPLEX DATA

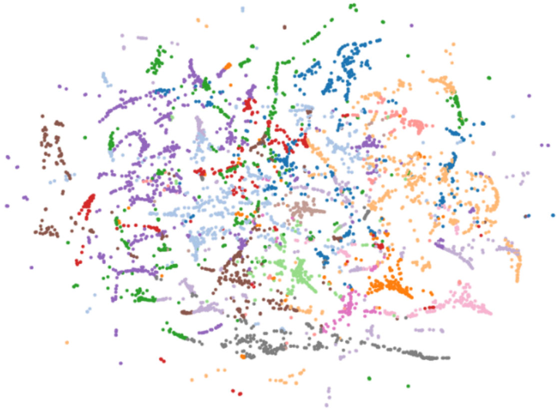
Example of the « Bios dataset », which was made public by linkedin/microsoft, to predict the job occupation using neural networks.

```
 ' Her areas of clinical expertise include arthritis, ba
ck injuries and shoulder disorders, among many others.D
r. Pichard-Encina obtained her undergraduate degree from
the University of Maryland in College Park. She complete
d her medical degree and orthopaedic surgery residency a
t Johns Hopkins. During her residency she was elected to
the American Orthopaedic Association resident leadership
forum.Her research interests include musculoskeletal edu
cation to non-orthopaedic surgery colleagues, as well as
conditions affecting the hand.Dr. Pichard-Encina was hon
ored to appear in the American Academy of Orthopaedic Su
rgery "Heroes" Public Service Announcement Campaign. She
is a member of several professional organizations, inclu
ding the American Academy of Orthopaedic Surgeons, the A
merican Orthopaedic Association and the Ruth Jackson Ort
hopaedic Society.']
```

Input data

(A biography on linkedin)

Data preparation (generally by using a generic pre-trained neural-network)



Optimal data representation

(embedding)

Prediction (using a specific neural-network)

"Surgeon"

Job recommendation

(out of a list of known jobs)

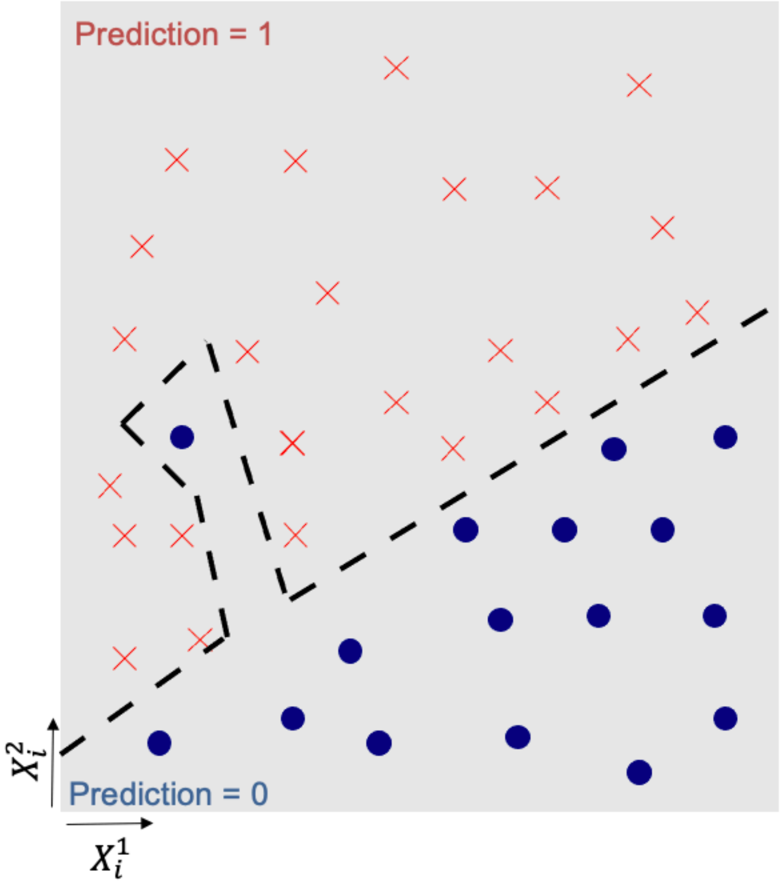
Mimics the recommendations made in a reference **training set** (here more than 400K recommendations)

1. Why neural-networks have become so popular in A.I.?
2. Why should we be cautious with algorithmic bias?
3. How the E.U. starts regulating A.I.?
4. Can we measure, explain or control algorithmic biases?

ALGORITHMIC BIAS IN ARTIFICIAL INTELLIGENCE

2 — Why should we be cautious with algorithmic bias?

CHOOSING A PREDICTION MODEL HAS AN IMPACT ON THE FUTURE PREDICTION ACCURACY



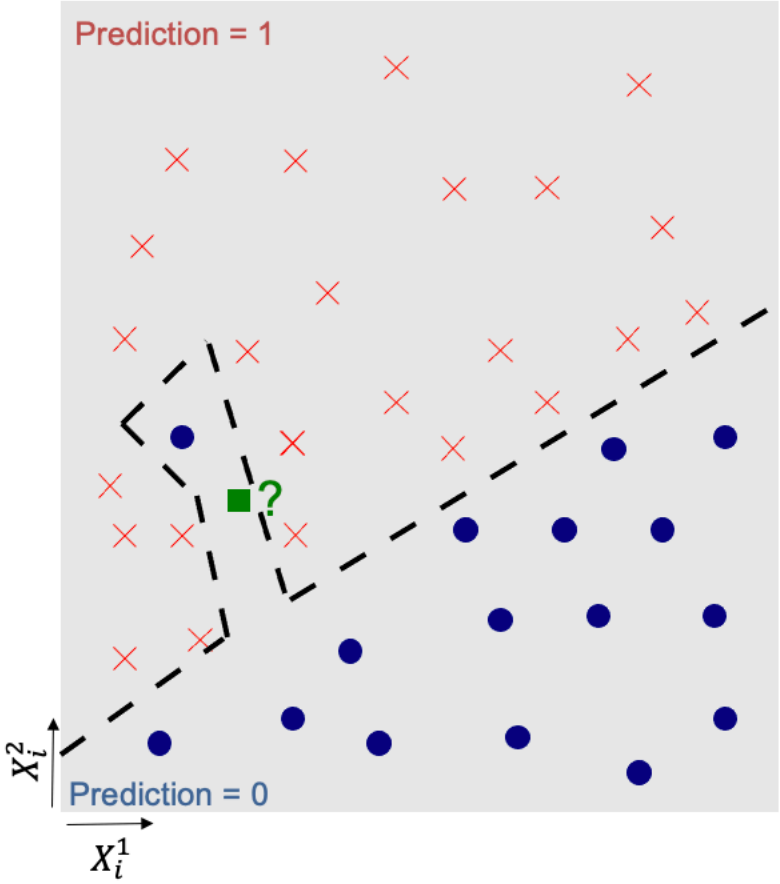
Suppose now that a training observation was improperly labelled!

→ We can use a more flexible model

ALGORITHMIC BIAS IN ARTIFICIAL INTELLIGENCE

2 — Why should we be cautious with algorithmic bias?

CHOOSING A PREDICTION MODEL HAS AN IMPACT ON THE FUTURE PREDICTION ACCURACY



Suppose now that a training observation was improperly labelled!

→ We can use a more flexible model

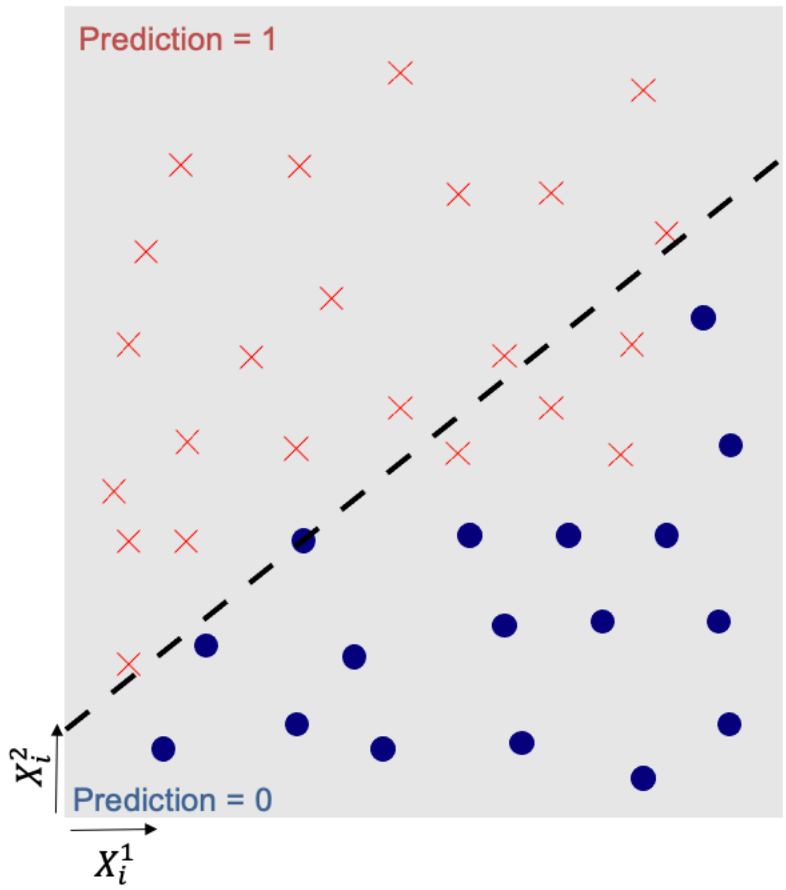
Poor generalisation here

Defining prediction models that are reasonably well constrained with regard to the data is very important for the data scientist!

ALGORITHMIC BIAS IN ARTIFICIAL INTELLIGENCE

2 — Why should we be cautious with algorithmic bias?

DISCRIMINATION BIAS MAY APPEAR, EVEN UNINTENTIONALLY

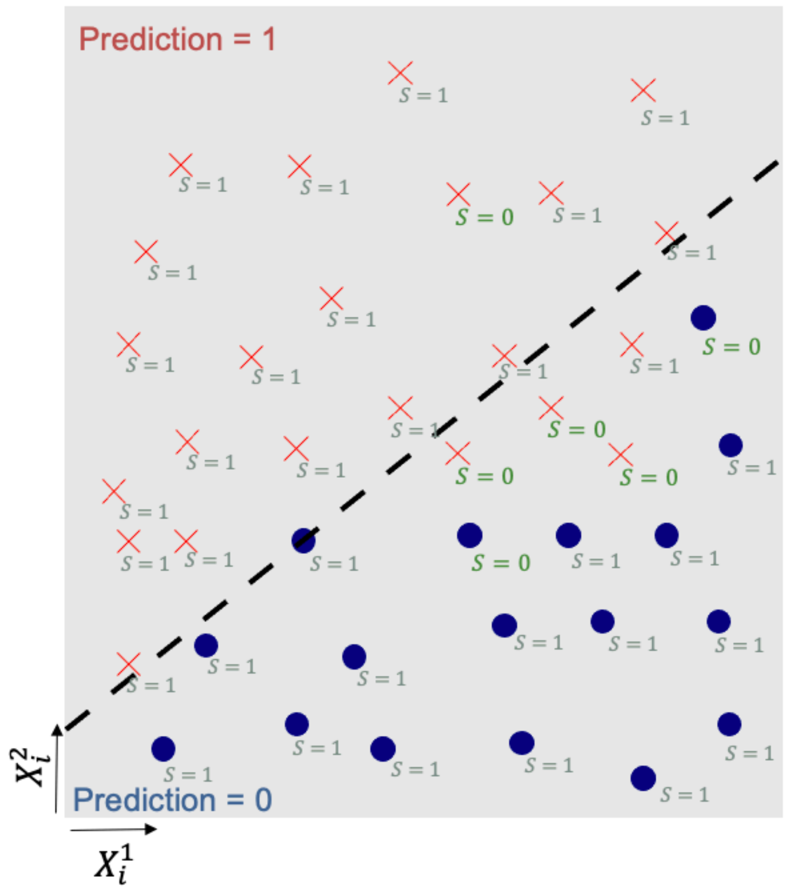


A linear model is used to split the data although it is not purely suited to their spatial distribution!

ALGORITHMIC BIAS IN ARTIFICIAL INTELLIGENCE

2 — Why should we be cautious with algorithmic bias?

DISCRIMINATION BIAS MAY APPEAR, EVEN UNINTENTIONALLY



A linear model is used to split the data although it is not purely suited to their spatial distribution!

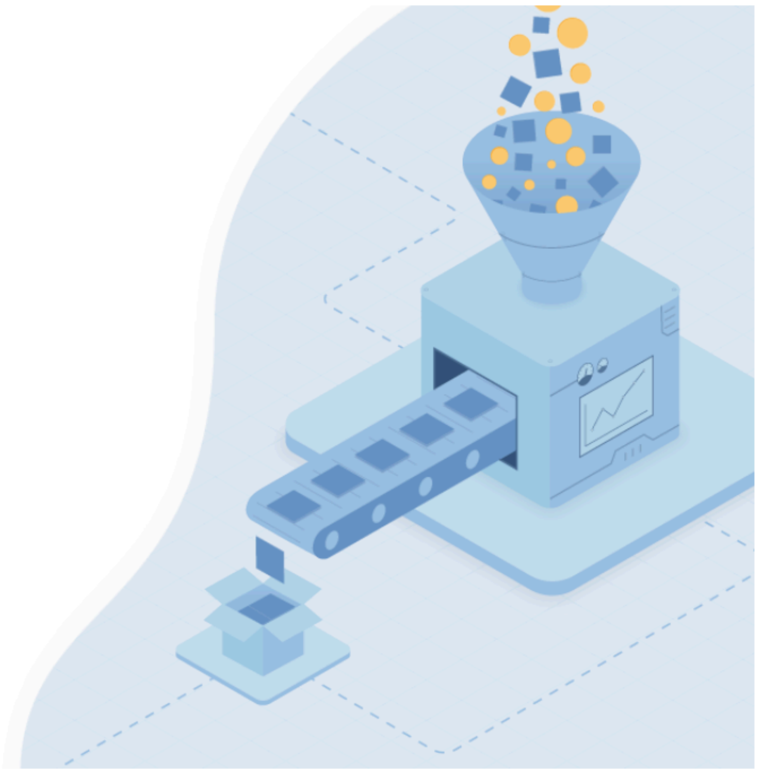
Now suppose that a group of subjects $S = 0$ (e.g. a geographic origin, a work context, a diet, …) is over-represented in the data with false predictions.

Unfair decisions although this is unintentional!

REMARK: THERE ARE VARIOUS POTENTIAL CAUSES FOR BIAS IN MACHINE LEARNING PREDICTORS

Main causes for bias in machine learning

- Poorly annotated data
- Unbalanced data
- Under- or over-fitting
- Confounding variables



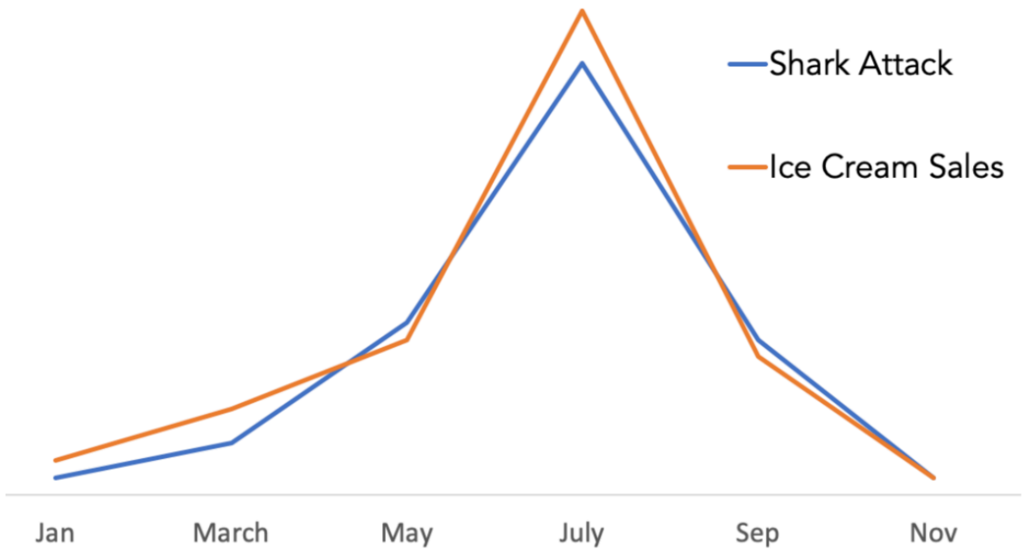
⇢ Most of them can be addressed by cautious data scientists

⇢ Confounding variables are those that are the trickiest ones to tackle and require a true expertise!

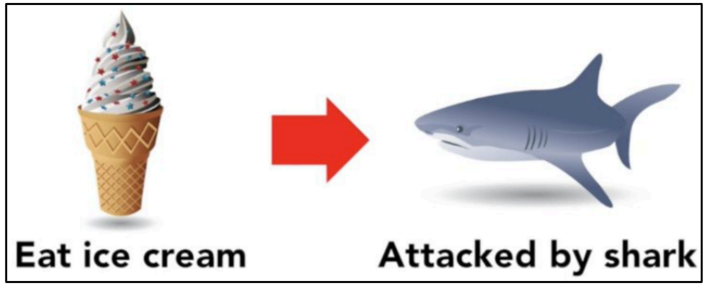
ALGORITHMIC BIAS IN ARTIFICIAL INTELLIGENCE

2 — Why should we be cautious with algorithmic bias?

CONFOUNDING VARIABLES – ICE CREAM AND SHARKS EXAMPLE



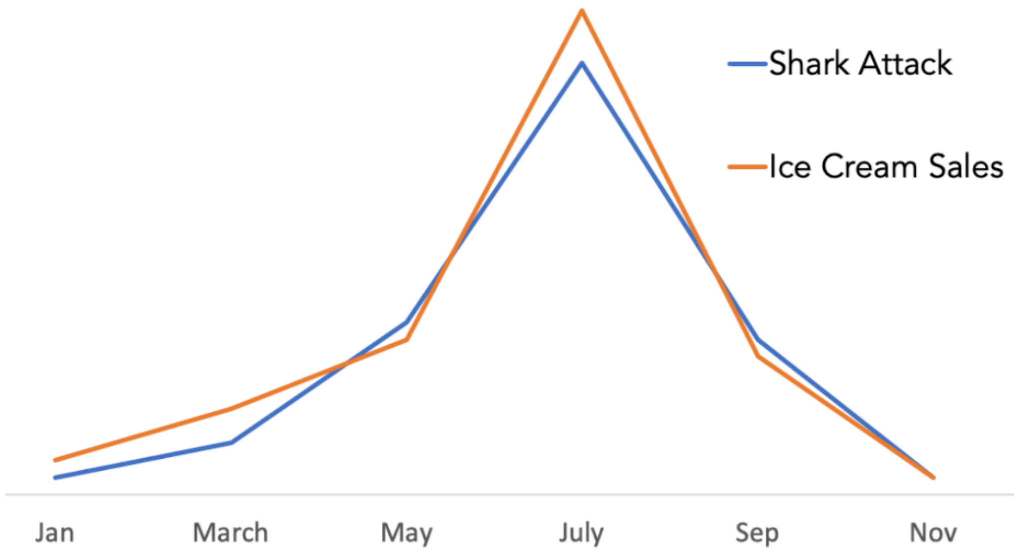
???



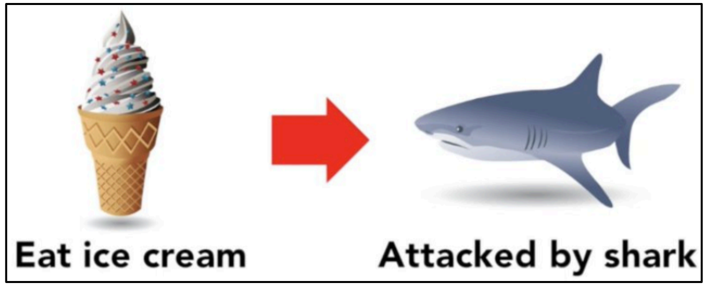
ALGORITHMIC BIAS IN ARTIFICIAL INTELLIGENCE

2 — Why should we be cautious with algorithmic bias?

CONFOUNDING VARIABLES – ICE CREAM AND SHARKS EXAMPLE



???



⇢ Correlation is <u>not</u> causality

⇢ Here the *hot temperature* is the confounding variable

ALGORITHMIC BIAS IN ARTIFICIAL INTELLIGENCE

2 — Why should we be cautious with algorithmic bias?

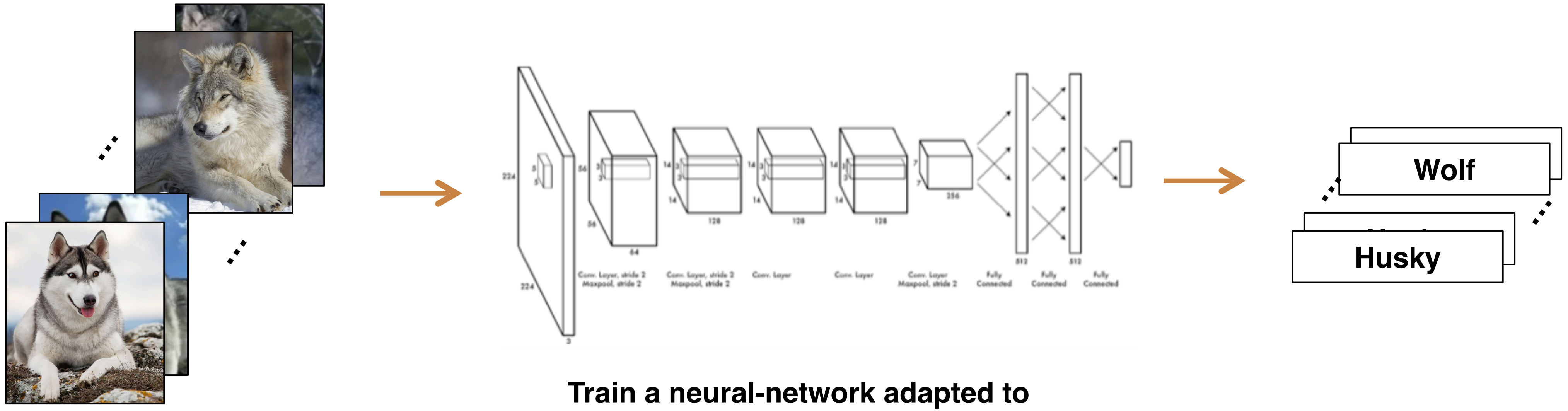
CONFOUNDING VARIABLES – HUSKIES AND WOLVES EXAMPLE (RIBEIRO ET AL, 2016)

Goal: Automatic recognition of a husky or a wolf based on a picture

ALGORITHMIC BIAS IN ARTIFICIAL INTELLIGENCE

2 — Why should we be cautious with algorithmic bias?

CONFOUNDING VARIABLES – HUSKIES AND WOLVES EXAMPLE (RIBEIRO ET AL, 2016)

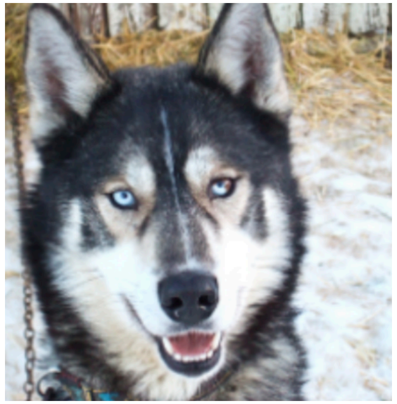


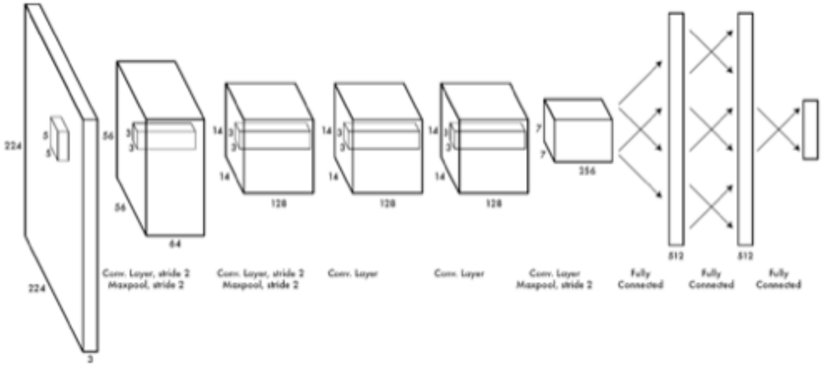
Train a neural-network adapted to images with labelled data

ALGORITHMIC BIAS IN ARTIFICIAL INTELLIGENCE

2 — Why should we be cautious with algorithmic bias?

CONFOUNDING VARIABLES – HUSKIES AND WOLVES EXAMPLE (RIBEIRO ET AL, 2016)





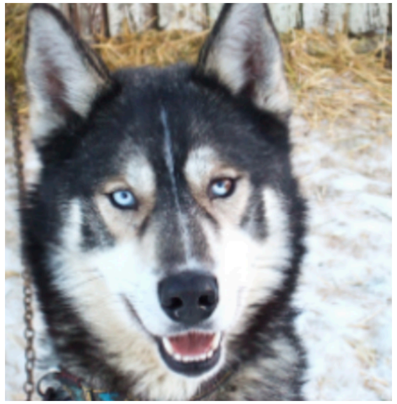
~~Wolf~~

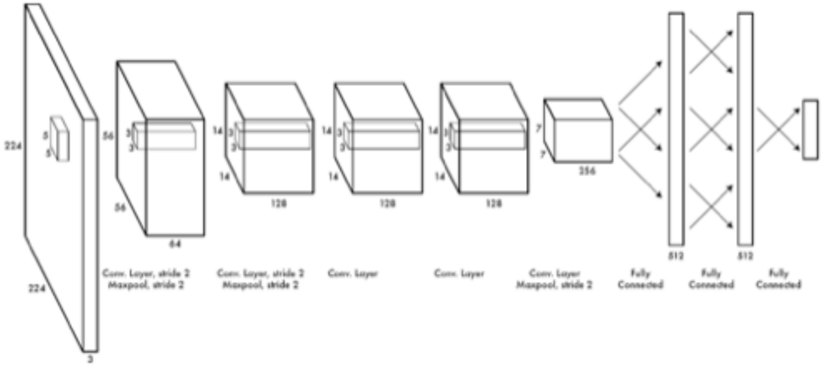
False prediction using the trained neural-network

ALGORITHMIC BIAS IN ARTIFICIAL INTELLIGENCE

2 — Why should we be cautious with algorithmic bias?

CONFOUNDING VARIABLES – HUSKIES AND WOLVES EXAMPLE (RIBEIRO ET AL, 2016)





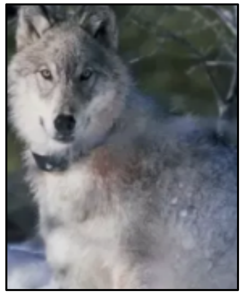
~~Wolf~~

False prediction using the trained neural-network

Why?

⇢ In the training set, most pictures representing a wolf also represent a snowy background, which is not the case for huskies.

⇢ The neural-network associated a snowy background to wolves



1. Why neural-networks have become so popular in A.I.?
2. Why should we be cautious with algorithmic bias?
3. How the E.U. starts regulating A.I.?
4. Can we measure, explain or control algorithmic biases?

ALGORITHMIC BIAS IN ARTIFICIAL INTELLIGENCE

3 — How the E.U. starts regulating A.I.?

THE ARTIFICIAL INTELLIGENCE ACT (EUROPEAN COMMISSION 2021 – CURRENTLY AMENDED)

Article 9.7 (Risk management system):
The testing of the high-risk AI systems shall be performed, as appropriate, at any point in time throughout the development process, and, in any event, prior to the placing on the market or the putting into service. Testing shall be made against preliminarily defined metrics and probabilistic thresholds that are appropriate to the intended purpose of the high-risk AI system.

Article 10.2: Training, validation and testing data sets shall be subject to
(c) relevant data preparation processing operations, such as annotation, labelling, cleaning, enrichment and aggregation;
(d) the formulation of relevant assumptions, notably with respect to the information that the data are supposed to measure and represent;
(f) examination in view of possible biases;

Article 13.1 (Transparency and provision of information to users):
High-risk AI systems shall be designed and developed in such a way to ensure that their operation is sufficiently transparent to enable users to interpret the system's output.

Article 71 (Sanctions):
71.3 (high risk systems and forbiden practice): 30 000 000 euros or up 6% of annual turnover
71.4 (others): 20 000 000 euros or up 4% of annual turnover

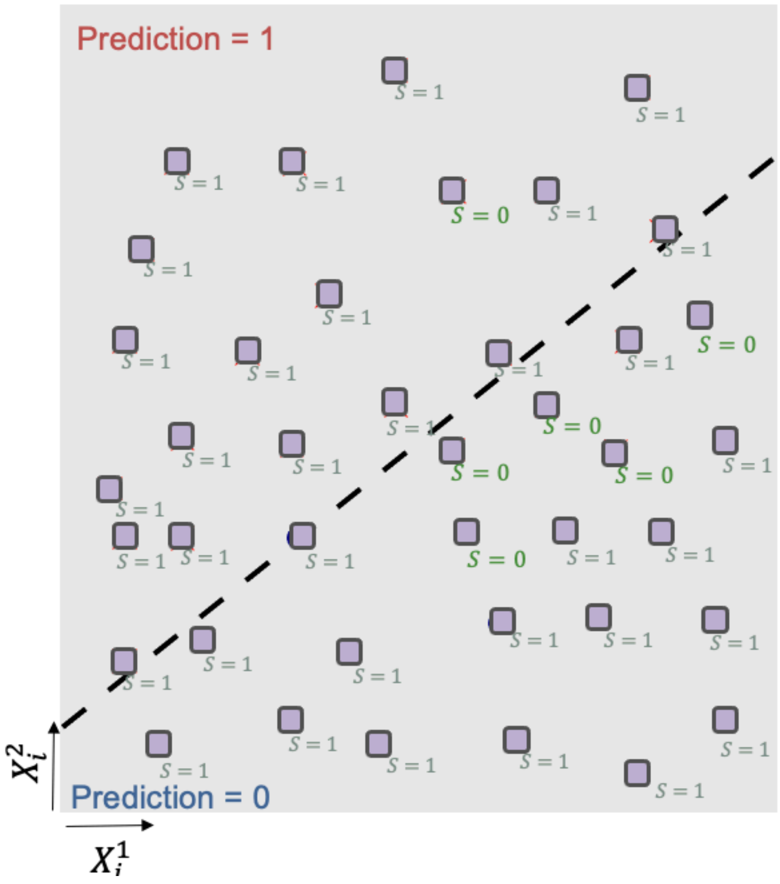
- ➢ Clear requirement to control algorithmic biases
- ➢ Need for appropriate “metrics and probabilistic thresholds” to assess the compliance of AI systems
- ➢ Need for explainable decision rules when using high risk systems

1. Why neural-networks have become so popular in A.I.?
2. Why should we be cautious with algorithmic bias?
3. How the E.U. starts regulating A.I.?
4. Can we measure, explain or control algorithmic biases?

ALGORITHMIC BIAS IN ARTIFICIAL INTELLIGENCE

4 — Can we measure, explain or control algorithmic biases?

POPULAR INDICES TO MEASURE THE BIASES IN A GROUP



$$\text{Disparate Impact : D.I.} = \frac{\text{Percentage of predictions } \hat{Y}=1 \text{ in group } S=0}{\text{Percentage of predictions } \hat{Y}=1 \text{ in group } S=1}$$

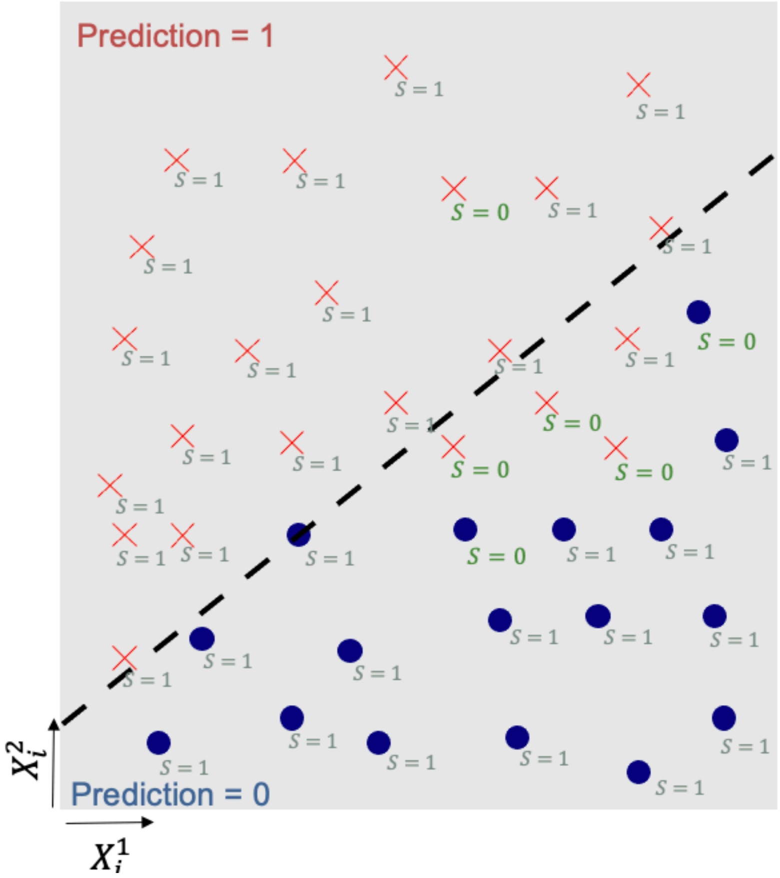
D.I.=0.33 here, which means that there are clearly more **positive predictions** in group S=1 than in group S=0.

- Makes sense for job recommendations
- Makes no sense for disease aid

ALGORITHMIC BIAS IN ARTIFICIAL INTELLIGENCE

4 — Can we measure, explain or control algorithmic biases?

POPULAR INDICES TO MEASURE THE BIASES IN A GROUP



$$\text{Equal Opportunity : E.O.} = \frac{\text{Percentage of predictions } \hat{Y} = 1 \text{ when } Y = 1 \text{ in group } S = 0}{\text{Percentage of predictions } \hat{Y} = 1 \text{ when } Y = 1 \text{ in group } S = 1}$$

E.O.=0.26 here, which means that there are clearly more **true positive predictions** in group S=1 than in group S=0.

- Makes sense for job recommendations
- Makes sense for disease aid
- Requires a ground truth

Many other indices exist, each of them explaining specific bias properties!

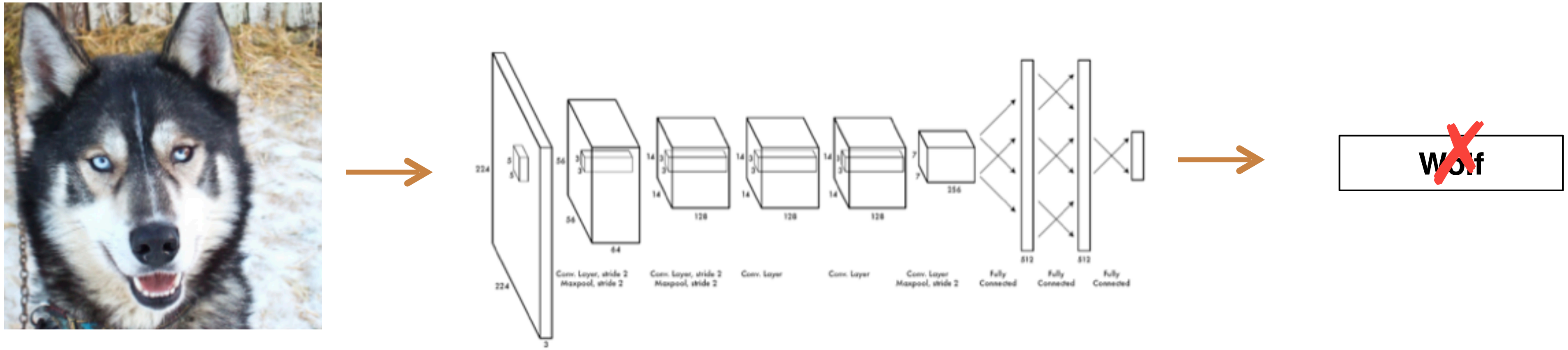
ALGORITHMIC BIAS IN ARTIFICIAL INTELLIGENCE

4 — Can we measure, explain or control algorithmic biases?

POPULAR TOOLS FOR EXPLAINABILITY

Discrimination indices like the « Disparate Impact » or the « Equal Opportunity » work on groups of test data

How to detect that a specific prediction is made for wrong reasons if the ideal prediction is unknown ⇢ use of explainability tools



POPULAR TOOLS FOR EXPLAINABILITY – LIME (RIBEIRO ET AL, 2016)

Model agnostic method

- No need to take into account the model architecture
- Observe how sensitive are the output predictions when the input variables are perturbed
- The prediction is explained by the input variables related to the strongest output changes

(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.

ALGORITHMIC BIAS IN ARTIFICIAL INTELLIGENCE

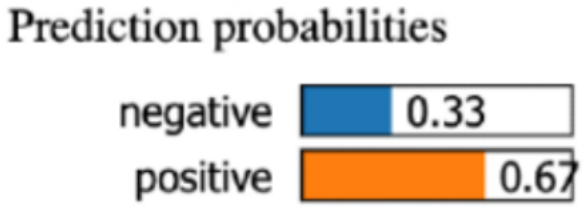
4 — Can we measure, explain or control algorithmic biases?

POPULAR TOOLS FOR EXPLAINABILITY – LIME (RIBEIRO ET AL, 2016)

Text with highlighted words

This amazing documentary gives us a glimpse into the lives of the brave women in Cameroun's judicial system-- policewomen, lawyers and judges. Despite tremendous difficulties-- lack of means, the desperate poverty of the people, multiple languages and multiple legal precedents depending on the region of the country and the religious/ethnic background of the plaintiffs and defendants-- these brave, strong women are making a difference.lbr /llbr /lThis is a rare thing-- a truly inspiring movie that restores a little bit of faith in humankind. Despite the atrocities we see in the movie, justice does get served thanks to these passionate, hardworking women.lbr /llbr /lI only hope this film gets a wide release in the United States. The more people who see this film, the better.

Works on various types of data!



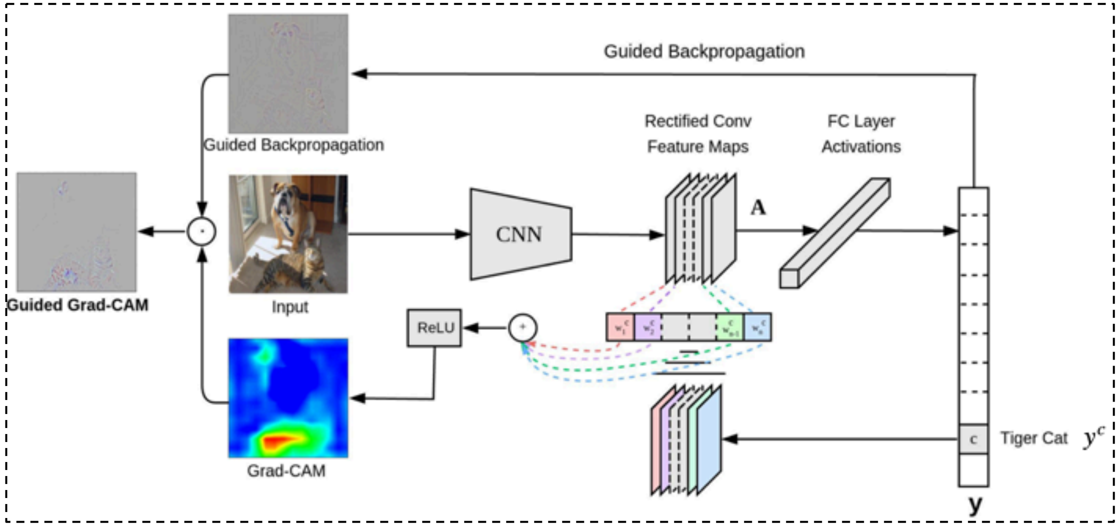
ALGORITHMIC BIAS IN ARTIFICIAL INTELLIGENCE

4 — Can we measure, explain or control algorithmic biases?

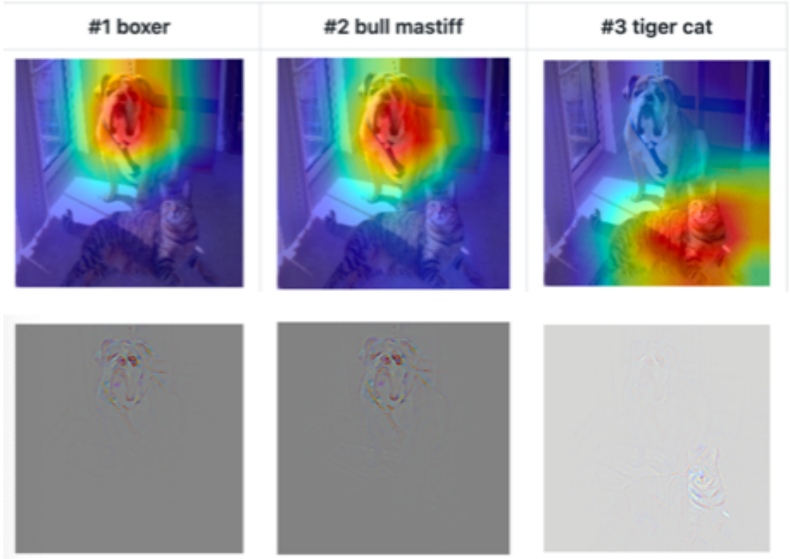
POPULAR TOOLS FOR EXPLAINABILITY – GRADCAM (SELVARAJU ET AL, 2016)

Specialised to images with specific neural-network architectures

- Much Faster than LIME
- Far less flexible



Method overview



Typical result

ALGORITHMIC BIAS IN ARTIFICIAL INTELLIGENCE

4 — Can we measure, explain or control algorithmic biases?

POPULAR TOOLS FOR EXPLAINABILITY – GROUP EXPLAINABILITY USING GEMS-AI (BACHOC ET AL, 2018)

Example: CelebA dataset (http://mmlab.ie.cuhk.edu.hk/projects/CelebA.html)
- >200K celebrity images with 40 binary annotations
- Y_i can be the *Attractive* feature

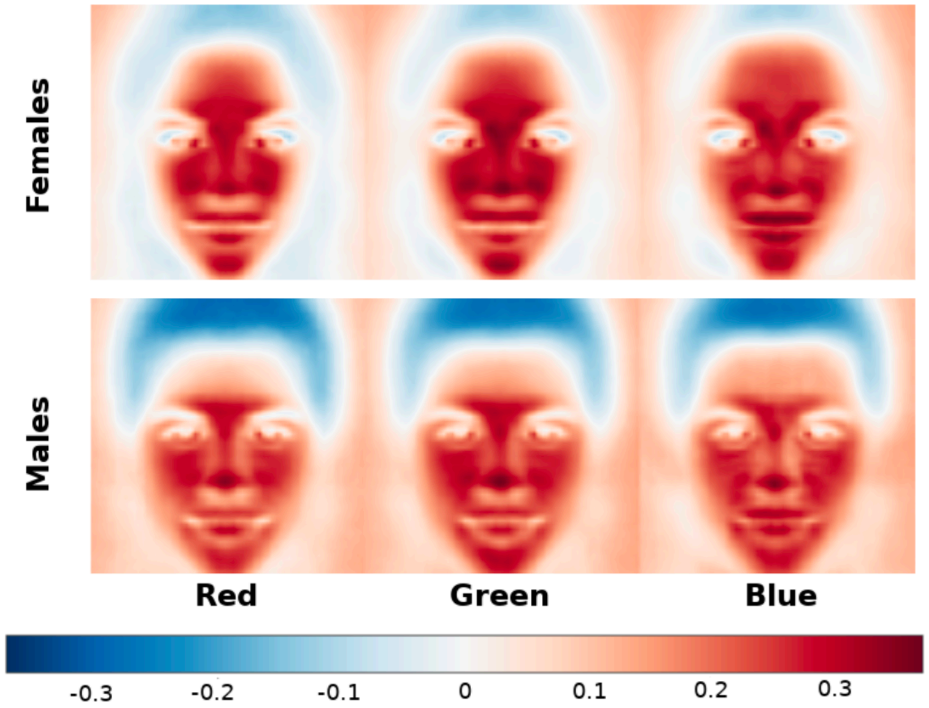
ALGORITHMIC BIAS IN ARTIFICIAL INTELLIGENCE

4 — Can we measure, explain or control algorithmic biases?

POPULAR TOOLS FOR EXPLAINABILITY – GROUP EXPLAINABILITY USING GEMS-AI (BACHOC ET AL, 2018)

Example: CelebA dataset (http://mmlab.ie.cuhk.edu.hk/projects/CelebA.html)

- ResNet 18 CNN trained to predict who is attractive → 87% of accurate predictions on the test set
- What-if the average impact of pixel intensities on Predictions == Attractive for different sub-groups of the test set



https://github.com/XAI-ANITI/ethik

CONTROLLING THE LEVEL OF BIAS IN AI

To sum-up on part 4 so far

- Biases can be quantified and *generally* explained
- In a sense, the explanations give to the data scientists the ability to evaluate the robustness of the neural-networks they train

A key question is

- How to tackle, or at least to reduce, the undesired biases when the are detected?

ALGORITHMIC BIAS IN ARTIFICIAL INTELLIGENCE

4 — Can we measure, explain or control algorithmic biases?

CONTROLLING THE LEVEL OF BIAS IN AI

Very active field of research with some solutions that start being mature!

$$\begin{aligned}\text{minimize} \quad & L(\boldsymbol{\theta}) \\ \text{subject to} \quad & \tfrac{1}{N}\textstyle\sum_{i=1}^{N}(\mathbf{z}_i - \bar{\mathbf{z}})\, d_{\boldsymbol{\theta}}(\mathbf{x}_i) \le \mathbf{c}, \\ & \tfrac{1}{N}\textstyle\sum_{i=1}^{N}(\mathbf{z}_i - \bar{\mathbf{z}})\, d_{\boldsymbol{\theta}}(\mathbf{x}_i) \ge -\mathbf{c},\end{aligned}$$

[Zafar et al., PMLR 2017]

https://github.com/mbilalzafar/fair-classification

$$\hat{\theta} = \arg\min_{\theta \in \Theta} \left\{ \mathcal{R}(\theta) + \lambda \mathcal{W}_2^2(\mu_{\theta,0}, \mu_{\theta,1}) \right\} ,$$ where

$$\mathcal{W}_2^2(\mu_{\theta,0}, \mu_{\theta,1}) = \int_0^1 \left({\mathcal{H}_0}^{-1}(\tau) - {\mathcal{H}_1}^{-1}(\tau) \right)^2 d\tau$$

[Risser et al., JMIV 2022]

https://github.com/lrisser/W2reg

...

- *Neural-networks outperform other A.I. models for advanced prediction/decision tasks.*

- *Neural-network predictions can be biased.*

- *Unreasonable biases will be soon sanctioned by law.*

- *Detection, explanation or reduction of biases is technically doable but requires an expertise in machine learning.*

Direction des Relations avec les Entreprises

CNRS FORMATION ENTREPRISES

L'organisme de formation continue du CNRS

Formation - Intelligence artificielle de confiance : biais en IA et explicabilité - Mise en oeuvre pratique

NOUVEAU

Environnement scientifique et technique de la formation

INSTITUT de MATHEMATIQUES de TOULOUSE

Institut de mathématiques de Toulouse - UMR 5219

RESPONSABLES

Laurent RISSER
Ingénieur de recherche
UMR 5219

Jean-Michel LOUBES
Professeur
UMR 5219

LIEU

TOULOUSE (31)

OBJECTIFS

- Comprendre les mécanismes à l'œuvre dans l'intelligence artificielle
- Comprendre la problématique du biais et de l'explicabilité dans les données et dans l'algorithme
- Détecter le biais et s'en prémunir
- Etre capable de définir les décisions algorithmiques et d'en comprendre l'explicabilité
- Connaître les nécessités juridiques liées aux réglementations nationales et européennes

PUBLICS

Techniciens et ingénieurs en production, traitement, analyse de données et enquêtes. Les stagiaires doivent avoir une expérience sur la manipulation de données, mais pas nécessairement en apprentissage automatique.

Merci !