ellis ALICANTE unit

Fairness in Algorithmic Decision Making

A general introduction about Fairness in Algorithmic ML

Adrián Arnaiz-Rodríguez

1y PhD Student

ELLIS Alicante

Talk in Seminarios del Doctorado en Tecnologías Industriales e Ingeniería Civil - UBU

3rd March 2022

European Union
European Regional Development Fund

GOBIERNO DE ESPAÑA
MINISTERIO DE CIENCIA E INNOVACIÓN

GENERALITAT VALENCIANA
Conselleria d'Innovació, Universitats, Ciència i Societat Digital

Sabadell Fundación

- Introduction to Algorithmic Fairness
- Fairness definitions
- Imposing Fairness
- Current prominent approaches
- General conclusions
- Resources

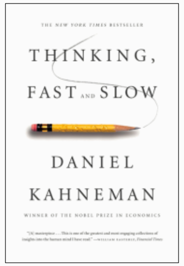
Introduction to algorithmic fairness

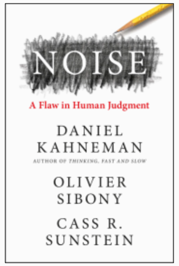
From biased decisions to algorithmic fairness

Human are imperfect decision-makers

- Value system
- Past experiences
- Culture
- Unconscious emotion
- Randomness
- Noise
- Biases
- Greed

Decision





Biases

- *Confirmation bias*
- *Decoy effect*
- *Framing effect*
- *Omission bias*
- *Survivorship bias*
- …

ML for critical Decision Making

- ML models are becoming the main tools for addressing complex societal problems
  → *Algorithms don't have human behaviors and not crooked*
  - Education
  - Justice: pretrial and detention
  - Security: Recidivism
  - Health
  - Child Maltreatment screening
  - Social Services
  - Hiring
  - Finance
  - Advertising

✓ Privacy	✓ Reliability
✓ Transparency	✓ Autonomy
✓ Accountability	✓ **Fairness**

Ethical implications
Universally accepted definitions?

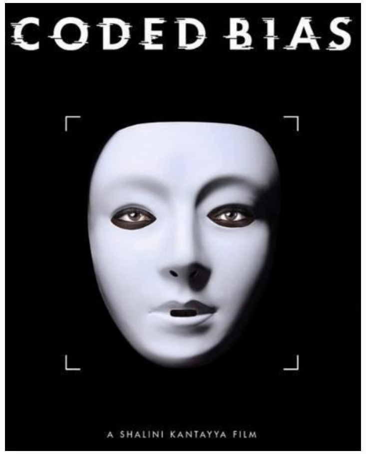
- Each one with its own objectives
  - Reduce cost
  - Maximize social benefit
  - …

Are models itself unbiased Decision-Makers?

Can the criminal justice system's artificial intelligence ever be truly fair?

Computer programs used in 46 states incorrectly label Black defendants as "high-risk" at twice the rate as white defendants

Natalia Mesa
Neuroscience
University of Washington



Forbes

Deliveroo Rating Algorithm Was Unfair To Riders, Italian Court Rules

Jonathan Keane Contributor
Consumer Tech
Freelance technology journalist covering the gig economy.
Follow

Personal and protected reasons → Shift Cancelation/Acceptation → Reliability index → Offered Shifts

SCIENTIFIC AMERICAN

Racial Bias Found in a Major Health Care Risk Algorithm

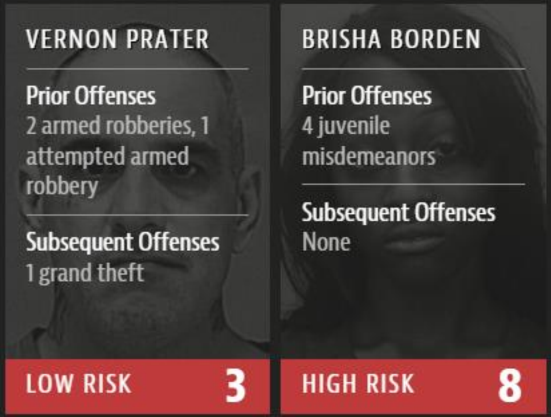
Black patients lose out on critical care when systems equate health needs with costs

The Guardian
For 200 years

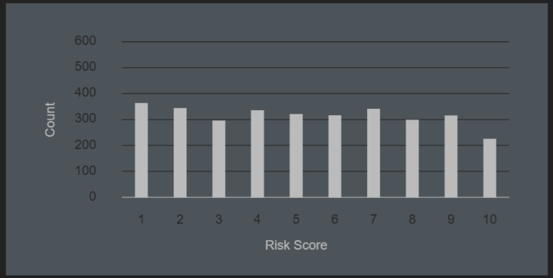
Amazon ditched AI recruiting tool that favored men for technical jobs

Specialists had been building computer programs since 2014 to review résumés in an effort to automate the search process

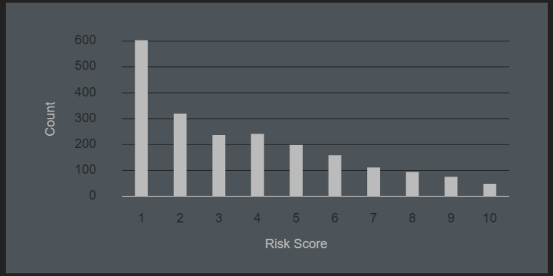
Two Petty Theft Arrests



Black Defendants' Risk Scores



White Defendants' Risk Scores



Two Drug Possession Arrests

Prediction Fails Differently for Black Defendants

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

Machine Bias

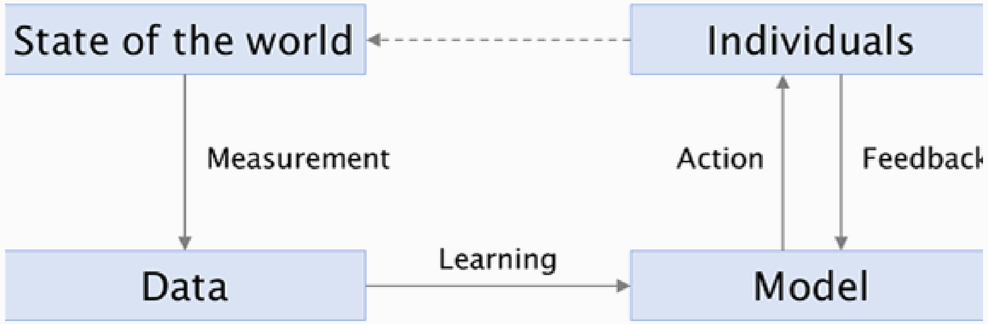
There's software used across the country to predict future criminals. And it's biased against blacks.

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

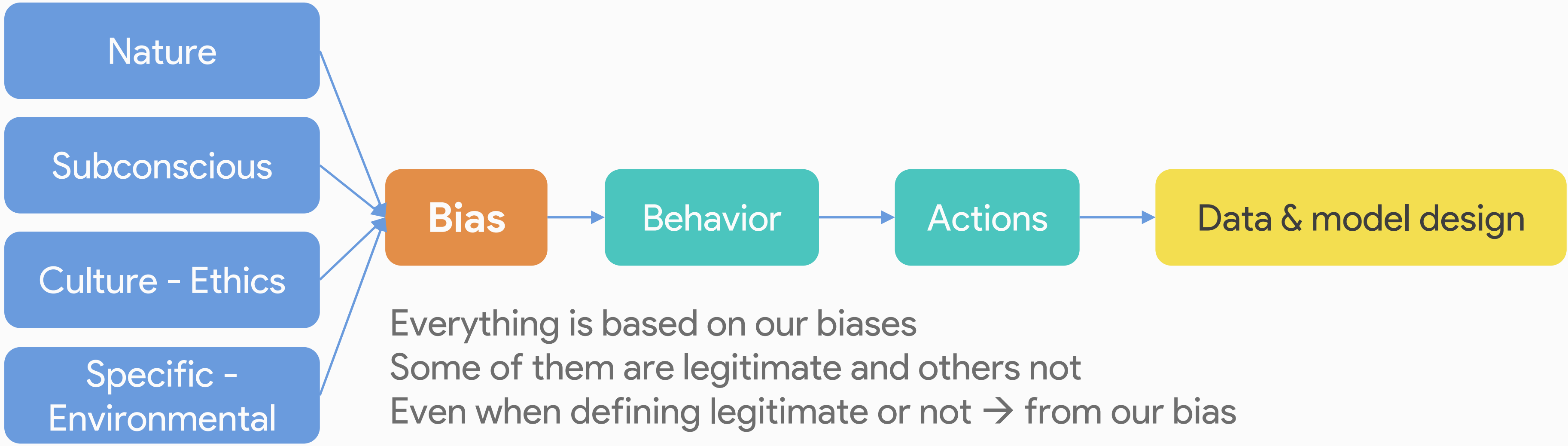
May 23, 2016

Correctional Offender Management Profiling for Alternative Sanctions - COMPAS

Why algorithms are biased?



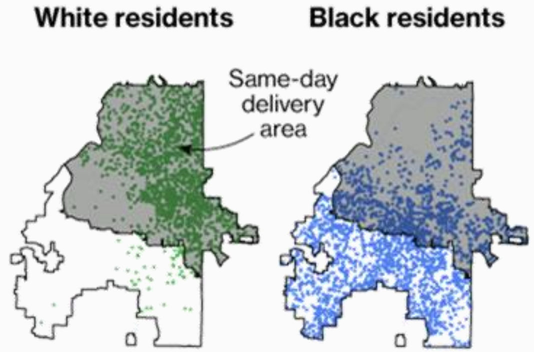
- Models learn from data → Bias in the loop
  - Skewed or imbalanced data features
  - Problems in labels: imbalanced, imperfect and selective

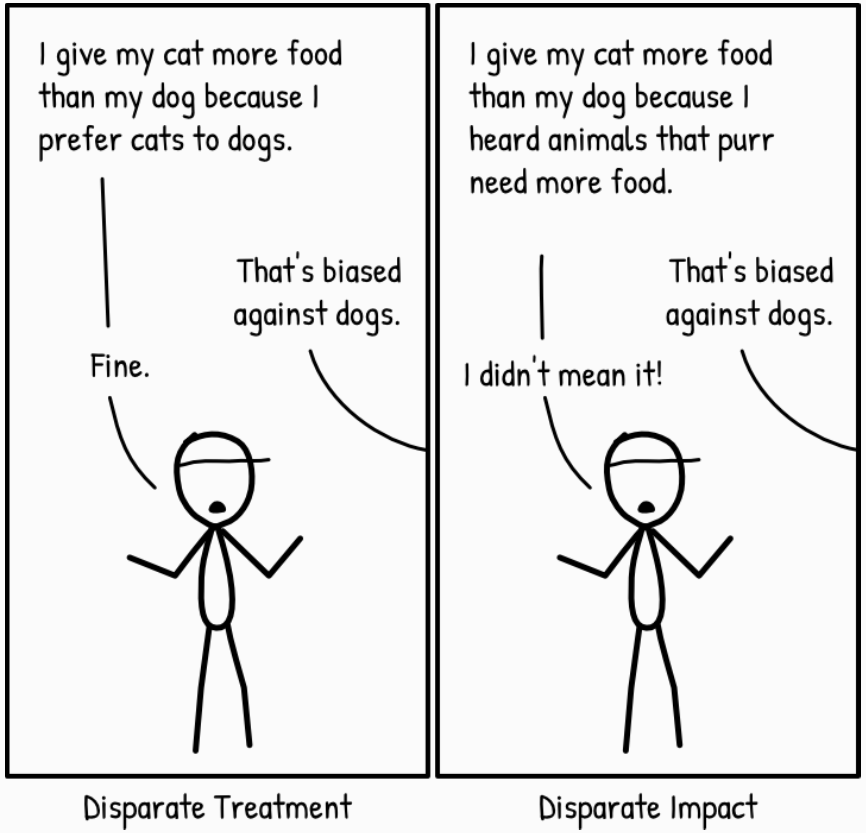


Everything is based on our biases
Some of them are legitimate and others not
Even when defining legitimate or not → from our bias

Disparate Treatment and Impact

- Anti-discrimination laws in various countries prohibit unfair treatment of individuals
- Legal or ethical support and formalize it quantitively
  - **Disparate treatment:**
    - Decisions are (partly) based on the subject's sensitive attribute
    - Explicit or intentional
  - **Disparate impact:**
    - Outcomes or implemented policy disproportionately hurt people with certain sensitive attribute
    - Implicit or unintentional





Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. Calif. L. Rev., 104, 671
Lim Swee Kiat. Retrieved December 2021. Machines go Wrong. https://machinesgonewrong.com/fairness/
Ingold, D. and Soper, S., 2016. Amazon doesn't consider the race of its customers. Should It?. Bloomberg News.

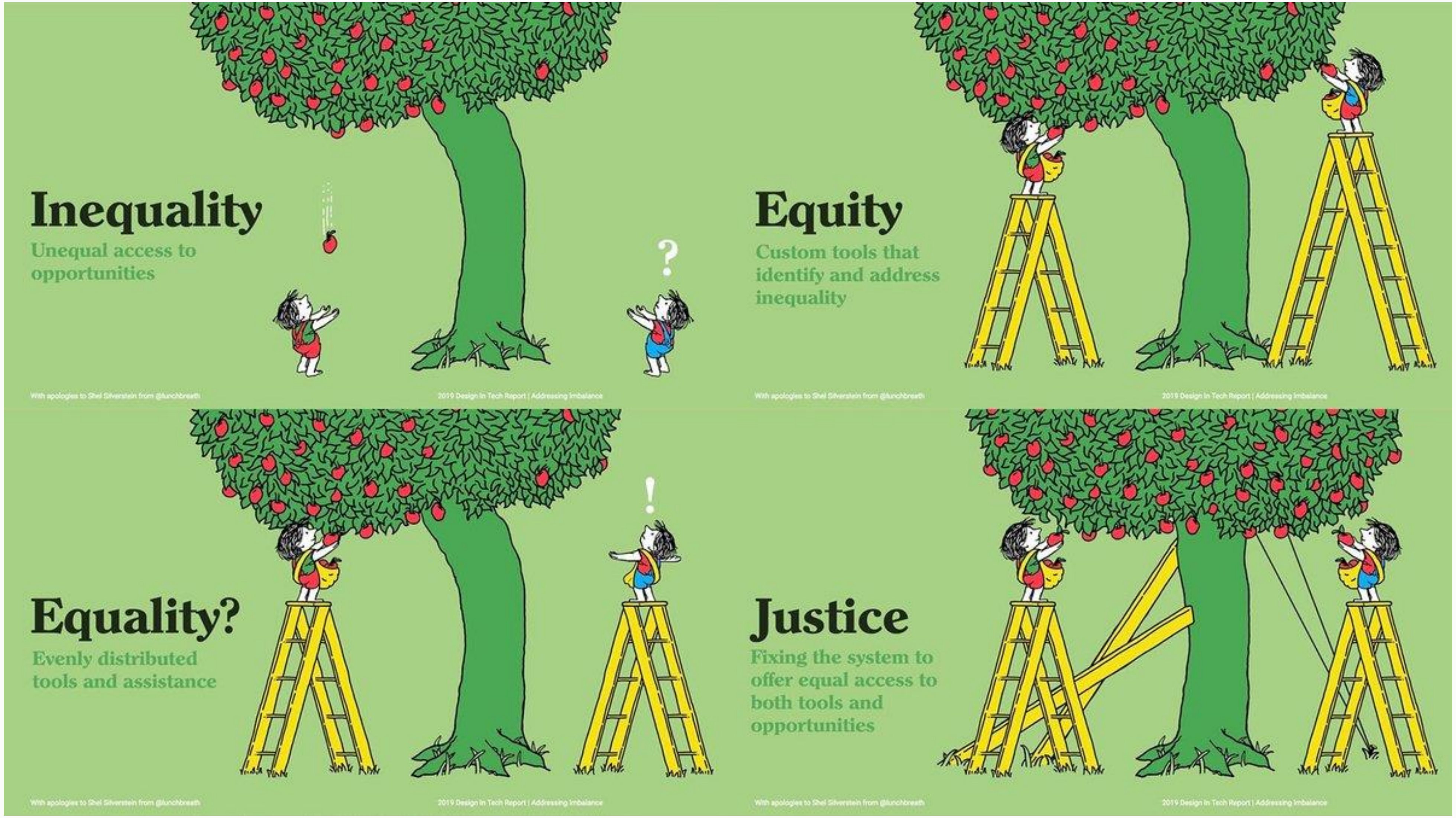
What are the effects of biased decision-making?

INDIVIDUAL HARMS		COLLECTIVE SOCIAL HARMS
ILLEGAL DISCRIMINATION	UNFAIR PRACTICES	
HIRING		LOSS OF OPPORTUNITY
EMPLOYMENT		
INSURANCE & SOCIAL BENEFITS		
HOUSING		
EDUCATION		
CREDIT		ECONOMIC LOSS
DIFFERENTIAL PRICES OF GOODS		
LOSS OF LIBERTY		SOCIAL STIGMATIZATION
INCREASED SURVEILLANCE		
STEREOTYPE REINFORCEMENT		
DIGNATORY HARMS		

Chart Contents Courtesy of Megan Smith, Former CTO of the United States

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In FAccT. PMLR. http://gendershades.org/overview.html

Justice, equality and equity



Human centric ML approaches

AI systems learning moral notions

*AI-based systems can **learn moral notions** or ethical behaviors and then **autonomously behave ethically***

- Comparative Moral Turing Test
- Ethical Turing Test

➢ Evaluate the morality of the choices of automated systems

➢ **Branch quite unexplored:** difficult connection between philosophy, ethic and technical problems

➢ AGI related

How humans should design AI systems to minimize harms

*Designing for **minimizing** harms derived from **poor design**, **bad applications** and **misuse** of the systems*

- **Algorithmic Fairness**
- Privacy Preserving Data Mining – Federated Learning
- Explainable AI [2] & Interpretable AI
- Adversarial Learning

➢ Many more examples due to many different ML methods and problems addressed

HCML Perspective: building responsible AI including human relevant requirements, but also considering broad societal issues [1]

- Safety, **Fairness**, privacy, accountability & interpretability - Ethics and legislation

Franco, D., Navarin, N., Donini, M., Anguita, D., & Oneto, L. (2022). **Deep fair models for complex data: Graphs labeling and explainable face recognition.** Neurocomputing, 470
1. A.F. Winfield, K. Michael, J. Pitt, V. Evers, **Machine ethics: the design and governance of ethical ai and autonomous systems**, Proceedings of the IEEE 107 (2019) 509–517
2. D. Gunning, **Explainable artificial intelligence (xai)**, Defense Advanced Research Projects Agency (DARPA), nd Web 2 (2).

What should we consider to formally defining fairness?

ML is used for critical decision making

Bias is in the humans & society, and it's transmitted to the algorithms

→

Challenges of ML

- **Uncover bias/unfairness**
- **Measure bias (definitions Fairness)**
- **Mitigate bias**
- **Real world applications**

How do we formulate the bias-fairness problem in every problem set up?

How do we detect the bias in our models and how to solve it?

How could we define and measure bias or fairness?

Which are the ethical principles that follows each definition of bias and fairness?

Which are the implications in the real-world problems and, specifically in our own value system?

What are the philosophical and ethical limitations of the current Fairness approach?

SPOILER: Everything depends on the CONTEXT

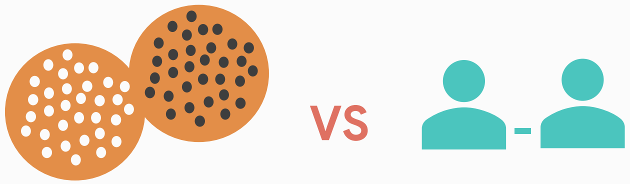
$P(S=s \mid A=a) = P(S=s \mid A=b)$

Fairness definitios and metrics

Several notions of fairness already exist in the literature

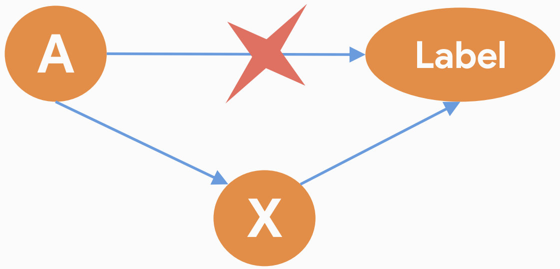
Algorithmic Fairness

- Algorithmic Fairness deals with the problem of developing AI-based systems able to treat:
  - **Subgroups in the population <u>equally</u>** → **Group fairness**
  - **Similar individuals in a <u>similar</u> way** → **Individual Fairness**
    - *Specifically, similar individuals from different subgroups*



How do we define equally? And similar?

- Subgroups → determined by means of **<u>sensitive attributes</u>**, considered for decisions
  - *Gender*, *incomes*, *ethnicity*, and *sexual* or *political orientation*…
- **Ensure that the outputs of a model DO NOT depend on sensitive attributes**
  - $F(X) = R,\ A \in X$ → $R \perp A$



$$Pr(\hat{Y} = y | Y = y)$$
$$Pr(Y = y | \hat{Y} = y)$$

Confusion matrix reminder

Event	Condition	Notion $P(event\|condition)$
$\hat{Y} = 0$	$Y = 0$	True Negative rate
$\hat{Y} = 1$	$Y = 0$	False Positive rate
$\hat{Y} = 0$	$Y = 1$	False Negative rate
$\hat{Y} = 1$	$Y = 1$	True Positive rate

Classical clf criteria

Event	Condition	Notion $P(event\|condition)$
$Y = 0$	$\hat{Y} = 0$	Positive predicted value
$Y = 1$	$\hat{Y} = 1$	Negative predicted value

Additional clf criteria

		Predicted Label		
		$\hat{y} = 1$	$\hat{y} = -1$	
True Label	$y = 1$	True positive	False negative	$P(\hat{y} \neq y \| y = 1)$ False Negative Rate
	$y = -1$	False positive	True negative	$P(\hat{y} \neq y \| y = -1)$ False Positive Rate
		$P(\hat{y} \neq y \| \hat{y} = 1)$ False Discovery Rate	$P(\hat{y} \neq y \| \hat{y} = -1)$ False Omission Rate	$P(\hat{y} \neq y)$ Overall Misclass. Rate

		Predicted Label	
		Positive	Negative
True Label	Positive	**True Positives** $PPV = \frac{TP}{TP + FP}$ $TPR = \frac{TP}{TP + FN}$	**False Negative** $FOR = \frac{FN}{FN + TN}$ $FNR = \frac{FN}{FN + TP}$
	Negative	**False Positive** $FDR = \frac{FP}{FP + TP}$ $FPR = \frac{FP}{FP + TN}$	**True Negatives** $NPV = \frac{TN}{TN + FN}$ $TNR = \frac{TN}{TN + FP}$

Confusion matrix allow us to go further accuracy in error explanations related with joint distributions of $(X, \hat{Y}, Y)$

Barocas, S., Hardt, M., & Narayanan, A. (2017). Fairness in machine learning. Nips tutorial, 1, 2017
Zafar, M. et al. (2017). Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. 26th WWW.
Verma, S., & Rubin, J. (2018). Fairness definitions explained. In 2018 ieee/acm fairware. IEEE.

Group fairness: Formal criteria

Different groups must have similar statistics overall in terms of predictions and errors

"Many fairness criteria have been proposed over the years, each aiming to formalize different desiderata. We'll start by jumping directly into the formal definitions of three representative fairness criteria that relate to many of the proposals that have been made." (Barocas, Hardt, Narayanan, Fairness in Machine Learning book, 2019)

$P(S\|A)$	$P(S\|Y,A)$	$P(Y\|S,A)$
Independence	*Separation*	*Sufficiency*
S⊥A	S⊥A | Y	A⊥Y | S

Demographic parity

P(d=1|A=a) = P(d=1|A=b)

Positive Predicted Ratio
Equal acceptance rate

A S

Equalized odds

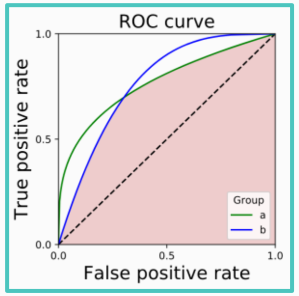
P(d=1 | Y=i , A=a) = P(d=1 | Y=i, A=b), i ∈ 0, 1

Equal opportunity

P(d=0 | Y=1, A=a) = P(d=0| Y=1, A=b)

TPR - FPR
Equal error rates

A — Y — S



Predictive Parity

P(Y=1 | d=1, A=a) = P(Y=1 | d=1 , A=b)

Calibration

P(Y=1 | S=s>t, A=a)= P(Y=1 | S=s>t, A=b)∀ t

PPV - NPV
Calibration by group

A — S — Y

Barocas, S., Hardt, M., & Narayanan, A. (2017). Fairness in machine learning. Nips tutorial, 1, 2017

Example of Group fairness metrics

SOME FAIRNESS DEFINITIONS CAN BE MUTUALLY EXCLUSIVE.

Group A	Qualified	Unqualified
Admitted	45	2
Rejected	45	8
Total	90	10

Group B	Qualified	Unqualified
Admitted	5	18
Rejected	5	72
Total	10	90

$P(d = 1 \mid Y = 1, A = a)\ \forall\, a \in A$

A *qualified* students *admitted*: 45/90 = **50%**

B *qualified* students *admitted*: 5/10 = **50%**

$P(d = 0 \mid Y = 0, A = a)\ \forall\, a \in A$

A *unqualified* students *rejected*: 8/10 = **80%**

B *unqualified* students *rejected*: 72/90 = **80%**

$P(d = 1 \mid A = a)\ \forall\, a \in A$

Total A students *admitted*: (45+2)/100 = **47%**

Total B students *admitted*: (5+18)/100 = **23%**

Equalized odds satisfied → Both groups 50% of being admitted (TPR) and 80% of being rejected (TNR)

Demographic parity not satisfied → 47% of A admitted and only 23% of B

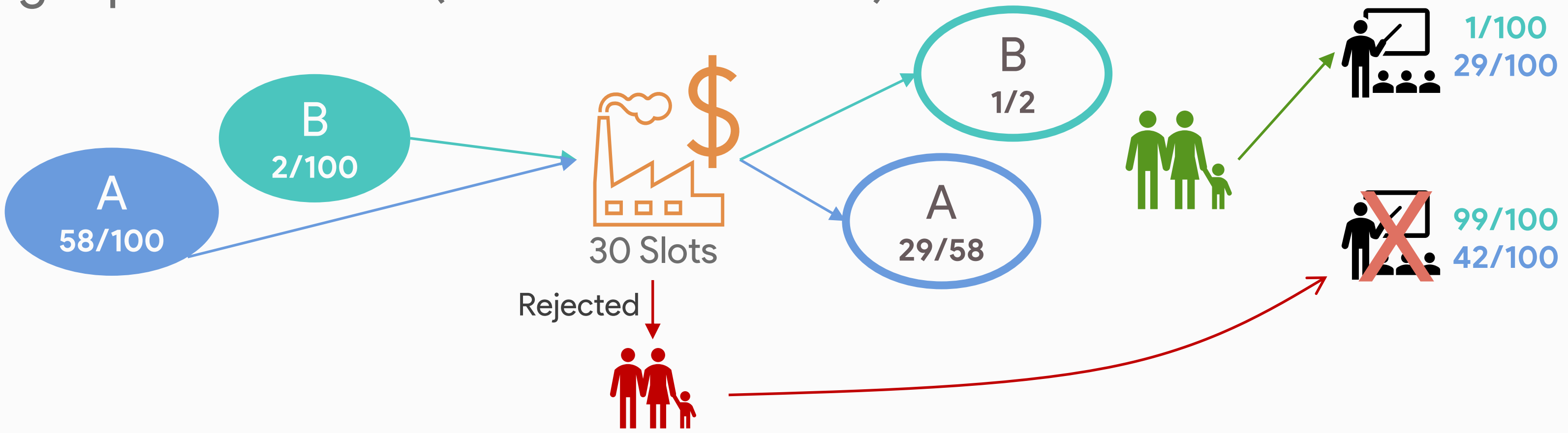
If base rates between groups are different → Impossible to achieve more than one fairness measure

Google Fairness Glosary [Link]

Societal Risks in the application of Group Fairness

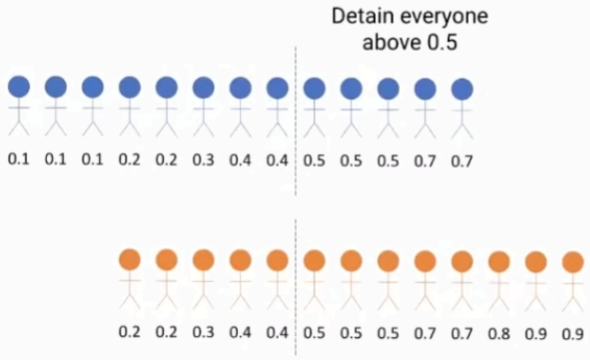
- Satisfying Demographic parity
  - E.g., Perfect predictor (S=Y) is not considered fair when base rates differ (i.e., P[Y=1 |A=a] ≠ P[Y=1|A=b])
  - **laziness**: if we hire the qualified from one group and random people from the other group, we can still achieve demographic parity.

- Satisfying Equalized Odds (same TPR and TNR)



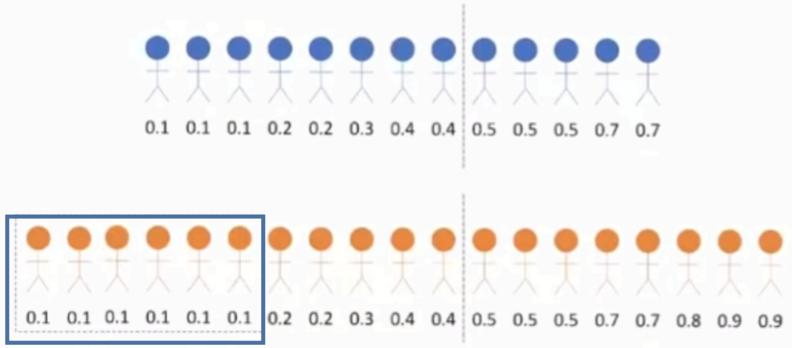
[1] Richard Berka, Hoda Heidaric, Shahin Jabbaric, Michael Kearnsc, and Aaron Rothc. 2017. **Fairness in Criminal Justice Risk Assessments: The State of the Art.**
[2] Alexandra Chouldechova. 2016. **Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments**. Big Data (2016)
[3] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. **Fairness Through Awareness**. 3rd Innovations in Theoretical CS Conference.
[4] Jon M. Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2017. **Inherent Trade-Offs in the Fair Determination of Risk Scores**. In ITCS

Societal Risks in the application of Group Fairness



Detention rate	False pos. rate
38%	25%
61%	42%

— Impendence and error rate parity [EO, FPR] **violated**



Statistical fairness criteria on their own cannot be a proof of fairness, **just a piece of it**

Detention rate	False pos. rate
38%	25%
~~61%~~ 42%	~~42%~~ 26%

Garg, P., Villasenor, J., & Foggo, V. (2020). **Fairness metrics: A comparative analysis**. In 2020 IEEE Big Data. IEEE.
del Barrio, E., Gordaliza, P., & Loubes, J. M. (2020). Review of mathematical frameworks for fairness in machine learning. arXiv
Castelnovo, A., Crupi, R., Greco, G., & Regoli, D. (2021). The zoo of Fairness metrics in Machine Learning. arXiv preprint arXiv:2106.00467
Chiappa, S., & Isaac, W. S. (2018). A causal bayesian networks viewpoint on fairness. In IFIP International Summer School on Privacy and Identity Management. Springer, Cham.Oneto, L., & Chiappa, S. (2020). Fairness in Machine Learning. ArXiv, abs/2012.15816.
Martin Wattenberg, Fernanda Viégas, and Moritz Hardt Attacking discrimination with smarter ML. https://research.google.com/bigpicture/attacking-discrimination-in-ml/
Moritz Hardt - MLSS 2020, Tübingen. https://youtu.be/Igq_S_7IfOU?t=4056
http://www-student.cse.buffalo.edu/~atri/algo-and-society/support/notes/fairness/index.html

Individual Fairness

- **Individual Fairness → treating similar individuals similarly**
  - Difference between individuals similar to difference in predictions
  - More fine-grained than any group-notion fairness: it imposes restriction on for each pair of i.

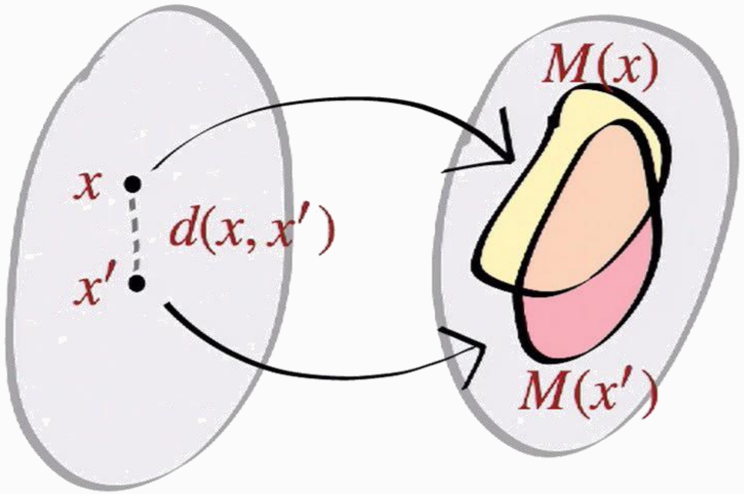
Our Dataset: $\boldsymbol{D} = \{(\boldsymbol{x_i}, \boldsymbol{y_i})\}_i^N$

Distance between $\boldsymbol{x_i}$pairs: $\boldsymbol{k}: \boldsymbol{V}\ \boldsymbol{x}\ \boldsymbol{V} \rightarrow \boldsymbol{R}$.

Mapping from $\boldsymbol{x_i}$ to probability distribution over outcomes $\boldsymbol{M}:\ \boldsymbol{V} \rightarrow \boldsymbol{\alpha A}$

Distance between distributions of outputs $\boldsymbol{D}$

Individual fairness $\boldsymbol{D}(\boldsymbol{M}(\boldsymbol{x}), \boldsymbol{M}(\boldsymbol{y})) \ =< \ \boldsymbol{k}(\boldsymbol{x}, \boldsymbol{y})$



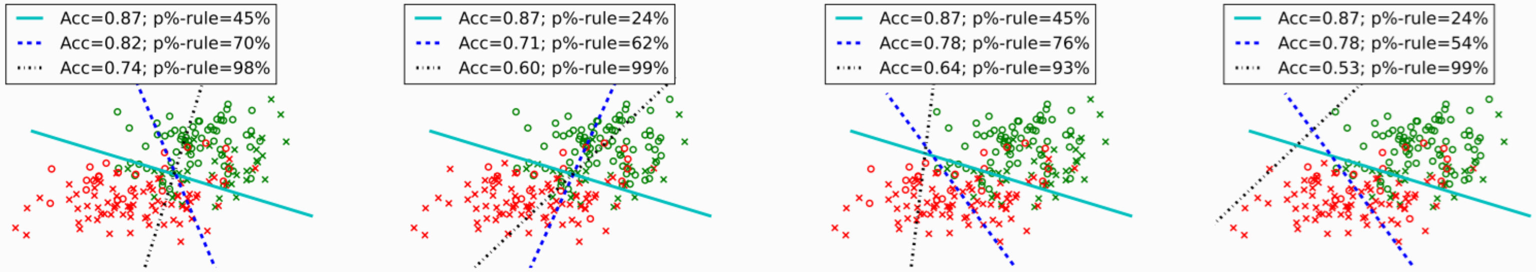
- Big dependence on similarity metric definition both samples and predictions
- How to define appropriate distance metrics for the specific problem and application?

Metric Learning	Graph Theory *More elaborated distances and relationship* *Cliques, communities etc*	Representation Learning *Narrow search space*

Dwork, C., et al.2012. **Fairness through awareness.** Proceedings of the 3rd innovations in theoretical computer science conference, pp. 214-226
Verma, S., & Rubin, J. (2018). **Fairness definitions explained.** In 2018 ieee/acm fairware. IEEE.

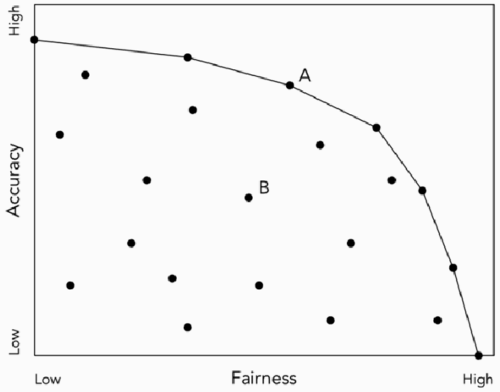
Group and individual flaws?

- **Tradeoffs**
  - Accuracy VS Fairness



$\phi = \pi/4$ $\phi = \pi/8$ $\phi = \pi/4$ $\phi = \pi/8$

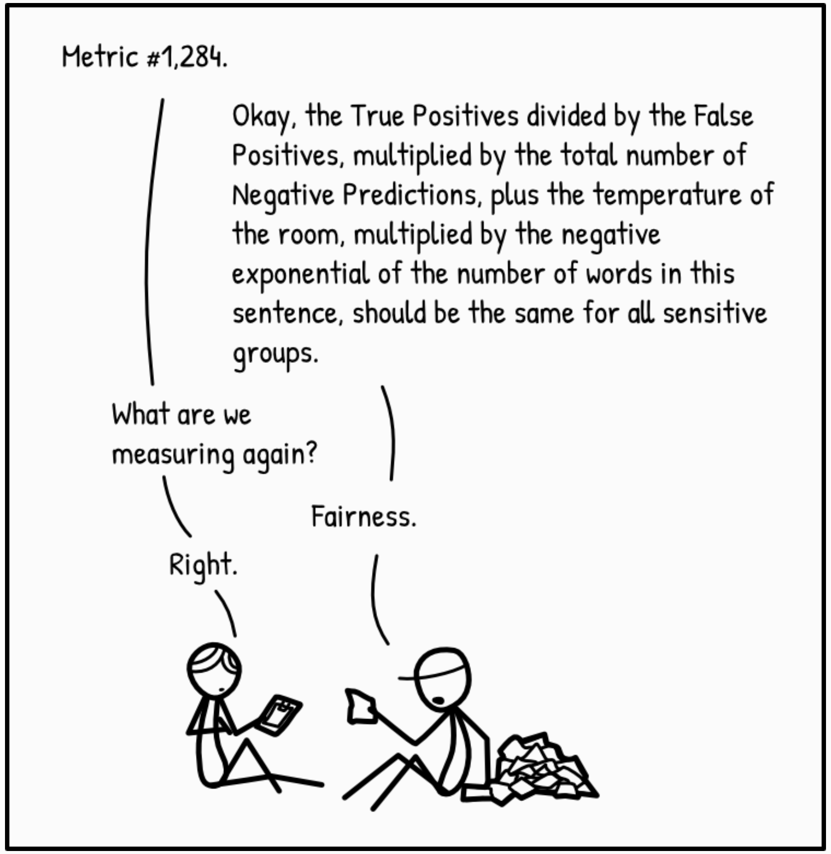
(a) Maximizing accuracy under fairness constraints (b) Maximizing fairness under accuracy constraints



  - Group Fairness Impossibility Theorem
  - Group vs Individual
- Sociological Criticism (Carey et al. 2022)
  - Protected attributes are not discrete. Besides, it's mostly based in social constructs.
  - There shouldn't be tradeoff between group and individual…
  - Be closer to the actual population beliefs

Carey, Alycia N., and Xintao Wu. "The Fairness Field Guide: Perspectives from Social and Formal Sciences." arXiv preprint arXiv:2201.05216 (2022) J. Kleinberg, S. Mullainathan, M. Raghavan, Inherent trade-offs in the fair determination of risk scores, Innovations in Theoretical Computer Science Conference
Barocas, S., Hardt, M., & Narayanan, A. (2017). Fairness in machine learning. Nips tutorial, 1, 2017
Menon, A. K., & Williamson, R. C. (2018, January). The cost of fairness in binary classification. In Conference on Fairness, Accountability and Transparency (pp. 107-118). PMLR
Zafar, M. B., Valera, I., Rogriguez, M. G., & Gummadi, K. P. (2017, April). **Fairness constraints: Mechanisms for fair classification.** In Artificial Intelligence and Statistics . PMLR.

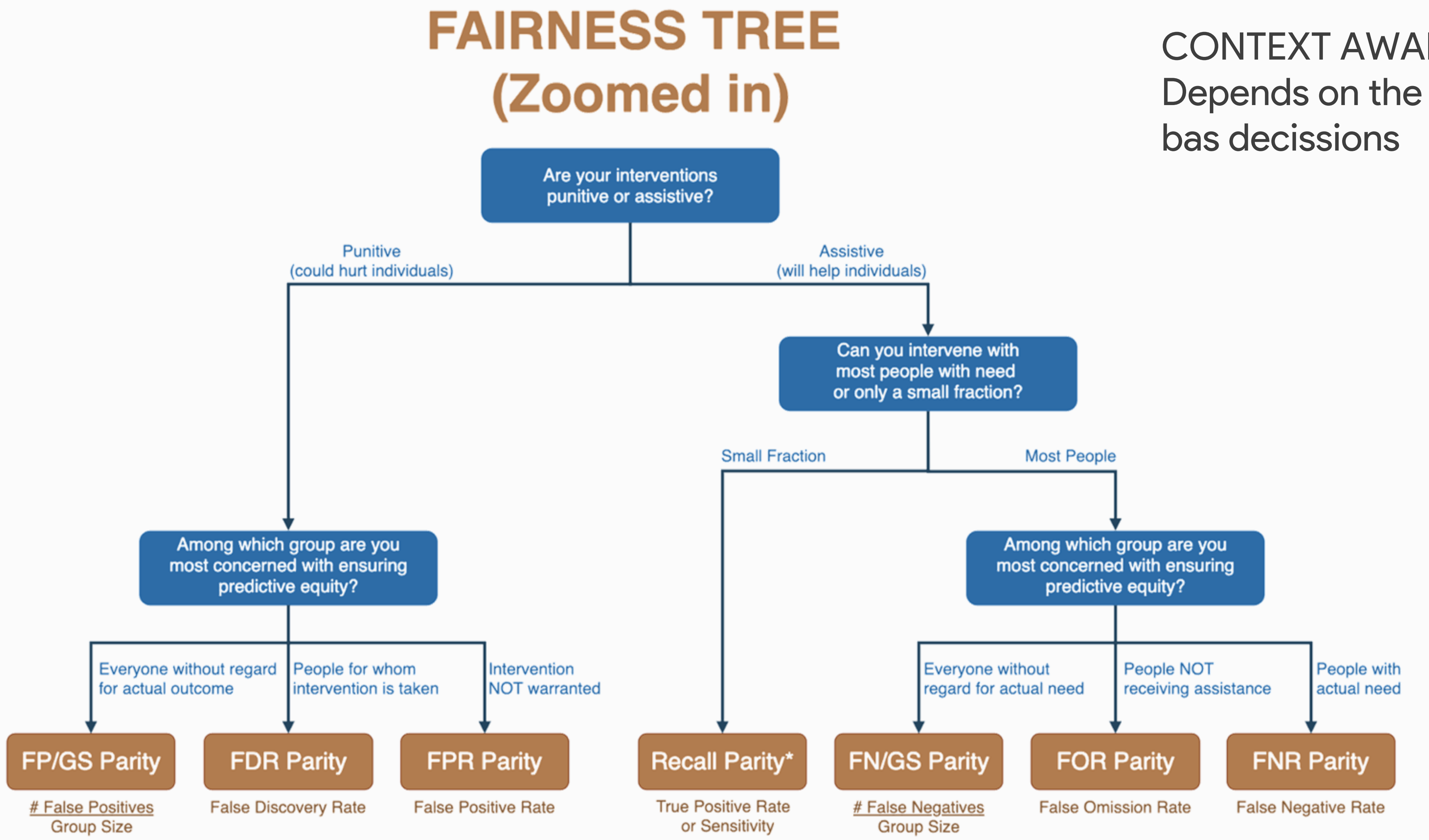
Metrics clarification



Cluster Id	MID	Metrics	Adult	Compas	German	Health	Bank	Student	Titanic	Metric Type
			Datasets							
0	C3	false_omission_rate_difference	Unfair	Fair	Fair	Unfair	Fair	Fair	Unfair	Mis-classification
0	C7	false_omission_rate_ratio	Unfair	Fair	Fair	Unfair	Fair	Unfair	Unfair	
0	C11	error_rate_difference	Unfair	Fair	Fair	Unfair	Fair	Fair	Fair	
0	C12	error_rate_ratio	Unfair	Fair	Fair	Unfair	Fair	Fair	Fair	
		Percentage of agreement	**100%**	**100%**	**100%**	**100%**	**100%**	**75%**	**50%**	
1	C10	average_abs_odds_difference	Unfair	Unfair	Unfair	Unfair	Unfair	Fair	Unfair	Differential Fairness
1	C25	differential_fairness_bias_amplification	Unfair	Unfair	Unfair	Unfair	Unfair	Fair	Unfair	
		Percentage of agreement	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	
2	C16	generalized_entropy_index	Fair	Unfair	Fair	Fair	Fair	Fair	Unfair	Individual Fairness
2	C19	theil_index	Unfair	Unfair	Fair	Unfair	Unfair	Fair	Unfair	
2	C20	coefficient_of_variation	Unfair	Unfair	Unfair	Unfair	Unfair	Unfair	Unfair	
		Percentage of agreement	**67%**	**100%**	**67%**	**67%**	**67%**	**67%**	**100%**	
3	C4	false_discovery_rate_difference	Fair	Fair	Fair	Fair	Fair	Fair	Unfair	Mis-classification
3	C8	false_discovery_rate_ratio	Fair	Fair	Fair	Fair	Fair	Unfair	Unfair	
		Percentage of agreement	**100%**	**100%**	**100%**	**65%**	**100%**	**50%**	**100%**	
4	C0	true_positive_rate_difference	Unfair	Unfair	Fair	Unfair	Unfair	Fair	Unfair	Confusion Matrix Based Group Fairness
4	C1	false_positive_rate_difference	Fair	Unfair	Unfair	Unfair	Unfair	Fair	Unfair	
4	C2	false_negative_rate_difference	Unfair	Unfair	Unfair	Unfair	Unfair	Fair	Unfair	
4	C5	false_positive_rate_ratio	Fair	Unfair	Unfair	Unfair	Unfair	Fair	Unfair	
4	C6	false_negative_rate_ratio	Unfair	Unfair	Unfair	Unfair	Unfair	Unfair	Unfair	
4	C9	average_odds_difference	Unfair	Unfair	Unfair	Unfair	Unfair	Fair	Unfair	
4	C14	disparate_impact	Unfair	Unfair	Unfair	Unfair	Unfair	Unfair	Unfair	
4	C15	statistical_parity_difference	Unfair	Unfair	Unfair	Unfair	Unfair	Fair	Unfair	
		Percentage of agreement	**75%**	**100%**	**88%**	**100%**	**100%**	**75%**	**100%**	
5	C17	between_all_groups_generalized_entropy_index	Fair	Fair	Fair	Fair	Fair	Fair	Fair	Between Group Individual Fairness
5	C18	between_group_generalized_entropy_index	Fair	Fair	Fair	Fair	Fair	Fair	Fair	
5	C21	between_group_theil_index	Fair	Fair	Fair	Fair	Fair	Fair	Fair	
5	C22	between_group_coefficient_of_variation	Fair	Fair	Fair	Fair	Fair	Fair	Unfair	
5	C23	between_all_groups_theil_index	Fair	Fair	Fair	Fair	Fair	Fair	Fair	
5	C24	between_all_groups_coefficient_of_variation	Fair	Fair	Fair	Fair	Fair	Fair	Unfair	
		Percentage of agreement	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**67%**	
6	C13	selection_rate	Unfair	Unfair	Unfair	Unfair	Unfair	Unfair	Unfair	Intermediate Metric
		Percentage of agreement	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	
		Percentage of metrics marking dataset as unfair	**58%**	**54%**	**34%**	**65%**	**50%**	**23%**	**77%**	

Majumder, S., Chakraborty, J., Bai, G. R., Stolee, K. T., & Menzies, T. (2021). Fair Enough: Searching for Sufficient Measures of Fairness. preprint arXiv:2110.13029.

Metrics clarification



CONTEXT AWARE
Depends on the harms of bas decissions

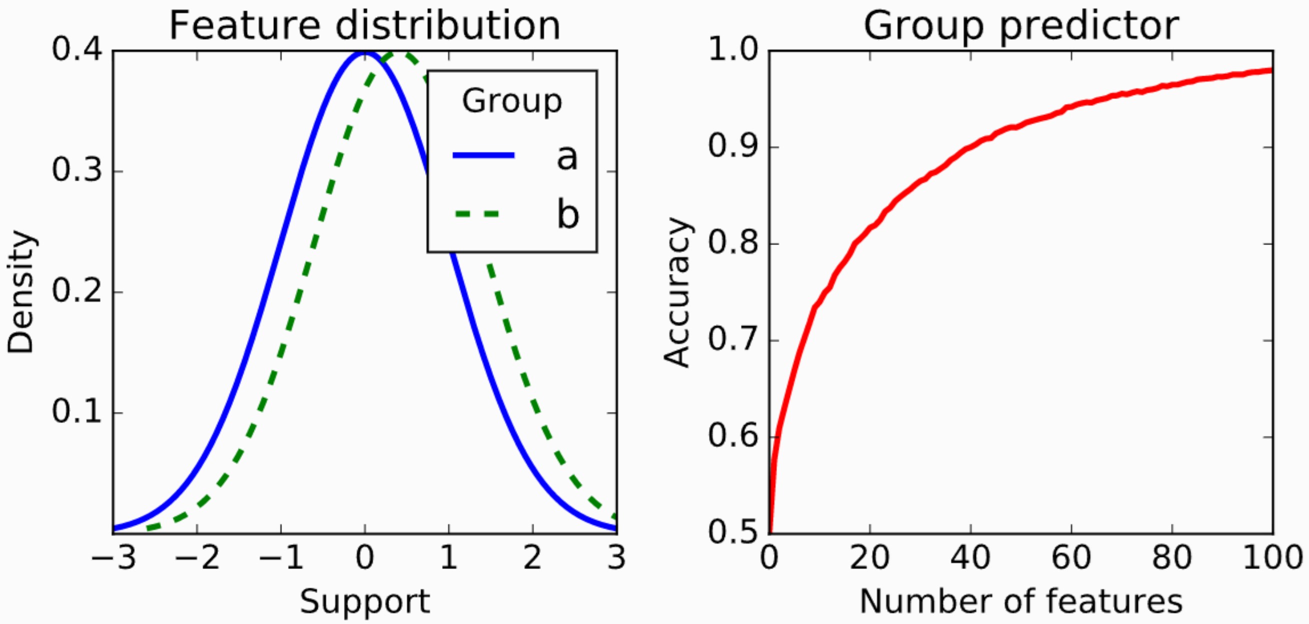
Saleiro, P., et al. (2018). Aequitas: A bias and fairness audit toolkit. arXiv:1811.05577
http://www.datasciencepublicpolicy.org/our-work/tools-guides/aequitas/

Imposing fairness

How to plug chosen fairness definition into the training on ML algorithms?

Fairness through Unawareness

- Does not work → several features may be slightly predictive of A
- Don’t take into account protected attribute → but proxies finally discover it



How to impose fairness

DATA

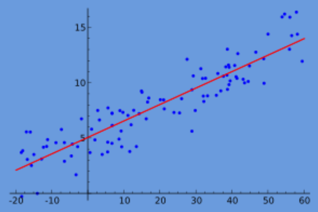
***Transform data** to remove the underlying discrimination in it*

Pre-processing

- Find biases in data – exploratory
- Re-sampling-labeling-weighting
- Data Selection & Valuation
- Fairness through awareness
- Learning Fair Representations

Model agnostic
Inherent learning
Information loss & huge search space

MODEL



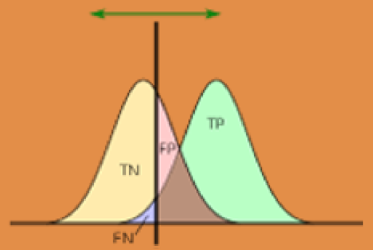
***Fairness constraints** in optimization to penalize discrimination*

In-processing

- Fairness regularizers in Loss
- Prejudice remover
- Adversarial debiasing

Fairness search during optimization process
Very model & problem specific

DECISSION



***Modify decision thresholds of model outputs** to ensure fairness*

Post-processing

- Assessing model fairness
- Equality of opportunity
- Calibration
- Threshold tunning

No retrain needed
We only need access to outcomes
Less efficient

Fair treatment

How to impose fairness

DATA

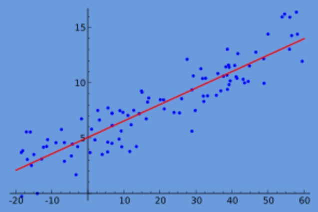
***Transform data** to remove the underlying discrimination in it*

Pre-processing

- Find biases in data – exploratory
- Re-sampling-labeling-weighting
- Data Selection & Valuation
- Fairness through awareness
- Learning Fair Representations

Model agnostic
Inherent learning
Information loss & huge search space

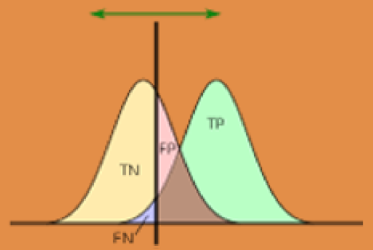
MODEL



***Fairness constraints** in optimization to penalize discrimination*

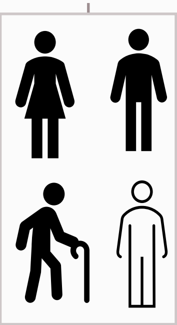
In-processing

DECISSION



***Modify decision thresholds of model outputs** to ensure fairness*

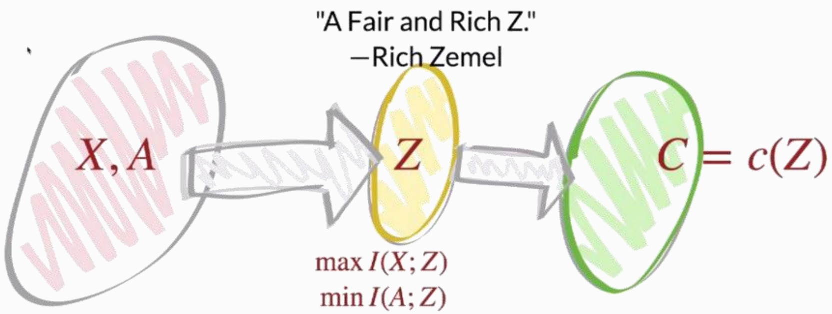
Post-processing



Fair treatment

Pre-processing: Fair Representation Learning

- Approaches
  - Awareness
  - Representation Learning
  - Re-weighting
  - Resampling → Over/Under – SMOTE, etc



- Z → Latent representation
  - $\max_{Z=g(X)} I(X;Z)$
  - subject to $I(A;Z) < e$
  - S⊥A

$$\alpha Loss_{similarity} + \beta Loss_{fairness} + \gamma Loss_{prediction}$$

$$D = \{(a_i, x_i, y_i)\}_{i=1}^{N}$$
$$x_i \in R^d$$
$$g: R^d \to R^r \text{ i.e., } g(x_i) = z_i$$
$$z_i \in R^z$$
$$z_i \perp a_i$$
$$Z \perp A$$

If model involved [hybrid]:
$f(g(X))$

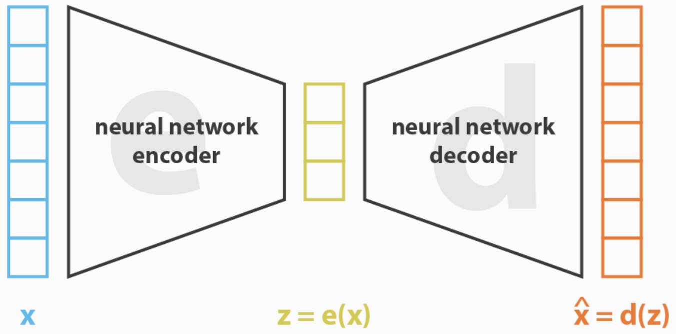
- Strict approach → Optimizes only Statistical Parity or Individual Fairness
  - Info of Y not used
- No need to access A at test time nor Y at representation time
- If Y is used → **hybrid** approach with potential better results [S⊥A|Y and Y⊥A|S]

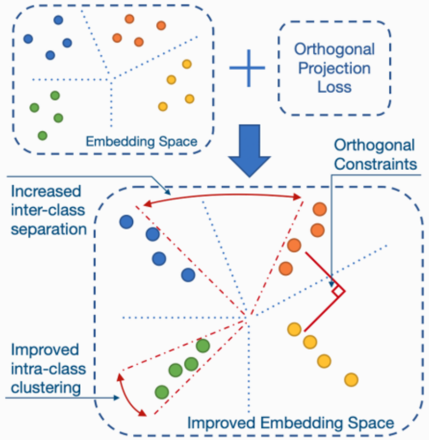
Zemel, R., Wu, Y., Swersky, K., Pitassi, T., & Dwork, C. 2013,. **Learning fair representations**. In International conference on machine learning
Cynthia Dwork,et al. 2012**. Fairness Through Awareness**. In Proceedings of the 3rd Innovations in Theoretical Computer Science Conference
F. Kamiran and T.G.K. Calders. 2012. **Data preprocessing techniques for classification without discrimination.** Knowledge and Information Systems 33

Pre-processing: Fair Representation Learning

Lots of works using NN
max I(A, g(X)) while *min I(A,g(X))* and may *max(g(X),Y)*

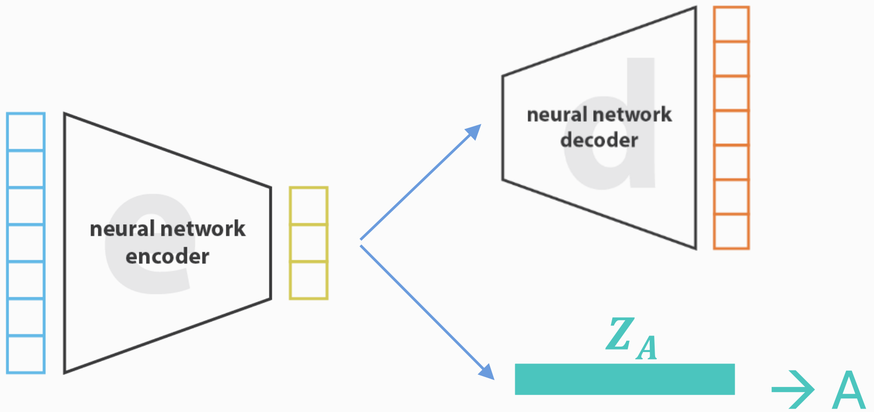
$$Loss_C = |x - x'|^2 - \lambda \, Loss_A(z)$$





$$Loss_C = \alpha|x - x'|^2 + \lambda \, Loss_A(Z_A) + \beta L\perp$$

$$\alpha Loss_{similarity} + \beta Loss_{fairness} + \gamma Loss_{prediction}$$



`aif360.algorithms.preprocessing` **.LFR**

class `aif360.algorithms.preprocessing.LFR`*(unprivileged_groups, privileged_groups, k=5, Ax=0.01, Ay=1.0, Az=50.0, print_interval=250, verbose=0, seed=None)* [source]

Learning fair representations is a pre-processing technique that finds a latent representation which encodes the data well but obfuscates information about protected attributes [2]. .. rubric:: References

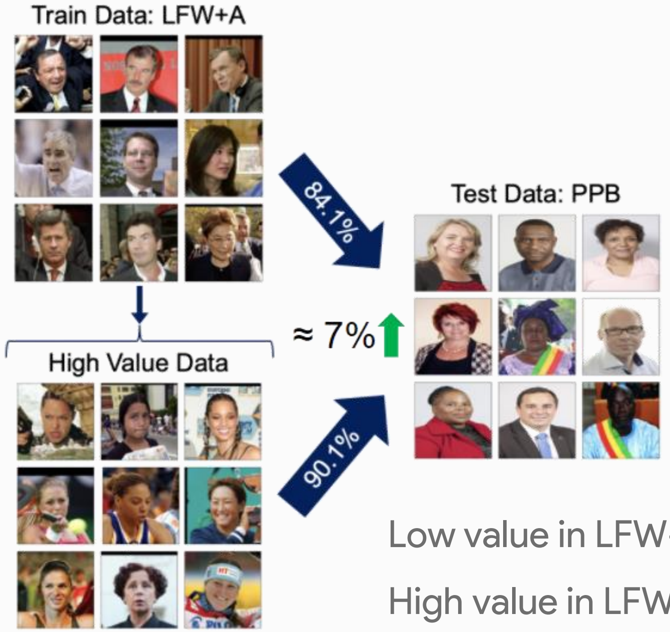
[2] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork, "Learning Fair Representations." International Conference on Machine Learning, 2013.

Bai, H.,et al.(2020). Decaug: Out-of-distribution generalization via decomposed feature representation and semantic augmentation. preprint arXiv:2012.09382
FRLTradeoffs: https://blog.ml.cmu.edu/2020/02/28/inherent-tradeoffs-in-learning-fair-representations/

Pre-processing: Reweighting

- Weight the examples (group, label) to ensure fairness in classification
- Unbalanced learning-related → e.g., Fair-SMOTE
- Advanced example → SHAPLEY values

Domain adaptation: gender detection



Low value in LFW+A - males – overrepresented

High value in LFW+A –women – underrepresented

`aif360.algorithms.preprocessing`.**Reweighing**

class `aif360.algorithms.preprocessing.Reweighing`(*unprivileged_groups, privileged_groups*) [source]

Reweighing is a preprocessing technique that Weights the examples in each (group, label) combination differently to ensure fairness before classification [4].

References

[4] F. Kamiran and T. Calders, "Data Preprocessing Techniques for Classification without Discrimination," Knowledge and Information Systems, 2012.

Ghorbani, A., & Zou, J. (2019, May). Data shapley: Equitable valuation of data for machine learning. In ICML. PMLR
Joymallya Chakraborty, et al. 2021. Bias in Machine Learning Software: Why? How? What to Do?. 29th ESEC/FSE 2021. ACM

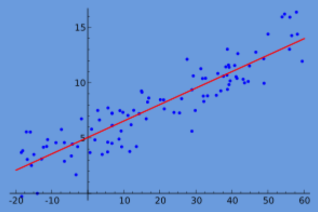
How to impose fairness

DATA

***Transform data** to remove the underlying discrimination in it*

Pre-processing

MODEL



***Fairness constraints** in optimization to penalize discrimination*

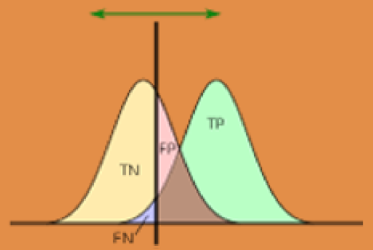
In-processing

- Fairness regularizers in Loss
- Prejudice remover
- Adversarial debiasing

Fairness search during optimization process

Very model & problem specific

DECISSION



***Modify decision thresholds of model outputs** to ensure fairness*

Post-processing

Fair treatment

In-processing

aif360.algorithms.inprocessing.PrejudiceRemover

class aif360.algorithms.inprocessing.PrejudiceRemover(*eta=1.0, sensitive_attr='', class_attr=''*) [source]

Prejudice remover is an in-processing technique that adds a discrimination-aware regularization term to the learning objective [6].

References

[6] T. Kamishima, S. Akaho, H. Asoh, and J. Sakuma, "Fairness-Aware Classifier with Prejudice Remover Regularizer," Joint European Conference on Machine Learning and Knowledge Discovery in Databases, 2012.

- Add penalty to objective function during learning → Regularizer
- Prior work: **Prejudice remover** (Kamishima et al., 2012)
  - Prejudice remover regularizer: Based on the degree of indirect prejudice (PI)

Mutual Information between Y and S

$$\boldsymbol{PI} = \sum_{(y,a)\in D} \hat{P}[y,s] \ln \frac{\hat{P}[y,s]}{\hat{P}[y]\hat{P}[s]}$$

S: *protected/sensitive attribute*

Prejudice remover regularizer

$$\mathrm{R}_{\mathrm{PR}}(\mathcal{D},\boldsymbol{\Theta}) = \sum_{(\mathbf{x}_i,s_i)\in\mathcal{D}} \sum_{y\in\{0,1\}} \mathcal{M}[y|\mathbf{x}_i,s_i;\boldsymbol{\Theta}] \ln \frac{\hat{\Pr}[y|s_i]}{\hat{\Pr}[y]}$$

$$\sum_{(y_i,\mathbf{x}_i,s_i)} \boxed{\ln \mathcal{M}[y_i|\mathbf{x}_i,s_i;\boldsymbol{\Theta}]} + \boxed{\eta\, \mathrm{R}_{\mathrm{PR}}(\mathcal{D},\boldsymbol{\Theta})} + \boxed{\frac{\lambda}{2}\sum_{s\in\mathcal{S}} \|\mathbf{w}_s\|_2^2},$$

Logistic Regression

Prejudice remover regularization

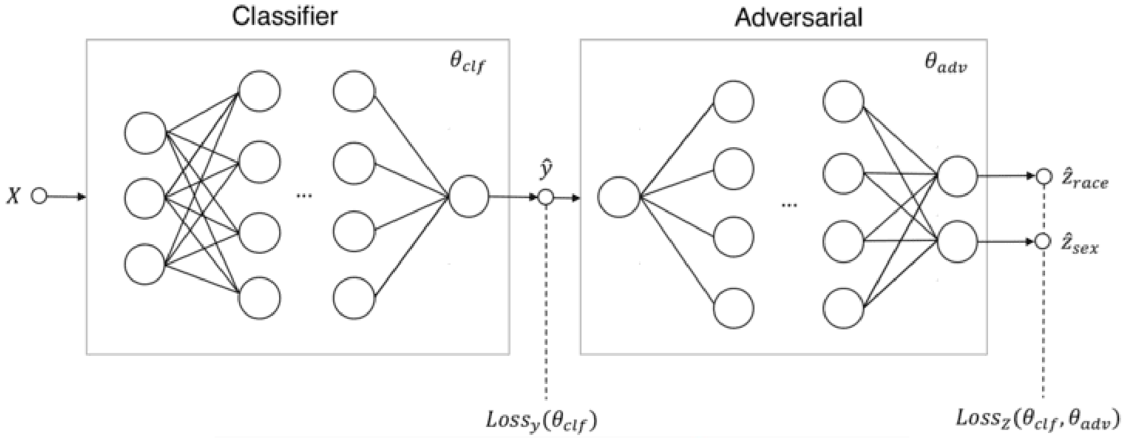
L2 Regularization

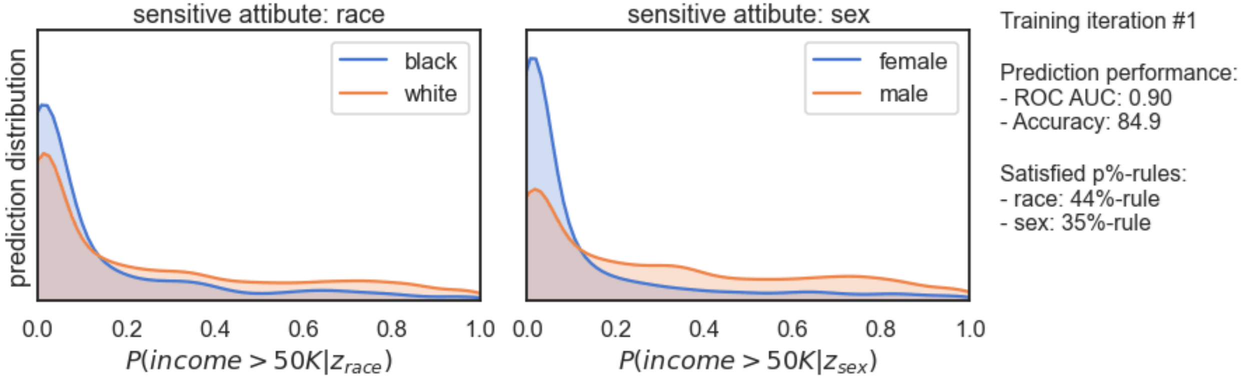
Kamishima, T., Akaho, S., Asoh, H., & Sakuma, J. 2012. Fairness-aware classifier with prejudice remover regularizer. Joint ECML-KDD.

In-processing: Adversarial debiasing

- Make the best possible predictions while ensuring that A cannot be derived from them
  - Demographic Parity
    - Adversary gets $\hat{Y}$
  - Equality Of Odds
    - Adversary gets $\hat{Y}$ and Y
  - Equality Of Opportunity
    - On a given class y → restrict adversary's training set to X where $Y = y$

$$\min_{\theta_{clf}} [Loss_y(\theta_{clf}) - \lambda Loss_Z(\theta_{clf}, \theta_{adv})]$$





aif360.algorithms.inprocessing.AdversarialDebiasing

```
class aif360.algorithms.inprocessing.AdversarialDebiasing(unprivileged_groups, privileged_groups, scope_name, sess, seed=None, adversary_loss_weight=0.1, num_epochs=50, batch_size=128, classifier_num_hidden_units=200, debias=True)
```
[source]

$$p\%rule = \min\left(\frac{P\{\hat{Y} = 1 \mid A = a\}}{P\{\hat{Y} = 1 \mid A = b\}}, \frac{P\{\hat{Y} = 1 \mid A = b\}}{P\{\hat{Y} = 1 \mid A = a\}}\right) \geq \frac{p}{100}$$

Zhang, B. H., et al (2018). Mitigating unwanted biases with adversarial learning. 2018 AAAI/ACM AI, Ethics, and Society (pp. 335-340). https://arxiv.org/pdf/1801.07593.pdf
Towards fairness in ML with adversarial networks. Stijn Tonk. 27 April 2018. URL: https://godatadriven.com/blog/towards-fairness-in-ml-with-adversarial-networks/

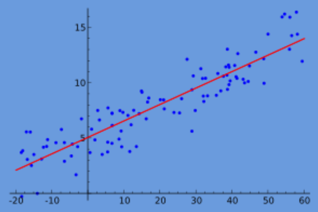
How to impose fairness

DATA

***Transform data** to remove the underlying discrimination in it*

Pre-processing

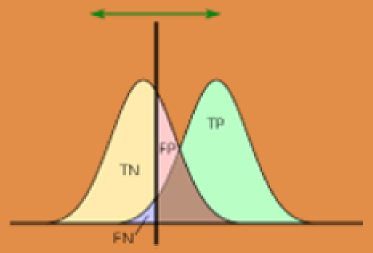
MODEL



***Fairness constraints** in optimization to penalize discrimination*

In-processing

DECISSION



***Modify decision thresholds of model outputs** to ensure fairness*

Post-processing

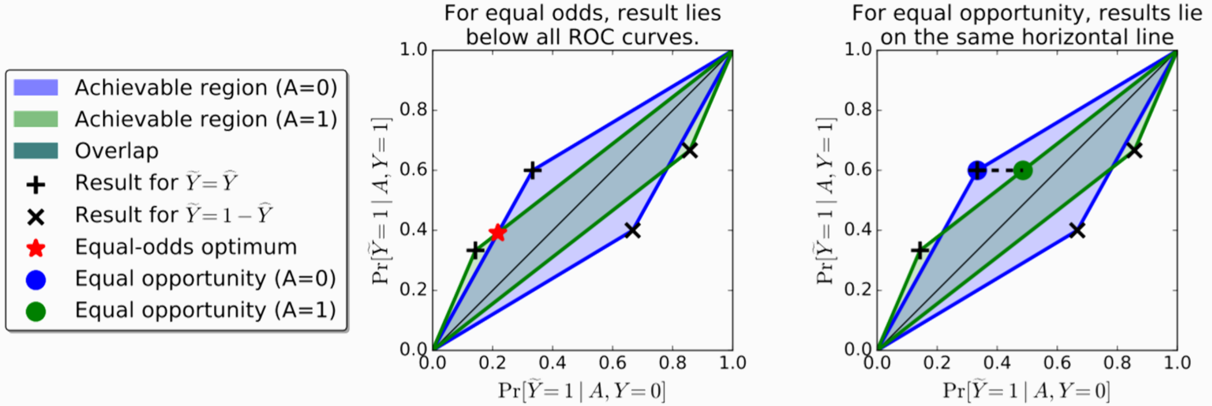
- Assessing model fairness
- Equality of opportunity
- Calibration
- Threshold tunning

No retrain needed
We only need access to outcomes
Less efficient

Fair treatment

Post-processing

- Deal with output predictions of the model
  - Useful in **black-box models** or if we don't have access to the train pipeline → NO retraining
  - **Find a proper threshold** using the output for each group
  - Require A to be available in testing → compliance risk



`aif360.algorithms.postprocessing`.**EqOddsPostprocessing**

class `aif360.algorithms.postprocessing.EqOddsPostprocessing`(*unprivileged_groups, privileged_groups, seed=None*) [source]

Equalized odds postprocessing is a post-processing technique that solves a linear program to find probabilities with which to change output labels to optimize equalized odds [8] [9].

`aif360.algorithms.postprocessing`.**RejectOptionClassification**

class `aif360.algorithms.postprocessing.RejectOptionClassification`(*unprivileged_groups, privileged_groups, low_class_thresh=0.01, high_class_thresh=0.99, num_class_thresh=100, num_ROC_margin=50, metric_name='Statistical parity difference', metric_ub=0.05, metric_lb=-0.05*) [source]

Reject option classification is a postprocessing technique that gives favorable outcomes to unpriviliged groups and unfavorable outcomes to priviliged groups in a confidence band around the decision boundary with the highest uncertainty [10].

Nengfeng Zhou, et al.. 2021. Bias, Fairness, and Accountability with AI and ML Algorithms. arXiv:2105.06558
F. Kamiran, A. Karim, and X. Zhang, 2012 "Decision Theory for Discrimination-Aware Classification," IEEE International Conference on Data Mining
G. Pleiss, M. Raghavan, F. Wu, J. Kleinberg, and K. Q. Weinberger, 2017 "On Fairness and Calibration," Conference on Neural Information Processing Systems
M. Hardt, E. Price, and N. Srebro, 2016 "Equality of Opportunity in Supervised Learning," Conference on Neural Information Processing Systems

More prominent approaches

Causality

Domain-specific
Images
Text
Graphs

Discriminatory Transfer
Multitask Fairness

XAI
Interpretability

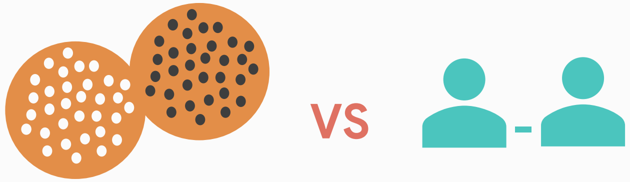
Game theoretical
approaches

Current situation

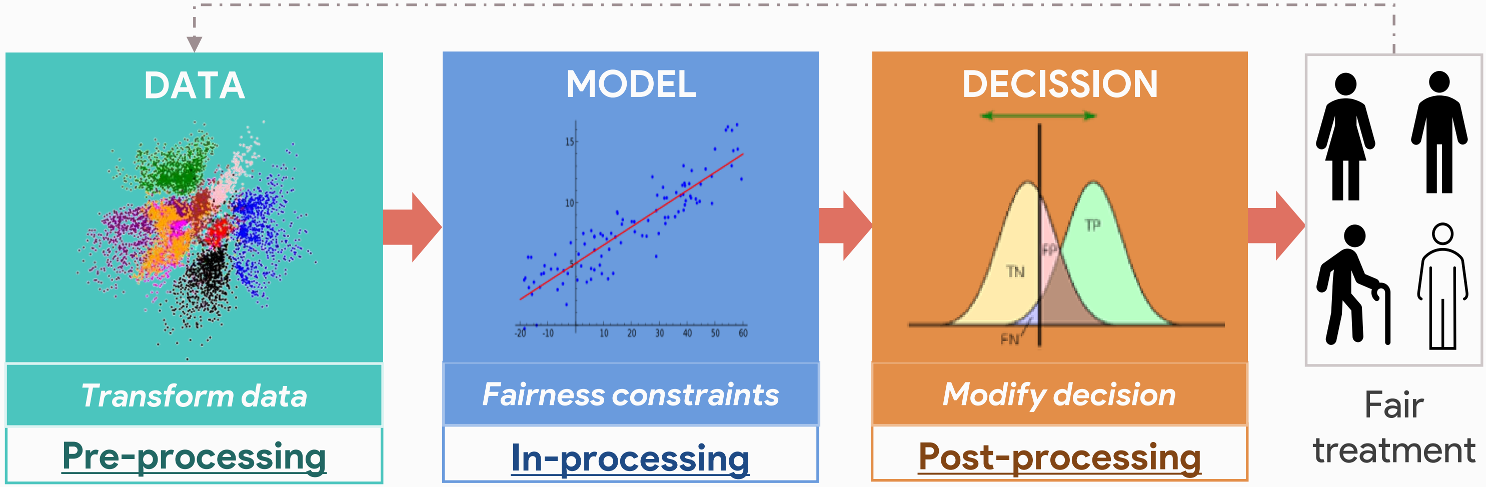
Quick view on graphs & causality

Recap

- Algorithmic Fairness deals with the problem of developing AI-based systems able to treat:
  - **Subgroups in the population <u>equally</u>** → **Group fairness**
  - **Similar individuals in a <u>similar</u> way** → **Individual Fairness**
    - *Specifically, similar individuals from different subgroups*



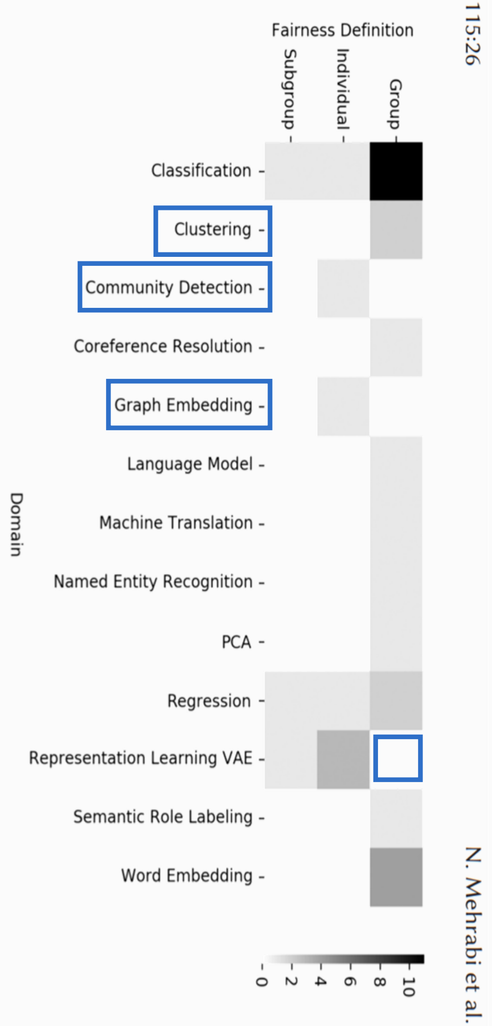
How do we define equally? And similar?

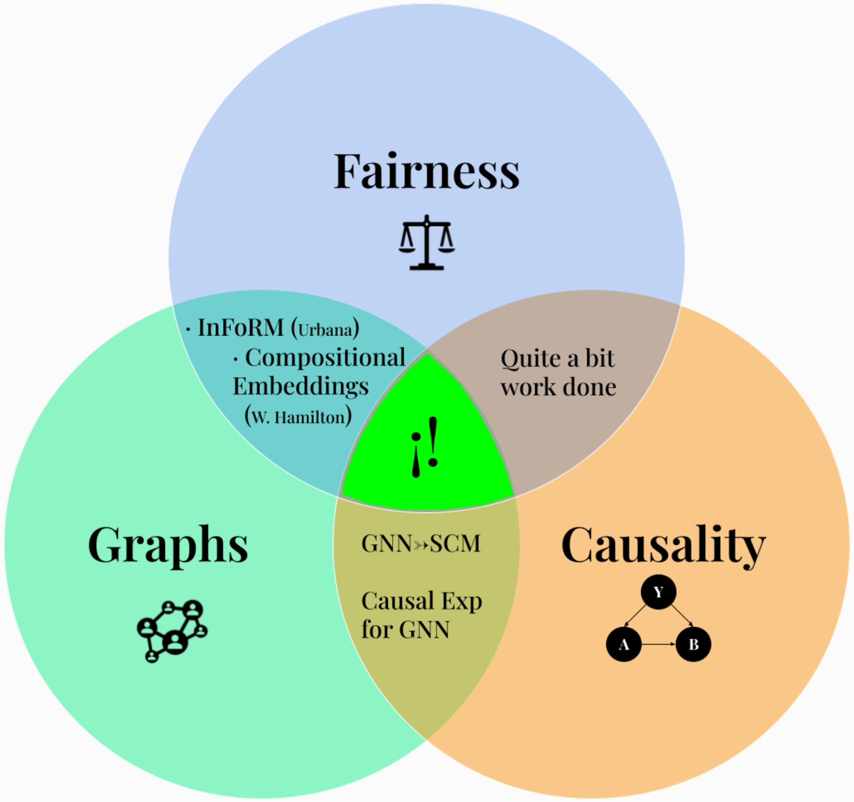


Current landscape

Table 2. List of Papers Targeting and Talking about Bias and Fairness in Different Areas

Area	Reference(s)
Classification	[25, 49, 57, 63, 69, 73, 75, 78, 85, 102, 118, 143, 150, 151, 155]
Regression	[1, 14]
PCA	[133]
Community detection	[101]
Clustering	[8, 31]
Graph embedding	[22]
Causal inference	[82, 95, 111, 112, 123, 156, 160, 161]
Variational auto encoders	[5, 42, 96, 108]
Adversarial learning	[90, 152]
Word embedding	[20, 58, 165] [23, 162]
Coreference resolution	[130, 164]
Language model	[21]
Sentence embedding	[99]
Machine translation	[52]
Semantic role labeling	[163]
Named Entity Recognition	[100]





Mehrabi, N., et al. (2021). **A survey on bias and fairness in machine learning.** ACM Computing Surveys (CSUR), 54(6), 1-35

Why causality or graphs?

- Beyond observational → **Causality**
  - Current only based on statistical based on joint probabilities of $(X, Y, \hat{Y}, A)$
  - Too observational approach, jus take the world as it is
  - What about all the inherent biases in labels?

- Towards robust distances and data relationship → **Graphs**
  - Metrics used in similarity are taken pairwise → not structural information
  - Groups are taken as a whole only regarding their sensitive attribute → not structural info
  - Distance is taken without any context → complex similarity of individuals
    - We should consider the energy and structure of the whole feature space

Graphs & Fairness → Improving robustness

What fairness need? *Defining – detecting – imposing - apply*	**How can Graphs help?**
Capture Individual similarity	– Natural node pairwise distance – Structural similarity – Role similarity – Graph Representation Learning *(for Nodes & Edges & Graphs)*
Capture Group Structure-Behavior	– Community detection – Inherent data structure in graphs – Structural Analysis (e.g., Laplacian)
Capture deeper relationships between data	– Node – Edge classification – Missing link prediction – Message passing – Information Flow – Rewiring – Changing graph structure
Different label bias problems	– Semi-Supervised Learning *i.e., help with labels we cannot see*
Causality	– Strong theory behind graphs – GNN → SCM
Applied to social problems	– Network is the natural structure of data – Also, everything can be modeled as a graph
XAI	– Interpretable by design – Friendly straightforward graph explanations – Great XAI graph-based

Yuan, H., Yu, H., Gui, S., & Ji, S. (2020). Explainability in graph neural networks: A taxonomic survey. arXiv preprint arXiv:2012.15445
Zecevic, M., Dhami, D. S., Velickovic, P., & Kersting, K. (2021). Relating graph neural networks to structural causal models. arXiv preprint arXiv:2109.04173
R. Ying, D. Bourgeois, J. You, M. Zitnik, J. Leskovec. 2019 GNNExplainer: Generating Explanations for Graph Neural Networks, NeurIPS
Bose, A., & Hamilton, W. (2019). Compositional fairness constraints for graph embeddings. ICML. PMLR.

Graphs & Fairness

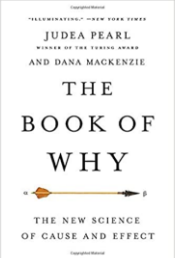
- Group fairness on graphs
  - Fair Graph Ranking → Fair PageRank
  - Fair Graph Clustering
  - Fair Graph embeddings
- Individual Fairness on graphs
  - Similar nodes → similar outcome
- Beyond Group and Individual
  - Degree Related
  - Counterfactual Fairness: Rewire graph to make it fair
- Graph XAI
  - GNN Explainer
  - DIG (Deep into graphs)
- Fairness in Influence Maximization and independent cascades

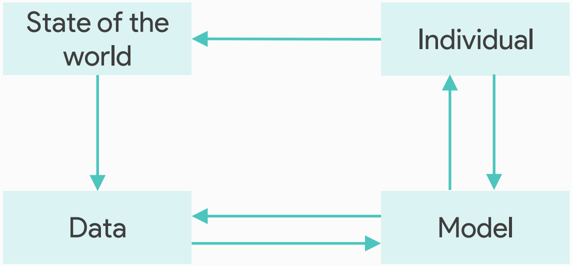
Venkatasubramanian, S., Scheidegger, C., Friedler, S., & Clauset, A. (2021). Fairness in Networks, a tutorial [Link]
Kang, Jian, and Hanghang Tong. "Fair Graph Mining." Proceedings of the 30th ACM International Conference on Information & Knowledge Management. 2021 [Link]

Causality

- Previous definitions relies on **Joint probabilities of (X,Y,S,A)**
  - Reactive vision: take everything as given about the world as it is → Observational
- Can we capture social context? **Let's use causal models**
  - How changes in variables propagate in a system, be it natural, engineered or social
  - What should we do when there's no direct effect?

Exploit Structural Causal Model properties to look for biases Neal, B. (2020)





Definition 4.2 (Structural Causal Model (SCM)) *A structural causal model is a tuple of the following sets:*

1. *A set of endogenous variables V*
2. *A set of exogenous variables U*
3. *A set of functions f, one to generate each endogenous variable as a function of other variables*

$$M: \quad \begin{aligned} B &:= f_B(A, U_B) \\ C &:= f_C(A, B, U_C) \\ D &:= f_D(A, C, U_D) \end{aligned}$$

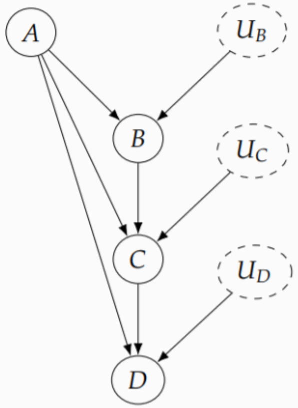
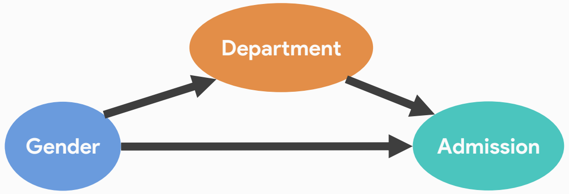


Figure 4.8: Graph for the structural equations in Equation 4.24.



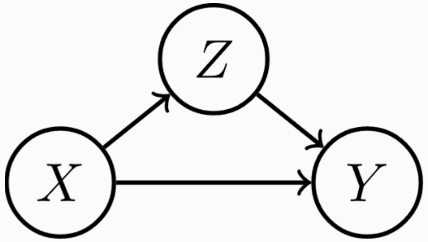
Causal fairness criteria and path-specific effects

J. Pearl, 2009 Causality: Models, Reasoning and Inference, 2nd ed. New York, NY, USA: Cambridge University Press,
Neal, B. (2020). Introduction to causal inference from a ML perspective. *Book (draft)*. https://www.bradyneal.com/Introduction_to_Causal_Inference-Dec17_2020-Neal.pdf
Kusner, M. J., Loftus, J. R., Russell, C., & Silva, R. (2017). Counterfactual fairness.
Loftus, J. R., Russell, C., Kusner, M. J., & Silva, R. (2018). Causal reasoning for algorithmic fairness
Makhlouf, K., Zhioua, S., & Palamidessi, C. (2020). Survey on Causal-based Machine Learning Fairness Notions. arXiv preprint arXiv:2010.09553.
Kilbertus, N., Rojas-Carulla, M., Parascandolo, G., Hardt, M., Janzing, D., & Schölkopf, B. (2017). Avoiding discrimination through causal reasoning
Zhang, J., & Bareinboim, E. (2018, April). Fairness in decision-making—the causal explanation formula. In Thirty-Second AAAI
Wu, Y. (2020). Achieving Causal Fairness in Machine Learning
S. Chiappa. 2019, Path-specific counterfactual fairness. Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19)
Chiappa, S., & Isaac, W. S. (2018,). A causal bayesian networks viewpoint on fairness. In IFIP International Summer School on Privacy and Identity Management
Fairness – Moritz Hardt – Part 2 – MLS2020 - https://www.youtube.com/watch?v=9oNVFQ9llPc&t=1449s

Counterfactual

- **Counterfactual** →*"Would I have been hired if I were non-black?" "Would I have avoided the traffic jam had I taken a different route this morning?"*
  - Decision does not depend on protected attribute

- The counterfactual $\boldsymbol{Y}_{\{X:=1, Z:=Z_{X:=0}\}}$ is the value that Y would obtain had X been set to 1 and had Z been set to the value Z would've assumed had X been set to 0



- Fair Causal graph → if Y don't depend on A, i.e., no A-Y way
  - Make decision only using non-descendants of A in the causal graph
  - PATH-SPECIFIC Fairness

- Difficult task of agreeing on which graph to build and validating it
- Impossible to test an existing classifier against **strict** causal definitions of fairness
- What should we do when not we are not able to built neither validate a causal graph?
  - **Counterfactual discrimination criteria → normative fairness criteria**

M.J. Kusner, J. Loftus, C. Russell and R. Silva, **Counterfactual fairness**, In Neural Information Processing Systems, (2017)
Barocas, S., Hardt, M., & Narayanan, A. (2017). **Fairness in machine learning.** Nips tutorial, 1, 2017
Shira Mitchell. 2018. **Reflection on quantitative fairness.** Web Book

Takeaways

Other cultural and conceptual challenges

Even we are looking for bias, **we are inducing bias**

CONTEXT MATTERS
Quantitative techniques
+ policy-level questions

Make methods flexible to **adapt to each situation, context and use**

PUBLIC'S NOTION OF FAIRNESS
Explicitly connect fairness criteria to different **socio-cultural and philosophical values**

Try to **unify fairness** definition and framework

Politics and law **implication**

Remind: Fairness and unfairness are related but different concepts

Make Fair ML research **accessible** to general public, other researchers

From equality to equity
Give each one the resources that each one need to reach to the same point

Example of conceptual bias: **Why groups should be treated as discrete categories?**

- Most definitions of protected attribute-group relies on **categoric division → implicit cultural bias & unstable social construct**
- Other possibility: intersectional modelling → **Protected attribute as continuous variables**
    - Quantify fairness along one dimension (e.g., age) conditioned on another dimension (e.g., skin tone)

e.g., Use Computer vision clustering of skin tones instead of pre-defined ethnics

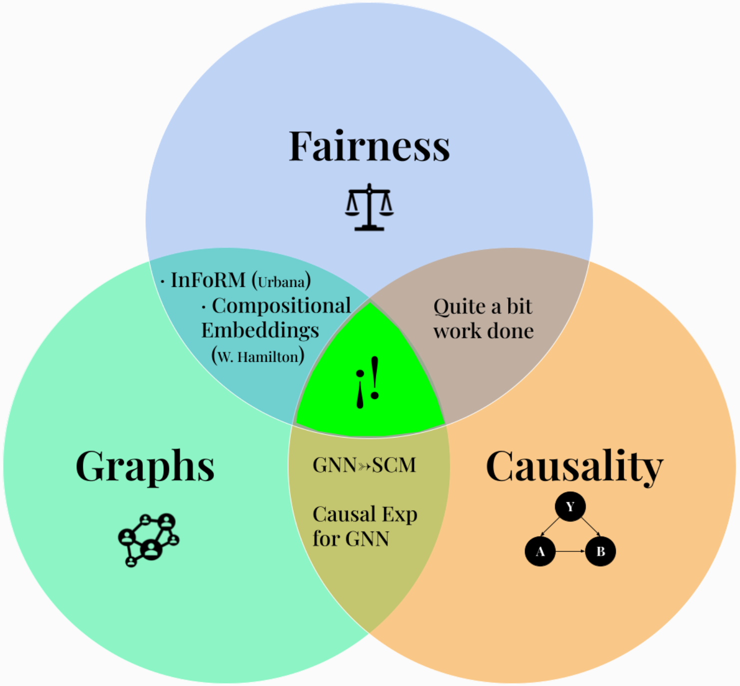
Hutchinson, B., & Mitchell, M. 2019. **50 years of test (un) fairness: Lessons for machine learning**. FAccT 2019
Nancy S Cole and Michael J Zieky. 2001. The new faces of fairness. Journal of Educational Measurement 38, 4
Rebecca Zwick and Neil J Dorans. 2016. Philosophical Perspectives on Fairness in Educational Assessment. In Fairness in Educational Assessment and Measurement
T. Anne Cleary. 1966. Test bias: Validity of the Scholastic Aptitude Test for Negro and white students in integrated colleges
Calders, Kamiran, and Pechenizkiy, "Building Classifiers with Independency Constraints," in In Proc. IEEE ICDMW, 2009, 13–18
Kamiran and Calders, "Classifying Without Discriminating," in Proc. 22Nd International Conference on Computer, Control and Communication, 2009.

Conclusion

- **Don't feel overwhelmed** by the big amount methods and measures!
  - Method depends on task, and technical context
  - Definitions and metrics depends on the context
  - Development and relationship of the measures with ethics
    → Now **you choose context** – experts – social and ethical analysis (Frameworks & Guidelines)
  - More work in create context-dependent

- More work needed in **ethical-cultural aspect**
  - Equity → Considering individual resources
  - Continual protected attributes
  - Social-Law-Political needs close relationship
  - Real impact of models: performative prediction (Hardt, 2010)

- **Technical takeaways**
  - Beyond observational → Causality
  - Deep structural data relationship → Graphs



Resources

Libraries

IBM Research Trusted AI

AI Fairness 360

FAIRENSICS

Fairlearn

FairKit

Aequitas

Bias & Fairness Audit

Benchmarking datasets

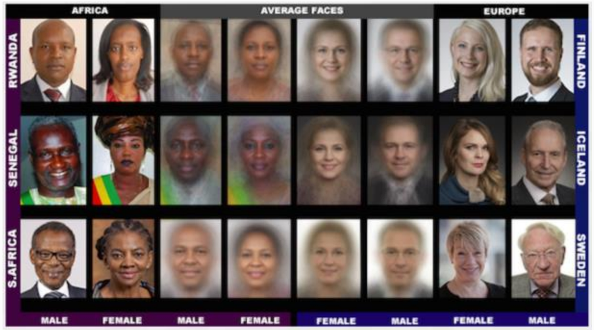
- Big amount of tabular dataset in all domains

- Every dataset may have intrinsic bias

Images

Text

School Effectiveness	[66]	15362	9	Ethnicity, Gender	R
Heart Disease	[90]	303	75	Age, Gender	MC, R
German Credit	[85]	1K	20	Age, Gender/Marital-Stat	MC
Census/Adult Income	[112]	48842	14	Age, Ethnicity, Gender, Native-Country	BC
Contraceptive Method Choice	[121]	1473	9	Age, Religion	MC
Law School Admission	[187]	21792	5	Ethnicity, Gender	R
Arrhythmia	[70]	452	279	Age, Gender	MC
Communities & crime	[169]	1994	128	Ethnicity	R
Wine Quality	[154]	4898	13	Color	MC, R
Heritage Health	[146]	≈60K	≈20	Age, Gender	MC, R
Stop, Question & Frisk	[45]	84868	≈100	Age, Ethnicity, Gender	BC, MC
Bank Marketing	[142]	45211	17-20	Age	BC
Diabetes US	[181]	101768	55	Age, Ethnicity	BC, MC
Student Performance	[38]	649	33	Age, Gender	R
CelebA Faces	[122]	≈200K	40	Gender Skin-Paleness, Youth	BC
xAPI Students Perf.	[6]	480	16	Gender, Nationality, Native-Country	MC
Chicago Faces	[127]	597	5	Ethnicity, Gender	MC
Credit Card Default	[195]	30K	24	Age, Gender	BC
COMPAS	[119]	11758	36	Age, Ethnicity, Gender	BC, MC
MovieLens	[77]	100K	≈20	Age, Gender	R
Drug Consumption	[54]	1885	32	Age, Ethnicity, Gender, Country	MC
Student Academics Perf.	[87]	300	22	Caste, Gender	MC
NLSY	[148]	≈10K		Birth-date, Ethnicity, Gender	BC, MC, R
Diversity in Faces	[140]	1 M	47	Age, Gender	MC, R



Pilot Parliaments Benchmark

Retiring Adult:
New Datasets for Fair Machine Learning

Frances Ding* UC Berkeley | Moritz Hardt* UC Berkeley | John Miller* UC Berkeley | Ludwig Schmidt* Toyota Research Institute

Quy, T. L., Roy, A., Iosifidis, V., & Ntoutsi, E. (2021). A survey on datasets for fairness-aware machine learning. arXiv
Oneto, L. (2020). Learning fair models and representations. Intelligenza Artificiale, 14(1), 125-152
Barocas, S., Hardt, M., & Narayanan, A. (2017). **Fairness in machine learning**. Nips tutorial, 1, 2017
Majumder, S., Chakraborty, J., Bai, G. R., Stolee, K. T., & Menzies, T. (2021). Fair Enough: Searching for Sufficient Measures of Fairness. preprint arXiv:2110.13029.
http://gendershades.org/overview.html - https://nips.cc/media/neurips-2021/Slides/26854.pdf

Bibliography

More references in each slide

- M. Hardt, E. Price, N. Srebro, Equality of opportunity in supervised learning, Advances in Neural Information Processing Systems (2016).
- Cynthia Dwork,et al. 2012. Fairness Through Awareness. In Proceedings of the 3rd Innovations in Theoretical Computer Science Conference
- Alexandra Chouldechova. 2016. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. Big Data.
- Verma, J. Rubin, **Fairness definitions explained**, IEEE/ACM International Workshop on Software Fairness (2018) 1–7.
- Carey, A. N., & Wu, X. (2022). **The Fairness Field Guide: Perspectives from Social and Formal Sciences**. https://arxiv.org/pdf/2201.05216.pdf
- Richard Berka, Hoda Heidaric, Shahin Jabbaric, Michael Kearnsc, and Aaron Rothc. 2017. Fairness in Criminal Justice Risk Assessments: The State of the Art.
- Alexandra Chouldechova. 2016. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. Big Data (2016)
- Jon M. Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2017. Inherent Trade-Offs in the Fair Determination of Risk Scores. In ITCS
- Ben Hutchinson and Margaret Mitchell. 2019. 50 Years of Test (Un)fairness: Lessons for Machine Learning.
- M.J. Kusner, J. Loftus, C. Russell and R. Silva**, Counterfactual fairness**, In Neural Information Processing Systems, (2017)
- Barocas, S., Hardt, M., & Narayanan, A. (2017). **Fairness in machine learning**. Nips tutorial, 1, 2017 and book
- Shira Mitchell. 2018. Reflection on quantitative fairness. Web Book
- Majumder, S., Chakraborty, J., Bai, G. R., Stolee, K. T., & Menzies, T. (2021). Fair Enough: Searching for Sufficient Measures of Fairness. arXiv preprint arXiv:2110.13029.
- L. Oneto, Learning fair models and representations, Intelligenza Artificiale 14 (1) (2020) 151–178.
- Castelnovo, A., Crupi, R., Greco, G., & Regoli, D. (2021). The zoo of Fairness metrics in Machine Learning. arXiv
- Franco, D., Navarin, N., Donini, M., Anguita, D., & Oneto, L. (2022). Deep fair models for complex data: Graphs labeling and explainable face recognition. Neurocomputing, 470
- A.F. Winfield, K. Michael, J. Pitt, V. Evers, Machine ethics: the design and governance of ethical ai and autonomous systems, Proceedings of the IEEE 107 (2019) 509–517
- D. Gunning, Explainable artificial intelligence (xai), Defense Advanced Research Projects Agency (DARPA), nd Web 2 (2)
- Majumder, S., Chakraborty, J., Bai, G. R., Stolee, K. T., & Menzies, T. (2021). Fair Enough: Searching for Sufficient Measures of Fairness. preprint arXiv:2110.13029.
- Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. Calif. L. Rev., 104, 671
- Lim Swee Kiat. Retrieved December 2021. Machines go Wrong. https://machinesgonewrong.com/fairness/
- Manuel Gomez Rodriguez et al. (2020). Human-Centric Machine Learning Feedback loops, Human-AI Collaboration and Strategic Behavior [Link]. Web
- Corbett-Davies & Goel. (2018). The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning
- Mehrabi, N., et al. (2021). **A survey on bias and fairness in machine learning**. ACM Computing Surveys (CSUR), 54(6), 1-35
- 2017. CS 294**: Fairness in Machine Learning**. https://fairmlclass.github.io (2017). Online; accessed February 2018
- **Google glossary** https://developers.google.com/machine-learning/glossary/fairness

ellis

ALICANTE unit

Talk in the scope of the project:

Achieving Fair, Accountable and Transparent Machine Learning Models through Graph Theory and Causality

Thesis in Progress by PhD Student Adrián Arnaiz Rodríguez

PhD Nuria Oliver

PhD Francisco Escolano

PhD Manuel Gómez Rodríguez

ellis
ALICANTE unit

Thank you!

Q's & feedback?

adrian@ellisalicante.org

@arnaiztech

AdrianArnaiz

European Union
European Regional
Development Fund

GOBIERNO DE ESPAÑA
MINISTERIO DE CIENCIA E INNOVACIÓN

GENERALITAT VALENCIANA
Conselleria d'Innovació, Universitats, Ciència i Societat Digital

Sabadell
Fundación