Fairness in ML

Data, algorithms and discrimination

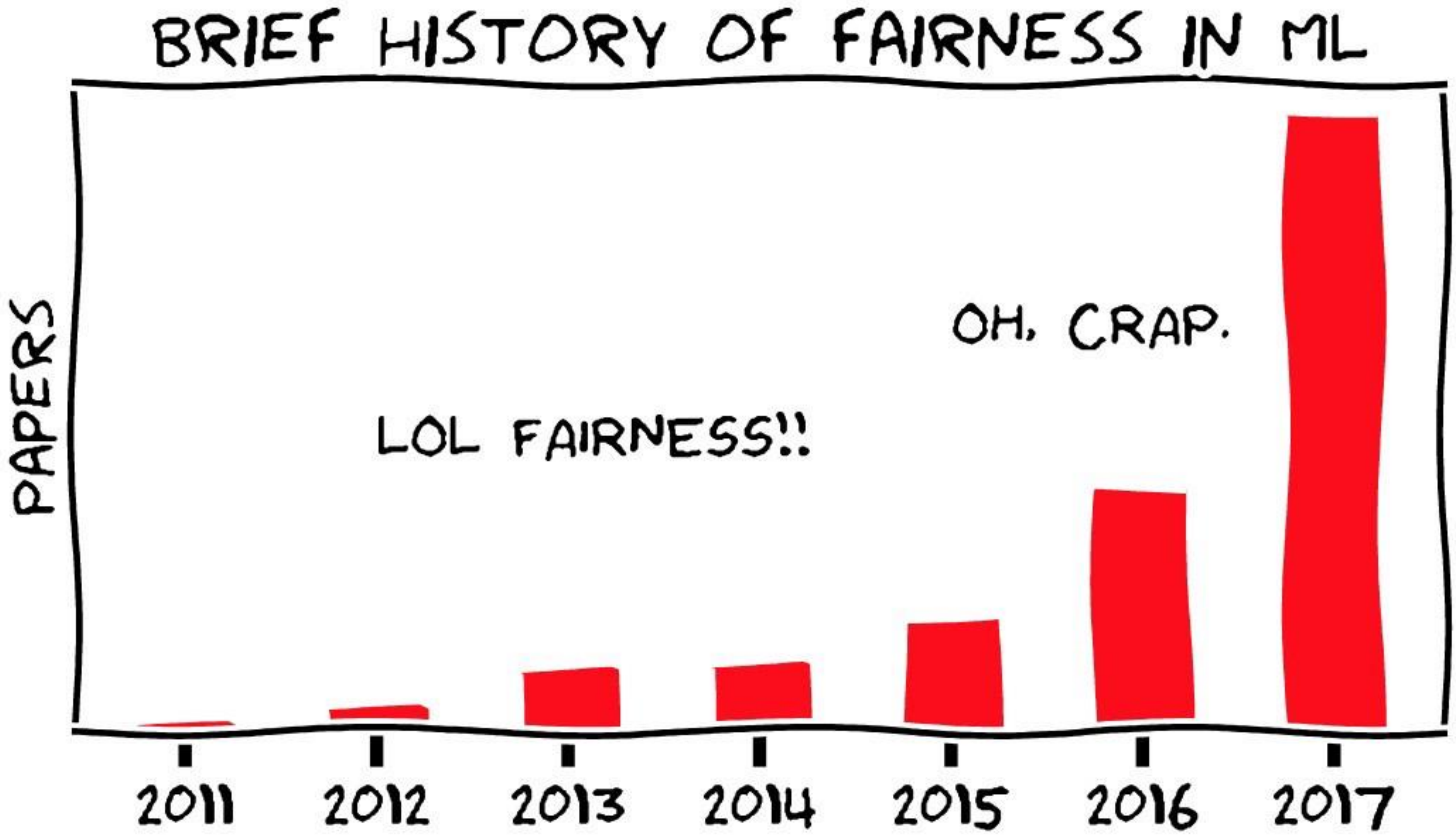
2023.07.06 Intesa Sanpaolo Academy
AI4CITIZENS_DATA & AI ETHICS

daniele regoli

Data Science & AI @ Intesa Sanpaolo

daniele.regoli@intesasanpaolo.com

Why we talk about fairness in ML



credits: Moritz Hardt

risks for people

data as a social mirror

ML could amplify and perpetuate biases already present in data, at large scale

sample size imbalances

ML could disgregard minority groups, effectively producing bias even if absent in the data

this can have a huge impact on people's lives
e.g. Recruiting / Loans approval
but also, in more indirect ways, in *recommendations*

«How big data is unfair», Hardt (2014)
«A survey of bias in Machine Learning» Mehrabi et al. *ACM Computing Surveys (2021)*

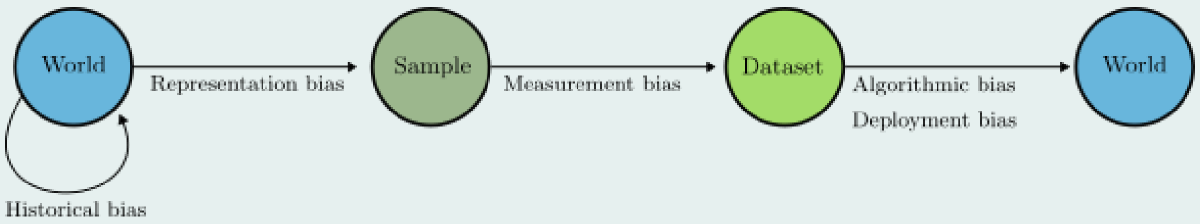
bias types

historical/life bias
when some group is systematically unfavoured e.g. for cultural reasons (gender bias)

measurement bias
when the variables we employ are a distorted version of what we really want (e.g. QI for intelligence)

Representation bias
when the data we use are skewed with respect to the whole population

...



[“Bias on Demand: A Modelling Framework that Generates Synthetic Data with Bias”, Baumann, Castelnovo, Crupi, Inverardi, Regoli, FAcct (2023)]

risks for companies

risks for companies

WILL KNIGHT BUSINESS 11.19.2019 09:15 AM

WIRED

The Apple Card Didn't 'See' Gender—and That's the Problem

The way its algorithm determines credit lines makes the risk of bias more acute.

risks for companies

WILL KNIGHT BUSINESS 11.19.2019 09:15 AM

WIRED

The Apple Card Didn't 'See' Gender—and That's the Problem

The way its algorithm determines credit lines makes the risk of bias more acute.

HOME > TECH

Amazon built an AI tool to hire people but had to shut it down because it was discriminating against women

Isobel Asher Hamilton Oct 10, 2018, 11:47 AM

INSIDER

risks for companies

WILL KNIGHT BUSINESS 11.19.2019 09:15 AM

WIRED

The Apple Card Didn't 'See' Gender—and That's the Problem

The way its algorithm determines credit lines makes the risk of bias more acute.

PROPUBLICA

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

May 23, 2016

HOME > TECH

Amazon built an AI tool to hire people but had to shut it down because it was discriminating against women

INSIDER

Isobel Asher Hamilton Oct 10, 2018, 11:47 AM

risks for companies

WILL KNIGHT BUSINESS 11.19.2019 09:15 AM

The Apple Card Didn't 'S Problem

The way its algorithm determines credit lines makes t

GOOGLE IS POISONING ITS REPUTATION WITH AI RESEARCHERS

The firing of top Google AI ethics researchers has created a significant backlash

By James Vincent | Apr 13, 2021, 9:30am EDT

THE VERGE

PROPUBLICA

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

May 23, 2016

HOME > TECH

Amazon built an AI tool to hire people but had to shut it down because it was discriminating against women

Isobel Asher Hamilton Oct 10, 2018, 11:47 AM

INSIDER

AI Regulation

EUROPEAN COMMISSION

Brussels, 21.4.2021

COM(2021) 206 final

2021/0106(COD)

Proposal for a

REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL

LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS

Article 10

Data and data governance

Training, validation and testing data sets shall be subject to appropriate data governance and management practices. Those practices shall concern in particular,

(a) the relevant design choices;

(b) data collection;

(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;

(e) a prior assessment of the availability, quantity and suitability of the data sets that are needed;

(f) examination in view of possible biases;

(g) the identification of any possible data gaps or shortcomings, and how those gaps and shortcomings can be addressed.

different concepts of fairness

Southern State Parkway bridges – Robert Moses

legal principles

- Discrimination is not a clear-cut concept
- Discrimination is domain specific
- Even given a very specific situation, reaching an agreement about what is fair is far from easy

«protected» attributes

Costituzione Italiana – Art.3

*Tutti i cittadini hanno pari dignità sociale e sono eguali davanti alla legge, senza distinzione di **sesso**, di **razza**, di **lingua**, di **religione**, di **opinioni politiche**, di **condizioni personali e sociali**.*

È compito della Repubblica rimuovere gli ostacoli di ordine economico e sociale, che, limitando di fatto la libertà e l'eguaglianza dei cittadini, impediscono il pieno sviluppo della persona umana e l'effettiva partecipazione di tutti i lavoratori all'organizzazione politica, economica e sociale del Paese.

Legally recognized 'protected classes'

Race (Civil Rights Act of 1964); **Color** (Civil Rights Act of 1964); **Sex** (Equal Pay Act of 1963; Civil Rights Act of 1964); **Religion** (Civil Rights Act of 1964); **National origin** (Civil Rights Act of 1964); **Citizenship** (Immigration Reform and Control Act); **Age** (Age Discrimination in Employment Act of 1967); **Pregnancy** (Pregnancy Discrimination Act); **Familial status** (Civil Rights Act of 1968); **Disability status** (Rehabilitation Act of 1973; Americans with Disabilities Act of 1990); **Veteran status** (Vietnam Era Veterans' Readjustment Assistance Act of 1974; Uniformed Services Employment and Reemployment Rights Act); **Genetic information** (Genetic Information Nondiscrimination Act)

credits: Solon Barocas & Moritz Hardt. NeurIPS 2017 tutorial

legal principles

DISPARATE TREATMENT

procedural / deonthological

don’t employ sensitive information

should decide which info is really relevant for the problem

DISPARATE IMPACT

focus on impact / consequentialist

final decision independent of sensitive information

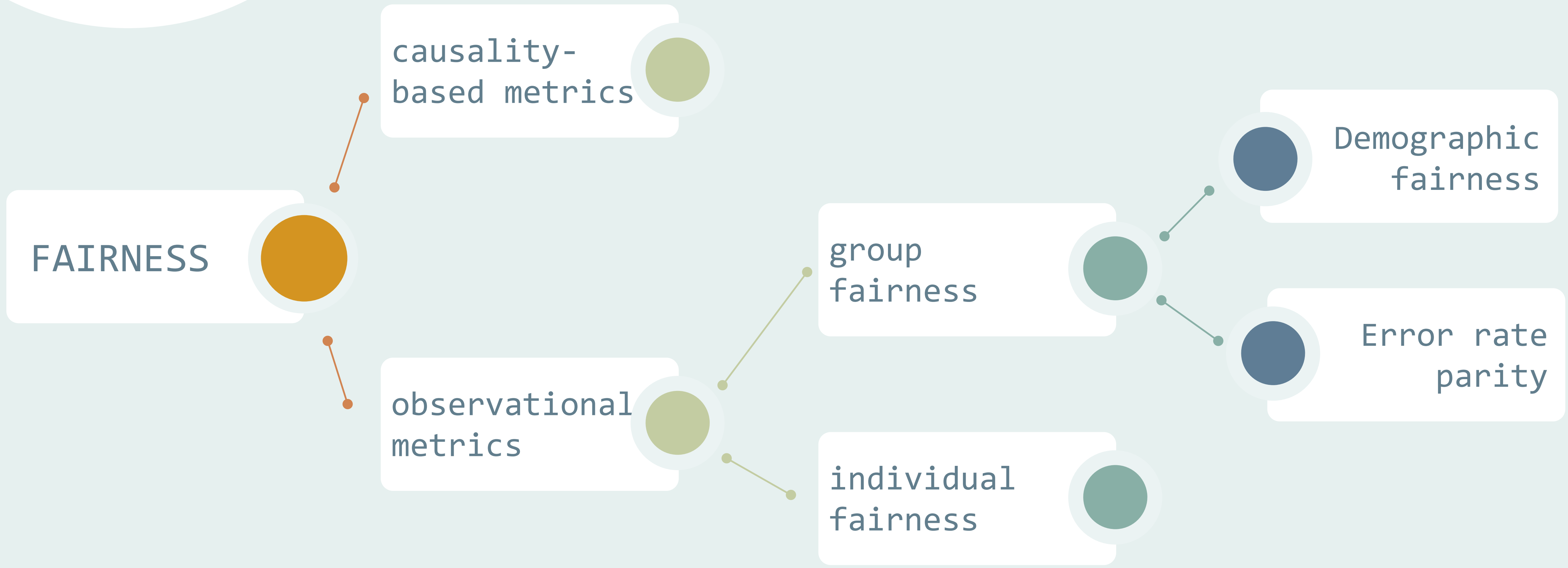
if not, justifications are needed

[“Big Data’s disparate impact”, Barocas and Selbst, Calif. L. Review (2016)]

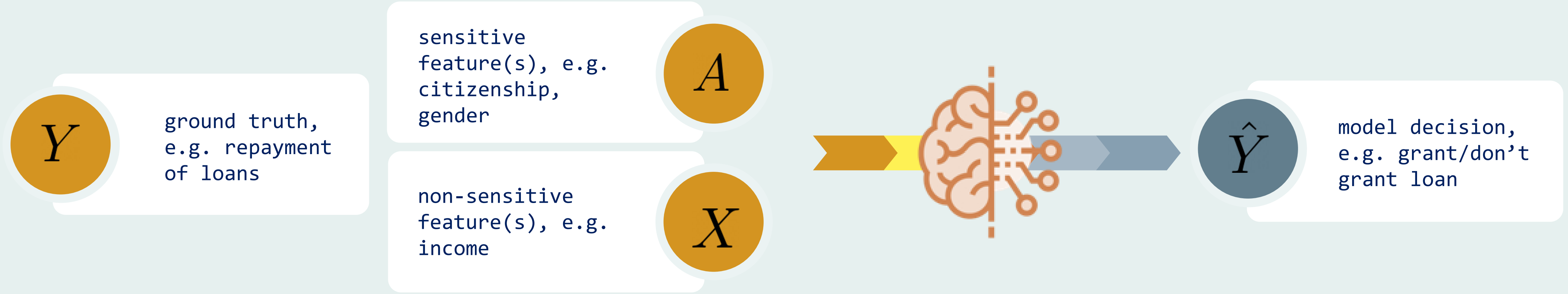
the zoo of fairness metrics

assessing fairness

there are **a lot of different definitions** of fairness, in general non compatible with one another



Machine Learning model



INDEPENDENCE $\hat{Y} \perp\!\!\!\perp A$

SEPARATION $\hat{Y} \perp\!\!\!\perp A \mid Y$

SUFFICIENCY $Y \perp\!\!\!\perp A \mid \hat{Y}$

group fairness criteria

Fairness and Machine Learning, Solon Barocas and Moritz Hardt and Arvind Narayanan (2019) https://fairmlbook.org/

Independence

~in line with the disparate impact principle

$$\hat{Y} \perp\!\!\!\perp A$$

$$P(\hat{Y}=1 \mid A=a) = P(\hat{Y}=1 \mid A=b), \quad \forall a, b$$

same percentage of loans granted to men and women

also known as **DEMOGRAPHIC PARITY (DP)** or **STATISTICAL PARITY**

$$\frac{P(Y=1 \mid A=a)}{P(Y=1 \mid A=b)} > 1 - \epsilon$$

DP ratio - *4/5 rule*

An important variant

CONDITIONAL DEMOGRAPHIC PARITY

$$\hat{Y} \perp\!\!\!\perp A \mid R$$

$$P(\hat{Y}=1 \mid R=r, A=a) = P(\hat{Y}=1 \mid R=r, A=b), \forall a, b, r$$

given some characteristics, same percentage of loans granted to men and women

Separation

$$\hat{Y} \perp\!\!\!\perp A \mid Y$$

$$P(\hat{Y} = 1 \mid A = a, Y = y) = P(\hat{Y} \mid A = b, Y = y), \quad \forall a, b, y$$

same error rates for men and women

related to **Equality of Opportunity / Predictive Equality / Equality of Odds,**

namely requires the parity of **recall** (true positive rate) and/or **false positive rate → ROC curve**

you need to put a lot of trust on the target Y!

Sufficiency

$$Y \perp\!\!\!\perp A \mid \hat{Y}$$

related to **Predictive Parity**

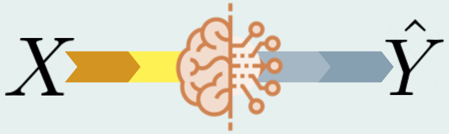
namely requires the parity of **precision**, i.e. it's the «other side of the coin» with respect to Equality of Odds

Sufficiency ***on score*** is implied by **calibration by group**

$$P(Y = 1 \mid \text{score} = s, A = a) = s, \quad s \in [0, 1], \forall a$$

~in line with the disparate treatment principle

FAIRNESS THROUGH UNAWARENESS / BLINDNESS



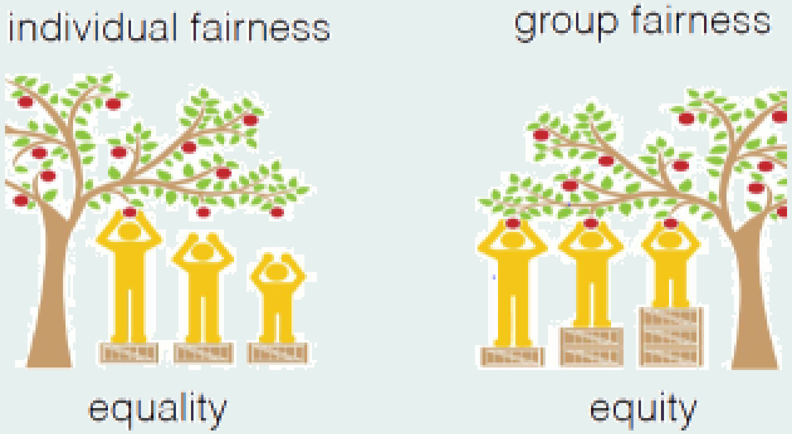
model's outcomes are functions of non-sensitive features only

$$\hat{Y} = f(X)$$

FAIRNESS THROUGH AWARENESS

$$D(x_1, x_2) \leq Cd(h(x_1), h(x_2)), \quad \forall x_1, x_2 \in \mathcal{X}$$

similar individuals are given similar decisions

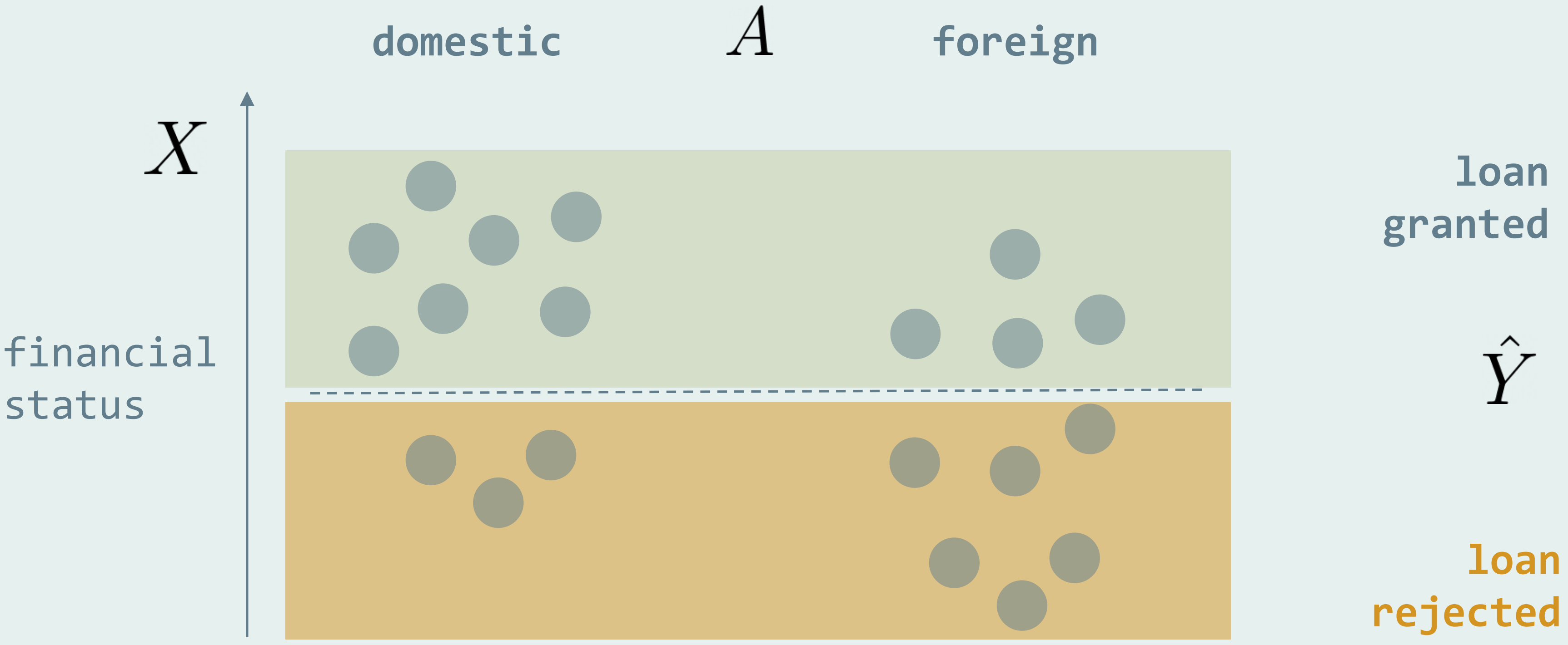


individual fairness criteria

"Fairness through awareness", Dwork, Cynthia, et al. *Proceedings of the 3rd innovations in theoretical computer science conference*. 2012

the devil is in the details

group fairness vs individual fairness



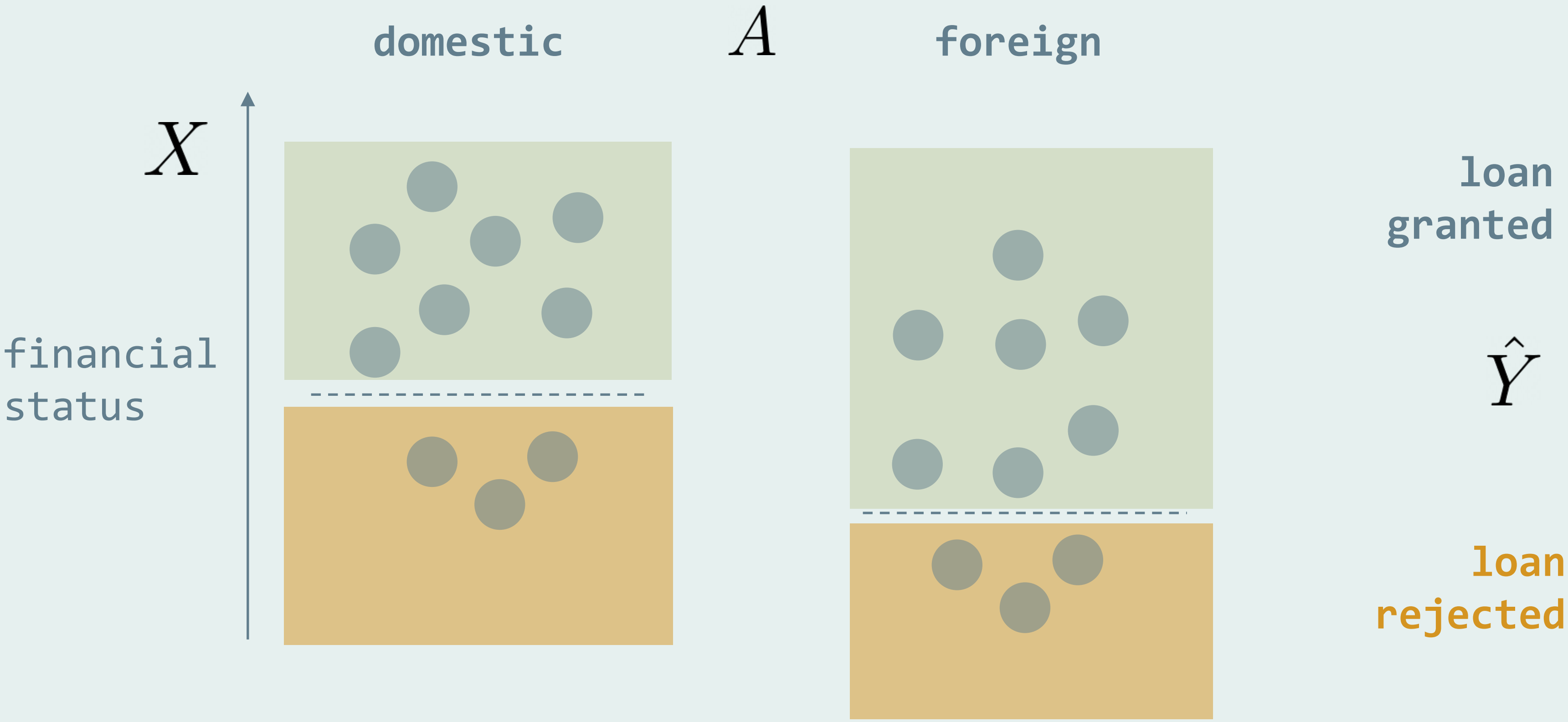
$\hat{Y} = f(X)$

fairness through unawareness:
don't use explicitely the sensitive variable

the devil is in the details

widely cited **4/5 rule** of the Uniform Guidelines on Employee Selection Procedures

group fairness vs individual fairness



$\hat{Y} \perp\!\!\!\perp A$

demographic parity:
same acceptance rate for different groups

LIMITS

INDEPENDENCE

in general, the perfect predictor is not compliant

incentivize laziness: accept random individuals from the unfavoured group

this could lead to an exacerbation of the bias!

this is reasonable when we want to break the status-quo, but we need to be very careful at consequences.
Need to distinguish the long-term goal (where we aim at independence) and algorithmic actions.
Maybe it is useless or even *harmful* to impose Demographic Parity

INDIVIDUAL FAIRNESS

Hard to define a task-based similarity

blindness has the obviuous problem of proxies

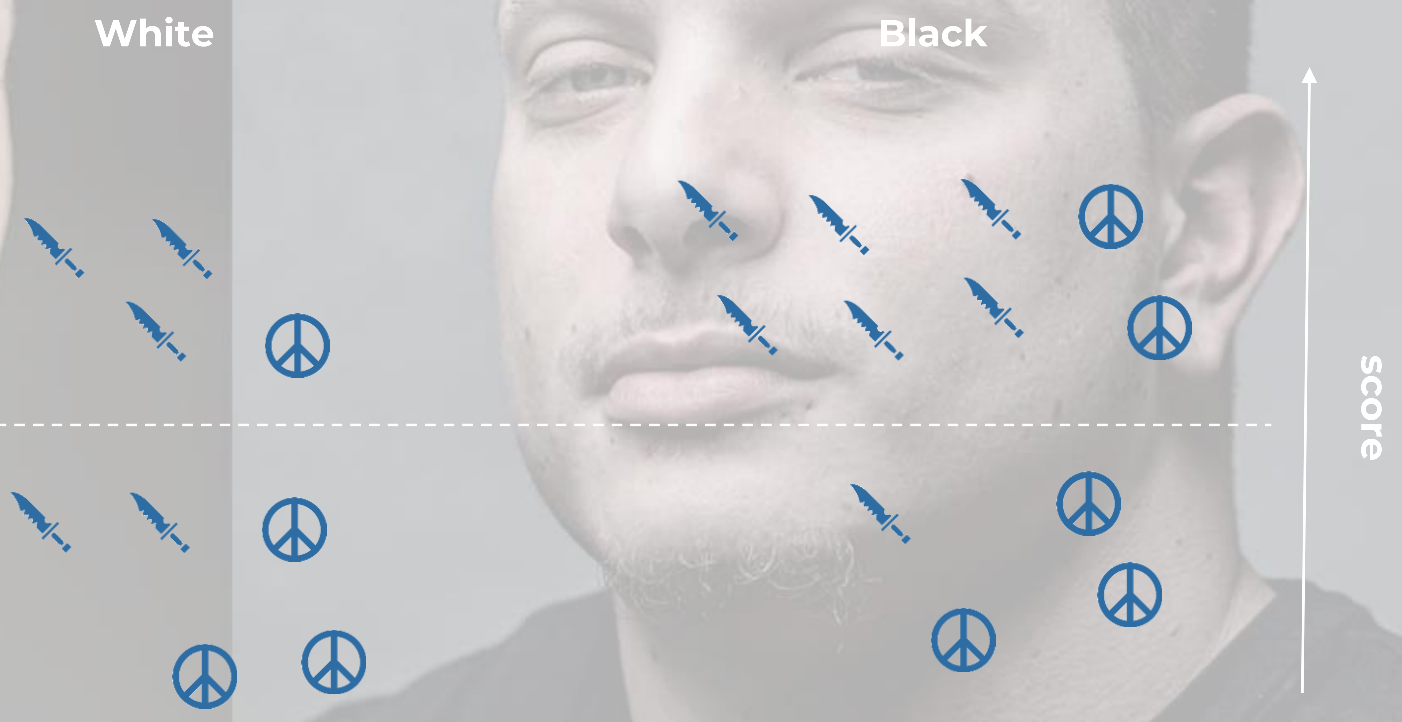
ultimately, we need to agree on what are the variables that we can «fairly» employ in the process

error rate parities are not all the same:
the Compas Debate

the
devil
is in
the
details

error rate parities are not all the same: the Compas Debate

the devil is in the details

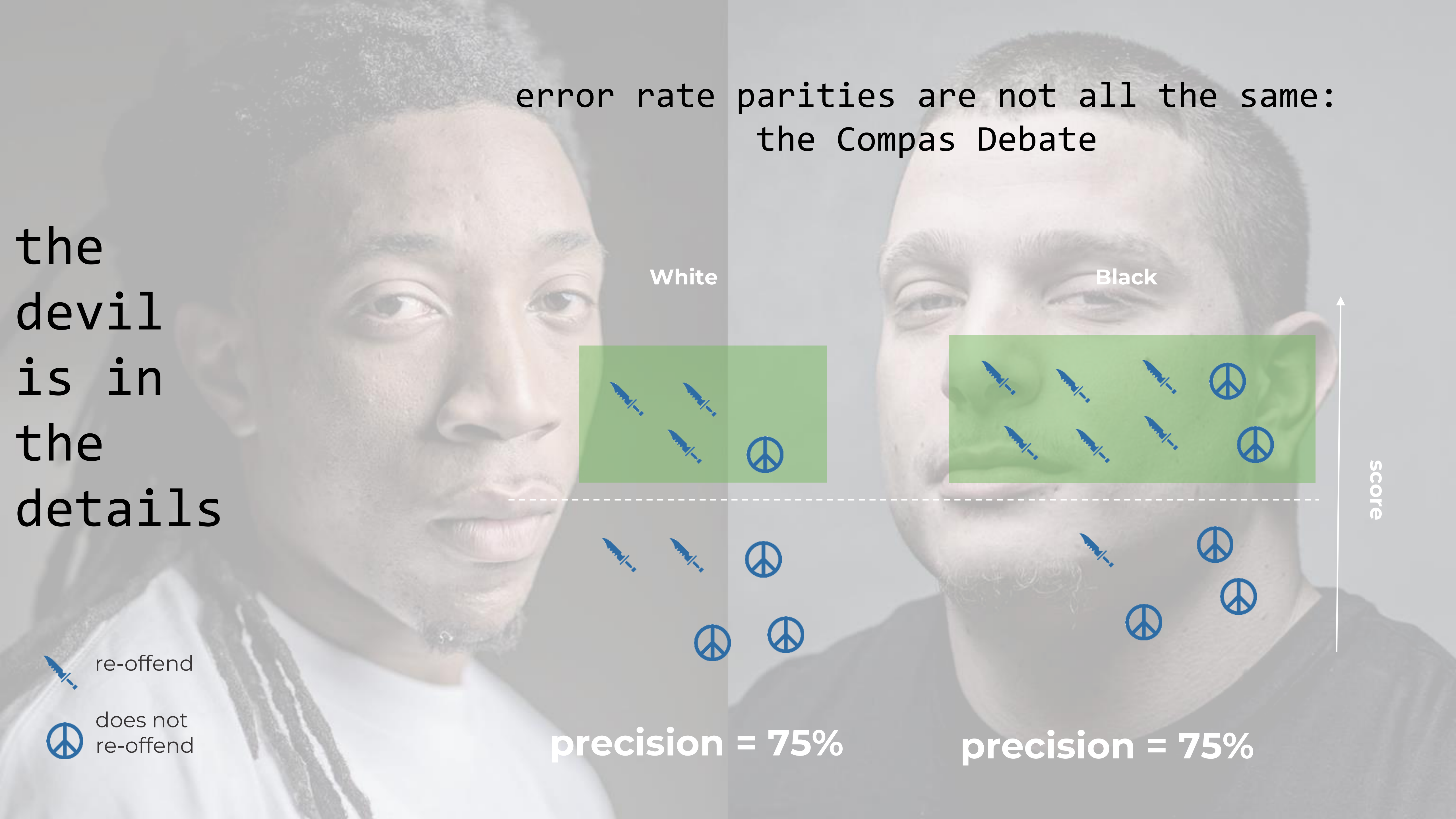


re-offend

does not re-offend

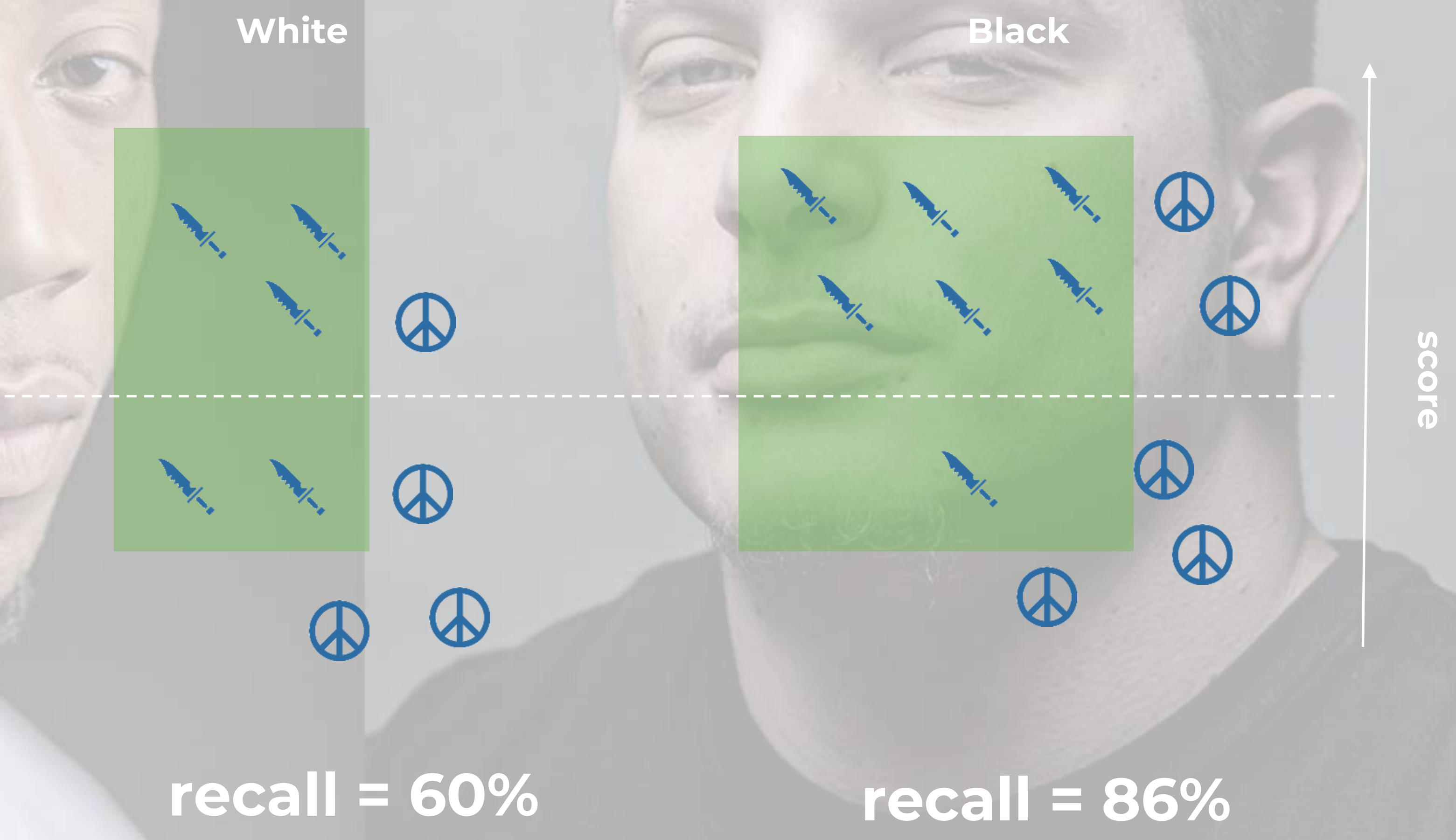
error rate parities are not all the same:
the Compas Debate

the devil is in the details



error rate parities are not all the same: the Compas Debate

the
devil
is in
the
details

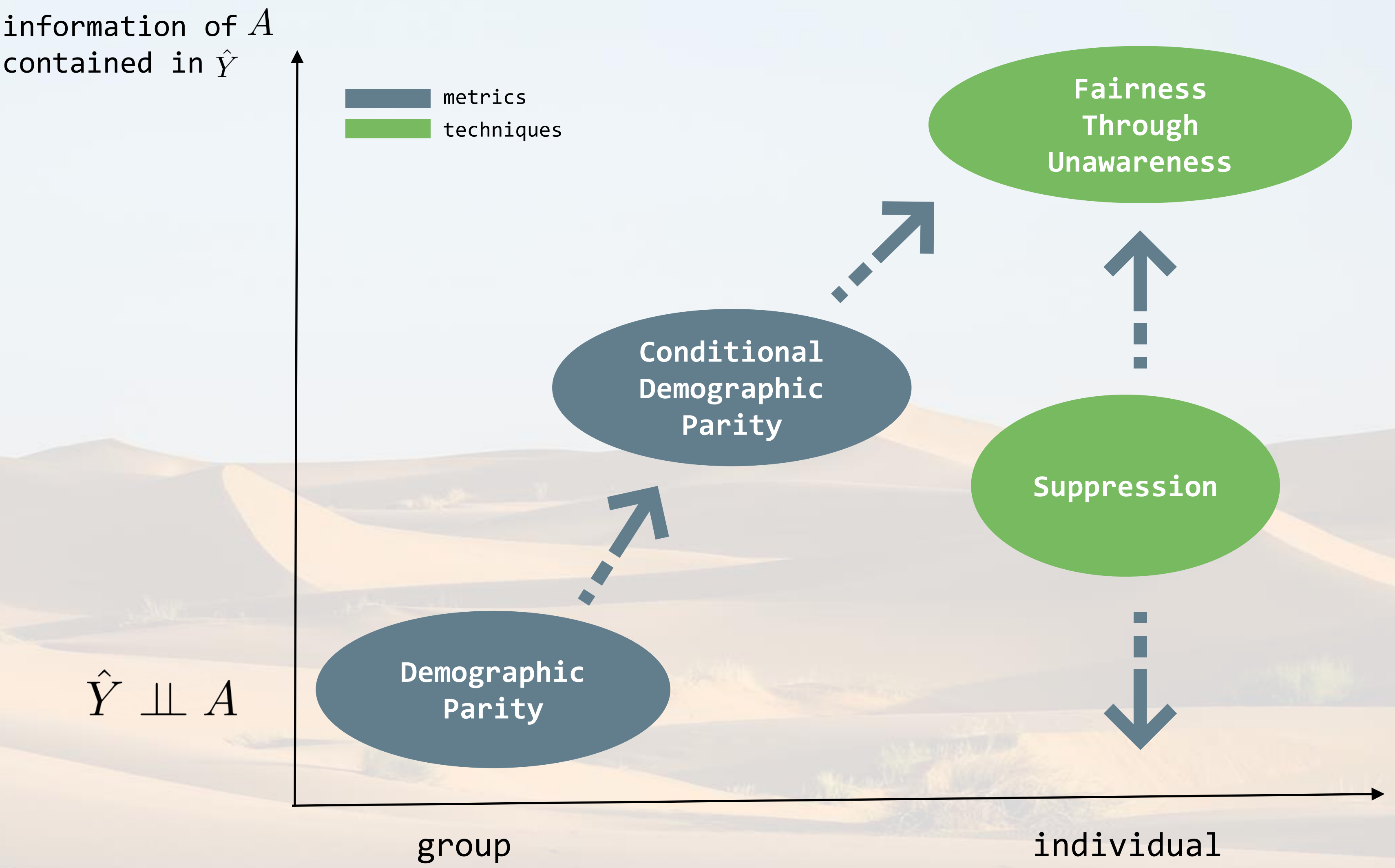


re-offend

does not re-offend

fairness metrics landscape

fairness metrics landscape



«A clarification of the nuances in the fairness metrics landscape», Castelnovo, Crupi, Greco, Regoli, Penco, Cosentini *Scientific Reports 2022*

Impossibility Theorem

~recall parity

~precision parity

Proposition 4. *Assume that all events in the joint distribution of (A, R, Y) have positive probability, and assume $A \not\perp Y$. Then, separation and sufficiency cannot both hold.*

Proof. A standard fact[27] about conditional independence shows

$$A \perp R \mid Y \quad \text{and} \quad A \perp Y \mid R \quad \Longrightarrow \quad A \perp (R, Y)\,.$$

Moreover,

$$A \perp (R, Y) \quad \Longrightarrow \quad A \perp R \quad \text{and} \quad A \perp Y\,.$$

Taking the contrapositive completes the proof. □

[27] See Theorem 17.2 in L. Wasserman, *All of Statistics: A Concise Course in Statistical Inference* (Springer, 2010)

Fairness and Machine Learning, Solon Barocas and Moritz Hardt and Arvind Narayanan (2019) https://fairmlbook.org/

Impossibility Theorems

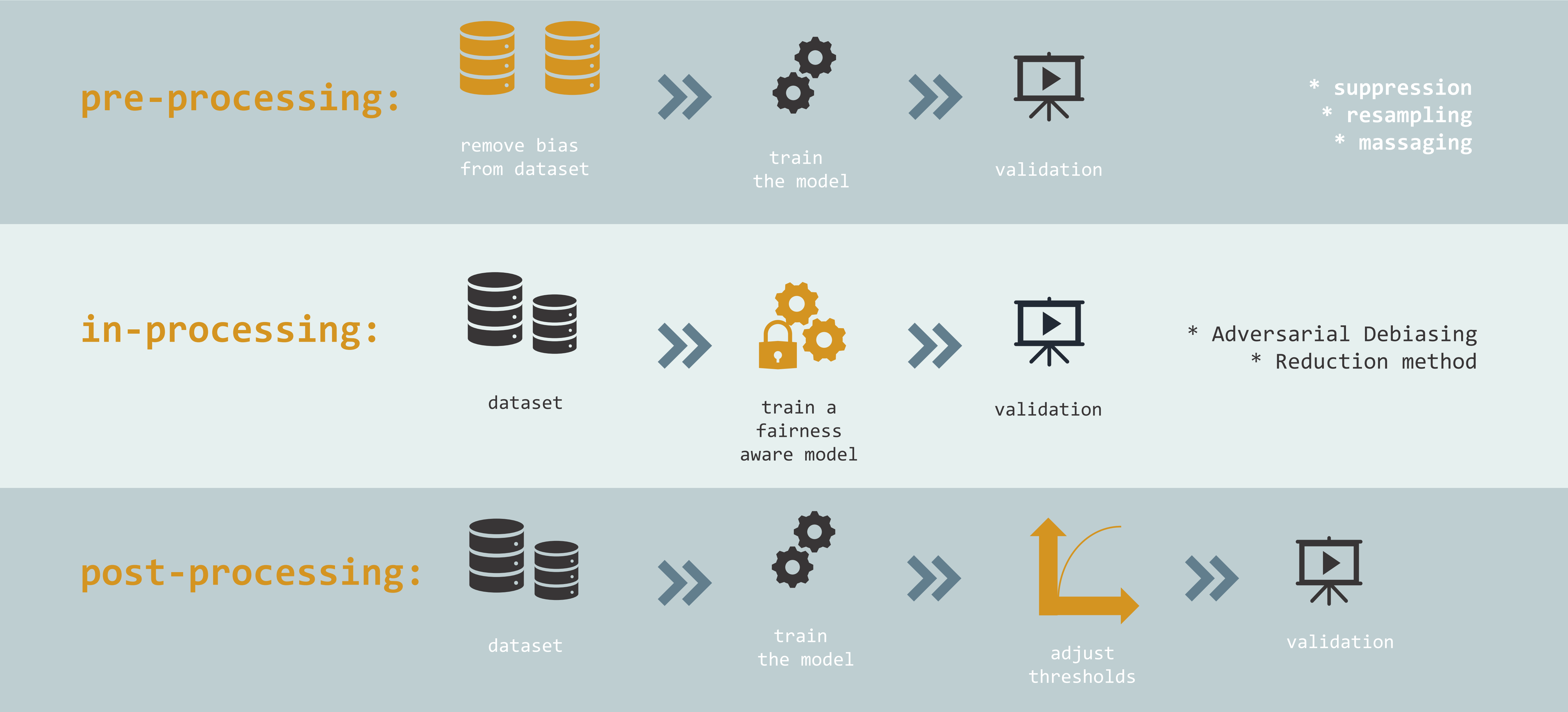
a part for degenerate cases, independence – separation – sufficiency are not mutually compatible

Fairness and Machine Learning, Solon Barocas and Moritz Hardt and Arvind Narayanan (2019) https://fairmlbook.org/

LIMITS OF SEPARATION

- you need to put a lot of trust in the target Y
- in some cases, you don't even have access to the full distribution of Y (you don't know if people you didn't give loan to would repay it back!)
- we are somehow letting the model learn bias from data (as long as Y justifies it), but that's what we wanted to avoid in the first place

Mitigation strategies


pre-processing:
remove bias
from dataset
train
the model
validation
* suppression
* resampling
* massaging
in-processing:
dataset
train a
fairness
aware model
validation
* Adversarial Debiasing
* Reduction method
post-processing:
dataset
train
the model
adjust
thresholds
validation

"Learning fair representations", Zemel, Rich, et al. International conference on machine learning. PMLR, 2013

pre-processing

Suppression

Remove sensitive variable(s) and features highly correlated with them

Fair Representation

Learn a representation of data such that sensitive information is removed while keeping as much information as possible from X

Sampling

Resample observations in order to reach Demographic Parity

Massaging

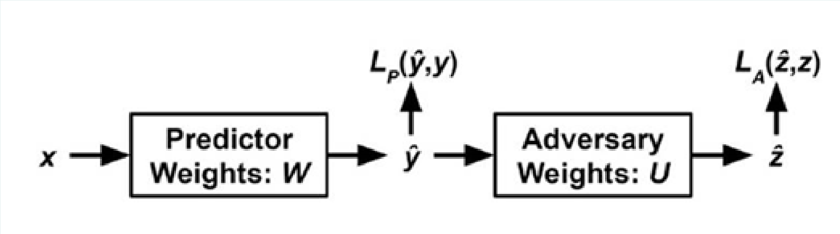
Re-label enough "minority" observations so that a new "massaged" training dataset is reached which satisfies demographic parity to begin with. The choice of which observations to be re-labeled is done via training an auxiliary model on the original target

“Data preprocessing techniques for classification without discrimination”, F. Kamiran and T. Calders, Knowledge and Information Systems, 2012

“Mitigating unwanted biases with adversarial learning”, AI B. H. Zhang, B. Lemoine, and M. Mitchell, Proceedings of the 2018 AAAI/ACM Conference on Ethics, and Society, 2018

in-processing

Idea: a model tries to maximize performance while an **Adversary** tries to reconstruct the protected variable *Z* from the model’s outputs



here *Z* is the protected variable

Goal: train using the following gradient steps

U step $\nabla_U L_A$

W step $\underbrace{\nabla_W L_P}_{\text{Estimates Y}} - \underbrace{\mathrm{proj}_{\nabla_W L_A} \nabla_W L_P}_{\text{reconstructs } Z \text{ and removes its effect}} - \underbrace{\alpha \nabla_W L_A}_{\text{pushes against the Adversary}}$

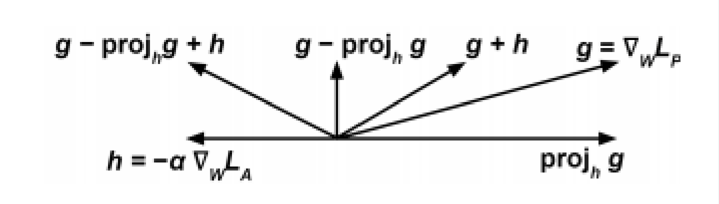


Diagram illustrating the gradients

Without the projection term, in the pictured scenario, the predictor would move in the direction labelled *g+h* in the diagram, which actually helps the adversary. With the projection term, the predictor will never move in a direction that helps the adversary.

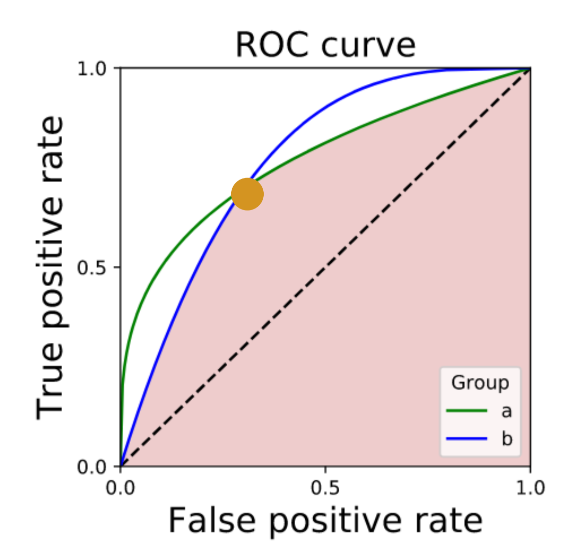
post-processing

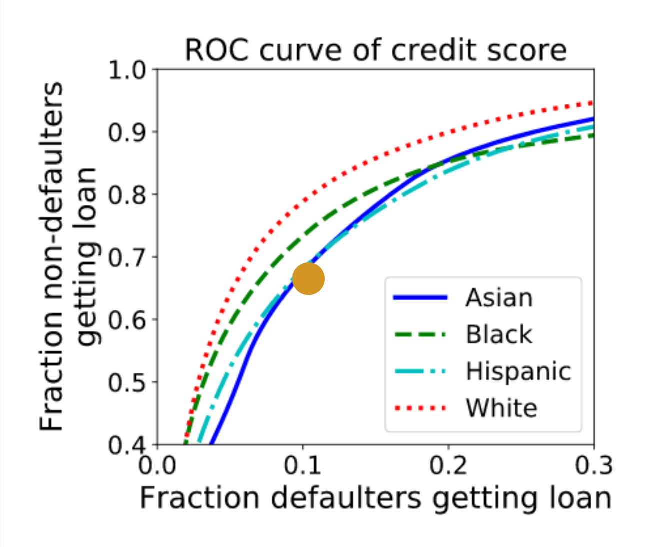
General idea

Tweak the threshold for different groups with respect to the sensitive variable

Equality of Odds

Need to look at the ROC



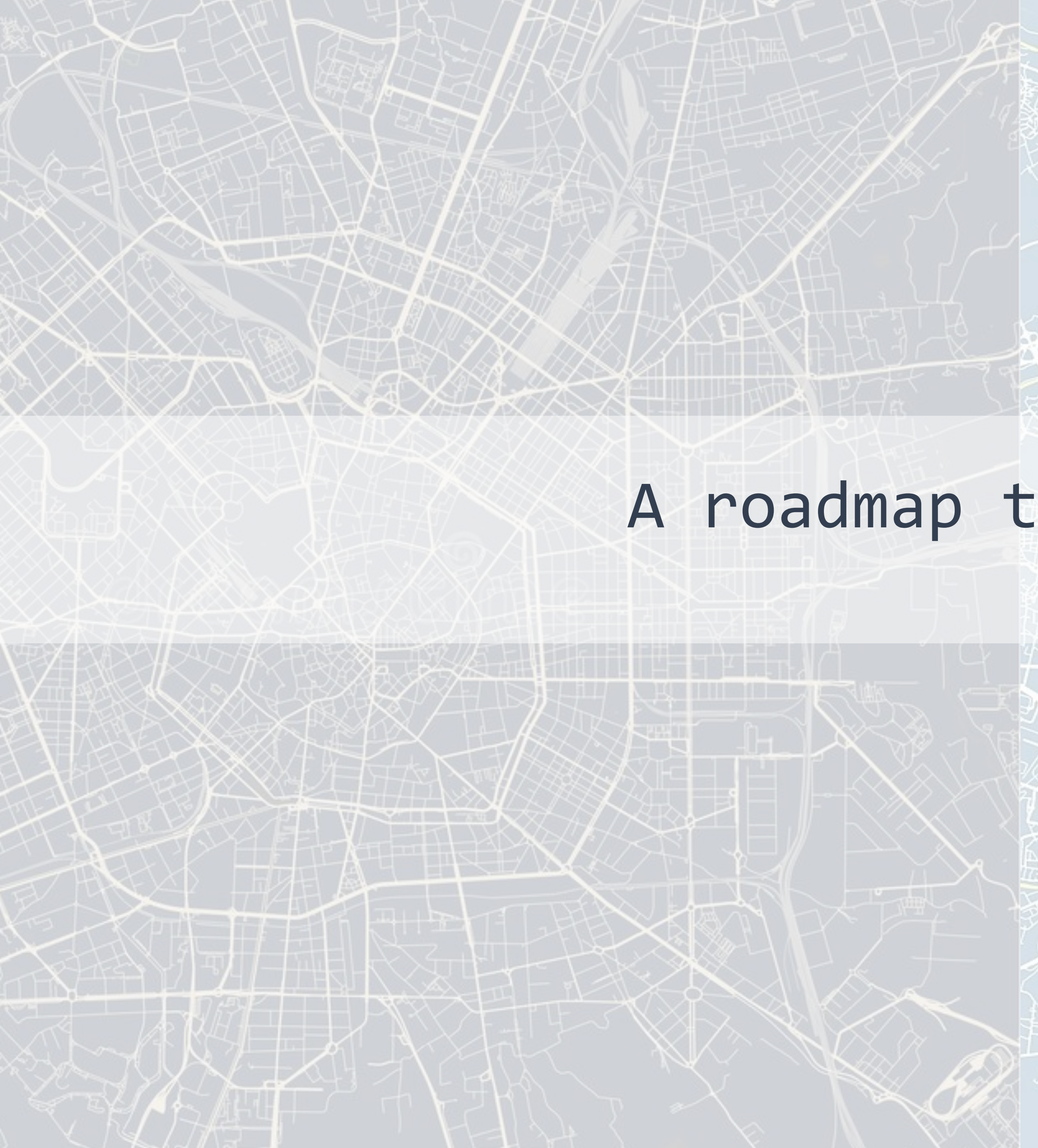
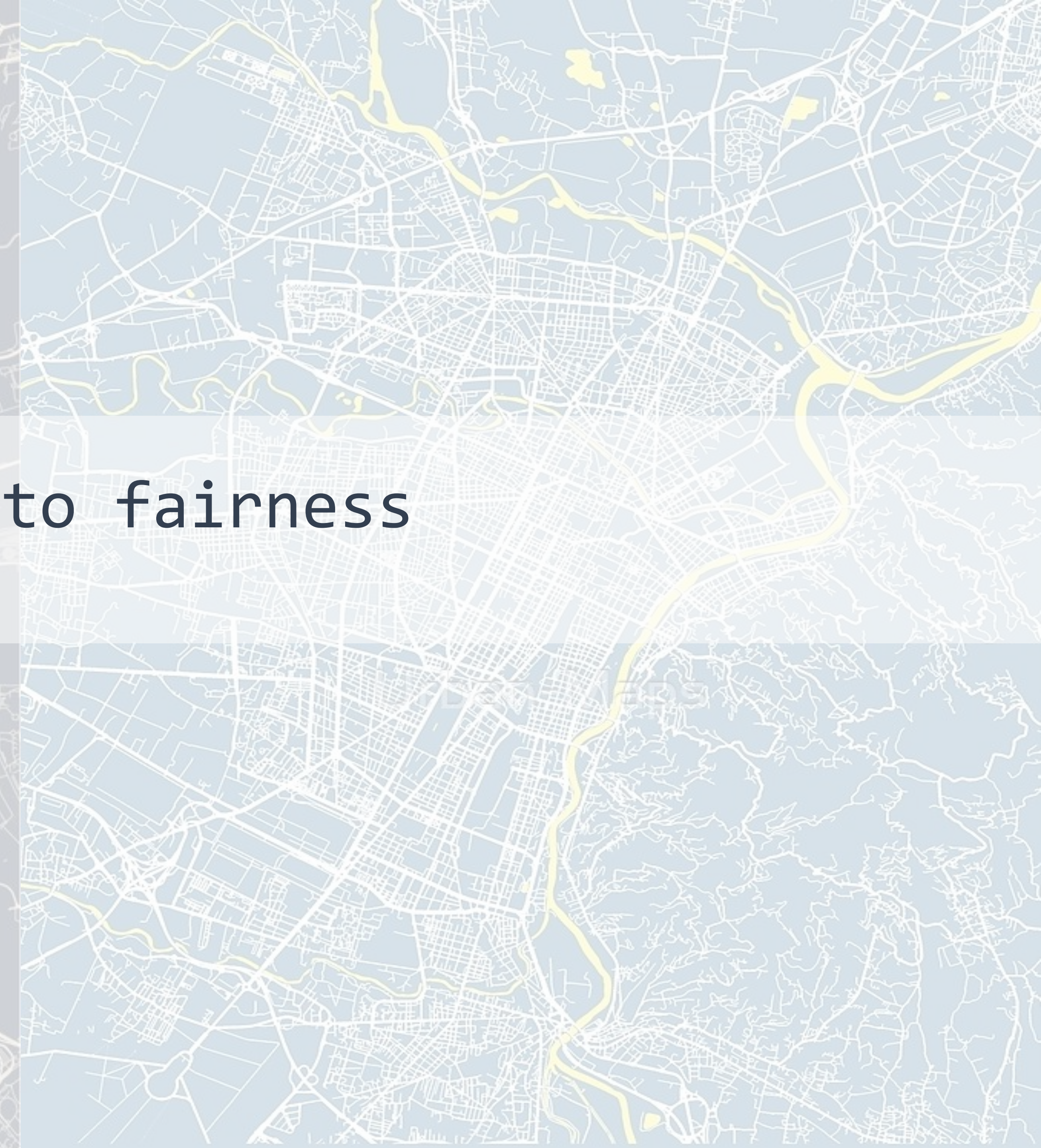


Plots from Barocas, Hardt, Narayanan https://fairmlbook.org

Demographic Parity

straightforward

“Equality of opportunity in supervised learning”, Hardt, Price, Srebro, NeurIPS 2016

A roadmap to fairness

Project description and objectives

Joint collaboration to research on Trustworthy AI in Financial domain.

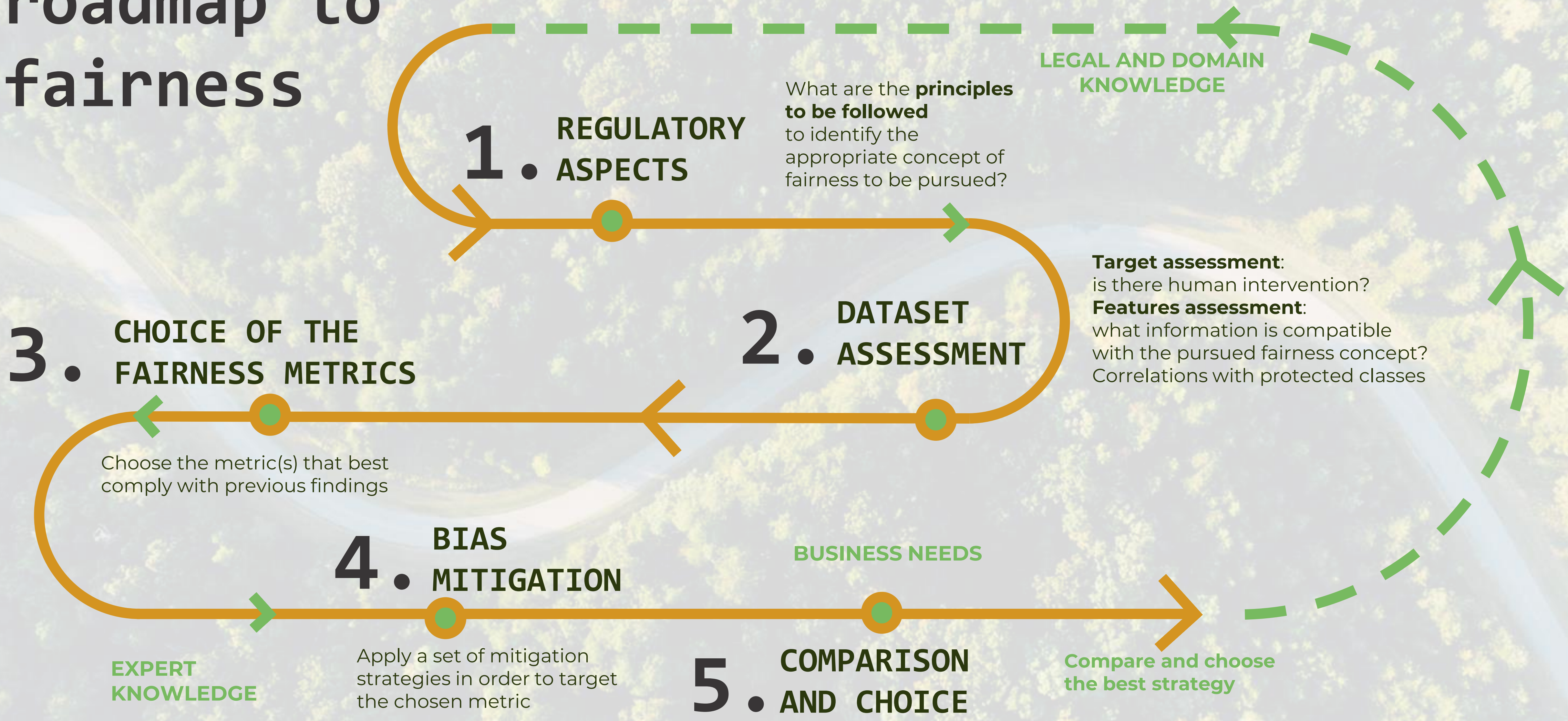
The **goal** is to overview the available metrics and techniques and to come up with a **roadmap to follow in order to pursue Fairness.**

To reach this goal, we carried out explorations on a **real-world financial use-case of credit lending.**

Collect tools of assessment, mitigation, visualization into a fairness toolbox called **BeFair.**

«BeFair: addressing Fairness in the Banking sector» Castelnovo, Crupi, Greco, Del Gamba, Naseer, Regoli, San Miguel Gonzalez, (2020 IEEE Big Data Conference)

roadmap to fairness



BeFair: addressing Fairness in the Banking sector. Castelnovo, Crupi, Greco, Del Gamba, Naseer, Regoli, San Miguel Gonzalez, 2020 IEEE Big Data Conference

Credit Lending use case

Dataset assessment

~200,000 loan applications
~50 predictors, including financial variables and personal information.
The target is the final decision of a human officer.

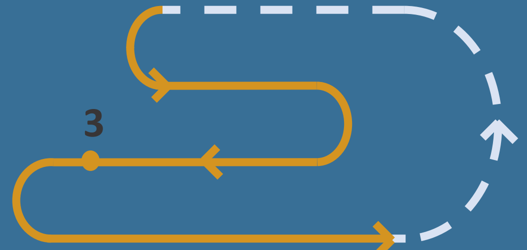
Throughout the analysis, we focus on

CITIZENSHIP = {0, 1}

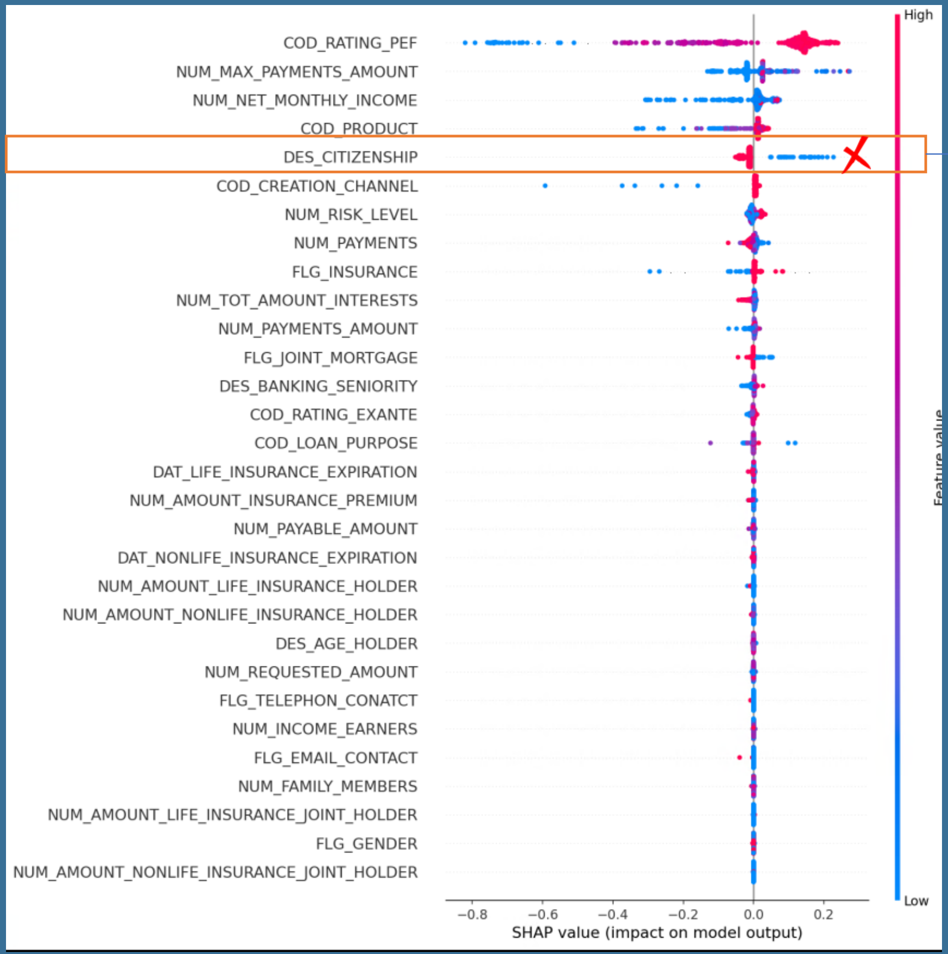
as **sensitive attribute** with respect to which assess fairness.

Bias, measured in terms of Demographic Parity, is negligible in the original target, but amplified by a the application of a ML model.

«BeFair: addressing Fairness in the Banking sector» Castelnovo, Crupi, Greco, Del Gamba, Naseer, Regoli, San Miguel Gonzalez, (2020 IEEE Big Data Conference)

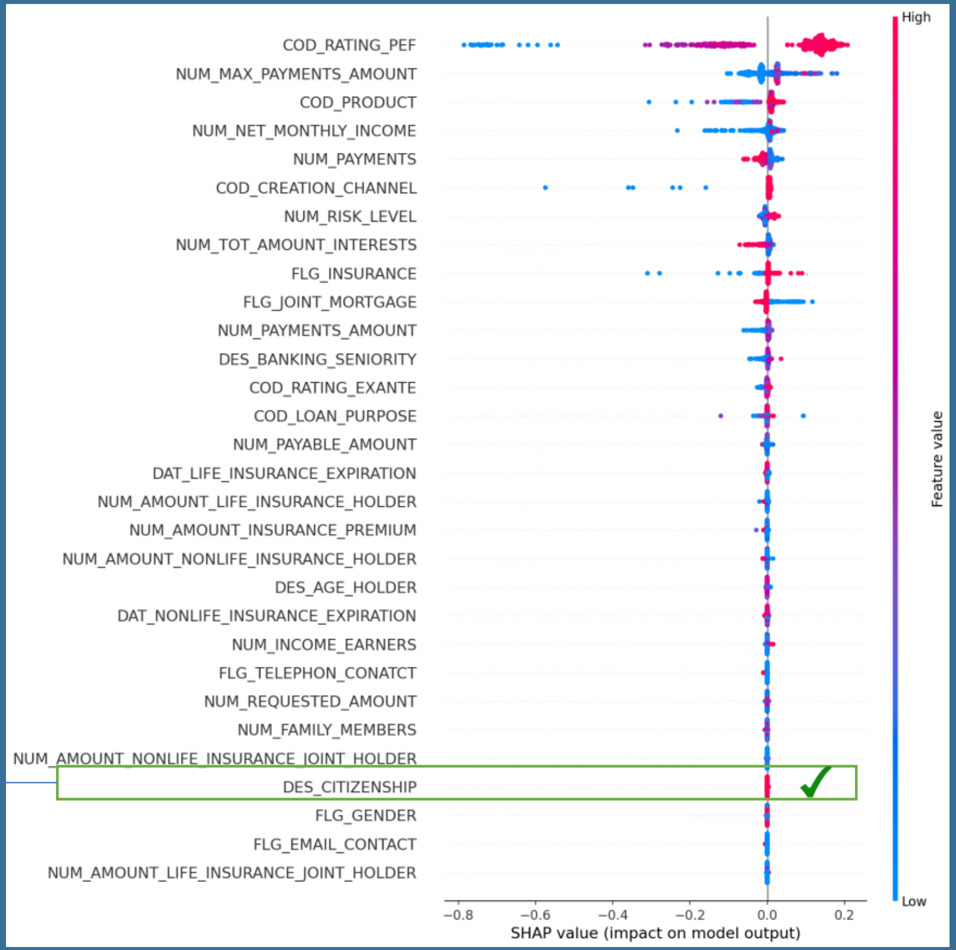


Mitigated Model – Group Level



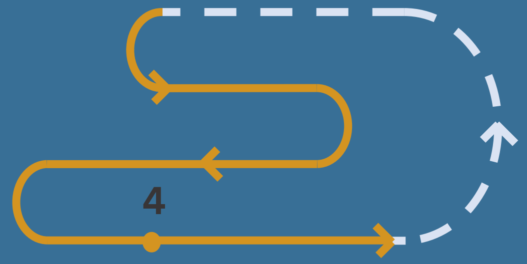
% loans allowed to citizens: 75,5%
% loans allowed to non-citizens: 75,1%

Mitigated Model – Individual Level



% loans allowed to citizens: 77,5%
% loans allowed to non-citizens: 51,7%

BeFair: developed methods and fairness goal



		Demographic Parity	Error Rate Parity	Individual Fairness
pre	FTU			✓
	Suppression	✓		
	Massaging	✓		
	Sampling	✓		
	CFF			✓
in	AdvDP	✓		
	AdvEO		✓	
	AdvCDP	✓		✓
	ReductionsGS	✓		
	ReductionsEG	✓		
post	ThreshDP	✓		
	ThreshEO		✓	
	ThreshEopp		✓	
	ThreshCDP	✓		✓

Bias mitigation

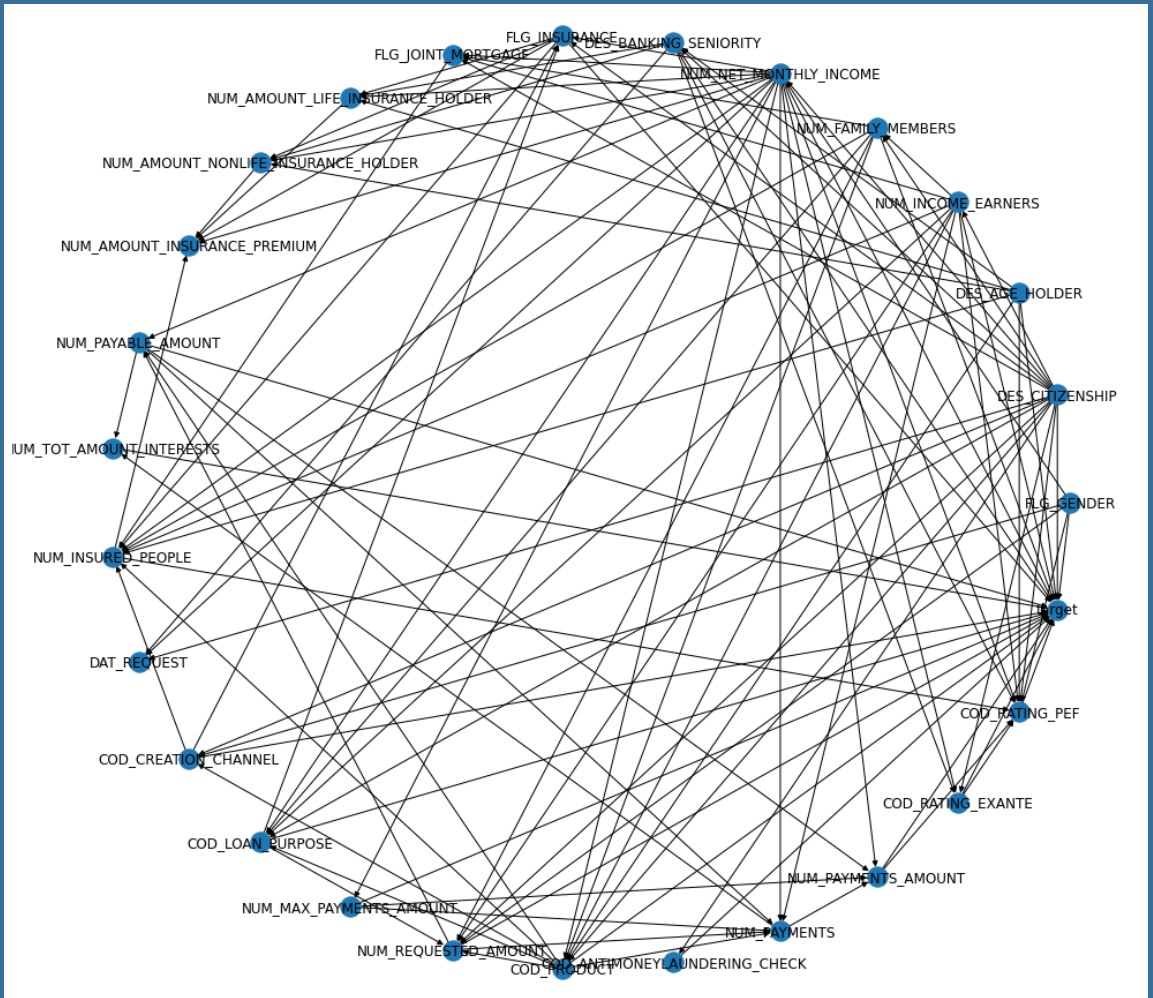
4

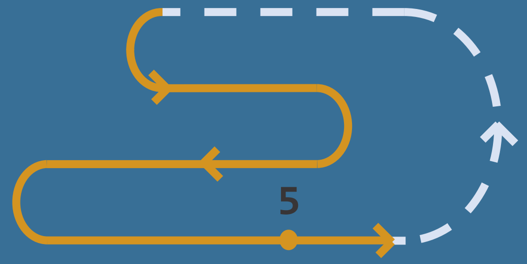
Counterfactual Fairness

Nodes are variables, while directed edges express the causal relationships among them.

Build **causal graph** with **causal discovery algorithms** and validate with **domain experts**.

Employ the causal graph to train a **counterfactually fair** model (Kusner et al. 2017): no causal flow from sensitive attribute to final decision.



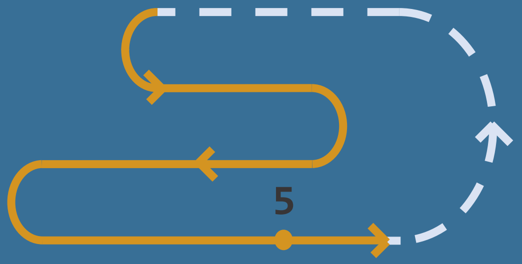


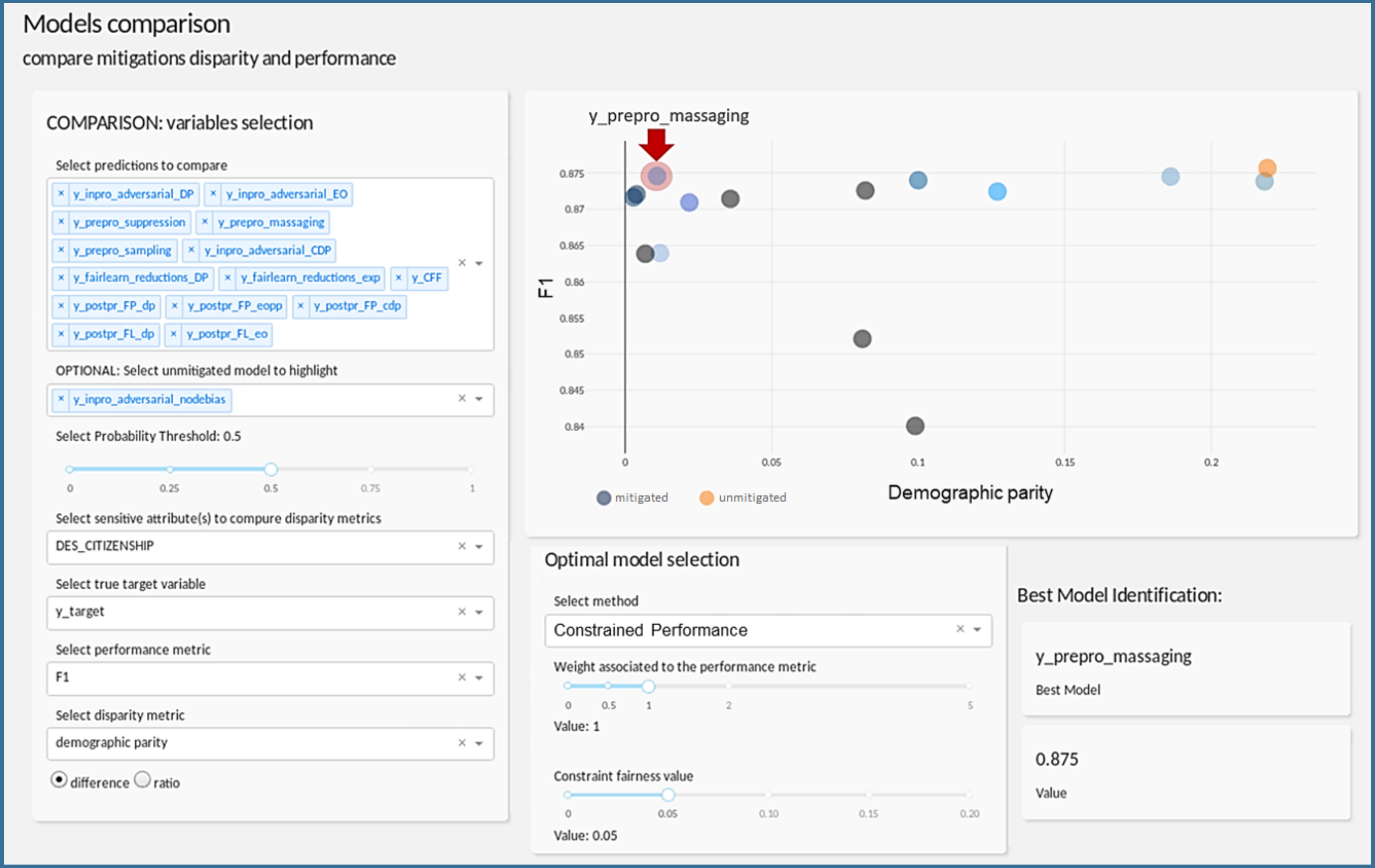
BeFair: bias mitigation results

family	type	fairness				performance		
		DP	EO	EOpp	PP	AUROC	Accuracy	F1
no mitigation	Logistic	0.324	0.272	0.272	**0.032**	0.817	0.761	0.823
	Random forest	0.221	0.202	-0.104	0.068	**0.838**	0.804	0.875
	Neural network	0.219	0.198	0.104	0.072	0.830	0.811	**0.876**
pre-process	FTU	0.164	0.124	0.058	0.095	**0.838**	0.812	**0.876**
	Suppression	0.099	-0.053	0.065	0.152	0.753	0.748	0.840
	Massaging	-0.004	0.062	0.062	0.163	0.818	**0.868**	0.803
	Sampling	0.080	0.012	0.012	0.115	0.835	0.791	0.851
	CFF	0.218	0.192	0.104	0.070	0.832	0.810	0.874
in-process	AdvDP	-0.034	0.073	0.063	0.176	0.823	0.802	0.869
	AdvEO	0.102	0.029	-0.010	0.148	0.819	0.805	0.871
	AdvCDP	0.147	0.101	-0.050	0.112	0.830	0.807	0.872
	ReductionsGS	0.012	0.077	0.049	0.159	0.812	0.794	0.864
	ReductionsEG	0.007	0.084	0.051	0.161	–	0.794	0.864
post-process	ThreshDP	**0.003**	0.099	0.056	0.164	–	0.805	0.872
	ThreshEO	0.082	**0.006**	0.006	0.138	–	0.812	0.873
	ThreshEOpp	0.100	0.048	**0.005**	0.119	–	0.809	0.874
	ThreshCDP	0.186	0.159	0.072	0.083	–	0.810	0.875

«BeFair: addressing Fairness in the Banking sector» Castelnovo, Crupi, Greco, Del Gamba, Naseer, Regoli, San Miguel Gonzalez, (2020 IEEE Big Data Conference)

Comparison and choice





Proposed methods to identify the best perfomance-fairness tradeoff:

Trade-off fairness-performance

$$(1+\beta^2)\frac{(1-|\phi|)*\pi}{\beta^2*(1-|\phi|)+\pi}$$

Constrained performance

$$\max_{\phi\leq\Phi}\pi$$

π and φ are the preferred performance and fairness metrics, respectively and beta is the weight associated with the performance metric.

other limits of current methodologies…

perimeter of application

most metrics are for **classifications**

most mitigation strategies target DP (sometimes Eodds) for **classification** only

sensitive attributes

what are the relevant sensitive attributes?

aggregation problems (e.g. age)

what about **intersectional bias**?

other types of bias

There are types of bias hardly captured by this framework, e.g. **bias in language models** (gender-profession correlations, etc.)

Ruggieri, Alvarez, Pugnana, State, Turini, Can We Trust Fair-AI? AAAI (2023)

summarizing…

Bias discrimination is a concrete risk for AI applications at scale.

Fairness concepts are manifold, and care should be taken in any specific situation.

…a crucial point is that Fairness in Machine Learning cannot be left to Data Scientists only.

More research is needed on the ethical and legal side to clarify the needs of specific domains.

More research is needed on the technical side, e.g. to understand the relationship among different fairness metrics, to find appropriate metrics for various tasks (besides classifications) and to find mitigation strategies enforcing a wider range of metrics.

SOME REFERENCES

surveys

Barocas, Hardt, Narayanan, *Fairness and machine learning*, (2019)

Barocas, Selbst, Big data's disparate impact, Calif. L. Rev. (2016)

Mehrabi et al. A survey on bias and fairness in machine learning, ACM Computing Surveys (2021)

Castelnovo, Crupi, Greco, Regoli, Penco, Cosentini, A clarification of the nuances in the fairness metrics landscape, Scientific Reports (2022)

Baumann, Castelnovo, Crupi, Inverardi, Regoli, Bias on Demand: A Modelling Framework that Generates Synthetic Data with Bias, FAcct (2023)

Ruggieri, Alvarez, Pugnana, State, Turini, Can We Trust Fair-AI? AAAI (2023)

mitigation strategies

Zhang, Hu, Mitchell, Mitigating unwanted biases with adversarial learning, Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (2018)

Kamiran, Calders, Data preprocessing techniques for classification without discrimination, Knowledge and Information Systems (2012)

Hardt, Price, Srebro, Equality of opportunity in supervised learning, Advances in neural information processing systems (2016)

Zemel, Rich, et al. Learning fair representations, International conference on machine learning. PMLR, 2013

thank you

daniele regoli

Data Science & AI @ Intesa Sanpaolo

daniele.regoli@intesasanpaolo.com