Freie Universität Berlin

Bias in data-driven AI systems

Eirini Ntoutsi

Free University Berlin

Data Science Lecture Series@University of Pireaus (online), 10.05.2021

Reality check: Can algorithms discriminate?

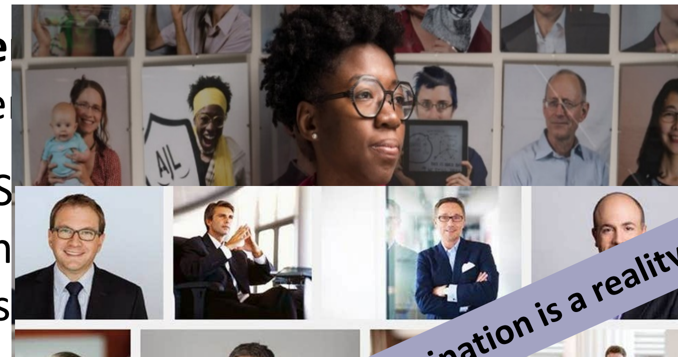
- Bloomberg analysts compared Amazon same-day delivery areas with U.S. Census Bureau data
- They found that in 6 major same-day delivery cities, the service area excludes predominantly black ZIP codes to varying degrees.

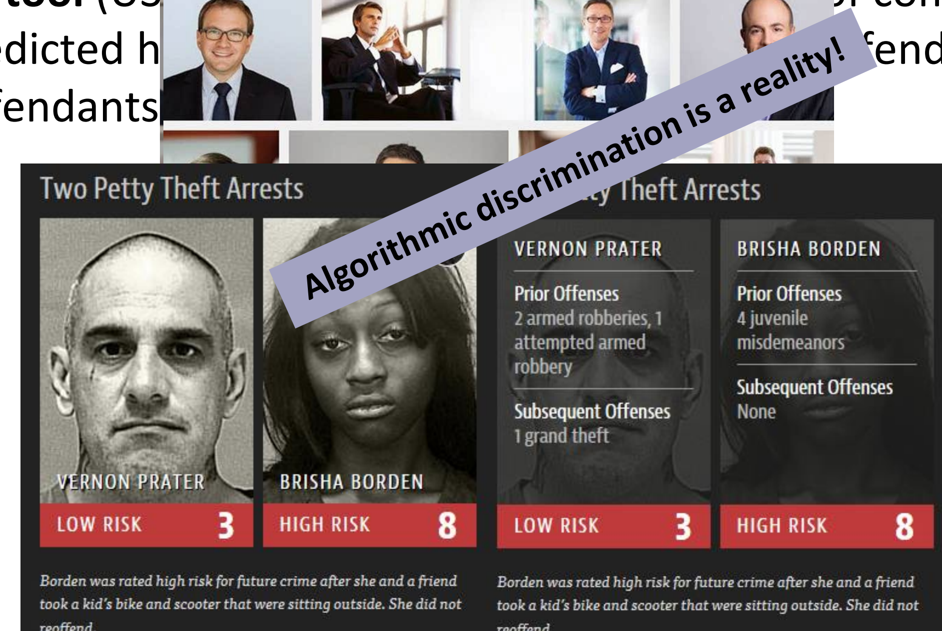
Source: https://www.bloomberg.com/graphics/2016-amazon-same-day/

- Shouldn’t this service be based on customer’s spend rather than race?
    - Amazon claimed that race was not used in their models.

Reality check cont': Can algorithms discriminate?

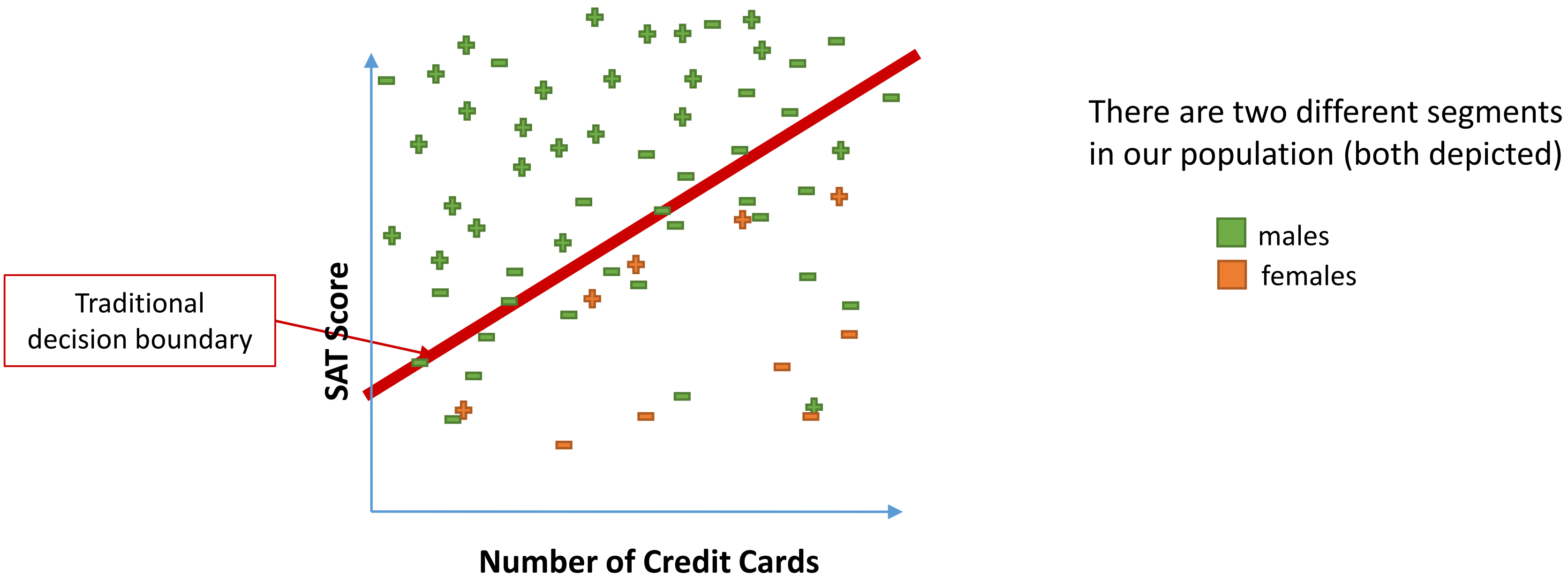
- There have been already plenty of cases of algorithmic discrimination
  - **State of the art visions systems** (used e.g. in autonomous driving) recognize better white males than black women (*racial and gender bias*)
  - **Google's AdFishe** was found to serve significantly fewe than men (*gender-bias*)
  - **COMPAS tool** (US of committing another crime predicted h fendants (and lower for white defendants





The myth of algorithmic objectivity and the need for fairness-aware machine learning

- Consider the following binary classification problem with classes: {+,-}. Consider also a binary protected attribute like gender {males, females}



- The goal of a traditional classifier (simple perceptron in this case) is to find the hypothesis (parameters of the line) that minimizes the empirical error.
  - This might incur discrimination (all female instances are rejected in our example)

The fairness-aware machine learning domain

- A young, fast evolving, multi-disciplinary research field
  - Bias/fairness/discrimination/… have been studied for long in philosophy, social sciences, law, …
- Don’t blame (only) the AI
  - “*Bias is as old as human civilization*” and “*it is human nature for members of the dominant majority to be oblivious to the experiences of other groups*”
  - Human bias: a prejudice in favour of or against one thing, person, or group compared with another usually in a way that’s considered to be unfair.
    - Bias triggers (protected attributes): ethnicity, race, age, gender, religion, sexual orientation …
  - Algorithmic bias: the inclination or prejudice of a decision made by an AI system which is for or against one person or group, especially in a way considered to be unfair.

Dealing with bias in data-driven AI systems

UNDERSTANDING BIAS

Socio-technical causes of bias
- Data generation
- Data collection
- Institutional bias

Bias manifestation in data
- Sensitive features & causal inferences
- Data representativeness
- Data modalities

Fairness definition
- Similarity-based
- Causal reasoning
- Predicted outcome
- Predicted & actual outcome
- Predicted probabilities & actual outcome

MITIGATING BIAS

Pre-processing
- Instance class modification
- Instance selection
- Instance weighting

In-processing
- Classification model adaptation
- Regularization / Loss function s.t. constraints
- Latent fair classes

Post-processing
- Confidence/probability score corrections
- Promoting/demoting boundary decisions
- Wrapping a fair classifier on top of a black-box baselearner

ACCOUNTING FOR BIAS

Bias-aware data collection
- Bias elicitation: individual assessors, mathematical pooling, group elicitation, consensus building
- Crowdsourcing

Describing and modelling bias
- Description and causal logics
- Ontological formalisms and reasoning

Explaining AI decisions
- Model explanation by approximation
- Inherently interpretable models
- Local behaviour explanation

LEGAL ISSUES

Regulations provisions
- Data accuracy (GDPR)
- Equality, prohibition of discrimination (CFR-EU)

Are data modifications legal?
- Intellectual Property issues
- Legal basis for data/model modification

Application of existing rules
- Applicability to algorithmic decision-making
- Limited scope of anti-discrimination law. Indirect discrimination

E. Ntoutsi, P. Fafalios, U. Gadiraju, V. Iosifidis, W. Nejdl, M.-E. Vidal, S. Ruggieri, F. Turini, S. Papadopoulos, E. Krasanakis, I. Kompatsiaris, K. Kinder-Kurlanda, C. Wagner, F. Karimi, M. Fernandez, H. Alani, B. Berendt, T. Kruegel, C. Heinze, K. Broelemann, G. Kasneci, T. Tiropanis, S. Staab"*Bias in data-driven artificial intelligence systems—An introductory survey*", WIREs Data Mining and Knowledge Discovery, 2020.

Outline

- Introduction
- Dealing with bias in data-driven AI systems
  - Understanding bias
  - Mitigating bias
  - Accounting for bias
- Case: bias-mitigation with sequential ensemble learners (boosting)
- Wrapping up

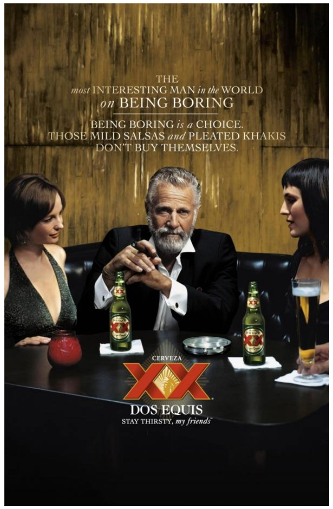
Understanding bias: Sociotechnical causes of bias

- AI-systems rely on data generated by humans (UGC) or collected via systems created by humans.

- As a result human biases
  - enter AI systems
    - e.g., bias in word-embeddings (Bolukbasi et al, 2016)
  - might be amplified by complex sociotechnical systems
    - e.g., the Web
  - new types of biases might be created

Understanding bias: How is bias manifested in data?

- Protected attributes and proxies
  - E.g., neighborhoods in U.S. cities are highly correlated with race
- Representativeness of data
  - E.g., underrepresentation of women and people of color in IT developer communities and image datasets
  - E.g., overrepresentation of black people in drug-related arrests
- Depends on data modalities



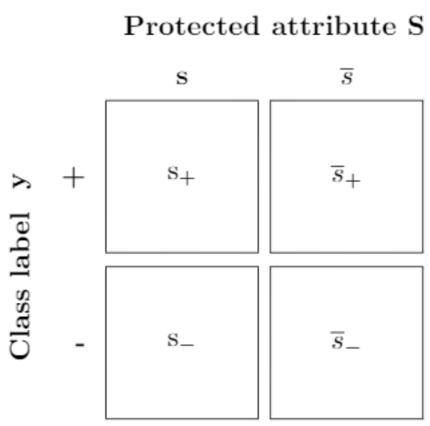
https://incitrio.com/top-3-lessons-learned-from-the-top-12-marketing-campaigns-ever/

https://ellengau.medium.com/emily-in-paris-asian-women-i-know-arent-like-mindy-chen-6228e63da333

Typical (batch) fairness-aware learning setup

- **Input:** D = training dataset drawn from a joint distribution P(F,S,y)
  - F: set of non-protected attributes
  - S: (typically: binary, single) protected attribute
    - *s* (*$\bar{s}$*): protected (non-protected) group
  - y = (typically: binary) class attribute {+,-} (+ for accepted, - for rejected)

	F1	F2	S	y
$User_1$	f_{11}	f_{12}	*s*	+
$User_2$	f_{21}			-
$User_3$	f_{31}	f_{23}	*s*	+
…	…	…	…	…
$User_n$	f_{n1}			+



- **Goal of fairness-aware classification**: Learn a mapping from *f(F, S)* → *y*
  - achieves good predictive performance → We know how to measure this
  - eliminates discrimination → According to some fairness measure

Measuring (un)fairness: some measures

	F1	F2	S	y	$\hat{y}$
User$_1$	f_{11}	f_{12}	s	+	-
User$_2$	f_{21}			-	+
User$_3$	f_{31}	f_{23}	s	+	-
…	…	…	…	…	…
User$_n$	f_{n1}			+	+

- Statistical parity: If subjects in both protected and unprotected groups should have *equal* probability *of being assigned* to the *positive class*

$$P(\hat{y} = +|S = s) = P(\hat{y} = +|S = \bar{s})$$

- Equal opportunity: There should be no difference in model's prediction errors regarding the positive class

$$P(\hat{y} \neq y|S = s_+) = P(\hat{y} \neq y|S = \bar{s}_+)$$

- Disparate Mistreatment: There should be no difference in *model's prediction errors* between protected and non-protected groups for *both classes*

$$\delta\mathrm{FNR} = P(\hat{y} \neq y|S = s_+) - P(\hat{y} \neq y|S = \bar{s}_+)$$
$$\delta\mathrm{FPR} = P(\hat{y} \neq y|S = s_-) - P(\hat{y} \neq y|S = \bar{s}_-)$$
$$\text{Disparate Mistreatment } = |\delta FNR| + |\delta FPR|$$

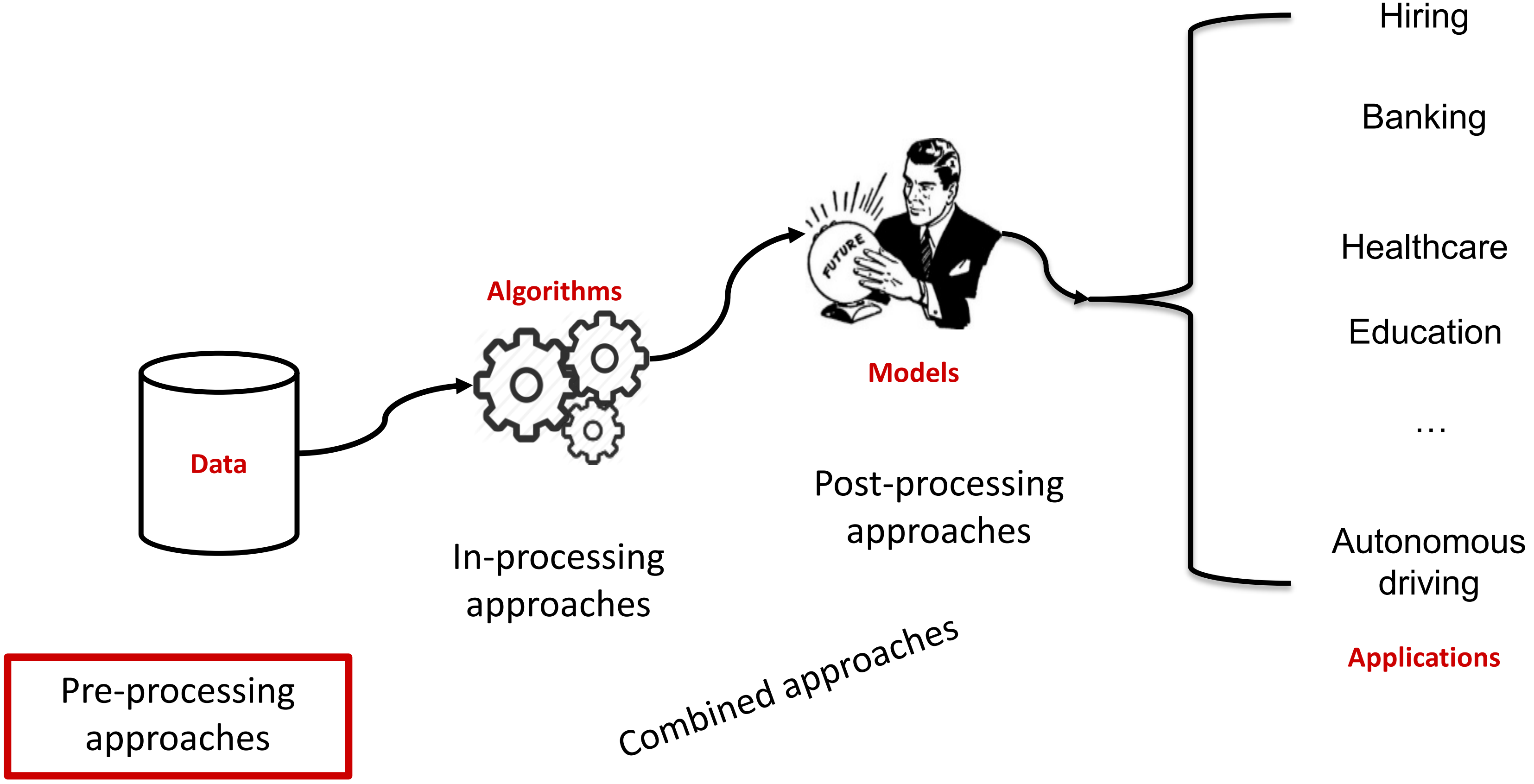
(Verma and Rubin, 2018)

Outline

- Introduction
- Dealing with bias in data-driven AI systems
  - Understanding bias
  - Mitigating bias
  - Accounting for bias
- Case: bias-mitigation with sequential ensemble learners (boosting)
- Wrapping up

Mitigating bias

- Bias can arise at any stage of the data-driven AI decision making



Mitigating bias: pre-processing approaches

- Intuition: making the data more fair will result in a less unfair model
- Idea: balance the protected and non-protected groups in the dataset
- Design principle: minimal data interventions (to retain data utility for the learning task)
- Different techniques:
  - Instance class modification (massaging), (Kamiran & Calders, 2009),(Luong, Ruggieri, & Turini, 2011)
  - Instance selection (sampling), (Kamiran & Calders, 2010) (Kamiran & Calders, 2012)
  - Instance weighting, (Calders, Kamiran, & Pechenizkiy, 2009)
  - Synthetic instance generation (Iosifidis & Ntoutsi, 2018)
  - …

Mitigating bias: pre-processing approaches: Massaging

- Change the class label of carefully selected instances (Kamiran & Calders, 2009).
  - The selection is based on a ranker which ranks the individuals by their probability to receive the favorable outcome.
  - The number of massaged instances depends on the fairness measure (group fairness)

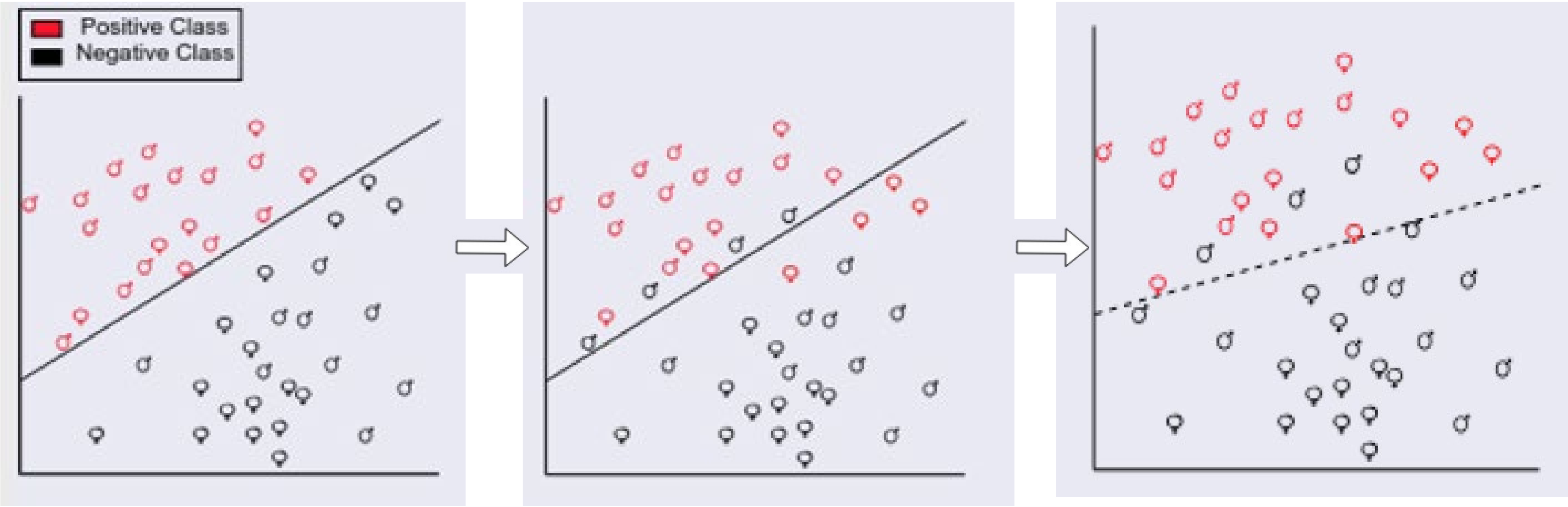
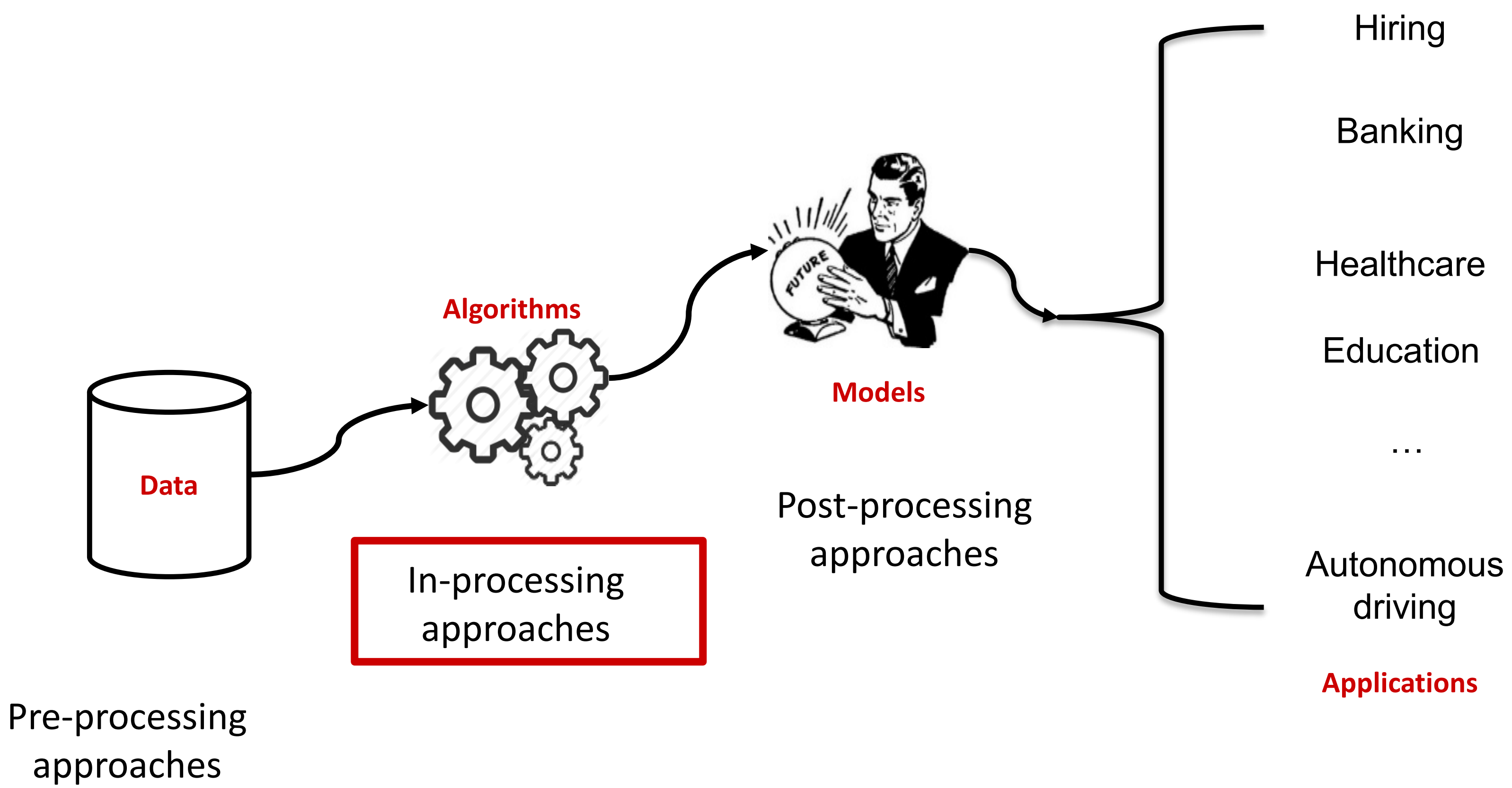


Image credit Vasileios Iosifidis

Mitigating bias

- Bias can arise at any stage of the data-driven AI decision making

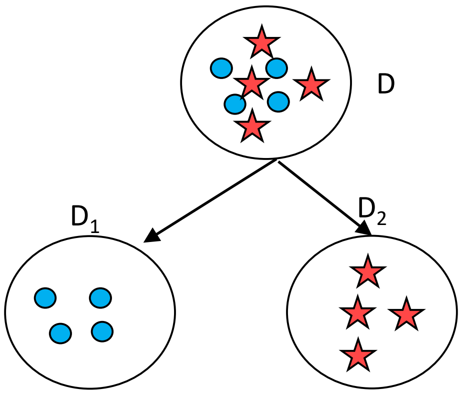


Mitigating bias: in-processing approaches

- Intuition: working directly with the algorithm allows for better control
- Idea: explicitly incorporate the model's discrimination behavior in the objective function
- Design principle: "balancing" predictive- and fairness-performance
- Different techniques:
  - Regularization (Kamiran, Calders & Pechenizkiy, 2010),(Kamishima, Akaho, Asoh & Sakuma, 2012), (Dwork, Hardt, Pitassi, Reingold & Zemel, 2012) (Zhang & Ntoutsi, 2019)
  - Constraints (Zafar, Valera, Gomez-Rodriguez & Gummadi, 2017)
  - Training on latent target labels (Krasanakis, Xioufis, Papadopoulos & Kompatsiaris, 2018)
  - In-training altering of data distribution (Iosifidis & Ntoutsi, 2019)
  - …

Mitigating bias: in-processing approaches: change the objective function

- We introduce the fairness gain of an attribute (FG)



$$FG(D,A) = |Disc(D)| - \sum_{v \in dom(A)} \frac{|D_v|}{|D|} |Disc(D_v)|$$

  - *Disc(D)* corresponds to statistical parity (group fairness)

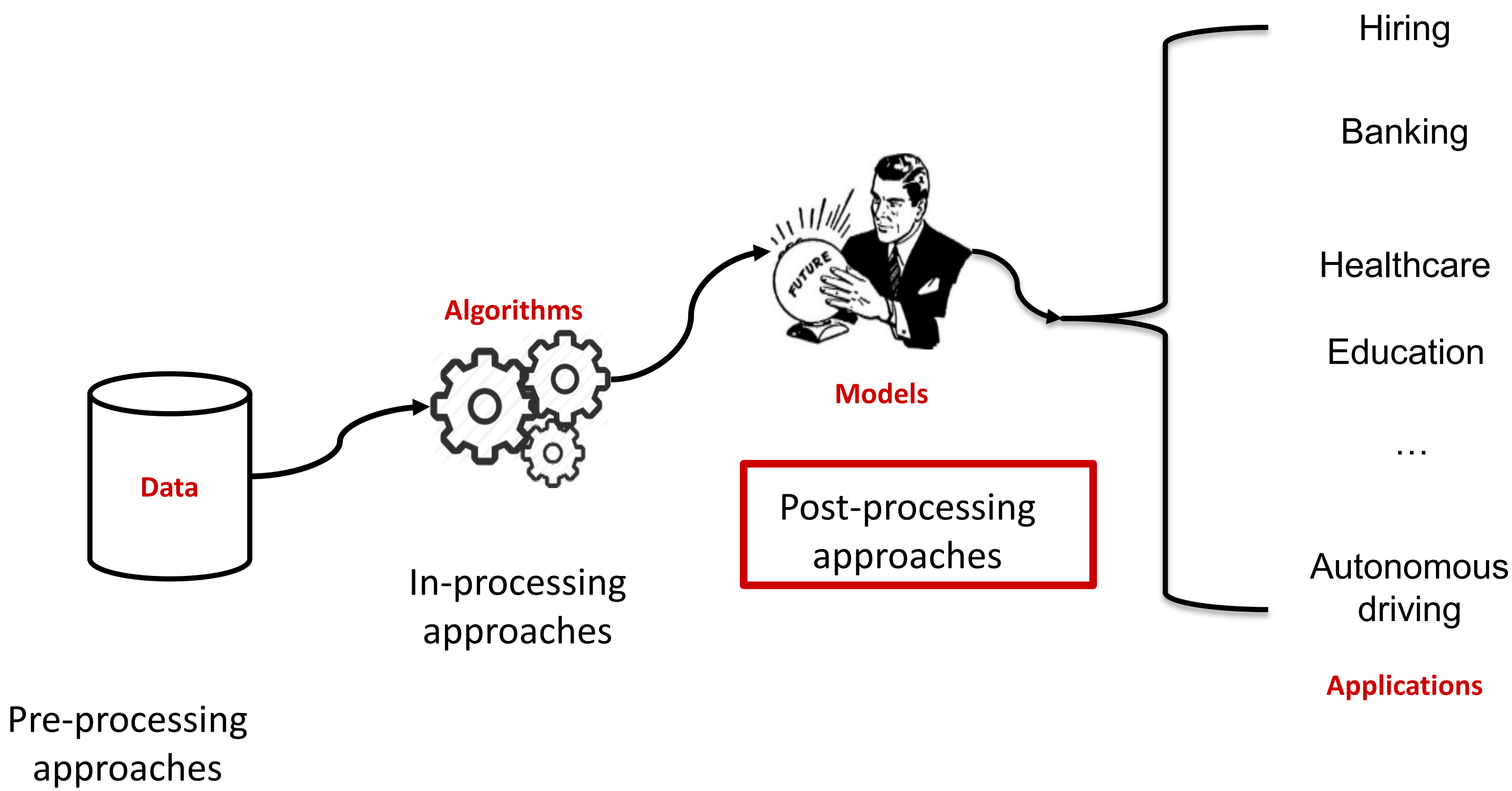
- We introduce the joint criterion, fair information gain (FIG) that evaluates the suitability of a candidate splitting attribute *A* in terms of both predictive performance and fairness.

$$FIG(D,A) = \begin{cases} IG(D,A) & , if FG(D,A) = 0 \\ IG(D,A) \times FG(D,A) & , otherwise \end{cases}$$

W. Zhang, E. Ntoutsi, "*An Adaptive Fairness-aware Decision Tree Classifier*", IJCAI 2019.

Mitigating bias

- Bias can arise at any stage of the data-driven AI decision making

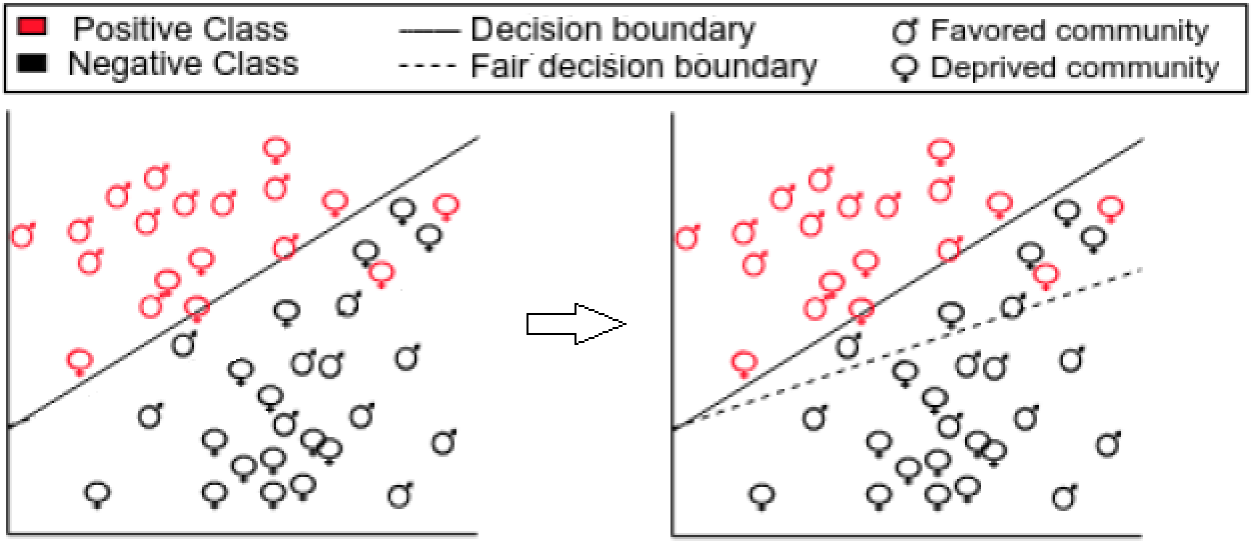


Mitigating bias: post-processing approaches

- Intuition: start with predictive performance
- Idea: first optimize the model for predictive performance and then tune for fairness
- Design principle: minimal interventions (to retain model predictive performance)
- Different techniques:
  - Correct the confidence scores (Pedreschi, Ruggieri, & Turini, 2009), (Calders & Verwer, 2010)
  - Correct the class labels (Kamiran et al., 2010)
  - Change the decision boundary (Kamiran, Mansha, Karim, & Zhang, 2018), (Hardt, Price, & Srebro, 2016)
  - Wrap a fair classifier on top of a black-box learner (Agarwal, Beygelzimer, Dudík, Langford, & Wallach, 2018)
  - …

Mitigating bias: post-processing approaches: shift the decision boundary

- An example of decision boundary shift



V. Iosifidis, H.T. Thi Ngoc, E. Ntoutsi, “*Fairness-enhancing interventions in stream classification*", DEXA 2019.

Outline

- Introduction
- Dealing with bias in data-driven AI systems
  - Understanding bias
  - Mitigating bias
  - Accounting for bias
- Case: bias-mitigation with sequential ensemble learners (boosting)
- Wrapping up

Accounting for bias

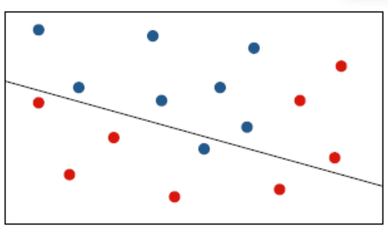
- Algorithmic accountability refers to the assignment of responsibility for how an algorithm is created and its impact on society (Kaplan et al, 2019).
- Many facets of accountability for AI-driven algorithms and different approaches
  - Proactive approaches:
    - bias-aware data collection, e.g., for Web data, crowd-sourcing
    - bias-description and modeling, e.g., via ontologies
    - ...
  - Retroactive approaches:
    - Explaining AI decisions in order to understand whether decisions are biased
      - What is an explanation? Explanations w.r.t. legal/ethical grounds?
      - Using explanations for fairness-aware corrections (inspired by Schramowski et al, 2020)

Outline

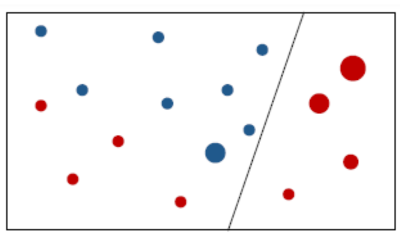
- Introduction
- Dealing with bias in data-driven AI systems
  - Understanding bias
  - Mitigating bias
  - Accounting for bias
- Case: bias-mitigation with sequential ensemble learners (boosting)
- Wrapping up

Fairness with sequential learners (boosting)

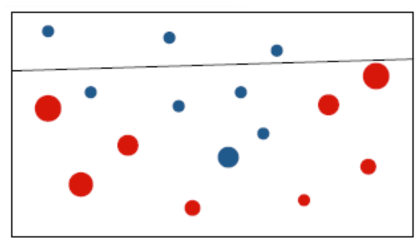
- Sequential ensemble methods generate *base learners* in a *sequence*
- The sequential generation of base learners promotes the dependence between the base learners.
  - Each learner learns from the mistakes of the previous predictor
- The *weak* learners are combined to build a *strong* learner
- Popular examples: Adaptive Boosting (AdaBoost), Extreme Gradient Boosting (XGBoost).
- Our base model is AdaBoost (Freund and Schapire, 1995), a sequential ensemble method that in each round, re-weights the training data to focus on misclassified instances.



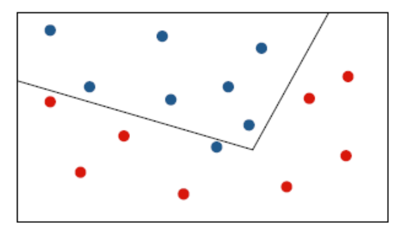
Round 1: Weak learner h_1



Round 2: Weak learner h_2



Round 3: Weak learner h_3

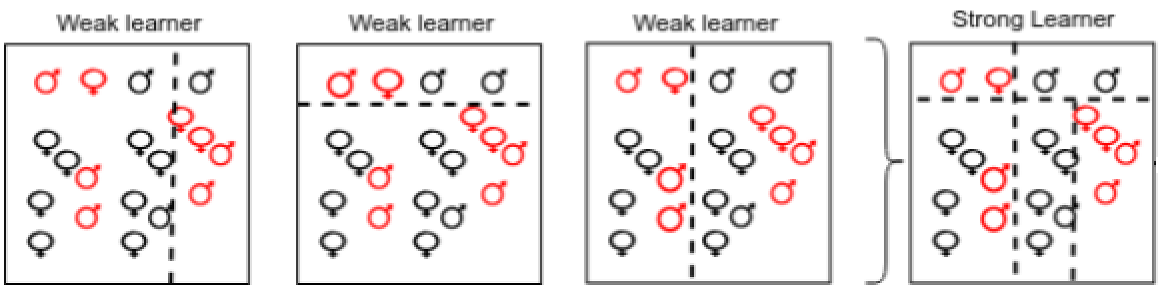


Final strong learner *H()*

$$H(x) = \sum_{j=1}^{T} \alpha_i h_j(x)$$

Intuition behind using boosting for fairness

1. It is easier to make "fairness-related interventions" in simpler models rather than complex ones
2. We can use the whole sequence of learners for the interventions instead of the current one



Limitations of related work

- Existing works evaluate predictive performance in terms of the *overall classification error rate (ER)*, e.g., [Calders et al'09, Calmon et al'17, Fish et al'16, Hardt et al'16, Krasanakis et al'18, Zafar et al'17]
- In case of class-imbalance, ER is misleading
  - Most of the datasets however suffer from imbalance

	Adult Census	Bank	Compass	KDD Census
#Instances	45,175	40,004	5,278	299,285
#Attributes	14	16	9	41
Sen.Attr.	Gender	Marit. Status	Gender	Gender
Class ratio (+:−)	1:3.03	1:7.57	1:1.12	1:15.11
Positive class	*>50K*	*subscription*	*recidivism*	*>50K*

- Moreover, *Dis.Mis. is* "oblivious" to the class imbalance problem

Example

- Positive class << Negative class e.g., $|s^+| + |\bar{s}^+| = 5\%, |s^-| + |\bar{s}^-| = 95\%$
- Model classifies everything as negative.
- Accuracy is still high (95%) and model is "fair" i.e., $\delta FNR = 0, \delta FPR = 0$

From Adaboost to AdaFair

- We tailor AdaBoost to fairness
  - We introduce the notion of *cumulative fairness* that assesses the fairness of the model *up* to the current boosting round (partial ensemble).
  - We directly incorporate fairness in the *instance weighting* process (traditionally focusing on classification performance).
  - We optimize the number of weak learners in the final ensemble based on *balanced error rate* thus directly considering *class imbalance* in the best model selection.

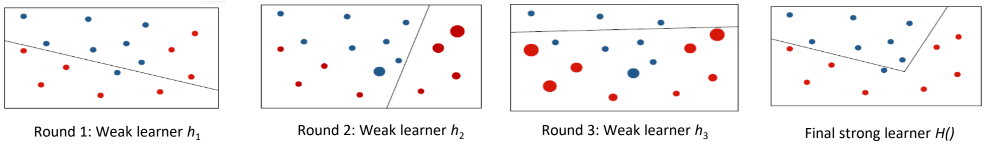
$$BER = 1 - \frac{1}{2} \cdot \left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP}\right) = 1 - \frac{1}{2} \cdot (TPR + TNR)$$

$$ER = 1 - \frac{TP + TN}{TP + FN + TN + FP}$$

V. Iosifidis, E. Ntoutsi, *"AdaFair: Cumulative Fairness Adaptive Boosting"*, ACM CIKM 2019.

AdaFair: Cumulative boosting fairness

- Let *j*: 1−*T* be the current boosting round, *T* is user defined
- Let $H_{1:j}(x) = \sum_{i=1}^{j} a_i h_i(x)$ be the *partial ensemble*, up to current round *j*.



- The cumulative fairness of the ensemble up to round *j,* is defined based on the parity in the predictions of the partial ensemble between protected and non-protected groups for both classes

$$\delta FNR^{1:j} = \frac{\sum_{i=1}^{|\bar{s}_+|} 1 \cdot \mathbb{I}[\sum_{k=1}^{j} a_k h_k(x_i^{\bar{s}+}) \neq y_i]}{|\bar{s}_+|} - \frac{\sum_{i=1}^{|s_+|} 1 \cdot \mathbb{I}[\sum_{k=1}^{j} a_k h_k(x_i^{s+}) \neq y_i]}{|s_+|}$$

$$\delta FPR^{1:j} = \frac{\sum_{i=1}^{|\bar{s}_-|} 1 \cdot \mathbb{I}[\sum_{k=1}^{j} a_k h_k(x_i^{\bar{s}-}) \neq y_i]}{|\bar{s}_-|} - \frac{\sum_{i=1}^{|s_-|} 1 \cdot \mathbb{I}[\sum_{k=1}^{j} a_k h_k(x_i^{s-}) \neq y_i]}{|s_-|}$$

- ``Forcing'' the model to consider ``historical'' fairness over all previous rounds instead of just focusing on current round $h_j()$ results in better classifier performance and model convergence.

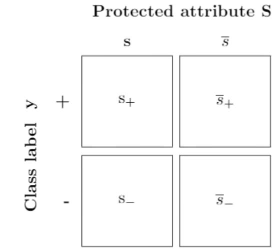
AdaFair: fairness-aware weighting of instances

- Vanilla AdaBoost already boosts misclassified instances for the next round
- Our weighting *explicitly* targets fairness by extra boosting discriminated groups for the next round
- The data distribution at boosting round *j*+1 is updated as follows

$$w_i \leftarrow \frac{1}{Z_j} w_i \cdot e^{\alpha_j \cdot \hat{h}_j(x) \cdot \mathbb{I}(y_i \neq h_j(x_i))} \cdot (1 + u_i)$$

- The fairness-related cost u_i of instances $x_i \in D$ which belong to a group that is discriminated is defined as follows:

$$u_i = \begin{cases} |\delta FNR^{1:j}|, & if\ \mathbb{I}((y_i \neq h_j(x_i)) \wedge |\delta FNR^{1:j}| > \epsilon), x_i \in s_+, \delta FNR^{1:j} > 0 \\ |\delta FNR^{1:j}|, & if\ \mathbb{I}((y_i \neq h_j(x_i)) \wedge |\delta FNR^{1:j}| > \epsilon), x_i \in \bar{s}_+, \delta FNR^{1:j} < 0 \\ |\delta FPR^{1:j}|, & if\ \mathbb{I}((y_i \neq h_j(x_i)) \wedge |\delta FPR^{1:j}| > \epsilon), x_i \in s_-, \delta FPR^{1:j} > 0 \\ |\delta FPR^{1:j}|, & if\ \mathbb{I}((y_i \neq h_j(x_i)) \wedge |\delta FPR^{1:j}| > \epsilon), x_i \in \bar{s}_-, \delta FPR^{1:j} < 0 \\ 0, & otherwise \end{cases}$$



AdaFair: optimizing the number of weak learners

- Typically, the number of boosting rounds/ weak learners *T* is user-defined
- We propose to select the optimal subsequence of learners 1 … θ, $\theta \leq T$ that minimizes the balanced error rate (BER)
- In particular, we consider both ER and BER in the objective function

$$argmin_{\theta}(c * BER_{\theta} + (1-c)ER_{\theta} + Mis.Dis.)$$

- The result of this optimization if a final ensemble model with *Mis.Dis.* fairness

$$H(x) = \sum_{i=1}^{\theta} a_i h_i(x)$$

Experimental evaluation

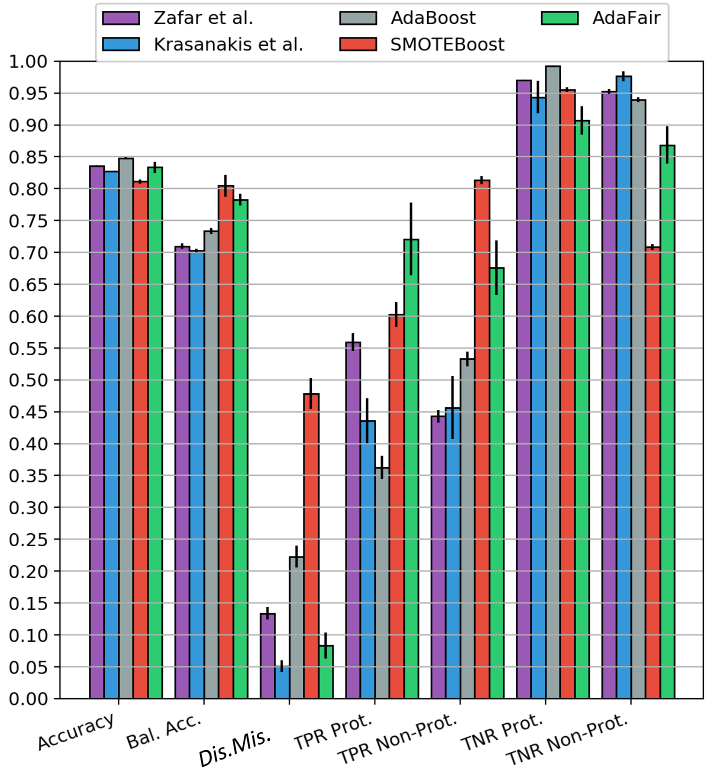
- Datasets of varying imbalance

	Adult Census	Bank	Compass	KDD Census
#Instances	45,175	40,004	5,278	299,285
#Attributes	14	16	9	41
Sen.Attr.	Gender	Marit. Status	Gender	Gender
Class ratio (+:−)	1:3.03	1:7.57	1:1.12	1:15.11
Positive class	*>50K*	*subscription*	*recidivism*	*>50K*

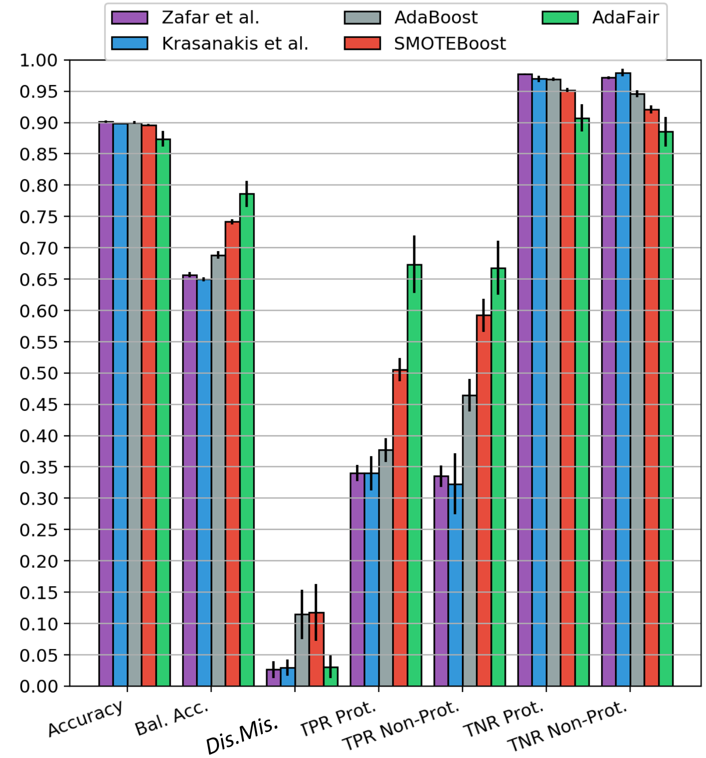
- Baselines
  - AdaBoost [Sch99]: vanilla AdaBoost
  - SMOTEBoost [CLHB03]: AdaBoost with SMOTE for imbalanced data.
  - Krasanakis et al. [KXPK18]: Boosting method which minimizes *Dis.Mis.* by approximating the underlying distribution of hidden correct labels.
  - Zafar et al.[ZVGRG17]: Training logistic regression model with convex-concave constraints to minimize *Dis.Mis.*
  - AdaFair NoCumul: Variation of AdaFair that computes the fairness weights based on individual weak learners.

Experiments: Predictive and fairness performance

- Adult census income (ratio 1+:3-)



- Bank dataset (ratio 1+:8-)



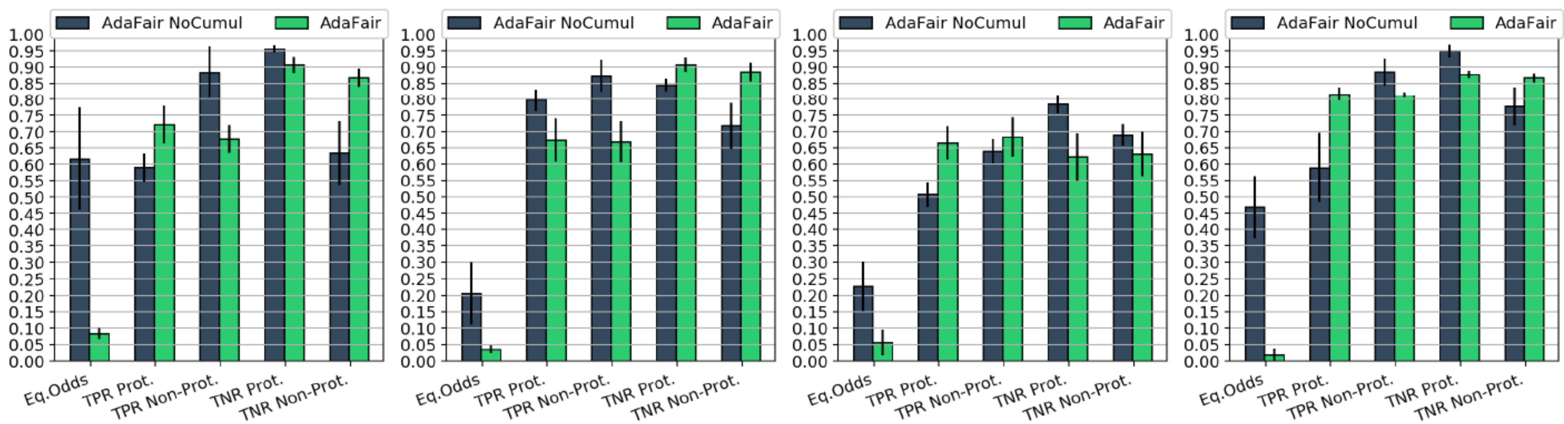
Larger values are better, for Dis.Mis. lower values are better

- Our method achieves high balanced accuracy and low discrimination (*Dis.Mis.*) while maintaining high TPRs and TNRs for both groups.
- The methods of Zafar et al and Krasanakis et al, eliminate discrimination by rejecting more positive instances (lowering TPRs).

Cumulative vs non-cumulative fairness

Please note: ~~Eq.Odds~~ Dis.Mis.

- Cumulative vs non-cumulative fairness impact on model performance



(a) Adult census (b) Bank (c) Compass (d) KDD census

  - Cumulative notion of fairness performs better
  - The cumulative model (AdaFair) is more *stable* than its non-cumulative counterpart (standard deviation is higher)

Outline

- Introduction
- Dealing with bias in data-driven AI systems
  - Understanding bias
  - Mitigating bias
  - Accounting for bias
- Case: bias-mitigation with sequential ensemble learners (boosting)
- Wrapping up

Wrapping-up, ongoing work and future directions

- In this talk I focused on the myth of algorithmic objectivity and
  - the reality of algorithmic bias and discrimination and how algorithms can pick biases existing in the input data and further reinforce them
- A large body of research already exists but
  - focuses mainly on fully-supervised batched learning with single-protected (and typically binary) attributes with binary classes
    - Moving from batch learning to online learning
  - targets bias in some step of the analysis-pipeline, but biases/errors might be propagated and even amplified (unified approached are needed)
    - Moving from isolated approaches (pre-, in- or post-) to combined approaches

T. Hu, V. Iosifidis, W. Liao, H. Zang, M. Yang, *E. Ntoutsi,B. Rosenhahn, "FairNN - Conjoint Learning of Fair Representations for Fair Decisions", DS 2020.*

V. Iosifidis, E. Ntoutsi, "FABBOO - Online Fairness-aware Learning under Class Imbalance", DS 2020.

Wrapping-up, ongoing work and future directions

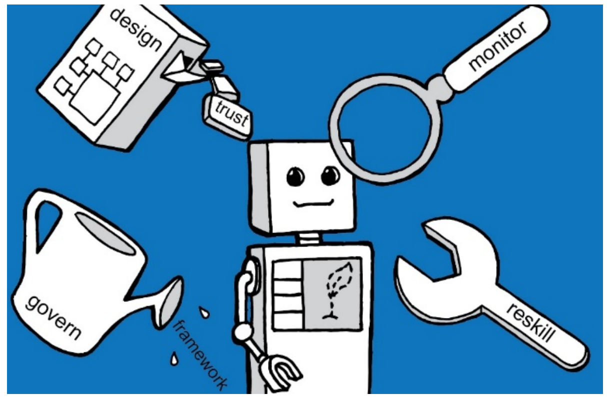
- Moving from single-protected attribute fairness-aware learning to multi-fairness
  - Existing legal studies define multi-fairness as compound, intersectional and overlapping [Makkonen 2002].
- Moving from fully-supervised learning to unsupervised and reinforcement learning
- Moving from myopic (maximize short-term effect/immediate performance) solutions to non-myopic ones (that consider long-term effects) [Zhang et al,2020]
- Actionable approaches (counterfactual generation)

A.Roy, V. Iosifidis, E. Ntoutsi, "Multi-Fair Pareto Boosting", arXiv

P. Naumann, E. Ntoutsi, "Consequence-aware Sequential Counterfactual Generation", arXiv

Thank you for you attention!

Questions?



NoBIAS
Artificial Intelligence without Bias

https://nobias-project.eu/
@NoBIAS_ITN

BIAS

https://www.bias-project.org/

LernMINT

https://lernmint.org/

Feel free to contact me:

- eirini.ntoutsi@fu-berlin.de
- @entoutsi
- https://www.mi.fu-berlin.de/en/inf/groups/ag-KIML/index.html