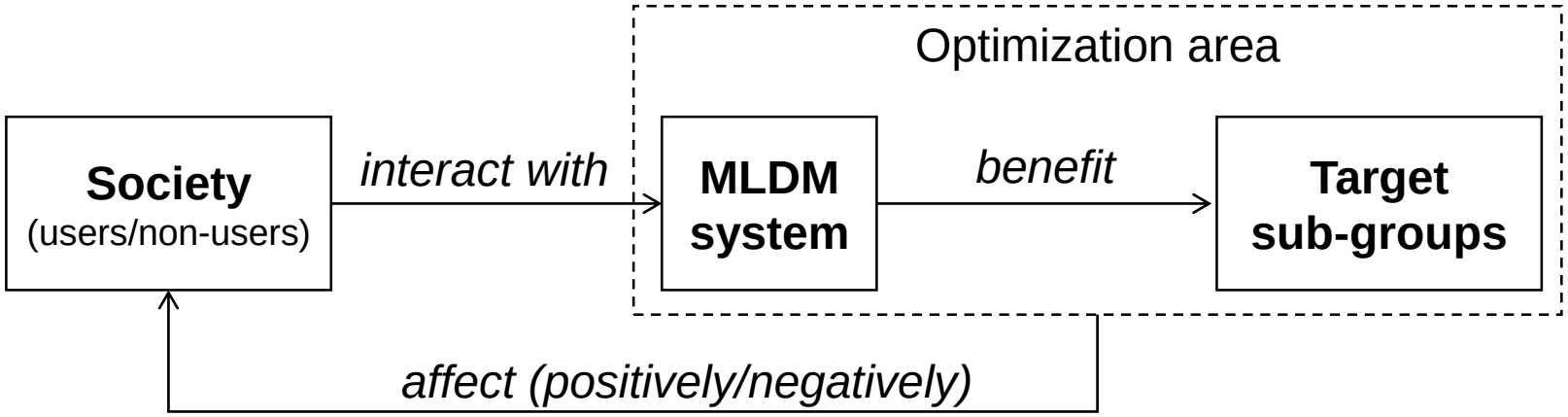
Algorithmic bias

Guilherme Alves

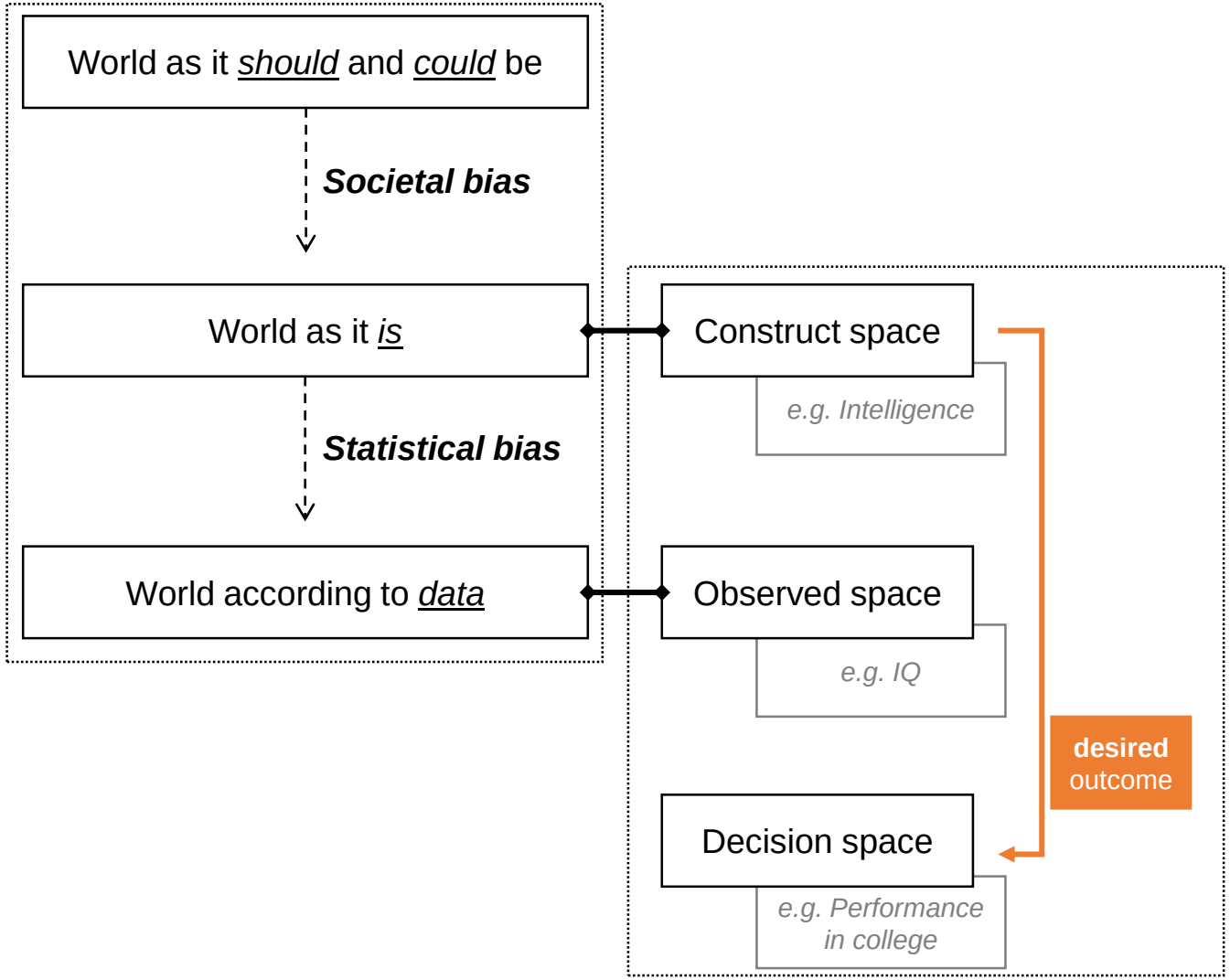
guilherme.alves-da-silva@univ-lorraine.fr

Society and Machine Learning (ML)

- Helping humans with ML-based Decision Making systems
  - **Humans**: subjective decisions
  - **Machines**: "objective" decisions



Biases, biases … everywhere!



Biases, biases … everywhere!

- **ML models**: *designed* to have some bias that *guide* them in their task

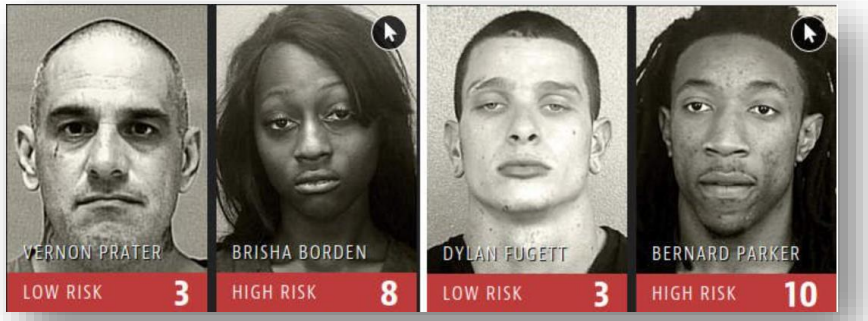
Expected bias		
Credit card default prediction	(good) credit payment history	↑
Hate speech prediction	(presence of) offensive terms	↑

Unintended bias		
Credit card default prediction	(minority) ethnicity	↓
Hate speech prediction	language variant	↓

- Unintended biases → **unfair** algorithmic decisions and **discrimination**
- **Discrimination:** "unjust or prejudicial treatment of different categories of people, especially, on the grounds of race, age, or sex"

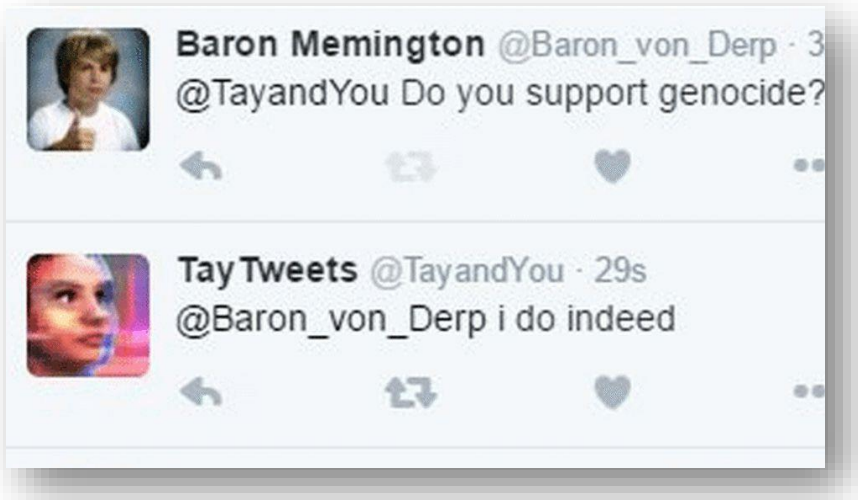
Some cases of unfair algorithmic decisions

- COMPAS (tabular data)
  - Public dataset contains information from Broward County, Florida
  - Goal: Predict two-year violent recidivism



Some cases of unfair algorithmic decisions

- Chatbot Tay, Microsoft (**textual** data)
  - Deployed on Twitter in 2016



Cases of unfair algorithmic decisions

- Other critical applications:
  - Loan requests
  - Job applications
  - Stop & Frisk
  - …

- **Need of fairness**
  - Unfair outcomes not only affect human rights…
  - …but they undermine public trust in artificial intelligence!

Algorithmic fairness

- **Assessing (un)fairness**
  - Based on decision outcomes
  - Based on the reliance of the model on “sensitive features”

- **Mitigating unfairness**
  - Enforce fairness contraints while training, e.g.:

$$P(Y = true|gender = female) = P(Y = true|gender = male)$$

    - **Drawback**: complexity

  - Exclude sensitive/salient features

    - **Drawback**: Decrease accuracy

Assessing (un)fairness

Group fairness notions

Individual fairness notions

Assessing (un)fairness

- **Group** fairness notions:
  - Separate instances into **two groups** w.r.t. a sensitive feature A
  - Unprivileged group (unp) *versus* privileged group ($priv$)
  - Example: female *versus* male

		Actual outcome	
		$Y = 1$	$Y = 0$
Pred	$\hat{Y} = 1$	TP (true positive)	FP (false positive)
	$\hat{Y} = 0$	FN (false negative)	TN (true negative)

- **Equal Opportunity:** focus on true positives (TP)

$$\frac{TP_{unp}}{TP_{unp} + FN_{unp}} = \frac{TP_{priv}}{TP_{priv} + FN_{priv}}$$

- **Predictive Equality:** focus on false positives (FP)

$$\frac{FP_{unp}}{FP_{unp} + TP_{unp}} = \frac{FP_{priv}}{FP_{priv} + TP_{priv}}$$

Assessing (un)fairness

- **Group** fairness notions. Example **Job Hiring** (Y)

(a) Dataset

Gender	Education Level	Job Expe-rience	Age	Marital Status	Y
Female 1	8	2	39	single	0
Female 2	8	2	26	married	1
Female 3	12	8	32	married	1
Female 4	11	3	35	single	0
Female 5	9	5	29	married	1
Male 1	11	3	34	single	1
Male 2	8	0	48	married	0
Male 3	7	3	43	single	1
Male 4	8	2	26	married	1
Male 5	8	2	41	single	0
Male 6	12	8	30	single	1
Male 7	10	2	28	married	1

(b) Prediction

Ŷ	S
1	0.5
0	0.1
1	0.5
0	0.2
0	0.3
1	0.8
0	0.1
0	0.1
1	0.5
1	0.5
1	0.8
0	0.3

Male

		Actual outcome	
		$Y = 1$	$Y = 0$
Pred	$\hat{Y} = 1$	$TP = 3$	$FP = 1$
	$\hat{Y} = 0$	$FN = 2$	$TN = 1$

Female

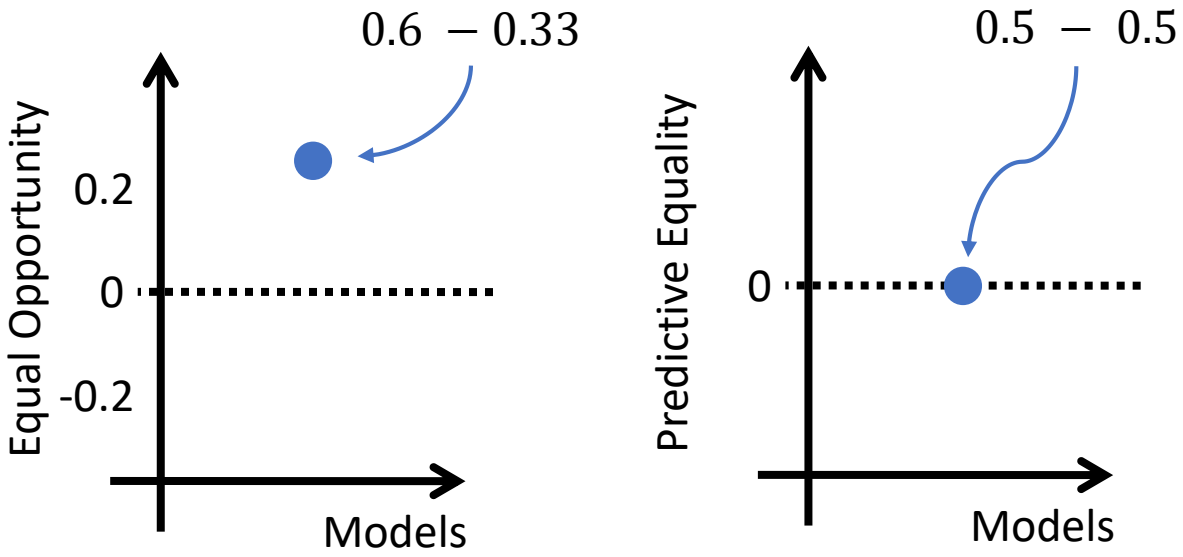
		Actual outcome	
		$Y = 1$	$Y = 0$
Pred	$\hat{Y} = 1$	$TP = 1$	$FP = 1$
	$\hat{Y} = 0$	$FN = 2$	$TN = 1$

TPR for male and female groups is 0.6 and 0.33
FPR is exactly the same (0.5) for both groups.

Makhlouf, K., et al. (2021). ***On the applicability of machine learning fairness notions****. ACM SIGKDD Explorations Newsletter, 23(1), 14-23.*

Assessing (un)fairness

- **Group** fairness notions. Example **Job Hiring** (Y)



Male		Actual outcome	
		$Y = 1$	$Y = 0$
Pred	$\hat{Y} = 1$	$TP = 3$	$FP = 1$
	$\hat{Y} = 0$	$FN = 2$	$TN = 1$

Female		Actual outcome	
		$Y = 1$	$Y = 0$
Pred	$\hat{Y} = 1$	$TP = 1$	$FP = 1$
	$\hat{Y} = 0$	$FN = 2$	$TN = 1$

TPR for male and female groups is 0.6 and 0.33
FPR is exactly the same (0.5) for both groups.

Makhlouf, K., et al. (2021). ***On the applicability of machine learning fairness notions****. ACM SIGKDD Explorations Newsletter, 23(1), 14-23.*

Assessing (un)fairness

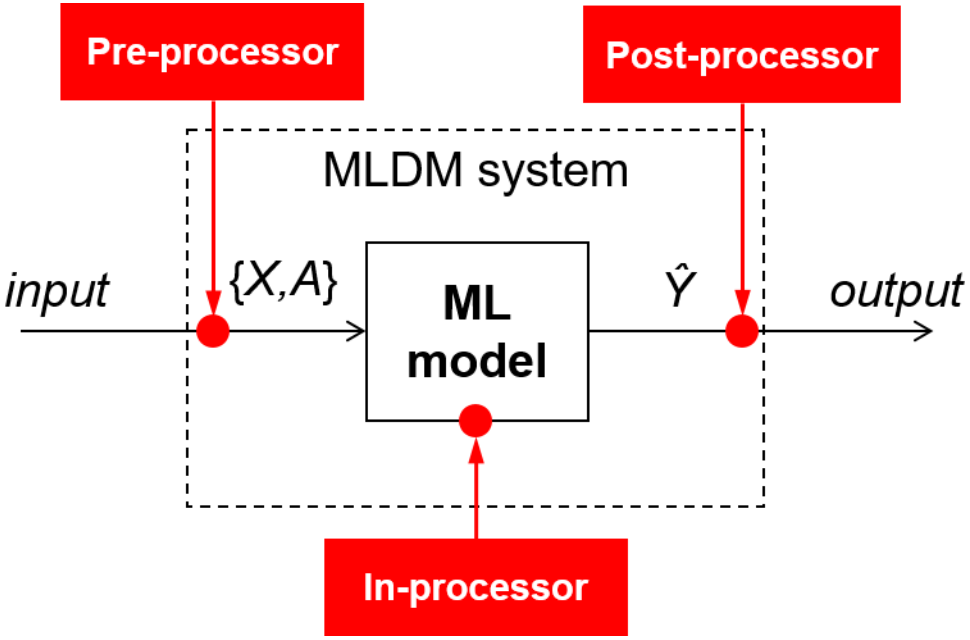
- **Group** fairness notions

  - Ignore **non-sensitive** features → **may hide unfairness**!
  - Several notions: **Impossible** to satisfy all of them at the same time!

- **Individual** fairness notions
  - Takes **non-sensitive** features into account
  - But it requires a measure of similarity between two **individuals**

Mitigating unfairness

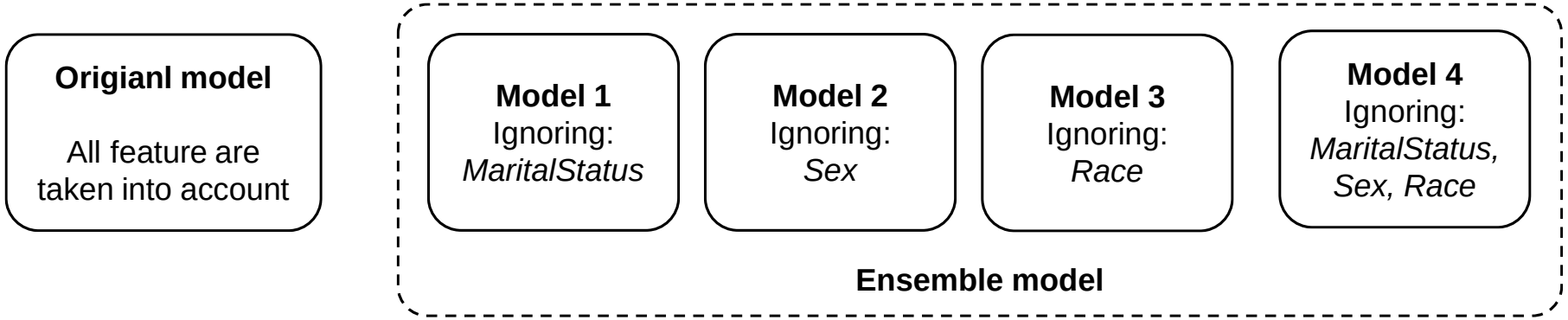
- **Pre-processing**
  - Modify input
- **In-processing**
  - Modify the algorithm to impose fairness during the training process
- **Post-processing**
  - Modify outputs
- **Hybrid-processing**
  - Combine several processors



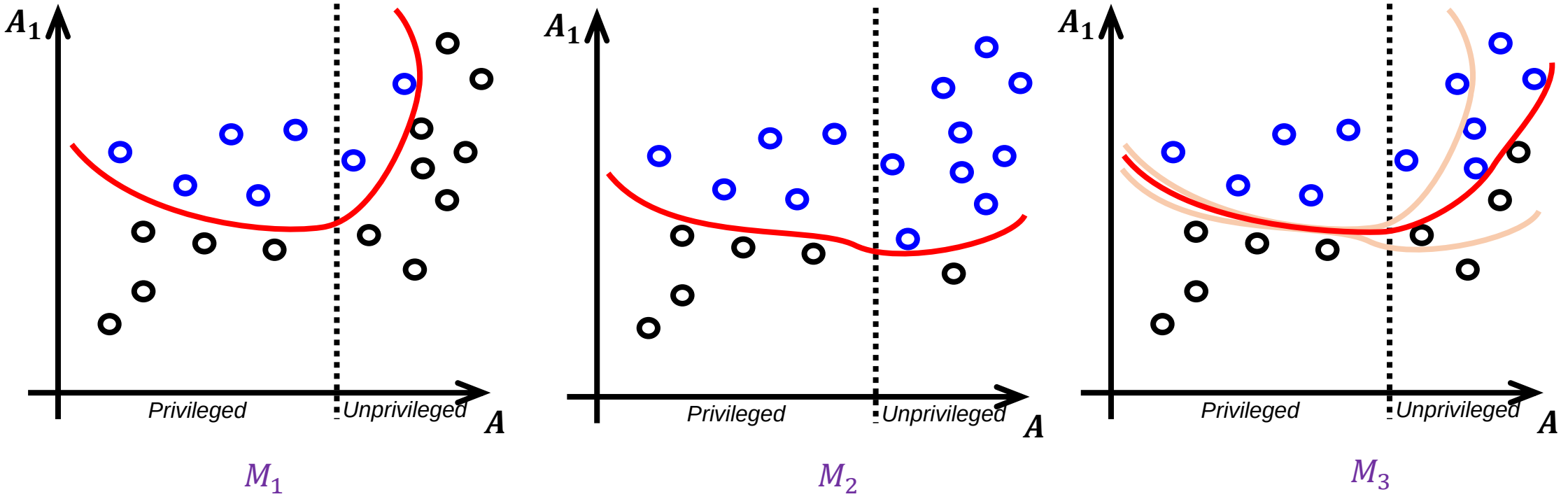
Example: pre-processing

- Illustration taken from the Adult dataset
  - **Goal**: predict if a person earns > US$ 50k per year
  - Profiles (demographic and socio-economic)
  - **Sensitive features**: '*MaritalStatus*', '*Sex*', and '*Race*'

- Which sensitive feature are removed before training



Example: post-processing



- M_1 was trained with the sensitive feature A
- M_2 was trained after removing A
- M_3 aggregates M_1 and M_2's outputs

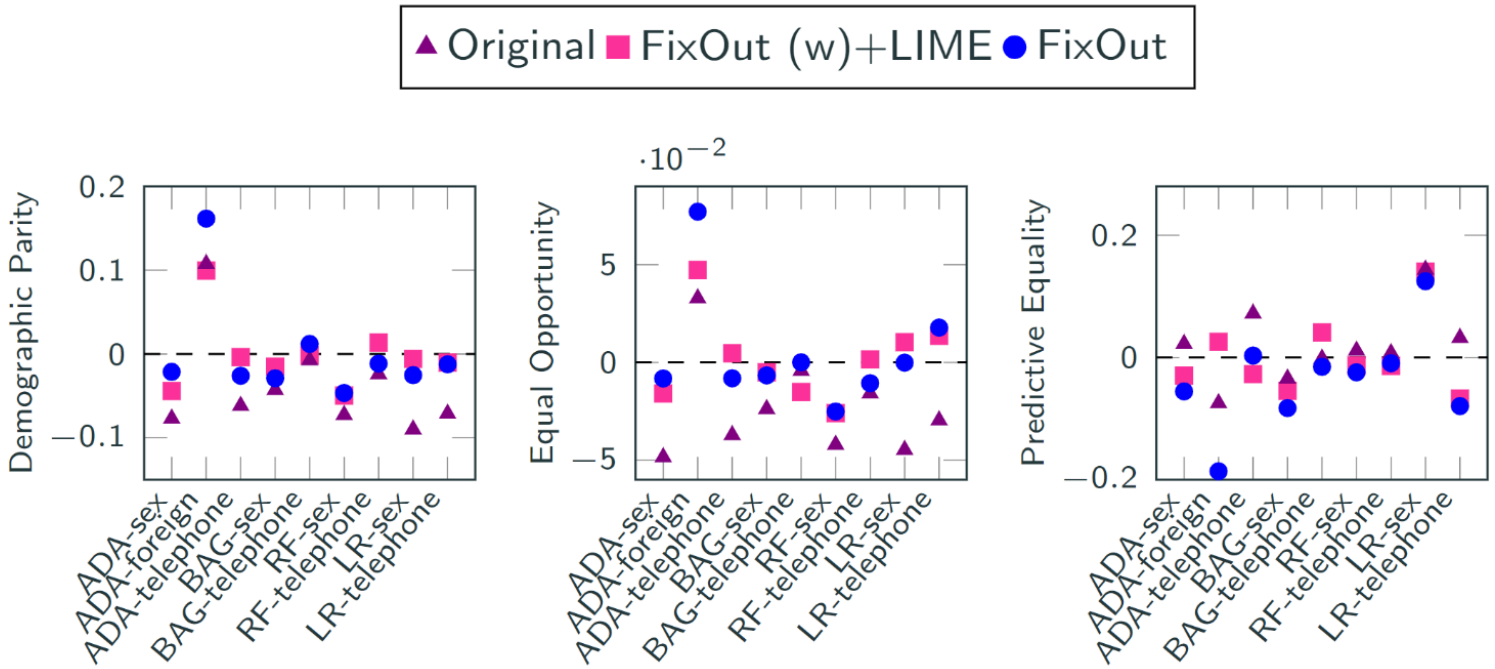
Example: Credit card default prediction

- **German Credit Card Score**
  - **Goal**: Predict credit risks (likely & unlikely to pay back)
  - Applicant profiles (demographic and socio-economic).
  - **Sensitive features**: '*Statussex*', '*telephone*', '*foreign worker*'

- **Empirical setting**
  - 4 different classifiers (models) trained: 70% training & 30% test data
  - Used: SMOTE oversampling

- Question: Are these models fair?

Assessment w.r.t. some groupe fairness notions (German)



- **Fairest model**: dashed line (zero) indicates the optimal value
- **Classifiers**: AdaBoost (ADA), Bagging (BAG), Random Forest (RF), Logistic Regression (LR)
- **Sensitive features**: '*sex*', '*foreign worker*', '*telephone*'

Another example: Hate speech detection

- **Goal**: Classify tweets as hate speech or not
- Data: *Hate speech* dataset (tweets)*
- Two language variants
  - Standard American English & African-American English
- Unfair outcomes w.r.t. tweets written in a particular language variant

- Sensitive words : '*nigga*', '*nigger*', …

- Idea: Bag of Words (BoW) (Or: Groups of words)

*Davidson et al. **Automated hate speech detection and the problem of offensive language**. AAAI. 2017

Another example: Hate speech detection

Word	Without grouping		With grouping	
	Rank	**Contrib.**	**Rank**	**Contrib.**
niggah	18	0.149	23	0.03
nigger	15	0.164	21	0.031
nigguh	22	0.13	83	0.008
nig	12	0.202	65	0.011
nicca	22	0.107	39	0.018
nigga	20	0.125	12	0.067
white	25	0.087	36	0.018

References

- **AI & ML**
  - Russell, et al. **Artificial intelligence a modern approach**. Pearson Education, 2010.
  - Bishop, et al. **Pattern recognition and machine learning**. New York: Springer, 2006.

- **ALGORITMIC FAIRNESS**
  - Barocas et al. **Fairness in machine learning**. 2017.
  - Makhlouf, K. et al. **On the applicability of machine learning fairness notions**. SIGKDD Explorations, 2021.
  - Alves et al. **Reducing unintended bias of ml models on tabular and textual data**. In : DSAA, 2021.