18734: Foundations of Privacy

Fairness in Classification

Anupam Datta

With many slides from Moritz Hardt

Fall 2018

Fairness in Classification

Advertising

Education

Financial aid

Banking

Health Care

Taxation

Insurance

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 ~ 2014

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

Overview

- Fairness as a (group) statistical property
- Individual fairness
- Achieving fairness with utility considerations

Discrimination arises even when nobody's *evil*

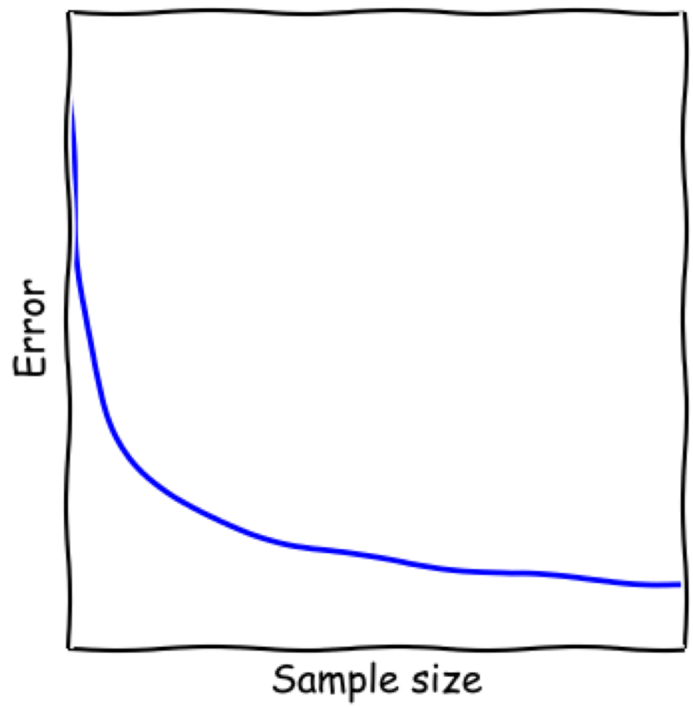
- Google+ tries to classify real vs fake names
- Fairness problem:
  - Most training examples standard white American names: John, Jennifer, Peter, Jacob, ...
  - Ethnic names often unique, much fewer training examples

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)"

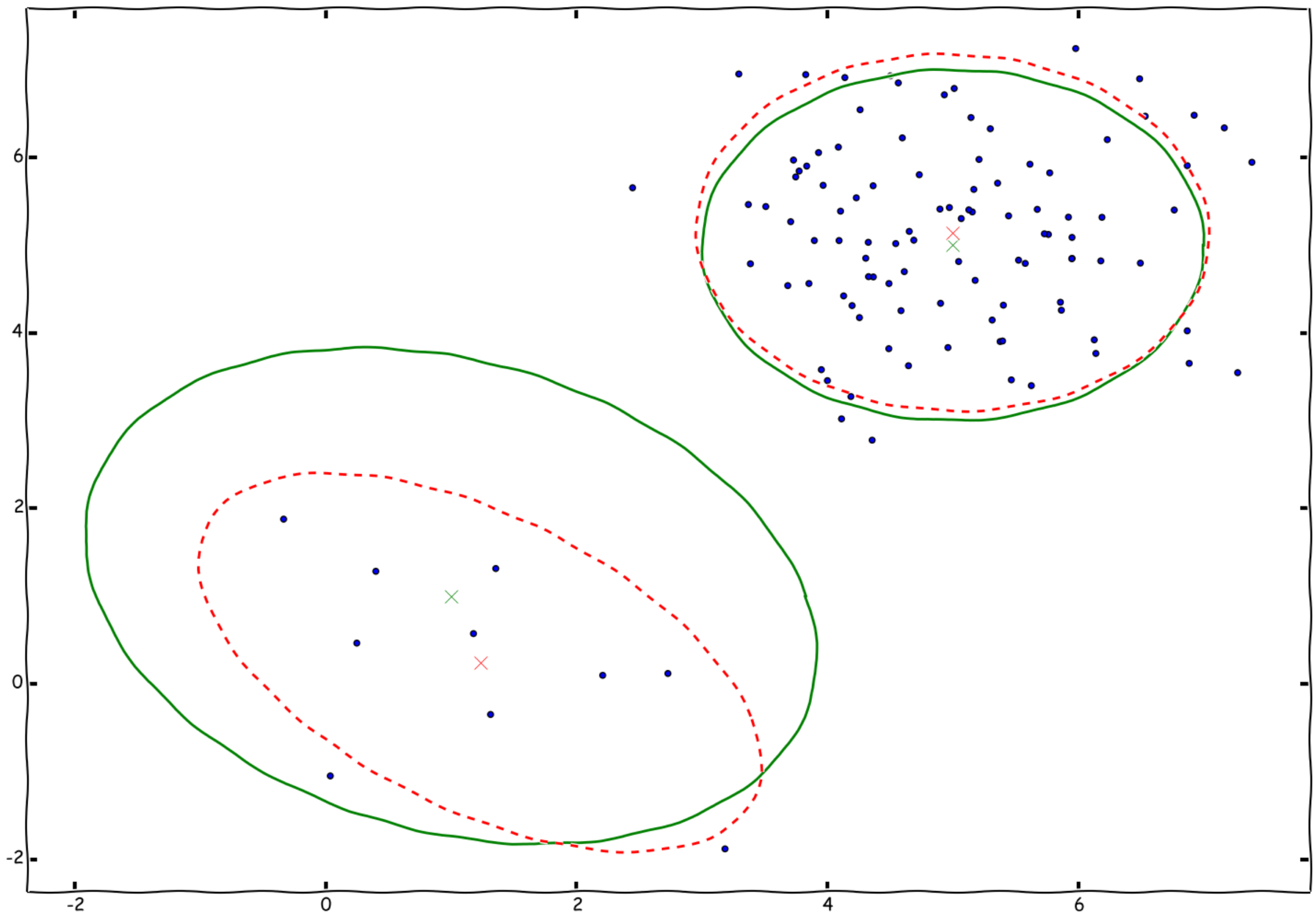
\- Katya Casio. Google Product Forums.

Error vs sample size



Sample Size Disparity:

In a heterogeneous population, smaller groups face larger error



-2
0
2
4
6
-2
0
2
4
6

Credit Application

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)

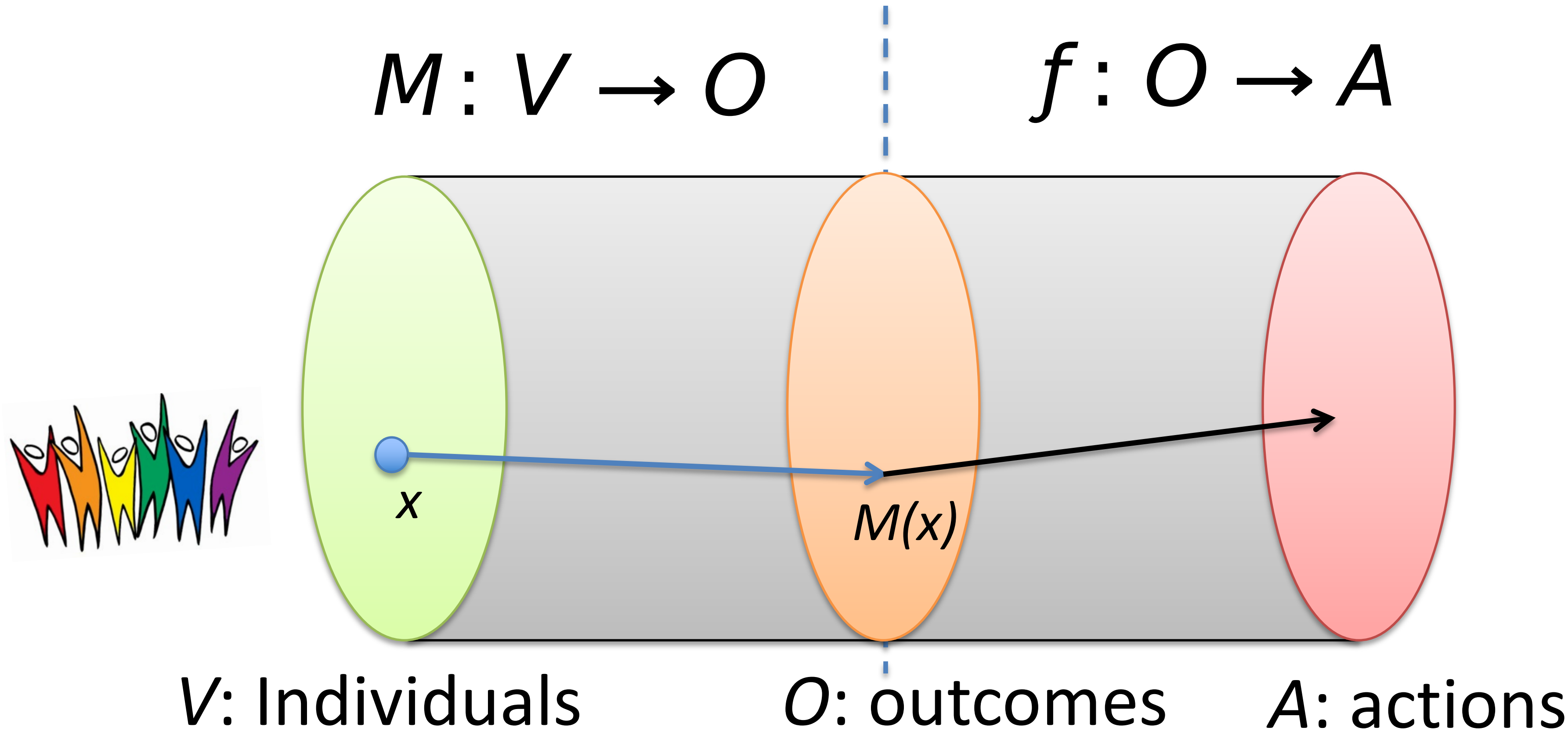
WSJ 2010

Classifier
(eg. ad network)

Vendor
(eg. capital one)

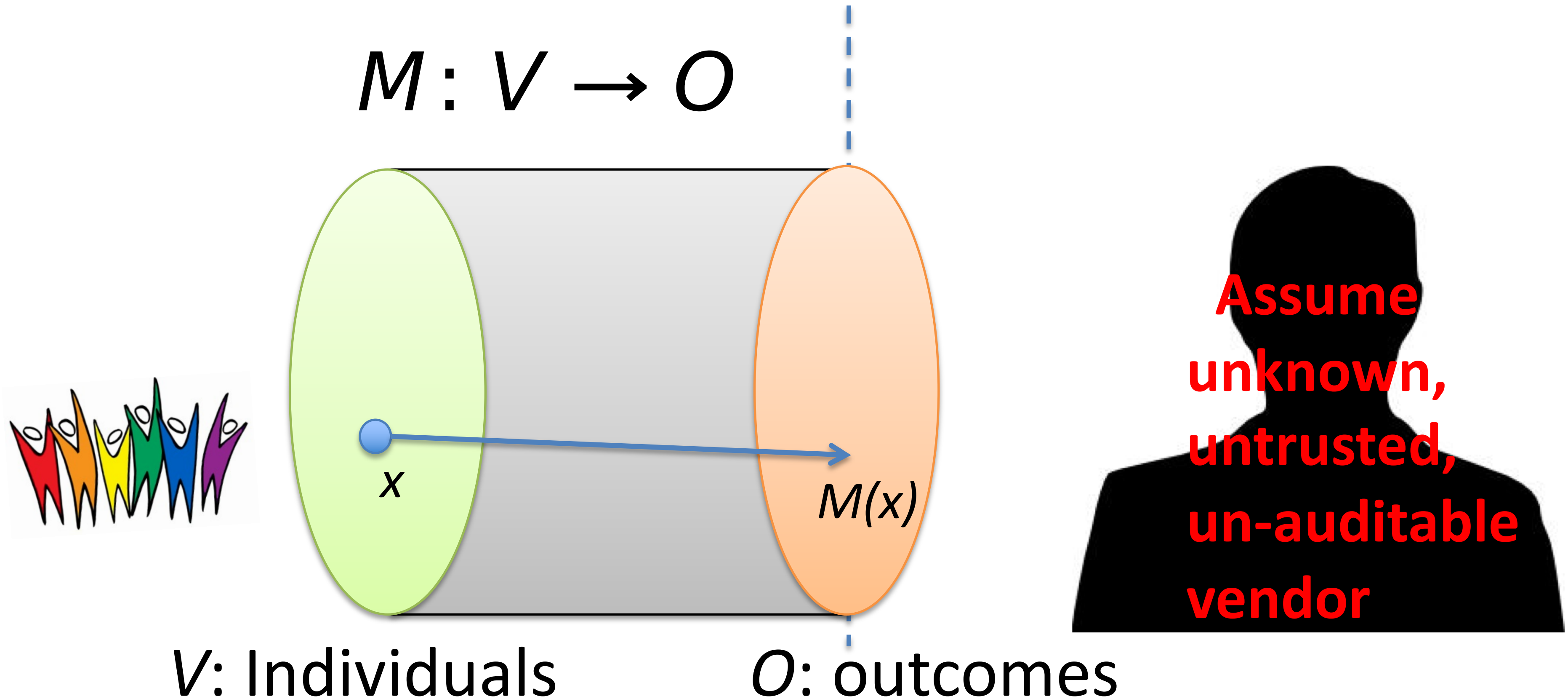
$M : V \rightarrow O$

$f : O \rightarrow A$



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 = 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."

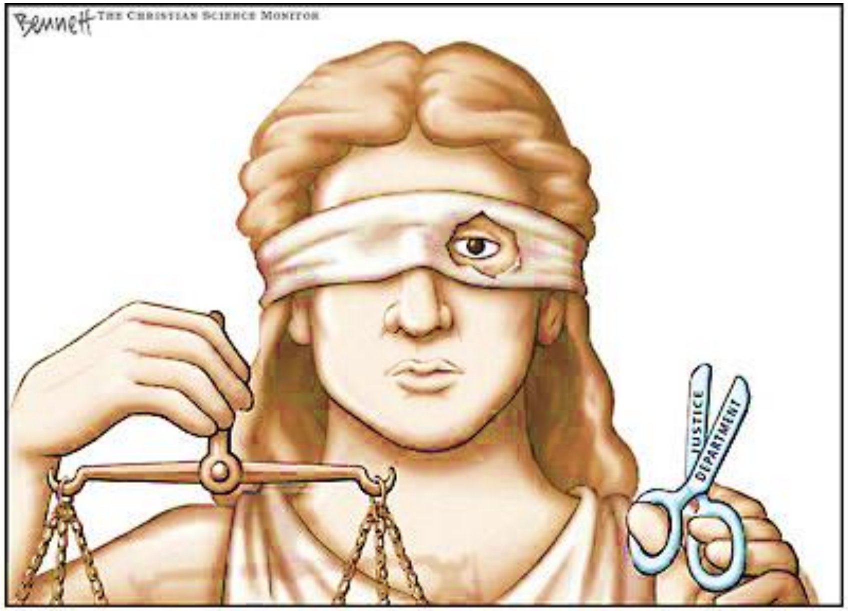
- **Not strong enough** as a notion of fairness
  - Sometimes desirable, but can be abused

- **Self-fulfilling prophecy**
  - Give credit offers to S persons deemed least credit-worth.
  - Give credit offers to those in S who are not interested in credit.

Lesson: Fairness is *task-specific*

Fairness requires understanding of classification task and protected groups

“Awareness”



- **Statistical property vs. individual guarantee**
  - Statistical outcomes may be ”fair”, but individuals might still be discriminated against

Individual Fairness Approach

Fairness Through Awareness. Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, Richard Zemel. 2011

Individual Fairness

Treat *similar* individuals *similarly*

Similar for the purpose of the classification task

Similar distribution over outcomes

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

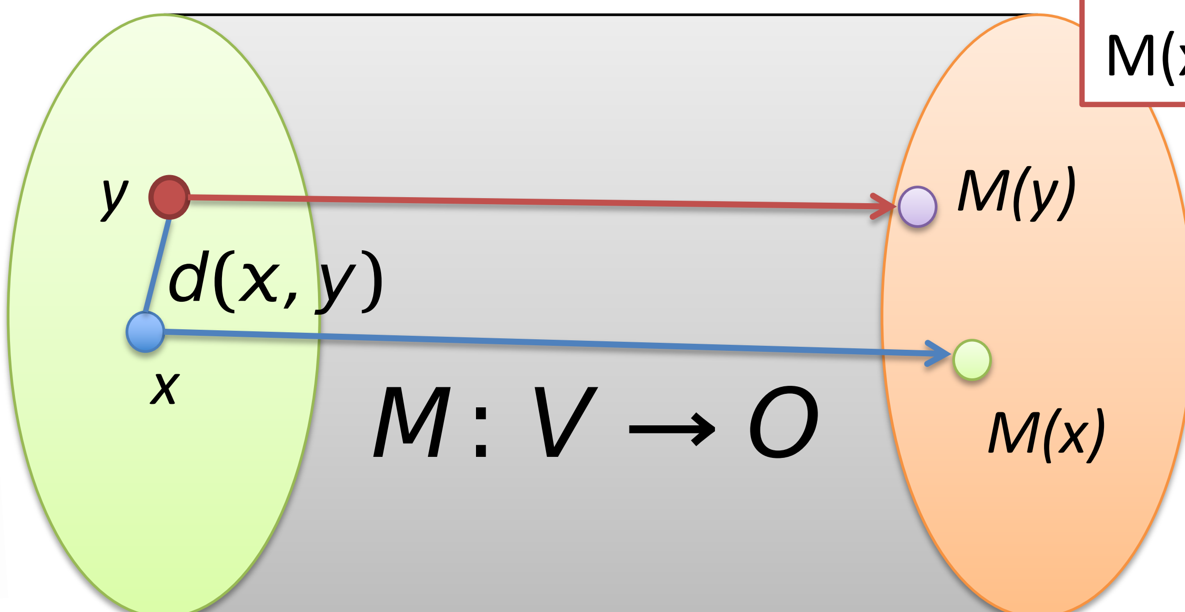
- Financial/insurance risk metrics
  - Already widely used (though secret)
- **AALIM health care metric**
  - health metric for treating similar patients similarly
- Roemer's relative effort metric
  - Well-known approach in Economics/Political theory

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)?



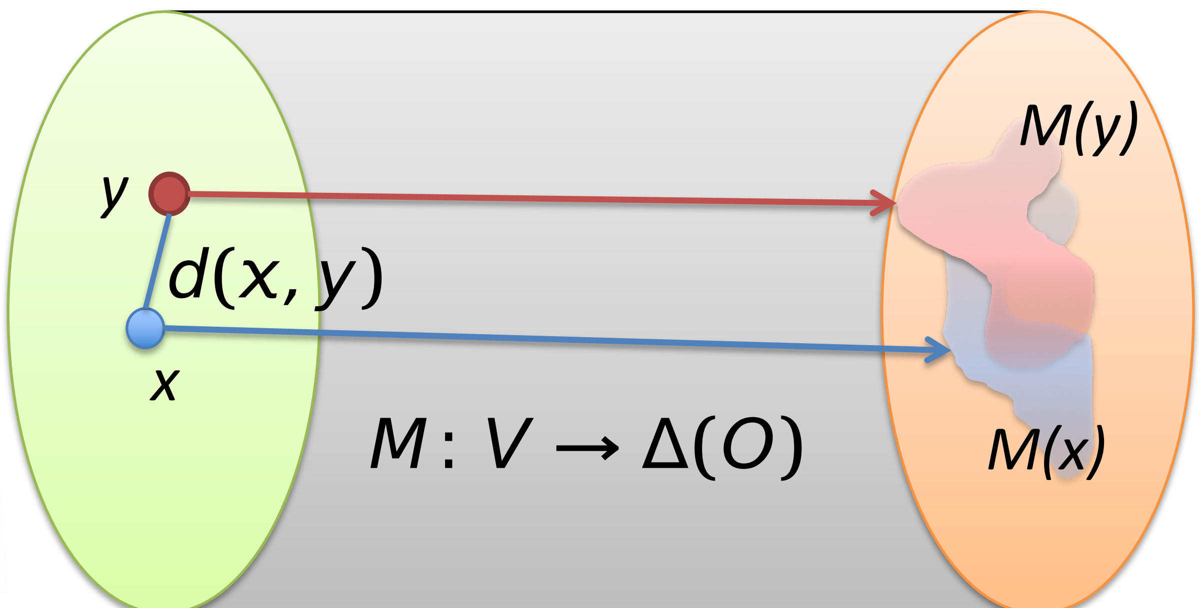
V: Individuals

O: outcomes

Distributional outcomes

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

Statistical distance!



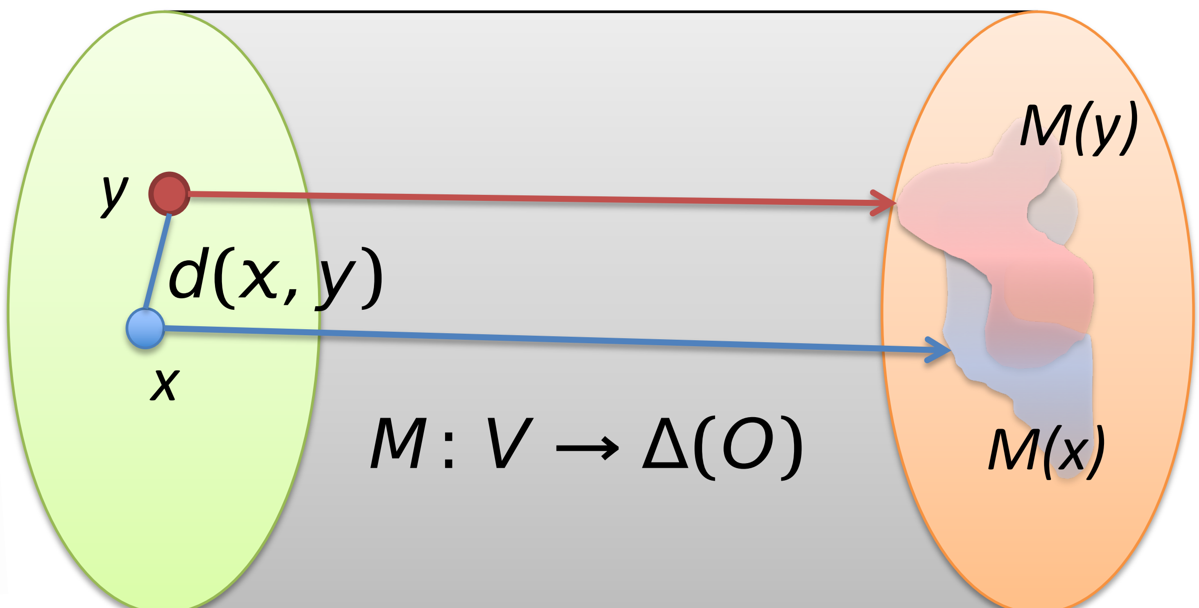
V: Individuals

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

Statistical Distance

P, Q denote probability measures on a finite domain A. The *statistical distance* between P and Q is denoted by

total variation norm / distance $$D_{\mathrm{tv}}(P, Q) = \frac{1}{2} \sum_{a \in A} |P(a) - Q(a)| \,.$$

Notation match:
M(x) = P
M(y) = Q
O = A

Statistical Distance

P, Q denote probability measures on a finite domain A. The *statistical distance* between P and Q is denoted by

$$D_{\text{tv}}(P, Q) = \frac{1}{2} \sum_{a \in A} |P(a) - Q(a)| .$$

Example: High D
A= {0,1}
P(0) = 1, P(1) = 0
Q(0) = 0, Q(1) = 1
D(P, Q) = 1

Statistical Distance

P, Q denote probability measures on a finite domain A. The *statistical distance* between P and Q is denoted by

$$D_{tv}(P, Q) = \frac{1}{2} \sum_{a \in A} |P(a) - Q(a)| .$$

Example: Low D
A= {0,1}
P(0) = 1, P(1) = 0
Q(0) = 1, Q(1) = 0
D(P, Q) = 0

Statistical Distance

P, Q denote probability measures on a finite domain A. The *statistical distance* between P and Q is denoted by

$$D_{tv}(P, Q) = \frac{1}{2} \sum_{a \in A} |P(a) - Q(a)| .$$

Example: Mid D
A= {0,1}
P(0) = P(1) = ½
Q(0) = ¾, Q(1) = ¼
D(P, Q) = ¼

Existence Proof

There exists a classifier that satisfies the Lipschitz condition

- Idea: Map all individuals to the same distribution over outcomes

- Are we done?

Key elements of approach…

Utility Maximization

Vendor can specify **arbitrary utility function**

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

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

Maximize vendor's expected utility subject to Lipschitz condition

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

s.t. M is d-Lipschitz

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

Linear Program Formulation

- Objective function is linear
  - U(x,o) is constant for fixed x, o
  - Distribution over V is known
  - Pr[M(x)=o] (for x in V, o in O) are only variables to be computed
- Lipschitz condition is linear when using statistical distance
  - Linear in number of instances times outcomes
- Linear program can be solved efficiently

Discrimination Harms

Information use

- Explicit discrimination
  - Explicit use of race/gender for employment
- Redundant encoding/proxy attributes

Practices

- Redlining
- Self-fulfilling prophesy
- Reverse tokenism

When does Individual Fairness imply Group Fairness?

Suppose we enforce a metric *d.*

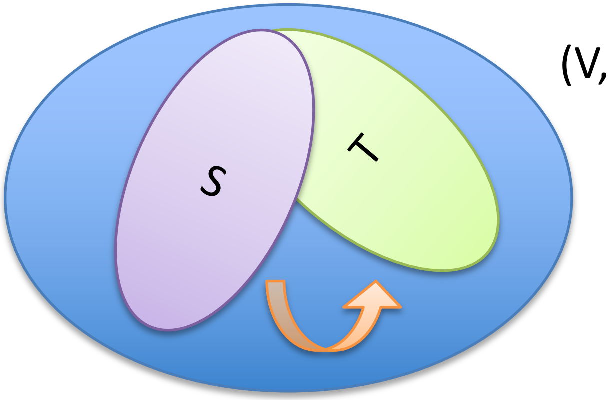
Question: Which *groups of individuals* receive (approximately) equal outcomes?

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 one distribution to another, by "moving" probability mass ("earth").



(V,d)

$$d_{EM}(S,T) \stackrel{\text{def}}{=} \min \sum_{x,y \in V} h(x,y)\, d(x,y)$$

$$\text{subject to} \quad \sum_{y \in V} h(x,y) = S(x)$$

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

$$h(x,y) \geq 0$$

h(x,y) – how much probability of x in S to move to y in T

$$d_{EM}(S,T) \stackrel{\text{def}}{=} \min \sum_{x,y \in V} h(x,y)\, d(x,y)$$

$$\text{subject to} \quad \sum_{y \in V} h(x,y) = S(x)$$

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

$$h(x,y) \geq 0$$

bias(d,S,T) = largest violation of statistical parity* between S and T that any d-Lipschitz mapping can create

Theorem:
bias(d,S,T) ≤ d_{EM}(S,T)

bias = $\max_M$ Pr[M(x)=*o* | x in *S*] - Pr[M(x)=o | x in *T*]
Max over all d-Lipschitz satisfying models

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

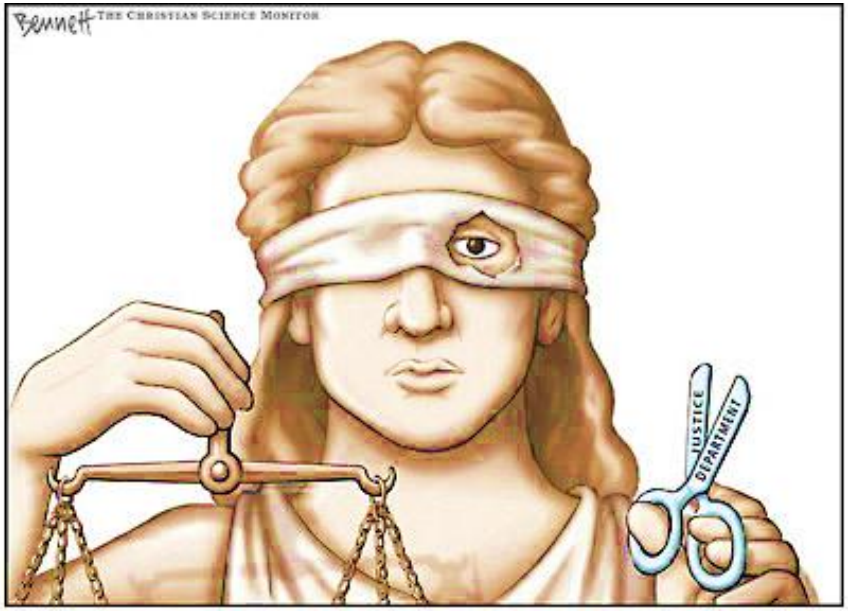
	Differential Privacy	Individual Fairness
Objects	Databases	Individuals
Outcomes	Output of statistical analysis	Classification outcome
Similarity	General purpose metric	Task-specific metric

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

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


Bennett THE CHRISTIAN SCIENCE MONITOR
JUSTICE
DEPARTMENT

Questions?

Metric

A metric on a set X is a function $d : X \times X \rightarrow R+$ (where R+ is the set of non-negative real numbers). For all x, y, z in X, this function is required to satisfy the following conditions:

- $d(x, y) \geq 0$ (non-negativity)
- $d(x, y) = 0$ if and only if $x = y$ (identity of indiscernibles. Note that condition 1 and 2 together produce positive definiteness)
- $d(x, y) = d(y, x)$ (symmetry)
- $d(x, z) \leq d(x, y) + d(y, z)$ (subadditivity / triangle inequality).

Another Statistical Distance

$$D_\infty(P, Q) = \sup_{a \in A} \log\left(\max\left\{\frac{P(a)}{Q(a)}, \frac{Q(a)}{P(a)}\right\}\right)$$

Partial Proof Idea

Theorem:
bias(d,S,T) <= d_{EM}(S,T)

- d_{EM}(S,T) cost of best coupling between the two distributions subject to the penalty function d(x,y) = E d(x,y)

Proof Sketch: LP Duality

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

max Pr(M(x) = 0 | x in S) – Pr(M(x) = 0 | x 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!

Fair Affirmative Action (1)

1. (a) First we compute a mapping from elements in S to distributions over T which transports the uniform distribution over S to the uniform distribution over T, while minimizing the total distance traveled. Additionally the mapping preserves the Lipschitz condition between elements within S.

   (b) This mapping gives us the following new loss function for elements of T: For $y \in T$ and $a \in A$ we define a new loss, $L'(y, a)$, as

   $$L'(y,a) = \sum_{x \in S} \mu_x(y) L(x,a) + L(y,a),$$

   where $\{\mu_x\}_{x \in S}$ denotes the mapping computed in step (a). L' can be viewed as a reweighting of the loss function L, taking into account the loss on S (indirectly through its mapping to T).

2. Run the Fairness LP only on T, using the new loss function L'.

Fair Affirmative Action (2)

Formally, we can express the first step of this alternative approach as a restricted Earthmover problem defined as

$$
\begin{aligned}
d_{\mathrm{EM+L}}(S,T) \stackrel{\text{def}}{=} \quad \min \quad & \mathop{\mathbb{E}}_{x\in S} \mathop{\mathbb{E}}_{y\sim\mu_x} d(x,y) \qquad (15)\\
\text{subject to} \quad & D(\mu_x,\mu_{x'}) \le d(x,x') \quad \text{for all} \quad x,x' \in S\\
& D_{\mathrm{tv}}(\mu_S, U_T) \le \epsilon\\
& \mu_x \in \Delta(T) \quad \text{for all} \quad x \in S
\end{aligned}
$$

Here, U_T denotes the uniform distribution over T. Given $\{\mu_x\}_{x\in S}$ which minimizes (15) and $\{\nu_x\}_{x\in T}$ which minimizes the original fairness LP (2) restricted to T, we define the mapping $M\colon V \to \Delta(A)$ by putting

$$
M(x) = \begin{cases} \nu_x & x \in T \\ \mathbb{E}_{y\sim\mu_x} \nu_y & x \in S \end{cases} . \qquad (16)
$$

Fair Affirmative Action (3)

Proposition 4.1. *The mapping M defined in (16) satisfies*

1. *statistical parity between S and T up to bias ϵ,*
2. *the Lipschitz condition for every pair $(x, y) \in (S \times S) \cup (T \times T)$.*

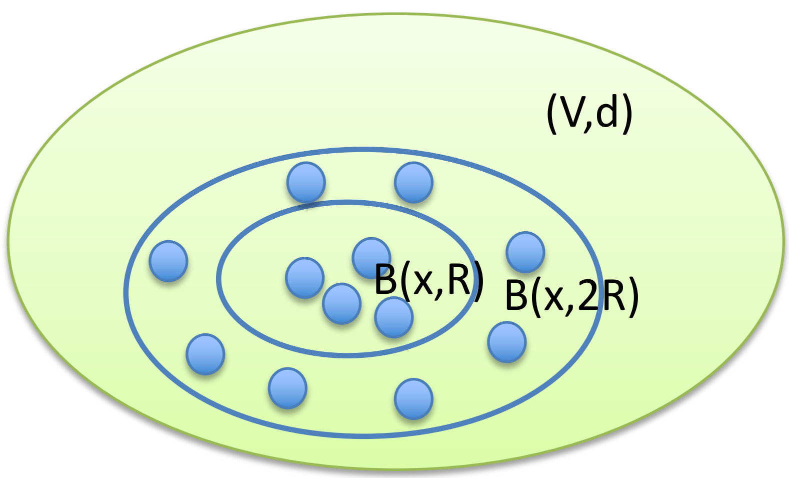
Proposition 4.2. *Suppose $D = D_{\mathrm{tv}}$ in (15). Then, the resulting mapping M satisfies*

$$\mathop{\mathbb{E}}_{x \in S} \max_{y \in T} \left| D_{\mathrm{tv}}(M(x), M(y)) - d(x, y) \right| \leq d_{\mathrm{EM+L}}(S, T).$$

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]

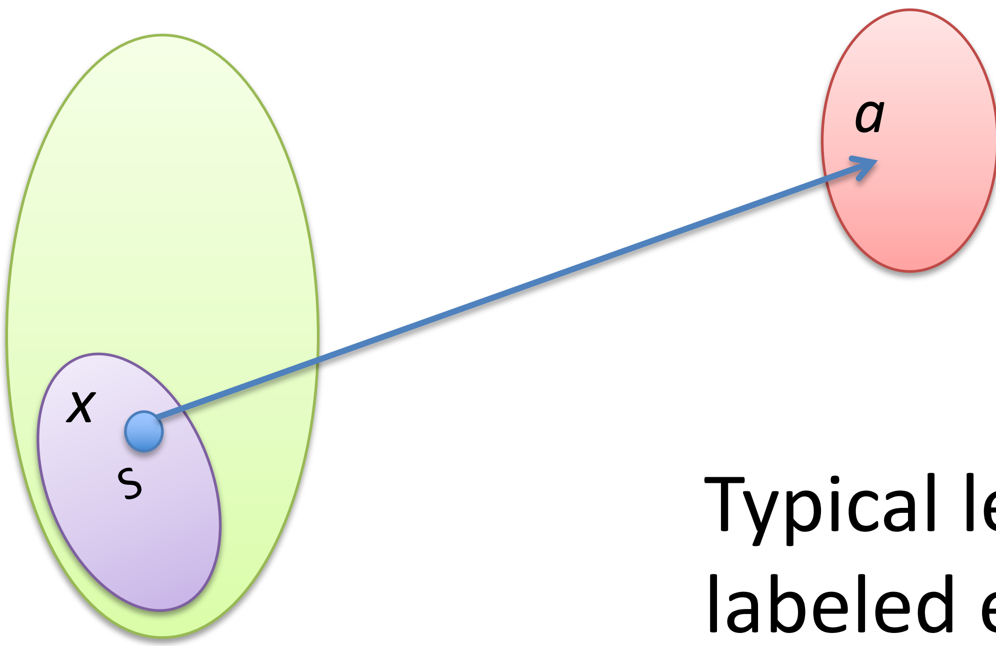


$|B(x,R)| \leq O(|B(x,2R))$

Some recent work

- Zemel-Wu-Swersky-Pitassi-Dwork

  "Learning Fair Representations" (ICML 2013)



A: labels

Typical learning task
labeled examples (x,a)

V: Individuals
S: protected set

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

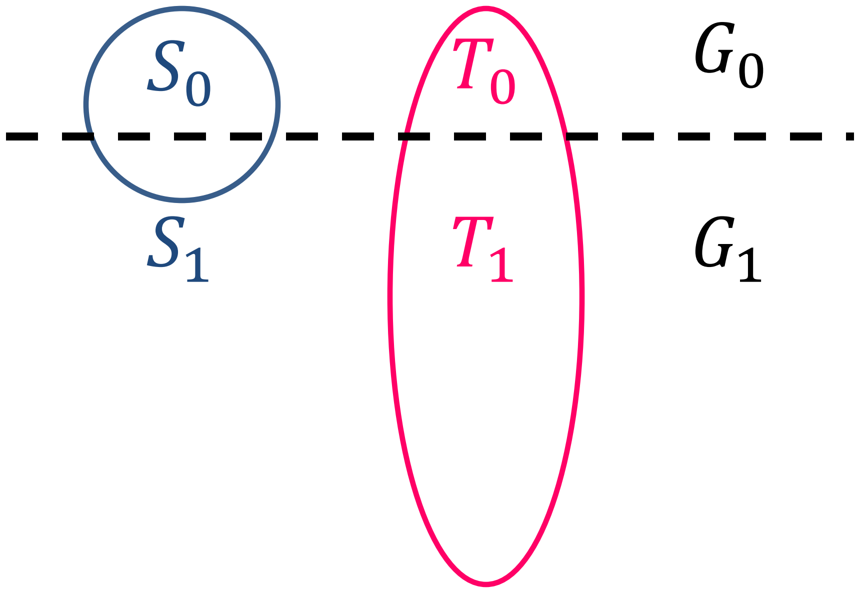
The Story So Far…

- Group fairness
- Individual fairness
- Group fairness does not imply individual fairness
- Individual fairness implies group fairness <u>if</u> earthmover distance small

- What if earthmover distance large?

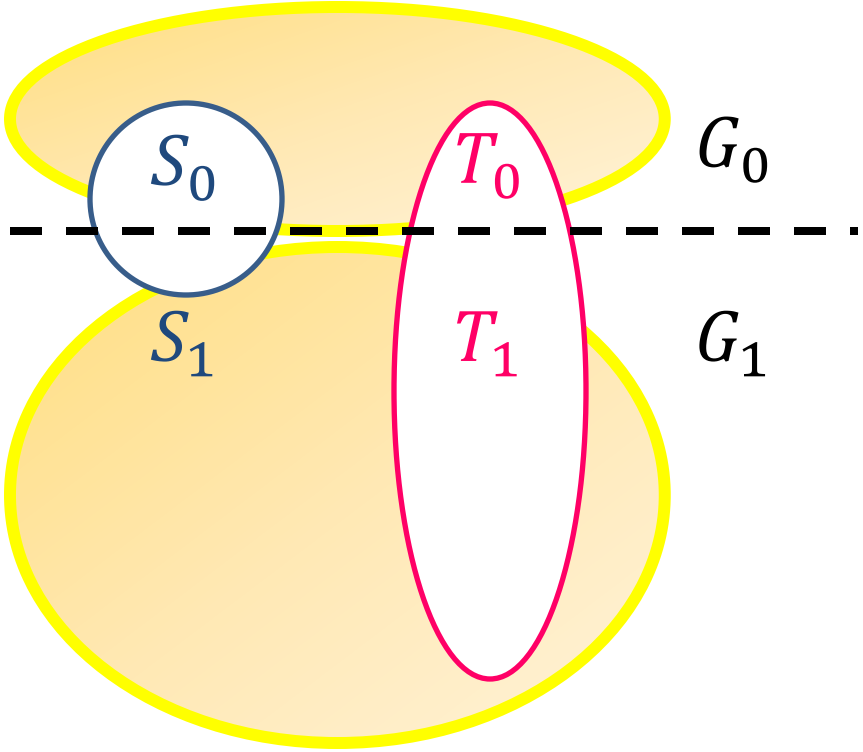
Toward Fair Affirmative Action: When EM(S,T) is Large

- G_0 is unqualified
- G_1 is qualified



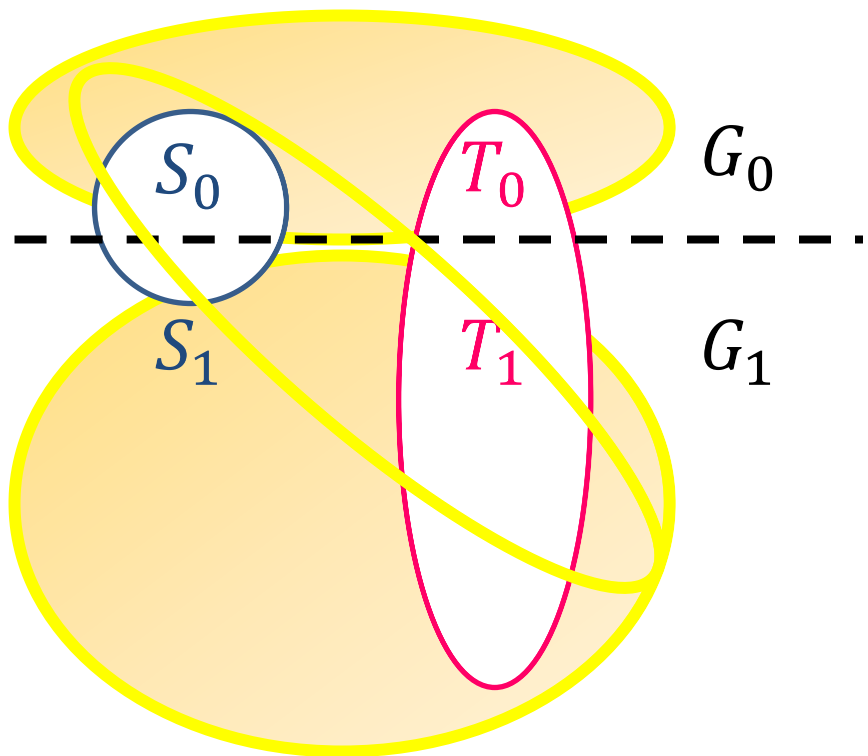
Toward Fair AA: When EM(S,T) is Large

- Lipschitz ⇒
  All in G_i treated same



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

- Lipschitz ⇒
  All in G_i treated same
- Statistical Parity ⇒
  much of S_0 must be treated the same as much of T_1

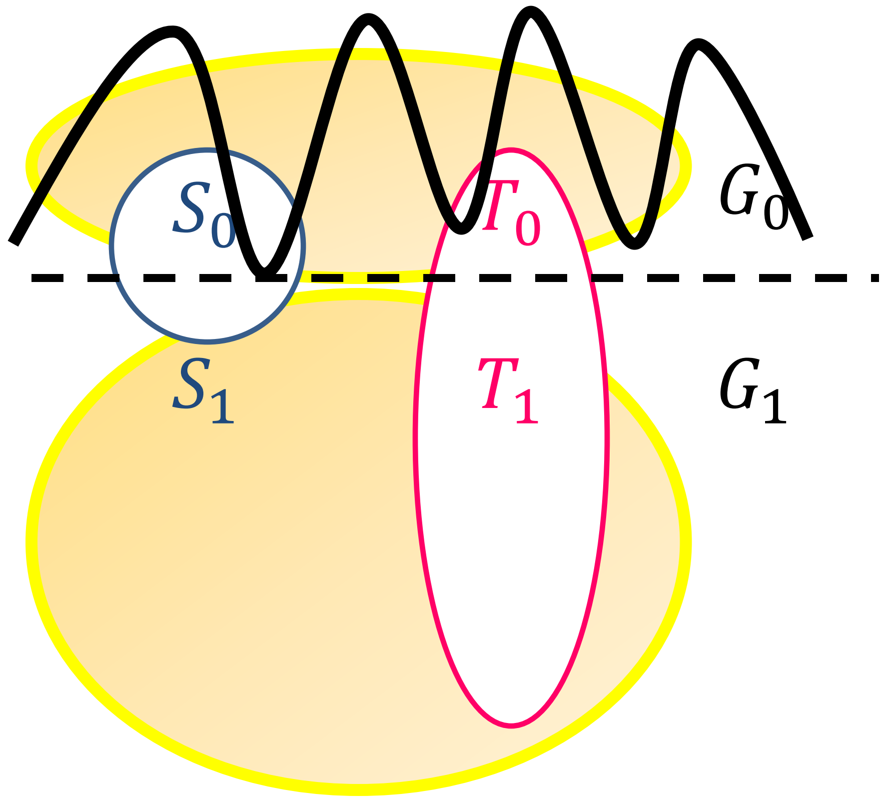


Toward Fair AA: When EM(S,T) is Large

- Lipschitz ⇒
  All in G_i treated same

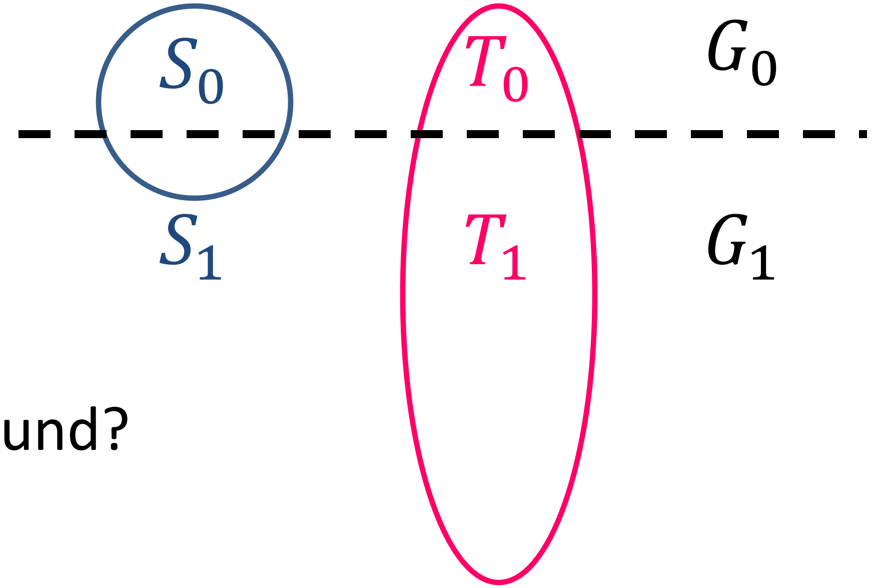
Failure to Impose Parity ⇒
anti-S vendor can target G_0
with blatant hostile ad f_u.

Drives away almost all of S
while keeping most of T.



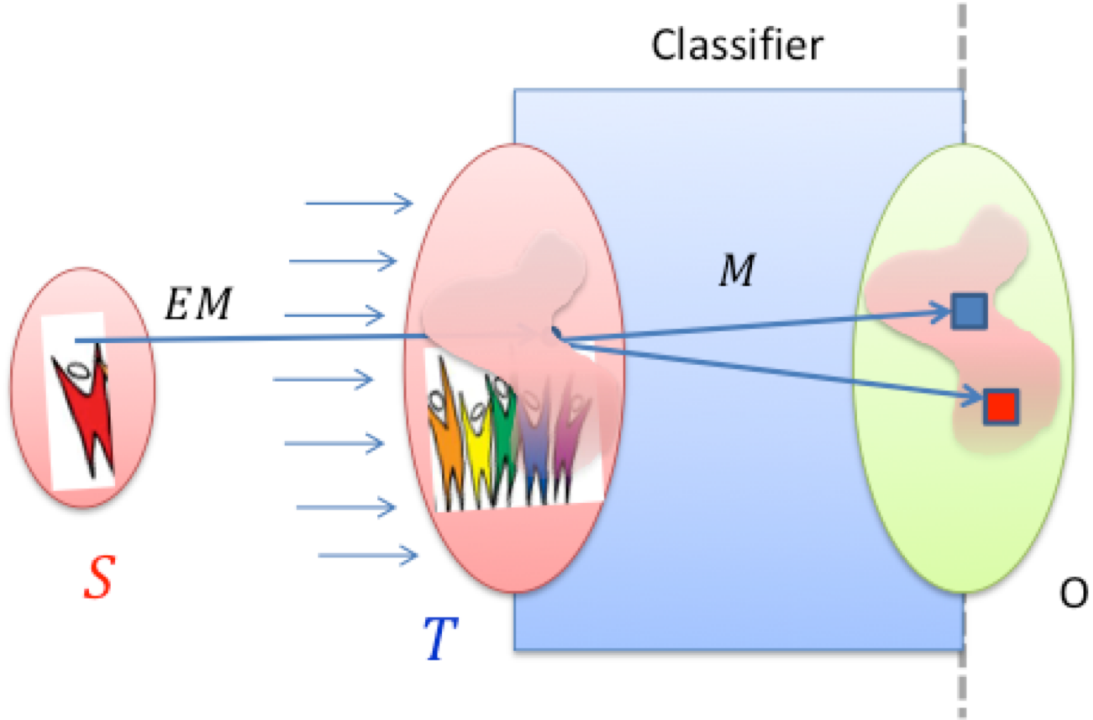
Dilemma: What to Do When EM(S,T) is Large?

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



How can we form a middle ground?

Fair Affirmative Action

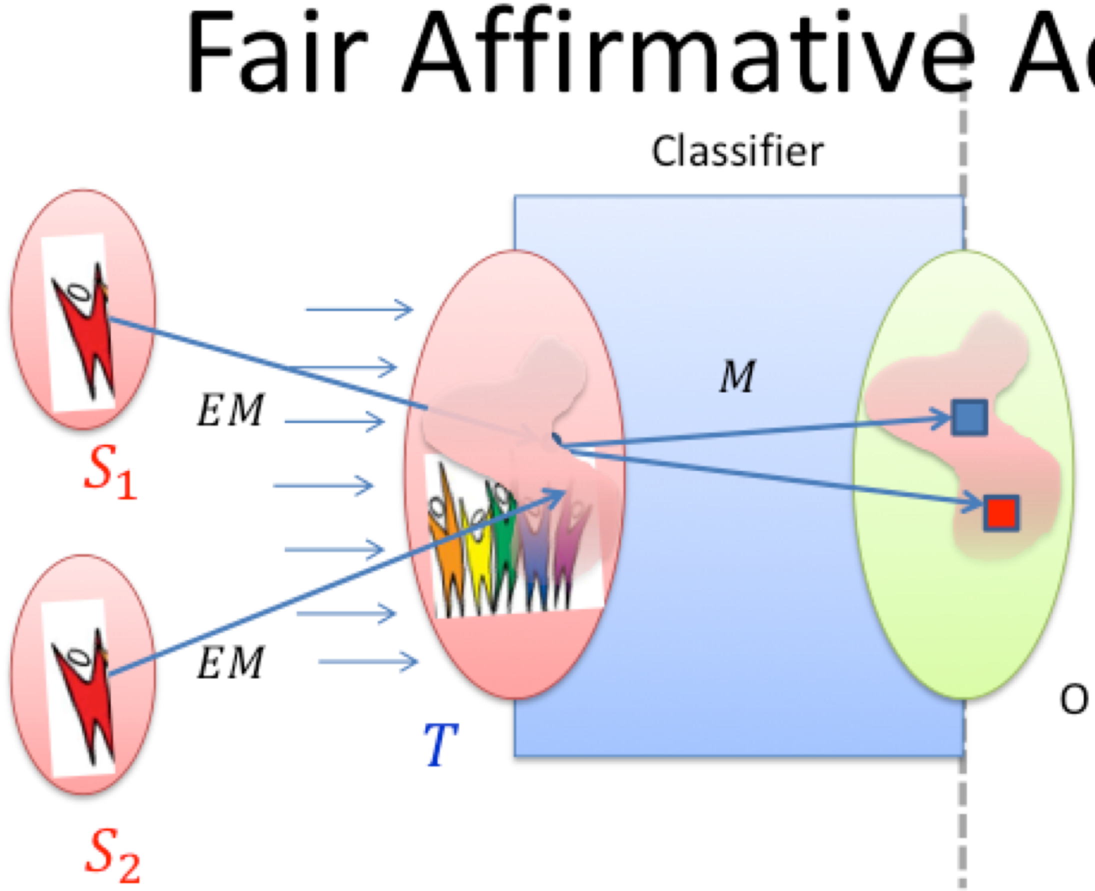


Earthmover mapping from S to T + Lipschitz mapping from T to O

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