APPROACHES TO FAIR CLASSIFICATION

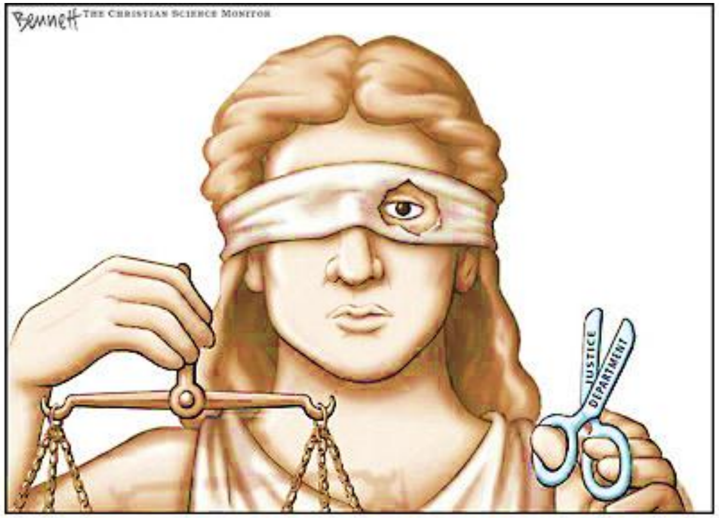
TONIANN PITASSI   RICHARD ZEMEL

CSC 2541

OCTOBER 1, 2019

FAIRNESS THROUGH AWARENESS

Dwork, Hardt, Pitassi, Reingold, Zemel, 2012

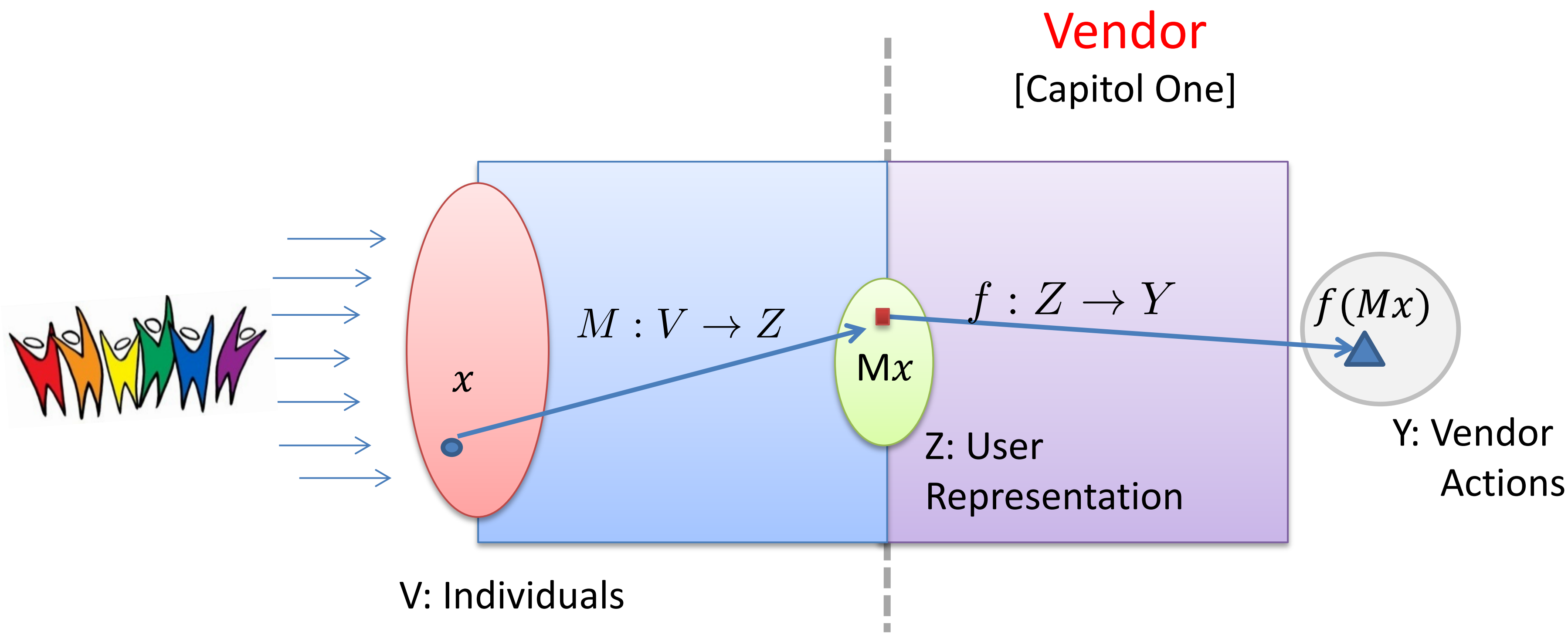


Goal: Assign each individual *a* representation *by being aware of membership in group A*

(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



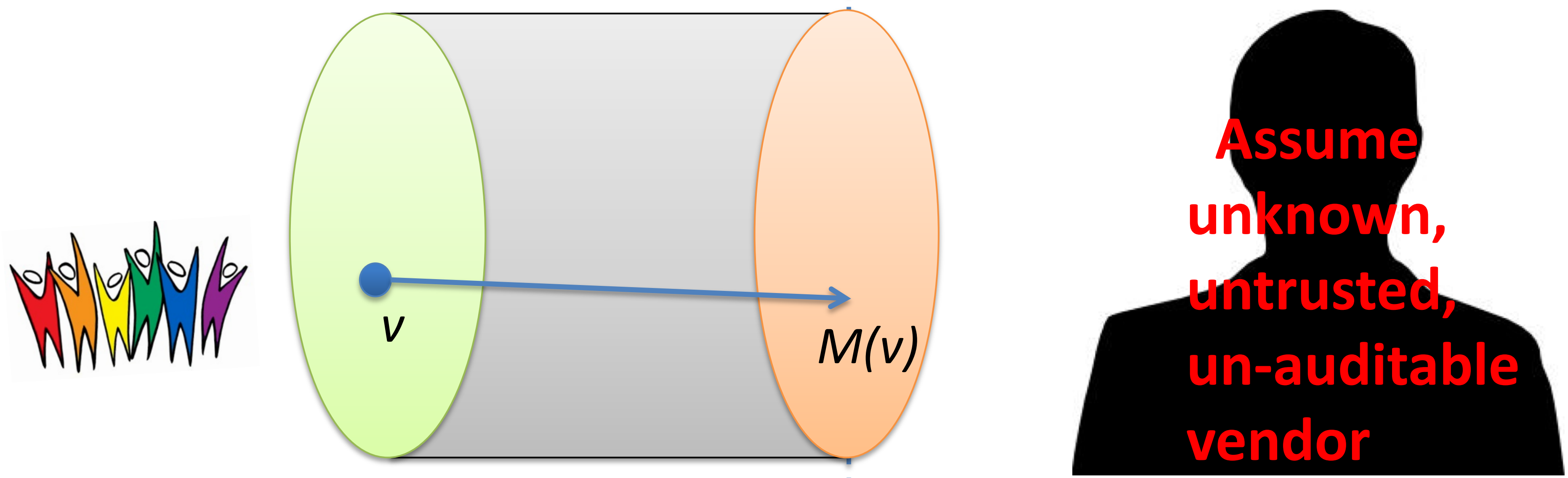
Our goal:

Achieve Fairness in the representation step

Ad Network (with society oversight)

Ad Network

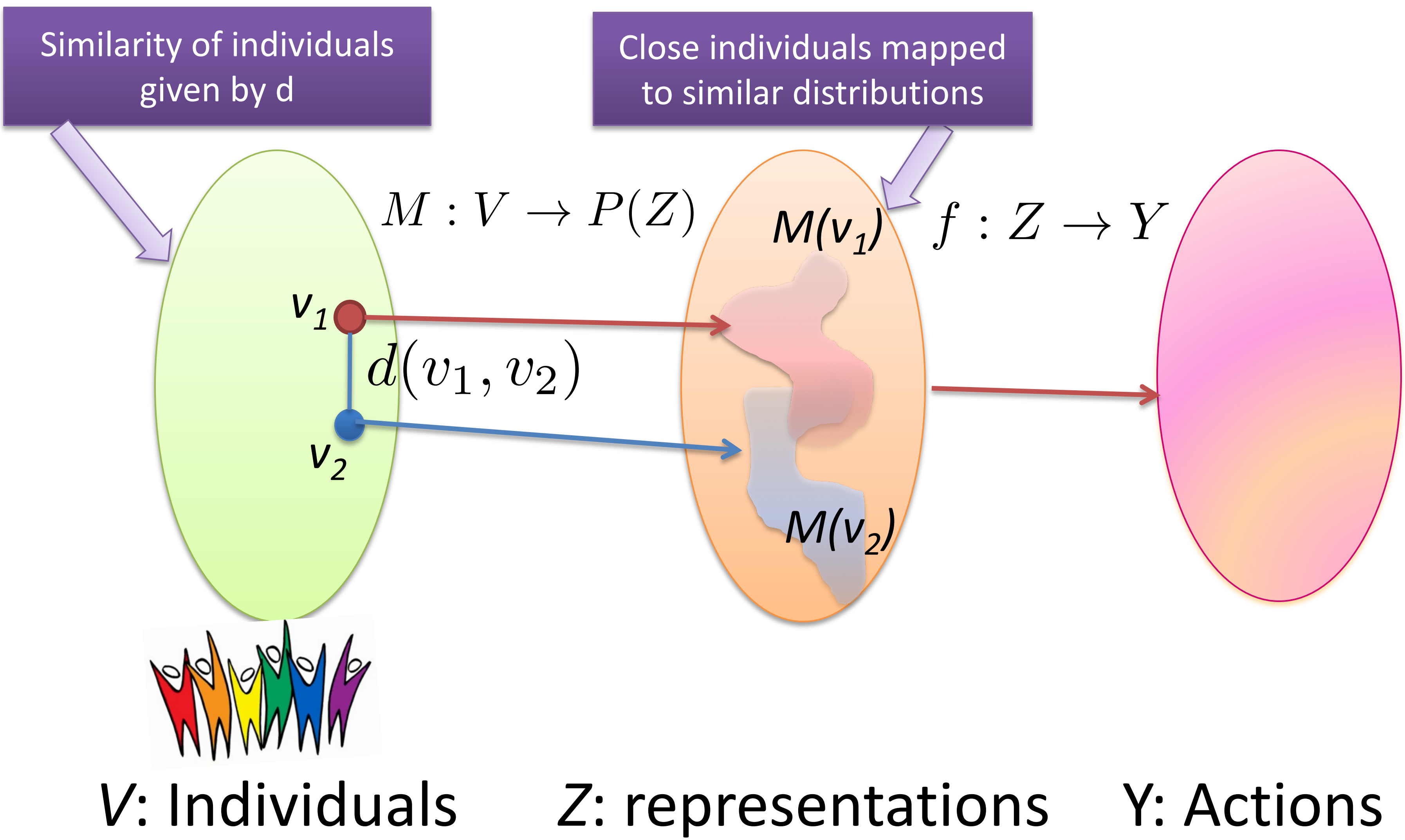
$M : V \to Z$



V: Individuals

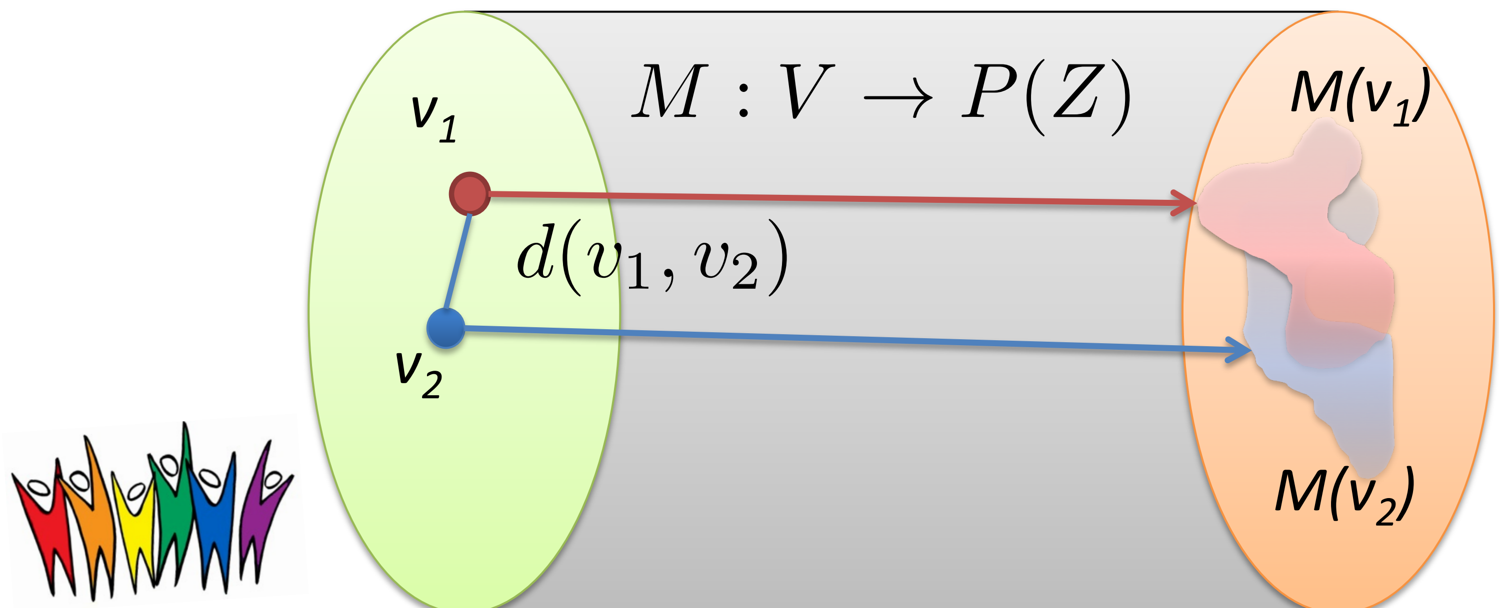
Z: representation

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



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

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



V: Individuals $\quad$ 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

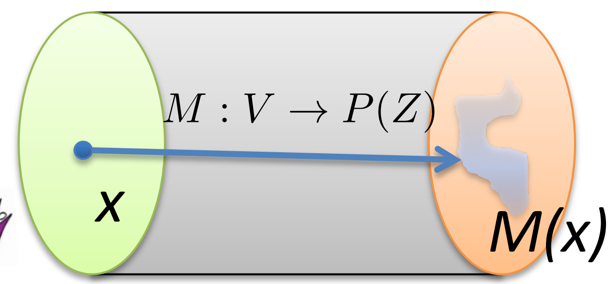
An Algorithm for Fair Classification

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

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

Efficient
Procedure

d-fair mapping M

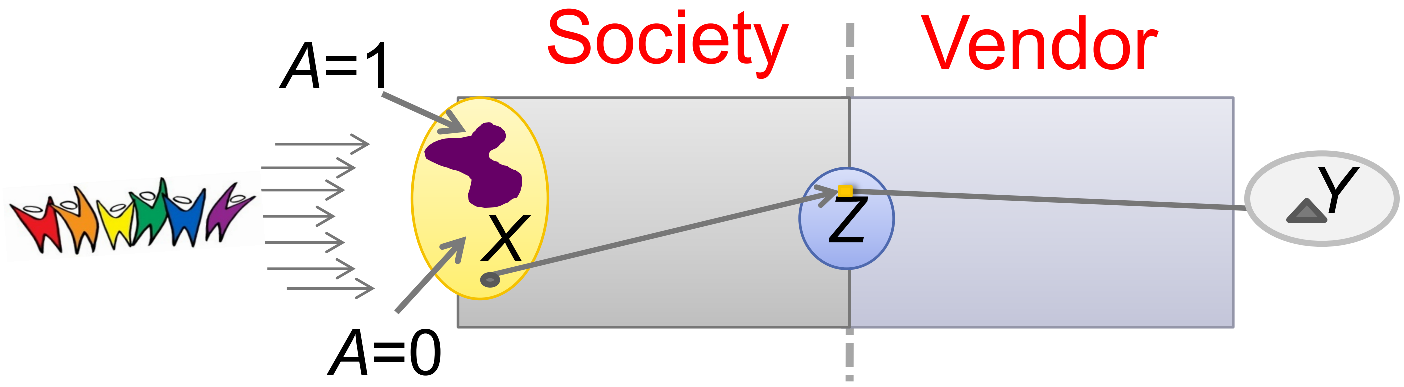


V: Individuals *Z:* Encodings

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:

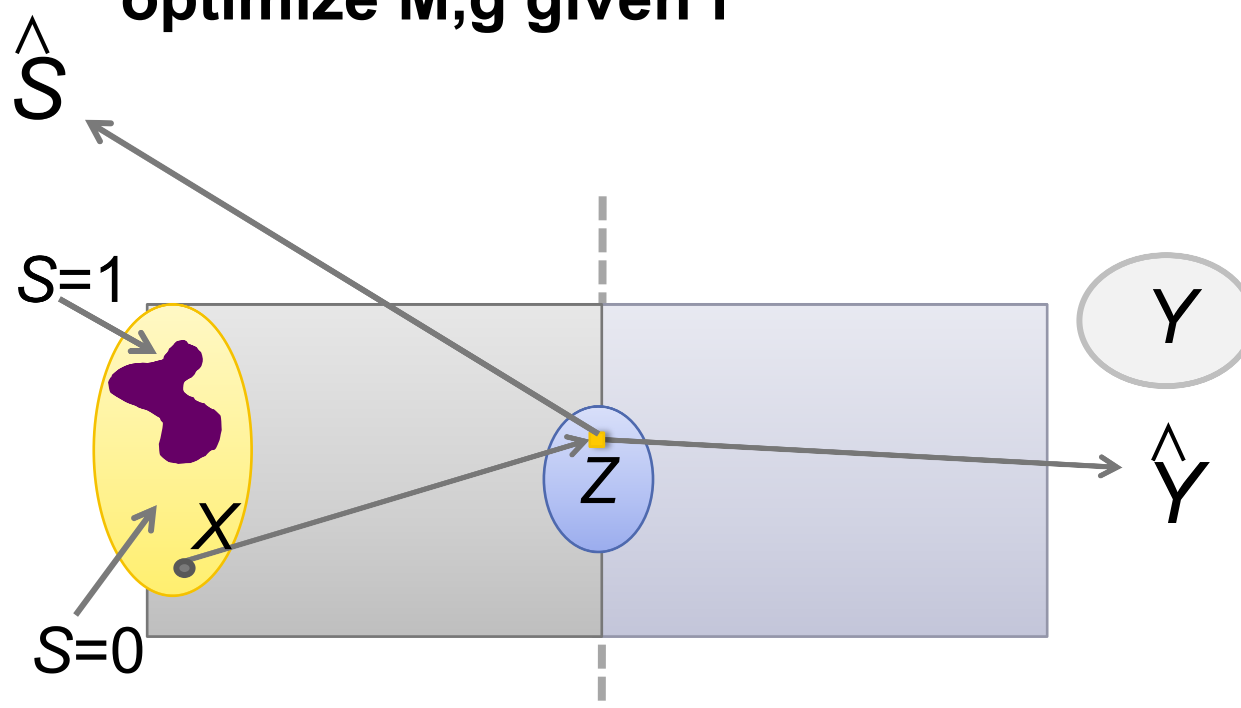
min. |f(Z) – Y|;   max. |g(Z) – S| (thwart adversary)

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

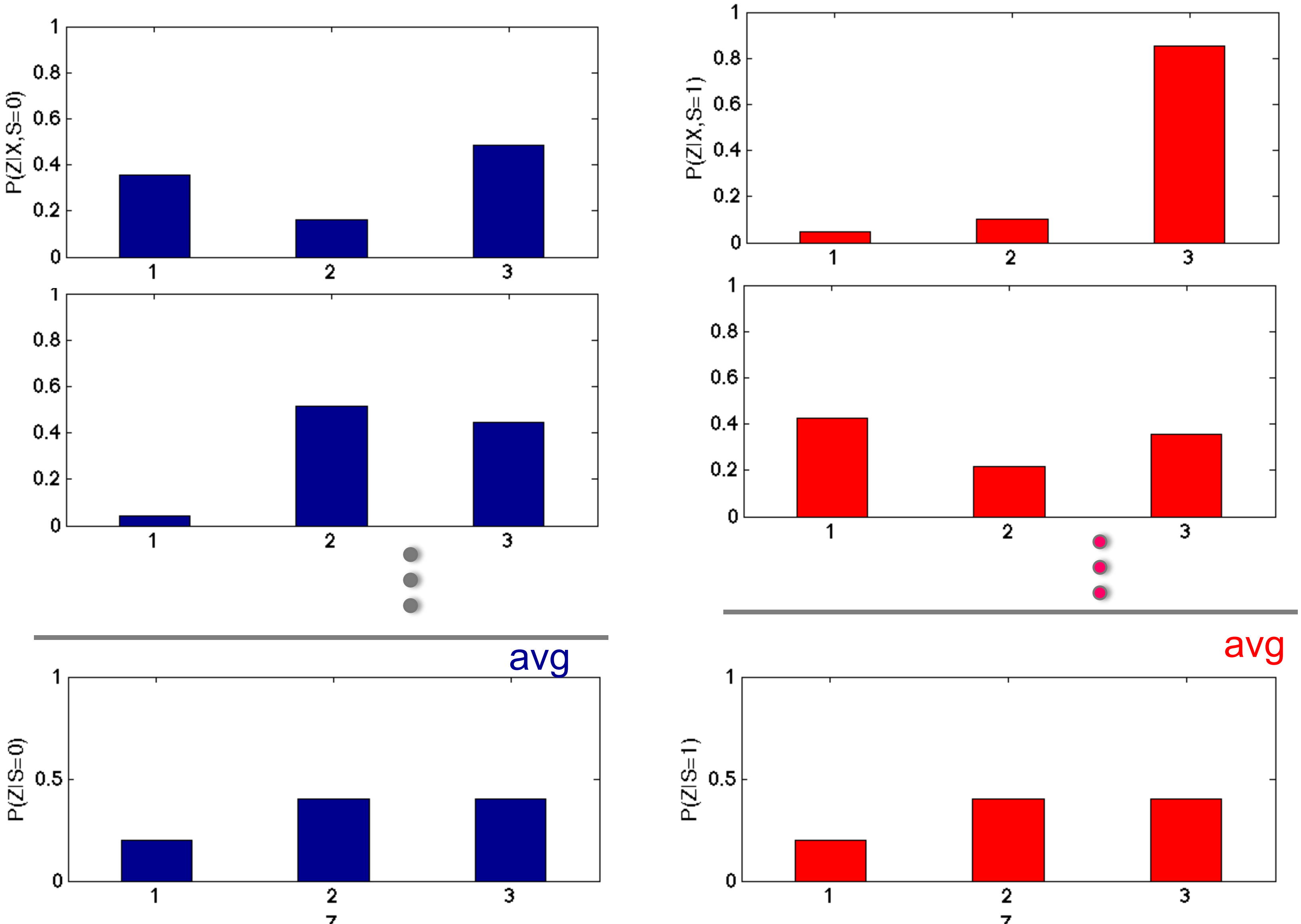
$$P^+_{n,k} = P(Z = k|\mathbf{x}, S = 1) = \frac{\exp(\mathbf{x}_n^T \mathbf{w}_k^+)}{\sum_{k'} \exp(\mathbf{x}_n^T \mathbf{w}_{k'}^+)}$$

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

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

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

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**

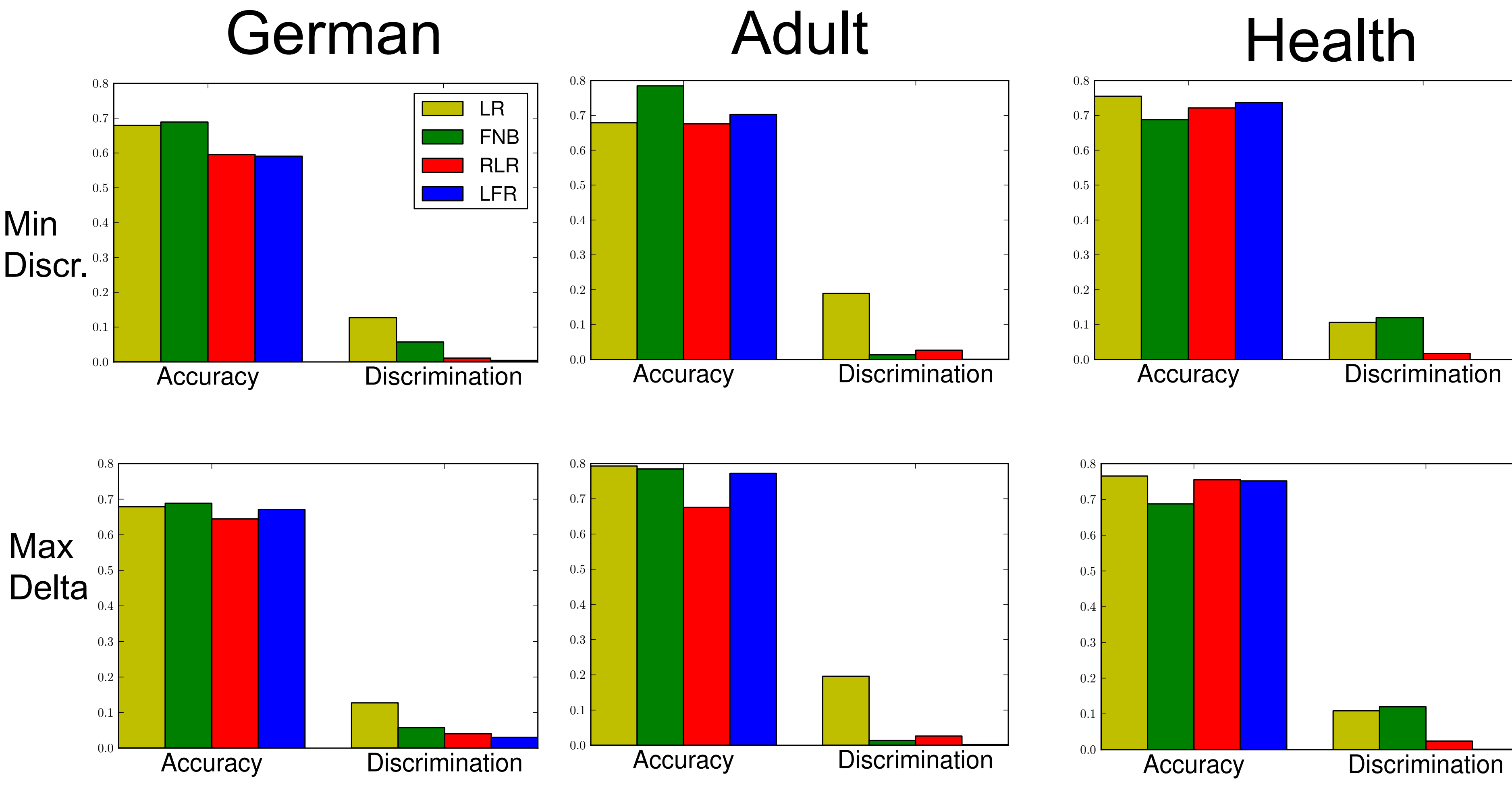
$$yDiscrim = |\frac{\sum_{n:s_n=1} \hat{y}_n}{\sum_{n:s_n=1} 1} - \frac{\sum_{n:s_n=0} \hat{y}_n}{\sum_{n:s_n=0} 1}|$$

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]

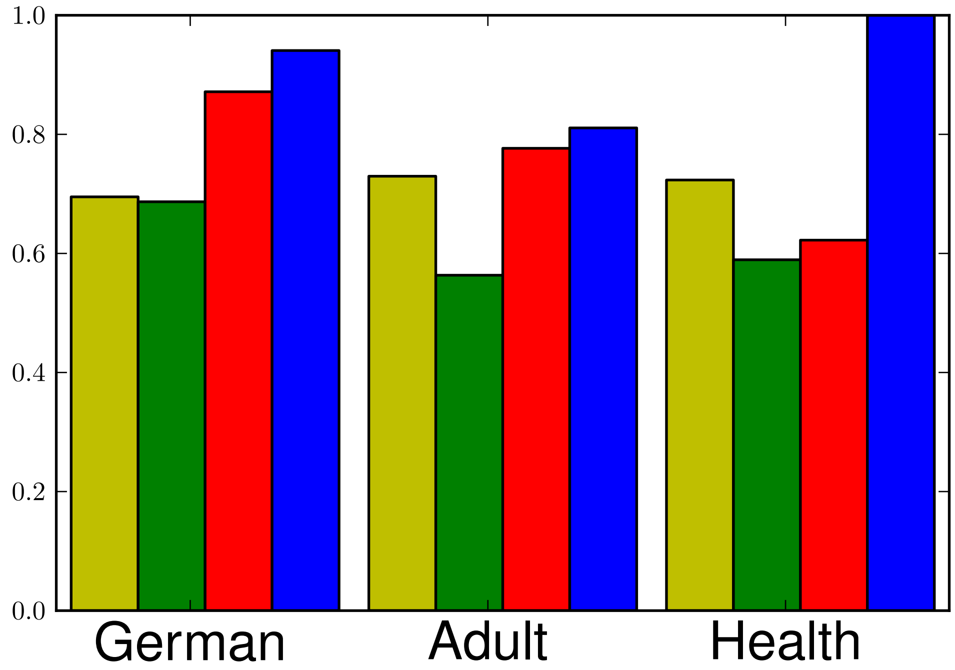
EXPERIMENTAL RESULTS



RESULTS: INDIVIDUAL FAIRNESS

Consistency:

$$yNN = 1 - \frac{1}{N}\sum_{n} |\, \hat{y}_n - \frac{1}{k}\sum_{j \in kNN(\mathbf{x_n})} \hat{y}_j \,|$$



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

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

REPRESENTATIONS BEYOND CLUSTERS

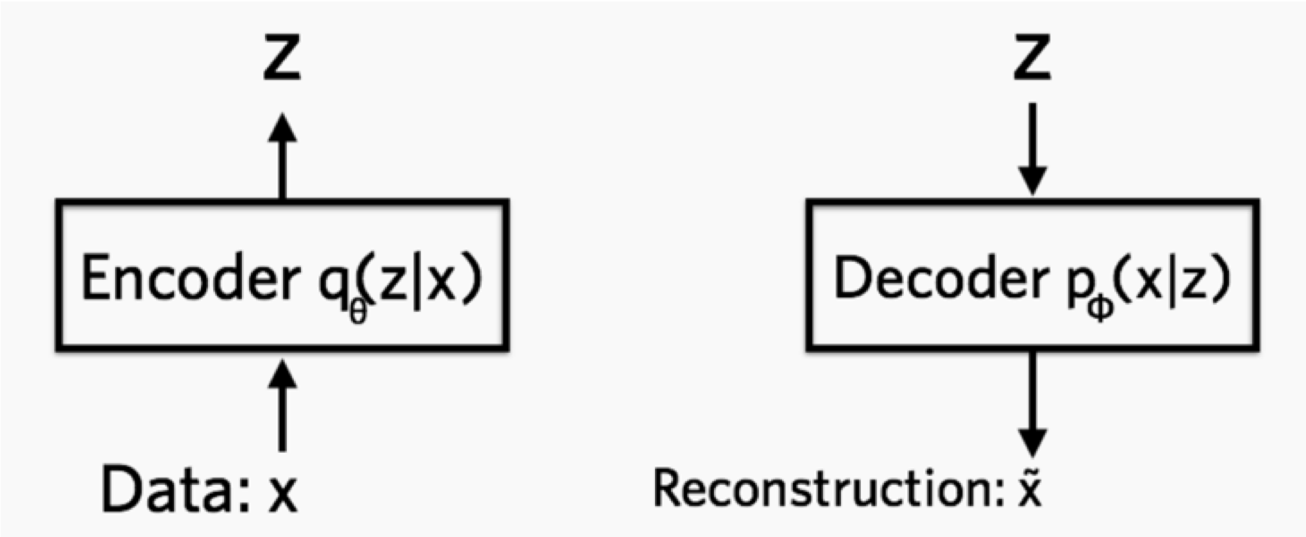
Aim: Replace discrete representation with continuous, multi-dimensional Z

Allow more flexible, nuanced representations

Bring ML arsenal to bear: powerful methods for mapping, embedding in vector spaces: Variational Auto Encoders (VAE)

How to maintain statistical parity in learned representations?

VAE



Re-formulation of autoencoders:

- Each input encoded into a distribution in latent space
- Output prediction obtained by sampling from distribution, mapping through decoder

Allows maximum-likelihood based density modelling:

$$\log p_\theta(\mathbf{x}) \geq \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x}\,|\,\mathbf{z})\right] - D_{KL}\left(q_\phi(\mathbf{z}\,|\,\mathbf{x}) \parallel p(\mathbf{z})\right)$$

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. [Gretton 2010]

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

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

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

- Our idea: learn to make two distributions indistinguishable

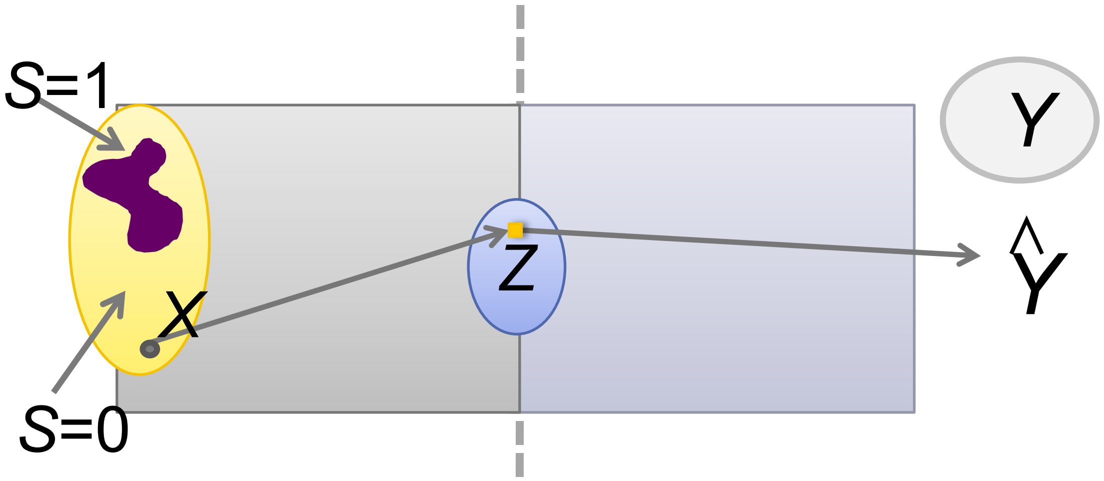
  ➔ small MMD!

VARIATIONAL FAIR AUTOENCODER

VAE with regularizer on latent representations

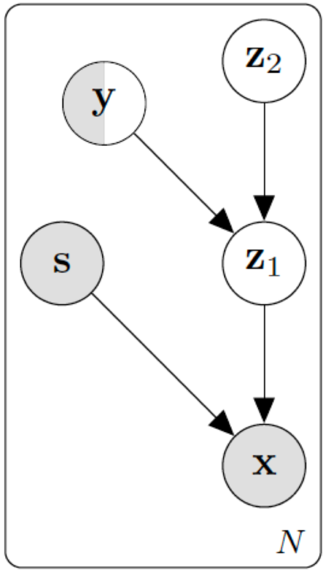
Match higher-order moments, continuous *Z*:

$$\ell_{\text{MMD}}(\mathbf{Z}_{1\mathbf{s}=0}, \mathbf{Z}_{1\mathbf{s}=1}) = \| \, \mathbb{E}_{\tilde{p}(\mathbf{x}|\mathbf{s}=0)}[\mathbb{E}_{q(\mathbf{z}_1|\mathbf{x},\mathbf{s}=0)}[\psi(\mathbf{z}_1)]] - E_{\tilde{p}(\mathbf{x}|\mathbf{s}=1)}[\mathbb{E}_{q(\mathbf{z}_1|\mathbf{x},\mathbf{s}=1)}[\psi(\mathbf{z}_1)]] \|^2$$



VARIATIONAL FAIR AUTOENCODER

Extend VAE to include some labels *y* (semi-supervised VAE [Kingma & Welling, 2014]) and "nuisance variable" *s*



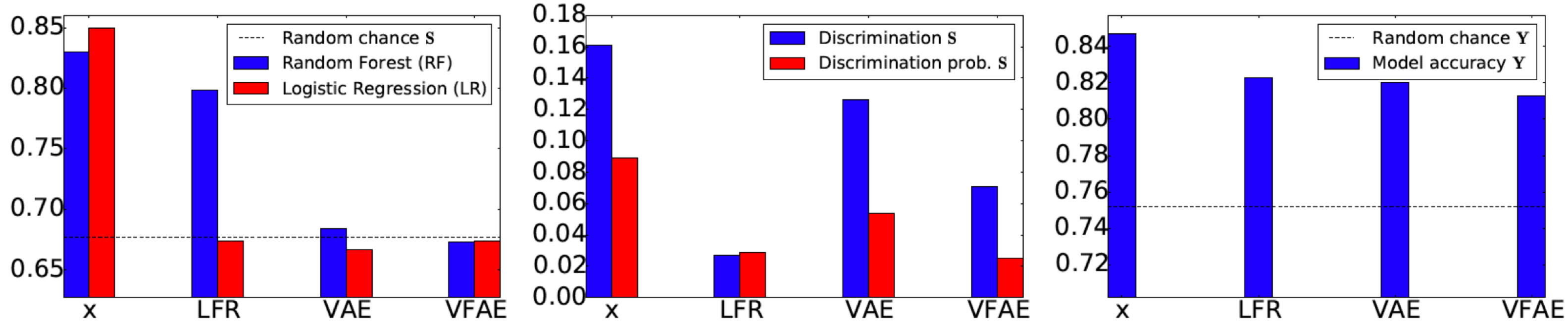
Objective -- maximize:

$$\sum_{n=1}^{N_s} \mathbb{E}_{q_\phi(\mathbf{z}_{1n}|\mathbf{x}_n,\mathbf{s}_n)}[-KL(q_\phi(\mathbf{z}_{2n}|\mathbf{z}_{1n},\mathbf{y}_n)||p(\mathbf{z}_2)) + \log p_\theta(\mathbf{x}_n|\mathbf{z}_{1n},\mathbf{s}_n)] + \\ + \mathbb{E}_{q_\phi(\mathbf{z}_{2n}|\mathbf{z}_{1n},\mathbf{y}_n)}[-KL(q_\phi(\mathbf{z}_{1n}|\mathbf{x}_n,\mathbf{s}_n)||p_\theta(\mathbf{z}_{1n}|\mathbf{z}_{2n},\mathbf{y}_n))]$$

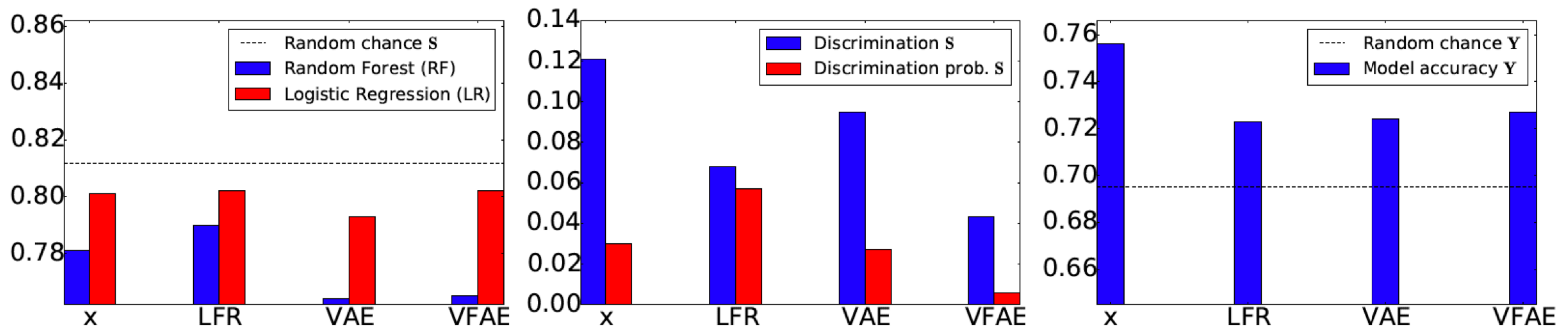
Add for labeled set: $\sum_{n=1}^{N} \mathbb{E}_{q(\mathbf{z}_{1n}|\mathbf{x}_n,\mathbf{s}_n)}[-\log q_\phi(\mathbf{y}_n|\mathbf{z}_{1n})]$

unlabeled set: $\sum_{m=1}^{M} \mathbb{E}_{q_\phi(\mathbf{z}_{1m}|\mathbf{x}_m,\mathbf{s}_m)}[-KL(q(\mathbf{y}_m|\mathbf{z}_{1m})||p(\mathbf{y}_m))$

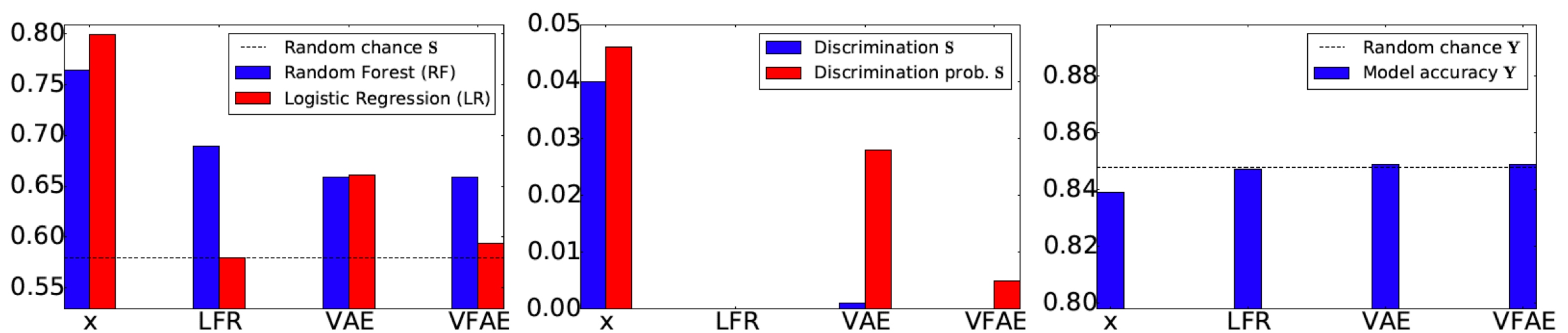
RESULTS



(a) Adult dataset



(b) German dataset



(c) Health dataset

RESULTS

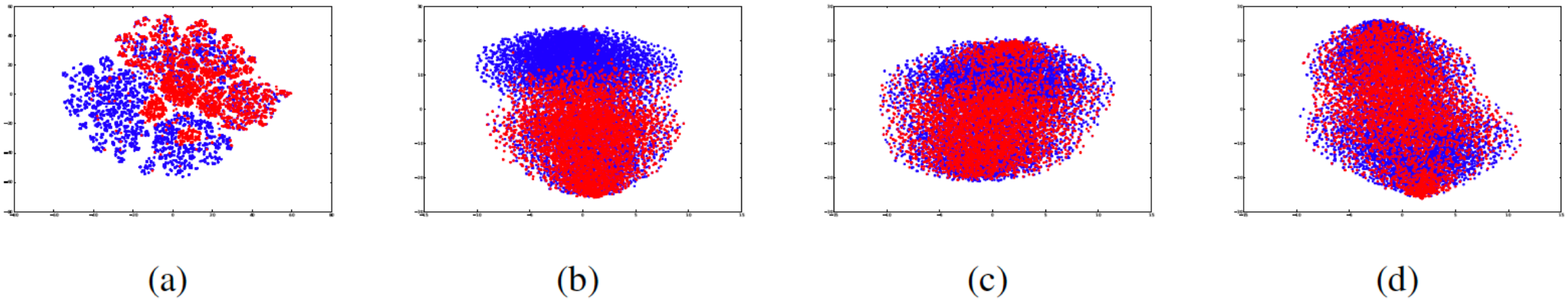


Figure 4: t-SNE (van der Maaten, 2013) visualizations from the Adult dataset on: (a): original $\mathbf{x}$, (b): latent $\mathbf{z}_1$ without s and MMD, (c): latent $\mathbf{z}_1$ with s and without MMD, (d): latent $\mathbf{z}_1$ with s and MMD. Blue colour corresponds to males whereas red colour corresponds to females.

ADAPTING THE FRAMEWORK

The same idea has many other useful applications, e.g.,

- Eliminating demographic discrimination in deciding who should get transplant surgery
- Removing confounds, such as which scanner produced a medical image

Key: Learning to make two (or more) distributions indistinguishable

DOMAIN ADAPTATION

Natural fit: domain adaptation

Make feature representations for source and target domain data indistinguishable

Sentiment classification

- Product reviews (text, tf-idf on words & bigrams)
- Labeled data from source domain, unlabeled data from target domain

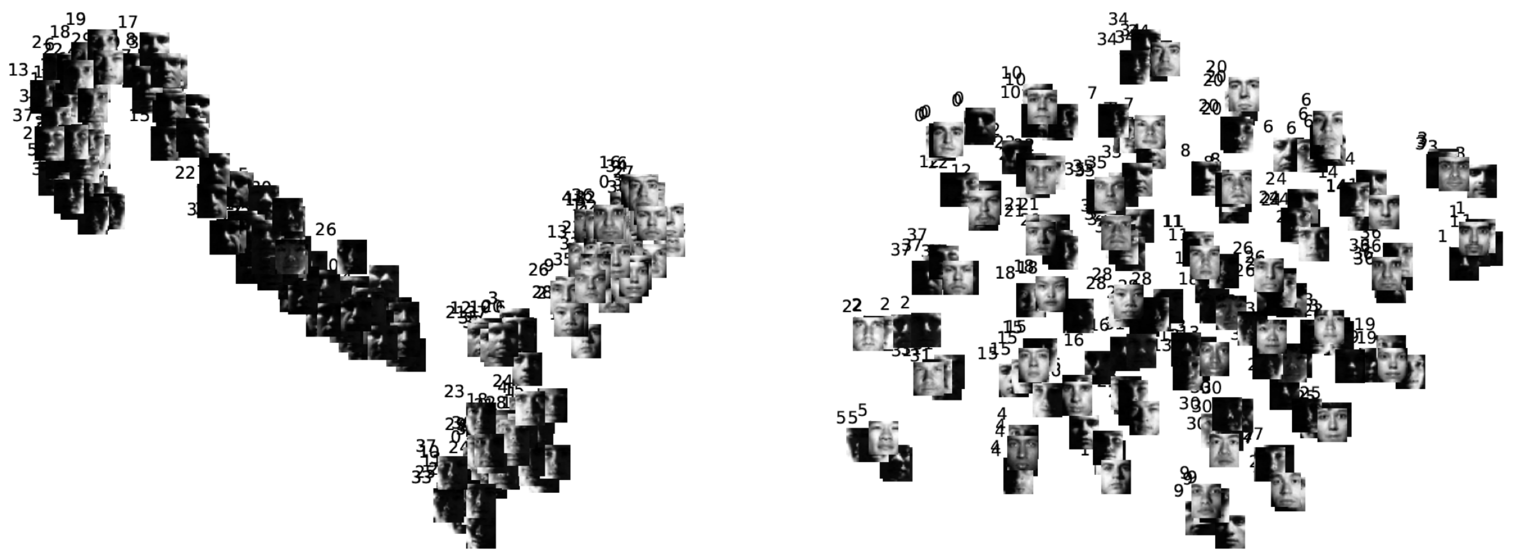
Source - Target	S		Y	
	RF	LR	VFAE	DANN
books - dvd	0.535	0.564	**0.799**	0.784
books - electronics	0.541	0.562	**0.792**	0.733
books - kitchen	0.537	0.583	**0.816**	0.779
dvd - books	0.537	0.563	**0.755**	0.723
dvd - electronics	0.538	0.566	**0.786**	0.754
dvd - kitchen	0.543	0.589	**0.822**	0.783
electronics - books	0.562	0.590	**0.727**	0.713
electronics - dvd	0.556	0.586	**0.765**	0.738
electronics - kitchen	0.536	0.570	0.850	**0.854**
kitchen - books	0.560	0.593	**0.720**	0.709
kitchen - dvd	0.561	0.599	0.733	**0.740**
kitchen - electronics	0.533	0.565	0.838	**0.843**

LEARNING INVARIANT FEATURES

If we have labeled data from all domains, factoring out unwanted domain bias still leads to better generalization.

Make the learned representations invariant to unwanted transformation / variation / bias.

Example: Face identification under different lighting conditions



ADVERSARIAL FAIR LEARNING

Rather than using MMD to ensure learned representation is fair, can use adversarial approach

Adversary takes latent representation (here *R*) as input and attempts to predict *S,* then model minimizes*:*

$$D_{\theta,\phi}(R,S) = \mathop{\mathbb{E}}_{X,S} S \cdot \log\left(\mathrm{Adv}(R)\right) + (1-S)\cdot \log\left(1-\mathrm{Adv}(R)\right)$$

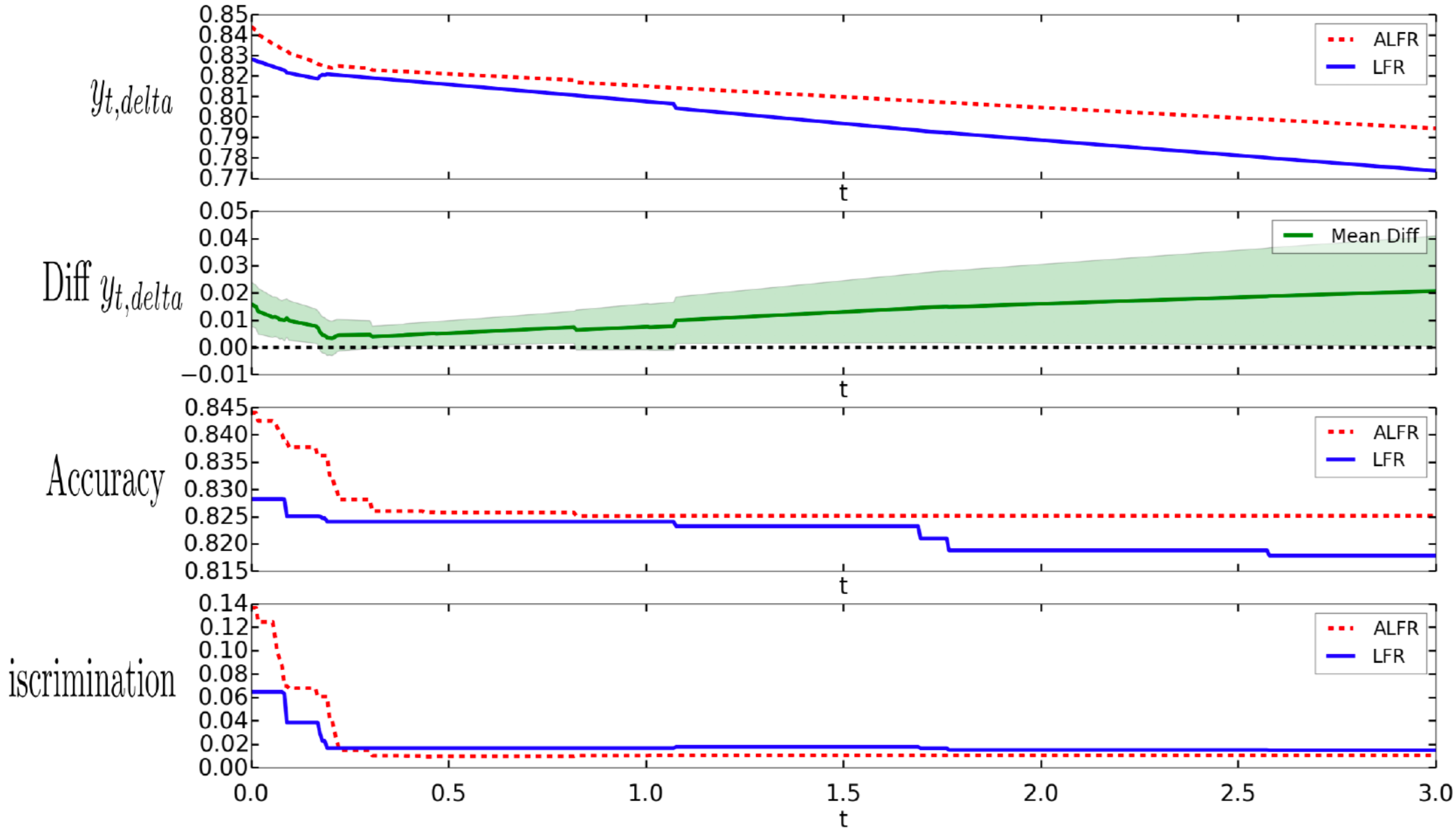
Combine with reconstruction and classification losses to ensure representation retains info about *X,Y*

$$C_\theta(X,R) = \mathop{\mathbb{E}}_{X} \|X - \mathrm{Dec}(R)\|_2^2$$

$$E_\theta(R,S) = -\mathop{\mathbb{E}}_{X,Y} Y \cdot \log\left(\mathrm{Pred}(R)\right) + (1-Y)\cdot \log\left(1-\mathrm{Pred}(R)\right)$$

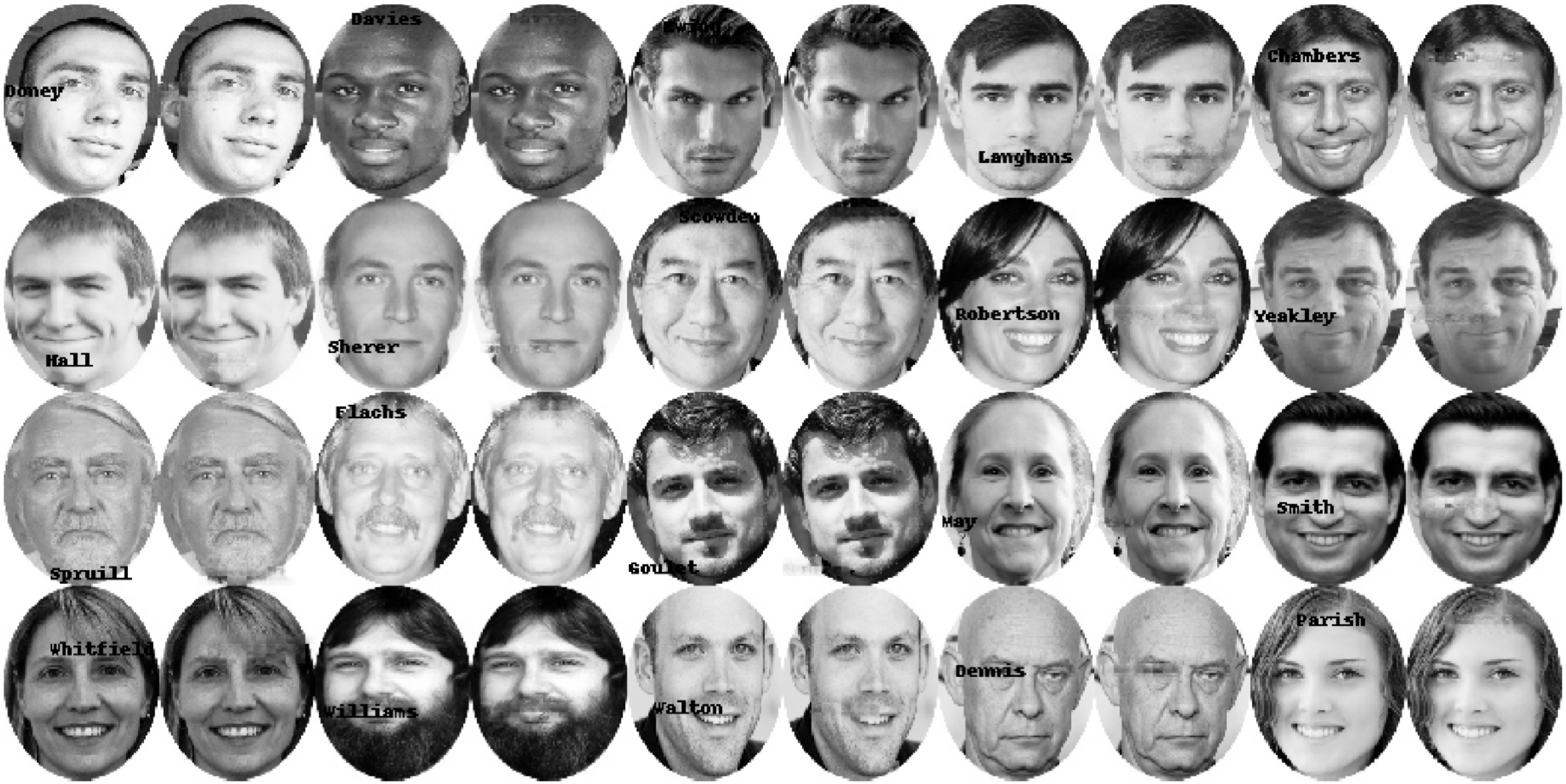
Censoring Representations with an Adversary: Edwards & Storkey, 2015

RESULTS



$y_{t,delta}$
0.85
0.84
0.83
0.82
0.81
0.80
0.79
0.78
0.77
ALFR
LFR
t
Diff $y_{t,delta}$
0.05
0.04
0.03
0.02
0.01
0.00
−0.01
Mean Diff
t
Accuracy
0.845
0.840
0.835
0.830
0.825
0.820
0.815
ALFR
LFR
t
iscrimination
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
ALFR
LFR
0.0
0.5
1.0
1.5
2.0
2.5
3.0
t

RESULTS


Doney
Davies
Langhans
Chambers
Hall
Sherer
Scowden
Robertson
Yeakley
Spruill
Flachs
Goulet
May
Smith
Whitfield
Williams
Walton
Dennis
Parish

EQUALIZED ODDS / OPPORTUNITY

Both VFAE and AFLR define fairness as statistical parity

Problems with demographic/statistical parity:

- Coarse measure, not about individuals
- May entail large loss in accuracy

Alternative definition: equal opportunity [Hardt, Price, Srebro, 2016]

- Encourage perfect prediction
- But ensure that the prediction errors are balanced between the groups

$$\Pr\left\{\widehat{Y} = 1 \mid A = 0, Y = y\right\} = \Pr\left\{\widehat{Y} = 1 \mid A = 1, Y = y\right\}, \quad y \in \{0, 1\}$$

BACK TO FAIR REPRESENTATIONS

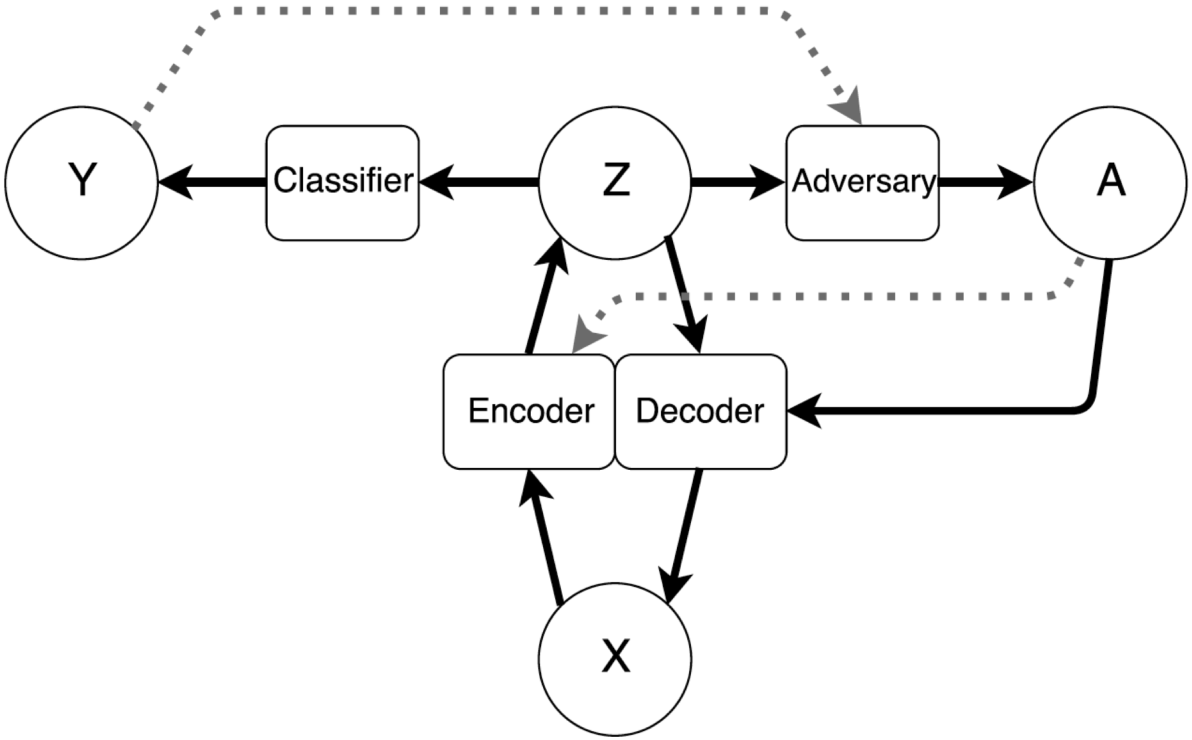
- Minimize unfair targeting of disadvantaged groups by vendors (worse lines of credit, lower paying jobs)
- Aim: form a data representation that ensures fair classifications downstream

- Consider two types of unfair vendors:
  1. The **indifferent** vendor: does not care about fairness, only maximizes utility
  2. The **malicious** vendor: doesn’t care about utility, discriminates unfairly

- Good fit to adversarial learning scheme

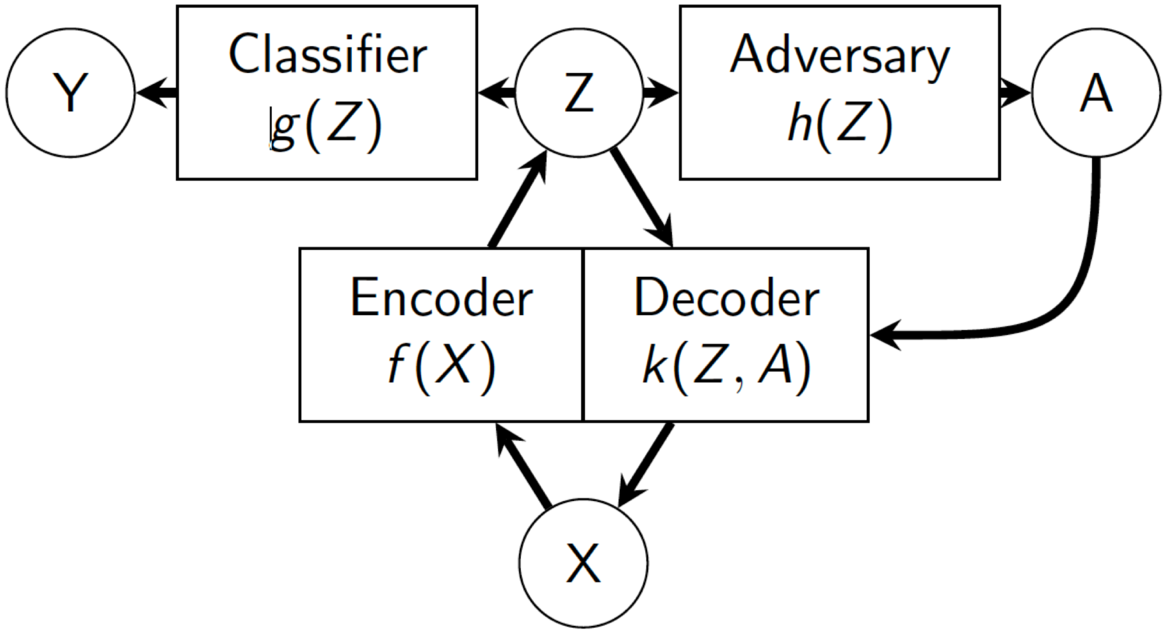
LEARNING ADVERSARIALLY FAIR TRANSFERABLE REPRESENTATIONS

Madras, Creager, Pitassi, Zemel, 2018



- The classifier is indifferent vendor, forcing the encoder to make the representations useful
- The adversary is the malicious vendor, forcing the encoder to hide the sensitive attributes in the representations

ADVERSARIAL LEARNING IN LAFTR

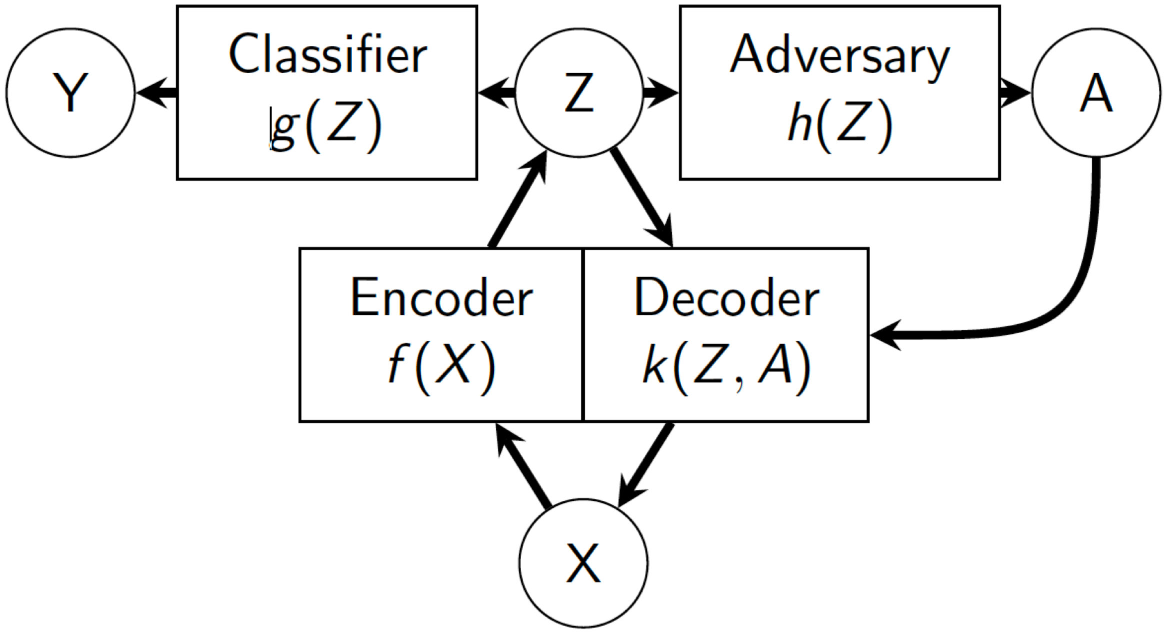


- Our game: encoder-decoder-classifier vs. adversary
- Aim: Learn fair encoder

$$\underset{f,g,k}{\text{minimize}}\ \underset{h}{\text{maximize}}\ \mathbb{E}_{X,Y,A}\left[\mathcal{L}(f, g, h, k)\right]$$

$$\mathcal{L}(f, g, h, k) = \alpha\mathcal{L}_{Class} + \beta\mathcal{L}_{Dec} - \gamma\mathcal{L}_{Adv}$$

ADVERSARIAL OBJECTIVES



Choice of adversarial objective depends on fairness desideratum

- Demographic parity: $\mathcal{L}_{DP}(h) = \sum_{i \in \{0,1\}} \frac{1}{|\mathcal{D}_i|} \sum_{(x,a) \in \mathcal{D}_i} |h(f(x)) - a|$
- Equalized odds: $\mathcal{L}_{EO}(h) = \sum_{i,j \in \{0,1\}^2} \frac{1}{|\mathcal{D}_i^j|} \sum_{(x,a,y) \in \mathcal{D}_i^j} |h(f(x), y) - a|$
- Equal Opportunity: $\mathcal{L}_{EOpp}(h) = \sum_{i \in \{0,1\}} \frac{1}{|\mathcal{D}_i^1|} \sum_{(x,a) \in \mathcal{D}_i^1} |h(f(x)) - a|$

FROM ADVERSARIAL OBJECTIVES TO FAIRNESS DEFINITIONS

In general: pick the right adversarial loss, encourage the right conditional independencies

- Demographic parity encourages $Z \perp A$ to fool adversary
- Equalized odds encourages $Z \perp A \mid Y$ to fool adversary
- Equal opportunity encourages $Z \perp A \mid Y = 1$ to fool adversary

Note that independencies of $Z = f(x)$ also hold for predictions $\hat{Y} = g(Z)$

We show: In the adversarial limit, these objectives guarantee these fairness metrics!

- The key is to connect predictability of A by the adversary $h(Z)$ to unfairness in the classifier $g(Z)$

EXPERIMENTS

Datasets

1. Adult Income

    Size: 45,222 instances, 14 attributes

    Task: predict whether or not annual income > 50K

    Sensitive feature: Gender

2. Heritage Health

    Size: 147,473 instances, 139 attributes

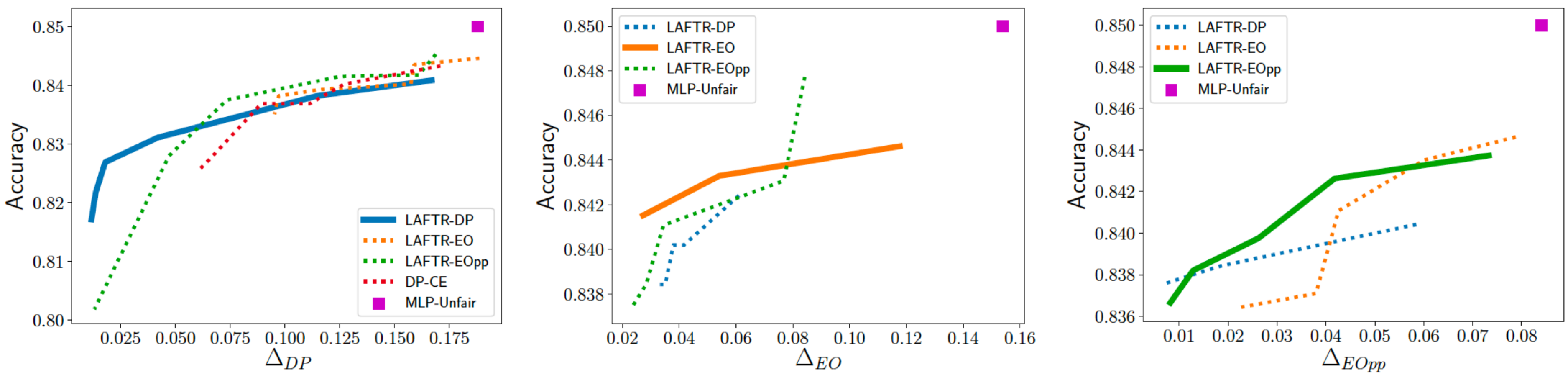
    Task: predict patient's Charlson Index (co-morbidity)

    Sensitive feature: Age

Models

Encoder, classifier, adversary: each single hidden-layer MLP (8; 20 hidden units)

RESULTS: FAIR CLASSIFICATION



- Train with 2-step process to simulate owner → vendor framework
- Tradeoffs between accuracy and fairness metrics produced by different LAFTR loss functions
- Achieves best solutions, wrt fairness-accuracy tradeoff

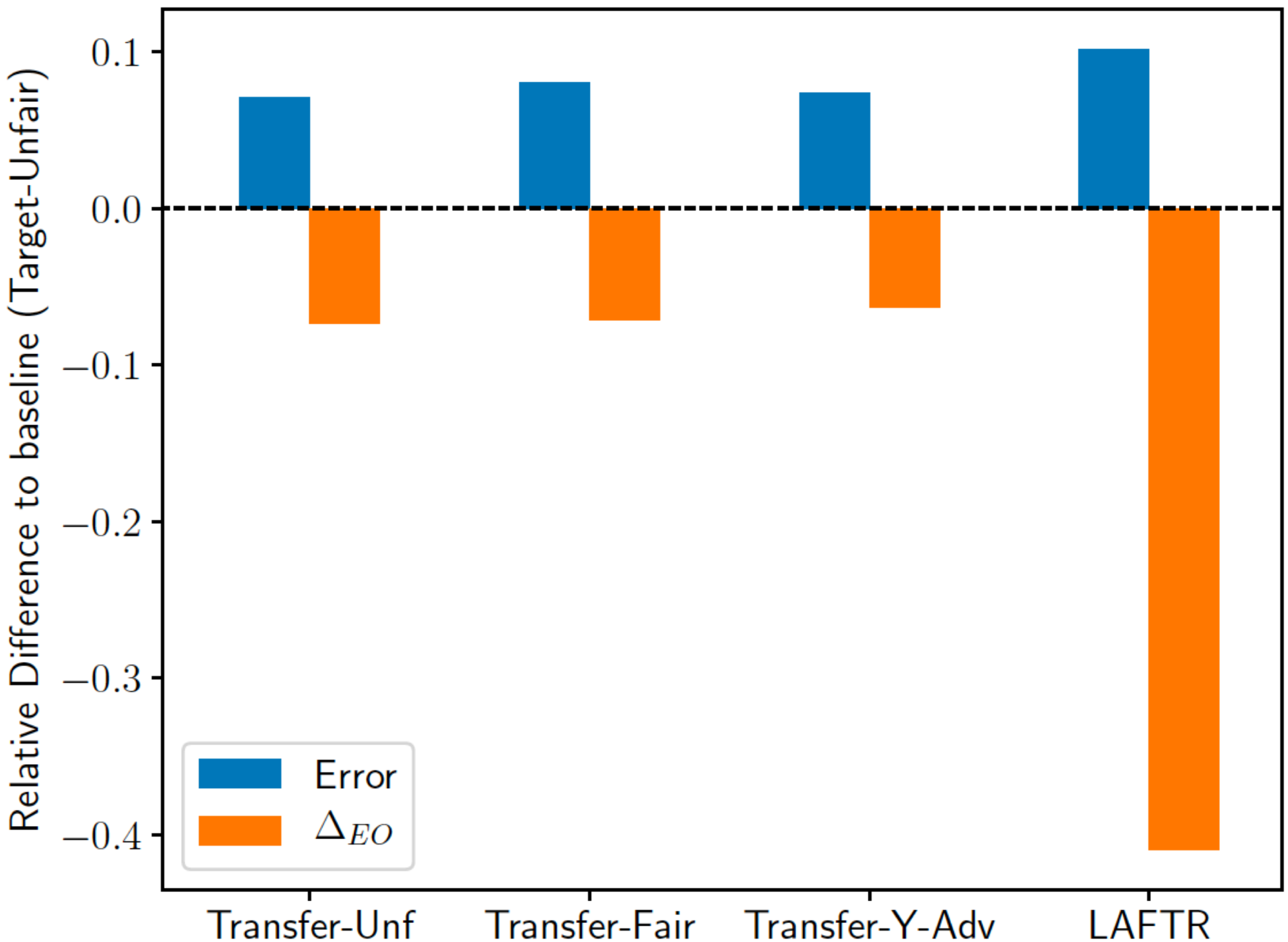
RESULTS: FAIRNESS METRICS

Method	Δ_{DP}	Δ_{EO}	Δ_{EOpp}	Acc.
MLP (unfair)	0.381	0.476	0.231	**0.785**
LAFTR-EO	0.152	**0.050**	0.036	0.763
	0.143	**0.052**	0.032	0.752
LAFTR-EOpp	0.087	0.092	**0.010**	0.742
	0.113	0.063	**0.024**	0.735
LAFTR-DP	**0.041**	0.140	0.025	0.731
	0.002	0.196	0.031	0.728

SETUP: FAIR TRANSFER LEARNING

- Downstream vendors will have unknown prediction tasks
- Does fairness transfer?
- We test this as follows:
  1. Train encoder f on data X, with label Y
  2. Freeze encoder f
  3. On new data X', train classifier on top of $f(X')$, with new task label Y'
  4. Observe fairness and accuracy of this new classifier on new task Y'
- Compare LAFTR encoder f to other encoders
- We use Heritage Health dataset
  - Y is Charlson comorbidity index > 0
  - Y' is whether or not a certain type of insurance claim was made
  - Check for fairness w.r.t. age

RESULTS : FAIR TRANSFER LEARNING



Fair transfer learning on Health dataset. Down is better in both metrics.

ALTERNATIVE FORMULATIONS

Rather than an (un)fairness regularizer, can set up as constrained optimization problem

$$\max_{\phi \in \Phi} I_q(\mathbf{x}; \mathbf{z} | \mathbf{u}) \qquad \text{s.t.} \;\; I_q(\mathbf{z}; \mathbf{u}) < \epsilon$$

Learning Controllable Fair Representations (2018) by Song et al.

- Hard to compute and optimize these mutual information terms
- Propose tractable approximations, bounds to optimize
- Solve the dual

ALTERNATIVE FORMULATIONS

Another popular approach is to adjust the input data, by removing features or pre-processing

- Data preprocessing techniques for classification without discrimination (2011), Kamiran & Calders
- Certifying and removing disparate impact (2015), Feldman et al.
- Optimized data pre-processing for discrimination prevention, Calmon et al.
- The case for process fairness in learning: Feature selection for fair decision ,aking, Grgić-Hlača et al.