(BIASED) HISTORY OF ALGORITHMIC FAIRNESS RESEARCH

VECTOR INSTITUTE

INSTITUT VECTEUR

UNIVERSITY OF TORONTO

RICHARD ZEMEL

MAY 24, 2019

CIFAR

CANADIAN INSTITUTE FOR ADVANCED RESEARCH

WHY WAS I NOT SHOWN THIS AD?

FAIRNESS IN AUTOMATED DECISIONS

Algorithmic unfairness: Algorithms are pervasive, high-stakes, high-impact

Need more than just ”accuracy”

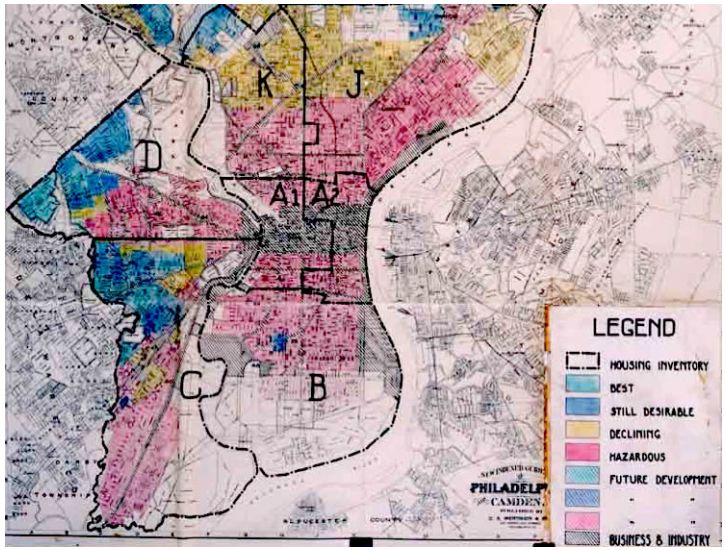
CONCERN: DISCRIMINATION

- Population includes minorities
  - Ethnic, religious, medical, geographic
- Protected by law, policy, ethics
- (If) we cannot completely control our data, can we regulate how it is used, how decisions are made based on it?

Forms of Discrimination

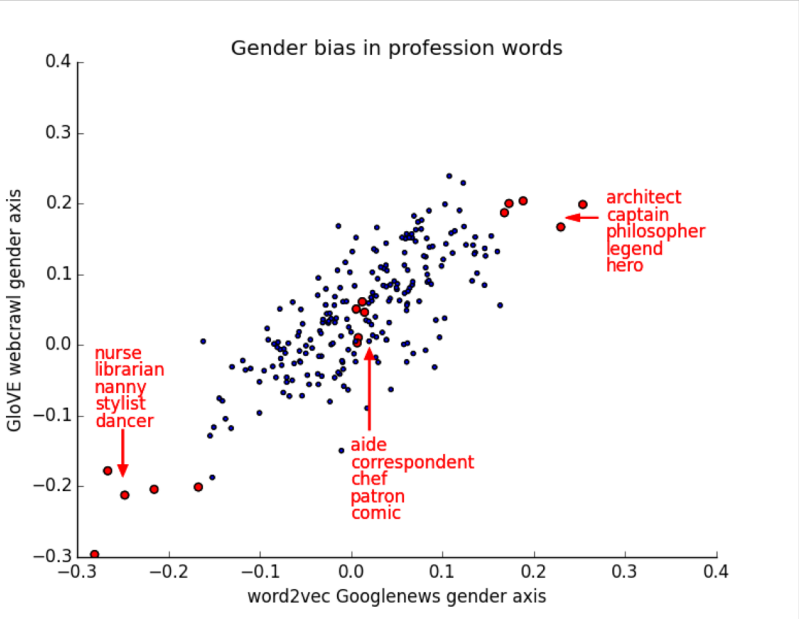
- *Steering* minorities into higher rates (advertising)

- *Redlining:* deny service, change rates based on area

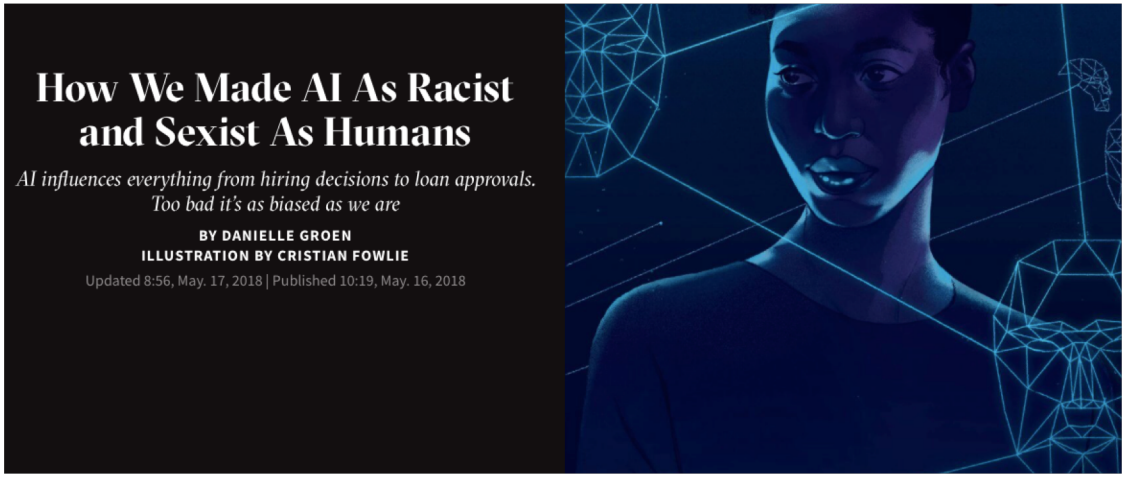


- *Self-fulfilling prophecy:* select less qualified to "justify" future discrimination

Unfairness in Machine Learning?

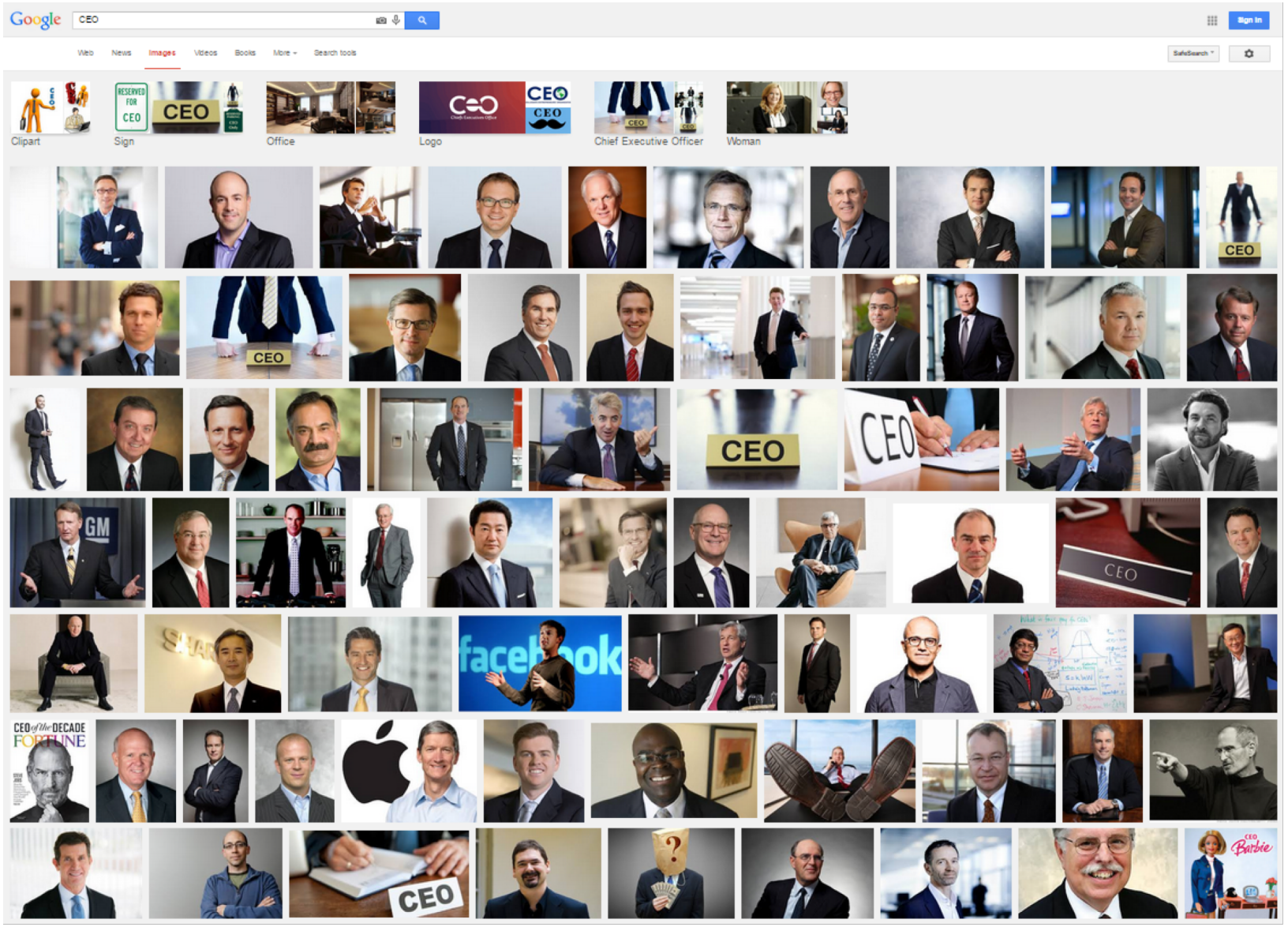


Joy Buolawmini



The Walrus, 2018

SUBTLER BIAS



SUBTLER BIAS

Fairness in ML: Goals

Identify and mitigate bias in ML-based decision-making, in all aspects of data pipeline

Sources of Bias/Discrimination

DATA

- Imbalanced / impoverished data
- Labeled data imbalance (more data on white recidivism outcomes)
- Labeled data incorrect / noisy (historical bias)

MODEL

- ML prediction error imbalanced
- Compound injustices (Hellman)

FAIR CLASSIFICATION

Explosion of fairness research over last five years

Fair classification is the most common setup, involving:

- X, some data
- Y, a label to predict
- $\hat{Y}$, the model prediction
- A, a sensitive attribute (race, gender, age, socio-economic status)

We want to learn a classifier that is:

- accurate
- fair with respect to A

FAIRNESS VIA S-BLINDNESS?

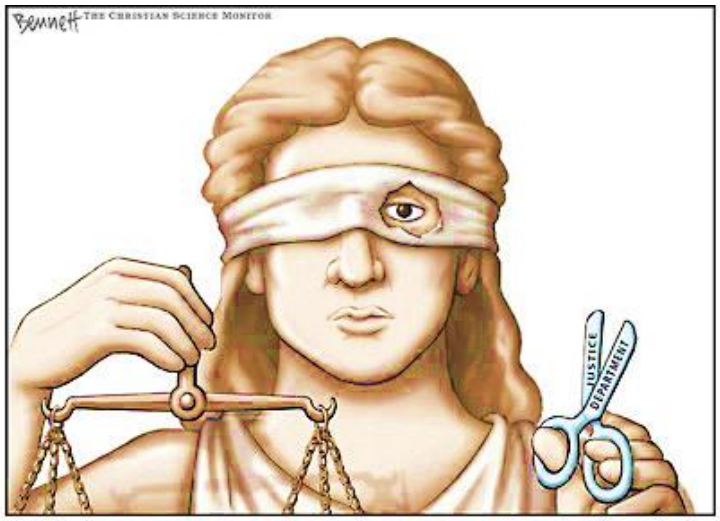
Remove or ignore the "membership in A" bit

- Fails: Membership in A may be encoded in other attributes

FAIRNESS THROUGH AWARENESS

Dwork, Hardt, Pitassi, Reingold, Zemel, 2012

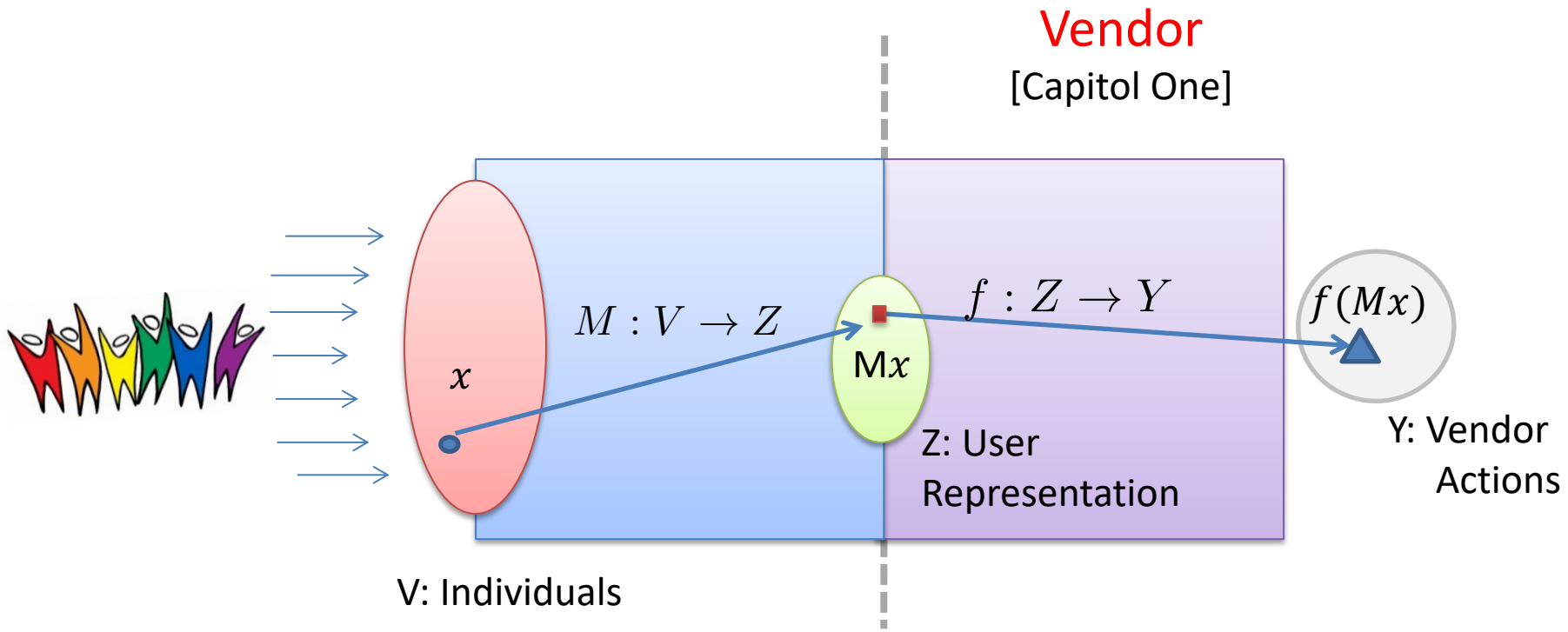
Goal: Assign each individual *a* representation *by being aware of membership in group A*



(1). **Individual Fairness**: Treat similar individuals similarly

(2). **Group Fairness:** equalize two groups (A=1 = minority; A=0 is majority) at the level of outcomes (statistical parity)

General Framework



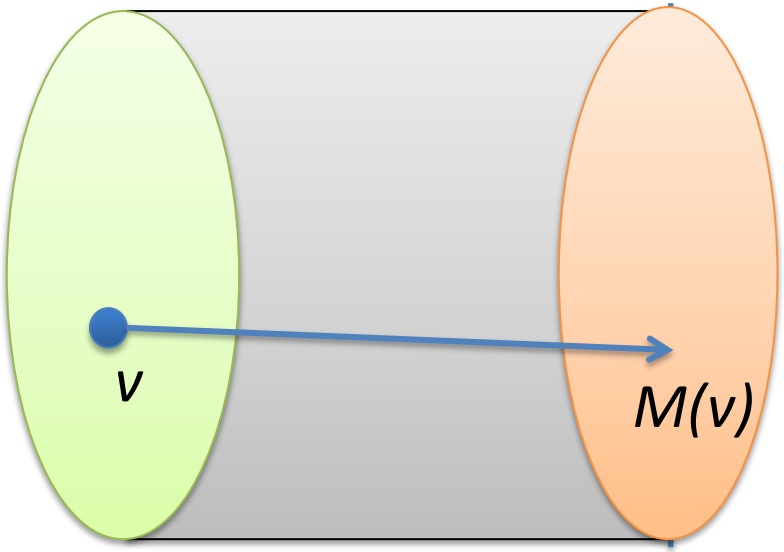
Our goal:

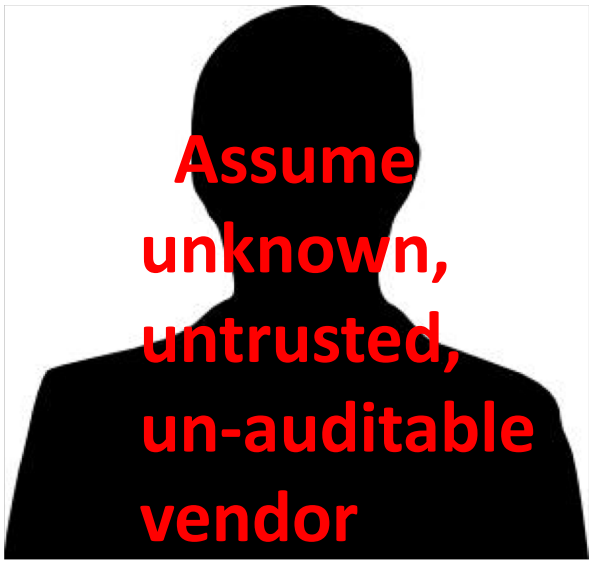
Achieve Fairness in the representation step

Ad Network (with society oversight)

Ad Network

$M : V \to Z$

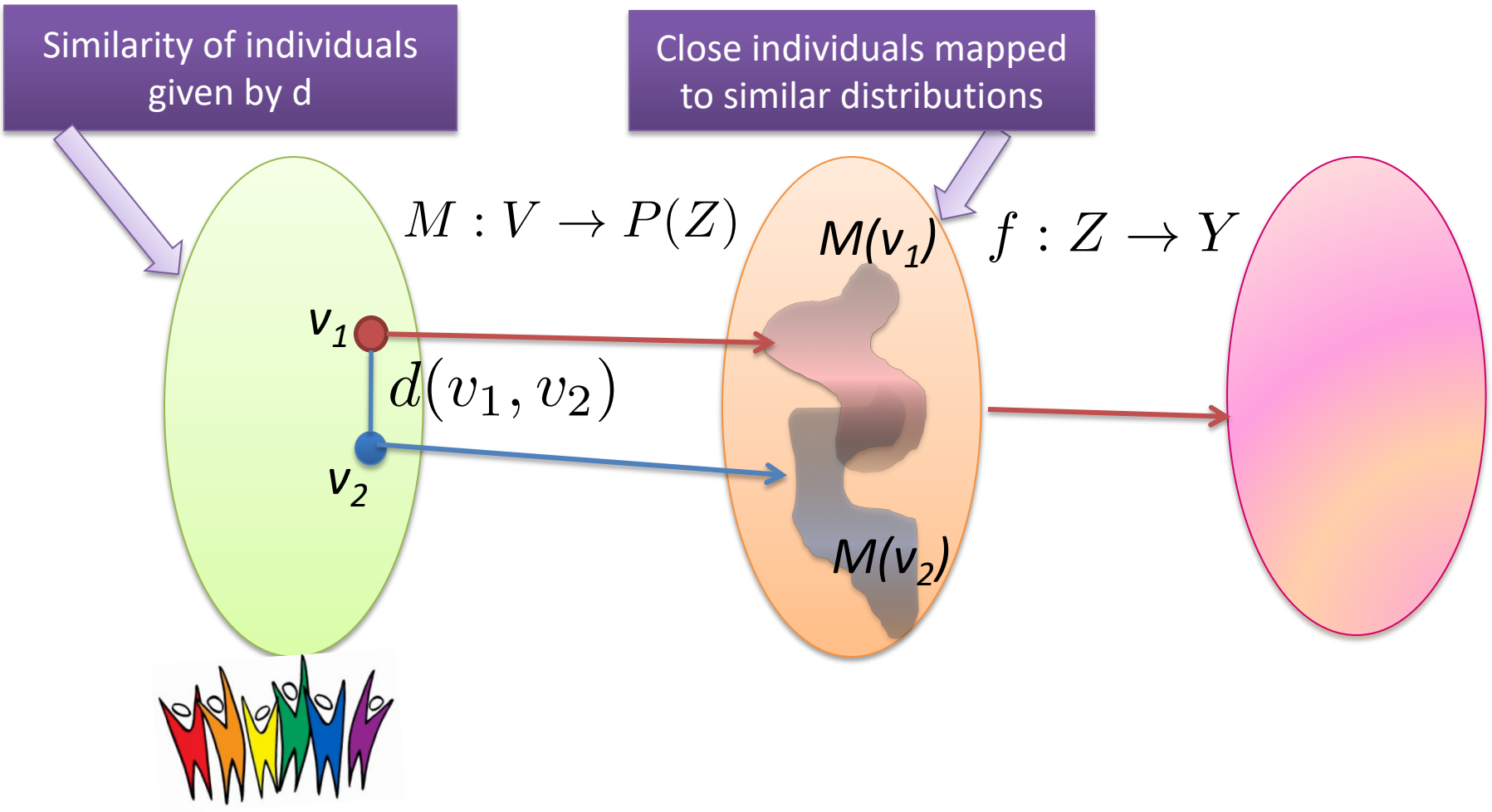




V: Individuals

Z: representation

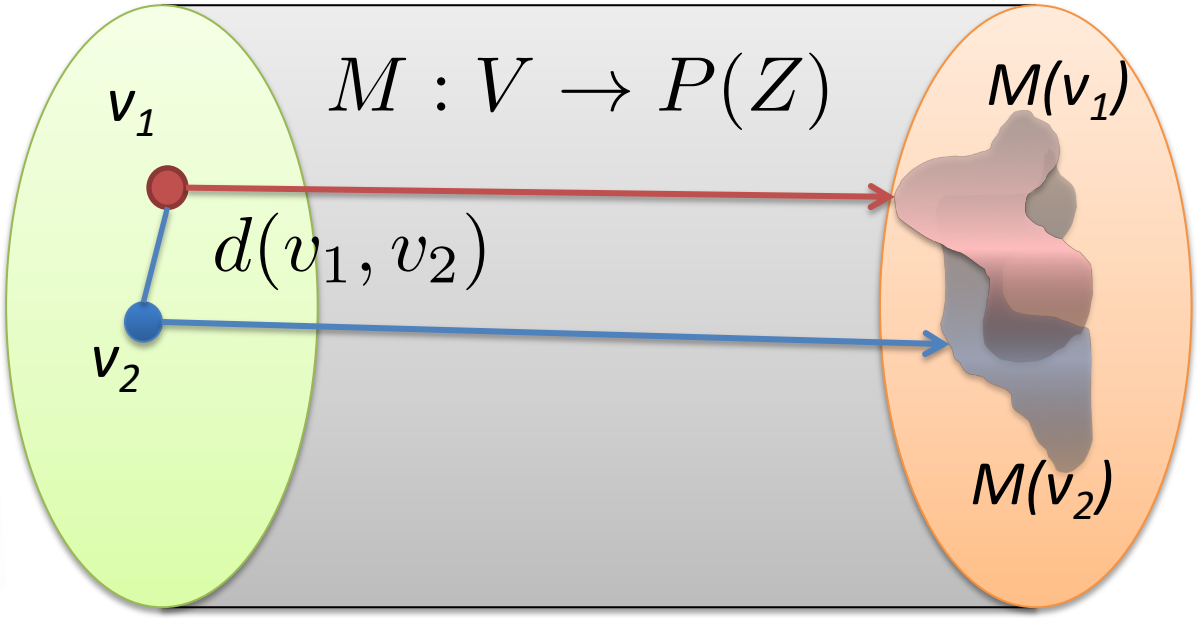
Our Approach: Define a randomized mapping that "blends people with the crowd"



V: Individuals *Z*: representations Y: Actions

Metric $d : V \times V \rightarrow \mathbb{R}$

Lipschitz condition $||M(v_1) - M(v_2)|| \leq d(v_1, v_2)$



V: Individuals

Z: Representations

The Metric

- Assume ***task-specific similarity metric***
  - Extent to which two individuals are similar w.r.t. the classification task at hand
- Ideally captures *ground truth*
  - Or, society's best approximation
- Open to public discussion, refinement

Examples: Financial/insurance risk metrics

  - Already widely used (though secret)
- AALIM health care metric
  - health metric for treating similar patients similarly
- Roemer's relative effort metric
  - Well-known approach in economics/political theory

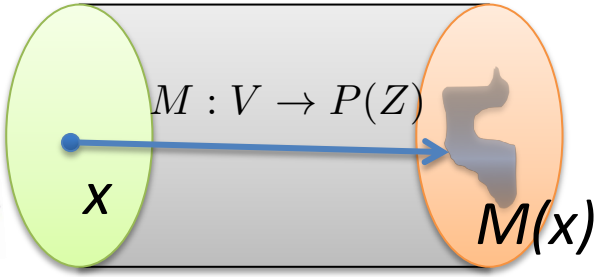
An Algorithm for Fair Classification

utility function
$U: V \times Z \to R$

Metric
$d: V \times V \to R$

Efficient Procedure

d-fair mapping M

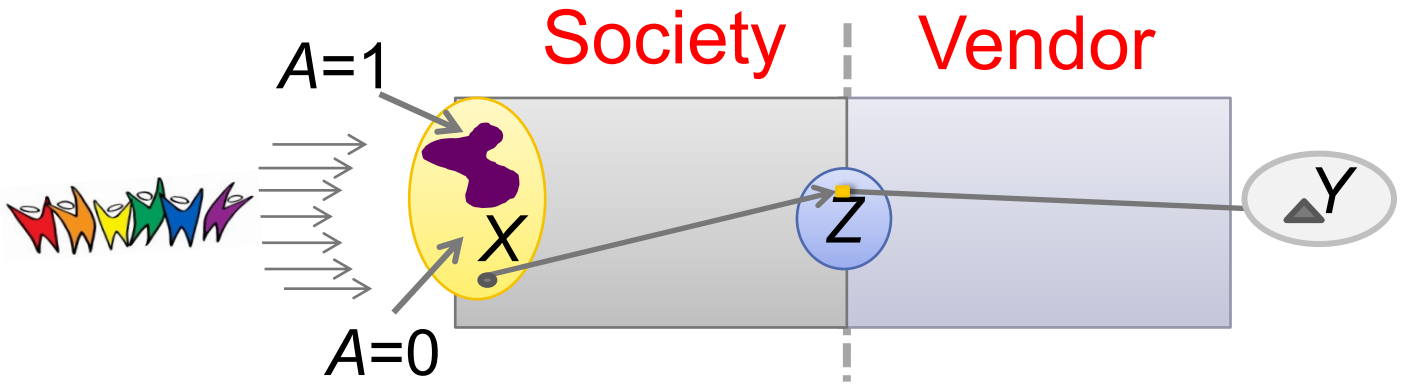


V: Individuals *Z:* Encodings

LP maximizes vendor's expected utility subject to fairness conditions

FAIR REPRESENTATION LEARNING: FRAMEWORK

Zemel. Wu, Swersky, Pitassi, Dwork, 2013



Goal: Learn a mapping from X to distributions over representations *Z that is fair*

Aims for Z:

1. Lose information about A:

$$P[Z=k \mid A=1] = P[Z=k \mid A=0]$$

2. Retain information about X
3. Preserve information for classification so vendor can max utility [decisions Y = g(Z)]

INITIAL FORMULATION

Difficult to jointly optimize:

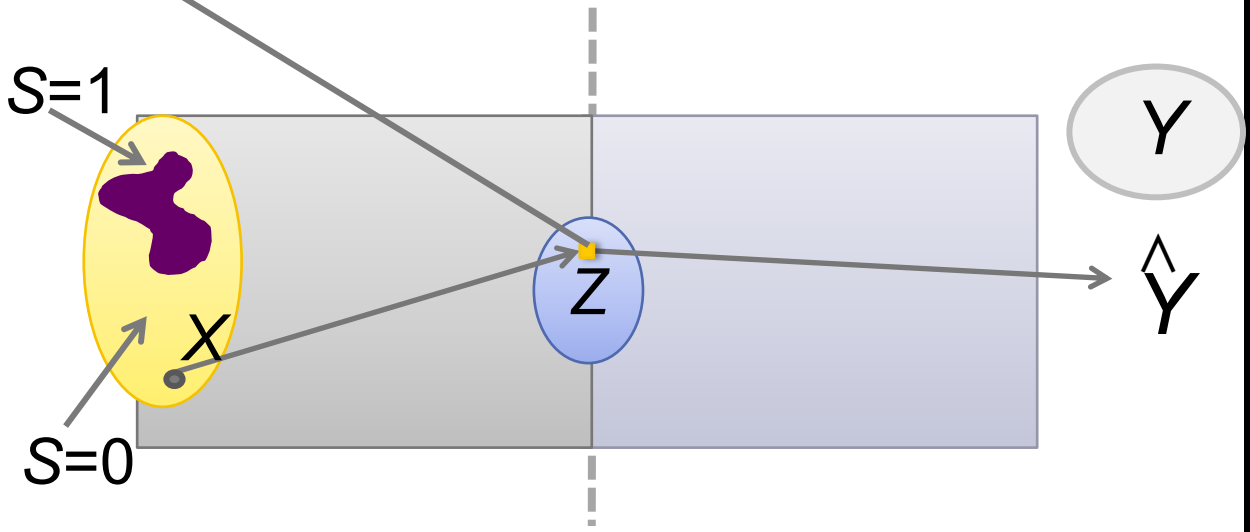
min. |f(Z) – Y|; max. |g(Z) – S| (thwart adversary)

Can alternate:

optimize M,f given g;

optimize M,g given f

But unstable



INSTANTIATING THE MODEL

Key: min. *MI(Z,S)* by forcing *P(Z|S+)* = *P(Z|S-)*

$$P(Z|S) = \int_X P(Z|X,S)P(X|S)dX$$

$$P(Z|S=1) \approx \frac{1}{N^+} \sum_{n=1}^{N^+} P(Z|X,S=1)$$

$$P(Z|S=1) = P(Z|S=0) = P(Z) \Rightarrow$$

Simple tractable formulation:

$$MI(Z,S) = 0$$

***Z* is a discrete latent variable**

FULL OBJECTIVE FUNCTION

Learn mapping M(X) to minimize L

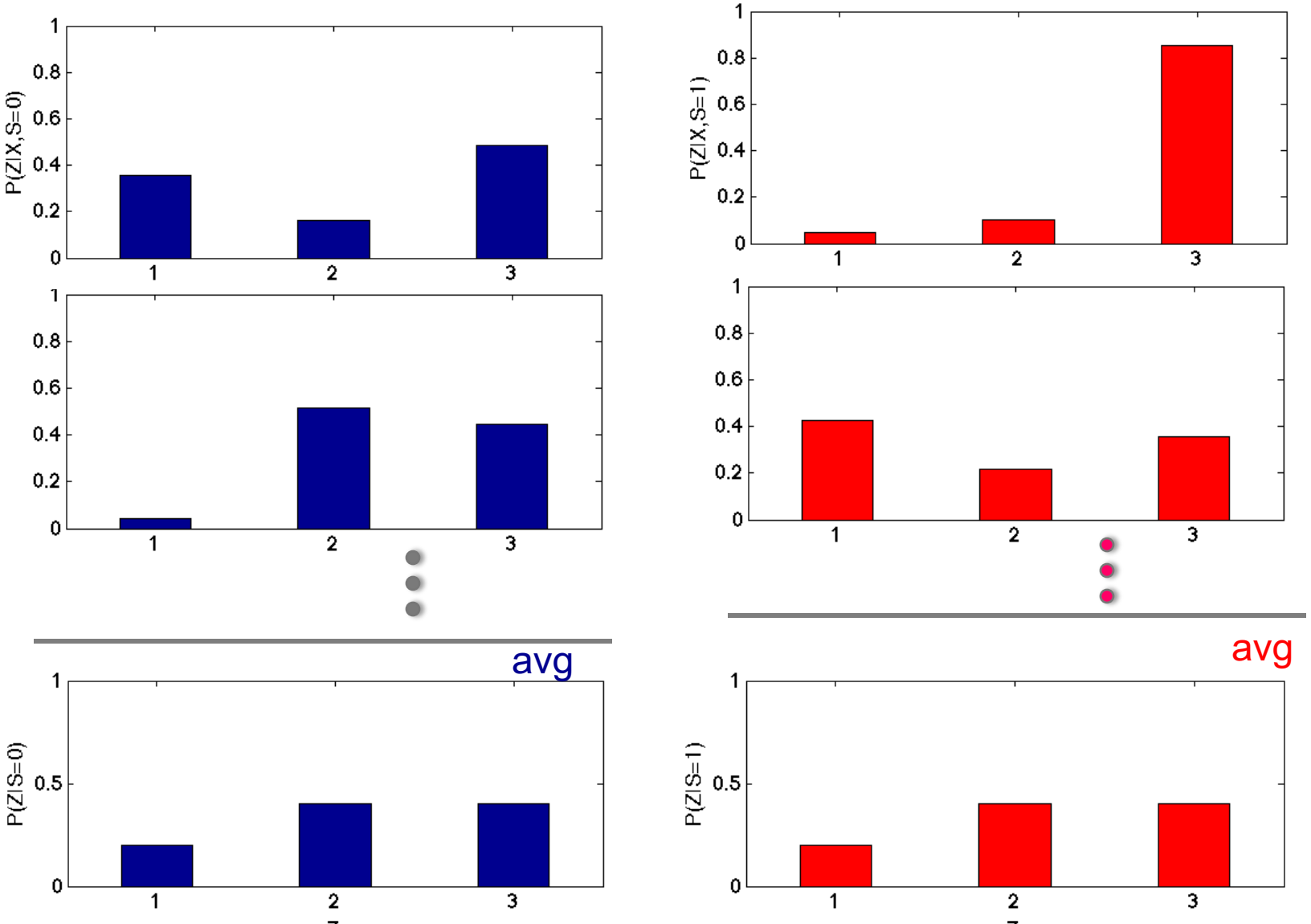
$$P^+_{n,k} = P(Z = k|\mathbf{x}, S = 1) = \frac{\exp(\mathbf{x}_n^T \mathbf{w}_k^+)}{\sum_{k'} \exp(\mathbf{x}_n^T \mathbf{w}_{k'}^+)}$$

$$L = A_y \cdot L_y + A_z \cdot L_z$$

$$L_z = \sum_k |P_k^+ - P_k^-| \qquad P_k^+ = P(Z = k|S = 1)$$

$$L_y = \sum_{n=1}^{N} -y_n \log \hat{y}_n - (1 - y_n) \log(1 - \hat{y}_n) \qquad \hat{y}_n = \sum_k P_{n,k} u_k$$

OBFUSCATING MEMBERSHIP



$$P(Z|S=1) = P(Z|S=0) \Rightarrow MI(Z,S) = 0$$

EXPERIMENTS

1. **German Credit**

   Size: 1000 instances, 20 attributes

   Task: classify as good or bad credit

   Sensitive feature: Age

2. **Adult Income**

   Size: 45,222 instances, 14 attributes

   Task: predict whether or not annual income > 50K

   Sensitive feature: Gender

3. **Heritage Health**

   Size: 147,473 instances, 139 attributes

   Task: predict whether patient spends any nights in hospital

   Sensitive feature: Age

PERFORMANCE METRICS

- **Accuracy**

$$yAcc = 1 - \frac{1}{N}\sum_{n=1}^{N}|y_n - \hat{y}_n|$$

- **Discrimination**

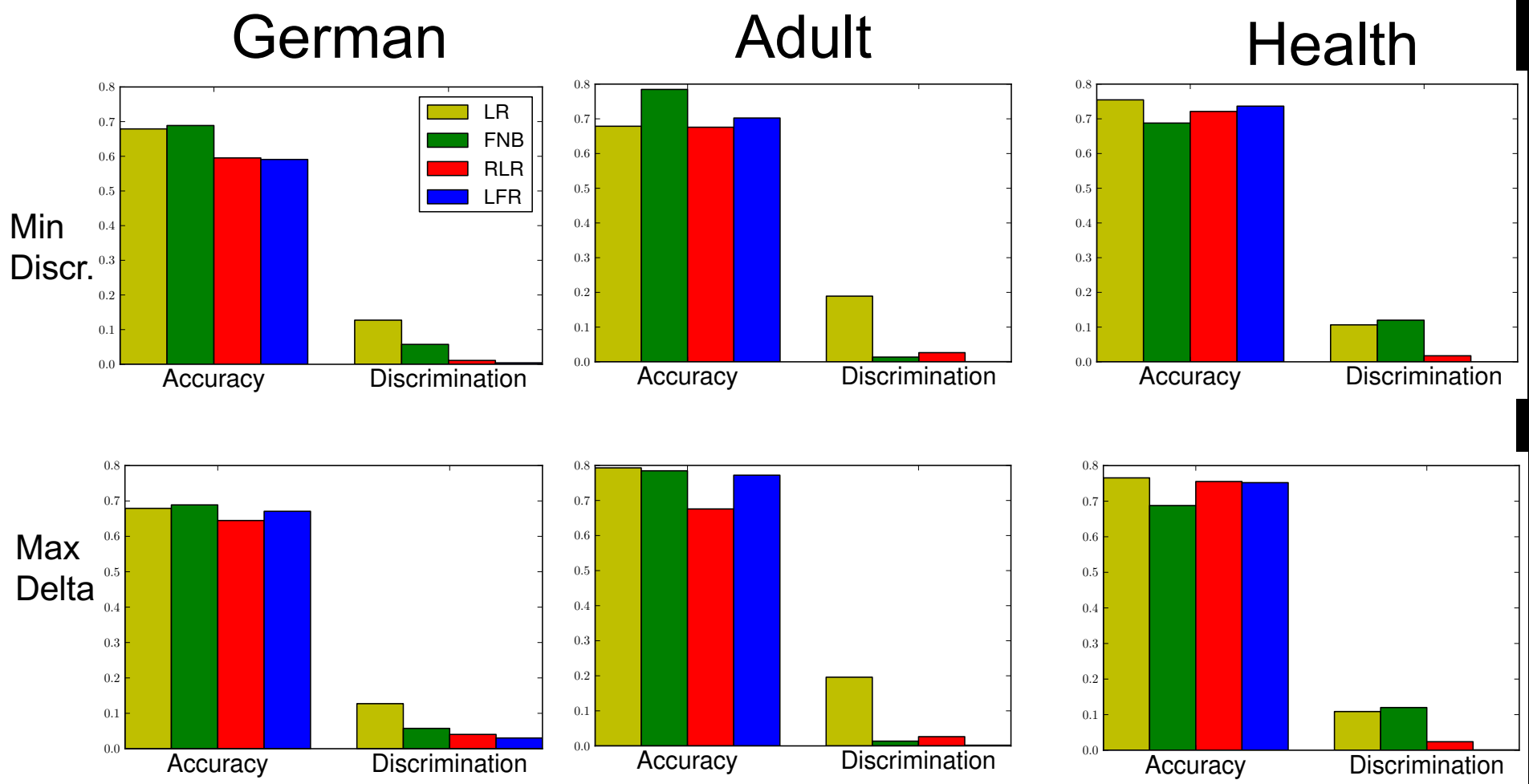
$$yDiscrim = |\frac{\sum_{n:s_n=1}\hat{y}_n}{\sum_{n:s_n=1}1} - \frac{\sum_{n:s_n=0}\hat{y}_n}{\sum_{n:s_n=0}1}|$$

ALTERNATIVE APPROACHES

Build fair classifier and force vendor to use it:

- Massage labels to achieve proportional access (FNB) [Kamiran & Calders, 2009]
- Trade off classification error vs. discrimination (RLR) [Kamishima et al, 2011]

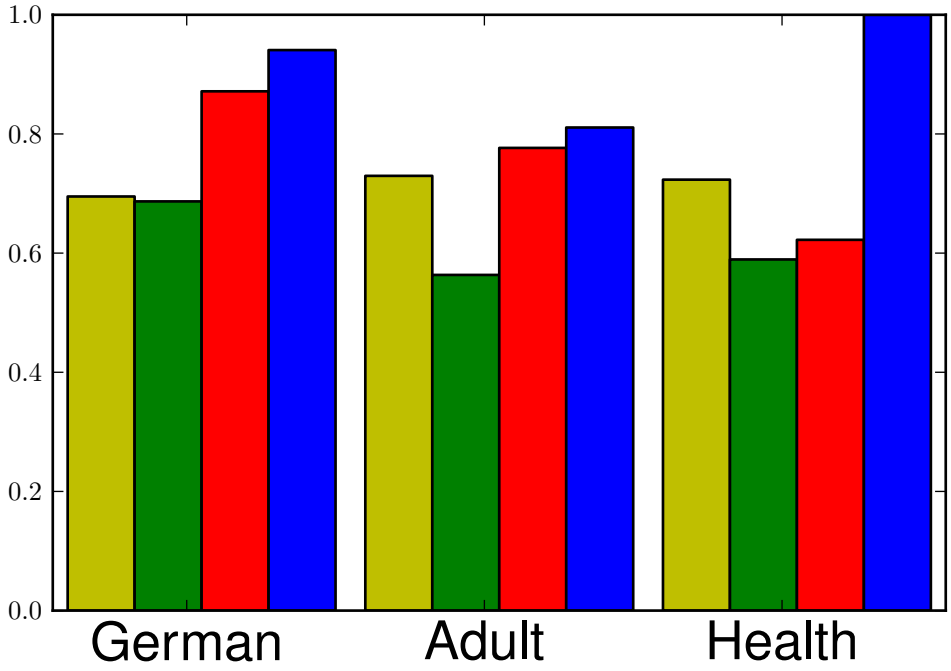
EXPERIMENTAL RESULTS



RESULTS: INDIVIDUAL FAIRNESS

Consistency:

$$yNN = 1 - \frac{1}{Nk} \sum_{n} |\hat{y}_n - \sum_{j \in kNN(\mathbf{x}_n)} \hat{y}_j|$$



EXAMPLE DOMAINS

1. Targeted search/advertising: How do different groups see internet content?
   - Males/females with equal interest, equal p(ad)?
   - (leisure interests; lower paying jobs; credit card rates)

2. Medical testing/diagnosis: decision-making based on tests, that affect p(diagnosis)
   - Applied uniformly to different groups
   - Medical tests for conditions that vary widely between groups

3. Recidivism: risk tools assess p(future-arrest) given history
   - Used in decisions about bail, sentencing, parole
   - Claims of bias based on race against COMPAS risk tool

Common:

1. Algorithm input to decision-maker
2. Attempting to classify individual possesses property: interest; condition; risk
3. Output is a probability

RECIDIVISM PREDICTION

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

COMPAS RECIDIVISM PREDICTION

ProPublica got COMPAS risk scores for 7K in Broward County 2013-14:

- Also recorded which were charged with crimes in next 2 years
- 20% predicted to commit violent crimes did
- 61% predicted to commit any crimes did

Key claim of bias:

- Black defendants who did not re-offend had higher averages scores than white defendants who did not re-offend
- White defendants who did re-offend had lower average scores than black defendants who re-offended

NORTHPOINTE RESPONDS

COMPAS scores are *well-calibrated in each group*

- Consider all white defendants assigned a score of 0.4
- A *v* fraction of them go on to re-offend
- The same is true for white defendants assigned a score of *v*

The score *v* means the same in each group

Calibration is a good measure – if it is not satisfied, then the sensitive attribute may be important to take into consideration

DESIRED PROPERTIES

1. Calibration within groups: For each group a v_b fraction of people in bin *b* are positive

2. Balance for the negative class: average score assigned to people of both groups who belong to the negative class should be the same (not more inaccurate for negative instances in one group than the other).

3. Balance for the positive class: average score assigned to people of both groups who belong to the negative class should be the same

Key question: **Can we achieve all of these at the same time?**

WHEN ARE THE PROPERTIES ACHIEVABLE?

Can achieve all three properties in two simple cases:

1. Perfect prediction; for each feature vector, either everyone is in the negative class or positive class
2. Equal base rates: the groups have the same fraction of positive instances

Theorem (Kleinberg et al, 2016) – In any assignment of risk scores, where all three properties can be achieved we must have either perfect predictions or equal base rates

- not about computational power, more about basic limitation wrt assigning estimates to equalize averages
- other similar theorems by Chouldechova; Corbett-Davies et al: discrete classification, not probabilities

WHEN ARE THE PROPERTIES ACHIEVABLE?

Perfect predictions or equal base rates

Calibration – total score in bin equals expected number of positives in that bin

- N_t – number of people in group t
- k_t -- number of people in positive class in group t
- Calibration – k_t is total score in group t (sum across all bins)

- X :avg score of people in negative class (both groups)
- Y: avg score of people in positive class (both groups)

Total score of group t:

- $(N_t - k_t)X + k_t Y = k_t$

- $X = (1-Y)\, k_t \,/\, (N_t - k_t)$

APPROXIMATE VERSION OF THEOREM

- Calibration within groups: For each group approx. v_b of people in bin *b* are positive
- Balance for positive class: Average score of positive members in A is approx. average score of positive members in B
- Balance for negative class: Average score of negative members in A is approx. average score of negative members in B
- Theorem: In any instance where all three properties can be approximately achieved, we must have either approximately perfect prediction or approximately equal base rates