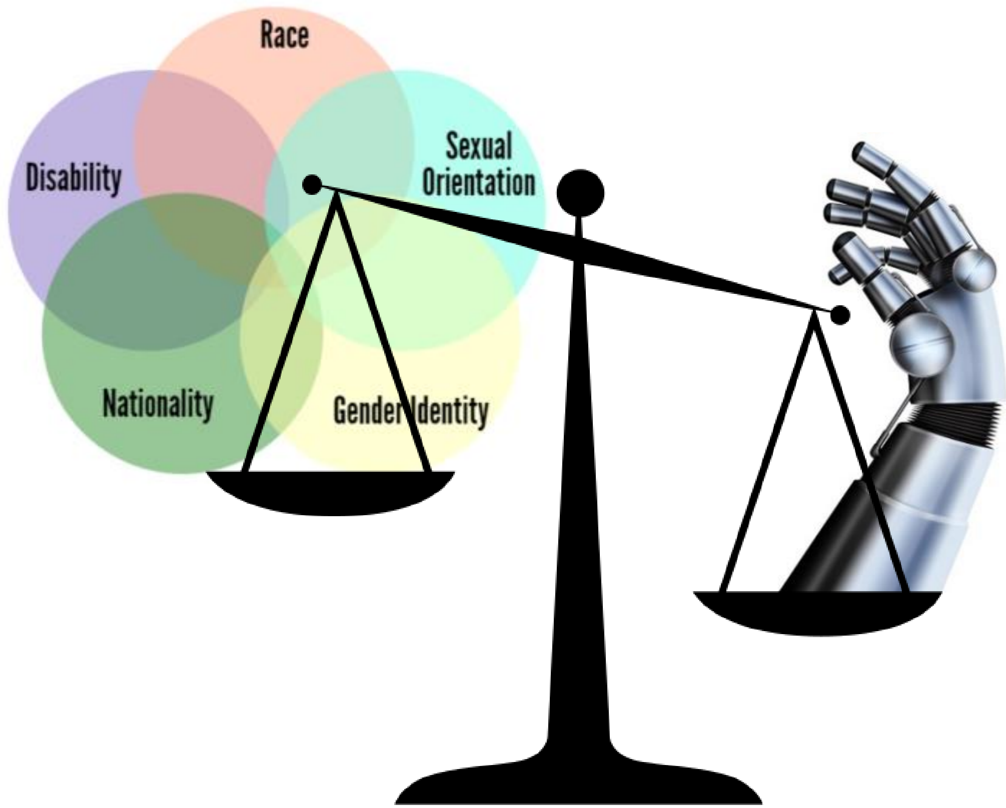
An Intersectional Definition of Fairness

James R. Foulds, Rashidul Islam, Kamrun Naher Keya, Shimei Pan
Department of Information Systems
University of Maryland, Baltimore County

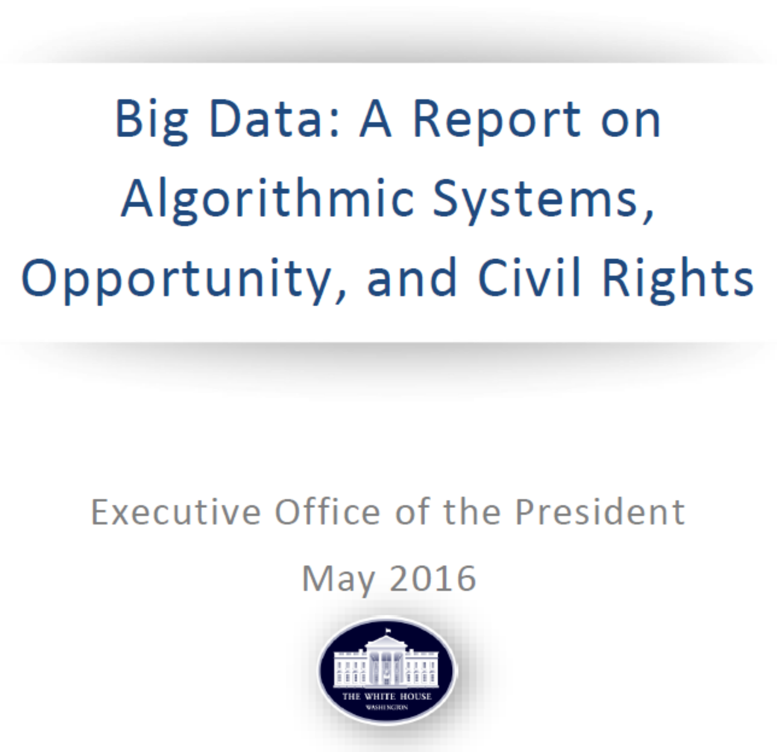
ICDE 2020 Conference, Remote Presentation, April 2020

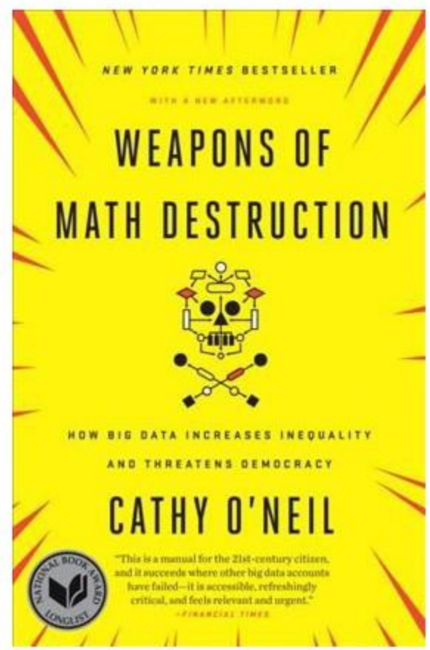


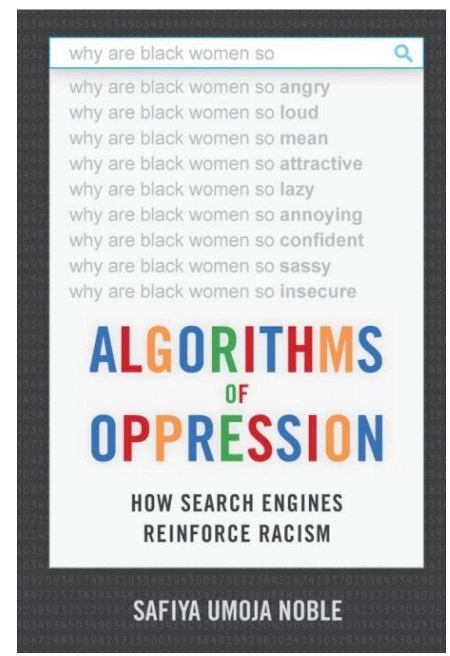
Work sponsored in part by NIST and NSF

Fairness in Machine Learning

- There is growing awareness that **biases inherent in data** can lead the behavior of machine learning algorithms to **discriminate against certain populations**.

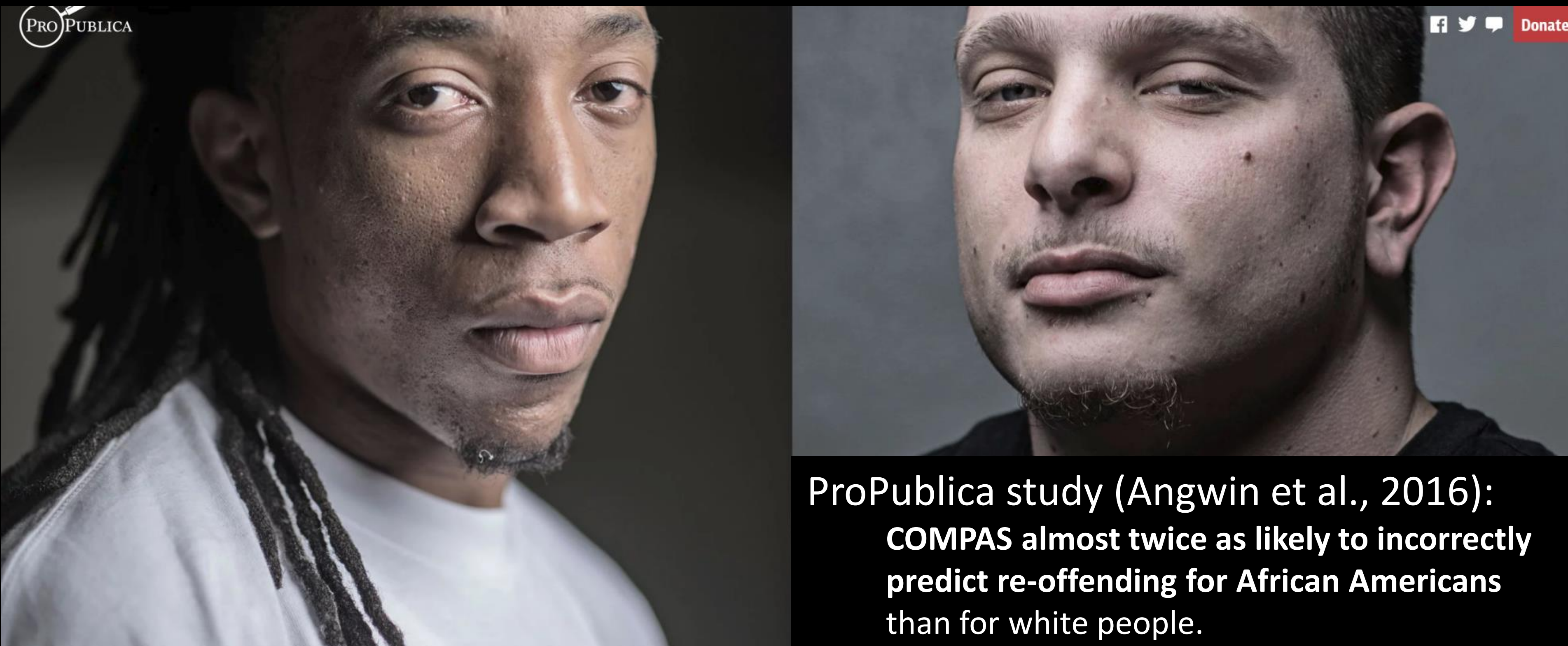






Bias in Criminal Justice Risk Assessments

- Correctional Offender Management Profiling for Alternative Sanctions (**COMPAS**), algorithm for risk assessment (Northpointe company)
  - Used for bail and sentencing decisions across the U.S.



ProPublica study (Angwin et al., 2016):

COMPAS almost twice as likely to incorrectly predict re-offending for African Americans than for white people.

Bernard Parker, left, was rated high risk; Dylan Fugett was rated low risk. (Josh Ritchie for 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

Amazon scraps secret AI recruiting tool that showed bias against women

Jeffrey Dastin

SAN FRANCISCO (Reuters) - Amazon.com Inc's (AMZN.O) machine-learning specialists uncovered a big problem: their new recruiting engine did not like women.

The team had been building computer programs since 2014 to review job applicants' resumes with the aim of mechanizing the search for top talent, five people familiar with the effort told Reuters.

Typical Philosophical Assumption for AI Fairness

- **Infra-marginality:**
  attributes used by algorithm may have **different distributions**, depending on the **protected attributes**.

C. Simoiu, S. Corbett-Davies, S. Goel, et al. The problem of infra-marginality in outcome tests for discrimination. The Annals of Applied Statistics, 11(3):1193–1216, 2017.

Typical Philosophical Assumption for AI Fairness

- **Infra-marginality:**

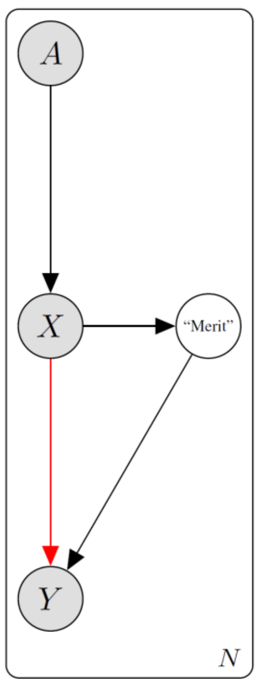
  attributes used by algorithm may have **different distributions**, depending on the **protected attributes**.

Causal assumption

Protected attributes

Features

Outcomes



C. Simoiu, S. Corbett-Davies, S. Goel, et al. The problem of infra-marginality in outcome tests for discrimination. The Annals of Applied Statistics, 11(3):1193–1216, 2017.

Typical Philosophical Assumption for AI Fairness

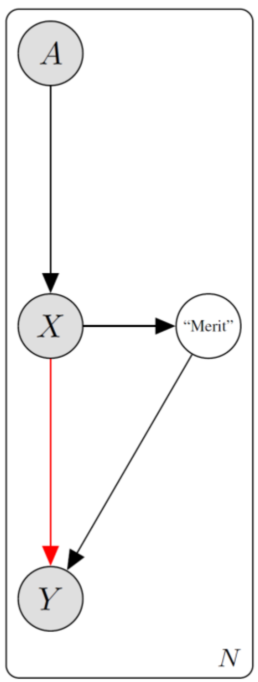
- **Infra-marginality:**
  attributes used by algorithm may have **different distributions**, depending on the **protected attributes**.

Causal assumption

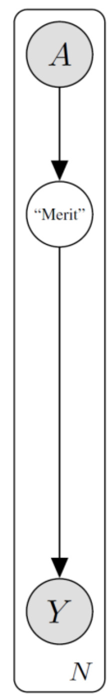
Protected attributes

Features

Outcomes



Ideal World



C. Simoiu, S. Corbett-Davies, S. Goel, et al. The problem of infra-marginality in outcome tests for discrimination. The Annals of Applied Statistics, 11(3):1193–1216, 2017.

Typical Philosophical Assumption for AI Fairness

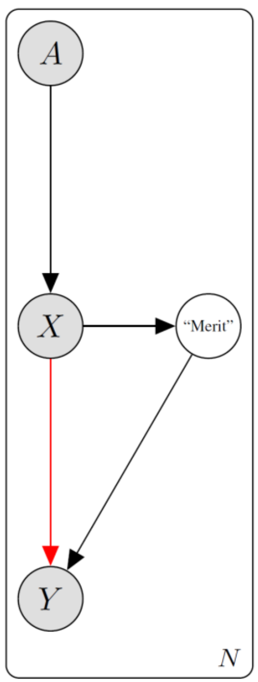
- **Infra-marginality:**
  attributes used by algorithm may have **different distributions**, depending on the **protected attributes**.

Causal assumption

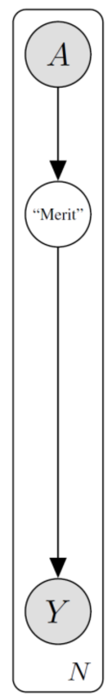
Protected attributes

Features

Outcomes



Ideal World

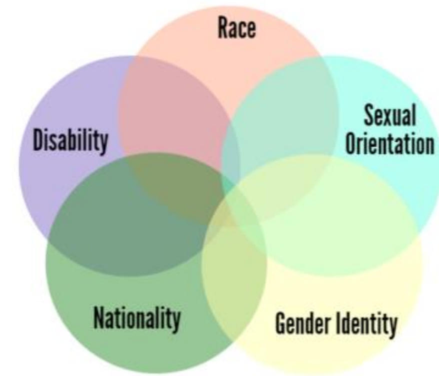


- Algorithm should behave differently for each group
- Individuals should get outcomes according to their "merit" or "risk"
- Algorithm is only biased if more inequitable than the data suggest

C. Simoiu, S. Corbett-Davies, S. Goel, et al. The problem of infra-marginality in outcome tests for discrimination. The Annals of Applied Statistics, 11(3):1193–1216, 2017.

Fairness and Intersectionality

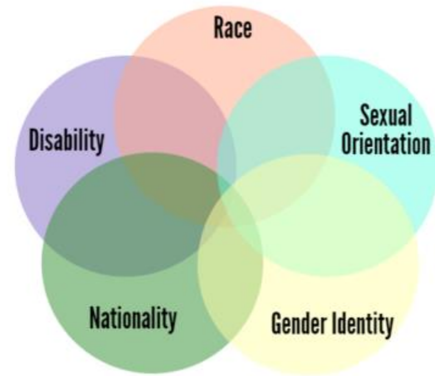
- **Intersectionality:**
  systems of oppression built into society lead to **systematic disadvantages** along **intersecting dimensions**
  - gender, race, nationality, sexual orientation, disability status, socioeconomic class, …



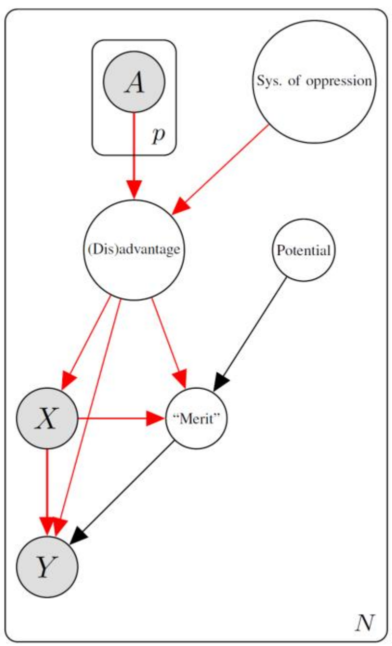
K. Crenshaw. Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. U. Chi. Legal F., pages 139–167, 1989.

Fairness and Intersectionality

- **Intersectionality:**
  systems of oppression built into society lead to **systematic disadvantages** along **intersecting dimensions**
  - gender, race, nationality, sexual orientation, disability status, socioeconomic class, …



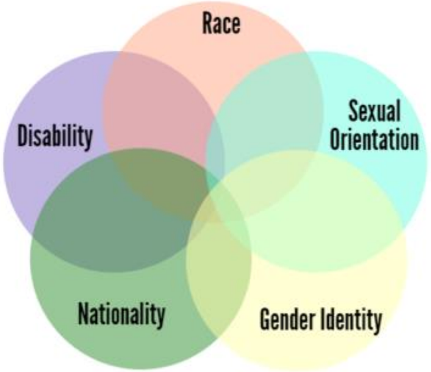
Causal assumption



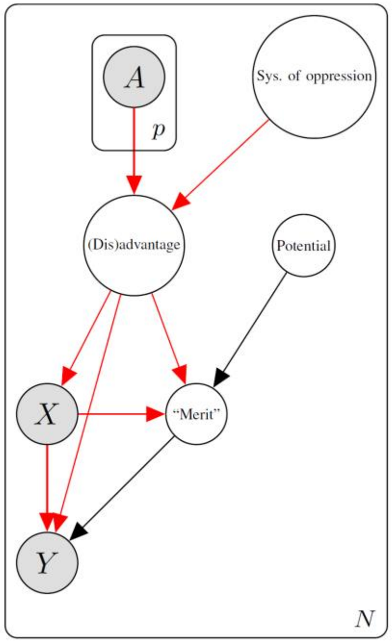
K. Crenshaw. Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. U. Chi. Legal F., pages 139–167, 1989.

Fairness and Intersectionality

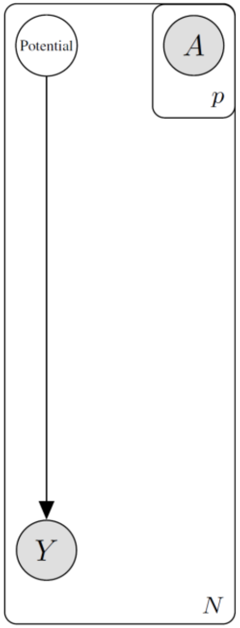
- **Intersectionality:**
  systems of oppression built into society lead to **systematic disadvantages** along **intersecting dimensions**
  - gender, race, nationality, sexual orientation, disability status, socioeconomic class, …



Causal assumption



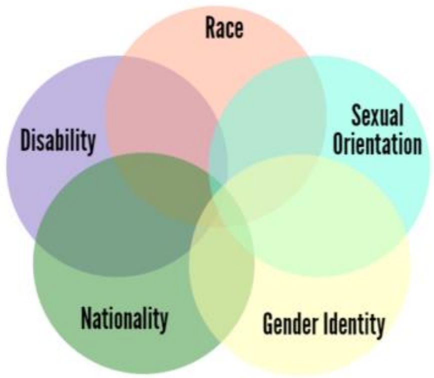
Ideal World



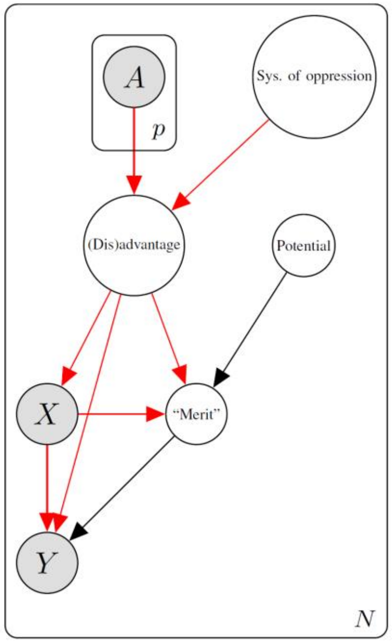
K. Crenshaw. Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. U. Chi. Legal F., pages 139–167, 1989.

Fairness and Intersectionality

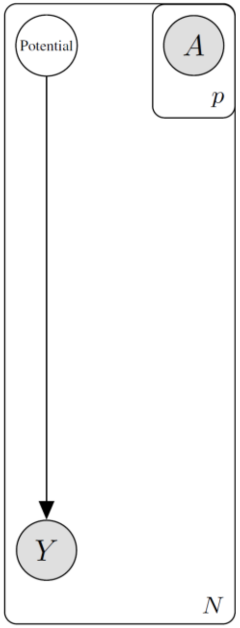
- **Intersectionality:**
  systems of oppression built into society lead to **systematic disadvantages** along **intersecting dimensions**
  - gender, race, nationality, sexual orientation, disability status, socioeconomic class, …



Causal assumption



Ideal World



- Rectify harmful effects of oppression
- Algorithm should **not** generally behave differently for each group

  (unless justified, e.g. confounder variables)

K. Crenshaw. Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. U. Chi. Legal F., pages 139–167, 1989.

Fairness and Intersectionality

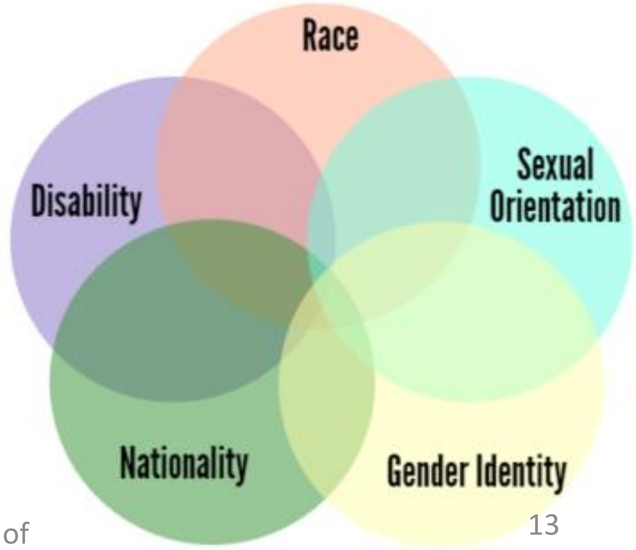
We argue that an **intersectional definition of fairness** should satisfy:

- **Multiple protected attributes** should be considered

- **All** of the **intersecting values** of the protected attributes, e.g. *black women*, should be protected
  - We should still ensure that the individual protected attributes are protected overall, e.g. *women* are protected

- Systematic differences, due to structural oppression, are **rectified, rather than codified**.

- Protects **minority groups**



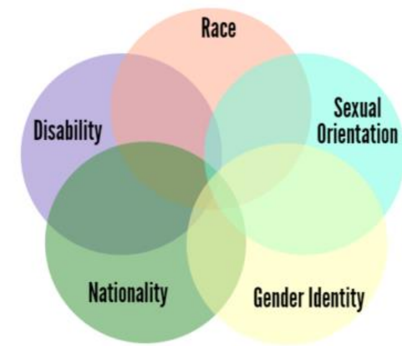
K. Crenshaw. Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. U. Chi. Legal F., pages 139–167, 1989.

Our contributions

- We address fairness in machine learning from an **intersectional perspective**
  - Fairness definitions that **respect intersectionality**
    - Also provide a more politically conservative option
  - Theoretical results on our definitions' properties
  - A learning algorithm to enforce our definitions
  - Experimental validation

Fairness and Intersectionality

- *Subgroup fairness* (Kearns et al., 2018)
  - Aims to prevent "*fairness gerrymandering*" a.k.a. *subset targeting*, by protecting specified subgroups



Definition 2.1 (Statistical Parity (SP) Subgroup Fairness). *Fix any classifier D, distribution $\mathcal{P}$, collection of group indicators $\mathcal{G}$, and parameter $\gamma \in [0,1]$. For each $g \in \mathcal{G}$, define*

$$\alpha_{SP}(g,\mathcal{P}) = \Pr_{\mathcal{P}}[g(x)=1] \quad \textit{and,} \quad \beta_{SP}(g,D,\mathcal{P}) = |\mathrm{SP}(D) - \mathrm{SP}(D,g)|,$$

where $\mathrm{SP}(D) = \Pr_{\mathcal{P},D}[D(X)=1]$ *and* $\mathrm{SP}(D,g) = \Pr_{\mathcal{P},D}[D(X)=1|g(x)=1]$ *denote the overall acceptance rate of D and the acceptance rate of D on group g respectively. We say that D satisfies* γ-statistical parity (SP) Fairness *with respect to $\mathcal{P}$ and $\mathcal{G}$ if for every $g \in \mathcal{G}$*

$$\alpha_{SP}(g,\mathcal{P})\,\beta_{SP}(g,D,\mathcal{P}) \leq \gamma.$$

  - punts on small groups (in order to prove generalization)

See also *multicalibration*, a similar definition but for calibration of probabilities (Hebert-Johnson et al., 2018)

M. Kearns, S. Neel, A. Roth, and Z. S. Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In J. Dy and A. Krause, editors, Proceedings of the 35th International Conference on Machine Learning (ICML)

Differential Fairness (DF)

We propose a fairness definition with the following properties:

- **Measures** the **fairness cost** of **algorithms** and **data**
  - Can measure difference in fairness between algorithms and data: **bias amplification**
- **Privacy** and **economic guarantees**
  - Privacy perspective provides an **interpretation** of definition, based on **differential privacy**
- Implements **intersectionality**: e.g. fairness for (gender, race) provably ensures fairness for gender and for race separately

Essentially, differential fairness extends the 80% rule to multiple protected attributes and outcomes, and provides a privacy interpretation

Fairness and the Law: Adverse Impact Analysis

- Title VII, other anti-discrimination laws prohibit employers from intentional discrimination against employees with respect to protected characteristics
  - gender, race, color, national origin, religion

- Uniform Guidelines for Employee Selection Procedures (Equal Employment Opportunity Commission)

Fairness and the Law: Adverse Impact Analysis

Uniform guidelines: the "four-fifths rule" (a.k.a. 80% rule)

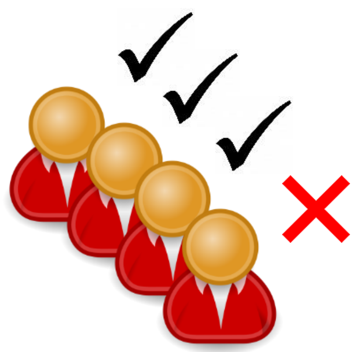
*"**A selection rate for any race, sex, or ethnic group which is less than four-fifths (4/5) (or eighty percent) of the rate for the group with the highest rate** will generally be regarded by the Federal enforcement agencies as evidence of adverse impact,*

while a greater than four-fifths rate will generally not be regarded by Federal enforcement agencies as evidence of adverse impact."

-Code of Federal Regulations 29 Part 1607 (1978)

Fairness and the Law: Adverse Impact Analysis

$$Pr(\text{hire}|\text{group A}) < 0.8 \times Pr(\text{hire}|\text{group B}) \ ?$$



If so, there is evidence of adverse impact

Differential Privacy vs the 80% Rule

Definition: $\mathcal{M}(\mathbf{X})$ is ϵ-differentially private if

$$e^{-\epsilon} \leq \frac{Pr(\mathcal{M}(\mathbf{X}) \in \mathcal{S})}{Pr(\mathcal{M}(\mathbf{X}') \in \mathcal{S})} \leq e^{\epsilon}$$

for all outcomes $\mathcal{S}$, and pairs of databases $\mathbf{X}$, $\mathbf{X}'$ differing in a single element.

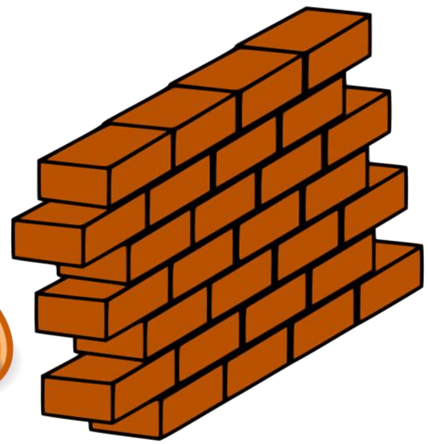
Follows from taking the reciprocal. We want ratios close to 1

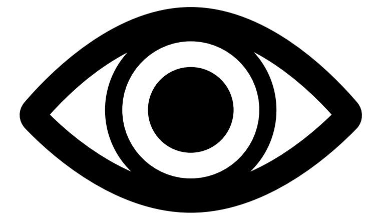
- 80% rule: Evidence of unfairness if:

$$\frac{Pr(\text{hire}|\text{group A})}{Pr(\text{hire}|\text{group B})} < 0.8$$

The ratio determines the degree of disparate impact between groups.
Like differential privacy, we want to bound a ratio to be somewhere near 1

Scenario for Differential Fairness





Vendor
(may be untrusted)

Individuals' data

$$x_i \sim \theta \ ,$$

$$\theta \in \Theta$$

Fair algorithm

$$M(x)$$

Outcomes y_i

Multiple protected attributes

$$a_i \in A$$

Randomness in data and mechanism

$$P_{M,\theta}(M(x) = y|\theta)$$

Our Proposed Fairness Definition: **Differential Fairness (DF)**

Protected attributes, e.g. gender, race

Classifier (e.g.)

Measures **fairness cost**

A mechanism $M(X)$ is ϵ-*differentially fair* in a framework (A, Θ) if for all $\theta \in \Theta$ with $X \sim \theta$, and $y \in \text{Range}(M)$,

$$e^{-\epsilon} \leq \frac{P_{M,\theta}(M(x) = y|s_i, \theta)}{P_{M,\theta}(M(x) = y|s_j, \theta)} \leq e^{\epsilon} \ ,$$

for all $(s_i, s_j) \in A \times A$ where $P(s_i|\theta) > 0$, $P(s_j|\theta) > 0$.

Distributions which could have generated the data

Probabilities w.r.t. **data and mechanism**

Key idea: ratios of probabilities of outcomes bounded for any pair of values of protected attributes

Interpreting ϵ: Bayesian Privacy

- Untrusted vendor/adversary can learn very little about the protected attributes of the instance}, relative to their prior beliefs, assuming their prior beliefs are in Θ:

$$e^{-\epsilon}\frac{P(s_i|\theta)}{P(s_j|\theta)} \leq \frac{P(s_i|M(x)=y,\theta)}{P(s_j|M(x)=y,\theta)} \leq e^{\epsilon}\frac{P(s_i|\theta)}{P(s_j|\theta)}$$

- E.g., if a loan was given to an individual, the vendor or adversary's Bayesian posterior beliefs about their race and gender will not be substantially changed

- This can **prevent subsequent discrimination**,
e.g. in retaliation for a correction against bias.

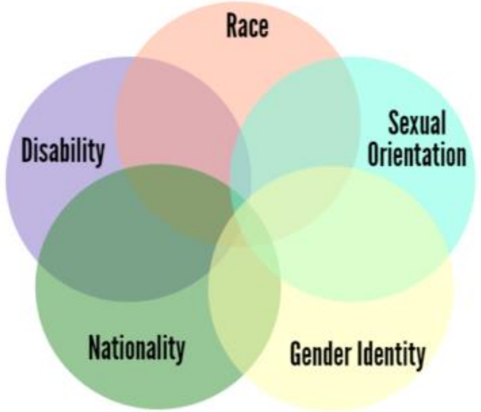
D. Kifer and A. Machanavajjhala. Pufferfish: A framework for mathematical privacy definitions. ACM Trans. on Database Systems, 39(1):3, 2014.

Intersectionality Property of DF: Fairness with Multiple Protected Attributes

- **Intersectionality theory**: gender is not the only dimension upon which power structures in society impose systems of oppression and marginalization.
  - **The intersection of a number of aspects must be considered**, including race, sexual orientation, class, and disability status

Theorem: Let M be an ϵ-*differentially fair* mechanism in (A, Θ), $A = S_1 \times S_2 \times \ldots \times S_p$, and let $D = S_a \times \ldots \times S_k$ be the Cartesian product of a nonempty proper subset of the variables included in A. Then M is ϵ-*differentially fair* in (D, Θ).

E.g., if M is differentially fair in (race, gender, nationality), it is differentially fair to a similar degree in gender alone



K. Crenshaw. Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. U. Chi. Legal F., pages 139–167, 1989.

Other Theoretical Properties

- Generalization Guarantee

THEOREM *Fix a class of functions $\mathcal{H}$, which without loss of generality aim to discriminate the outcome $y = 1$ from any other value, denoted here as $y = 0$. For any conditional distribution $P(y, \mathbf{x}|\mathbf{s})$ given a group $\mathbf{s}$, let $S \sim P^m$ be a dataset consisting of m examples $(\mathbf{x}_i, y_i)$ sampled i.i.d. from $P(y, \mathbf{x}|\mathbf{s})$. Then for any $0 < \delta < 1$, with probability $1 - \delta$, for every $h \in \mathcal{H}$, we have:*

$$|P(y = 1|\mathbf{s}, h) - P_S(y = 1|\mathbf{s}, h)| \leq \tilde{O}\left(\sqrt{\frac{VCDIM(\mathcal{H}) \log m + \log(1/\delta)}{m}}\right).$$

Enough data per intersection

Empirical estimates will converge on true values

Note: SF only needs enough data *overall*

- Economic guarantee

An ϵ-differentially fair mechanism admits a disparity in expected utility of as much as a factor of $\exp(\epsilon) \approx 1 + \epsilon$ (for small values of ϵ) between pairs of protected groups with $\mathbf{s}_i \in A$, $\mathbf{s}_j \in A$, for any utility function that could be chosen.

Protected groups have similar economic outcomes

Measuring Bias Amplification

- We can measure the extent to which an algorithm **increases the bias over the original data**
- Calculate differential fairness of data, ϵ_1
- Calculate differential fairness of algorithm, ϵ_2
- Bias amplification: $\epsilon_2 - \epsilon_1$

This is a more politically conservative fairness definition: implements infra-marginality

Learning with DF Penalty

- Objective: $$\min_{\mathbf{W}}[L_{\mathbf{X}}(\mathbf{W}) + \lambda R_{\mathbf{X}}(\epsilon)]$$

$$R_{\mathbf{X}}(\epsilon) = max(0, \epsilon_{M_{\mathbf{W}}(\mathbf{x})} - \epsilon_1)$$

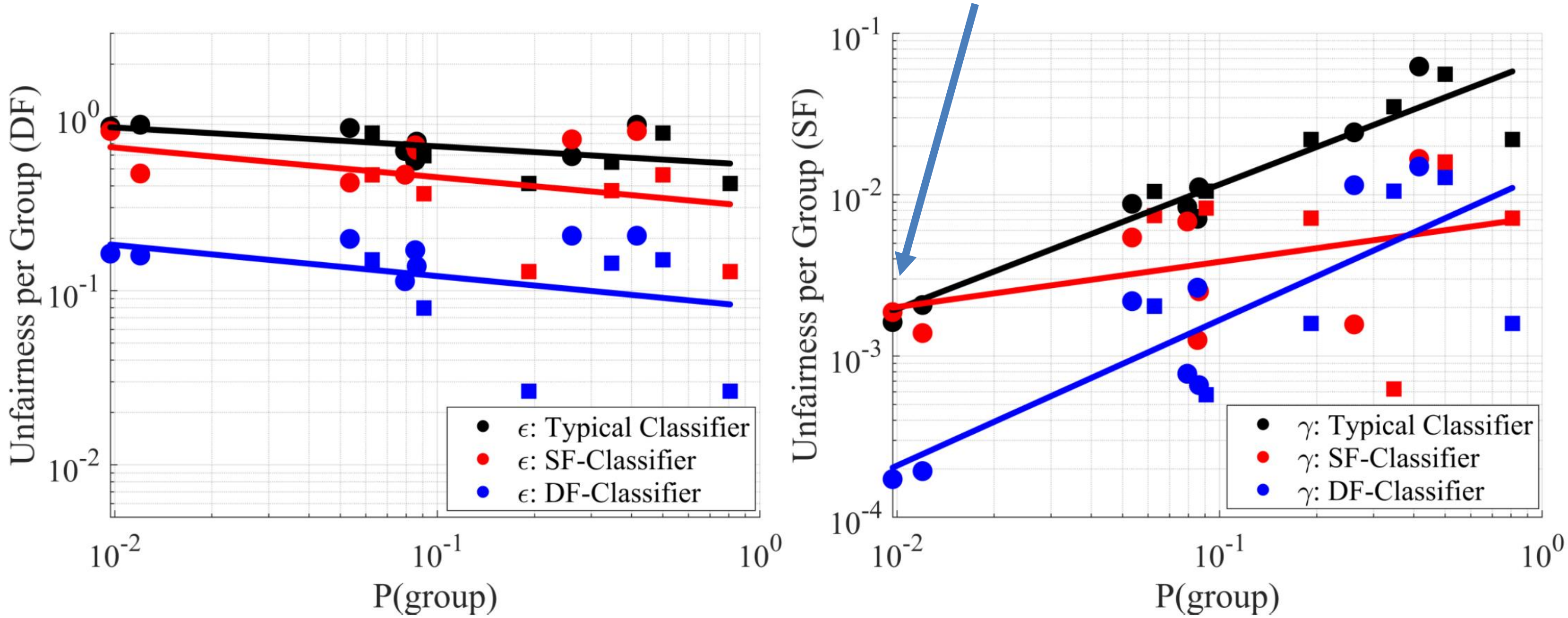
Fairness penalty term

Determines whether to penalize DF or DF-bias amplification

- Optimize via **gradient descent**: backprop + auto-diff (DF-Classifier)
- We use a similar algorithm to enforce subgroup fairness (SF-Classifier)

Learning Results

SF-Classifier ignores minority groups



(a) Improvement in DF measures

(b) Improvement in SF measures

- Both algorithms improve both metrics, both per-group and overall
- DF-classifier improves fairness for minority groups, even under SF metric

Thank you!

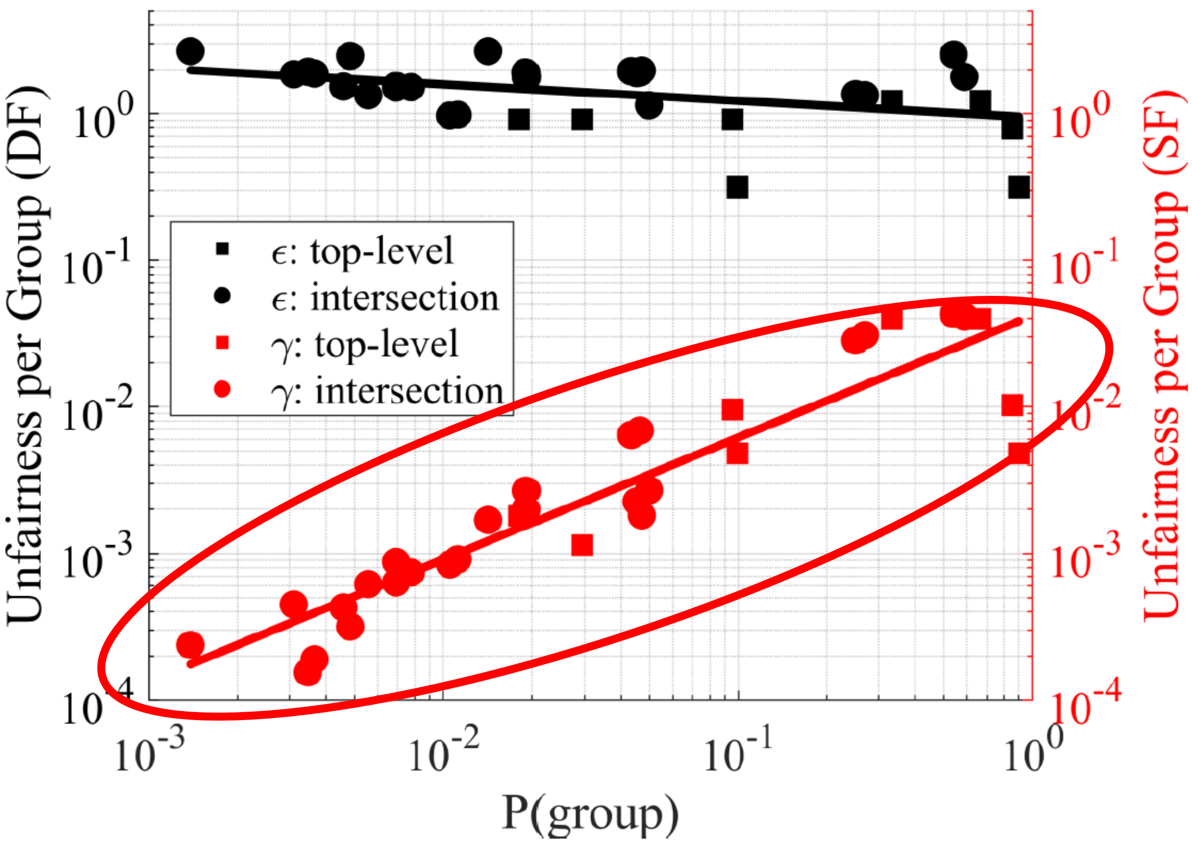
- Contact information:
  - James Foulds
    Assistant professor
    Department of Information Systems
    UMBC
    Email: jfoulds@umbc.edu
    Webpage: http://jfoulds.informationsystems.umbc.edu

- An extended version of our work is online at arxiv.org:
  - J. R. Foulds and S. Pan. **An Intersectional Definition of Fairness.** ArXiv preprint arXiv:1807.08362 [CS.LG]

- An accepted SDM 2020 paper on modeling uncertainty in estimating DF:
  - J. R. Foulds, R. Islam, K. Keya, S. Pan. **Bayesian Modeling of Intersectional Fairness: The Variance of Bias.** SIAM International Conference on Data Mining (**SDM**), ArXiv preprint arXiv:1811.07255 [cs.LG], 2020.

Bonus Slides

Subgroup Fairness and Intersectionality

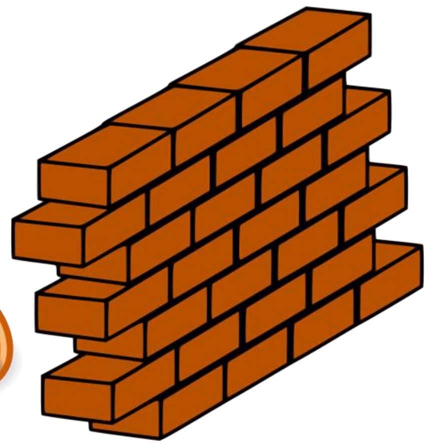


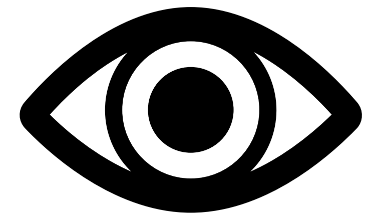
Size of group, as a proportion of the population

Our metric does not down-weight small intersectional groups

Subgroup fairness down-weights small intersectional groups

Fairness and Privacy: the Untrusted Vendor





Vendor
(may be untrusted)

Individuals' data x_i, including protected attribute(s)

Fair algorithm

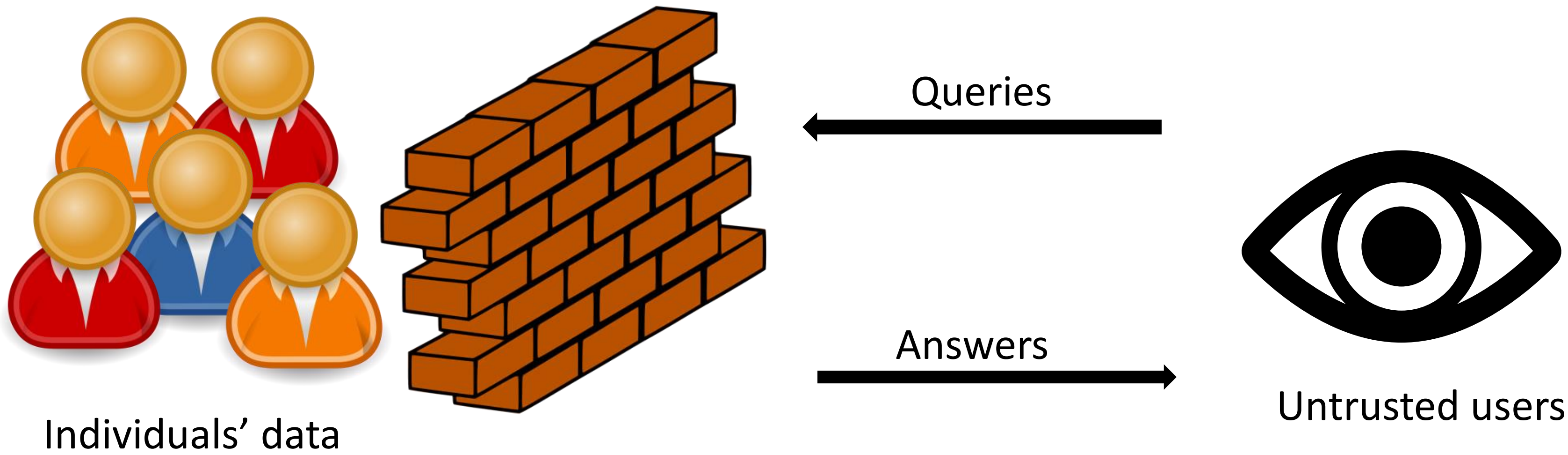
$M(x)$

Outcomes y

The user of the algorithm's outputs (the *vendor*) may discriminate, e.g. in retaliation for a fairness correction (Dwork et al., 2012)

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. In Proceedings of the 3rd innovations in theoretical computer science conference (pp. 214-226). ACM.

Interlude: Differential Privacy (Dwork et al., 2006)

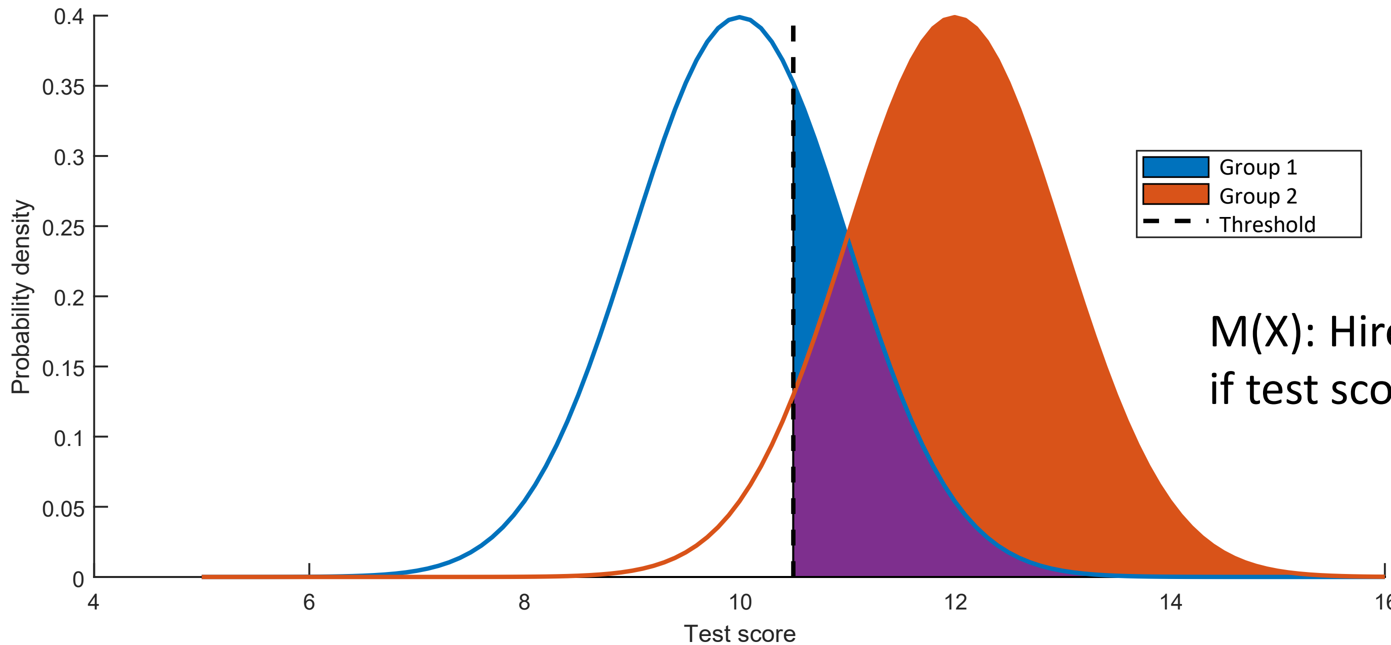


Privacy-preserving interface: randomized algorithms

- DP is a promise:
  - "If you add your data to the database, you will not be affected much"

C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In Theory of Cryptography, pages 265–284. Springer, 2006.

Differential Fairness Example

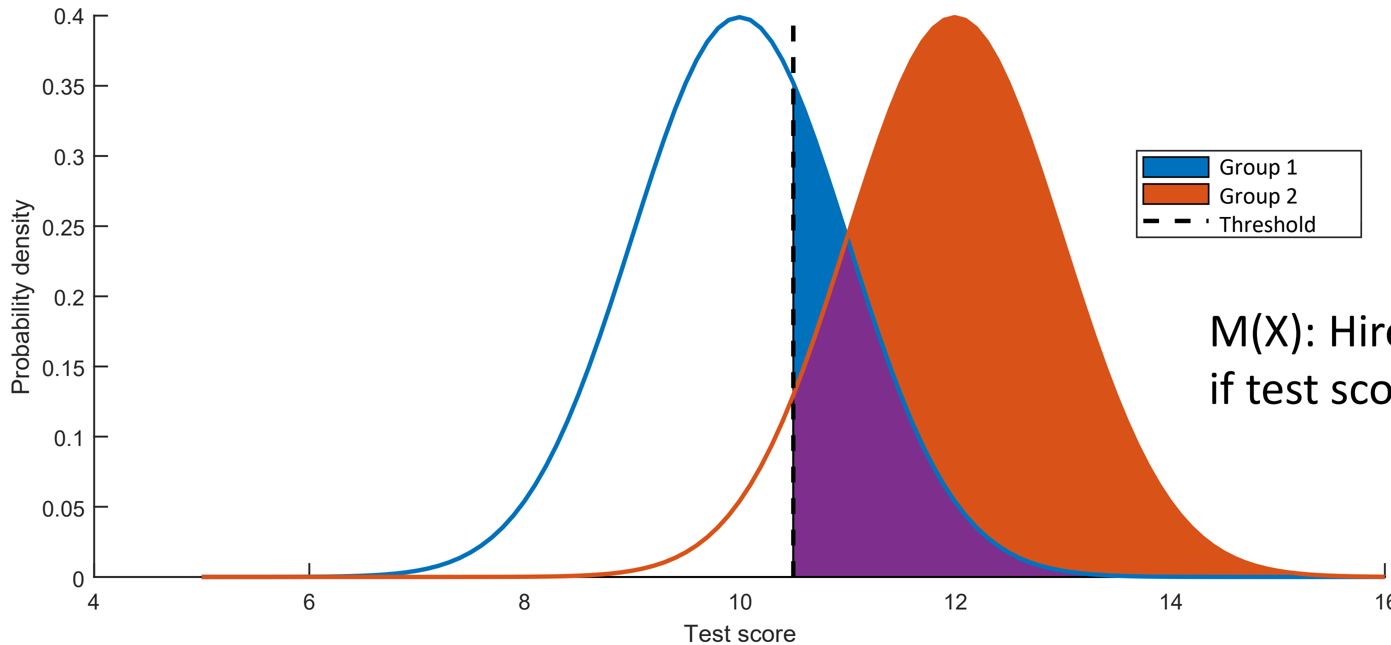


M(X): Hire applicant if test score > 10.5

Scenario: Given an applicant's score on a standardized test, an applicant is hired if there test score is greater than a threshold *t*. Here, t = 10.5.
Each group of applicant has a different distribution over scores:

$$\text{Group 1: } N(X; \mu_1 = 10, \sigma = 1)$$
$$\text{Group 2: } N(X; \mu_2 = 12, \sigma = 1)$$

Differential Fairness Example



M(X): Hire applicant if test score > 10.5

Probability of Hiring Outcome Given Group

		Group	
		1	2
Outcome	yes	0.3085	0.9332
	no	0.6915	0.0668

Log Ratios of Probabilities

y	s_i	s_j	$\log \frac{P_{M,\theta}(M(X)=y\|s_i,\theta)}{P_{M,\theta}(M(X)=y\|s_j,\theta)}$
no	1	2	2.337
	2	1	-2.337
yes	1	2	-1.107
	2	1	1.107

Find the worst case: $\epsilon = 2.337$

Measuring Bias in Data

- Can **measure bias in a dataset**

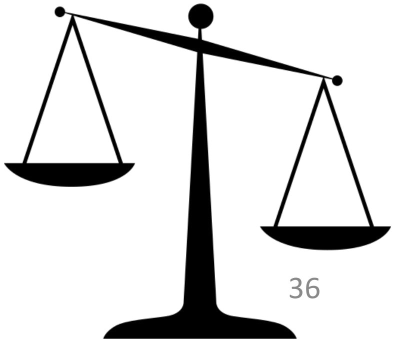
Special case of differential fairness, in which the **algorithm is the data distribution**

Empirical differential fairness (EDF) of a labeled dataset:

Corresponds to verifying that for any y, s_i, s_j, we have

$$\boxed{e^{-\epsilon} \leq \frac{N_{y,s_i}}{N_{s_i}} \frac{N_{s_j}}{N_{y,s_j}} \leq e^{\epsilon}}$$

- Also applies to a probabilistic model of the data



Learning Results

Models		DF-Classifier			SF-Classifier		Typical Classifier
		$\epsilon_1 = 0.0$	$\epsilon_1 = 0.2231$	$\epsilon_1 = \epsilon_{data}$	$\gamma_1 = 0.0$	$\gamma_1 = \gamma_{data}$	
Performance Measures	Accuracy	0.686	0.684	0.692	0.690	0.697	0.700
	F1 Score	0.633	0.642	0.643	0.622	0.647	0.641
	ROC AUC	0.730	0.723	0.734	0.719	0.739	0.734
Fairness Measures (using soft counts)	ϵ-DF	**0.180**	0.281	0.410	0.404	0.468	0.773
	γ-SF	**0.006**	0.021	0.033	0.007	0.028	0.035
	Bias Amp-DF	**-0.360**	-0.259	-0.130	-0.136	-0.072	0.233
	Bias Amp-SF	**-0.015**	0.000	0.012	-0.014	0.007	0.014
Fairness Measures (using hard counts)	ϵ-DF	**0.207**	0.671	0.884	0.825	0.860	0.897
	γ-SF	**0.015**	0.045	0.060	0.017	0.048	0.062
	Bias Amp-DF	**-0.339**	0.125	0.338	0.279	0.314	0.351
	Bias Amp-SF	**-0.025**	0.005	0.020	-0.023	0.008	0.022

Table 3: Comparison of intersectionally fair classifiers with the typical classifier on the COMPAS dataset ($\epsilon_1 = 0.2231$ is the 80% rule).

- Little to no loss in accuracy metrics when trained to prevent bias amplification
- Differential fairness is protected or improved vs training data (“bias de-amplification”)

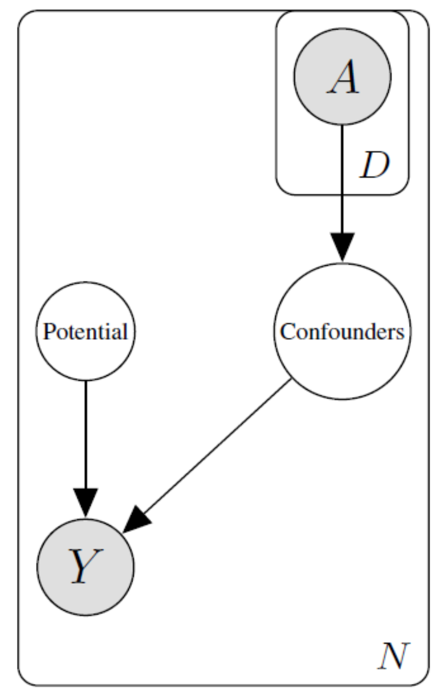
Dealing With Confounders

- UC Berkeley admissions: Simpson's paradox
  - *"Department applied to"* is a confounder
  - Demographic parity no longer ideal

- Solution:
  protect DF **per department**

  DFC: Differential fairness w/ confounders

- **Theorem:** overall admissions
  DF no worse than DFC
  (i.e. DF of the "worst" dept)



Proof of Intersectionality Theorem

Proof. Define $E = S_1 \times \ldots \times S_{a-1} \times S_{a+1} \ldots \times S_{k-1} \times S_{k+1} \times \ldots \times S_p$, the Cartesian product of the protected attributes included in A but not in D. Then for any $\theta \in \Theta$, $y \in \mathrm{Range}(M)$,

$$\begin{aligned}
&\log \max_{\mathbf{s} \in D : P(\mathbf{s}|\theta) > 0} P_{M,\theta}(M(\mathbf{x}) = y | D = s, \theta) \\
&= \log \max_{\mathbf{s} \in D : P(\mathbf{s}|\theta) > 0} \sum_{e \in E} P_{M,\theta}(M(\mathbf{x}) = y | E = e, \mathbf{s}, \theta) P_\theta(E = e | \mathbf{s}, \theta) \\
&\leq \log \max_{\mathbf{s} \in D : P(\mathbf{s}|\theta) > 0} \sum_{e \in E} \max_{e' \in E : P_\theta(E = e' | \mathbf{s}, \theta) > 0} \\
&\qquad \big(P_{M,\theta}(M(\mathbf{x}) = y | E = e', \mathbf{s}, \theta)\big) \times P_\theta(E = e | \mathbf{s}, \theta) \\
&= \log \max_{\mathbf{s} \in D : P(\mathbf{s}|\theta) > 0} \max_{e' \in E : P_\theta(E = e' | \mathbf{s}, \theta) > 0} P_{M,\theta}(M(\mathbf{x}) = y | E = e', \mathbf{s}, \theta) \\
&= \log \max_{\mathbf{s}' \in A : P(\mathbf{s}'|\theta) > 0} P_{M,\theta}(M(\mathbf{x}) = y | \mathbf{s}', \theta)
\end{aligned}$$

By a similar argument, $\log \min_{\mathbf{s} \in D : P(\mathbf{s}|\theta) > 0} P_{M,\theta}(M(\mathbf{x}) = y | D = \mathbf{s}, \theta) \geq \log \min_{\mathbf{s}' \in A : P(\mathbf{s}'|\theta) > 0} P_{M,\theta}(M(\mathbf{x}) = y | \mathbf{s}', \theta)$. Applying Lemma 7.1, we hence bound ϵ in (D, Θ) as

$$\begin{aligned}
&\log \max_{\mathbf{s} \in D : P(\mathbf{s}|\theta) > 0} P_{M,\theta}(M(\mathbf{x}) = y | D = \mathbf{s}, \theta) \\
&\quad - \log \min_{\mathbf{s} \in D : P(\mathbf{s}|\theta) > 0} P_{M,\theta}(M(\mathbf{x}) = y | D = \mathbf{s}, \theta) \\
&\leq \log \max_{\mathbf{s}' \in A : P(\mathbf{s}'|\theta) > 0} P_{M,\theta}(M(\mathbf{x}) = y | \mathbf{s}', \theta) \\
&\quad - \log \min_{\mathbf{s}' \in A : P(\mathbf{s}'|\theta) > 0} P_{M,\theta}(M(\mathbf{x}) = y | \mathbf{s}', \theta) \leq \epsilon \; .
\end{aligned}$$

Lemma 7.1

ϵ-DF criterion can be rewritten as: for any $\theta \in \Theta$, $y \in$ Range(M),

$$\begin{aligned}
&\log \max_{\mathbf{s} \in A : P(\mathbf{s}|\theta) > 0} P_{M,\theta}(M(\mathbf{x}) = y | \mathbf{s}, \theta) \\
&\quad - \log \min_{s \in A : P(\mathbf{s}|\theta) > 0} P_{M,\theta}(M(\mathbf{x}) = y | \mathbf{s}, \theta) \leq \epsilon \; .
\end{aligned}$$

References

- **J. Angwin**, J. Larson, S. Mattu, and L. Kirchner. Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. ProPublica, May, 23, 2016.
- Barocas, S., & Selbst, A. D. Big data's disparate impact. Cal. L. Rev., 104, pp. 671-732, 2016.
- Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., and Kalai, A. T. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In Advances in Neural Information Processing Systems, 2016.
- **K. Crenshaw**. Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. U. Chi. Legal F., pages 139–167, 1989.
- **C. Dwork**, M. Hardt, T. Pitassi, O. Reingold, & R. Zemel. Fairness through awareness. In Proceedings of the 3rd innovations in theoretical computer science conference (pp. 214-226). ACM, 2012.
- **C. Dwork**, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In Theory of Cryptography, pages 265–284. Springer, 2006.
- Executive Office of the President. Big Data: A Report on Algorithmic Systems, Opportunity, and Civil Rights. White House, Executive Office of the President, 2016.
- M. Hardt, E. Price, N. Srebro, et al. Equality of opportunity in supervised learning. In Advances in NIPS, pages 3315–3323, 2016.
- D. Kifer and A. Machanavajjhala. Pufferfish: A framework for mathematical privacy definitions. ACM Trans. on Database Systems, 39(1):3, 2014.
- M. J. Kusner, J. Loftus, C. Russell, & R. Silva. Counterfactual fairness. In Advances in Neural Information Processing Systems (pp. 4069-4079), 2017.
- **S. Lowry** & G. Macpherson, A Blot on the Profession, BRIT. MED. J. 296, pp. 657-658 (1988).
- **S. U. Noble**. Algorithms of Oppression: How Search Engines Reinforce Racism. NYU Press, 2018.
- **C. Simoiu**, S. Corbett-Davies, S. Goel, et al. The problem of infra-marginality in outcome tests for discrimination. The Annals of Applied Statistics, 11(3):1193–1216, 2017.
- R. Speer. How to Make a Racist AI Without Really Trying. ConceptNet blog.2017. http://blog.conceptnet.io/posts/2017/how-to-make-a-racist-ai-without-really-trying/
- **C. Starmans**, M. Sheskin, and P. Bloom. Why people prefer unequal societies. Nature Human Behaviour, 1(4):0082, 2017.
- **Zhao, J.**, Wang, T., Yatskar, M., Ordonez, V., & Chang, K. W. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. EMNLP, 2017.

(9/16 female first authors, indicated in bold)

Conclusion

“The rise of big-data optimism is here, and if ever there were a time when politicians, industry leaders, and academics were enamored with artificial intelligence as a superior approach to sense-making, it is now.

This should be a wake-up call for people living in the margins, and people aligned with them, to engage in thinking through the interventions we need.”

-**Safiya Umoja Noble**. *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York University Press, 2018