

# **Fairness in Machine Learning**

**Moritz Hardt**

# Fairness in Classification

Advertising



Education



Financial aid

Health

Care



Banking

Insurance



Taxation

*many more...*

# Concern: Discrimination

- Certain attributes should be *irrelevant!*
- Population includes minorities
  - Ethnic, religious, medical, geographic
- Protected by law, policy, ethics



# “Big Data: Seizing Opportunities, Preserving Values”

THE 90-DAY REVIEW  
FOR BIG DATA



"big data technologies can cause societal harms beyond damages to privacy"

# Other notions of “fairness” in CS

- Fair scheduling
- Distributed computing
- Envy-free division (cake cutting)
- Stable matching





# Discrimination arises even when nobody's *evil*

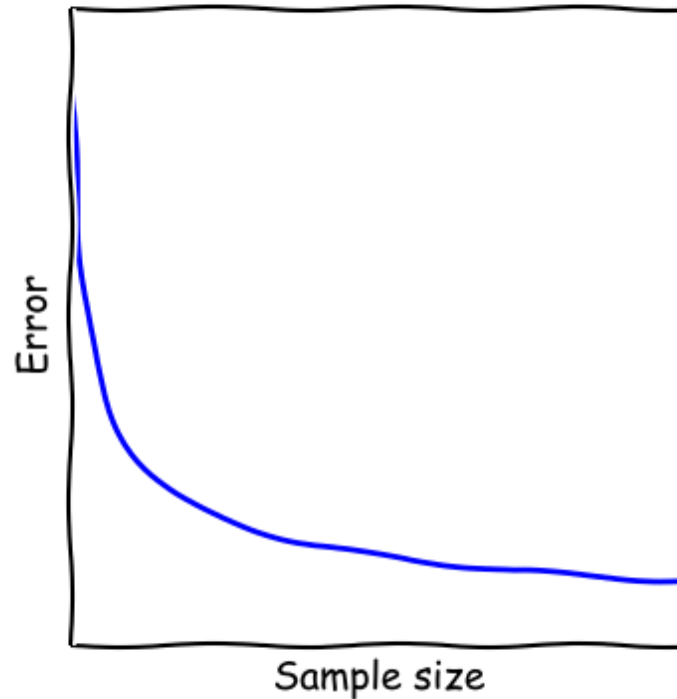
- Google+ tries to classify real vs fake names
- Fairness problem:
  - Most training examples standard white American names: John, Jennifer, Peter, Jacob, ...
  - Ethnic names often unique, much fewer training examples

Likely outcome: Prediction accuracy  
*worse on ethnic names*

*“Due to Google's ethnocentricity I was prevented from using my real last name (my nationality is: Tungus and Sami)”*

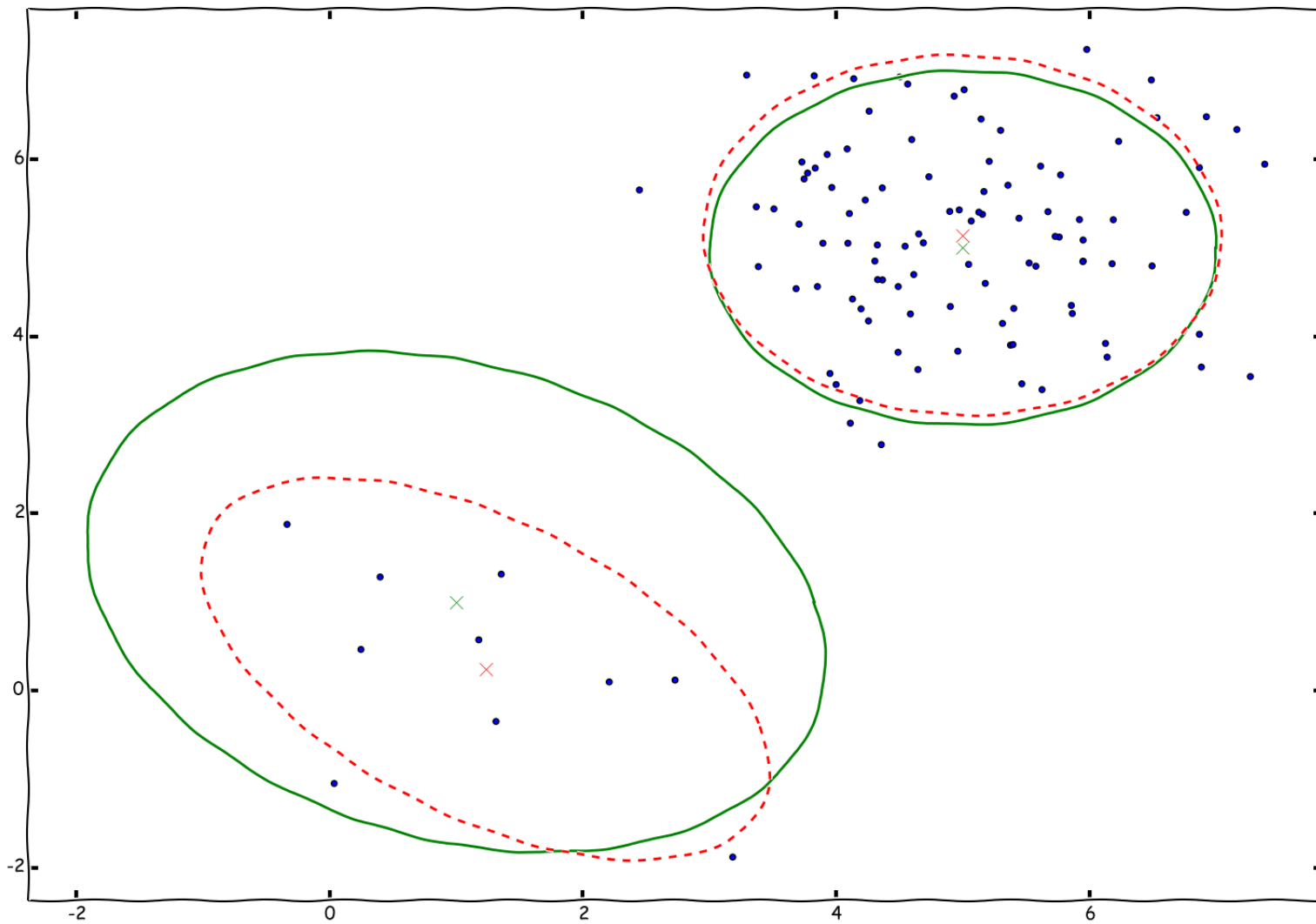
- Katya Casio. Google Product Forums.

# Error vs sample size



## Sample Size Disparity:

In a heterogeneous population, smaller groups face larger error





# Credit Application



**More miles  
and no annual fee**

Earn trips faster with VentureOne<sup>SM</sup>

Get Started 

only at **Capital One CARD LAB**

**Capital One Card Lab  
Platinum Prestige Credit Card**

**Capital One Card Lab  
VentureOne Card**

**Savings Accounts  
Earn With Great Rates**

The advertisement features a yellow Capital One VentureOne Visa Signature credit card. The card displays the name 'VENTURE', the number '4000 1234 5678 9010', the expiration date '12/12', and the name 'DER'. The card is set against a background of a tropical island with palm trees and a blue sky. The text 'More miles and no annual fee' is prominently displayed in blue and orange. Below this, it says 'Earn trips faster with VentureOne<sup>SM</sup>'. A green button with a white arrow and the text 'Get Started' is positioned below the text. At the bottom, there are three white boxes with blue text: 'Capital One Card Lab Platinum Prestige Credit Card', 'Capital One Card Lab VentureOne Card', and 'Savings Accounts Earn With Great Rates'. The Capital One logo is visible in the top right corner of the card image.

User visits `capitalone.com`

Capital One uses tracking information provided by the tracking network [x+1] to personalize offers

**Concern:** Steering minorities into higher rates (illegal)

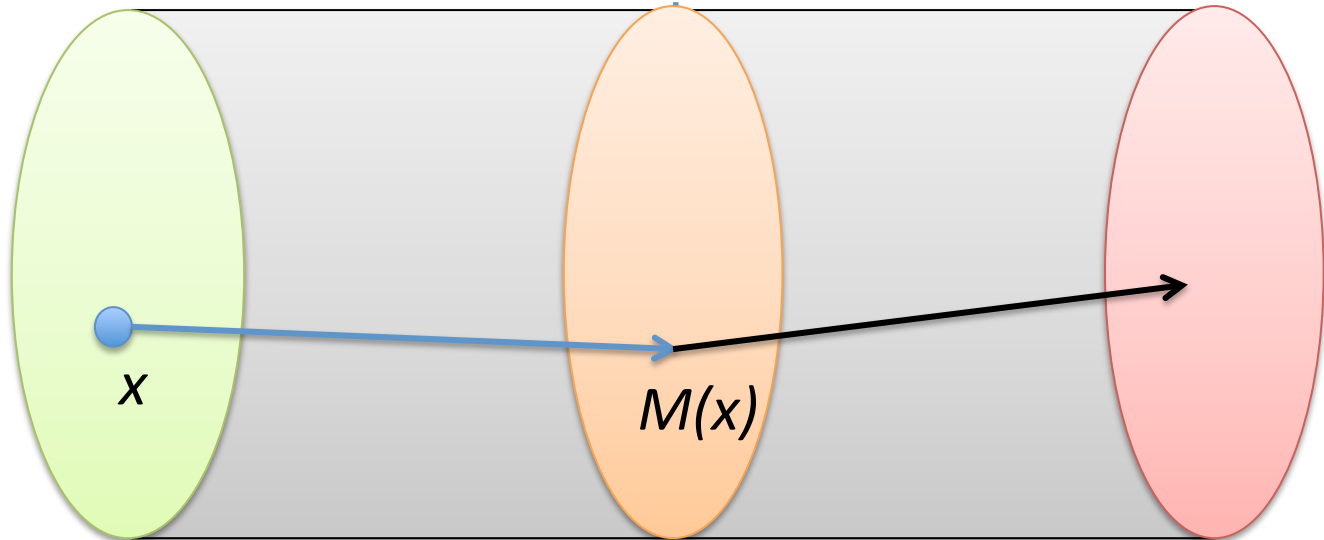
WSJ 2010

Classifier  
(eg. ad network)

$$M: V \rightarrow O$$

Vendor  
(eg. capital one)

$$f: O \rightarrow A$$



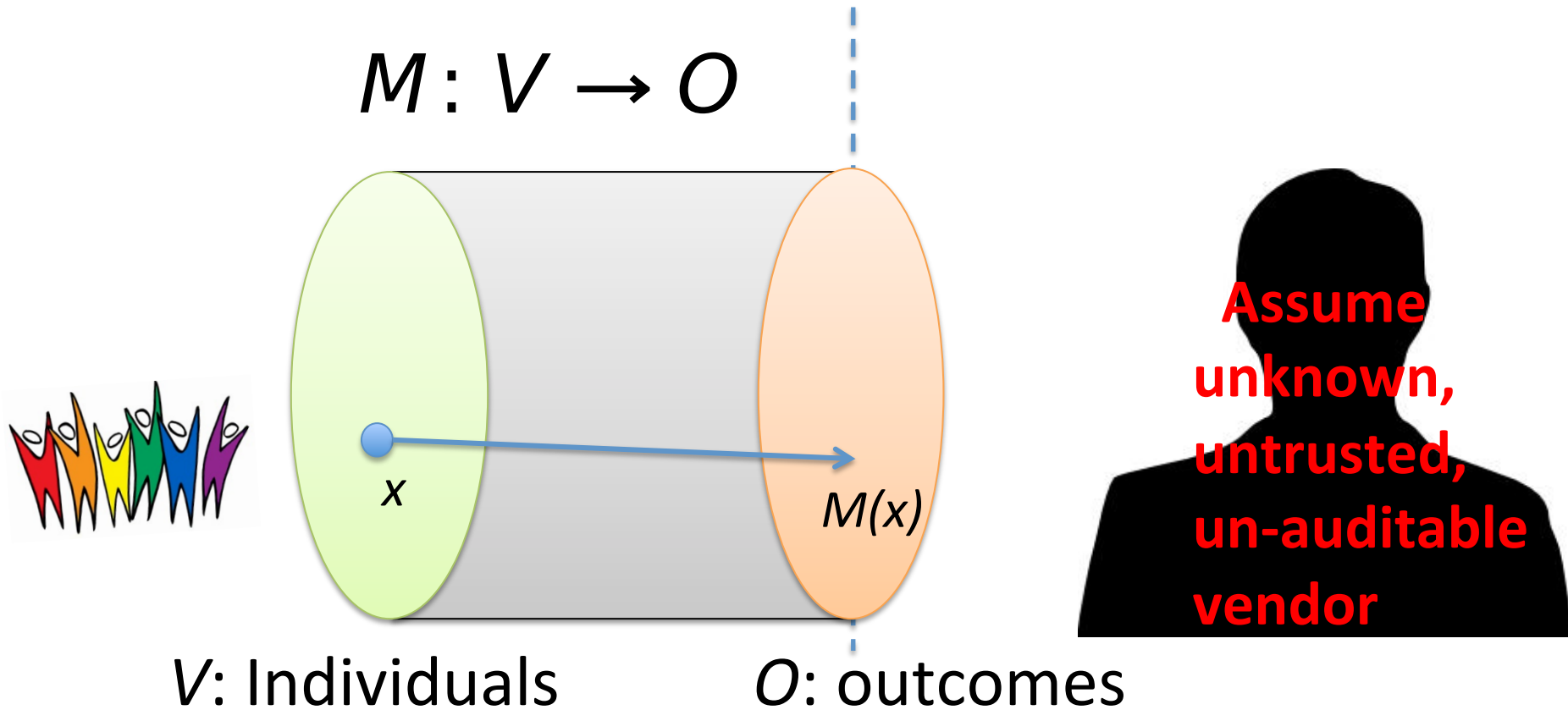
V: Individuals

O: outcomes

A: actions

Our goal:

Achieve Fairness in the classification step



First attempt...

# Fairness through Blindness



# Fairness through Blindness

Ignore all irrelevant/protected attributes

*“We don’t even look at ‘race’!”*

# Point of Failure

You don't need to *see* an attribute to be able to *predict* it with high accuracy

E.g.: User visits `artofmanliness.com`  
... 90% chance of being male

# Fairness through Privacy?

“It's Not Privacy, and It's Not Fair”

Cynthia Dwork & Deirdre K. Mulligan. Stanford Law Review.

Privacy is no Panacea: Can't hope to have privacy solve our fairness problems.

“At worst, **privacy solutions can hinder efforts to identify classifications that unintentionally produce objectionable outcomes**—for example, differential treatment that tracks race or gender—by limiting the availability of data about such attributes.”



Second attempt...

# Statistical Parity (Group Fairness)

Equalize two groups  $S$ ,  $T$  at the level of outcomes

– E.g.  $S = \text{minority}$ ,  $T = S^c$

$$\Pr[\text{outcome } o \mid S] = \Pr[\text{outcome } o \mid T]$$

“Fraction of people in  $S$  getting credit same as in  $T$ .”

**Not strong enough** as a notion of fairness

– Sometimes desirable, but can be abused

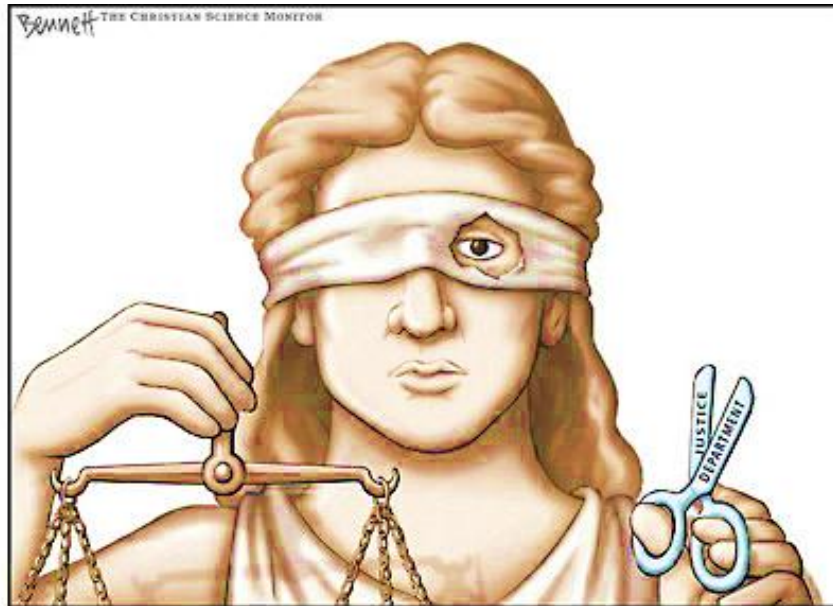
- **Self-fulfilling prophecy:** Select smartest students in  $T$ , random students in  $S$

– *Students in  $T$  will perform better*

# Lesson: Fairness is *task-specific*

Fairness requires understanding of classification task and protected groups

“Awareness”



# **Individual Fairness Approach**

# Individual Fairness

Treat *similar* individuals *similarly*



Similar for the purpose of  
the classification task



Similar distribution  
over outcomes



# The Similarity Metric

# 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
  - In the spirit of Rawls
- Typically, does not suggest classification!



# Examples

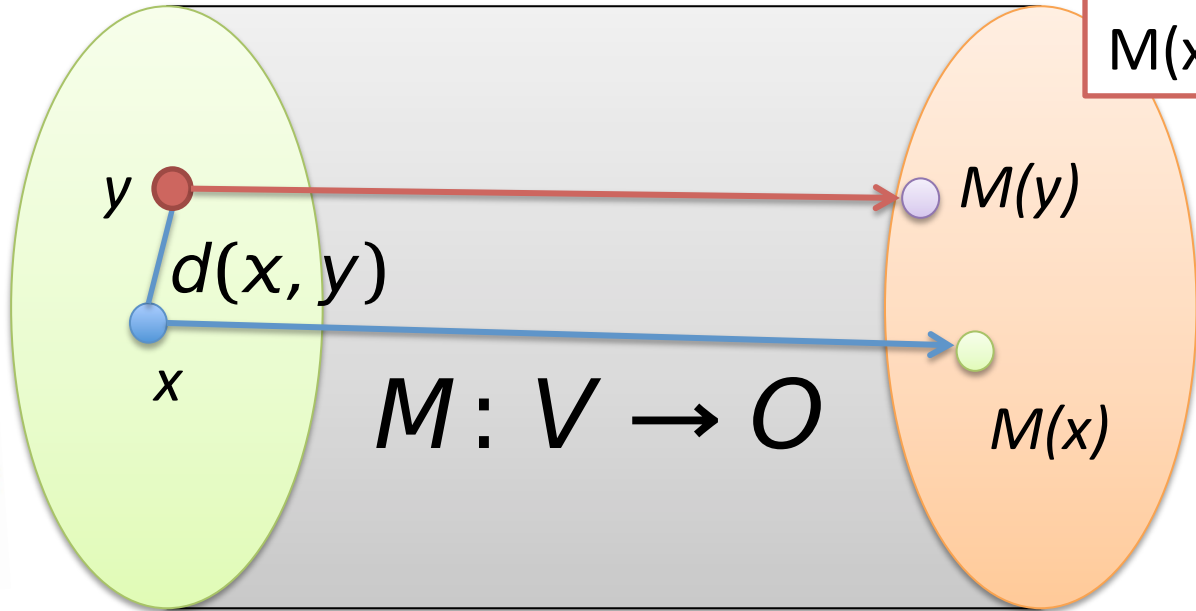
- Financial/insurance risk metrics
  - Already widely used (though secret)
- **AALIM health care metric**
  - health metric for treating similar patients similarly
- Roemer's relative effort metric
  - Well-known approach in Economics/Political theory

Maybe not so much science fiction after all...

# How to formalize this?

Think of  $V$  as space  
with metric  $d(x,y)$   
similar = small  $d(x,y)$

How can we  
compare  
 $M(x)$  with  $M(y)$ ?

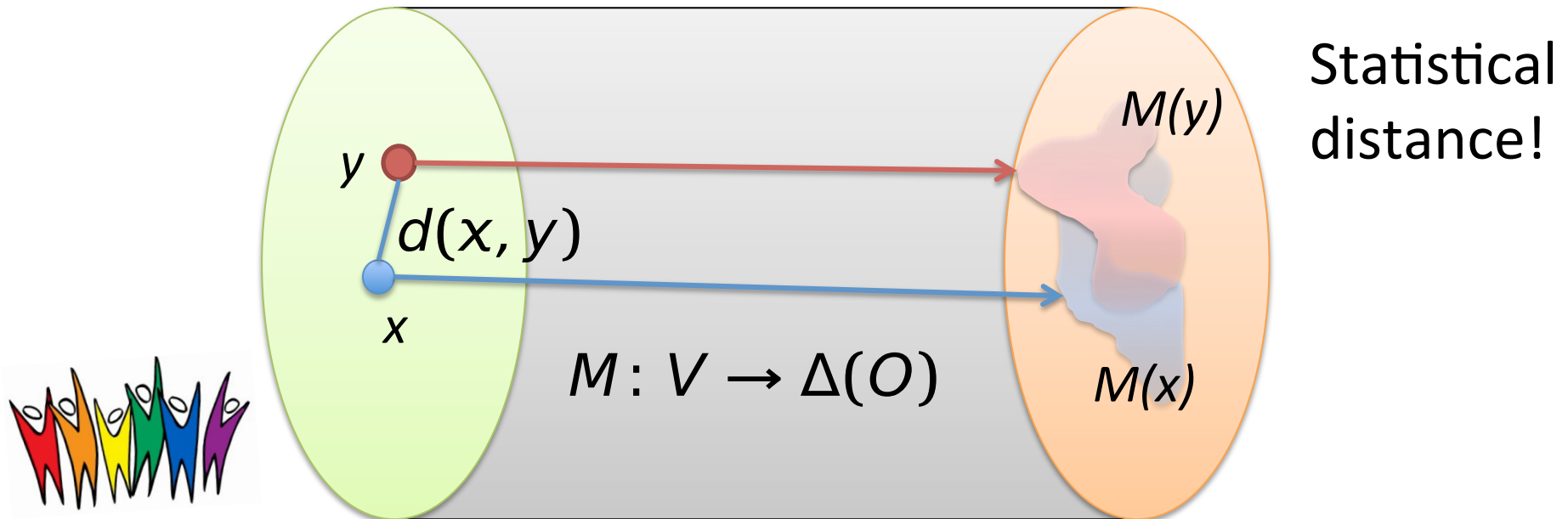


$V$ : Individuals

$O$ : outcomes

# Distributional outcomes

How can we compare  $M(x)$  with  $M(y)$ ?



$V$ : Individuals

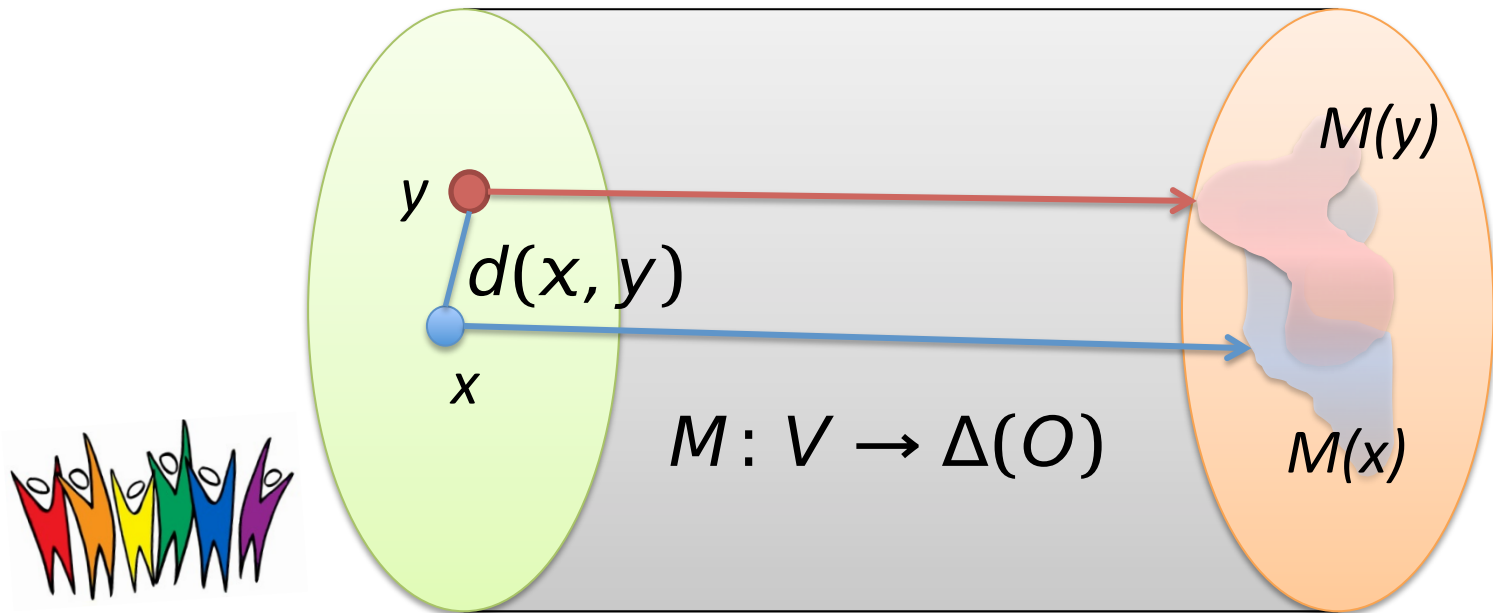
$O$ : outcomes

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

Lipschitz condition  $\|M(x) - M(y)\| \leq d(x, y)$

This talk: Statistical distance

in  $[0,1]$



$V$ : Individuals

$O$ : outcomes

Key elements of our approach...

# Utility Maximization

Vendor can specify **arbitrary utility function**

$$U: V \times O \rightarrow \mathbb{R}$$

$U(v,o)$  = Vendor's utility of giving individual  $v$   
the outcome  $o$

Can efficiently maximize vendor's expected utility subject to Lipschitz condition

$$\max_{x \in V} \mathbb{E}_{o \sim M(x)} U(x, o)$$

s.t.  $M$  is  $d$ -Lipschitz

**Exercise:**  
Write this as  
an LP

# When does Individual Fairness imply Group Fairness?

Suppose we enforce a metric  $d$ .

**Question:** Which *groups of individuals* receive (approximately) equal outcomes?

**Theorem:**

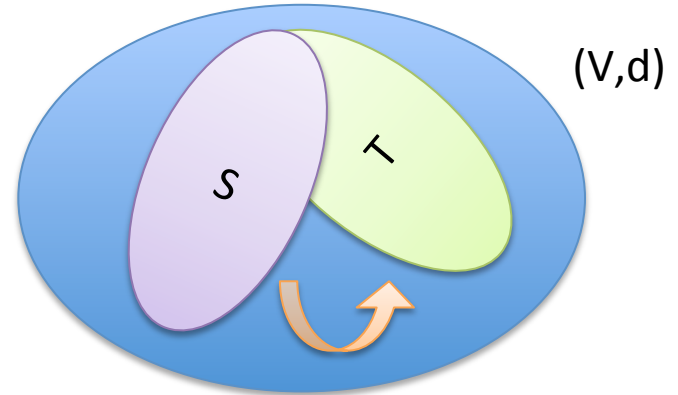
Answer is given by **Earthmover distance** (w.r.t.  $d$ ) between the two groups.





# How different are $S$ and $T$ ?

Earthmover Distance:  
Cost of transforming  
uniform distribution on  $S$  to  
uniform distribution on  $T$



$$EM_d(S, T) = \min \sum_{x, y \in V} h(x, y) d(x, y)$$

s.t.

$$\sum_{x \in V} h(x, y) = S(x)$$
$$\sum_{y \in V} h(x, y) = T(y)$$
$$h(x, y) \geq 0$$

$$EM_d(S, T) = \min \sum_{x, y \in V} h(x, y) d(x, y)$$

$$\begin{aligned} \text{s.t.} \quad & \sum_{x \in V} h(x, y) = S(x) \\ & \sum_{y \in V} h(x, y) = T(y) \\ & h(x, y) \geq 0 \end{aligned}$$

$\text{bias}(d, S, T) =$  largest violation of statistical parity between  $S$  and  $T$  that any  $d$ -Lipschitz mapping can create

### Theorem:

$$\text{bias}(d, S, T) = EM_d(S, T)$$



# Proof Sketch: LP Duality

- $EM_d(S,T)$  is an LP by definition
- Can write  $\text{bias}(d,S,T)$  as an LP:

$\max \Pr( M(x) = 0 \mid x \text{ in } S ) - \Pr( M(x) = 0 \mid x \text{ in } T )$

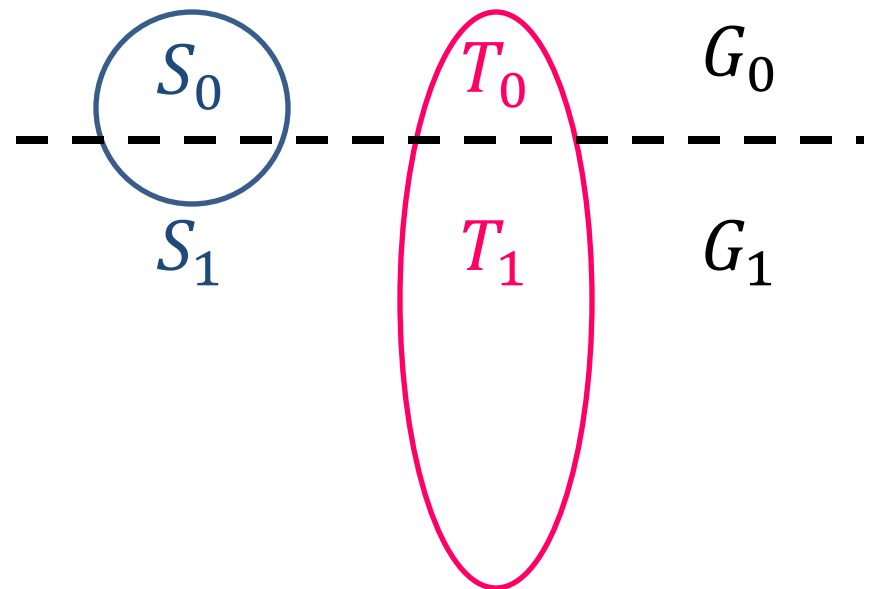
subject to:

- (1)  $M(x)$  is a probability distribution for all  $x$  in  $V$
- (2)  $M$  satisfies all  $d$ -Lipschitz constraints

Program dual to Earthmover LP!

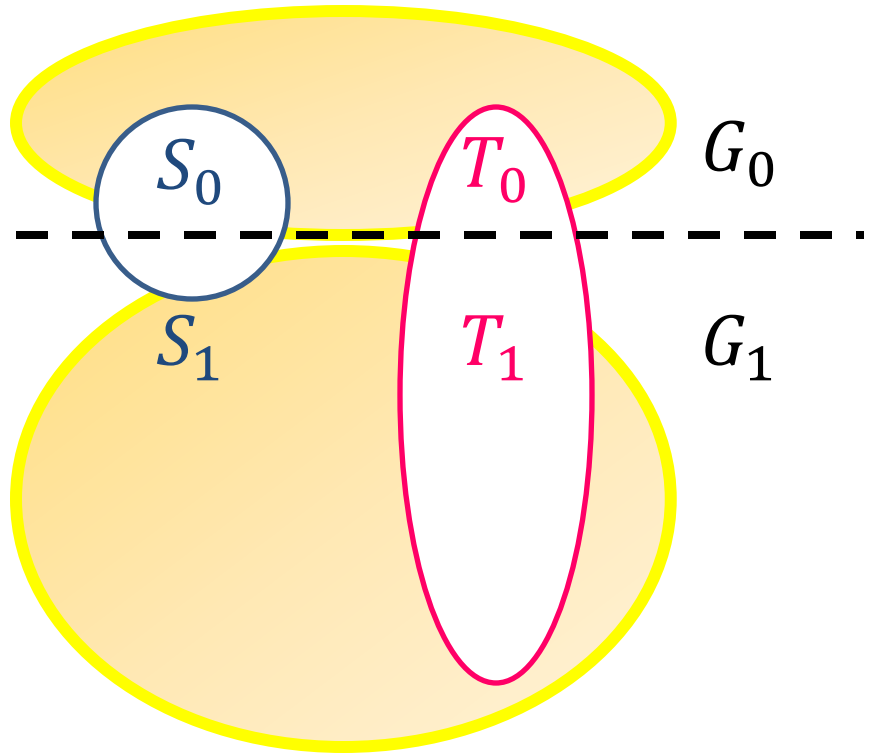
# Toward Fair Affirmative Action: When $EM(S,T)$ is Large

- $G_0$  is unqualified
- $G_1$  is qualified



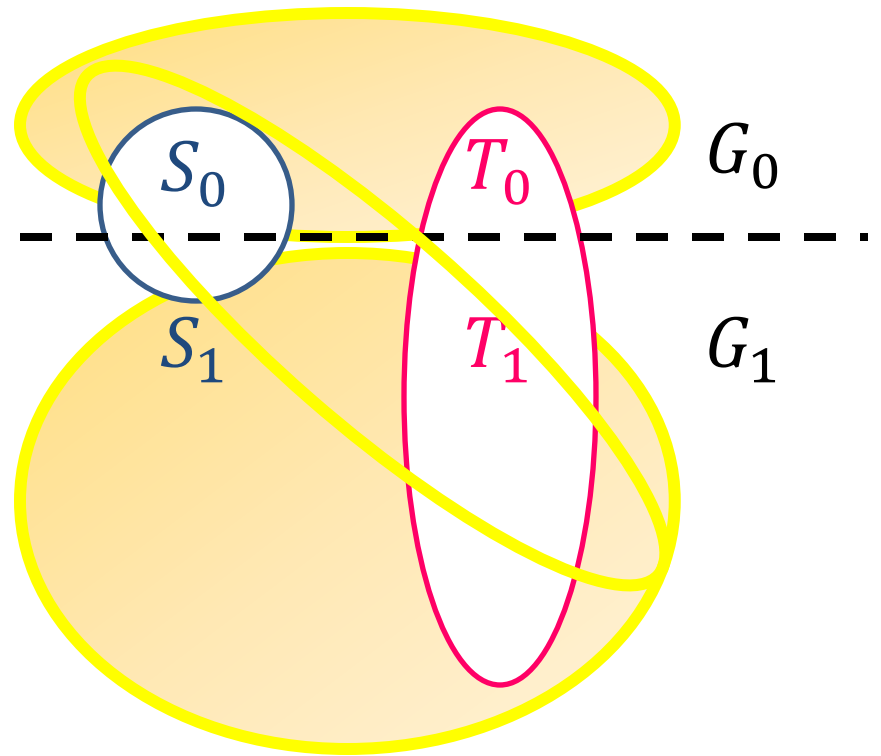
# Toward Fair AA: When $EM(S,T)$ is Large

- Lipschitz  $\Rightarrow$   
All in  $G_i$  treated same



# Toward Fair AA: When $EM(S,T)$ is Large

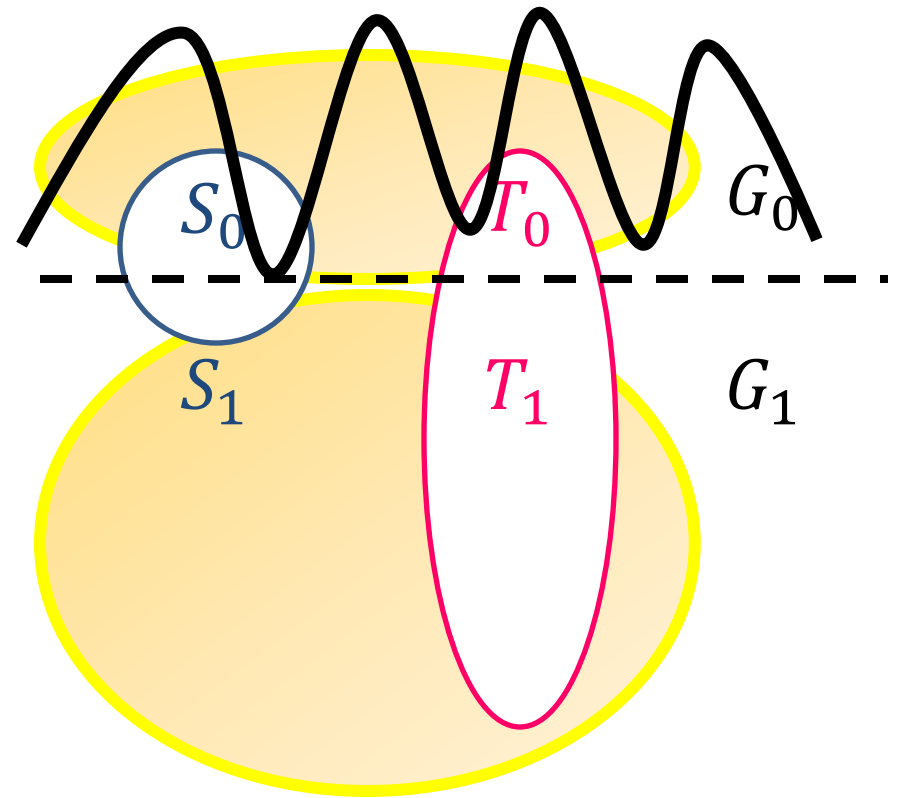
- Lipschitz  $\Rightarrow$   
All in  $G_i$  treated same
- **Statistical Parity**  $\Rightarrow$   
much of  $S_0$  must be  
treated the same as  
much of  $T_1$



# Toward Fair AA: When $EM(S,T)$ is Large

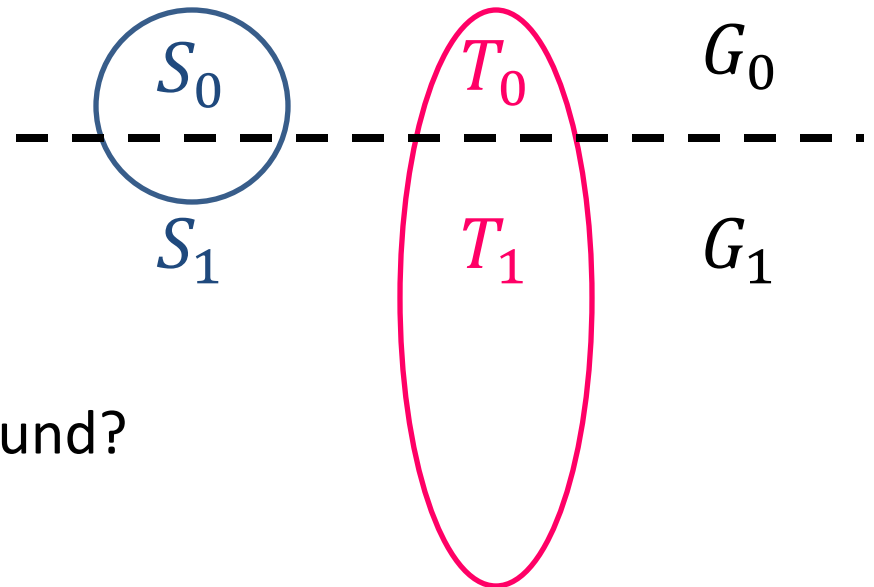
- Lipschitz  $\Rightarrow$   
All in  $G_i$  treated same

**Failure to Impose Parity**  $\Rightarrow$   
anti- $S$  vendor can target  $G_0$   
with blatant hostile ad  $f_u$ .  
Drives away almost all of  $S$   
while keeping most of  $T$ .



# Dilemma: What to Do When $EM(S,T)$ is Large?

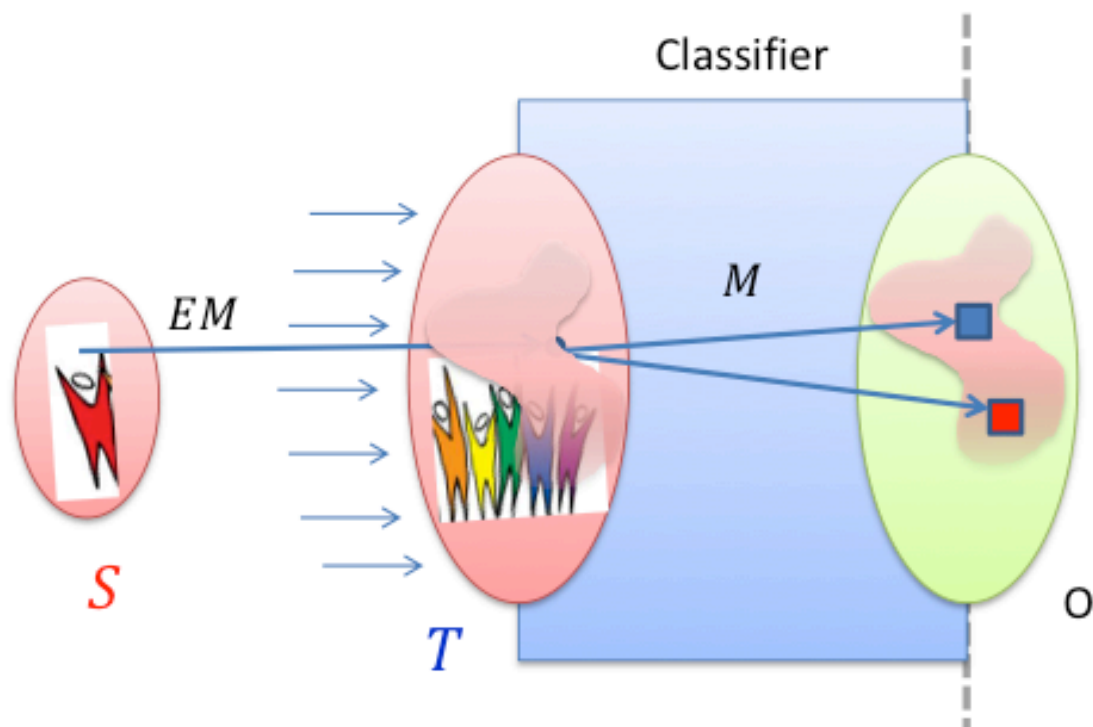
- Imposing parity causes collapse
- Failing to impose parity permits blatant discrimination



How can we form a middle ground?



# Fair Affirmative Action

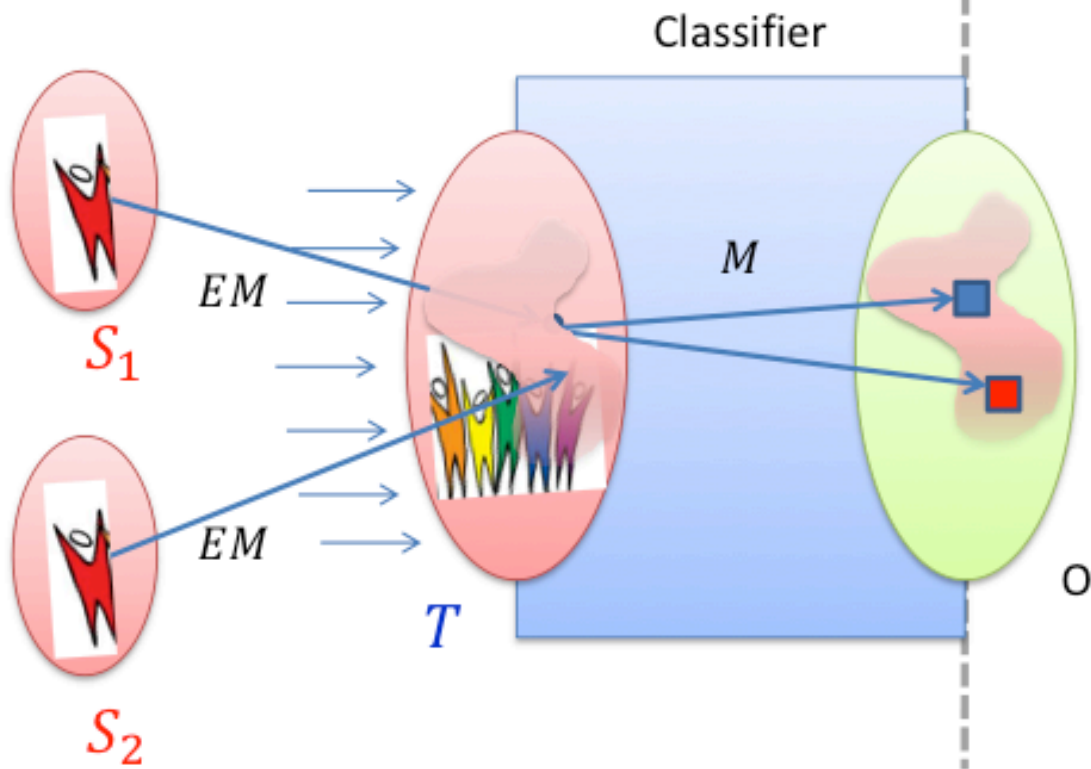


Earthmover mapping from  $S$  to  $T$  + Lipschitz mapping from  $T$  to  $O$

Achieves:

- Lipschitz on  $S \times S, T \times T$ , on average on  $S \times T$
- statistical parity between  $S$  and  $T$
- no collapse

# Fair Affirmative Action



- ▶ Immediately suggests a method of dealing with multiple disjoint  $S$ 's

# Connection to differential privacy

- Close connection between individual fairness and **differential privacy** [Dwork-McSherry-Nissim-Smith'06]

DP: Lipschitz condition on set of databases

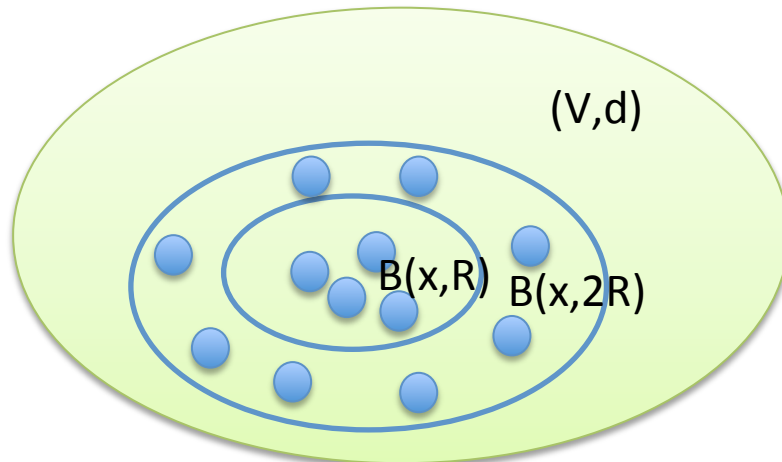
IF: Lipschitz condition on set of individuals

	Differential Privacy	Individual Fairness
Objects	Databases	Individuals
Outcomes	Output of statistical analysis	Classification outcome
Similarity	General purpose metric	Task-specific metric

# Can we import techniques from Differential Privacy?

**Theorem:** Fairness mechanism with “high utility” in metric spaces  $(V, d)$  of bounded doubling dimension

Based on exponential mechanism [MT'07]



$$|B(x, R)| \leq O(|B(x, 2R)|)$$

# Summary: Individual Fairness

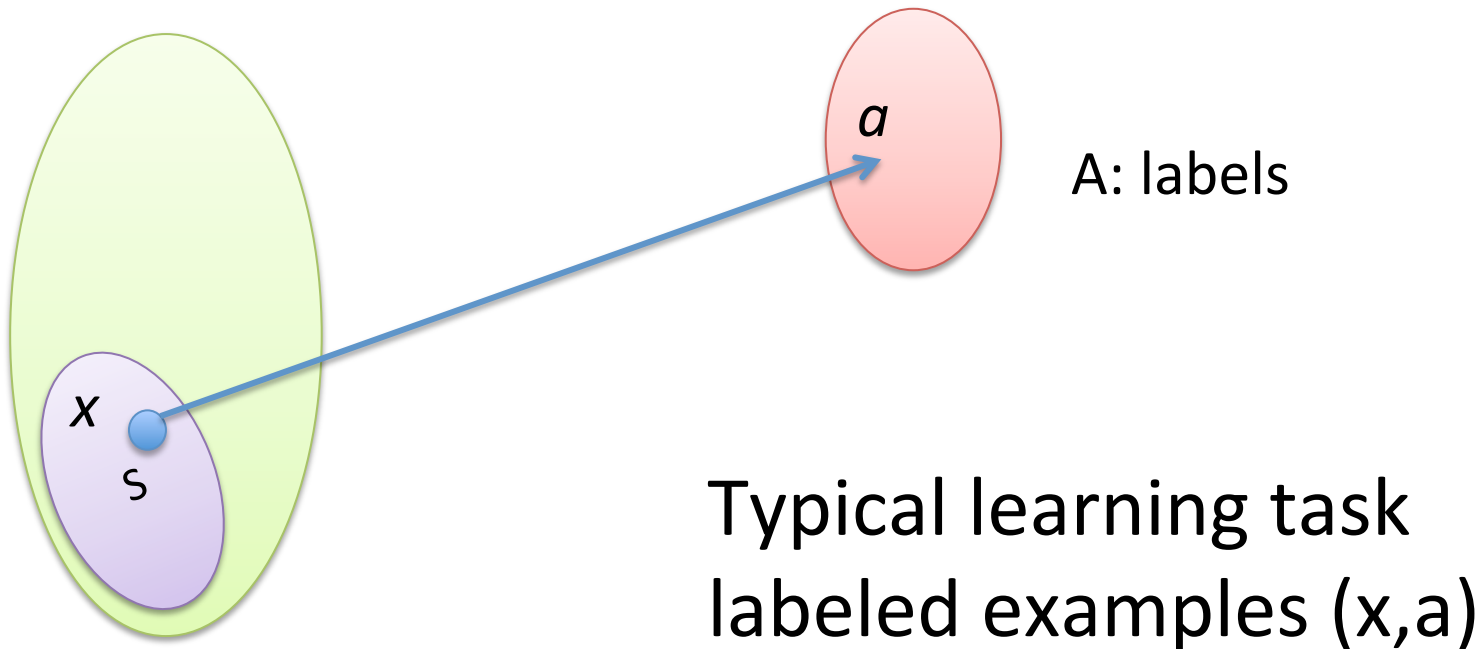
- Formalized fairness property based on treating similar individuals similarly
  - Incorporates vendor's utility
- Explored relationship between individual fairness and group fairness
  - Earthmover distance
- Approach to fair affirmative action based on Earthmover solution

# Lots of open problems/direction

- **Metric**
  - Social aspects, who will define them?
  - How to generate metric (semi-)automatically?
- **Earthmover characterization** when probability metric is not statistical distance (but infinity-div)
- Explore connection to **Differential Privacy**
- Connection to **Economics** literature/problems
  - Rawls, Roemer, Fleurbaey, Peyton-Young, Calsamiglia
- **Case Study**
- **Quantitative trade-offs** in concrete settings

# Some recent work

- Zemel-Wu-Swersky-Pitassi-Dwork  
“Learning Fair Representations” (ICML 2013)

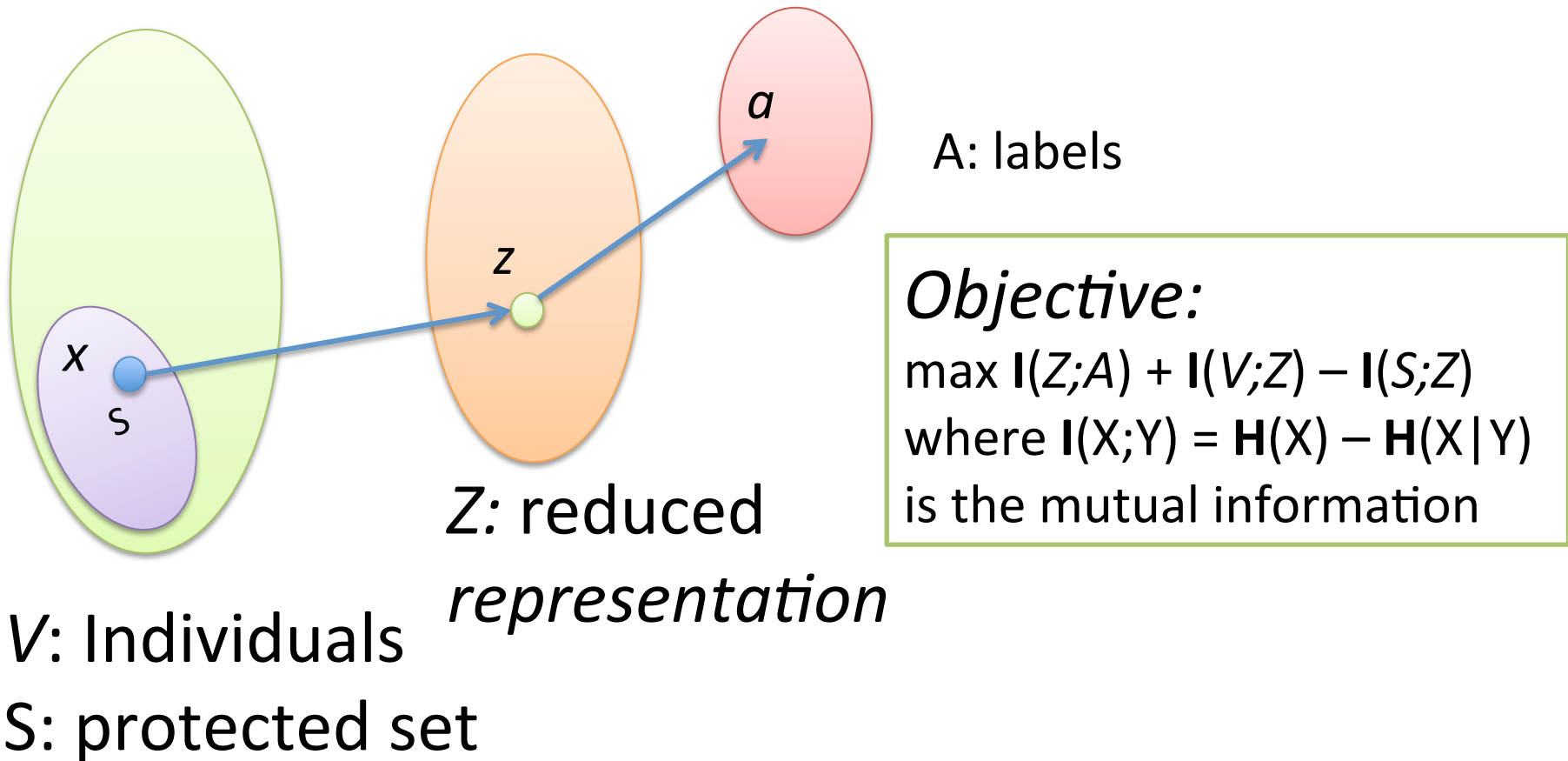


V: Individuals

S: protected set

# Some recent work

- Zemel-Wu-Swersky-Pitassi-Dwork  
“Learning Fair Representations” (ICML 2013)

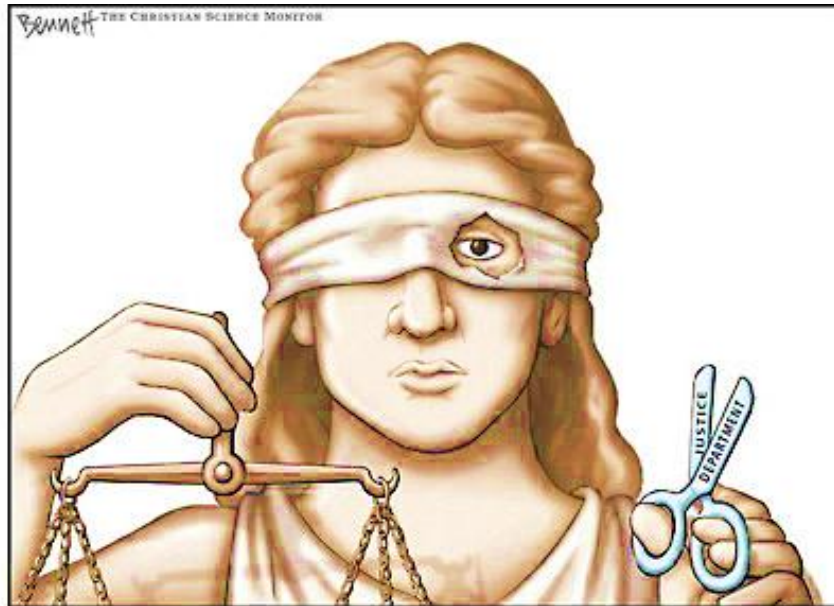




# Web Fairness Measurement

How do we measure the **“fairness of the web”**?

- Need to model/understand user browsing behavior
- Evaluate how web sites respond to different behavior/attributes
- Cope with noisy measurements
- Exciting progress by Datta, Datta, Tschantz



Questions?