Sprawiedliwość w uczeniu maszynowym

Fairness in machine learning

Adam Żychowski

Definicja

Uczciwość

postępowanie zgodnie z przyjętymi zasadami lub prawem

Sprawiedliwość

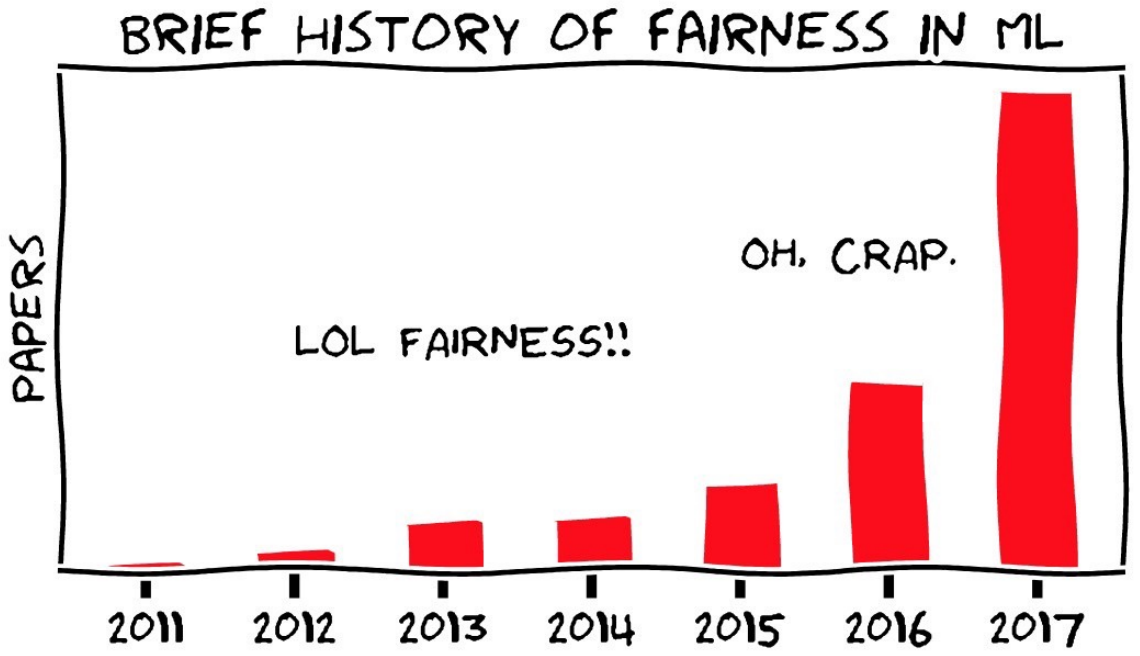
stała i niezmienna wola przyznania każdemu należnego mu prawa

Fairness

brak jakichkolwiek uprzedzeń lub faworyzowania jednostki lub grupy ze względu na jej wrodzone lub nabyte cechy

Popularność

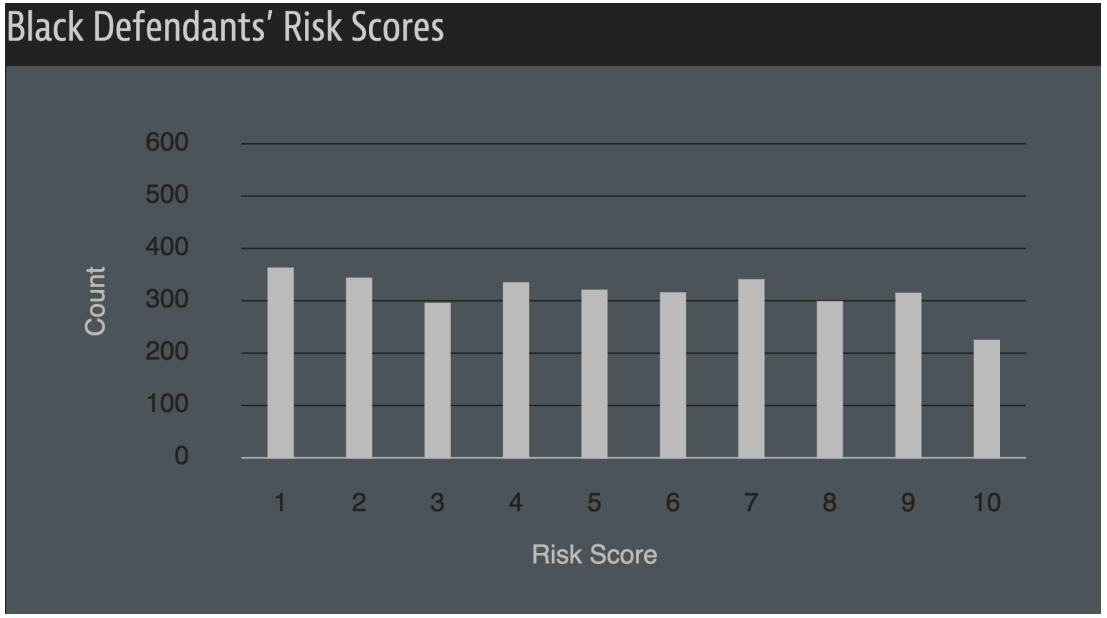
- znaczny wzrost popularności od 2016 roku
- ACM Conference on Fairness, Accountability, and Transparency (od 2018)
- ICML 2018 – dwie z 5 nagrodzonych prac były o sprawiedliwości
- każdego tygodnia kilka nowych prac

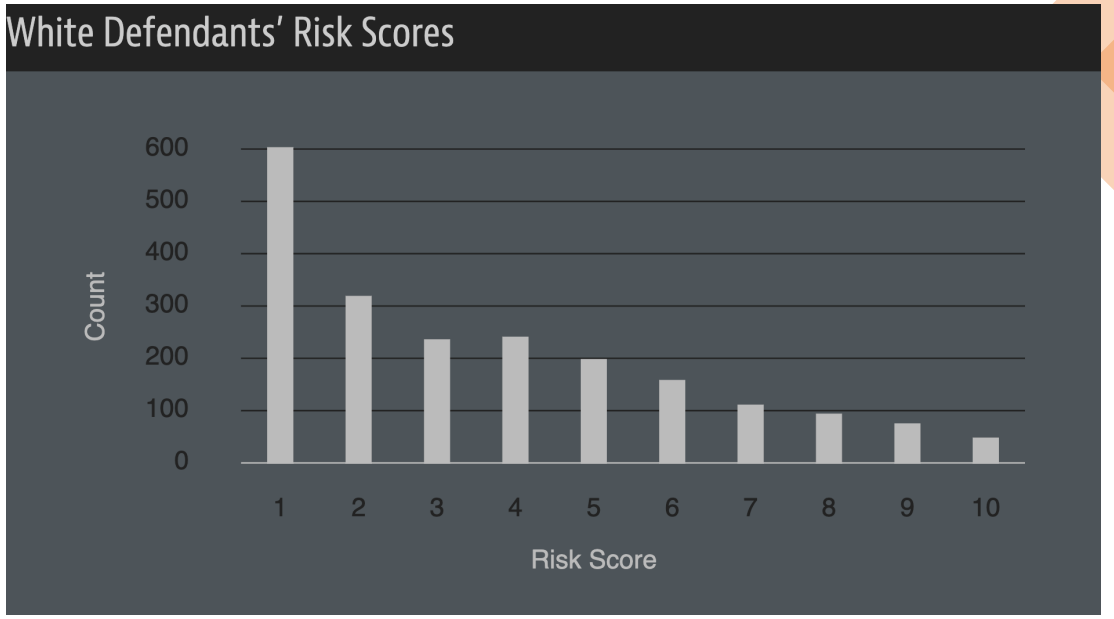


https://towardsdatascience.com/a-tutorial-on-fairness-in-machine-learning-3ff8ba1040cb

COMPAS

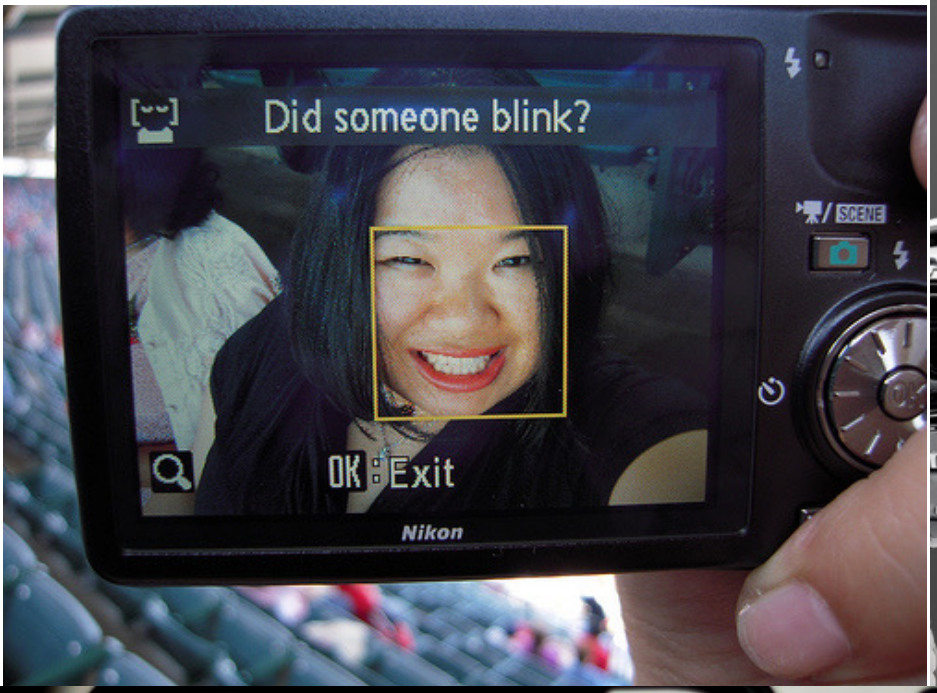
- Correctional Offender Management Profiling for Alternative Sanctions (COMPAS)

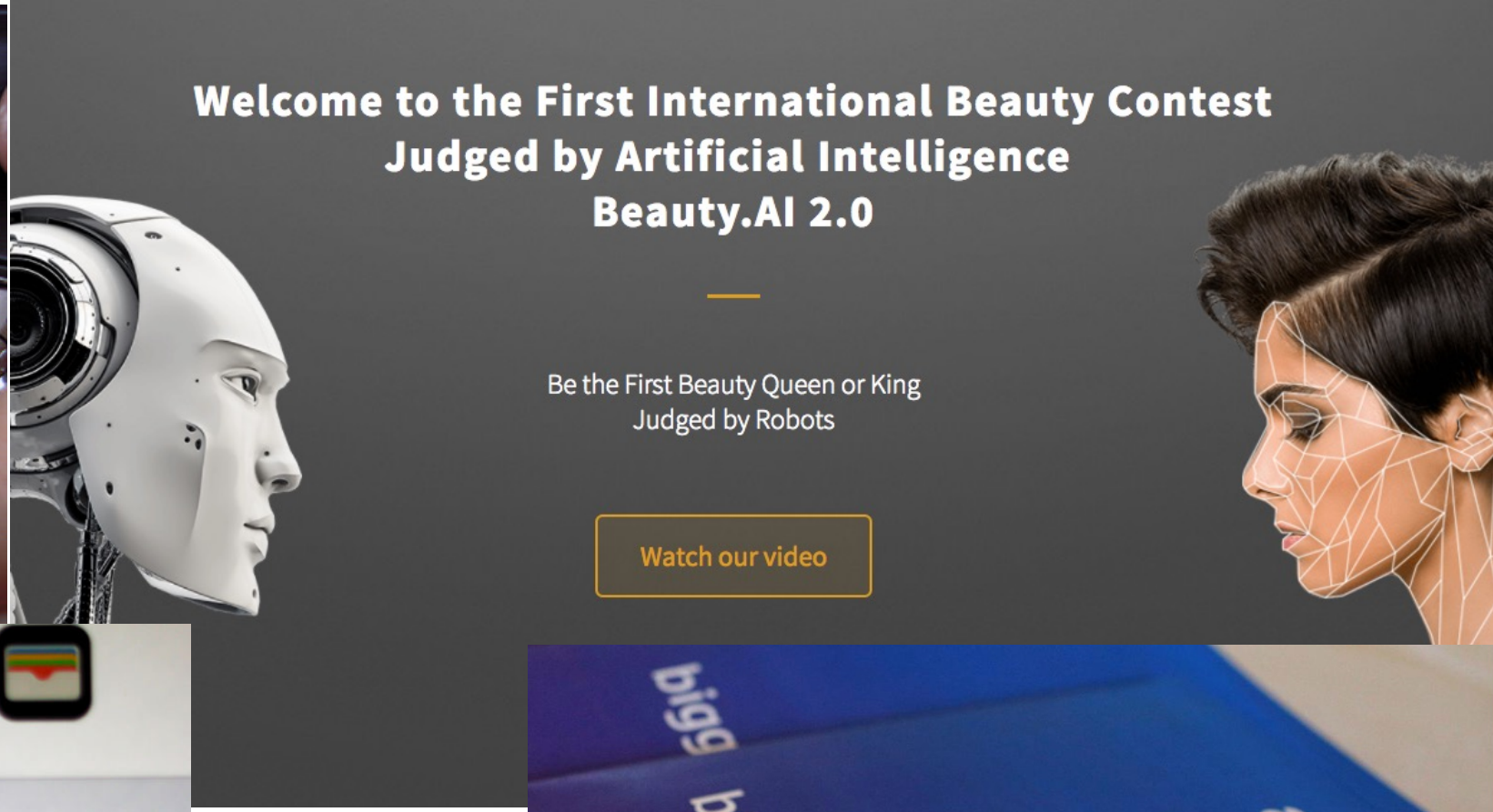




	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing


Did someone blink?
OK Exit
Nikon


Welcome to the First International Beauty Contest
Judged by Artificial Intelligence
Beauty.AI 2.0
Be the First Beauty Queen or King
Judged by Robots
Watch our video


9:41
Done
apple.com/apple-card/
Mac
iPhone
Watch
Apple Card
Card


amazon
Amazon's biggest hiring event of the year
AUG 2


a lawyer


a builder


ceo


a nurse

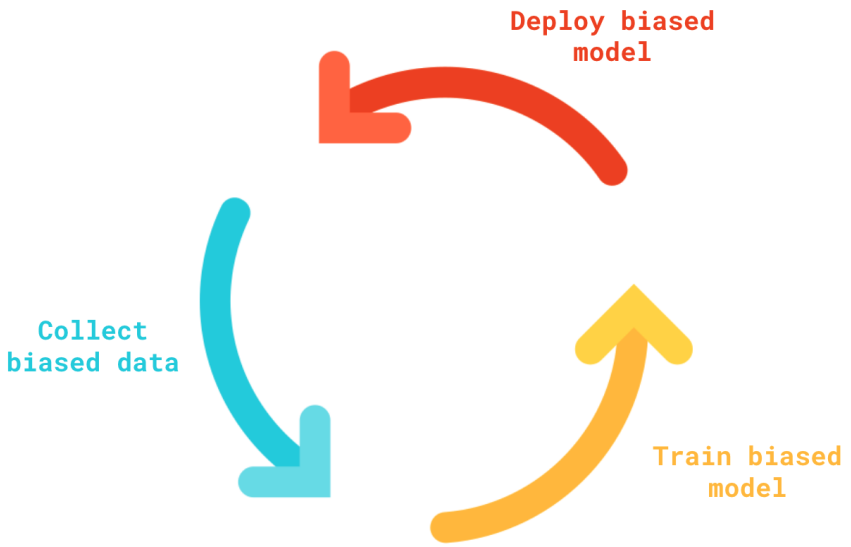

a flight attendant

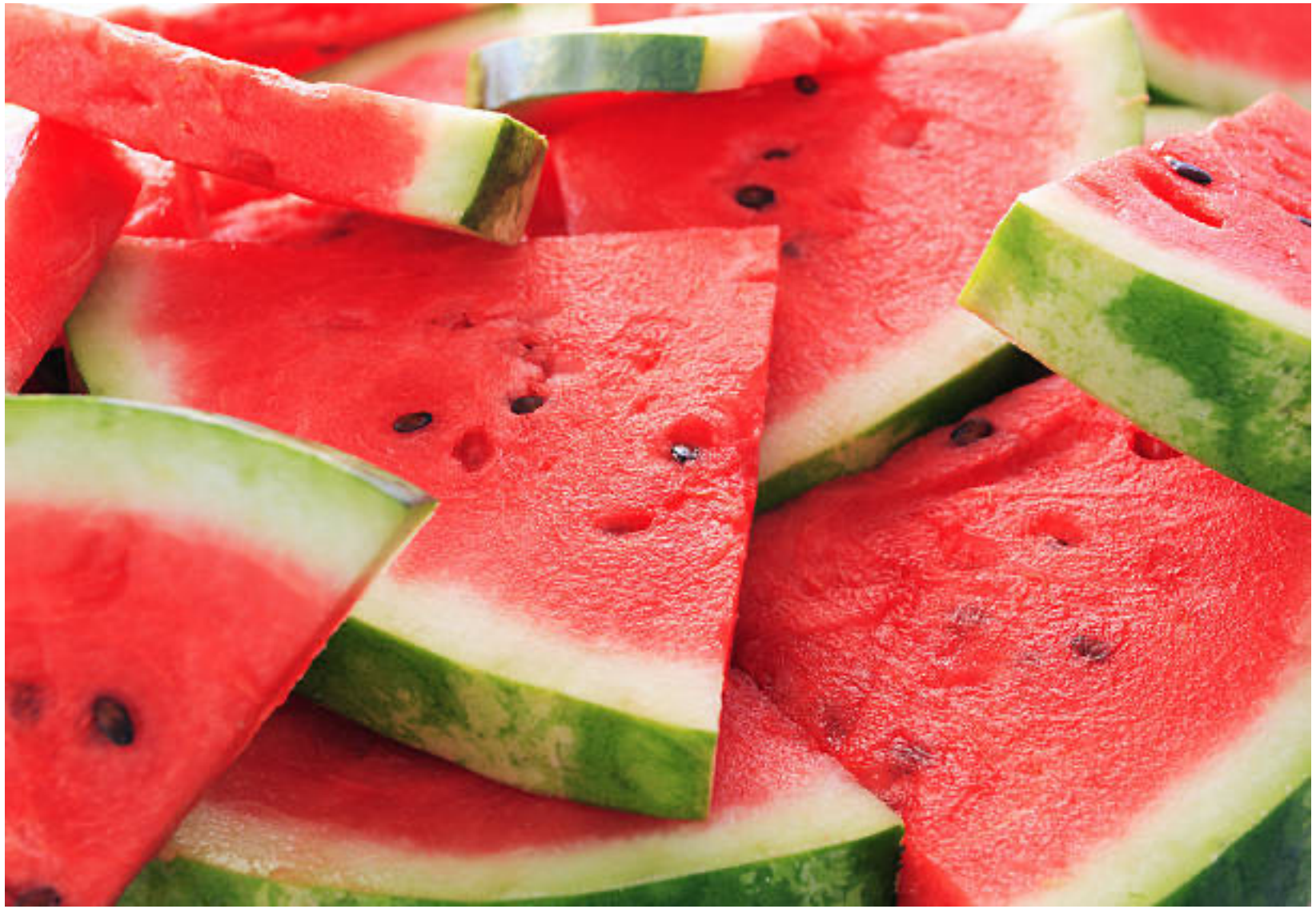

a personal assistant

https://github.com/openai/dalle-2-preview/blob/eeec5a1843b1d17cb9ed113117a2fcaa9206a564/system-card.md

Przyczyny

- historyczne zaszłości (np. prawo, które obowiązywało)
- stereotypy, nieuświadomione uprzedzenia
- nieróżnorodne, niezbalansowane dane
- interwencja człowieka w dane, wyniki modelu, jego parametry





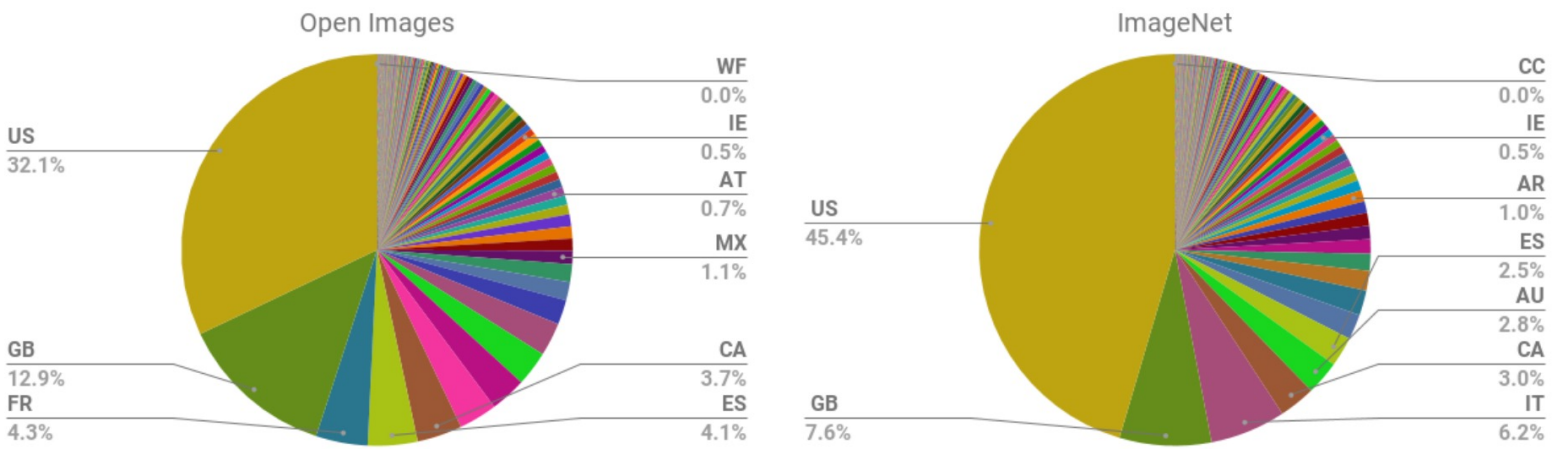


Figure 1: Fraction of Open Images and ImageNet images from each country. In both data sets, top represented locations include the US and Great Britain. Countries are represented by their two-letter ISO country codes. [1]

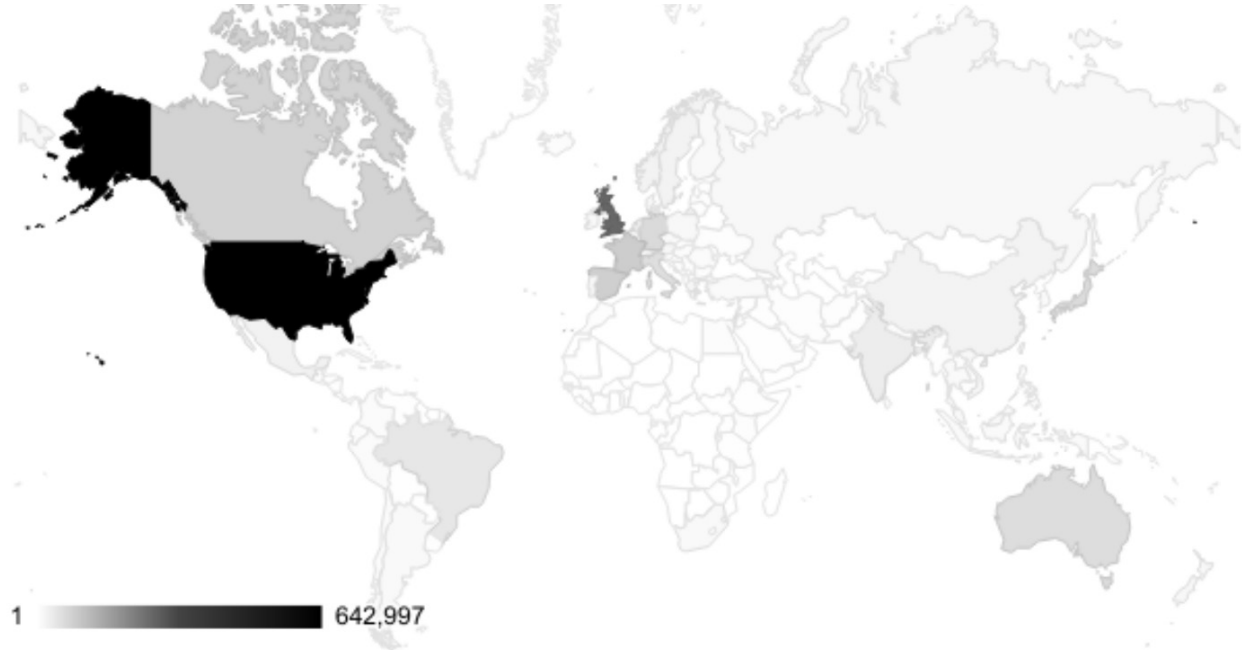


Figure 2: Distribution of the geographically identifiable images in the Open Images data set, by country. Almost a third of the data in our sample was US-based, and 60% of the data was from the six most represented countries across North America and Europe. [1]


ceremony, wedding, bride, man, groom, woman, dress
bride, ceremony, wedding, dress, woman
ceremony, bride, wedding, man, groom, woman, dress
person, people

VS

Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%

Classifier	Metric	All	F	M	Darker	Lighter	DF	DM	LF	LM
MSFT	PPV(%)	93.7	89.3	97.4	87.1	99.3	79.2	94.0	98.3	**100**
	Error Rate(%)	6.3	10.7	2.6	12.9	0.7	**20.8**	6.0	1.7	0.0
	TPR (%)	93.7	96.5	91.7	87.1	99.3	92.1	83.7	**100**	98.7
	FPR (%)	6.3	8.3	3.5	12.9	0.7	**16.3**	7.9	1.3	0.0
Face++	PPV(%)	90.0	78.7	99.3	83.5	95.3	65.5	**99.3**	94.0	99.2
	Error Rate(%)	10.0	21.3	0.7	16.5	4.7	**34.5**	0.7	6.0	0.8
	TPR (%)	90.0	98.9	85.1	83.5	95.3	98.8	76.6	**98.9**	92.9
	FPR (%)	10.0	14.9	1.1	16.5	4.7	**23.4**	1.2	7.1	1.1
IBM	PPV(%)	87.9	79.7	94.4	77.6	96.8	65.3	88.0	92.9	**99.7**
	Error Rate(%)	12.1	20.3	5.6	22.4	3.2	**34.7**	12.0	7.1	0.3
	TPR (%)	87.9	92.1	85.2	77.6	96.8	82.3	74.8	**99.6**	94.8
	FPR (%)	12.1	14.8	7.9	22.4	3.2	**25.2**	17.7	5.20	0.4

Dlaczego jest to ważne?

- pogłębianie nierówności społecznych
- niesprawiedliwe traktowanie osób, wpływ na życie człowieka
- straty wizerunkowe
- problemy prawne

Prawo

USA

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)

Unia Europejska

Dyrektywa Rady 2000/43/WE wprowadzająca w życie zasadę równego traktowania osób bez względu na **pochodzenie rasowe lub etniczne** , dyrektywa Rady 2000/78/WE ustanawiająca ogólne warunki ramowe równego traktowania w zakresie zatrudnienia i pracy – chroniąca przed dyskryminacją ze względu na „**religię lub przekonania, niepełnosprawność, wiek lub orientację seksualną**”, dyrektywy Rady 2004/113/WE oraz 2006/54/WE wprowadzające w życie zasadę **równego traktowania mężczyzn i kobiet** w zakresie dostępu do towarów i usług oraz w dziedzinie zatrudnienia i pracy

Polska

Art. 32 i 33 Konstytucji
1. Nikt nie może być dyskryminowany w życiu politycznym, społecznym lub gospodarczym z jakiejkolwiek przyczyny.

1. Kobieta i mężczyzna w Rzeczypospolitej Polskiej mają równe prawa w życiu rodzinnym, politycznym, społecznym i gospodarczym.

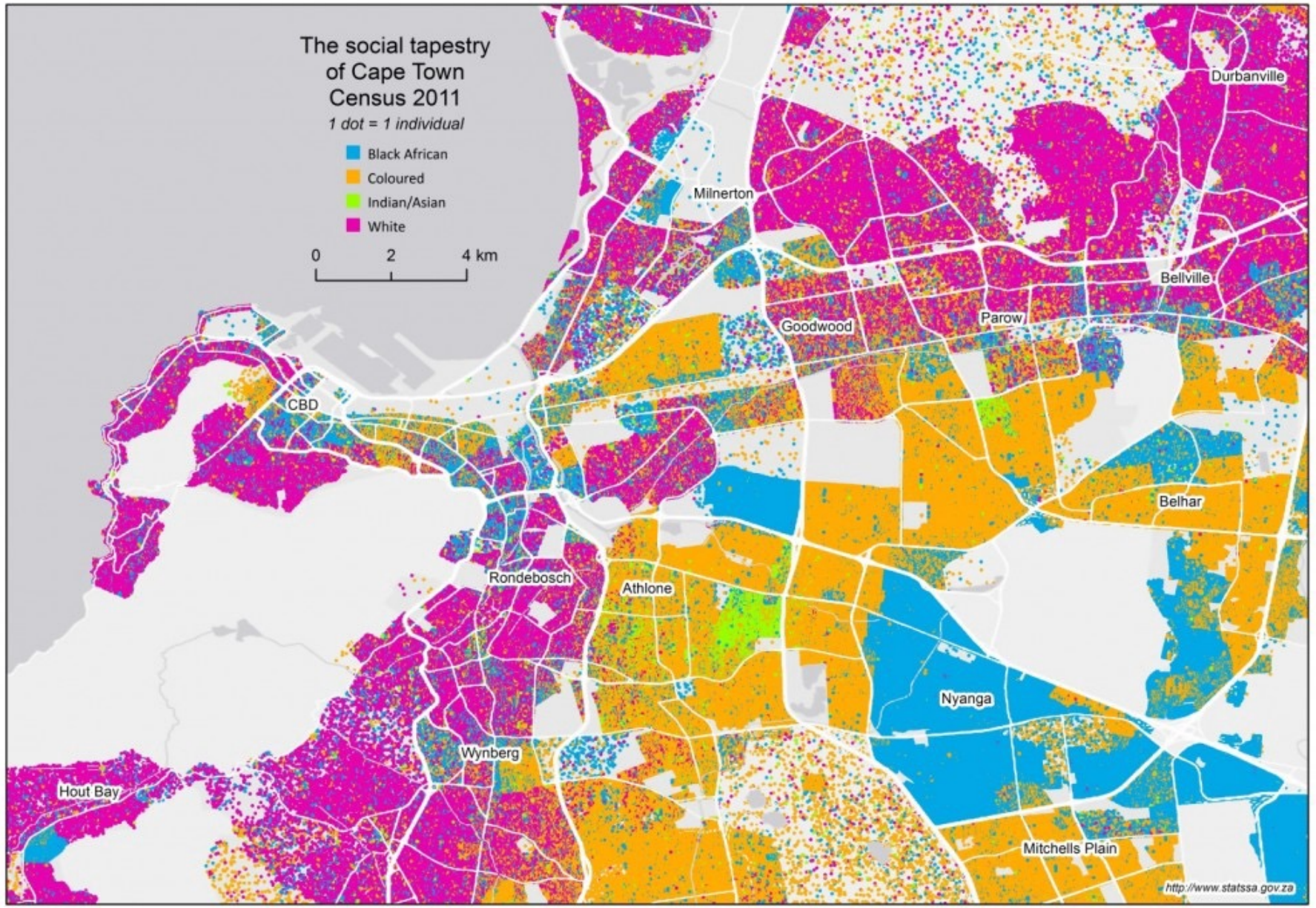
Dane wrażliwe/chronione

X - zbiór zmiennych wejściowych

$\hat{X} \subset X$ - podzbiór danych wrażliwych

dane wrażliwe - dane określające jakąś grupę, która nie powinna podlegać dyskryminacji, np. płeć, pochodzenie, wiek, niepełnosprawność

- zależą od kontekstu, rozwiązywanego problemu
- czasami mogą być podane nie wprost, np. płeć – poziom edukacji, wysokość zarobków, słowa kluczowe w CV, wydział


The social tapestry
of Cape Town
Census 2011
1 dot = 1 individual
Black African
Coloured
Indian/Asian
White
0
2
4 km
Durbanville
Milnerton
Bellville
Goodwood
Parow
CBD
Belhar
Rondebosch
Athlone
Nyanga
Wynberg
Hout Bay
Mitchells Plain
http://www.statssa.gov.za

Najpopularniejsze zbiory danych

Dataset Name	Reference	Size	Area
UCI adult dataset	[7]	48,842 income records	Social
German credit dataset	[47]	1,000 credit records	Financial
Pilot parliaments benchmark dataset	[24]	1,270 images	Facial images
WinoBias	[168]	3,160 sentences	Coreference resolution
Communities and crime dataset	[129]	1,994 crime records	Social
COMPAS Dataset	[89]	18,610 crime records	Social
Recidivism in juvenile justice dataset	[28]	4,753 crime records	Social
Diversity in faces dataset	[107]	1 million images	Facial images

Metryki

nie istnieje jedna uniwersalna, ogólnie przyjęta miara sprawiedliwości (fairness)

Główny podział

- **Indywidualne** – podobne jednostki powinny być traktowane podobnie (sprawiedliwie)
- **Grupowe** – różne grupy powinny być traktowane jednakowo (sprawiedliwie), np. kobiety, mężczyźni
- **Wielogrupowe** – jednostki z każdą kombinacją cech wrażliwych powinny być traktowane jednakowo (sprawiedliwie), np. młode kobiety, czarnoskórzy mężczyźni

Oznaczenia:

$y \in 0, 1$ - oczekiwana (prawdziwa) odpowiedź

$\hat{y} \in 0, 1$ - przewidywana wartość

$s = Pr(\hat{y}_i = 1)$ - prawdopodobieństwo przypisania wartości 1 obserwacji *i*

g_i, g_j - identyfikatory grup (podział na podstawie zmiennych wrażliwych)

Metryki

Statistical/Demographic Parity $Pr(\hat{y} = 1|g_i) = Pr(\hat{y} = 1|g_j)$

Disparate Impact $\frac{Pr(\hat{y} = 1|g_1)}{Pr(\hat{y} = 1|g_2)}$

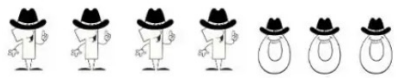
Equal Opportunity $Pr(\hat{y} = 1|y = 1\&g_i) = Pr(\hat{y} = 1|y = 1\&g_j)$

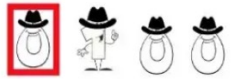
Equalized Odds $Pr(\hat{y} = 1|y = 1\&g_i) = Pr(\hat{y} = 1|y = 1\&g_j)$ & $Pr(\hat{y} = 1|y = 0\&g_i) = Pr(\hat{y} = 1|y = 0\&g_j)$

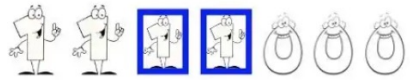
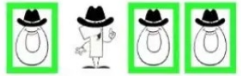
Overall accuracy equality $Pr(\hat{y} = 0|y = 0\&g_i) + Pr(\hat{y} = 1|y = 1\&g_i) = Pr(\hat{y} = 0|y = 0\&g_j) + Pr(\hat{y} = 1|y = 1\&g_j)$

Fairness Through Awareness – “podobne” dane wejściowe dawać powodować “podobny” wynik

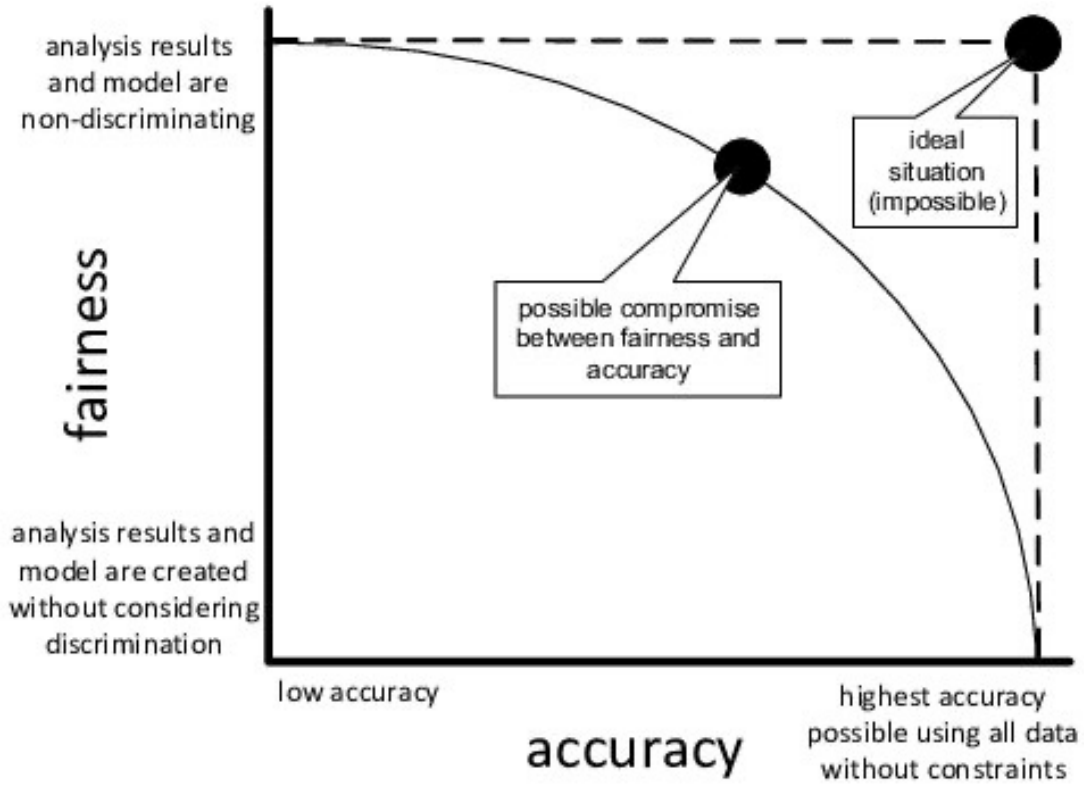
Fairness Through Unawareness – wynik powinien być niezależny od danych wrażliwych (ich usunięcie nie zmienia wyniku)

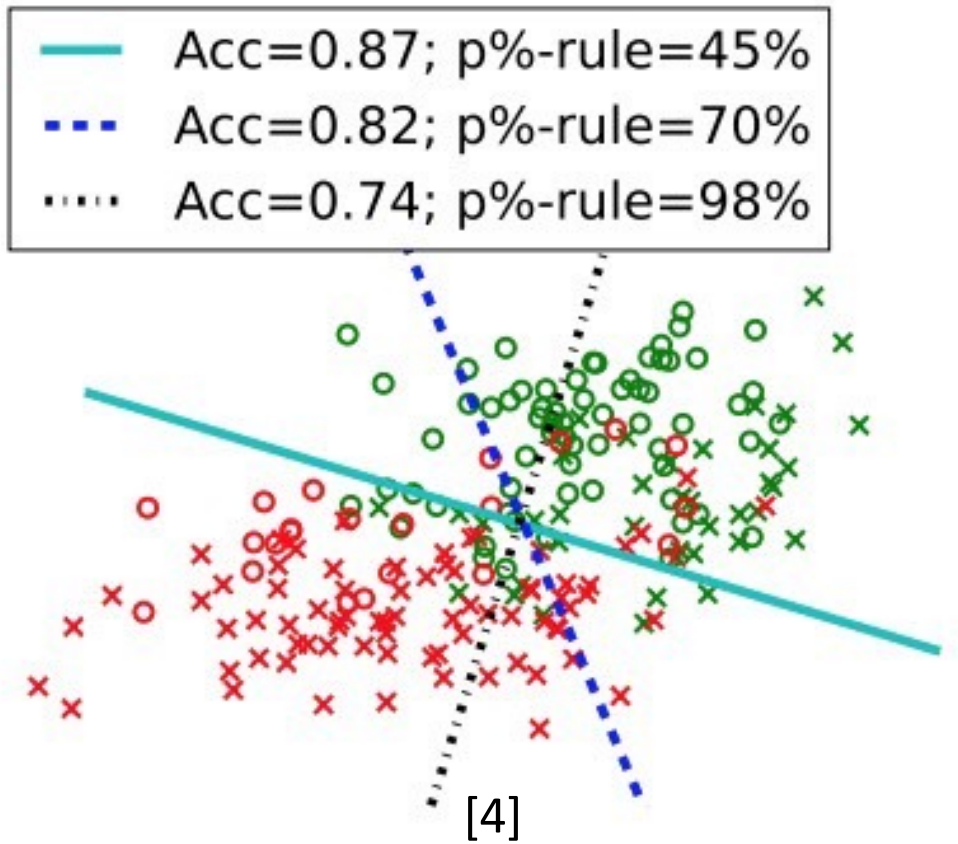
Group	a	b	
Outcome			Unequal base rates
Predictor			

Group	a	b	
Outcome			Unequal base rates
Predictor			

Group	a	b	
Outcome			Unequal base rates
Predictor			
NPV	2/5	1/3	

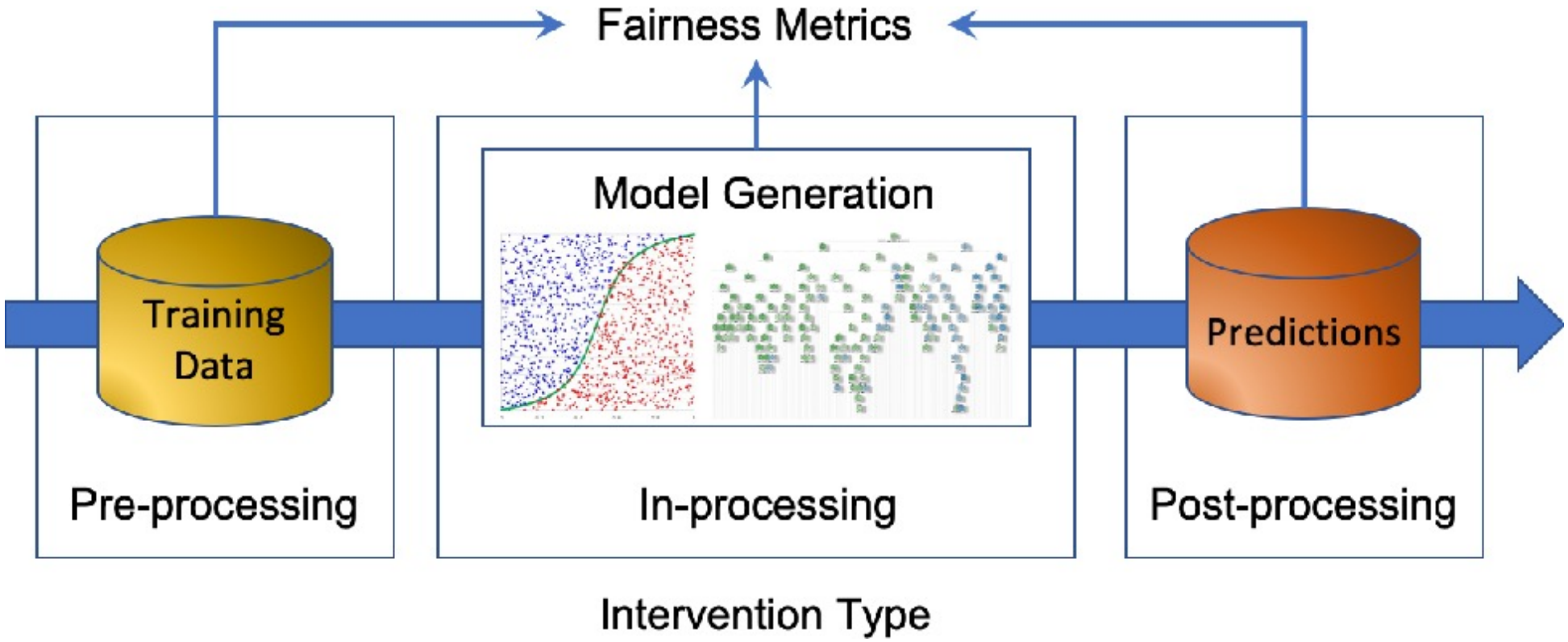
Accuracy vs fairness tradeoff





[4]

Podział technik



[2]

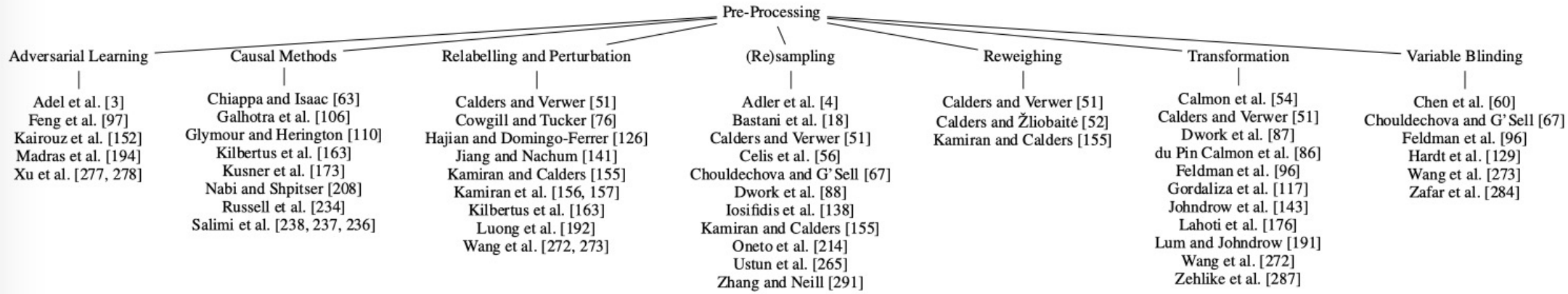


Figure 3: Pre-processing Methods

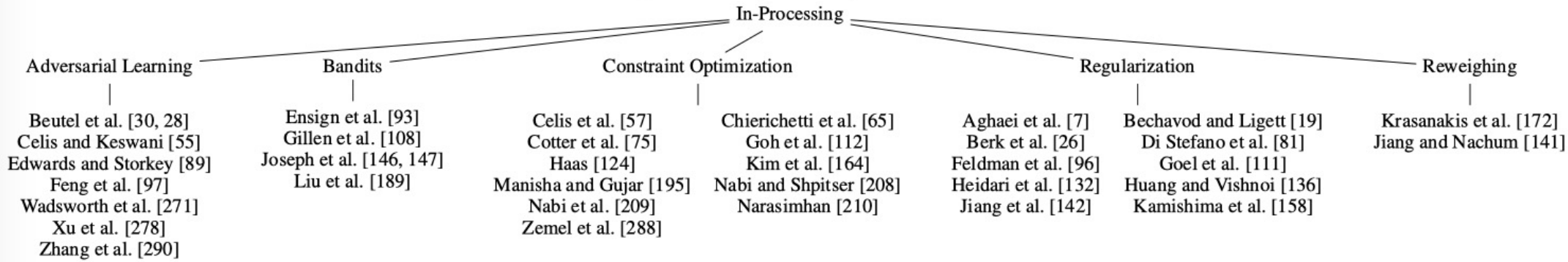


Figure 4: In-processing Methods

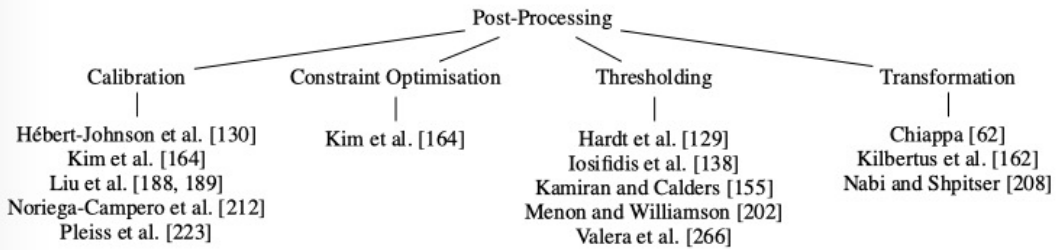


Figure 5: Post-processing methods

Pre-processing

- Ukrywanie (blinding) – z danych uczących usuwane są zmienne wrażliwe,

  wady: zmniejszenie accuracy, dane skorelowane

- Naprawa – ”ręczne” poprawienie danych na podstawie wiedzy eksperckiej,

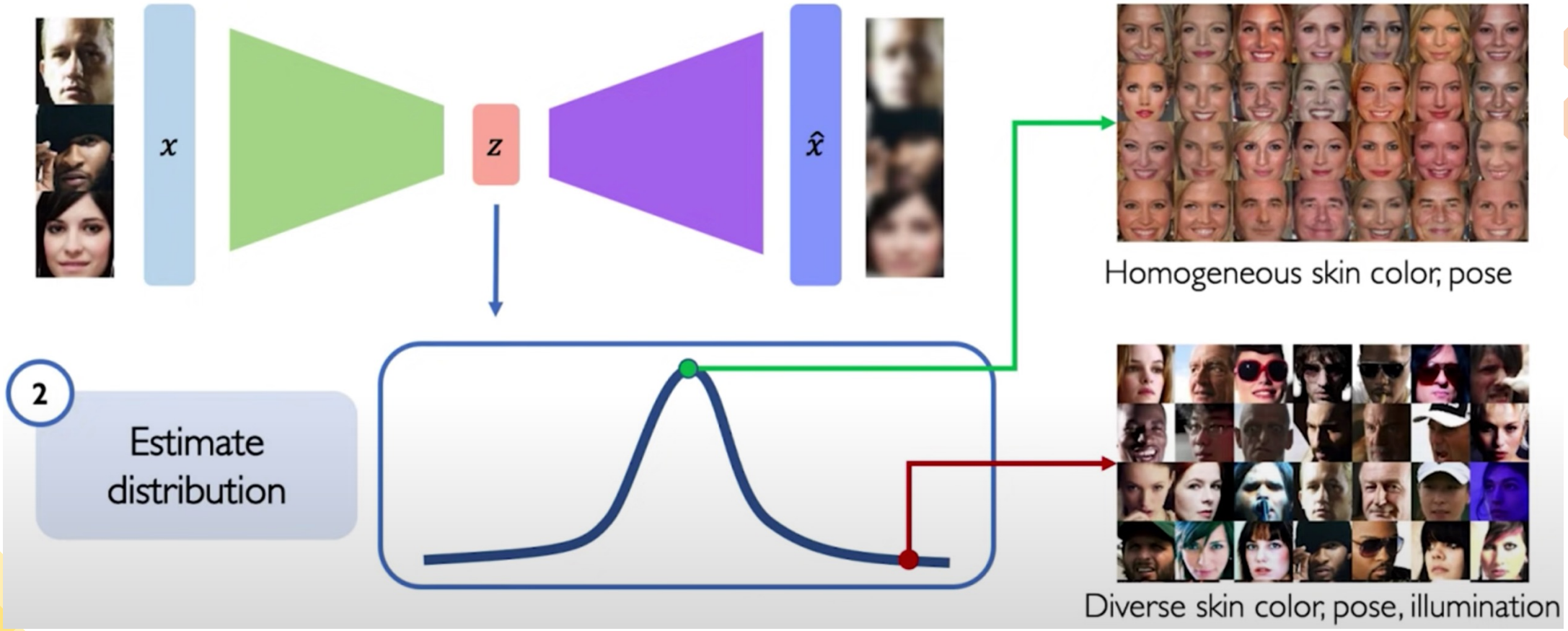
  wady: trudne w praktyce

- Rozszerzenie danych – dodanie nowych obserwacji, zdywersyfikowanie zbioru danych

  wady: trudne w praktyce

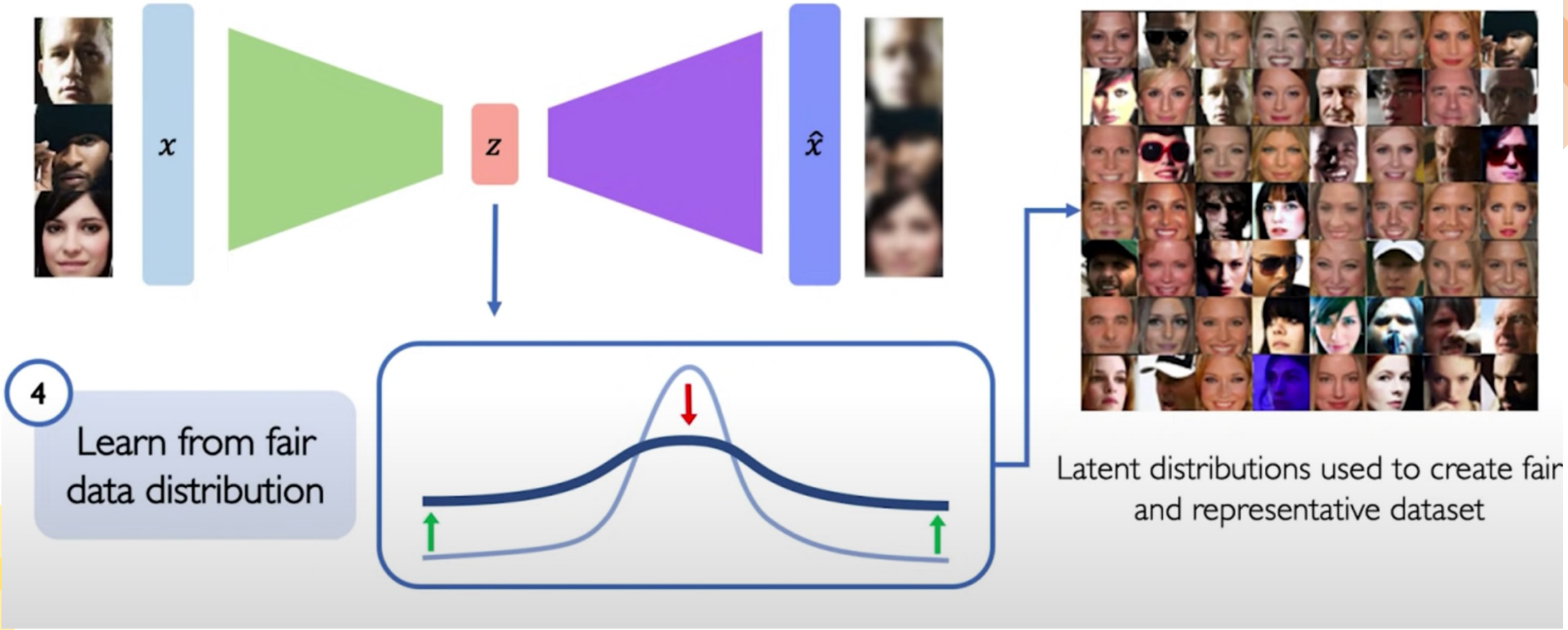
- Resampling – usunięcie, zduplikowanie danych, aby wyrównać liczbę danych w poszczególnych grupach
- Transformacja – jakaś transformacja danych wejściowych, która usuwa/zmniejsza czynnik dyskryminujący

Autoenkoder

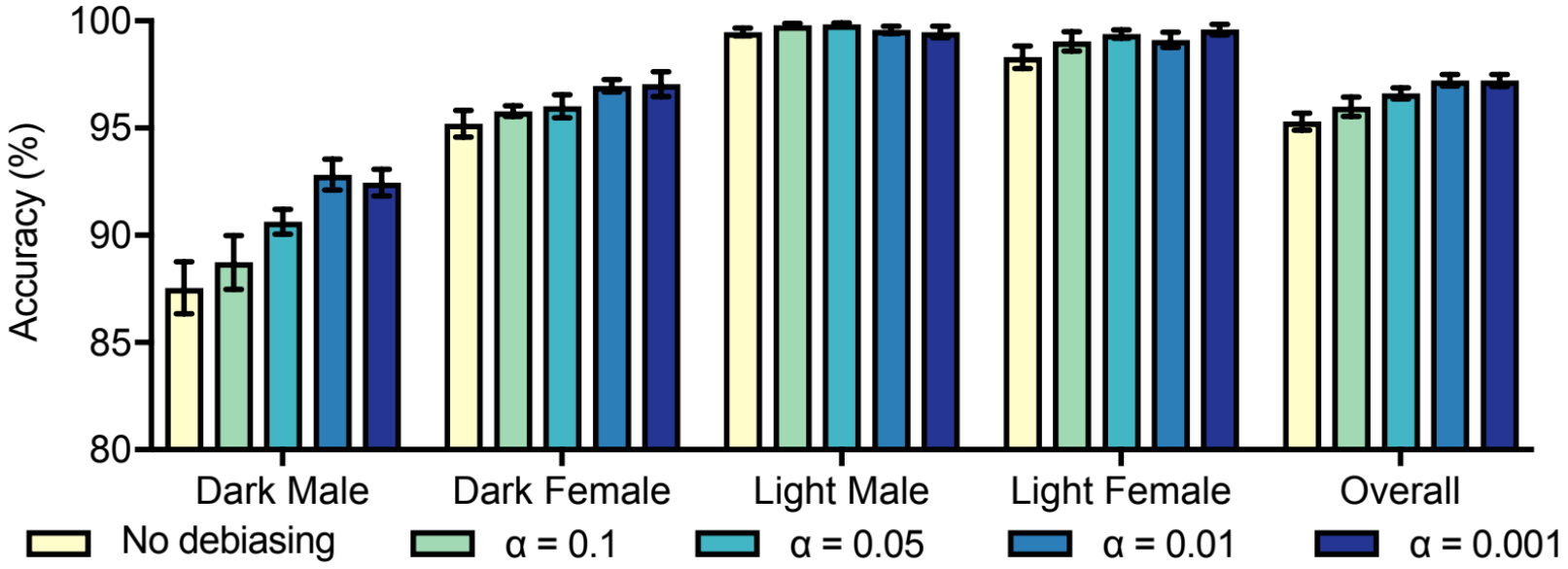


Amini, Alexander, et al. *Uncovering and mitigating algorithmic bias through learned latent structure.* AAAI 2019

Autoenkoder

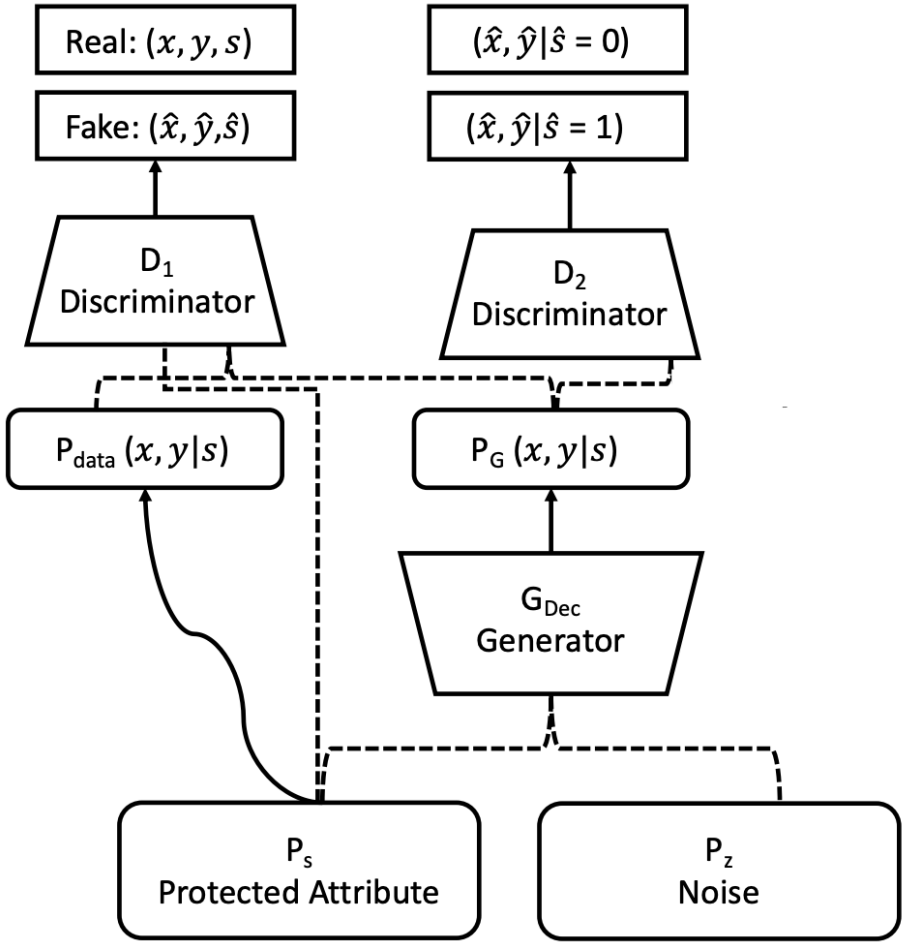


Autoenkoder - wyniki



FairGAN [6]

- model generuje nowe, sztuczne dane, które pozbawione są czynnika dyskryminującego
- Generator: generuje nowe dane
- Dyskryminator: ocenia czy dane są prawdziwe



In-processing

- Modyfikacja funkcji celu/błędu

Dodanie do funkcji błędu/funkcji celu czynnika uwzględniającego fairness i połączenie go (np. poprzez średnią ważoną) z accuracy/loss.

$$\mathcal{L}_t + \lambda \mathcal{L}_f$$

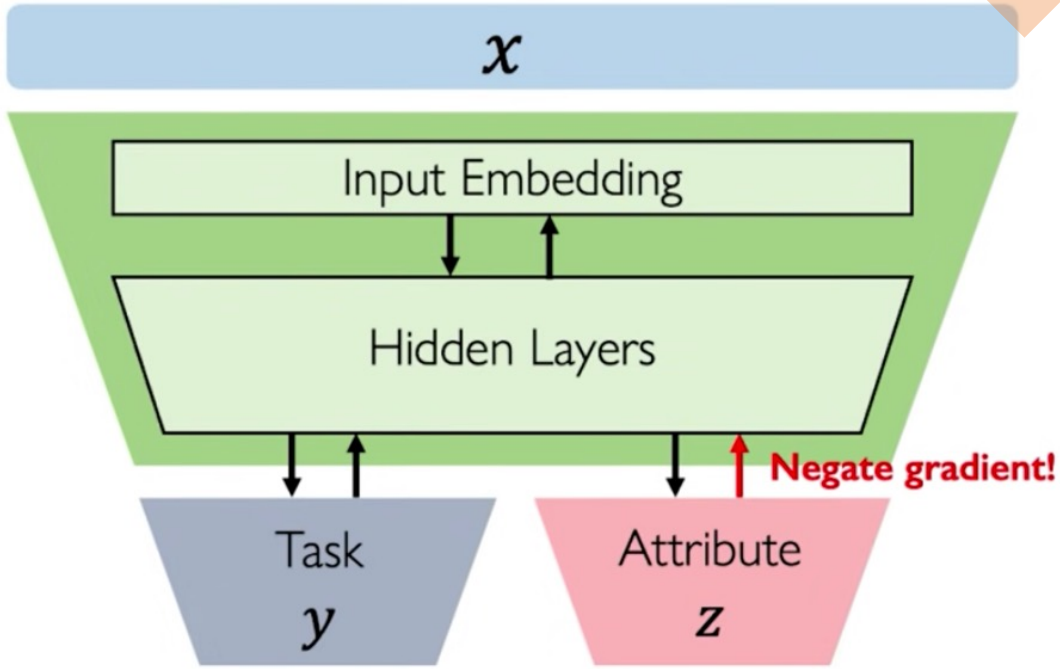
- Oddzielne modele dla każdej z grup + transfer learning

Multi-task learning

Nauka dwóch celów:

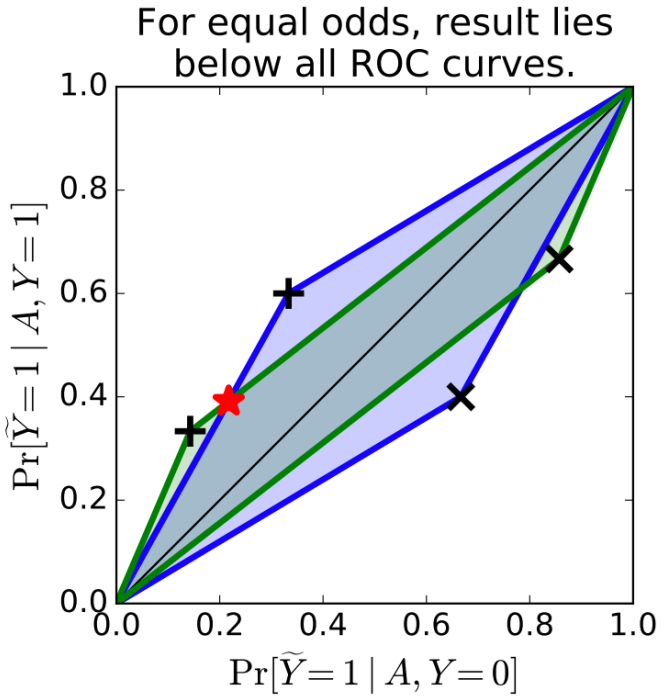
1. oryginalny problem klasyfikacji
2. przewidywanie danych wrażliwych

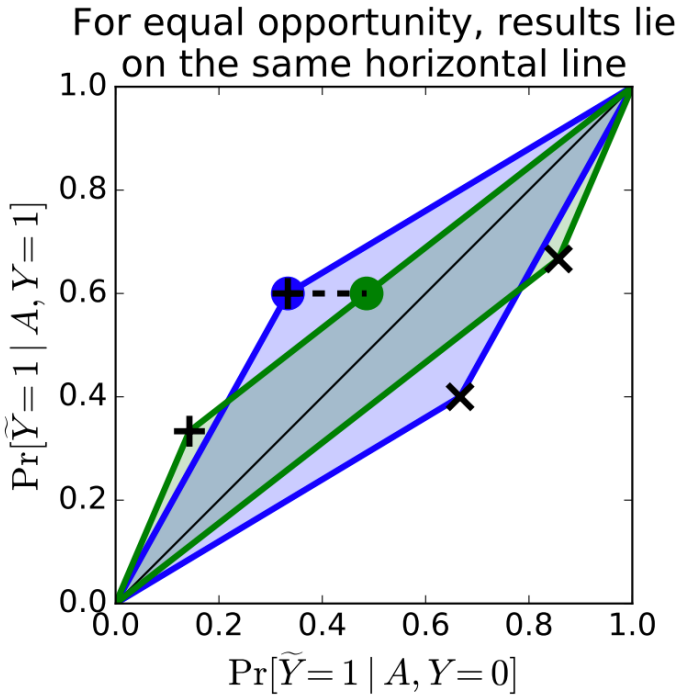
Wada: konieczność podania do modelu wprost informacji, które dane są wrażliwe



Post-processing

- „ręczna” zmiana wyników predykcji
- dostosowanie punktu odcięcia





[3]

Equalized Odds $Pr(\hat{y}=1|y=1\&g_i) = Pr(\hat{y}=1|y=1\&g_j) \;\; \& \;\; Pr(\hat{y}=1|y=0\&g_i) = Pr(\hat{y}=1|y=0\&g_j)$

Equal Opportunity $Pr(\hat{y}=1|y=1\&g_i) = Pr(\hat{y}=1|y=1\&g_j)$

Badania

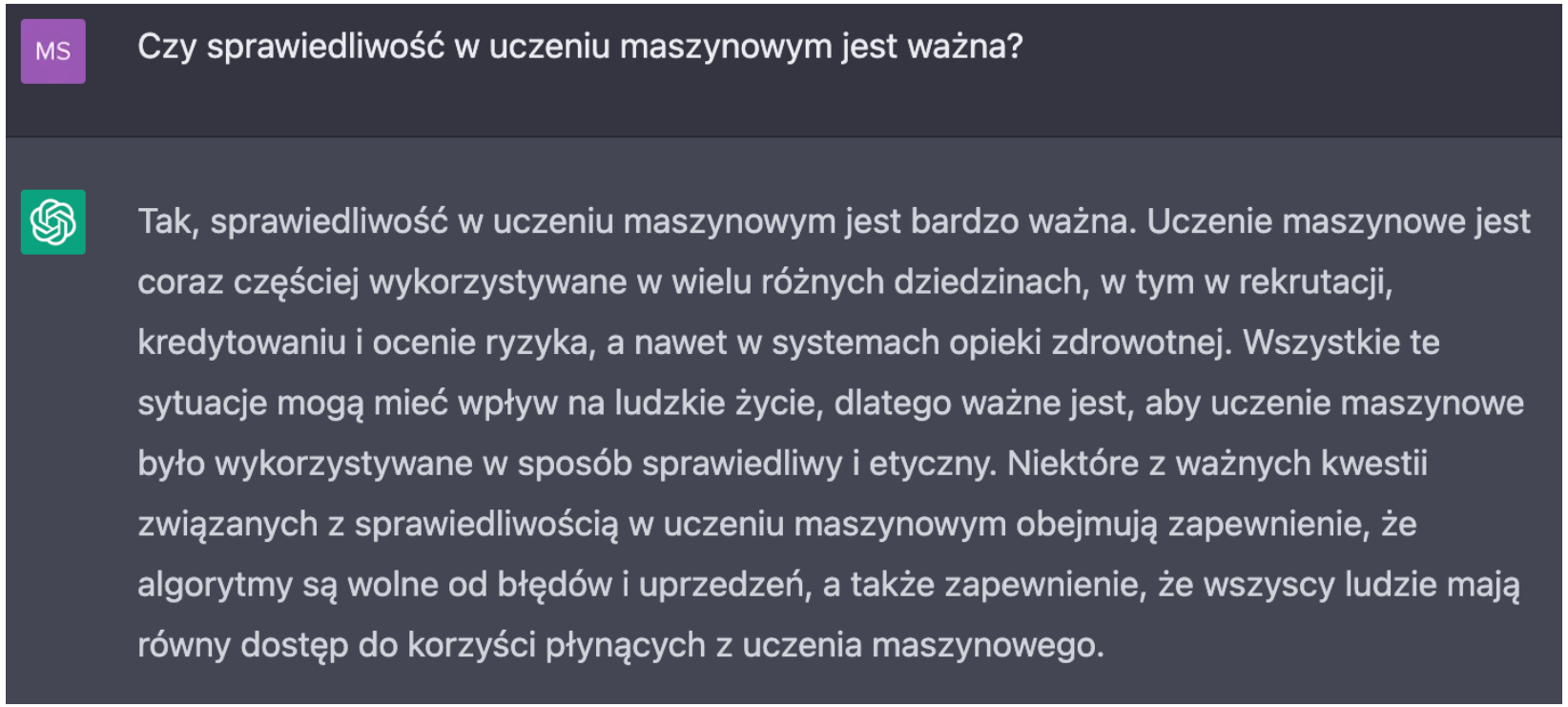
Learning Optimal and Fair Decision Trees for Non-Discriminative Decision-Making

Sina Aghaei, Mohammad Javad Azizi, Phebe Vayanos
CAIS Center for Artificial Intelligence in Society
University of Southern California, Los Angeles, CA 90007
{saghaei,azizim,phebe.vayanos}@usc.edu

- MILP budujący optymalne drzewa pod względem sprawiedliwości i skuteczności
- propozycja: alternatywny (szybszy) algorytm ewolucyjny budujący drzewa decyzyjne o zadanych kryteriach

Podsumowanie

- popularny temat w pracach związanych z uczeniem maszynowym
- w niektórych zastosowaniach bardzo istotne znaczenie
- brak jednoznacznej definicji i miary powoduje trudność w porównywaniu metod oraz wyników



https://chat.openai.com/chat

Bibliografia

[1] Mehrabi, Ninareh, et al. "A survey on bias and fairness in machine learning." *ACM Computing Surveys (CSUR)* 54.6 (2021): 1-35.

[2] Caton, Simon, and Christian Haas. "Fairness in machine learning: A survey." arXiv preprint arXiv:2010.04053 (2020).

[3] Hardt, Moritz, Eric Price, and Nati Srebro. "Equality of opportunity in supervised learning." Advances in neural information processing systems 29 (2016).

[4] Zafar, Muhammad Bilal, et al. "Fairness constraints: Mechanisms for fair classification." Artificial intelligence and statistics. PMLR, (2017).

[5] Amini, Alexander, et al. "Uncovering and mitigating algorithmic bias through learned latent structure." Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (2019).

[6] Xu, Depeng, et al. "Fairgan: Fairness-aware generative adversarial networks." 2018 IEEE International Conference on Big Data (Big Data). IEEE (2018).

[7] Kamiran, Faisal, and Toon Calders. "Data preprocessing techniques for classification without discrimination." Knowledge and information systems 33.1 (2012).

[8] Corbett-Davies, Sam, and Sharad Goel. "The measure and mismeasure of fairness: A critical review of fair machine learning." arXiv preprint arXiv:1808.00023 (2018).

[9] Chouldechova, Alexandra, and Aaron Roth. "The frontiers of fairness in machine learning." arXiv preprint arXiv:1810.08810 (2018).