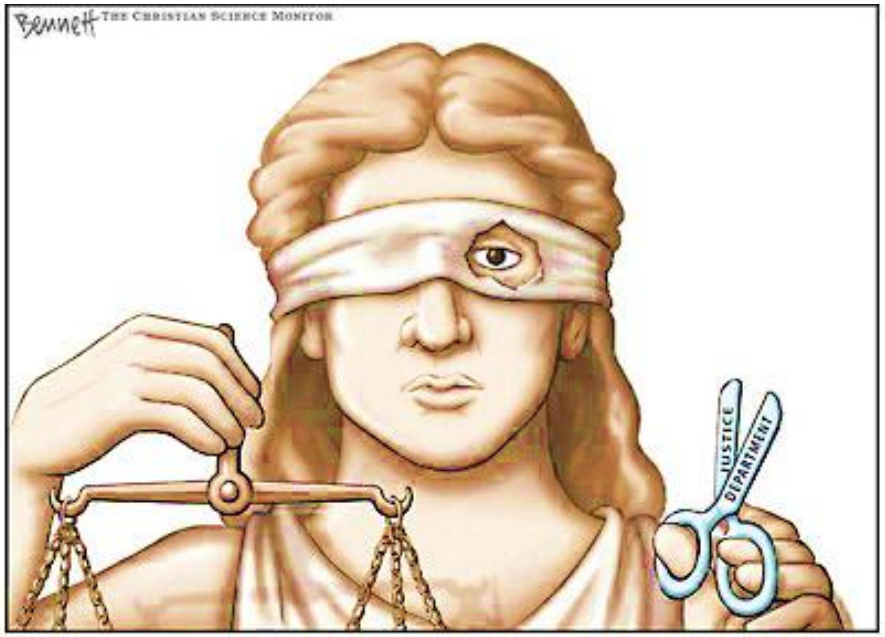
Learning Rich but Fair Representations



Richard Zemel

December 12, 2014

UNIVERSITY OF TORONTO

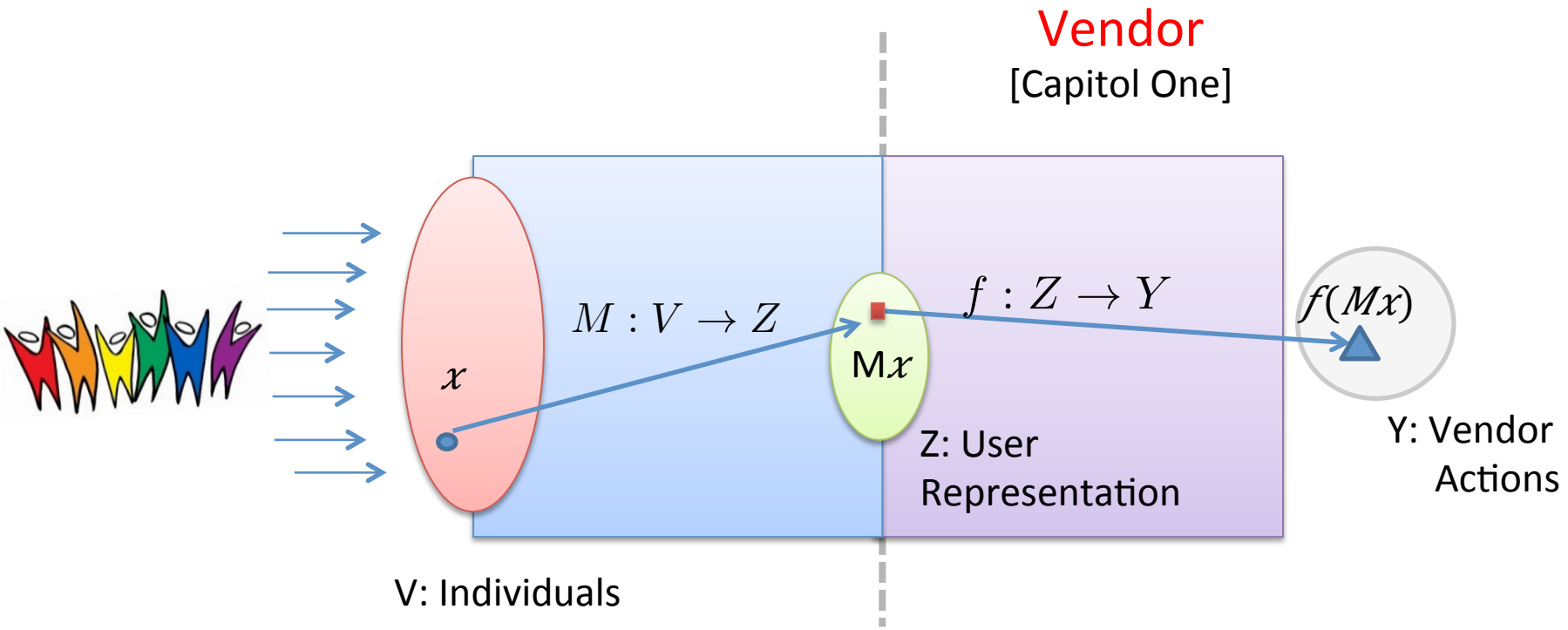
Kevin Swersky

Toni Pitassi

Cynthia Dwork

Yu Wu

General Framework

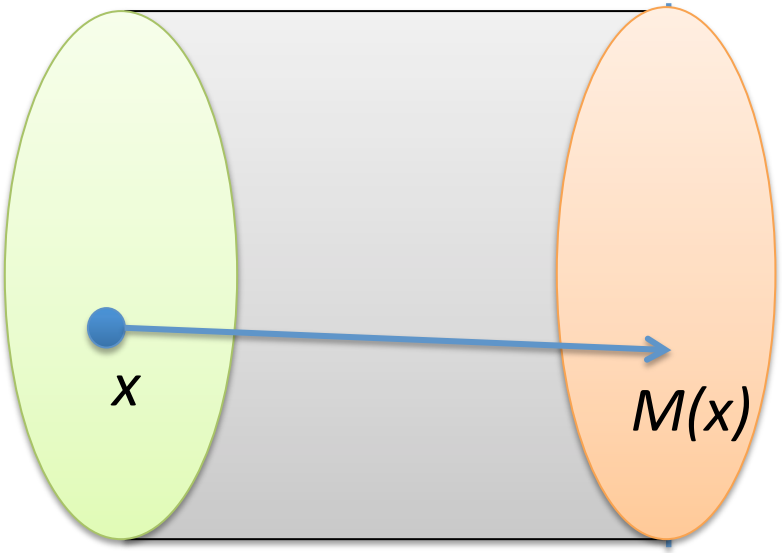


Our goal:

Achieve Fairness in the representation step

Ad Network (with society oversight)

$$M : V \to Z$$



V: Individuals

Z: representation

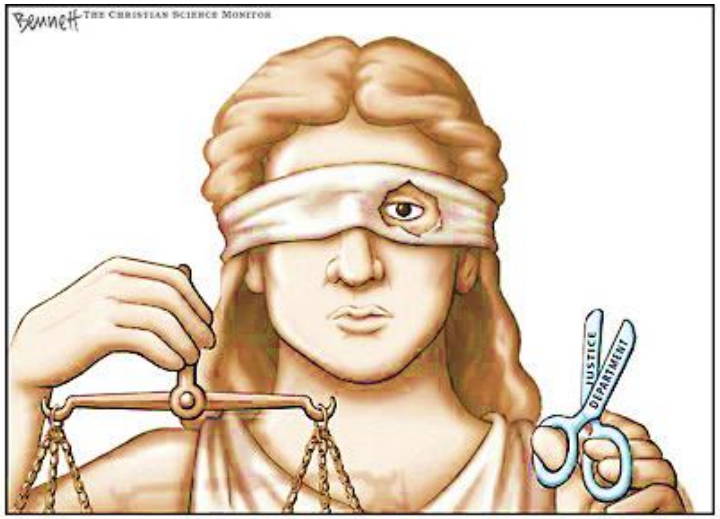
Fairness Through Awareness

Dwork, Hardt, Pitassi, Reingold, Zemel
Innovations in Theoretical Computer Science, 2012

(1). Individual Fairness: Treat similar individuals similarly

(2). Group Fairness: equalize two groups S+, S- at the level of outcomes (statistical parity)

- S+: S=1 = minority, S-: S=0
- P[Z=k | S=1] = P[Z=k | S=0]



- **Insufficient** as a notion of fairness
  - Has good properties, but can be abused

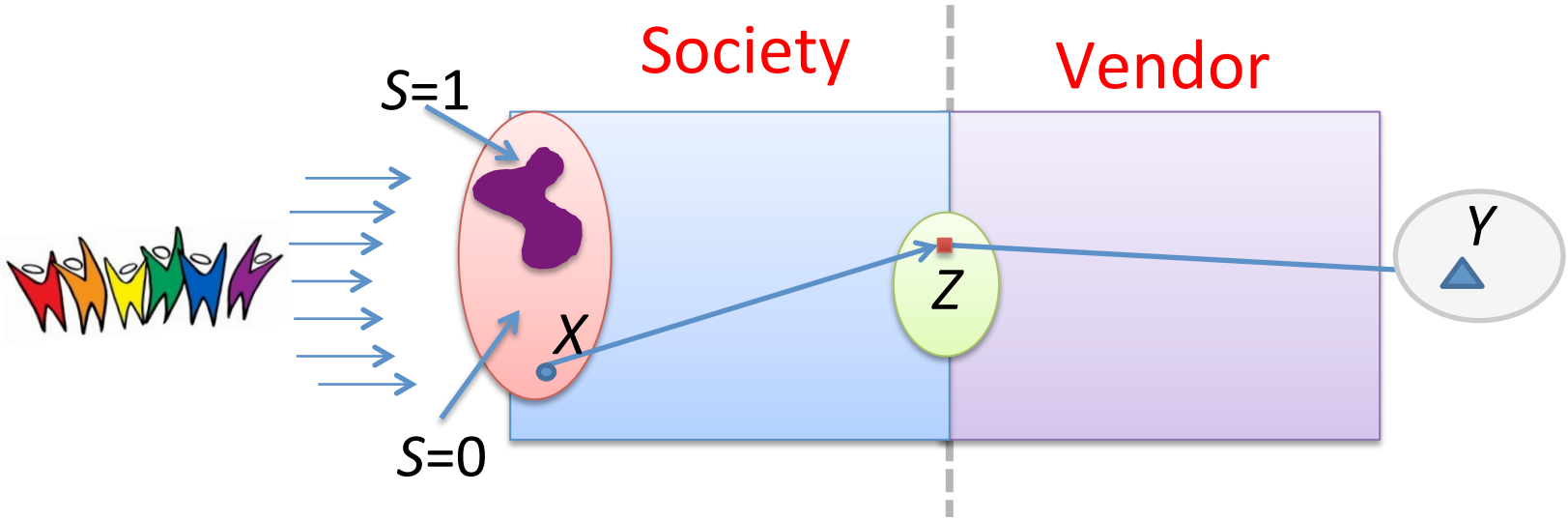
- Fairness requires understanding of classification task
  Cultural understanding of protected groups

Learning Fair Representations

Zemel, Yu, Swersky, Pitassi, Dwork
ICML, 2013

- Generalizes to new data: learn general mapping, applies to any individual
- Mapping should satisfy fairness criteria, vendor utility
- Learn representations of individuals, distances
- Use fair representation for additional classification tasks (transfer learning)
- Working example: dataset of bank loan decisions, protected group (S+) is women

Model Overview



Aims for Z:

1. Lose information about S
   Group Fairness/Statistical Parity: P(Z|S=0) = P(Z|S=1)
2. Preserve information so vendor can max utility

Maximize MI(*Z, Y*); Minimize MI(*Z, S*)

Initial Formulation

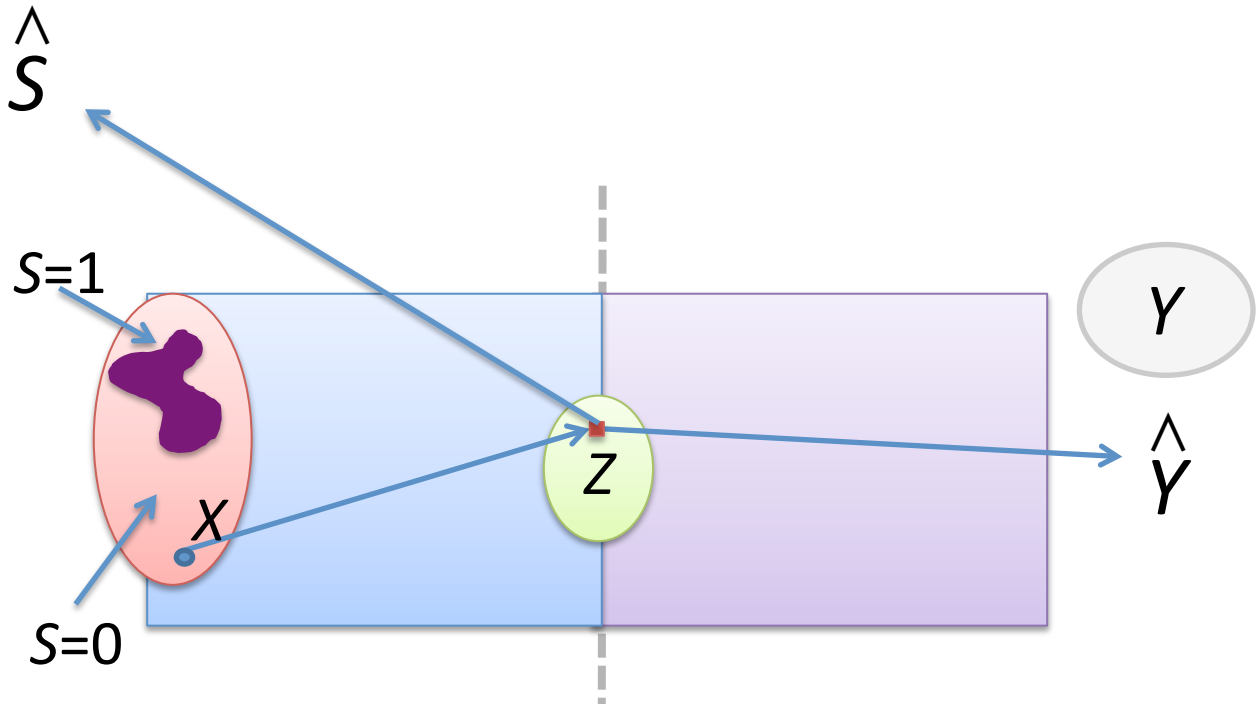
Difficult to jointly optimize:

$$\min. \ |f(Z) - Y|; \quad \max. \ |g(Z) - S|$$

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-)*

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

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

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

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

Simple tractable formulation:

Z is a discrete latent variable

Full Objective Function

Learn mapping M(X) to minimize L

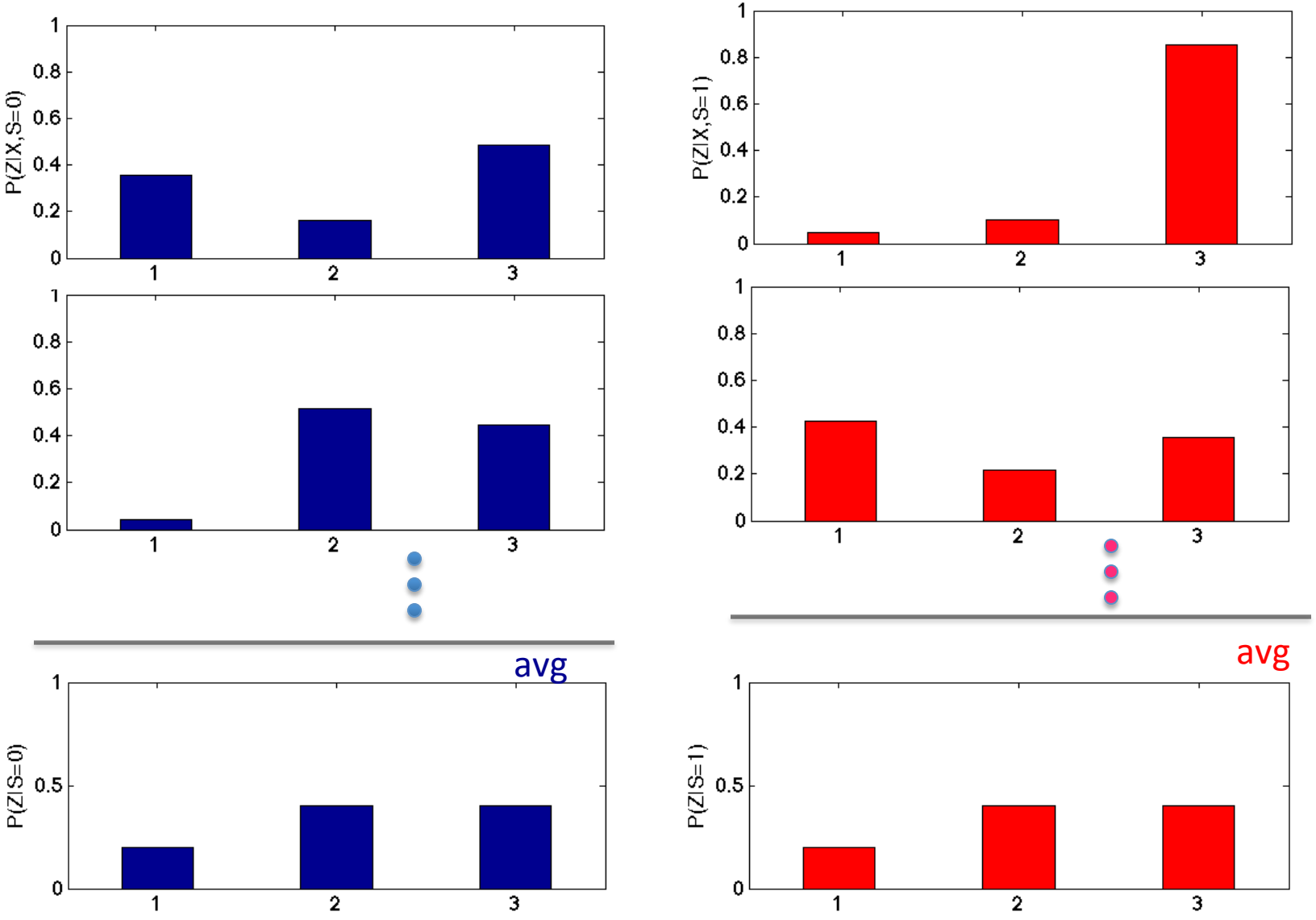
$$P^+_{n,k} = P(Z = k|\mathbf{x}, S = 1) = \frac{\exp(\mathbf{x}_n^T \mathbf{w}_k^+)}{\sum_{k'} \exp(\mathbf{x}_n^T \mathbf{w}_{k'}^+)}$$

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

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

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

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**

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

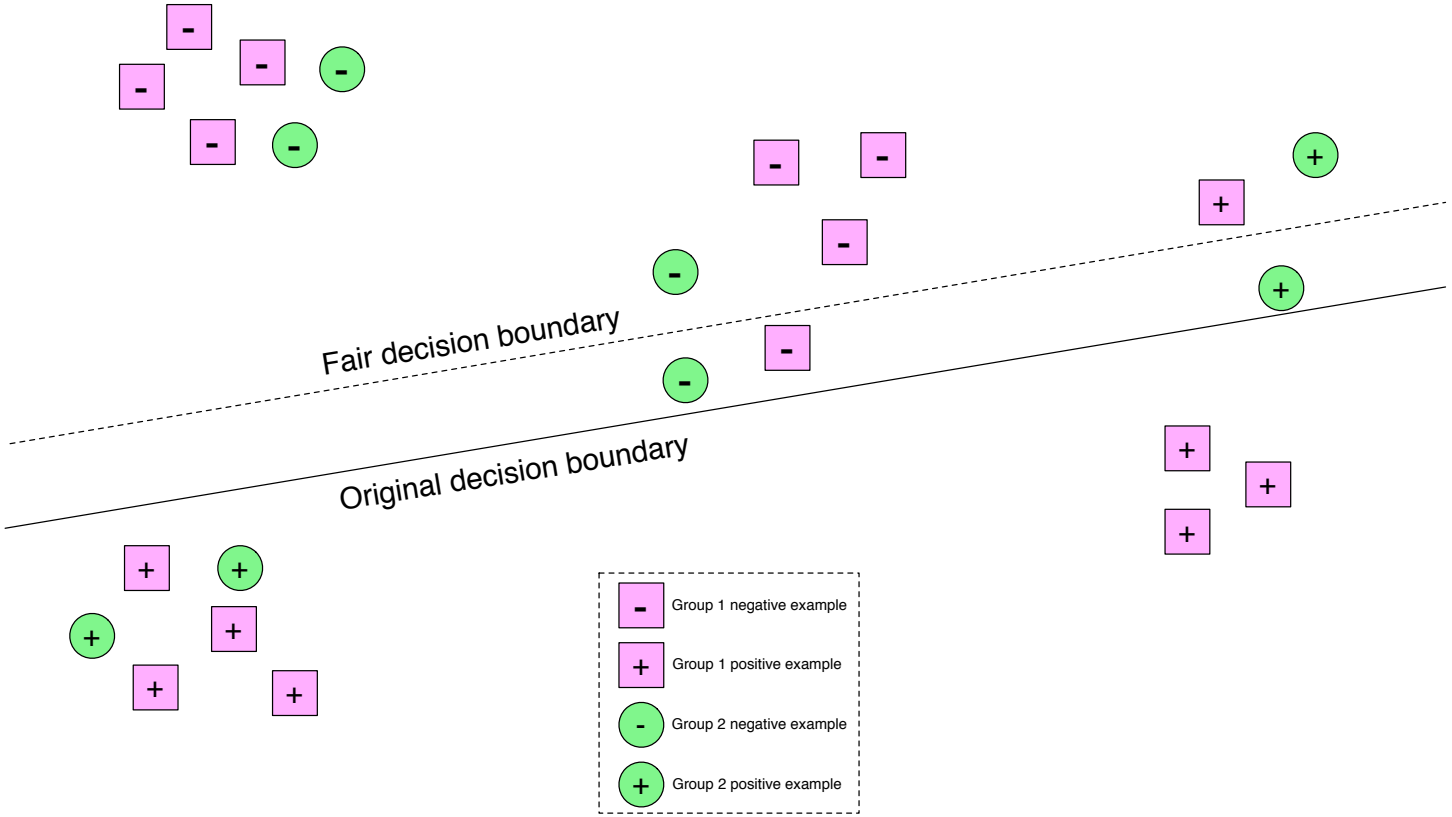
- **Discrimination**

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

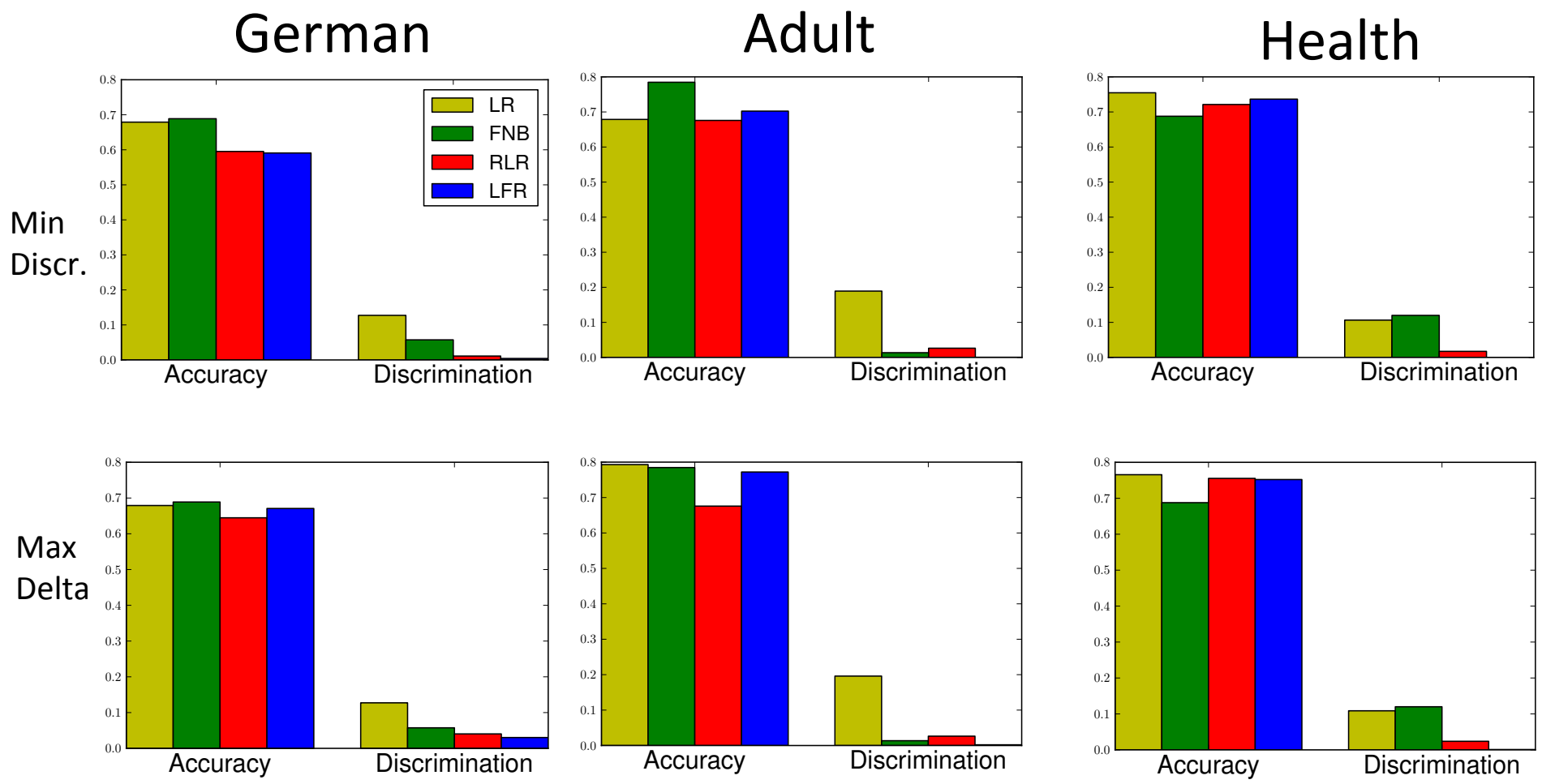
Alternative Approaches

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

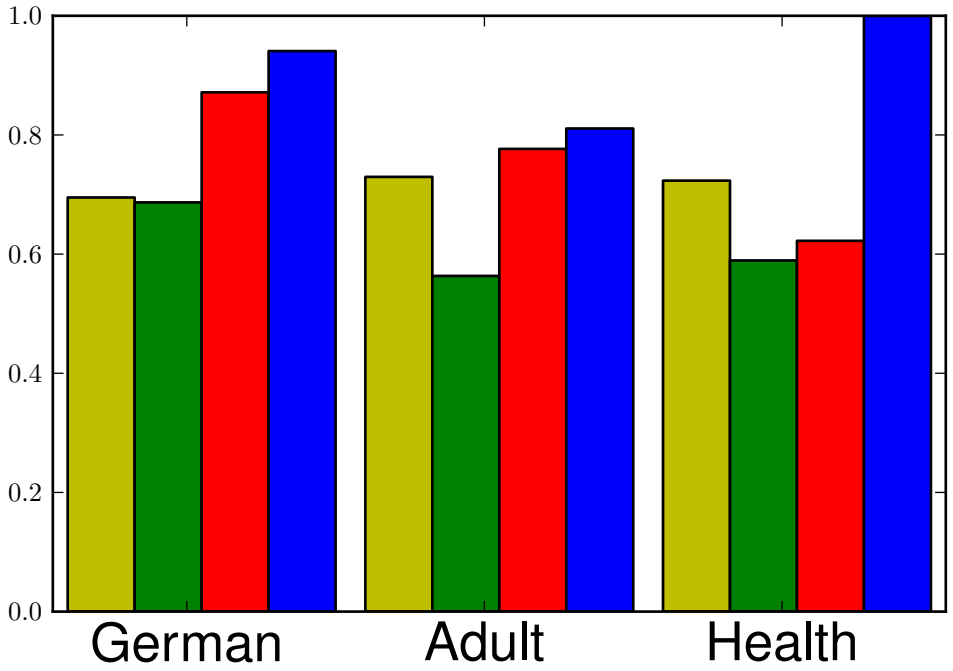


Results: Individual Fairness

Consistency:

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

yNN



Current Direction: Richer Z

- Aim: Replace discrete representation with continuous, multi-dimensional Z
- Allow more flexible, nuanced representations
- Bring ML arsenal to bear: powerful methods for mapping, embedding in vector spaces
- How to maintain statistical parity?
- Minimize

$$d(P(Z|S=1); P(Z|S=0))$$

MMD

- Suppose we have access to samples from two probability distributions $X \sim P_A$ and $Y \sim P_B$, how can we tell if $P_A = P_B$?

- Maximum Mean Discrepancy (MMD) is a measure of distance between two distributions given only samples from each.

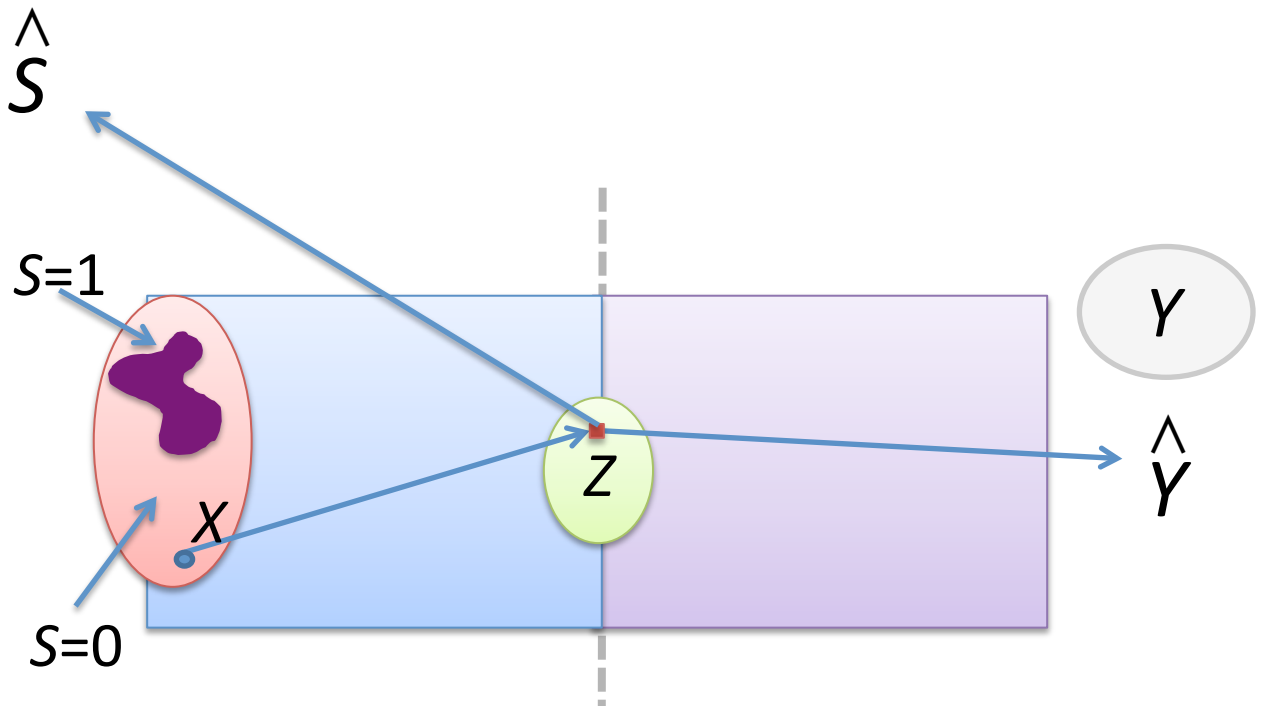
$$\left\| \frac{1}{N} \sum_{n=1}^{N} \phi(X_n) - \frac{1}{M} \sum_{m=1}^{M} \phi(Y_m) \right\|^2$$

$$= \frac{1}{N^2} \sum_{n=1}^{N} \sum_{n'=1}^{N} \phi(X_n)^\top \phi(X_{n'}) + \frac{1}{M^2} \sum_{m=1}^{M} \sum_{m'=1}^{M} \phi(Y_m)^\top \phi(Y_{m'}) - \frac{2}{NM} \sum_{n=1}^{N} \sum_{m=1}^{M} \phi(X_n)^\top \phi(Y_m)$$

$$= \frac{1}{N^2} \sum_{n=1}^{N} \sum_{n'=1}^{N} k(X_n, X_{n'}) + \frac{1}{M^2} \sum_{m=1}^{M} \sum_{m'=1}^{M} k(Y_m, Y_{m'}) - \frac{2}{MN} \sum_{n=1}^{N} \sum_{m=1}^{N} k(X_n, Y_m)$$

Sequence of Models

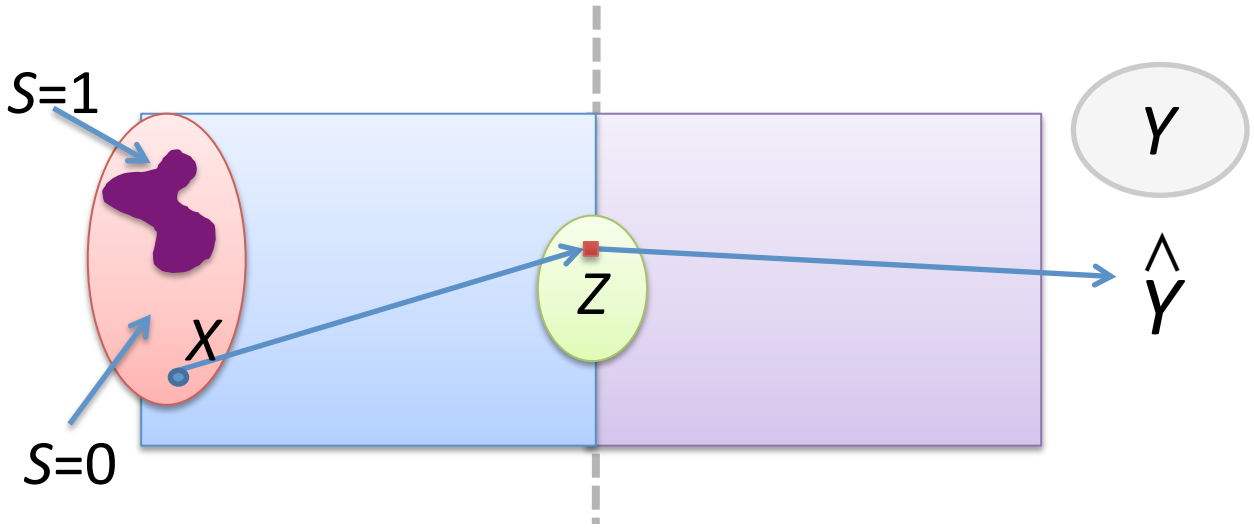
(1). Achieve fairness by preventing adversary from predicting S



Sequence of Models

(2). Match means, using k-dim. multinomial Z:

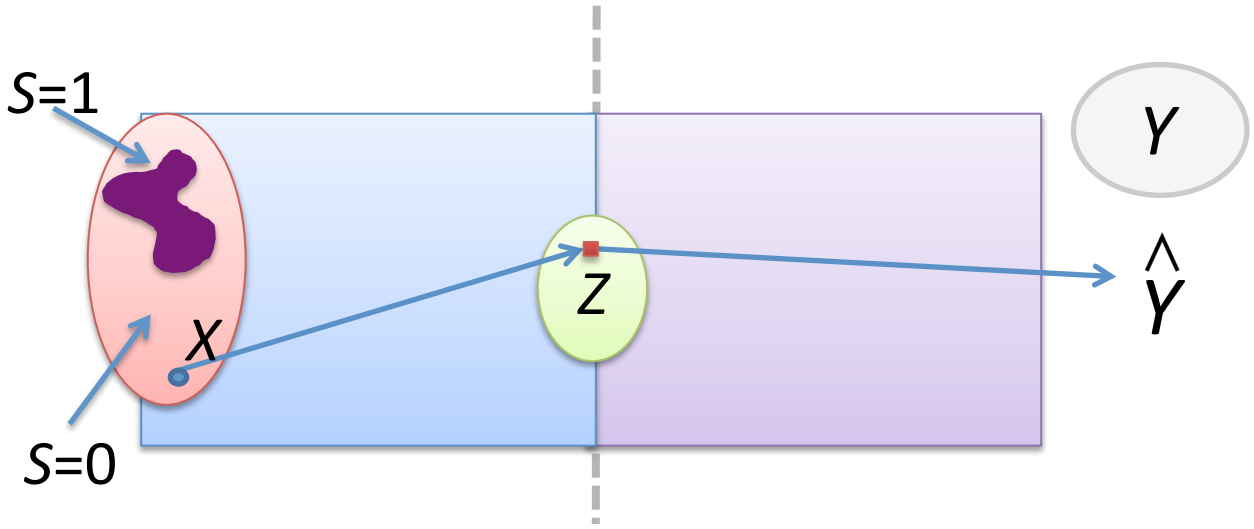
$$L_z = \sum_k |P_k^+ - P_k^-| \qquad P_k^+ = P(Z = k|S = 1)$$



Sequence of Models

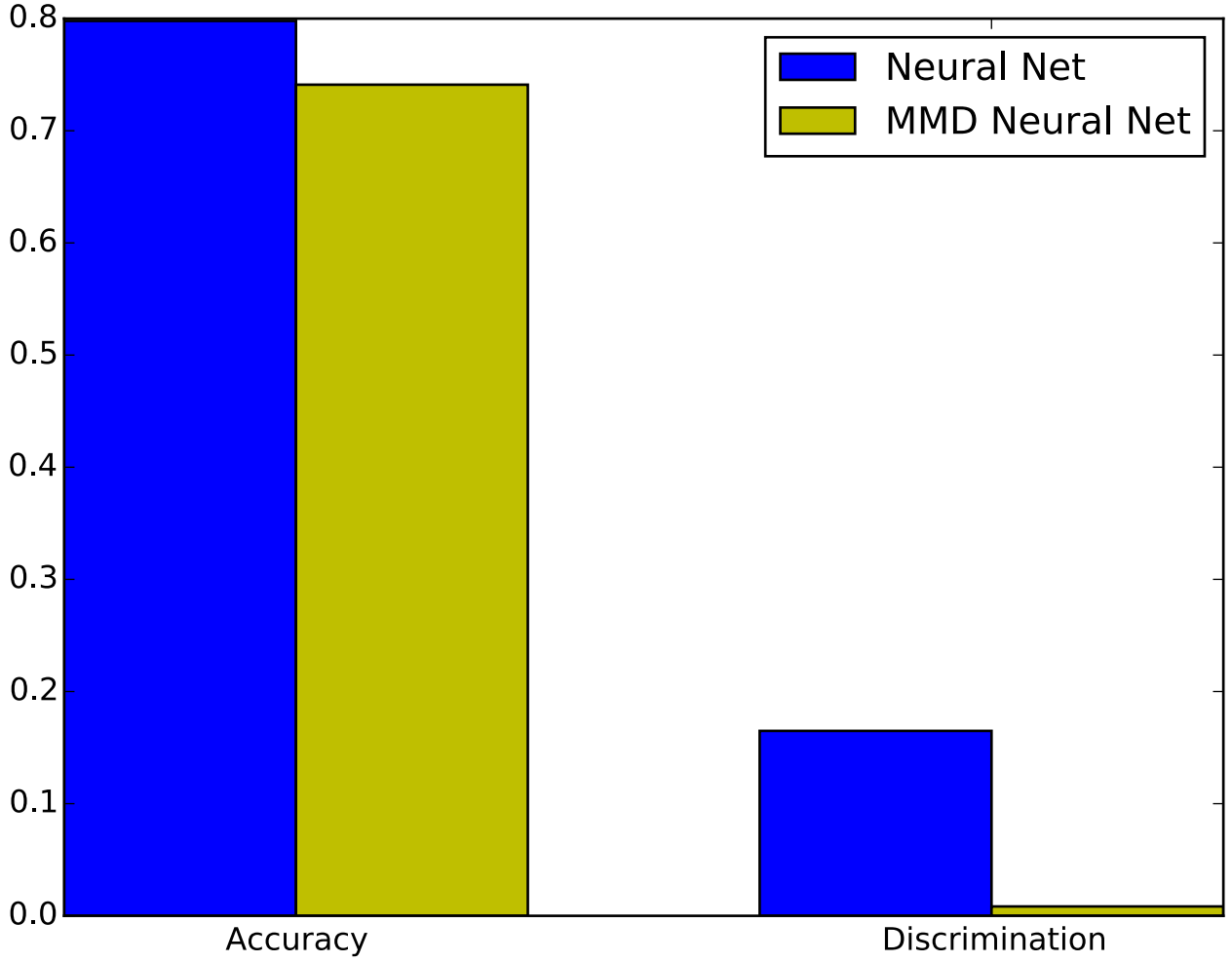
(3). Match higher-order moments, continuous Z:

$$MMD(P(Z|S=1); P(Z|S=0))$$



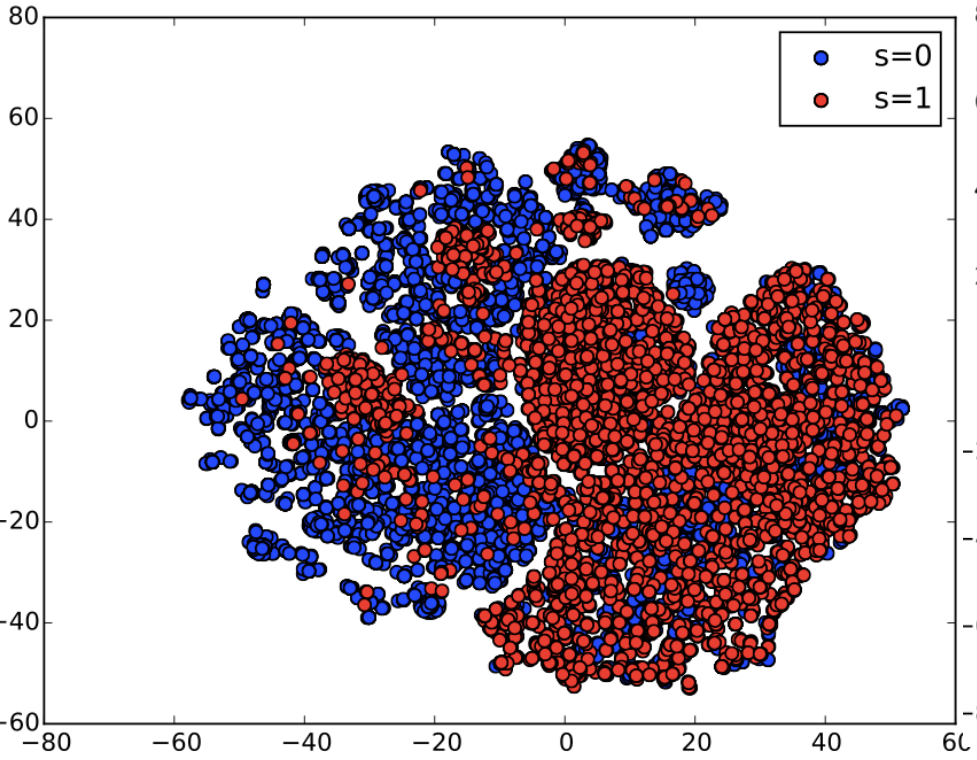
Results: Richer Z

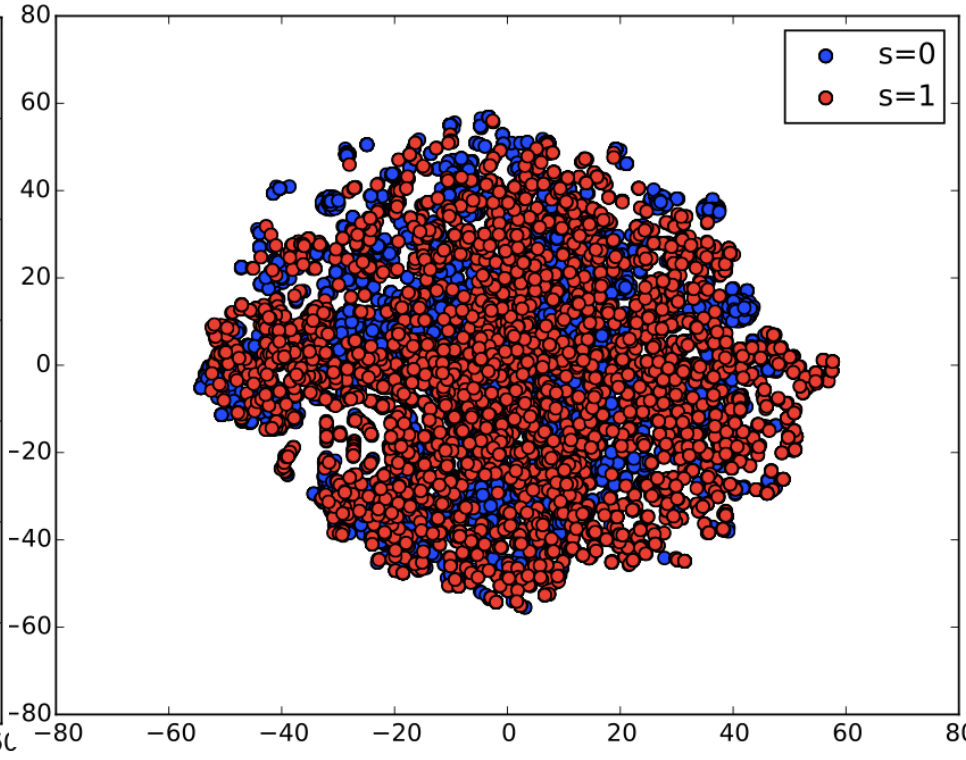
Compare neural network (50 dim Z) with/without MMD:



Results: Richer Z

Compare neural network (50 dim Z) with/without MMD:





Open Problems

- Further extensions of intermediate representations: build more expressive mappings, but still preserve information in X while losing information about S

- Other applications: Eliminating demographic discrimination is deciding who should get transplant surgery

- Refining definition, objectives of fairness:
  work with legal scholars, public policy experts
  - Is statistical parity, or quotas, the right goal?
  - Would individual fairness, with appropriate metric, suffice?

Thanks!