AI ACT DAY

Hands-on workshop
Bias in AI: how to measure & mitigate

événement co-organisé par

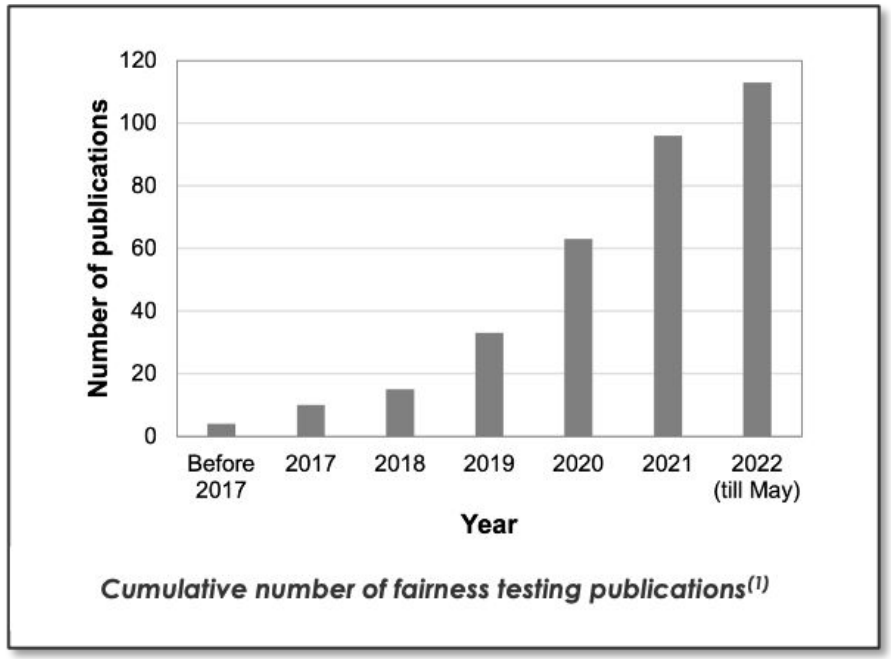
datacraft* IMPACT AI

Summary

- Introduction to bias
- Metrics & tools for measuring bias
- Mitigating Bias
- Workshop presentation

Introduction to bias

Multiple AI scandals related to biased algorithms triggered actions in the scientific community and the EU



Fairness is one of the 7 requirements in the **EU Ethics guidelines for trustworthy AI**[(2)] and it will certainly become a prerequisite for high-risk AI applications after the enforcement of the EU AI Act

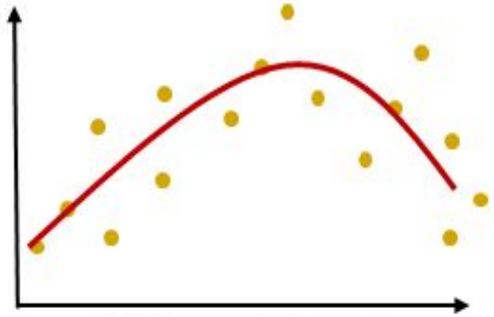
> Diversity, non-discrimination and fairness: Unfair bias must be avoided, as it could could have multiple negative implications, from the marginalization of vulnerable groups, to the exacerbation of prejudice and discrimination. Fostering diversity, AI systems should be accessible to all, regardless of any disability, and involve relevant stakeholders throughout their entire life circle.

(1) Fairness Testing: A Comprehensive Survey and Analysis of Trends, (2) Ethics guidelines for trustworthy AI

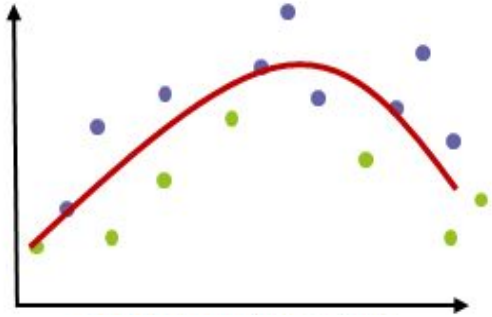
Definition and origin of Bias

Definition of a bias

Systematic difference between "prediction" made by an algorithm and the actual expected value. This may concern all predictions or some having common points:



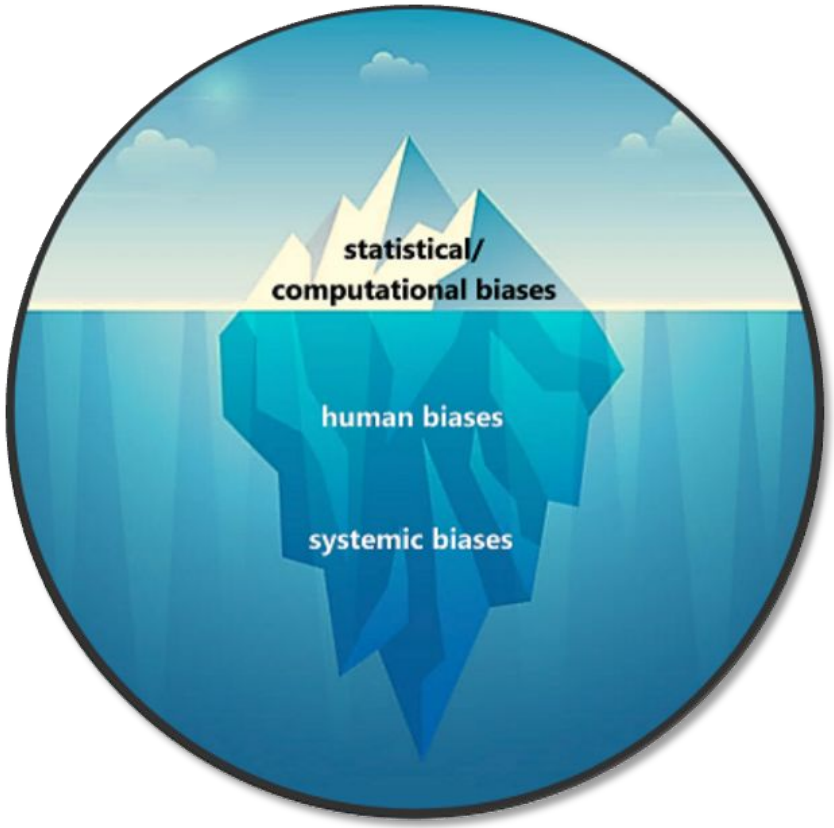
Model seems to have low bias



But it is biased considering a specific variable

Man

Woman



Origin of bias (unsubstantiated figures, but this is to give an order of magnitude)

- The data - 95% of the time
- Algorithmic processing - 5% of the data

(1) Towards a Standard for Identifying and Managing Bias in Artificial Intelligence (NIST).

First source of bias is the dataset itself and its (pre)processing

Different types of bias in data

Sampling bias
Dataset is not representative of the population or the phenomena, some societal groups are excluded. This data is not generalizable to the entire population.
Example: Non-uniform methods, technological differences

Availability bias
Already available dataset are not fully representative of the target population. Disadvantaged groups are always underrepresented
Example: Effects related to survivor bias (Survey, Fraud), recency bias (Rewriting data).

Label bias
Bias introduced during the creation of training data or the labelling phase
Example: people labelling bring their implicit personal biases into their labels. A not enough diverse group of people can lead to skewed labels and systematic disadvantages to certain groups

Some guidance to minimize bias related to data

Ensure accountablility for socio-technical factor during the AI lifecycle

Design phase
- What are the social variations to take into account ? Are there specific characteristics of the phenomena?
- Does the dataset fit with the identified characteristics ?

Development phase
- Select data sources and attributes accordingly
- Integrate impact assessment together with algorithmic accuracy

Deployment phase
- Make sure the AI systems is used in the intended social context

Be aware of human and systemic biases while collecting and pre-processing data

- Identify and pay attention to vulnerable demographics
- Make sure that data is representative of all groups and analyse their statistical representation
- Take great precautions while using benchmark datasets

Towards a Standard for Identifying and Managing Bias in Artificial Intelligence (NIST)

Risks related to bias

Performance issues

When undesired, bias can lead to performance degradation (algorithmic and/or financial)

Legal issues

The criteria of discrimination, according to gender, ethnic origin etc., retained by the defender of rights, Claire Hédon in her recommendations

Ethical issues

Not always legally condemnable, but it can have serious consequences on the public image of the company like bad buzz, boycott, etc.

Management and arbitration

Take into account business context

Resolving bias issues is not only technical and requires information and business context to be taken into consideration

Separate auditor vs audited roles

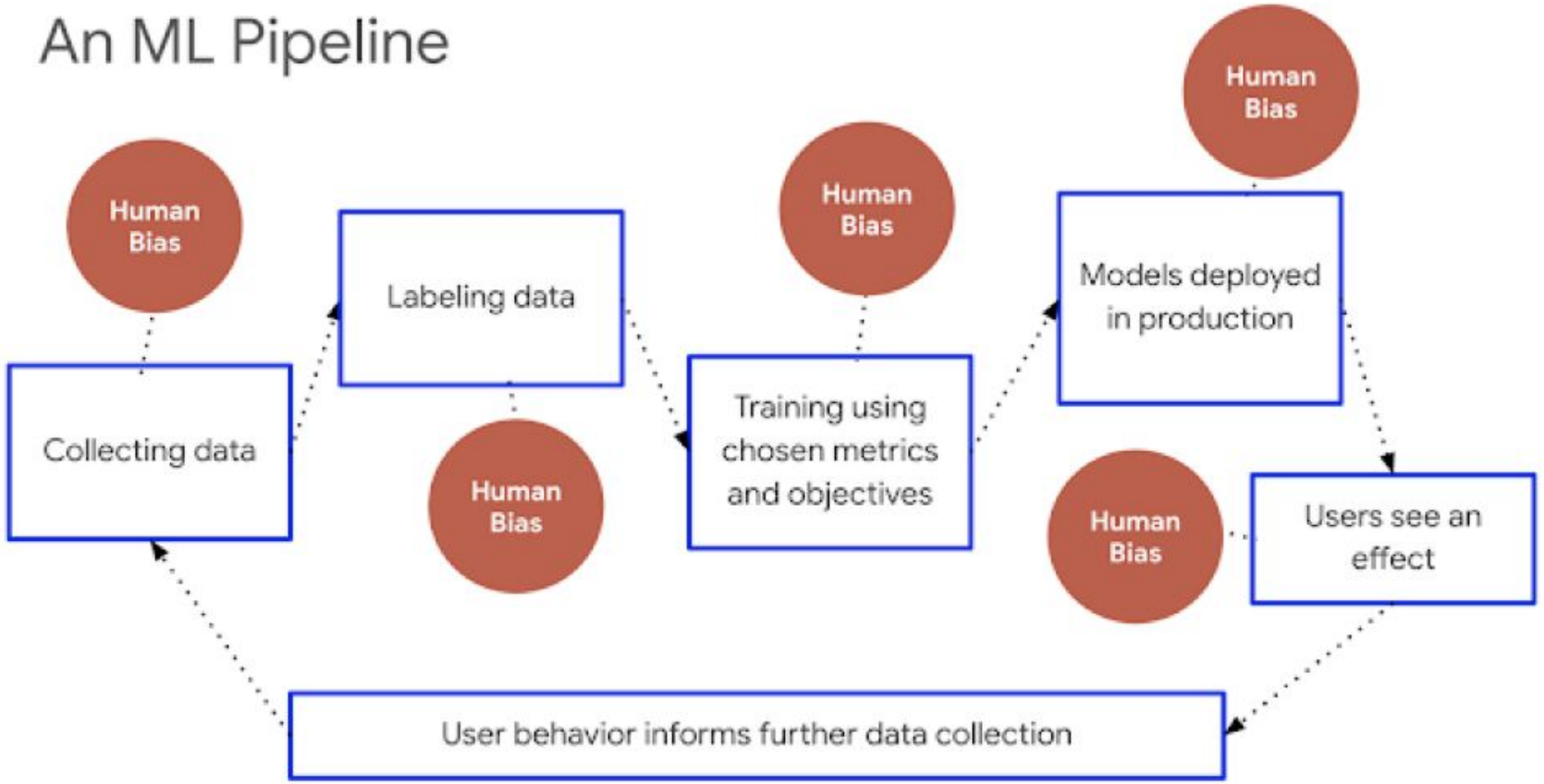
Looking for biases in one's own work is not an effective configuration: questioning one's own work, presence of unconscious biases, lack of differences in viewpoints, etc.

Trade-off mitigation <> cost

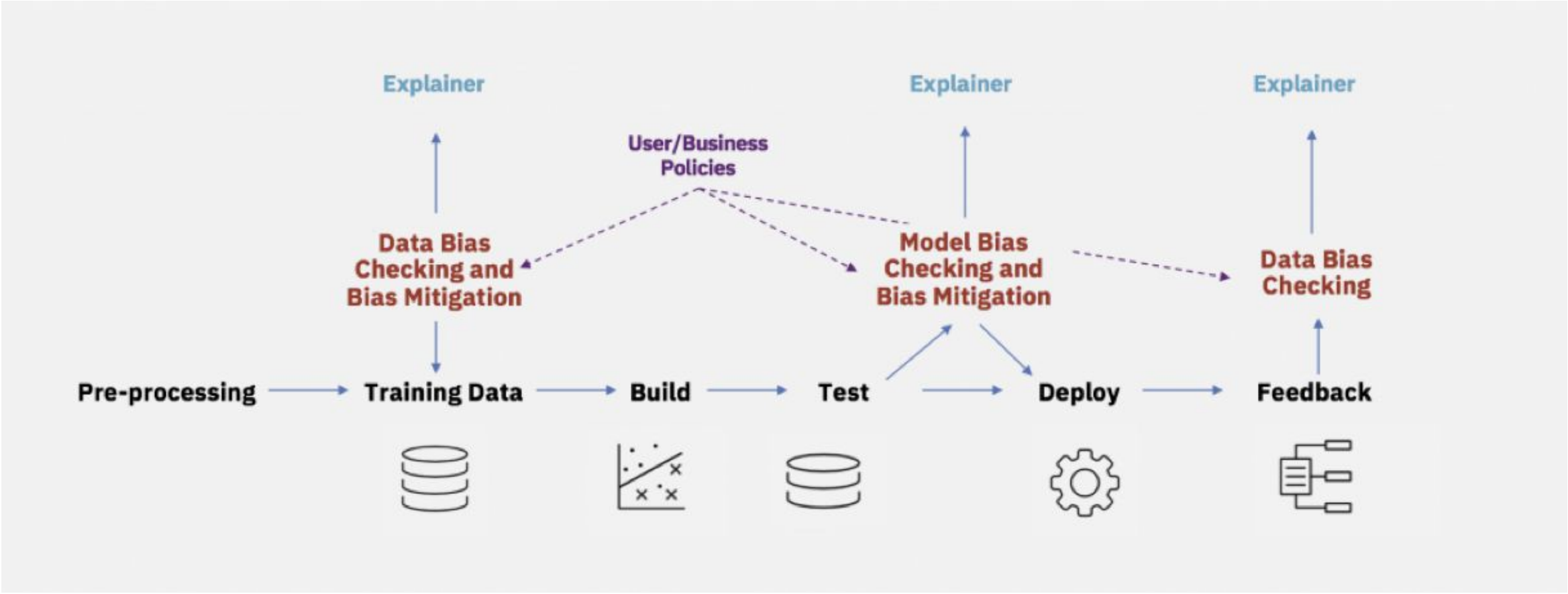
Being more ethical by degrading the performance of the model may result in lost revenue (in addition to the time allocated to finding and resolving such biases).

Metrics & tools for measuring bias

Biases at every levels

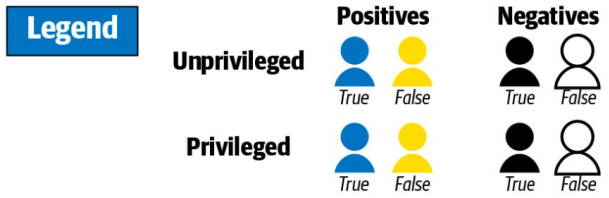


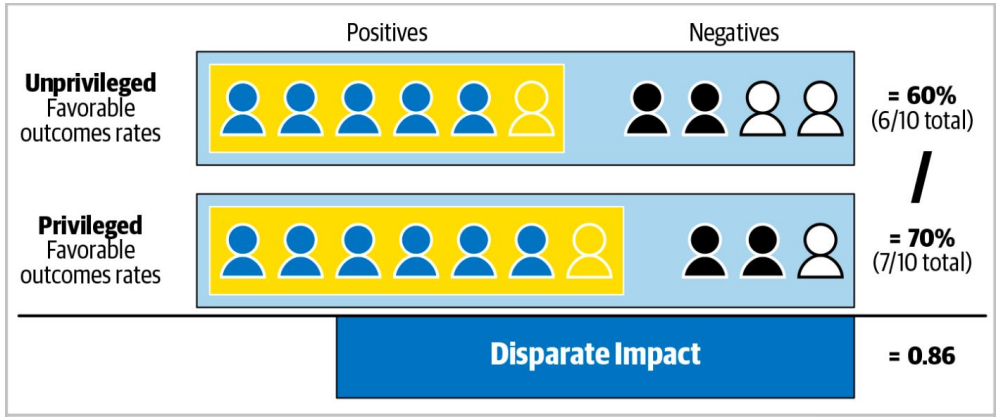
Checks at every levels



Bias Metrics

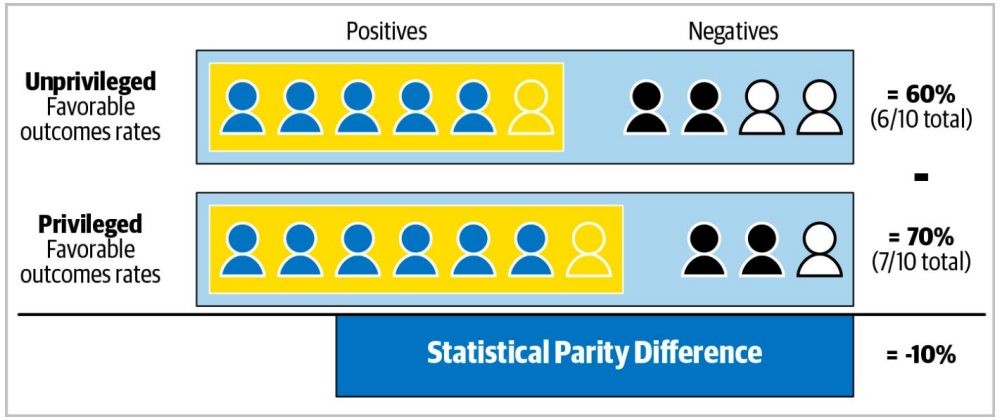
Here are 3 of the most common metrics that are considered the most commonly used :

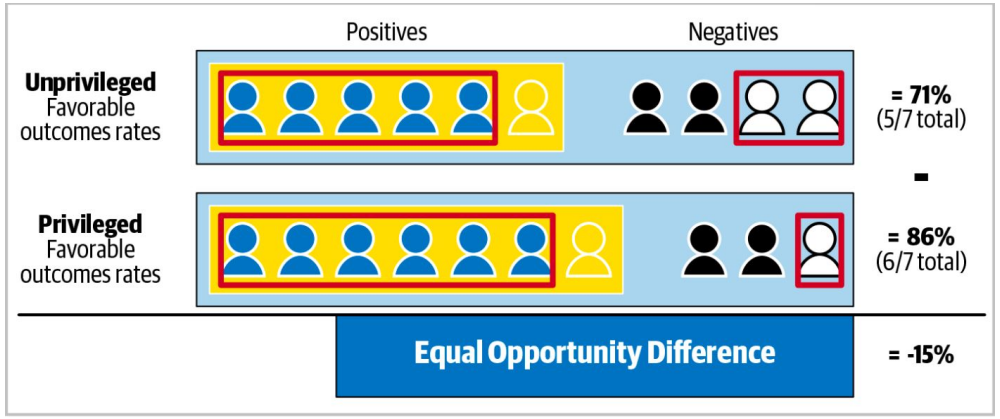




$$\frac{\mathbb{P}(Positive\ outcome \mid unprivilege)}{\mathbb{P}(Positive\ outcome \mid privilege)}$$

from here





Here is a research paper list a lot more fairness metrics (still not an exhaustive list !).

Choosing a fairness metric depends on the bias to be treated and it includes its business aspect.

Bias Metrics

Here are 2 metrics that are considered the most commonly used in the literature :

Equalized Opportunity Ratio (or Disparate Impact)	The ratio of the proportion of unprivileged elements that receive a positive outcome over the proportion of privileged elements that receive a positive outcome.

$$\frac{\mathbb{P}(Positive\ outcome \mid unprivilege)}{\mathbb{P}(Positive\ outcome \mid privilege)}$$

Equalized Odds Ratio	The ratio of proportion of unprivileged elements that are predicted correctly (True positive & True negative) over the proportion of privilege elements that are predicted correctly.

$$\frac{\mathbb{P}(True\ Prositive_{unprivilege}\ or\ True\ Negative_{unprivilege})}{\mathbb{P}(True\ Prositive_{privilege}\ or\ True\ Negative_{privilege})}$$

Here is a research paper list a lot more fairness metrics (still not an exhaustive list !).

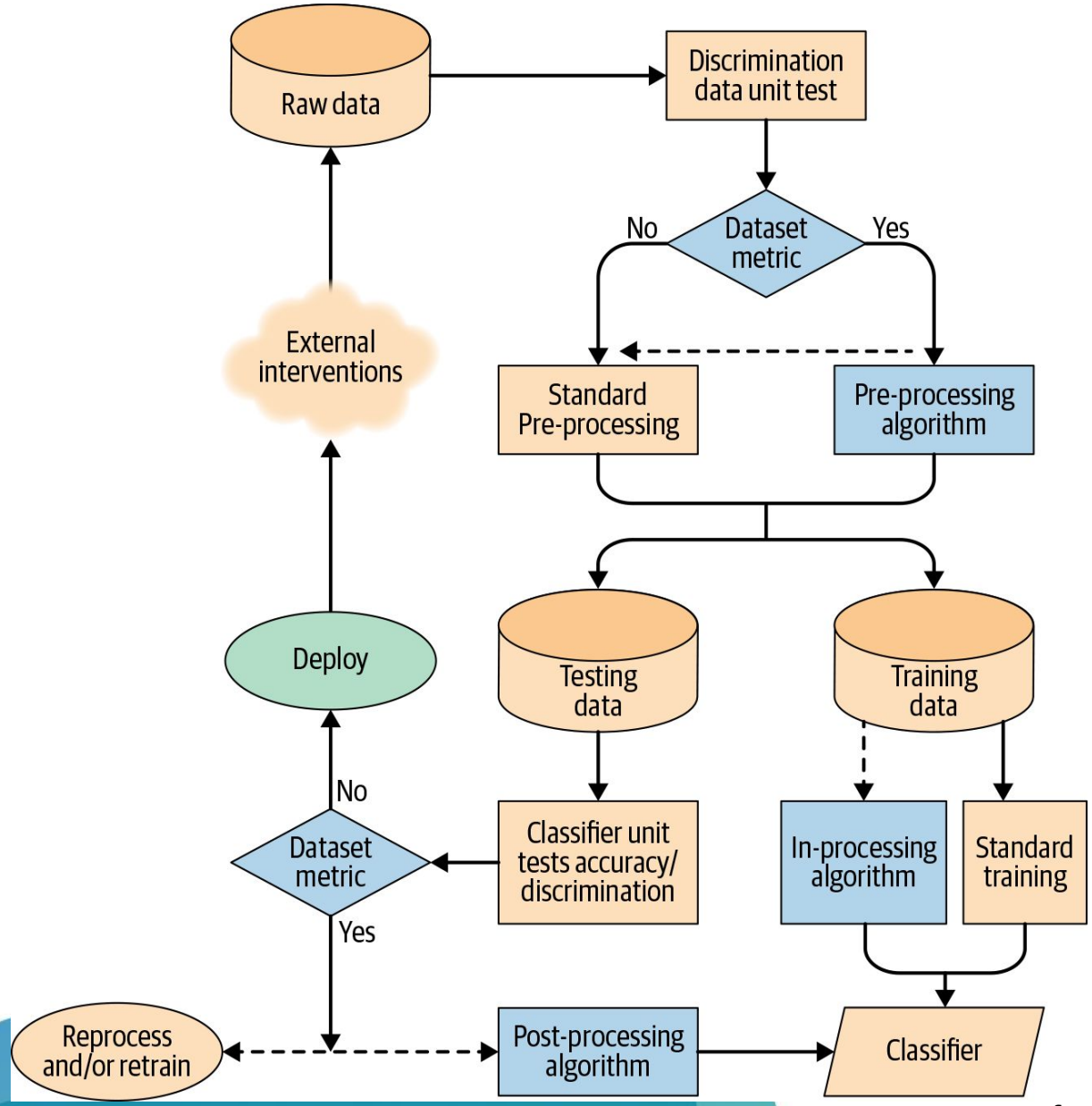
Choosing a fairness metric depends on the bias to be treated and it includes its business aspect.

Tools for measuring Bias (non-exhaustive list)

DALEX	Dalex : Developed by MI² (members of Warsaw University, PL). Python & R package. Model explanation and bias measurement.	 
AI Fairness 360	AI Fairness 360 : Created by IBM Research and donated by IBM to the Linux Foundation AI & Data. Python package.	
Aequitas Bias & Fairness Audit	Aequitas : Center for Data Science and Public Policy at University of Chicago. Open source bias audit toolkit. Python package.	
What-If Tool	What if tool : Open source project (Maintained by Google Research). Available in python (jupyter implementation). Descriptive statistics on data and model performances & fairness.	
Fairlearn	Fairlearn : Open source project (Maintained by Microsoft Research). Python package. Bias measurement and mitigation.	

Mitigating Bias

Bias management (technically)



from here

Approaches

- **Pre-processing**: Re-process the data used by the models by removing (as much as possible) the biases: Resampling, data removal, manual corrections, reweighting….

- **In-processing**: Re-develop machine learning algorithms to integrate counter-metrics constraining the initial optimization algorithms.

- **Post-processing:** Analyze the predicted results and change (manually or automatically) some of the results in order to debiased the overall results according to the desired criteria. Be careful, this strategy is not applicable in all cases.

Mitigation techniques (there are many)

Optimized Pre-processing

Use to mitigate bias in training data. Modifies training data features and labels.

Reweighing

Use to mitgate bias in training data. Modifies the weights of different training examples.

Adversarial Debiasing

Use to mitigate bias in classifiers. Uses adversarial techniques to maximize accuracy and reduce evidence of protected attributes in predictions.

Reject Option Classification

Use to mitigate bias in predictions. Changes predictions from a classifier to make them fairer.

Disparate Impact Remover

Use to mitigate bias in training data. Edits feature values to improve group fairness.

Learning Fair Representations

Use to mitigate bias in training data. Learns fair representations by obfuscating information about protected attributes.

Prejudice Remover

Use to mitigate bias in classifiers. Adds a discrimination-aware regularization term to the learning objective.

Calibrated Equalized Odds Post-processing

Use to mitigate bias in predictions. Optimizes over calibrated classifier score outputs that lead to fair output labels.

Equalized Odds Post-processing

Use to mitigate bias in predictions. Modifies the predicted labels using an optimization scheme to make predictions fairer.

Meta Fair Classifier

Use to mitigate bias in classifier. Meta algorithm that takes the fairness metric as part of the input and returns a classifier optimized for that metric.

Workshop presentation

Workshop principles: Three use cases

1. **Titanic: Practical introduction to Responsible AI implementation techniques (using the AIF360 library), led by Annabelle**

2. **HR: Analysis of recruitment bias (using the Dalex library), led by Adeline, Céline and Francis**

3. **Banking: Bias analysis when granting a credit (using the AIF360 library) led by Julien and Zied**

CHOOSE A USE CASE **SMALL GROUPS** **INTERACTIVE & COLLABORATIVE**

YOUR OWN PYTHON ENVIRONNEMENT **OR** **GOOGLE COLAB**

https://github.com/datacraft-paris/2212-AIActDay-BiasWorkshop/tree/main/notebooks

Presentation of the 'Titanic' use case (1/2)

- On April 15, 1912, the largest ocean liner ever built collided with an iceberg on its maiden voyage. The sinking of the Titanic killed 1502 of the 2224 passengers and crew. This sensational tragedy shocked the international community and led to better regulation of ship safety. One of the reasons the sinking resulted in such a high loss of life was that there were not enough lifeboats for the passengers and crew. Although there is an element of chance in survival, it seems that some groups of people have a better chance of survival than others.
- In this use case, the goal is to build a predictive model that answers the question, "What kinds of people are more likely to survive?" using passenger data (name, age, gender, socioeconomic class, etc.).
- The dataset contains data from actual Titanic passengers. Each row represents a person. The columns describe different attributes of the person, including whether they survived, their age, passenger class, gender, and the price they paid.

	Survived	Pclass	Sex	Age	SibSp	Parch	Fare	Embarked	relatives	not_alone	Deck	Title	Age_Class	Fare_Per_Pers
0	0	3	0	2	1	0	0	0	1	0	8	1	6	0
1	1	1	1	5	1	0	3	1	1	0	3	3	5	1
2	1	3	1	3	0	0	0	0	0	1	8	2	9	0

Presentation of the 'Titanic' use case (2/2)

We will go through the different stages of pre-processing and model training applying best practices in terms of :

- Data Exploration
- Data Cleaning
- Biais detection in the dataset (AIF360)
- Pre-processing mitigation
- Directly interpretable model (Decision tree)
- Biais detection in prediction (AIF360)
- Non-directly interpretable model (Random Forest)
- Interpretability method (Shap)

Presentation of ‘RH’ use case (1/3)

- StackOverflow's annual user-generated survey (over 70,000 responses from over 180 countries) of developers examines all aspects of the developer experience, from learning code to preferred technologies, version control and work experience.

- The survey provides a comprehensive dataset that illustrates the different backgrounds, technologies used, and education of developers around the world, a perfect HR use case for recruitment.

Res	MainBranch	Employment	RemoteWorl	CodingActivi	EdLevel	LearnCode	LearnCodeOnline	LearnCodeCours
1	None of these	NA	NA	NA	NA	NA	NA	NA
2	I am a developer by	Employed, full-time	Fully remote	Hobby;Contri	NA	NA	NA	NA
3	I am not primarily a	Employed, full-time	Hybrid (some	Hobby	Master’s degree	Books / Physical media;Friend	Technical documentatic	NA

Presentation of ‘RH’ use case (2/3)

We started from the original dataset and built a simplified dataset with the following columns:

Age: age range <35 or >35

EdLevel: education level Undergraduate, Master, PhD...

Gender: Man, Woman, NonBinary

MainBranch: if the person is a Dev developer, NotDev

YearsCode: how long the person has been coding (int)

YearsCodePro: how long the person has been coding in a professional context (int)

PreviousSalary: salary of the previous job (float)

ComputerSkills: number of computer languages mastered (int)

Employed: if the person has been hired (target 0, 1)

A second, more complex dataset has other columns and can be explored and used in a second time.

Presentation of ‘RH’ use case (3/3)

In this use case we perform different tasks:

- Dataset n° 1
  - Exploratory analysis of the data
  - Analysis of different bias mitigation methods mainly with the DALEX library
    - Do nothing
    - Remove sensitive attributes
    - Re-sampling
    - Adversarial training with the AIF360 library
- Dataset n° 2
  - Complete analysis to be performed: this time it is up to each participant to choose the sensitive variables, the mitigation approach... it's up to you!

Presentation of the banking use case

In this use case we perform different tasks:

- Term deposits are a cash investment held at a financial institution that is invested at an agreed-upon interest rate over a set period of time, or term.
- Telephone marketing campaigns remain one of the most effective ways to reach customers to sell term deposits. However, they require considerable investment, which is why it is crucial to identify in advance the customers most likely to convert so that they can be specifically targeted.
- The data at our disposal are related to the direct marketing campaigns (phone calls) of a Portuguese banking institution. The objective of the classification is to predict whether the customer will subscribe to a term deposit.

# age	A job	A marital	A education	✓ default	# balance	✓ housing	✓ loan	A contact	# day
58	management	married	tertiary	no	2143	yes	no	unknown	5
44	technician	single	secondary	no	29	yes	no	unknown	5
33	entrepreneur	married	secondary	no	2	yes	yes	unknown	5

MERCI !