Learning Fair Representations

Richard Zemel Ledell Yu Wu Kevin Swersky

Toniann Pitassi Cynthia Dwork

ICML Test of Time

July 27, 2023

Origin Story

- Summer 2010: Toni & I visited Cynthia at MSR Silicon Valley (RIP)
- We met with Omer Reingold, Moritz Hardt
- Many hours discussing computational fairness, reading related philosophy, economics papers
- Struggled mightily with definition

Fairness via S-Blindness?

- Remove or ignore the "membership in S" bit

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

Fairness Through Awareness

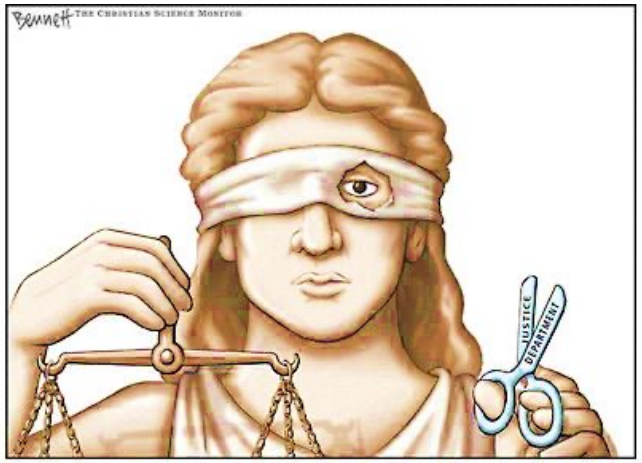
Dwork, Hardt, Pitassi, Reingold, Zemel
Innovations in Theoretical Computer Science, 2012

Key: Fairness requires awareness of membership in protected group(s)

(1). Individual Fairness: Treat similar individuals similarly

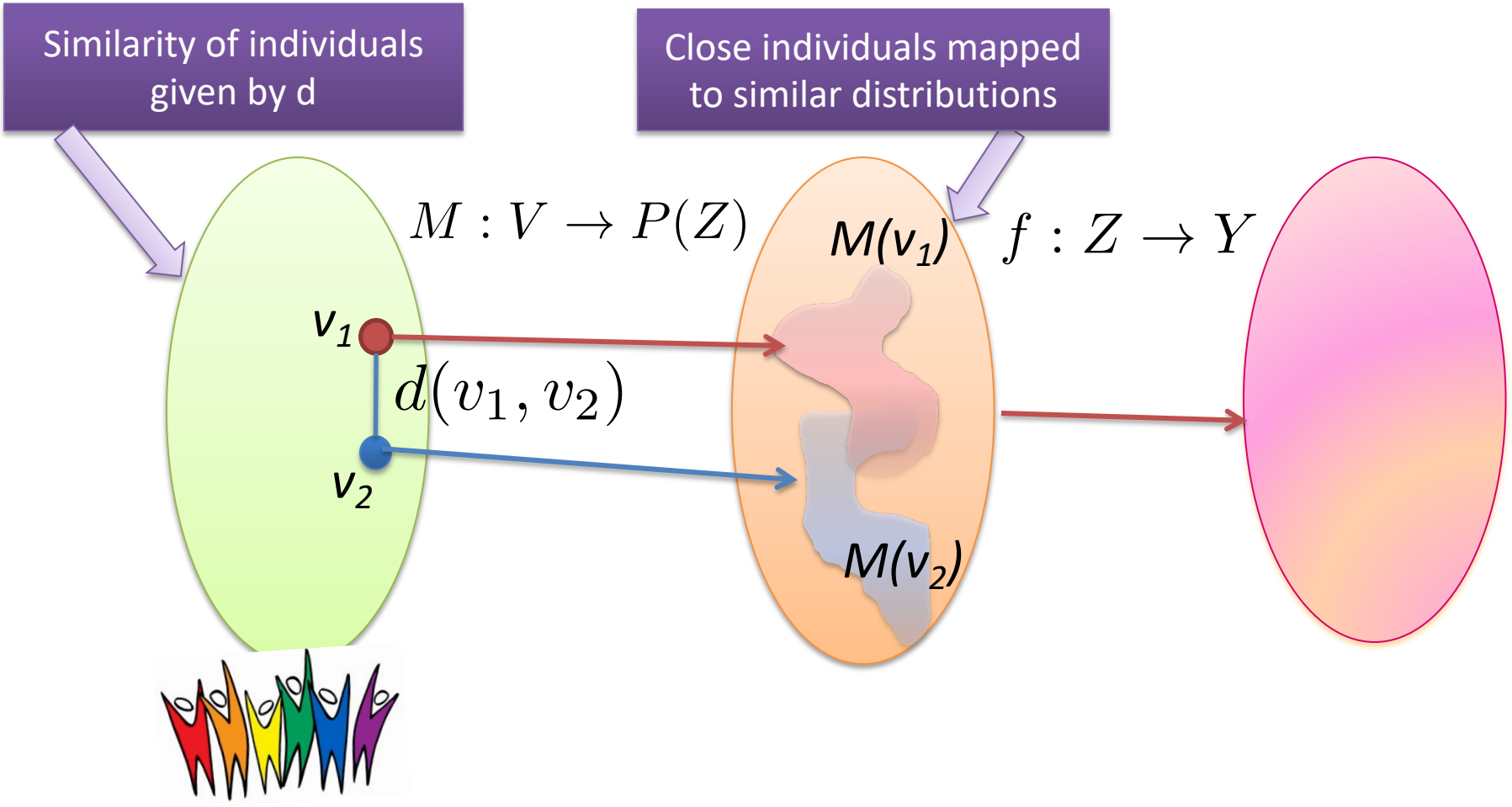
(2). Group Fairness: equalize two groups at the level of outcomes (statistical parity)

- S=1 : minority; S=0 : majority
- P[Y=1 | S=1] = P[Y=1 | S=0]



- **Insufficient** as a notion of fairness
- Fairness requires understanding of classification task
  Cultural understanding of protected groups

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



V: Individuals *Z*: representations Y: Actions

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, healthcare metrics
- Roemer's relative effort metric

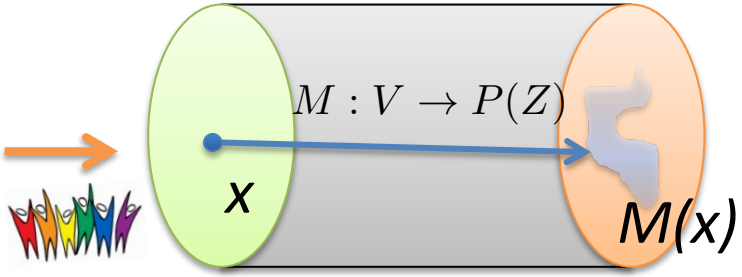
A Fair Optimization Algorithm

utility function
$U: V \times Z \to \mathbb{R}$

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

Efficient Procedure

d-fair mapping M



V: Individuals *Z:* Encodings

LP maximizing vendor's expected utility subject to fairness conditions

Workshop Proposal: NeurIPS 2012

Fairness in Machine Learning: Solon Barocas, Moritz Hardt, me

The aim of this workshop is to bring together people from computer science, philosophy, policy, and the law who have tried to tackle issues of fairness in information systems that rely on machine learning and statistical inference. One goal is to inform the NIPS community about this intellectually interesting and practically important area, in an attempt to develop machine learning approaches that give greater or more precise effect to existing anti-discrimination law, and tackle some important computational challenges. These include constructing a framework to represent what is meant by fairness, and develop algorithms for performing classifications that balance accuracy and fairness.

Invited speakers: Toon Calders, Dino Pedreschi; Toniann Pitassi; Omer Reingold; Mireille Hildebrandt; Deirdre Mulligan; Helen Nissenbaum

Workshop Chair: "We did discuss a lot about it and our we major concern was about the amount of audience that the workshop will get."

Learning Fair Representations

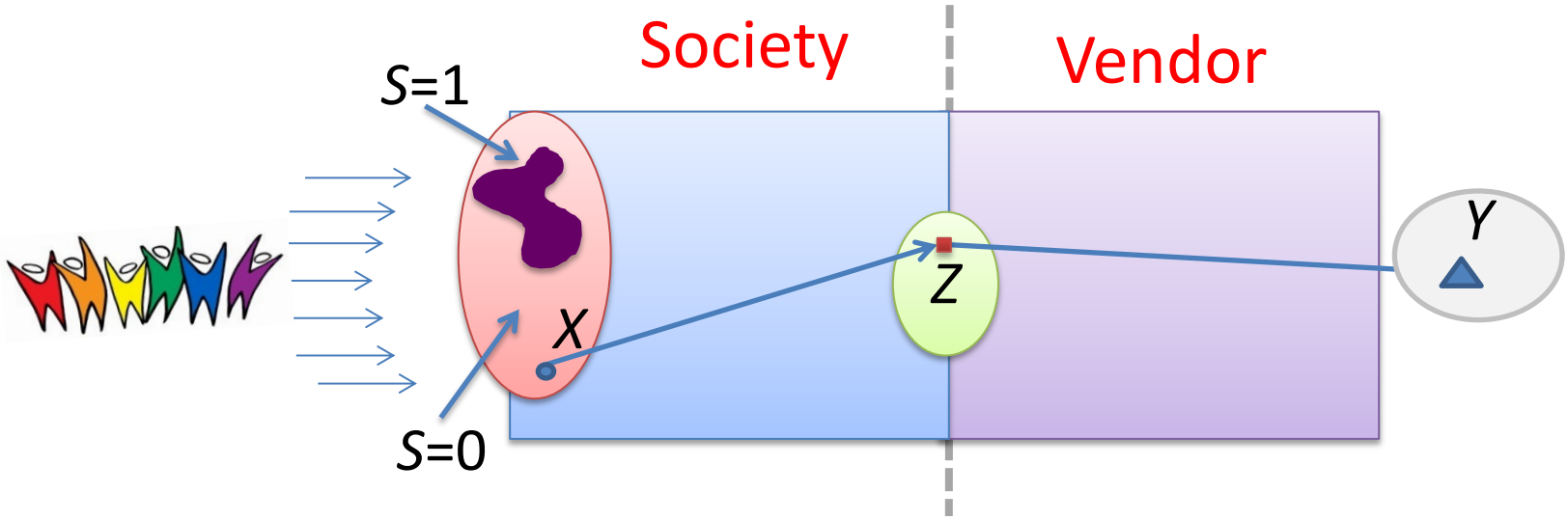
Zemel, Wu, Swersky, Pitassi, Dwork
ICML, 2013

- Cast as learning problem
  - Generalizes to new data: learn general mapping, applies to any individual

- Key idea: learn representations, such that distance in embedding space captures metric relevant to task at hand -- learning the metric

- Use fair representation for additional classification tasks (transfer learning)

Model Overview



Aims for Z:

1. Preserve information so vendor can max utility
2. Preserve information in X
3. Lose information about S

Group Fairness/Statistical Parity: P(Z|S=0) = P(Z|S=1)

Max: MI(*Z*, *Y*) + MI(*Z*, *X*) – MI(*Z*, *S*)

Initial Formulation

True Test-of-Time: started off with my Matlab code!

Aim to jointly optimize:

max. [MI($h(X),X$) – MI($h(X),S$)]

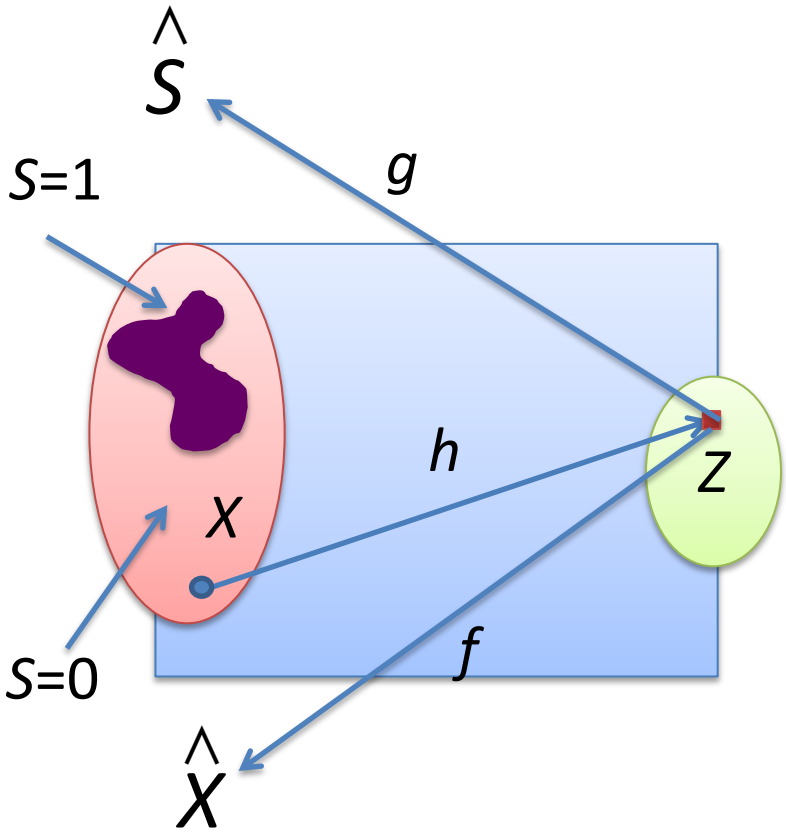
$$E_{X \sim P(X)} [f(h(X))-X]^2 - [g(h(X))-S]^2$$

Can alternate:

optimize h given f, g;

optimize f, g given h

→ unstable



Instantiating the Model

Simple tractable formulation:

Z is a discrete latent variable

Key: min. *MI(Z,S)* by forcing *P(Z|S=1) = P(Z|S=0)*

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

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

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

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

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**

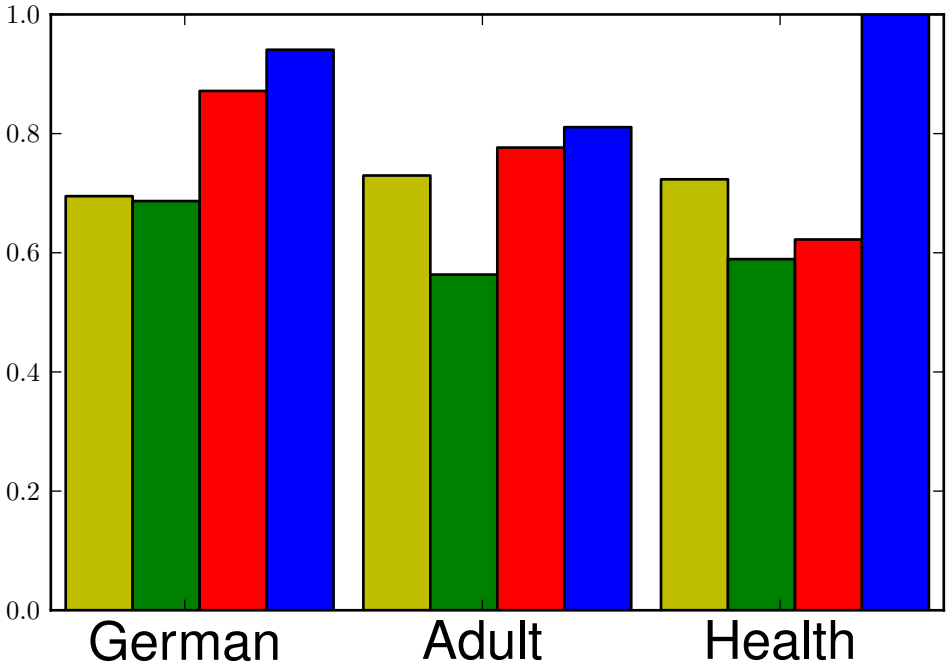
3. **Heritage Health**
   Size: 147,473 instances, 139 attributes
   Task: predict whether patient spends any nights in hospital
   Sensitive feature: **Age**

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



Open Problems (2013)

- Further extensions of intermediate representations: more expressive mappings → preserve information in X while losing information about S
  - Kernel formulation, multi-layer neural network
  - For more general mappings, need to utilize other methods of matching posteriors (sample tests?)

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

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

Shifting Landscape

"Big Data: Seizing Opportunities, Preserving Values" (2014)

- concluded that "big data technologies can cause societal harms beyond damages to privacy"
- concern about the possibility that decisions informed by big data could have discriminatory effects, even in absence of discriminatory intent
- could subject already disadvantaged groups to less favorable treatment
- expressed alarm about the threat that an "opaque decision-making environment" and "impenetrable set of algorithms" pose to autonomy
- called for additional "technical expertise to stop discrimination", and for further research into the dangers of "encoding discrimination in automated decisions"

Field Matures

- **FAT/ML Workshop, NeurIPS 2014:** ran for 5 years, spawned Fairness Accountability and Transparency Conference (FAccT)

- Other conferences:
  - AI Ethics and Society (AIES)
  - Foundations of Responsible Computing (FORC)

- Prominent research components in Computer Vision, NLP, AI conferences

Conference on Data & Civil Rights: A New Era of Policing and Justice (2015)

The purpose of the conference:

1. to better understand new data-related criminal justice practices from different perspectives
2. to help create connections between different constituencies, all to help assure that technology can be used as a force for good in criminal justice.

Discovered:

- deep distrust of algorithms
- warranted in many cases: Pennsylvania's recidivism scoring system

Richer Z

Replace discrete representation with continuous, multi-dimensional latent representation *Z*

- Formulate objective of matching group-based latent distributions using MMD

  Fair Variational Autoencoder

  [Louizos-Swersky-Li-Welling-Zemel-2015]

- Adversarial formulation, capable of handling fairness metrics beyond statistical parity

  Learning Adversarially Fair and Transferrable Representations

  [Madras-Creager-Pitassi-Zemel-2018]

Current Achievements & Challenges

Broader view

- Intersectional fairness: many sensitive attributes, often overlapping
- Unknown sensitive attributes: privacy, not observed; suboptimal when known [Tomasev-Kay-McKee-Mohamed-2021]
- Links with robust ML

Fairness ⟺ Domain Generalization

Statistic to match/optimize	e known?	DG method	Fairness method
match $\mathbb{E}[\ell\|e]\ \forall e$	yes	REx (Krueger et al., 2021),	CVaR Fairness (Williamson & Menon, 2019)
$\min \max_e \mathbb{E}[\ell\|e]$	yes	Group DRO (Sagawa et al., 2020)	
$\min \max_q \mathbb{E}_q[\ell]$	no	DRO (Duchi et al., 2021)	Fairness without Demographics (Hashimoto et al., 2018; Lahoti et al., 2020)
match $\mathbb{E}[y\|\Phi(x), e]\ \forall\ e$	yes	IRM (Arjovsky et al., 2019)	Group Sufficiency (Chouldechova, 2017; Liu et al., 2019)
match $\mathbb{E}[y\|\Phi(x), e]\ \forall\ e$	no	**EIIL (ours)**	**EIIL (ours)**
match $\mathbb{E}[\hat{y}\|\Phi(x), e, y = y']\ \forall\ e$	yes	C-DANN (Li et al., 2018) PGI (Ahmed et al., 2021)	Equalized Odds (Hardt et al., 2016)
match $\left\|\mathbb{E}[y\|S(x), e] - \mathbb{E}[\hat{y}(x)\|S(x), e]\right\|\ \forall\ e$	no		Multicalibration (Hébert-Johnson et al., 2018)
match $\left\|\mathbb{E}[y\|e] - \mathbb{E}[\hat{y}(x)\|e]\right\|\ \forall\ e$	no		Multiaccuracy (Kim et al., 2019)
match $\left\|\mathbb{E}[y \neq \hat{y}(x)\|y = 1, e]\right\|\ \forall\ e$	no		Fairness Gerrymandering (Kearns et al., 2018)

Environment Inference for Invariant Learning [Creager-Jacobsen-Zemel-2021]

Current Achievements & Challenges

Broader view

- Intersectional fairness: many sensitive attributes, often overlapping
- Unknown sensitive attributes: privacy, not observed; suboptimal when known [Tomasev-Kay-McKee-Mohamed-2021]
- Links with robust ML

Increased scrutiny of ML algorithms

Large body of work since

- multi-calibration [Hebert-Johnson+-2018}
- participatory ML - collective recourse, bias bounties
  - Queer In AI: A Case Study in Community-Led Participatory AI

Thanks!

Ledell Yu Wu

Kevin Swersky

Toni Pitassi

Cynthia Dwork